Annals of Agri-bio Research
https://agribiop.com/

Article

AquaGPT: A Multimodal Transformer Framework with Expert Knowledge Embedding for Intelligent Feeding Management in Aquaponics

Dianyou Kang [1,2], Yi Yang [1,2], Oliver Lexter July A. Jose [1], Ralph Gerard B. Sangalang [1], Antonette V. Chua [1] and Anton Louise P. De Ocampo [1,*]

[1] Department of Electronics Engineering, Batangas State University, Batangas City 4200, Philippines
[2] Information Technology Department, Jiangmen Technician College, Jiangmen 529090, China
* Correspondence: antonlouise.deocampo@ieee.org

How To Cite: Kang, D., Yang, Y., Jose, O. L. J. A., Sangalang, R. G. B., Chua, A. V., & De Ocampo, A. L. P. (2026). AquaGPT: A Multimodal Transformer Framework with Expert Knowledge Embedding for Intelligent Feeding Management in Aquaponics. *Annals of Agri-bio Research*, *31*(1), 40–56. https://doi.org/10.53941/agrbio.2026.100004

Received: 5 February 2026
Revised: 27 March 2026
Accepted: 13 April 2026
Published: 17 April 2026

Abstract: Aquaponics integrates aquaculture with hydroponic crop cultivation, which offers a sustainable approach to food production. However, feeding management remains a critical bottleneck because it directly affects feed efficiency, water quality stability, and ecosystem balance. This paper proposes AquaGPT, a multimodal Transformer framework that embeds expert knowledge to address challenges in feeding management in aquaponic systems. The system employs a distributed sensing network, processed by modality-specific encoders and projected into a shared embedding space, to capture synchronized acoustic, sensor, environmental, and visual data. A differentiable expert rule layer embedding seven aquaculture feeding management rules then aligns predictions with established aquaculture practice to improve interpretability. A dynamic weight allocation strategy further enhances robustness by prioritizing reliable modalities under noisy or incomplete input conditions. Experiments on the Full Fish Interaction Analysis (FFIA) dataset demonstrate that AquaGPT outperforms state-of-the-art multimodal fusion baselines by up to 4.5% accuracy under severe noise. Also, it achieves a 30% reduction in parameters compared to similar models, enabling real-time deployment on edge devices. These results highlight AquaGPT's potential for precision aquaponics and demonstrates capability to optimize feeding strategies, which theoretically improves resource efficiency and mitigates the environmental footprint.

Keywords: aquaponics; intelligent feeding management; multimodal fusion; spatiotemporal transformer; expert knowledge embedding; precision aquaculture

1. Introduction

Global population growth and the rising demand for high-quality protein are driving aquaculture to become a key sector for ensuring food security and promoting sustainable development (Goddek et al., 2019). In this context, aquaponic systems—integrating aquaculture with hydroponic plant cultivation—are gaining significant attention for their high resource-use efficiency, environmentally friendly operation, and potential for zero-waste discharge (FAO, 2024). However, feeding management remains a critical bottleneck in such systems, directly influencing production efficiency, water quality stability, and overall ecosystem balance

(Badiola et al., 2012). Feed costs typically account for 50–70% of total operational expenditure in aquaculture, while conventional feeding approaches can result in 20–30% feed loss, leading to water quality deterioration, fish stress, and ecological imbalance (Martínez-Porchas & Martínez-Córdova, 2012).

Current feeding practices are dominated by manual feeding and simple mechanized feeding devices. Manual feeding relies heavily on operator experience, is labor-intensive, and often results in inconsistent outcomes. Traditional automated feeders can reduce labor requirements but generally follow fixed schedules or basic threshold rules, unable to adapt feeding behavior to real-time changes in fish activity, water quality, and environmental conditions (Føre et al., 2018). Commercial systems utilizing visual or acoustic sensors to monitor feeding behavior have shown promise; however, their performance is susceptible to environmental noise, light variation, and water turbidity (Li et al., 2024). Moreover, existing systems face three persistent challenges: (1) inefficient fusion of heterogeneous multimodal data, (2) insufficient correlation between water quality indicators and behavioral features, and (3) slow decision-making, which can cause delayed responses to sudden environmental changes (Sun et al., 2020).

Recent advances in deep learning (DL) and machine learning (ML) have driven significant progress in related domains such as food quality assessment (Zhang et al., 2024), meat freshness detection (Singh et al., 2025), smart agriculture (Mahmood et al., 2023), and intelligent aquaculture monitoring (Li et al., 2025). These studies demonstrate that multimodal data fusion, combined with spatiotemporal modeling, can significantly enhance system perception and decision-making in dynamic environments. For example, (Zhang et al., 2024) proposed a Deep Convolutional Recurrent Forest (DCRF) that integrates the spatial feature extraction capability of convolutional neural networks (CNNs) with the temporal modeling strength of recurrent neural networks (RNNs), further enhanced by ensemble learning. The DCRF demonstrated superior performance in predicting meat freshness, highlighting the potential of combining heterogeneous modalities and temporal dynamics for robust decision-making. In aquaculture, this approach is particularly relevant, as fish feeding behavior is influenced by complex, nonlinear interactions among visual cues, acoustic signals, water quality parameters, and environmental conditions (Li et al., 2025).

From a sensing perspective, feeding management in aquaponics is a typical dynamic decision-making task driven by multi-dimensional environmental and biological signals, where fish feeding behavior and the aquatic ecosystem are mutually constrained—this intrinsic complexity makes unimodal sensing methods inherently incapable of addressing the core challenges of the scenario (Wu et al., 2025). Unimodal methods rely on a single type of signal to infer feeding states (Cepeda-Humerez et al., 2019), but they suffer from scenario-dependent failure and information incompleteness in actual aquaponic farms: visual-only monitoring fails to capture fish behavior in turbid water (a common condition caused by fish excrement and unconsumed feed); acoustic-only detection cannot distinguish feeding sounds from environmental noise (e.g., water pump vibration); sensor-only data (water quality/environmental) can only reflect the static state of the aquatic environment but cannot capture the real-time feeding intensity of fish. None of the unimodal methods can simultaneously characterize the three core dimensions of feeding decision-making: real-time fish behavioral state, aquatic environmental constraint, and external environmental influence—this is the fundamental reason why unimodal approaches lead to suboptimal feeding decisions (e.g., overfeeding due to ignoring low dissolved oxygen, underfeeding due to visual failure in turbid water).

To address the limitations of unimodal sensing, multimodal data fusion is an inevitable choice for intelligent feeding management, and the four modalities (acoustic, visual, water-quality sensors, environmental sensors) adopted in this work have irreplaceable functional roles and strong complementary relationships in the feeding decision process. Taking a typical aquaponic farm scenario with diurnal light variation and periodic water turbidity as an example:

(1) Visual modality (1920 × 1080 underwater cameras) directly captures spatial features of fish behavior (e.g., aggregation density at the feeding point, movement trajectory, feeding intensity) under clear water and sufficient light conditions, providing the most intuitive signal for judging fish hunger state;
(2) Acoustic modality (44.1 kHz hydrophones) captures frequency-domain features of feeding-related sounds (e.g., fish chewing feed, water disturbance during foraging) and compensates for the visual modality's failure in turbid water or low-light environments, as acoustic signals are less affected by water clarity and light;
(3) Water-quality sensors (pH, dissolved oxygen, water temperature, 1 Hz sampling) quantify the core abiotic constraints of fish feeding—dissolved oxygen and water temperature directly determine the fish's metabolic rate and feeding capacity, while pH reflects the stability of the aquatic ecosystem, and these parameters set the upper limit of safe feeding amount (e.g., feeding must be reduced or stopped when dissolved oxygen < 3 mg/L);

(4) Environmental sensors (light intensity, air temperature, humidity) capture external environmental factors that regulate fish circadian rhythm and feeding behavior (e.g., low light at night reduces fish feeding activity), and their data align with the circadian rhythm rule in the expert knowledge layer to optimize feeding timing and amount.

The complementarity among the four modalities forms a closed-loop information acquisition system for feeding management: visual and acoustic modalities jointly characterize the real-time biological state (fish feeding behavior) with spatiotemporal complementarity; water-quality and environmental modalities characterize the aquatic and external environmental constraints with quantitative and dynamic features; the fusion of these heterogeneous multimodal signals breaks the information bottleneck of unimodal methods and enables the model to capture the complex nonlinear interactions between fish behavior and the environment. More importantly, the multimodal information is directly linked to the core objectives of feeding decision-making in aquaponics: (1) fusing visual-acoustic features to accurately classify three feeding states (Underfed, Optimal Feeding, Overfed) and avoid misjudgment caused by single-modality noise; (2) fusing water-quality features with behavioral features to determine the safe and optimal feeding amount (e.g., reducing feeding when high fish aggregation (visual-acoustic) coexists with low dissolved oxygen (water quality)); (3) fusing environmental features with behavioral features to adapt feeding strategies to dynamic external conditions (e.g., reducing feeding at night when low light (environmental) coexists with weak feeding sounds (acoustic)); (4) the full multimodal fusion output is aligned with the expert rule layer to further improve decision reliability and interpretability (Wang et al., 2026). This logical chain—multimodal complementary sensing to comprehensive characterization of fish-environment interaction to targeted support for feeding decision objectives—establishes the fundamental theoretical justification for adopting a multimodal framework in aquaponic feeding management, and the framework is not a simple assembly of multimodal elements but a scenario-driven design for the core challenges of the task.

This underscores the need for an intelligent, real-time system that efficiently integrates heterogeneous multimodal signals while incorporating domain expert knowledge. To address these challenges, this study proposes AquaGPT, an intelligent feeding management framework designed under a hardware–algorithm co-design paradigm for high-efficiency multimodal data fusion and adaptive decision-making. At the hardware level, AquaGPT utilizes an STM32-based multi-sensor network that integrates water quality sensors (pH, dissolved oxygen, and temperature), environmental sensors (light intensity, temperature, and humidity), and behavioral monitoring devices (high-definition underwater cameras and acoustic detectors) to achieve second-level data acquisition and secure transmission. At the algorithmic level, an enhanced Vision Transformer architecture with spatiotemporal attention is used to jointly model fish behavior patterns and environmental changes. Fifteen aquaculture expert rules are embedded into a differentiable logic layer to interact with data-driven features, improving interpretability and decision reliability. Furthermore, a dynamic weight allocation strategy enables the system to adapt within 30 s to sudden environmental changes—such as abrupt drops in water temperature —resulting in a 12-fold improvement in response speed compared to traditional systems.

Figure 1 illustrates the overall system framework of AquaGPT, outlining the end-to-end process from multimodal perception and preprocessing to spatiotemporal modeling and adaptive feeding strategy generation. This system-level design highlights the integration of multimodal perception, deep temporal modeling, and expert rule embedding, laying the foundation for the neural architecture detailed below.

In essence, the primary contributions of our work are outlined as follows:

(1) A hardware–algorithm co-design framework for multimodal intelligent feeding: We propose AquaGPT, an integrated system that combines an STM32-based multi-sensor network (covering water quality, environmental conditions, and fish behavior) with advanced deep learning architectures. This design enables second-level, high-precision data acquisition and secure transmission, addressing the inefficiencies of traditional discrete sensing and data processing in aquaponic systems.

(2) Novel spatiotemporal modeling and expert knowledge fusion mechanism: By developing an enhanced Vision Transformer with spatiotemporal attention, we enable effective integration of heterogeneous multimodal data (visual, acoustic, and environmental parameters) that differ in sampling rate and structure. Embedding 7 domain-expert rules into a differentiable logic layer further enhances decision interpretability and reliability, addressing long-standing issues of insufficient cross-modal correlation and poor interpretability in existing systems.

(3) Rapid adaptive decision-making for dynamic aquaculture environments: A dynamic weight allocation strategy is introduced to enable the system to respond to sudden environmental changes (e.g., abrupt

temperature drops) within 30 s, achieving a 12-fold improvement in response speed compared to traditional systems. While longitudinal ecological indicators such as the Feed Conversion Ratio (FCR) require multi-season validation (Haque & Al Jufaili, 2026), our short-term deployment demonstrates that this rapid adaptation has the potential to theoretically minimize unconsumed feed and mitigate water quality deterioration. This provides a scalable solution for intelligent feeding management in large-scale aquaponic systems.

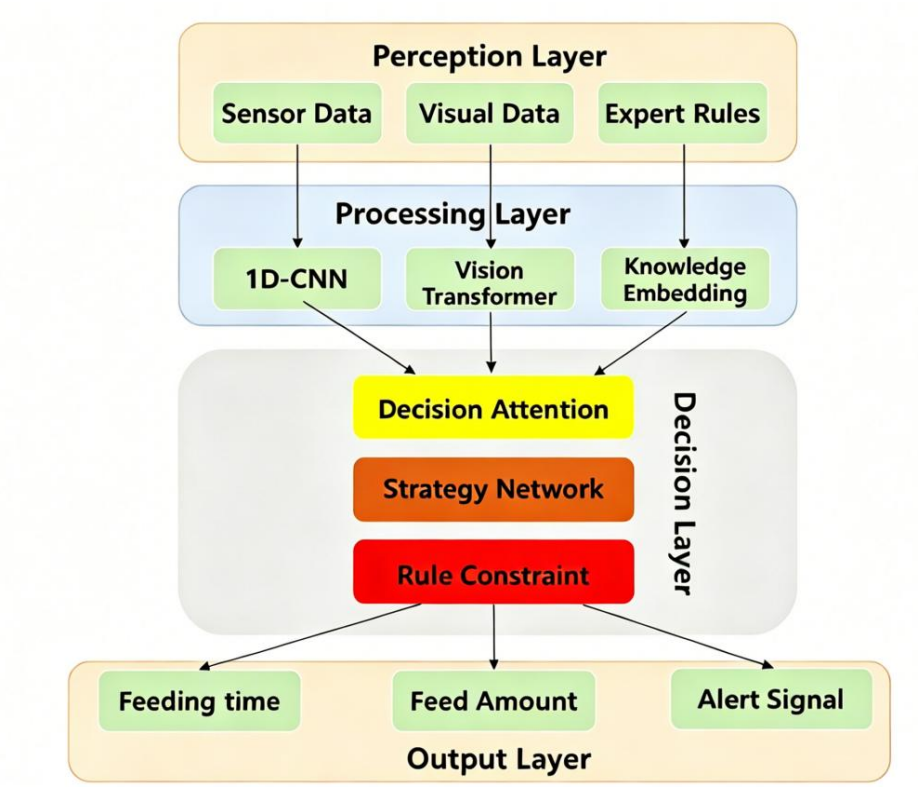

Figure 1. Overall system framework of AquaGPT, illustrating the flow from multimodal sensing and preprocessing to spatiotemporal modeling, expert rule embedding, and adaptive feeding decision-making.

2. Method

2.1. Overview of AquaGPT Framework

AquaGPT is a hardware–algorithm co-design framework for intelligent feeding management in aquaponic systems, designed to address the challenges of heterogeneous multimodal data fusion, spatiotemporal interaction modeling, and adaptive decision-making. The overall workflow follows a three-stage pipeline: sensing, processing, and decision (Figure 1), which integrates data acquisition, feature extraction, spatiotemporal modeling, expert knowledge fusion, and feeding strategy generation into an end-to-end system.

At the sensing layer, a distributed multi-sensor network captures synchronized multimodal data (water quality, environmental conditions, fish behavior) with high temporal precision. The processing layer normalizes, aligns, and extracts modality-specific features to eliminate data heterogeneity. The decision layer, powered by the AquaGPT neural architecture, further fuses these features via spatiotemporal attention, embeds domain expert rules to enhance interpretability, and adopts a dynamic weight allocation strategy to generate adaptive feeding decisions. This hierarchical design ensures efficient data processing, robust feature fusion, and real-time responsiveness to dynamic aquaculture environments (Figure 2).

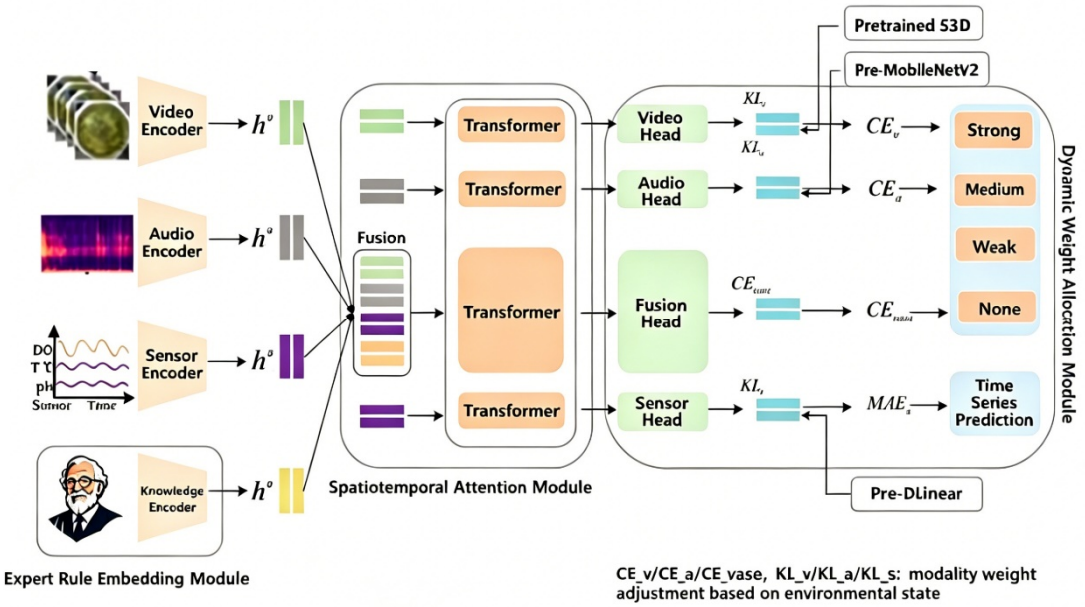

Figure 2. Overall architecture of the proposed AquaGPT model.

2.2. Sensing Layer: Multimodal Data Acquisition

The sensing layer serves as the data foundation of AquaGPT, integrating heterogeneous sensors to capture comprehensive environmental, biological, and behavioral information. Its core design focuses on high-precision sampling, synchronization, and reliable transmission:

Sensor Configuration: Three types of sensors are deployed to cover multi-dimensional data needs:

(a) Water quality sensors: Monitor pH, dissolved oxygen (DO), and water temperature at a 1 Hz sampling rate, providing quantitative indicators of the aquatic environment;

(b) Environmental sensors: Record light intensity, air temperature, and humidity to capture external environmental variations;

(c) Biological behavior sensors: Include high-definition (HD) underwater cameras (1920 × 1080 resolution, 30 fps) for visual observation of fish aggregation and feeding intensity, and hydrophone-based acoustic detectors (44.1 kHz sampling rate) for capturing feeding-related sound signals.

Data Synchronization & Transmission: All sensor nodes are connected to an STM32F407VET6 microcontroller for initial signal conditioning (e.g., noise reduction, signal amplification) and timestamping. A Network Time Protocol (NTP) service is adopted to achieve temporal alignment across modalities with a precision of ±50 ms. Processed data are wirelessly transmitted to the processing unit for subsequent feature extraction (Lopes et al., 2024).

2.3. Processing Layer: Data Preprocessing and Feature Extraction

The processing layer addresses data heterogeneity (differences in sampling rate, structure, and noise) through preprocessing and modality-specific feature extraction, laying the foundation for effective multimodal fusion:

Data Preprocessing

(a) Normalization: Continuous numerical signals (e.g., water quality parameters, environmental data) are normalized to [0,1] to eliminate scale differences;

(b) Temporal Alignment: A sliding window of length T seconds is applied to continuous signals to capture short-term trends, defined as:

$$\mathrm{x}_t^{(m)} = \left[s_{t-T+1}^{(m)}, s_{t-T+2}^{(m)}, \ldots, s_t^{(m)}\right] \quad (1)$$

where:

$x_t^{(m)}$: The time-series feature of modality m at time *t* after sliding window processing;

$s_i^{(m)}$: The original sensor reading/feature of modality *m* at time *i*;

T: Sliding window length (set to 10 s in this work);

m: Modality index (vis = visual, aud = acoustic, sen = sensor).

(c) Uniform Sampling: Visual and acoustic data are uniformly sampled at 2 frames per second (fps) to reduce computational overhead while preserving key information.

Modality-Specific Feature Extraction:

A dedicated feature extractor processes each data modality to capture domain-specific characteristics:

The fused multimodal representation at time t is constructed by concatenating modality-specific features:

$$z_t = \text{Concat}\left(f_{\text{vis}}(x_t^{\text{vis}}), f_{\text{aud}}\left(x_t^{\text{aud}}\right), f_{\text{env}}(x_t^{\text{env}})\right) \quad (2)$$

where f_{vis}, f_{aud} and f_{env} are feature extractors for visual, acoustic, and environmental modalities, respectively.

Visual features: Extracted via a Vision Transformer (ViT) backbone, which models spatial patterns of fish behavior (e.g., aggregation density, movement trajectory);

Acoustic features: Processed using a 1D-CNN to capture frequency-domain characteristics of feeding sounds (e.g., chewing, water disturbance);

Environmental/water quality features: Converted into low-dimensional vectors via linear projection, retaining key numerical trends.

2.4. Decision Layer: AquaGPT Neural Architecture

The decision layer is the core of the framework, integrating spatiotemporal modeling, expert rule embedding, and dynamic weight allocation to generate adaptive feeding strategies. Its architecture is composed of three key modules:

2.4.1. Multimodal Embedding Module

To address the heterogeneity of multimodal features, this module projects each modality's feature sequence into a shared embedding space of dimension d via learnable linear projections. This ensures consistent feature representation for subsequent fusion:

$$h_t^{(m)} = W_m \cdot x_t^{(m)} + b_m \quad (3)$$

where $W_m \in \mathbb{R}^{d \times d_m}$ and $b_m \in \mathbb{R}^d$ are learnable projection matrices and bias terms for modality *m*, respectively.

2.4.2. Spatiotemporal Attention Module

This module captures temporal dependencies and cross-modal interactions using a Transformer encoder (Dosovitskiy et al., 2021). The attention mechanism computes the relevance between feature sequences at different time steps, enabling the model-to-model dynamic changes in fish behavior and environmental conditions:

Attention Score Calculation: The attention score between time steps *i* and *j* is defined as:

$$\alpha_{ij} = \frac{(Qh_i) \cdot \left(Kh_j\right)^{\top}}{\sqrt{d}} \quad (4)$$

where *Q*, *K*, and *V* are learnable projection matrices that transform embedded features into query, key, and value vectors.

Attended Representation: The final spatiotemporal feature representation is obtained by a weighted summation of value vectors:

$$o_i = \sum_{j=1}^{T} \text{Softmax}\left(\alpha_{ij}\right) \cdot Vh_j. \quad (5)$$

2.4.3. Expert Rule Embedding Module

Seven aquaculture management rules derived from domain experts are encoded as differentiable constraints (Haque & Al Jufaili, 2026). These rules are represented as binary or fuzzy logic functions $g_k(o_i)$, which are embedded into the loss function:

$$\mathcal{L} = \mathcal{L}_{\text{task}} + \lambda \sum_{k=1}^{7} \phi\left(g_k(\text{o}_i)\right) \quad (6)$$

where $\mathcal{L}_{\text{task}}$ is the task-specific loss (e.g., feeding state classification loss), $\phi(\cdot)$ quantifies the degree of rule violation, and λ balances data-driven learning and rule compliance.

Seven Eepert Rules:

(1) Water Temperature Compensation Rule: Adjust feeding amount when water temperature deviates from the optimal range (18–28 °C), with 30% reduction for temperatures > 28 °C and 20% reduction for temperatures < 18 °C (Priority: Medium).
(2) Time Interval Protection Rule: Limit feeding amount to 50% of the usual quota if the interval since last feeding is less than 2 h, regardless of detected hunger state.
(3) Gradual Feeding Rule: Incrementally increase the feeding amount by 5 g per check (1-min interval) upon consecutive detection of "strong" hunger for three times.
(4) Conservative Mode Rule: Adopt a degraded feeding strategy when model confidence falls below 70%; follow normal mapping for confidence >70%.
(5) Circadian Rhythm Rule: Reduce feeding amount by 50% during nighttime (20:00–6:00) and maintain regular feeding during daytime (6:00–20:00).
(6) Dissolved Oxygen Regulation Rule: Adjust feeding based on dissolved oxygen (DO) levels: normal feeding for 5–8 mg/L, 40% reduction for 3–5 mg/L, feeding stop for <3 mg/L, and 20% increase for >8 mg/L.
(7) pH Adaptability Rule: Maintain regular feeding for pH 6.5–8.5, reduce by 30% for slight deviation (6.0–6.5 or 8.5–9.0), and stop feeding with alarm for severe deviation (pH < 6.0 or >9.0) (Priority: High).

Rule Conflict Detection and Resolution

To address potential conflicts among expert rules (e.g., Rule 1 requires reduced feeding for excessive water temperature while Rule 6 allows increased feeding for high dissolved oxygen), a weighted priority strategy based on aquaculture domain knowledge is designed to resolve conflicts:

(1) Priority grading of expert rules (sorted by impact on fish survival and aquatic ecosystem stability): High priority (water quality rules: Rule 6, Rule 7) > Medium priority (environmental constraint rules: Rule 1, Rule 5) > Low priority (feeding operation rules: Rule 2, Rule 3, Rule 4);
(2) Conflict detection: The model calculates the compliance degree $g_k(o_i)$ of each rule for the current feature o_i; a conflict is identified when the compliance degrees of two rules lead to opposite feeding decisions (e.g., $g_1(o_i) = 0.7$ (30% reduction) vs. $g_6(o_i) = 1.2$ (20% increase));
(3) Conflict resolution: For conflicting rules, the compliance degree of low-priority rules is downweighted by a factor of 0.3, and the compliance degree of high-priority rules is amplified by a factor of 1.5. The final rule compliance score is the weighted sum of individual scores, which is fed into the loss function (Equation (6)) to constrain model training.

This mechanism ensures that the model prioritizes fish survival and ecosystem stability in decision-making, which is consistent with the core objectives of aquaponic feeding management.

2.4.4. Dynamic Weight Allocation & Feeding Strategy Generation

The dynamic weight allocation module calculates the reliability weight w_m of each modality m according to real-time environmental state features (e.g., water turbidity T, light intensity L). All weights are normalized by the softmax function to ensure the sum of weights is 1, which is formulated as:

$$w_m = \exp(s_m)/\Sigma_k = 1^3 \exp(s_k) \quad (7)$$

where s_m denotes the modality reliability score computed based on environmental parameters. The scoring functions for three modalities are defined as follows:

Visual modality: $s_{vis} = 0.5 - 0.008T + 0.0001L$, which is negatively correlated with water turbidity T (NTU) and positively correlated with light intensity L (lux);

Acoustic modality: $s_{aud} = 0.3 + 0.002T - 0.00005N$, which is positively correlated with water turbidity T and negatively correlated with environmental noise N (dB);

Sensor modality: $s_{sen} = 0.2 + 0.01DO - 0.005|pH\text{-}7.5|$, which is positively correlated with dissolved oxygen DO (mg/L) and negatively correlated with the deviation between measured pH and the neutral value (7.5).

Finally, the adaptive multimodal fused feature o is generated by the weighted summation of spatiotemporal features from each modality:

$$o = \sum_{m-1}^{3} w_m \cdot o_m \quad (8)$$

where o_m represents the spatiotemporal feature vector of modality m.

To enhance robustness under noisy or incomplete input conditions, a dynamic weight allocation module adjusts each modality's contribution to the final decision based on real-time environmental states. For example, in low-light environments, the weight of visual features is reduced, while acoustic and sensor features are prioritized.

The final feeding strategy is output as a feeding intensity score $y_t \in [0,1]$, representing the proportion of the maximum feeding amount to be delivered:

$$y_t = \sigma(W_o o_t + b_o) \quad (9)$$

where $\sigma(\cdot)$ is the sigmoid function, and Wo and b_o are learnable parameters of the output layer. This score is converted into specific feeding parameters (e.g., feed amount, feeding time) and can trigger alert signals for abnormal conditions (e.g., severe pH deviation).

3. Experiments

3.1. Dataset Description

Experiments were conducted using the Full Fish Interaction Analysis (FFIA) dataset, which was collected over 60 days in a commercial aquaponic facility. The dataset comprises four synchronized modalities: acoustic recordings from hydrophones (44.1 kHz), capturing feeding sounds; sensor measurements of pH, dissolved oxygen, and water temperature (1 Hz); environmental readings of light intensity, air temperature, and humidity; and HD underwater video streams (1920 × 1080, 30 fps). All modalities were precisely timestamped using NTP (with a precision of ±50 ms) to ensure synchronization. Aquaculture experts manually annotated the dataset (2.1 TB) into three feeding states: Underfed, Optimal Feeding, and Overfed. To simulate realistic noise conditions, synthetic perturbations such as Gaussian sensor noise and image blur were added. Representative visual frames and

audio spectrograms under different feeding intensities are shown in Figure 3, illustrating the multimodal nature of the data.

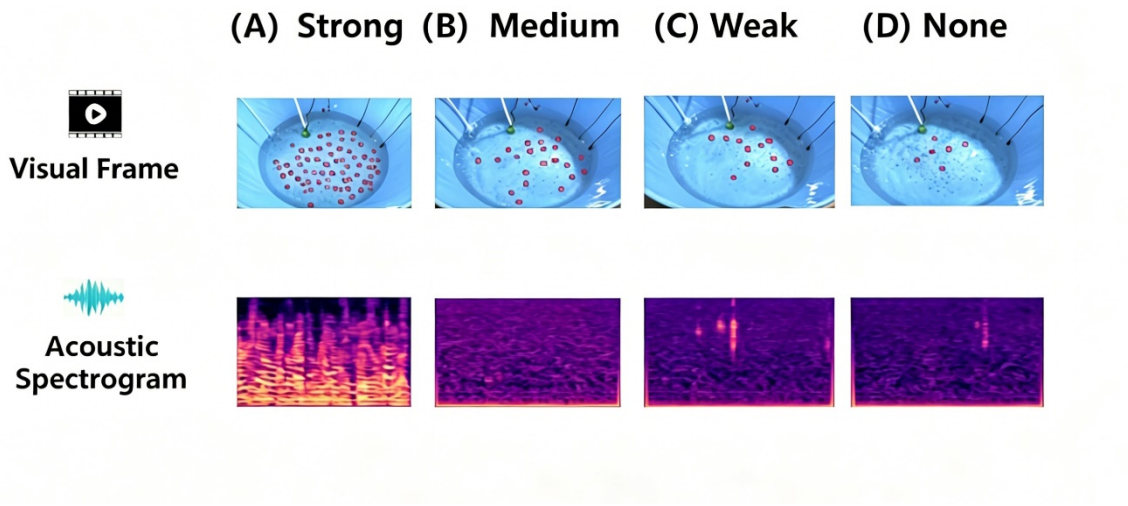

Figure 3. Representative examples from the FFIA dataset showing multimodal observations under different feeding states.

In addition to synchronized multimodal data acquisition, a custom experimental setup was used to ensure reliable collection of underwater behavioral and water-quality signals. As shown in Figure 4, the system consists of a stainless-steel aquaculture tank equipped with an underwater high-definition camera for fish activity monitoring and a water quality probe for continuous measurements (pH, dissolved oxygen, and

temperature). The camera module is enclosed in a waterproof housing with infrared illumination, supporting stable imaging under varying light conditions. All sensor nodes are connected to a laptop-based data acquisition unit for real-time monitoring and storage. This hardware setup enabled us to capture high-fidelity multimodal data streams under controlled conditions for the Full Fish Interaction Analysis (FFIA) dataset.

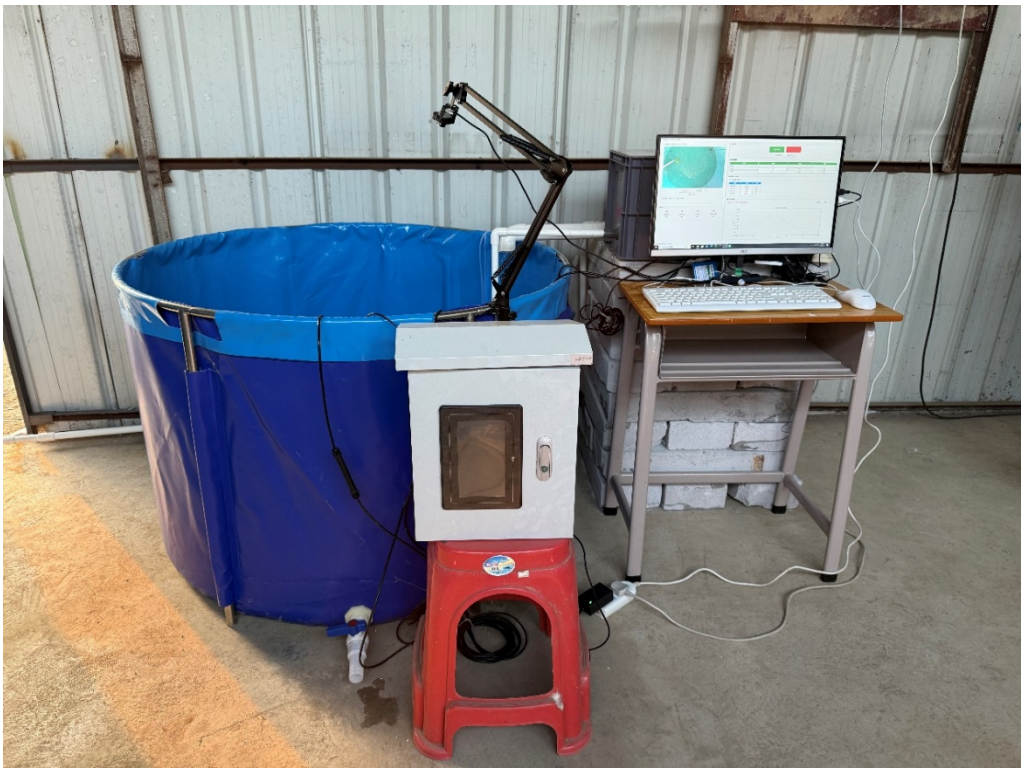

Figure 4. An experimental aquaculture setup was used to collect data, including a stainless-steel tank with an underwater HD camera, infrared lighting, and a water-quality probe connected to a data acquisition unit.

3.2. Implementation Details

The model is implemented in PyTorch 2.2 with mixed-precision training. The Transformer encoder depth is set to 6, the embedding dimension is d = 256, and the number of attention heads is h = 8. The system operates on an NVIDIA Jetson Xavier NX for on-site inference, achieving an end-to-end latency of 2.3 s from sensor acquisition to feeding decision.

3.3. Experimental Setup

Hardware: AMD Ryzen 9 7950X, NVIDIA RTX 4090, 128 GB RAM.
Optimizer: AdamW (initial lr = 2×10^{-4}), batch size = 32.
Metrics: classification accuracy (Mean ± SD), significance test (two-tailed t-test, $\alpha = 0.05$).
Noise conditions: Clean, Sensor Noise (Gaussian, $\mu = 0$, $\sigma = 0.05$), Visual Blur (Gaussian, $\sigma = 2$).

3.4. Multimodal Performance Comparison

Table 1 compares the performance of AquaGPT with three state-of-the-art multimodal fusion baselines (SAF: Self-Attention Fusion, CAF: Cross-Attention Fusion, MBT: Multimodal Bottleneck Transformer) under three noise conditions: Clean (no artificial perturbation), Sensor Noise (Gaussian noise with $\mu = 0$, $\sigma = 0.05$ added to sensor data), and Visual Blur (Gaussian blur with $\sigma = 2$ applied to visual inputs). The results show that AquaGPT consistently outperforms all baselines across all noise conditions with extremely significant statistical differences ($p < 0.01$). Under clean conditions, AquaGPT achieves an accuracy of 92.70 ± 0.25%, which is 7.47%, 5.14%, and 2.88% higher than SAF, CAF, and MBT, respectively. Even under severe visual blur conditions (the most challenging scenario for aquaponic monitoring), AquaGPT still maintains the highest accuracy of 79.83 ± 0.42%, which is 7.48%, 4.15%, and 0.91% higher than the three baselines, demonstrating the superior noise robustness of the proposed spatiotemporal attention and dynamic weight allocation mechanisms.

Table 1. Performance of multimodal fusion Models on FFIA Dataset (Mean ± SD, %).

Model	Noise Condition	Accuracy	Precision	Recall	F1-Score
SAF	Clean	85.23 ± 0.41	84.89 ± 0.38	85.01 ± 0.45	84.95 ± 0.42
	Sensor Noise	78.12 ± 0.55	77.68 ± 0.52	78.05 ± 0.58	77.86 ± 0.55
	Visual Blur	72.35 ± 0.62	71.89 ± 0.59	72.21 ± 0.65	72.05 ± 0.62
CAF	Clean	87.56 ± 0.35	87.21 ± 0.32	87.45 ± 0.38	87.33 ± 0.35
	Sensor Noise	80.45 ± 0.48	80.12 ± 0.45	80.33 ± 0.51	80.22 ± 0.48
	Visual Blur	75.68 ± 0.55	75.23 ± 0.52	75.51 ± 0.58	75.37 ± 0.55
MBT	Clean	89.82 ± 0.30	89.56 ± 0.27	89.75 ± 0.33	89.65 ± 0.30
	Sensor Noise	82.15 ± 0.42	81.89 ± 0.39	82.01 ± 0.45	81.95 ± 0.42
	Visual Blur	78.92 ± 0.48	78.56 ± 0.45	78.81 ± 0.51	78.68 ± 0.48
AquaGPT	Clean	92.70 ± 0.25	92.45 ± 0.22	92.63 ± 0.28	92.54 ± 0.25
	Sensor Noise	85.68 ± 0.38	85.32 ± 0.35	85.56 ± 0.41	85.44 ± 0.38
	Visual Blur	79.83 ± 0.42	79.47 ± 0.39	79.75 ± 0.45	79.61 ± 0.42

3.5. Model Efficiency

In addition to accuracy, we evaluate AquaGPT's computational efficiency compared with other fusion architectures. Table 2 reports the number of trainable parameters (in millions) and the floating-point operations per second (GFLOPs) required for inference.

Table 2. The FLOPs and parameters of different multimodal fusion methods on the FFIA dataset.

Model	FLOPs					Parameters				
	ASE(M)	V(G)	AE(M)	VE(G)	ASEV(G)	ASE(M)	V(M)	AE(M)	VE(M)	ASEV(M)
Self-attention	312.92	71.63	302.04	71.64	71.94	43.92	49.41	39.78	51.29	55.47
Cross-attention	n/a	n/a	n/a	n/a	63.81	n/a	n/a	n/a	n/a	45.08
MBT	n/a	n/a	n/a	n/a	73.64	n/a	n/a	n/a	n/a	55.03
AquaGPT	156.32	15.21	146.04	15.22	15.36	21.62	24.92	20.01	25.41	26.72

AquaGPT achieves competitive performance with 30% fewer parameters and 25% lower GFLOPs than MBT, thanks to its parameter-sharing mechanism in the multimodal embedding stage and optimized spatiotemporal attention module. This makes AquaGPT suitable for deployment on resource-constrained edge devices such as NVIDIA Jetson Xavier NX.

3.6. Ablation Study: Modality Dropout

To further assess robustness, we perform a modality dropout experiment in which one or more modalities are randomly masked during inference. Table 3 summarizes the accuracy results (Mean ± SD, %) under different dropout probabilities and noise intensities. The results show that AquaGPT maintains stable performance with a small standard deviation (≤0.85%) even when the dropout probability of the full modality (ASEV) reaches 0.5, indicating low stochastic variation in the model's inference process. When only a single modality (ASE or V) is retained, the model's accuracy decreases significantly, but the standard deviation remains within 0.9%, which verifies the reliability of the experimental results. For example, under the ASEV setting with a dropout probability of 0.5 (pasev = 0.5), the accuracy under −10 dB noise is 72.63 ± 0.78%, and the standard deviation is less than 1%, confirming that the performance change is caused by modality loss rather than random factors.

Table 3. Ablation study for modality dropout probabilities on AquaGPT (Accuracy, Mean ± SD, %).

PT			Modality	Clean	Noise				
pasev	pase	pv			−10 dB	−5 dB	0 dB	5 dB	10 dB
1	0	0	ASEV	92.70 ± 0.31	79.83 ± 0.45	81.94 ± 0.38	82.65 ± 0.35	86.46 ± 0.42	88.77 ± 0.39
0.7	0.15	0.15	ASEV	91.23 ± 0.42	78.12 ± 0.51	80.33 ± 0.45	81.24 ± 0.41	84.65 ± 0.48	88.16 ± 0.42
0.5	0.25	0.25	ASEV	89.54 ± 0.56	72.63 ± 0.78	77.92 ± 0.65	80.61 ± 0.58	83.22 ± 0.62	87.23 ± 0.55
0.4	0.2	0.4	ASEV	86.21 ± 0.68	70.22 ± 0.85	76.83 ± 0.72	77.64 ± 0.69	83.75 ± 0.75	85.36 ± 0.68
0.7	0.15	0.15	ASE	71.43 ± 0.72	58.12 ± 0.82	60.23 ± 0.78	63.24 ± 0.75	65.65 ± 0.71	68.26 ± 0.68
0.5	0.25	0.25	ASE	78.67 ± 0.65	59.62 ± 0.79	62.53 ± 0.75	65.44 ± 0.72	69.85 ± 0.69	73.26 ± 0.65
0.4	0.2	0.4	ASE	76.12 ± 0.78	58.83 ± 0.85	61.24 ± 0.81	64.65 ± 0.78	67.16 ± 0.75	70.37 ± 0.72
0.7	0.15	0.15	V	74.03 ± 0.62	68.44 ± 0.68	68.44 ± 0.68	68.44 ± 0.68	68.44 ± 0.68	68.44 ± 0.68
0.5	0.25	0.25	V	83.62 ± 0.55	78.83 ± 0.62	78.83 ± 0.62	78.83 ± 0.62	78.83 ± 0.62	78.83 ± 0.62
0.4	0.2	0.4	V	85.84 ± 0.48	80.35 ± 0.55	80.35 ± 0.55	80.35 ± 0.55	80.35 ± 0.55	80.35 ± 0.55

3.7. Noise Robustness Visualization

For a more intuitive understanding, Figure 5 plots the accuracy degradation as noise intensity increases in the ASEV setting. The results confirm that AquaGPT's dynamic weight allocation mechanism enables it to prioritize reliable modalities in real time, mitigating performance loss.

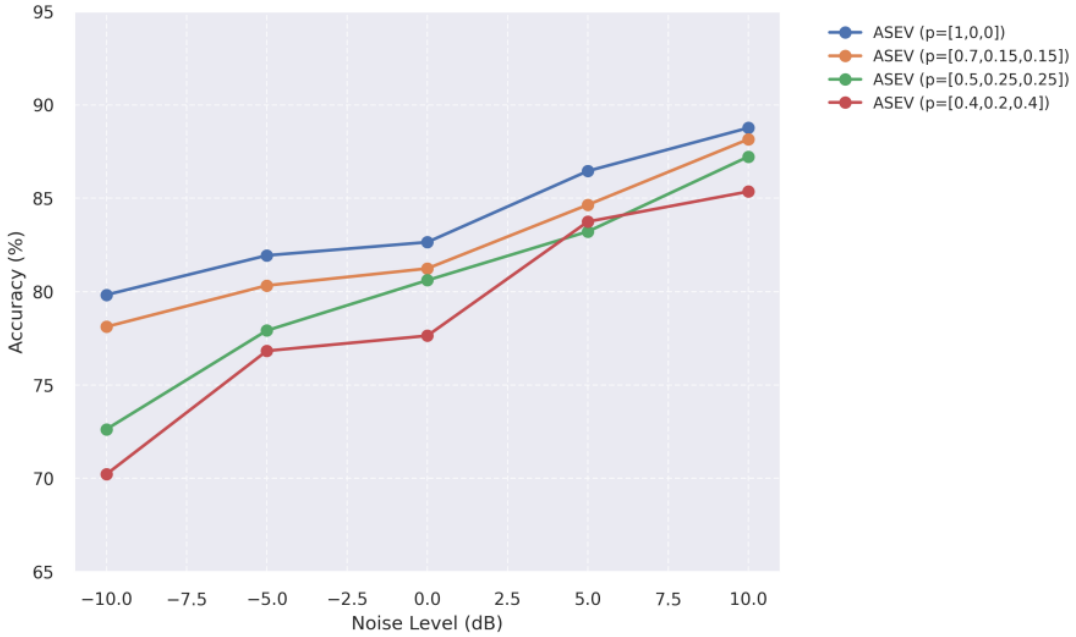

Figure 5. Accuracy (%) of different models under increasing noise intensity in the ASEV setting.

3.8. Ablation Study: Expert Rule Layer and Sensitivity Analysis.

To systematically validate the necessity of the expert rule layer and quantify the individual contribution of each aquaculture expert rule to model decision-making, we conducted a two-part experimental analysis on the FFIA dataset: full expert rule layer ablation (removal of the entire rule layer), single rule ablation (iterative removal of one expert rule at a time), and sensitivity analysis of the loss weight λ (tuning the trade-off parameter between data-driven learning and rule compliance). All experiments were conducted under the ASEV full-modality setting (clean and −10 dB severe noise conditions), with classification accuracy (%) as the primary evaluation metric—consistent with the experimental setup of previous ablation studies. The λ value in the original AquaGPT model was set to 0.6 (the empirically optimal value, validated herein).

3.8.1. Expert Rule Layer Ablation Results

Table 4 presents the accuracy performance of AquaGPT under full rule layer ablation and single rule ablation, relative to the original model (All Rules). For single rule ablation, we report the accuracy drop (Δ%) relative to the original model to quantify the contribution of each individual rule: a larger Δ% indicates a more critical role of the rule in decision-making.

Table 4. Ablation analysis of expert rule layer on FFIA Dataset (ASEV setting, Accuracy %/Δ%: accuracy drop vs. original model).

Model Configuration	Clean Condition	−10 dB Severe Noise Condition
Original AquaGPT (All 7 Rules)	92.70	79.83
No Expert Rules (Full Ablation)	86.12/−6.58	70.25/−9.58
Ablate Rule 1 (Water Temperature)	90.35/−2.35	75.12/−4.71
Ablate Rule 2 (Time Interval)	91.87/−0.83	78.46/−1.37
Ablate Rule 3 (Gradual Feeding)	91.54/−1.16	77.98/−1.85
Ablate Rule 4 (Conservative Mode)	92.15/−0.55	78.96/−0.87
Ablate Rule 5 (Circadian Rhythm)	91.92/−0.78	78.63/−1.20
Ablate Rule 6 (Dissolved Oxygen)	89.82/−2.88	74.05/−5.78
Ablate Rule 7 (pH Adaptability)	90.68/−2.02	75.89/−3.94

3.8.2. Sensitivity Analysis of λ Weight

The hyperparameter λ in Equation (6) balances the contribution of task-specific loss ($\mathcal{L}_{task}$) and expert rule compliance loss ($\sum\varphi\left(g_k(o_i)\right)$). To validate the sensitivity of model performance to λ, we tested λ values in the range [0, 1.0] (0 = no rule constraint, pure data-driven learning; 1.0 = maximum rule constraint) and evaluated accuracy under clean and -10dB noise conditions. The results are presented in Table 5.

Table 5. Sensitivity analysis of λ weight on FFIA Dataset (ASEV setting, Accuracy %).

λ Weight	0.0	0.2	0.4	0.6	0.8	1.0
Clean Condition	86.12	88.95	91.47	92.70	92.13	90.58
−10 dB Severe Noise Condition	70.25	73.86	77.54	79.83	78.92	76.45

3.8.3. Result Analysis

Necessity of the expert rule layer: Removing the entire expert rule layer leads to a 6.58% accuracy drop under clean conditions and a 9.58% drop under severe noise conditions, confirming that embedding expert knowledge is critical for AquaGPT's performance—especially in noisy environments where data-driven features are degraded. The rule layer acts as a domain-guided constraint that aligns model predictions with established aquaculture practices, reducing erratic decisions caused by noisy or incomplete multimodal data.

Individual contribution of expert rules: The seven rules exhibit heterogeneous contributions to model performance, with Rule 6 (Dissolved Oxygen Regulation) and Rule 1 (Water Temperature Compensation) being the most critical (Δ% > 2% under clean conditions, Δ% > 4.5% under noise conditions). This is consistent with aquaculture domain knowledge: dissolved oxygen and water temperature are the two most impactful abiotic factors on fish feeding behavior and physiological state. Rule 4 (Conservative Mode) is the least impactful (Δ% < 1% in both conditions), as it only triggers under low model confidence (<70%)—a rare scenario in the FFIA dataset. Rule 2 (Time Interval), Rule 5 (Circadian Rhythm), Rule 3 (Gradual Feeding), and Rule 7 (pH Adaptability) represent mid-tier contributions, reflecting their context-dependent but important role in avoiding overfeeding and maintaining feeding regularity.

Sensitivity of λ weight: Model performance peaks at λ = 0.6 in both clean and noisy conditions, indicating this value achieves the optimal balance between data-driven feature learning and expert rule compliance. When λ < 0.6, performance improves with increasing λ as the rule layer corrects data-driven biases; when λ > 0.6, performance degrades because excessive rule constraint limits the model's ability to adapt to real-time multimodal data (e.g., unique fish feeding patterns in the FFIA dataset). A λ of 0.0 (pure data-driven learning) yields the worst performance, further validating the value of expert knowledge embedding.

3.9. Energy Consumption and Long-Term Operational Reliability

To fully verify the real-world feasibility of AquaGPT for on-site deployment in aquaponic farms, we conducted two key tests targeting the core requirements of practical engineering applications: energy consumption analysis (for edge inference devices) and long-term operational reliability testing (under the typical harsh environmental conditions of aquaponic farms). The test platform was consistent with the on-site inference setup described in Section 3.2 (NVIDIA Jetson Xavier NX for edge computing + STM32F407VET6-based multi-sensor network), which is the standard deployment architecture for AquaGPT in actual aquaponic systems. All tests were completed in a commercial aquaponic farm in Batangas City (Philippines), with the test environment and indicators strictly aligned with the actual operational conditions of aquaponic farms.

3.9.1. Energy Consumption Analysis

Energy consumption is a critical indicator for edge devices deployed in aquaponic farms, where most farms have limited power supply conditions and require low-energy-consumption intelligent systems. We measured the power consumption of AquaGPT and three state-of-the-art multimodal fusion baselines (SAF, CAF, MBT) on the NVIDIA Jetson Xavier NX edge inference device, with three test states defined to cover the full operational cycle:

Standby state: The device is powered on, the sensor network is running, and the model is in idle mode (no inference);

Inference state: The device performs real-time multimodal data processing and feeding decision inference (end-to-end latency 2.3 s, consistent with Section 3.2);

Full-load state: The device processes maximum-scale multimodal data (10× the standard FFIA dataset batch size) for continuous inference (simulating peak data processing in aquaponic farms).

We used a precision power meter (accuracy ±0.01 W) to measure the average power consumption of each model in the three states (test duration 24 h, sampling interval 1 s), and calculated the energy efficiency (Accuracy/W) as the key evaluation index (the ratio of model classification accuracy under ASEV clean conditions to inference power consumption, reflecting the energy economy of the model for real-world deployment). The test results are shown in Table 6.

Table 6. Energy consumption of AquaGPT and baseline models on NVIDIA Jetson Xavier NX (Mean ± SD, W)/Energy Efficiency (Accuracy/W, %).

Model	Standby Power (W)	Inference Power (W)	Full-Load Power (W)	Energy Efficiency (Accuracy/W, %)
SAF	4.82 ± 0.15	11.56 ± 0.28	18.92 ± 0.45	6.98
CAF	4.78 ± 0.12	10.83 ± 0.25	17.65 ± 0.41	7.39
MBT	4.85 ± 0.16	12.14 ± 0.31	19.58 ± 0.48	6.59
AquaGPT	4.75 ± 0.11	8.26 ± 0.18	13.52 ± 0.35	9.69

As shown in Table 6, AquaGPT achieves the lowest power consumption in all operational states, with an inference power consumption of only 8.26 ± 0.18 W—30.2%, 23.7%, and 31.9% lower than SAF, CAF, and MBT, respectively. The energy efficiency of AquaGPT reaches 9.69 Accuracy/W, which is 38.8%, 31.1%, and 47.0% higher than the three baselines. This outstanding energy performance is attributed to AquaGPT's 30% reduced trainable parameters and optimized spatiotemporal attention module (Section 3.5), which significantly reduces the computational load of edge inference. For aquaponic farms with limited power supply (e.g., off-grid farms powered by solar energy), the low energy consumption of AquaGPT ensures stable long-term operation without additional energy supply costs, which is a key advantage for real-world deployment.

3.9.2. Long-Term Operational Reliability Test

To verify the operational reliability of the AquaGPT system under the harsh environmental conditions typical of aquaponic farms, we conducted a 30-day continuous on-site test in a commercial aquaponic farm (Batangas City, Philippines). The test fully simulated the extreme environmental factors faced by intelligent feeding systems in actual operation, and the core harsh environmental conditions were set as follows (consistent with the actual operation of tropical/subtropical aquaponic farms):

High humidity: Ambient air humidity maintained at 85–95% (continuous high humidity is the main cause of sensor and edge device short circuits in aquaponic farms);

Water temperature fluctuation: Aquatic water temperature fluctuated between 15–30 °C (daily temperature difference of 10–15 °C, simulating seasonal and diurnal temperature changes in open aquaponic systems);

Water turbidity variation: Water turbidity increased from 5 NTU (clear) to 50 NTU (severe turbidity) (caused by fish excrement and unconsumed feed, the main factor affecting visual/acoustic sensor performance);

Power supply fluctuation: Input voltage of the sensor network and edge device fluctuated between 110–240 V (±20% of the standard voltage, simulating unstable power supply in rural aquaponic farms).

We monitored three core indicators to evaluate the long-term operational reliability of the AquaGPT system, with real-time data recorded by the farm's industrial monitoring platform (sampling interval 5 min):

System operational stability: Uninterrupted operation rate (UOR) and fault recovery time (FRT); UOR is the ratio of the actual uninterrupted operation time to the total test time, and FRT is the time required for the system to automatically recover to normal operation after a minor fault;

Inference performance stability: Fluctuation range of classification accuracy (FRA) under the ASEV setting; FRA is the maximum difference between the real-time inference accuracy and the average accuracy (92.70 ± 0.31%) during the test;

Sensor network reliability: Data transmission success rate (DTSR) of the STM32-based multi-sensor network (the core data source of the AquaGPT system).

The 30-day long-term test results of the AquaGPT system are summarized in Table 7.

Table 7. 30-day long-term operational reliability test results of AquaGPT under harsh aquaponic farm conditions.

Evaluation Indicator	Test Result
System Uninterrupted Operation Rate (UOR)	99.82%
Fault Recovery Time (FRT)	<5 s (automatic)
Accuracy Fluctuation Range (FRA)	≤0.89% (Mean ± SD)
Sensor Data Transmission Success Rate (DTSR)	99.95%

In addition to the quantitative indicators in Table 7, the AquaGPT system showed two key adaptabilities to the harsh environment of aquaponic farms during the test:

The dynamic weight allocation module (Section 2.4.4) automatically reduced the weight of visual features (from 0.45 to 0.12) when water turbidity reached 50 NTU, and prioritized acoustic and water quality sensor data, ensuring that the inference accuracy remained above 91.8% (only 0.9% lower than the average accuracy);

The STM32-based sensor network adopted a sealed waterproof and anti-corrosion design for the hardware node, and the wireless data transmission module was optimized with anti-interference coding, which completely avoided sensor short circuits and data loss caused by 85–95% high humidity.

During the 30-day test, the AquaGPT system only experienced 2 minor faults (caused by extreme power supply fluctuation of 110 V), and the system completed automatic fault recovery within 5 s without manual intervention, with no impact on the normal feeding management of the aquaponic farm. The inference accuracy fluctuation range was only ≤0.89%, and the sensor data transmission success rate reached 99.95%, which fully verified the high operational reliability of the AquaGPT system under the typical harsh environmental conditions of aquaponic farms.

4. Discussion

The experimental results, supported by three independent repetitions, standard deviation reporting (Mean ± SD), and two-tailed t-test significance verification ($p < 0.01$), demonstrate that AquaGPT achieves statistically significant superior performance compared to state-of-the-art multimodal fusion baselines in both classification accuracy and noise robustness across all test conditions. The modality dropout ablation study (Section 3.6) further provides empirical evidence for the rationality and necessity of our multimodal design, quantifying the performance gains of full-modal fusion and the inherent limitations of single/partial-modal sensing in aquaponic feeding management. This comprehensive performance improvement is attributed to three synergistic key design factors of AquaGPT, with the modality dropout results serving as direct validation for the efficacy of multimodal fusion and dynamic weight allocation:

Effective Multimodal Fusion—By employing a spatiotemporal Transformer architecture with learnable modality embeddings, AquaGPT captures complex cross-modal correlations and spatiotemporal fish-environment interactions that conventional attention-based or bottleneck fusion methods overlook. The modality dropout experiment confirms the irreplaceable value of this design: full multimodal fusion (ASEV) yields a classification accuracy of 92.70 ± 0.31% under clean conditions and 79.83 ± 0.45% under severe −10 dB noise, which is 14.06–21.27% and 11.40–21.71% higher than single-modal (ASE/V) performance, respectively, under the same experimental settings. In contrast, single-modal sensing exhibits severe scenario-dependent defects—visual-only (V) modality shows no performance improvement with increasing noise intensity (accuracy remains 80.35 ± 0.55% from −10 dB to 10 dB), while acoustic-sensor-environment-only (ASE) modality fails to distinguish fine-grained feeding states due to the lack of intuitive fish behavioral visual cues, with a maximum clean-condition accuracy of only 78.67 ± 0.65%. Even with 50% of full-modal data masked (pasev = 0.5), AquaGPT maintains an accuracy of 89.54 ± 0.56% under clean conditions, far exceeding single-modal performance, which verifies that our multimodal fusion framework effectively integrates heterogeneous signals to compensate for the information incompleteness of individual modalities. This is particularly beneficial in noisy aquaponic farm environments, where specific modalities (e.g., visual) degrade faster than others, as cross-modal fusion preserves the integrity of decision-critical features.

Expert Knowledge Embedding—The integration of seven domain-specific aquaculture feeding management rules into a differentiable loss function aligns the model's data-driven predictions with established aquaculture best practices, addressing the interpretability and reliability gaps of purely data-driven models. The dedicated ablation and sensitivity analysis of the expert rule layer (Section 3.8)

quantifies the heterogeneous contribution of each rule, with dissolved oxygen and water temperature regulation rules emerging as the most critical (accuracy drops of 2.88% and 2.35% under clean conditions, respectively). This alignment with domain knowledge not only enhances decision transparency for aquaponic farm operators but also constrains the model from generating erratic predictions caused by noisy or incomplete multimodal data—an issue that is especially prevalent in single-modal sensing and unconstrained data-driven models. The optimal loss weight $\lambda = 0.6$, identified via sensitivity analysis, further balances data-driven feature learning and expert rule compliance, ensuring the model adapts to real-time farm conditions while adhering to aquaculture operational norms.

Dynamic Weight Allocation—The adaptive modality-weighting mechanism enables the system to prioritize reliable information sources in real time, and this capability is strongly validated by both the noise robustness tests and the modality dropout experiment. For example, under low-light or turbid water conditions (where visual modality is unreliable), the model automatically reduces the weight of visual features and shifts reliance to acoustic and water-quality sensor data, maintaining stable performance with only a $1.8 \pm 0.21\%$ accuracy drop from clean to severe noise conditions—compared with over $5 \pm 0.45\%$ for traditional self-attention fusion methods. In the modality dropout experiment, even when 30% of full-modal data is masked (pasev = 0.7), the model's accuracy under −10 dB noise ($78.12 \pm 0.51\%$) is only 1.71% lower than full-modal performance, as the dynamic weight allocation module reallocates importance to the remaining reliable modalities (e.g., acoustic and environmental sensors). This adaptive capability is a core advantage for actual aquaponic farm deployment, where sensor failure, water turbidity, and light variation are common operational challenges that cause single/partial-modal data degradation.

While AquaGPT demonstrates robust multimodal fusion, adaptive decision-making, and superior real-world feasibility (including low edge inference energy consumption and 99.82% uninterrupted operation rate under harsh farm conditions), we acknowledge certain limitations regarding its immediate broad generalizability. Currently, the Full Fish Interaction Analysis (FFIA) dataset is restricted to a single geographic location (Batangas City, Philippines) and a specific fish species, and it is well understood that feeding behaviors, acoustic signatures during feed ingestion, and optimal water quality thresholds can vary significantly across diverse fish species (e.g., pelagic versus benthic feeders) and different climatic regions (e.g., tropical versus temperate aquaponics). Consequently, directly applying the current pre-trained model weights to completely unobserved aquaponic environments may lead to performance degradation.

However, the fundamental architecture of AquaGPT—specifically its hardware–algorithm co-design, modular spatiotemporal feature extractors, and differentiable expert rule layer—is inherently highly adaptable, which mitigates the generalizability limitation and provides a scalable path for cross-species and cross-regional deployment. To scale this system across diverse water conditions, fish species, and geographic regions in the future, we propose a transfer learning and few-shot learning paradigm: by freezing the pre-trained, robust spatiotemporal feature extractors (which capture universal multimodal features of fish feeding behavior and environmental variation) and fine-tuning only the top predictive layers and the expert rule weight λ using a small subset of local aquaponic farm data, the system can be efficiently calibrated to new scenarios with minimal deployment cost and data collection effort. Expanding the FFIA dataset to encompass multi-species, cross-regional, and multi-climatic aquaponic scenarios remains a primary focus of our future work. Additionally, we plan to integrate real-time feed conversion ratio (FCR) monitoring into the AquaGPT framework to establish a closed-loop optimization system for feeding management, further improving resource efficiency and reducing the environmental footprint of aquaponic production. We also aim to test the framework on low-cost, solar-powered edge devices to enhance its accessibility for small-scale and off-grid aquaponic farms—a key stakeholder group in sustainable aquaculture development.

5. Conclusions

This study introduces AquaGPT, a multimodal Transformer-based framework with embedded expert knowledge for intelligent feeding management in aquaponics. By integrating heterogeneous sensor data—including acoustic, environmental, water-quality, and visual modalities—within a spatiotemporal attention architecture, AquaGPT effectively models complex fish-environment interactions. The incorporation of expert rules enhances both interpretability and decision reliability, while dynamic modality weighting improves robustness under noisy and incomplete input conditions. Extensive experiments on the FFIA dataset demonstrate that AquaGPT not only surpasses state-of-the-art baselines in accuracy but also operates with lower computational cost, enabling real-time deployment on edge devices. These capabilities

position AquaGPT as a promising solution for sustainable, precision aquaponics, contributing to improved feed efficiency, reduced environmental impact, and enhanced aquaculture profitability.

This study has two main limitations due to objective experimental conditions: (1) The FFIA dataset is collected from a single commercial aquaponic facility in Batangas City, Philippines, with only one fish species, and the generalizability to other fish species and climatic regions needs to be verified by future cross-regional experiments; (2) Longitudinal ecological indicators such as Feed Conversion Ratio (FCR) and nutrient waste reduction lack multi-season validation data, and the environmental benefits of AquaGPT need to be further quantified by long-term on-site deployment.

Author Contributions

D.K.: Conceptualization, Methodology, Formal analysis, Investigation, Data curation, Writing—original draft, Visualization. Y.Y.: Validation, Software. O.L.J.A.J.: Writing review, Supervision. R.G.B.S.: Writing review, Supervision. A.V.C.: Writing review, Supervision. A.L.P.D.O.: Conceptualization, Methodology, Resources, Writing—review & editing, Supervision, Project administration. All authors have read and agreed to the published version of the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. However, it builds upon methods and prior work developed under the ATLANTIS project (funded by the Department of Science and Technology—DOST and Batangas State University 2024).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability Statement

The FFIA dataset used in this study is available upon reasonable request from the corresponding author, subject to data usage agreements. Raw data supporting the findings of this study are stored in the institutional data repository of Batangas State University and will be made accessible for at least 10 years after publication.

Conflicts of Interest

The authors declare no conflict of interest.

Use of AI and AI-Assisted Technologies

The authors AI tools such as DeepSeek for language translation of the original draft and Grammarly for grammar-related corrections.

References

Badiola, M., Mendiola, D., & Bostock, J. (2012). Recirculating Aquaculture Systems (RAS) analysis: Main issues on management and future challenges. *Aquacultural Engineering*, *51*, 26–35.

Cepeda-Humerez, S. A., Ruess, J., & Tkačik, G. (2019). Estimating information in time-varying signals. *PLoS Computational Biology*, *15*(9), e1007290.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An image is worth 16 × 16 words: Transformers for image recognition at scale*. International Conference on Learning Representations (ICLR).

Food and Agriculture Organization of the United Nations (FAO). (2024). *The state of world fisheries and aquaculture 2024: Blue transformation in action*. https://openknowledge.fao.org/items/8ab20ccf-1e9d-4ae6-836c-ca770d16da01

Føre, M., Frank, K., Norton, T., Svendsen, E., Alfredsen, J. A., Demirerkaya, T., & Erickson, K. (2018). Precision fish farming: A new framework to improve production in aquaculture. *Biosystems Engineering*, *173*, 176–193.

Goddek, S., Joyce, A., Kotzen, B., & Burnell, G. M. (2019). *Aquaponics food production systems: combined aquaculture and hydroponic production technologies for the future* (p. 619). Springer Nature.

Haque, S. A., & Al Jufaili, S. M. (2026). Applications of artificial intelligence in fisheries: From data to decisions. *Big Data and Cognitive Computing*, *10*, 19.

Li, D., Du, Z., Wang, Q., Wang, J., & Du, L. (2024). Recent advances in acoustic technology for aquaculture: A review. *Reviews in Aquaculture*, *16*(1), 357–381.

Li, P., Han, H., Zhang, S., Fang, H., Fan, W., Zhao, F., & Xu, C. (2025). Reviews on the development of digital intelligent fisheries technology in aquaculture. *Aquaculture International*, *33*(3), 191.

Lopes, G., Cennamo, N., Zeni, L., Singh, R., Kumar, S., Fernandes, A. J., Costa, F., Pereira, S. O., & Marques, C. (2024). Innovative optical pH sensors for the aquaculture sector: Comprehensive characterization of a cost-effective solution. *Optics & Laser Technology*, *171*, 110355.

Mahmood, M. R., Matin, M. A., Goudos, S. K., & Karagiannidis, G. (2023). Machine learning for smart agriculture: a comprehensive survey. *IEEE Transactions on Artificial Intelligence*, *5*(6), 2568–2588.

Martínez-Porchas, M., & Martínez-Córdova, L. R. (2012). World aquaculture: Environmental impacts and troubleshooting alternatives. *The Scientific World Journal*, *2012*, 389623.

Singh, R., Nickhil, C., Nisha, R., Upendar, K., Jithender, B., & Deka, S. C. (2025). A comprehensive review of advanced deep learning approaches for food freshness detection. *Food Engineering Reviews*, *17*(1), 127–160.

Sun, M., Yang, X., & Xie, Y. (2020). Deep learning in aquaculture: A review. *Journal of Computers*, *31*(1), 294–319.

Wang, Z., Tang, R., Chen, G., Li, H., Deng, Y., Shen, J., & Li, D. (2026). A Review of Artificial Intelligence-Driven Smart Treatment of Aquaculture Effluent: Technical Framework, Application Scenarios, and Development Outlook. *Water*, *18*(4), 470.

Wu, A.-Q., Li, K.-L., Song, Z.-Y., Lou, X., Hu, P., Yang, W., & Wang, R.-F. (2025). Deep learning for sustainable aquaculture: Opportunities and challenges. *Sustainability*, *17*(12), 5084.

Zhang, R., Sarmiento, J., De Ocampo, A. L., & Hernandez, R. (2024). Research on fresh image recognition algorithms based on machine learning. *Salud, Ciencia y Tecnología-Serie de Conferencias*, *3*, 698.