OPEN ACCESS

International Journal of Innovations in Science & Technology

Communication-Aware Autonomous Underwater Vehicle Framework with Multimodal Temporal Transformer for Integrated Perception, Routing, and Predictive Maintenance

Zaheer Ahmad[1], Adnan Karamat[1], Naveed Ilyas[2]
[1]Capital University of Science and Technology, Islamabad, Pakistan
[2]University of California, Riverside, USA
*Correspondence: zaheer.ahmad01@cust.edu.pk, adnan.karamat@cust.edu.pk, naveedilyas.ucr@gmail.com

Citation | Ahmad. Z, Karamat. A, Ilyas. N, "Communication-Aware Autonomous Underwater Vehicle Framework with Multimodal Temporal Transformer for Integrated Perception, Routing, and Predictive Maintenance", IJIST, Special Issue pp 656-673, May 2026

Received | April 07, 2026 **Revised |** May 13, 2026 **Accepted |** May 16, 2026 **Published |** May 18, 2026.

The growing applicability of Autonomous Underwater Vehicles (AUVs) in oceanographic studies, environmental monitoring, and infrastructure inspections, and infrastructure inspections is limited by the problem of underwater acoustic communication, which includes low bandwidth, high propagation delay, and high-power limitations. Current AUV networks are based on some form of static routing and reactive fault management, which are not suitable to large-scale, long-range underwater missions. This paper introduces a new communication-aware architecture for AUV networks which is based on a multimodal temporal transformer (MMTT)-based framework. In this framework, communication, routing and predictive maintenance are considered a multi-task sequential learning problem, to be optimized as a joint problem on a common temporal representation. It is based on a time-varying dynamic graph model, and it considers the mobility of nodes, environmental perturbations, and varying channel conditions. Evaluations of representative underwater scenarios through simulations show that our solution is improves packet delivery, energy savings, and route efficiency by factors of 20, 15, and 30, respectively than baseline protocols. Our framework is also able to reduce end-to-end latency by a quarter, ensuring timely data delivery in a dynamic underwater setting. These advances allow the proposed system to be more reliable and energy-efficient thereby supporting long-duration, scalable, and autonomous underwater missions.

Keywords: Autonomous Underwater Vehicles (AUVs), Adaptive Routing, Predictive Maintenance, Artificial Intelligence (AI), Energy Efficiency

Introduction:

Autonomous Underwater Vehicles (AUVs) have emerged as a ubiquitous technology in a variety of underwater tasks, such as seabed mapping, offshore structure inspection, oceanography data acquisition and military intelligence monitoring [1][2][3]. The rising complexity of underwater operations has driven the use of multiple cooperative AUVs to work as a networked system, instead of being individual platforms of operation [4]. The coordination between AUVs needs good communication, efficient routing, and maintenance of the system so that their missions are successful in difficult and unpredictable underwater conditions [5][6].

The main method of underwater communication is acoustic because the radio frequency waves are highly attenuated in water and therefore, underwater communication is mainly based on acoustic signals only [1]. The acoustic communication is prone to numerous intrinsic drawbacks, including low data rate, high propagation delays, Doppler effects, multipath fading, and frequent link outages [5]. These are the features that have a major impact on the performance of underwater networks and render the traditional terrestrial networking solutions inapplicable [7]. Consequently, underwater networks have to be developed with special protocols capable of withstanding high latency, unreliable connectivity, and low bandwidth [8][9].

The majority of current underwater networking solutions use either fixed or semi-fixed routing models that are not particularly resilient to dynamic conditions in the environment and nodes mobility [9]. Existing underwater networking solutions utilize either fixed or semi-fixed routing models which are not particularly adaptable to the dynamic environmental conditions and mobility of nodes [10]. Most of the current underwater networking solutions use fixed or semi-fixed routing models that are not Because AUVs change location either as a result of mission demands or ocean currents, network topology can shift regularly, breaking previously established routes in the process of doing so [11]. This in turn leads to high packet loss, retransmissions as well as an increase in control overhead in such protocols [12]. These inefficiencies do not only reduce the performance of communication, but also hasten the battery depletion, thus reducing mission duration [13]. Energy efficiency is important especially to AUVs because the process of replacing or recharging the batteries is usually impractical when they are on a long-term mission [10]. In comparison to sensor nodes on the ground, AUVs require substantial energy to propel, sense, communicate, and process onboard computationally intensive functions that consume large amounts of energy. The inefficiency of communication and routing decisions may result in the unjustified use of energy that restrains the range of operations and mission endurance [11]. Consequently, the networking mechanisms will require energy awareness to ensure that underwater operations can be sustainable in the future [14].

Besides communication issues, system reliability is also a significant issue in multi AUV deployments [6]. Conventional AUV systems are often based on a reactive maintenance approach, in which faults are only identified once they have happened [3]. Such a reactive model may lead to unforeseen mission failures, the loss of valuable information, and expensive recovery efforts [2]. Unexpected failures in a system may be disastrous in mission-critical environments like deep-sea exploration or defense uses of the system [4]. Predictive maintenance is a promising solution as it allows detecting the degradation of the components and possible failures early on before they become critical issues [6]. Predictive models can predict impending failures by constantly tracking system health indicators and using historical data to predict failures before they happen [15]. Despite the successful deployment of predictive maintenance in industrial and land-based cyber-physical systems, it has not been widely used in underwater autonomous systems because of the environmental factors and the absence of integrated designs [16].

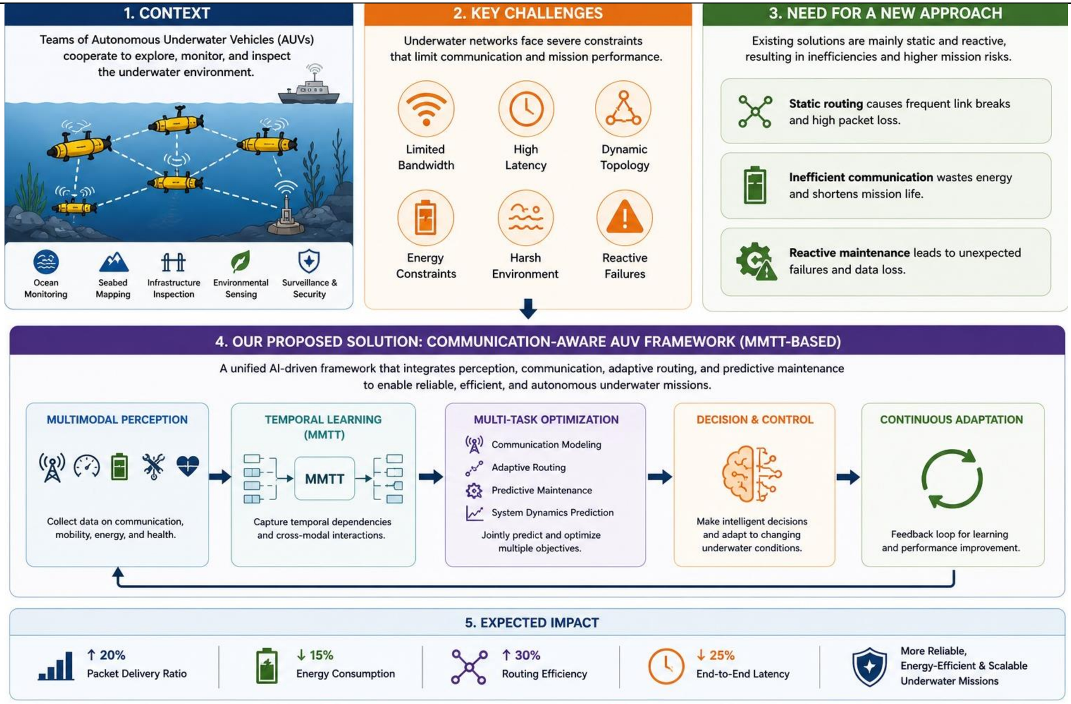


Figure 1. Introduction to the proposed MMTT-based communication-aware multi-AUV framework, showing the mission context, key underwater networking challenges, need for an adaptive AI-driven approach, integrated workflow for perception, routing, and predictive maintenance, and expected performance improvements.

Moreover, the current underwater communication solutions tend to assume networking, perception, and maintenance as separate entities human—¿More to the point, the current underwater communication systems tend to assume that networking, perception, and maintenance are different entities. Such a disjointed design restricts the system from making informed decisions holistically [7]. A combined method of environmental perception, communication-sensitive routing, and predictive maintenance could be an effective way to improve operational efficiency and mission robustness to a great extent of effectiveness and efficiency of operations and missions [12]. However, recent progress has seen the development of routing efficiency under changing underwater conditions while considering factors such as residual energy and node position, e.g. Drift-aware Fuzzy Cluster-based Routing, proposed. On the same note, have proposed a dynamic cluster head routing protocol, which improves the ratio of packet delivery and longevity of the network by creating dynamic yet energy efficient clusters [11].

There have been monumental advances in the field of AI-driven routing [14]. writes about the implementation of Reinforcement Learning (RL) in dynamic underwater settings, where routing choices are made by AI-based routing systems that predict link quality and optimize routing decisions in an adaptive way [14]. These improvements indicate the potential of AI-based solutions to enhance the adaptability and resilience of UWSNs [6][9].

Lastly, energy-efficient routing algorithms, including APC-EEDBR, have been demonstrated to have high stability of links as well as reduced energy expenditure by incorporating adaptive power control [10]. The MDTBOR protocol, proposed employs the use of multi-dimensional dynamic trust models to provide secure and efficient routing during multi-attack situations, which is a dependable alternative to the traditional routing mechanisms. In order to overcome the mentioned challenges and close the research gaps existing in the literature on multi-AUV communication systems, the research objectives of this study are the following: The main contributions of this work are summarized as follows:

Unified Multimodal Temporal Learning Framework: We present a new Communication-Aware Multimodal Temporal Transformer (MMTT) framework which learns underwater AUV networks as a sequence of learning. The given solution is a unified representation of communication dynamics, mobility patterns, energy variations, as well as system health that overcomes the short-comings of previous modular and rule-based approaches.

Cross-Modal Attention-Based Fusion Mechanism: We present a dynamic cross-modal attention model that dynamically combines heterogeneous features (communication, mobility, energy, and health) that are modalities. This allows context-based fusion of features and helps to be robust in the extremely dynamic and uncertain underwater conditions.

End-to-End Multi-Task Optimization: We come up with a joint multi-task learning policy that optimizes communication adaptation, routing, predictive maintenance and system state prediction in a common representation. This eliminates the need for individual task-specific models, and provides system-level coordination.

Learning-Based Communication and Routing Strategy: In contrast to traditional heuristic-based methods, we create communication control and routing decisions as data-driven functions that are learned through time variations. The suggested framework facilitates the control of adaptive transmission power, optimization of data rate and efficient routing of time varying underwater networks.

Predictive Maintenance via Temporal Degradation Modeling: To predict failure probability and remaining useful life (RUL) of the AUV components, we use a predictive maintenance module as part of the temporal transformer framework. This enables early fault detection and promotes proactive maintenance to enhance reliability of the system.

Simulation-Driven Multimodal: We create a semi-realistic multimodal underwater dataset that is built on a realistic simulation and run large-scale experiments on the state-of-the-art baselines, such as LSTM, Informer, and TimesNet. The findings show that the proposed framework is effective in enhancing packet delivery ratio (PDR), minimizing latency, increasing energy efficiency, and making accurate predictions of RUL.

Literature Review:

Recent developments in underwater wireless sensor networks (UWSNs) and autonomous underwater vehicle (AUV) systems have aimed to enhance communications reliability, routing efficiency, and sustainability in harsh underwater environments [1][2]. The conventional routing schemes have transformed and advanced to intelligent routing schemes, dynamic adaptive routing schemes and machine learning based routing schemes over the last ten years that can manage dynamic underwater routing schemes [10][16]. In more recent times, there has been a research focus on integrated architectures that incorporate communication optimization with predictive system monitoring and intelligent routing schemes [4][5]. The section is a critical review of literature that exists in the area, attempting to classify the related work into routing and communication strategies, intelligent and cross-layer optimization techniques, and predictive maintenance structures, and address the key research gaps that drive the proposed research.

Communication and Routing Strategies in Underwater Networks:

Early studies in underwater communication were mainly concerned with tackling the special properties of the acoustic media such as high latency, low bandwidth and severe attenuation. Early research provided the theoretical foundation for underwater acoustic networking and identified the necessity for specialized routing protocols in underwater conditions

Many of the current routing designs were based on traditional routing schemes like Depth-Based Routing (DBR) and vector-based forwarding schemes on the basis of their simplicity and minimal overhead. Nevertheless, it has been demonstrated in contemporary analyses that such strategies have serious constraints in dynamic underwater conditions with moving nodes and changing channel conditions in the underwater environment in dynamic conditions of node motion and channel variation [10][16]. Recent routing proposals proposed between 2024 and 2026 have aimed at enhancing flexibility through clustering, trust assessment, and mobility-aware routing. As an example, drift-aware fuzzy clustering algorithms enhance the stability of routes in dynamic underwater environments with the ability to dynamically choose cluster heads as per the environmental parameters [13]. Likewise, dynamic cluster-head routing strategies can be used to make inter-cluster communication more efficient and minimize the loss of packets in the changing network conditions [11]. Alternatively, trust-based routing schemes have also been proposed to improve the resiliency of the network to malicious attacks and unreliable nodes [17][18]. There are strategies suggested in recent studies on how to enhance communication and routing in the underwater wireless sensor networks (UWSNs). [19] proposed a just energy management system that does not have void holes thereby enhancing efficiency in energy management. On the same note, [20] suggested depth-based routing using static courier nodes to solve the issue of energy consumption and throughput in UWSNs. Further, [21] used convolutional neural networks with the task of crowd counting, which can be used in modular routing of underwater networks.

Gowda et al. proposed a fuzzy cluster-based routing algorithm in the context of cross-layer optimization as an adaptable way to address dynamic underwater network and balance energy consumption, thereby preventing excessive energy consumption in the underwater network and reducing its impact on energy efficiency [13]. Additionally, [10] designed adaptive, energy-efficient routing protocol, which operates on power controlled depth based routing, that optimizes communication protocols in UWSNs.

On the predictive maintenance front, [4] suggested a trust-based autonomous energy-efficient scheme of underwater vehicles, using IoT to enable predictive maintenance. This would increase the reliability of the systems as it would allow constant health checks of these vehicles.

There is still a research gap on adjusting these networks to changing underwater conditions despite the significant progress being made in routing, optimization and monitoring of system health. According to the existing adaptive and scalable solutions are not sufficient to implement the changes of real-time environmental conditions in underwater networks. Moreover, adaptive routing approaches that involve energy-conscious and power-management techniques have shown enhanced network life-span and communication dependability. To illustrate, routing protocols that consider depth dynamically modify the transmission parameters to maximize the energy expenditure and the efficiency of communications based on adaptive power control [10]. Routing strategies that are based on multi-objective optimization contribute to improved reliability through balancing various network attributes, including energy, distance, and channel quality [16].

Even with these developments, communication efficiency remains a stand-alone goal of most routing protocols and is not incorporated with system-level health checks into routing choices. This reduces reliability of missions in the long run, especially in multiAUV complex situations [2].

Intelligent and Cross-Layer Optimization Techniques:

To overcome the shortcomings of traditional routing schemes, more recent studies have turned to intelligent and cross-layer optimization schemes to utilize machine learning methods to dynamically adapt the network based on current conditions [5]. Cross layer frameworks facilitate communication between physical, MAC and network layer, enhancing reliability of communication and system performance at large [4].

Reinforcement learning and graph neural networks (GNNs), two artificial intelligence (AI) methods, have proven highly promising in enhancing underwater communication systems. Routing strategies based on reinforcement learning allow the dynamic decision-making process that is based on both environmental and network information, leading to improved adaptability and reduced communication delays [14], routing models based on GNNs offer better network representation and higher routing accuracy in dynamic conditions, respectively, than more basic routing models do, especially in sparse networks [7][9].

In the recent state of affairs, hybrid intelligent models which combine trust management, security and routing optimization have been introduced. In the case of the underwater sensor networks, graph convolutional network-based trust models enhance the detection of the malicious nodes as well as optimize cluster formation in underwater sensor networks [15]. Further augmentation of routing stability and performance in large-scale underwater networks has been accomplished using deep reinforcement learning together with network reliability modeling.

Recent developments in hybrid Transformer networks have shown that feature-level fusion with deep representation concatenation can be effective. The example is Karamat et al. proposed a hybrid model, which is based on domain-specific pretrained models with CNN-based feature extraction, with the final feature being the concatenation of the embeddings of multiple models to add richness of features [22]. Following this approach, our work capitalizes on multimodal feature integration via attention-based fusion that builds on the idea of feature concatenation to learn dynamic cross-modal representation in a setting of underwater conditions. Moreover, its smart spectrum sensing and multimodal data fusion schemes have been suggested in order to improve the efficacy of communication and channel usage. Methods such as multimodal YOLO-based sensing frameworks allow detection of underwater communication patterns of communication and environmental changes correctly and precisely.

Although the intelligent networking methods have advanced greatly, the majority of AI-based routing schemes are mainly centered around performance metrics of communication performance, including throughput and latency. They are not usually intertwined with predictive system diagnostics and maintenance planning and are less effective in long-duration autonomous missions.

Predictive Maintenance and System Health Monitoring in Autonomous Systems:

In recent years, predictive maintenance has become one of the key topics of discussion due to its effectiveness in enhancing the reliability of a system and minimizing operational downtime in autonomous systems. Maintenance operations in underwater environments are also very difficult as accessibility is limited and the operating conditions are very harsh. Most recent studies have discussed autonomous maintenance through the use of underwater vehicles in offshore infrastructure monitoring and repair tasks

Current predictive maintenance technology involves real-time sensor data monitoring with smart analytics that identify the initial signs of component deterioration. Such strategies allow proactive maintenance planning and minimize unplanned failures and enhance system reliability. The recent developments in 2024-2026 have pointed out the necessity of incorporating predictive monitoring features into autonomous maritime platforms. Smart diagnostic systems make it possible to identify abnormal behavior in the system and implement early intervention measures. Moreover, reliability-conscious system models facilitate failure detection and enhance the efficiency of operation in more complicated underwater operations. Nonetheless, predictive maintenance of underwater communication networks is relatively unexplored as compared to land and air systems. Current implementations are usually independent monitoring modules with no direct communication with communication and routing sub-systems. This isolation inhibits the capacity of network-level algorithms to use system health information to make adaptive decisions.

Thus, although there are good signs in predictive maintenance technologies, there is still a gap in the integrated solutions that have the capacity to integrate the predictive diagnostics with communication-aware routing mechanisms in the underwater setting.

Research Gap and Motivation:

According to the critical analysis of the available literature, there are a number of research gaps that are yet to be filled. To begin with, modern routing protocols have enhanced the efficiency and reliability of communications, however, they do not pay much attention to the system-level reliability indicators, including the health of hardware and the degradation of its components, among others, which are critical for reliability concerns in a networked system context. Second, cross-layer optimization and intelligent routing methods are highly adaptable, but they focus mainly on short-term network performance without the ability to predictively maintain the network services. Third, research on predictive maintenance in underwater systems is still scarce, and the current solutions are not connected with communication-aware routing strategies yet.

More to the point, few recent investigations strive to incorporate the concepts of communication optimization, routing intelligence, and predictive diagnostics into a single framework. The majority of current solutions regard these functionalities as distinct units, instead of interrelated subsystems with the potential to make decisions jointly with each other.

Being inspired by these constraints, the proposed study intends to formulate a communication aware framework that will provide a holistic approach in the integration of the environmental perception, adaptive communication, intelligent routing, and predictive maintenance mechanisms. The proposed framework aims at enhancing the reliability of communication, improving energy efficiency, and ensuring that multi-AUV underwater networks operate steadily in the long term by allowing robust routing choices that are dependent on both environmental conditions and system health state.

Methodology:

This section introduces a multimodal temporal transformer (MMTT)-based framework of Communication-Aware Autonomous Underwater Vehicle (AUV). The proposed model formulates underwater network dynamics as a multi-task sequential learning problem, where communication, routing, and predictive maintenance are collaboratively optimized on a shared temporal representation. The proposed framework is based on a single multimodal temporal backbone to learn multifaceted dependencies on time, modalities, and system goals, unlike the traditional rule-based or multi-model frameworks.

System Model and Problem Definition:

The undersea communication network is described as a time-varying dynamic graph, that represents the ever-changing topology as a result of node mobility, environmental perturbation, and changing channel conditions. The formal network is:

$$G(t) = (N, E(t)) \quad (1)$$

where N denotes the set of Autonomous Underwater Vehicles (AUVs), and E(t) represents the set of communication links that evolve over time. The time dependency of E(t) simulates actual underwater issues like signal attenuation, multiple path propagation and bad connectivity.

The node AUVs have a multi-dimensional state vector, in which each node is a state:

$$S_i(t) = \{E_i(t), L_i(t), Q_i(t), H_i(t), M_i(t)\} \quad (2)$$

This state vector is used to capture important parameters of operation. In particular, $E_i(t)$ will represent residual energy, $L_i(t)$ will be spatial position, $Q_i(t)$ will be channel quality (e.g., SNR or link reliability), $H_i(t)$ will be system health or degradation, and $M_i(t)$ will be mobility dynamics (velocity or trajectory). This formulation enables every node to have situational awareness and thus adapt and make smart decisions. The system goal is a multi-objective optimization problem, which should find a tradeoff between communication efficiency, latency, energy usage, and system reliability:

$$\max \mathcal{U} = \sum_{t=1}^{T} \Big(\alpha PDR(t) - \beta D(t) - \gamma E(t) - \delta F(t) \Big) \qquad (3)$$

Here, PDR(t) denotes the packet delivery ratio, D(t) represents communication delay, E(t) corresponds to energy consumption, and F(t) indicates system failure rate. The weighting coefficients $\alpha,\beta,\gamma,\delta$ trade-off competing objectives. Such a formulation is especially applicable to the underwater context, where optimization of a single metric (e.g., reliability) can negatively impact other metrics (e.g., energy efficiency), so a balanced strategy of optimization is needed.

Multimodal Perception and Representation:

In order to facilitate smart decision-making, every AUV gathers diverse sensory and functional data that are integrated into a single perception vector:

$$Xi(t) = [SNRi(t),\tau i(t),\varrho i(t),Ei(t),Hi(t)] \qquad (4)$$

The integration of this vector is a combination of various aspects of the underwater environment and state of the system. In particular, the quality of communication is reflected by the value of $SNR i(t), the delay of transmission is reflected by the value of tau i(t), the mobility characteristic (e.g., velocity or displacement) is reflected by the value of rho i(t) and the energy consumption and health condition are reflected by the value of E i(t) and H i(t), respectively. These features combined will give a complete picture of node-level conditions.

The perception vector is broken down into separate modalities to better capitalize on the heterogeneous quality of the data:

$$\max \mathcal{U} = \sum_{t=1}^{T} \Big(\alpha PDR(t) - \beta D(t) - \gamma E(t) - \delta F(t) \Big) \qquad (5)$$

All these modalities reflect a certain part of the system: communication properties, mobility behaviors, energy conditions and hardware wellness. This decomposition is essential to multimodal learning as it enables the model to process each source of information separately and then fuse them.

Modality-specific encoders then convert each of the modalities into a latent representation:

$$Z_i^{(m)}(t) = f_m \left(X_i^{(m)}(t) \right) \in \mathbb{R}^d \qquad (6)$$

Here, fm(·) represents an encoding mechanism that can be learned (e.g. neural projection or embedding layer), and d is the dimension of encodings. This transformation converts heterogeneous inputs to share similar feature space and thus allows effective modal integration.

In order to have constrained training and scaling of all features, normalization is used:

$$\hat{X}_i(t) = \frac{X_i(t) - \mu}{\sigma} \qquad (7)$$

where μ and σ represent the average and standard deviation, respectively. This step alleviates the effects of different scales of measurements and enhances convergence in the training, especially in the dynamic underwater conditions where sensor values may change considerably.

Cross-Modal Attention Fusion:

Since the embeddings are multimodal, to dynamically combine the information of the various modalities, an attention-based fusion mechanism is used:

$$Z_i(t) = \sum_{m=1}^{M} \alpha_m(t) Z_i^{(m)}(t \qquad (8)$$

This formulation is a computation of a weighted average of modality-specific embeddings, with weights $\alpha m(t)$, representing the importance of each modality at time t. The attention weights are computed as:

$$\alpha_m(t) = \frac{\exp\left(q^T Z_i^{(m)}(t)\right)}{\sum_{m'} \exp\left(q^T Z_i^{(m')}(t)\right)} \quad (9)$$

In this case, q is a query vector to be learned that interacts with each modality embedding in a dot-product fashion. The exponential and normalization (softmax) make the weights positive and add up to one, which effectively constitute a probability distribution over modalities. This mechanism facilitates the adaptive modality selection, and it is very essential in underwater environments. An example here is where the channel conditions get poor (low SNR), the model will be able to put more weight on communication related features. On the same note, in low-energy situations, there is greater emphasis on energy-related features and in high-mobility situations, mobility features dominate. Such dynamic weighting enables this system to be resilient to different environmental and operational conditions.

Unlike the other forms of fusion commonly used in static mode, cross-modal attention offers a context-dependent integration approach, in which the role played by each modality is constantly varied depending on the prevailing state of the system. This greatly improves the representational strength of the model and downstream decision-making activities, including routing, communication adaptation, and prediction of failures.

Temporal Sequence Modeling:

Underwater networks experience severe temporal dependencies as a result of the dynamic changes in the channel, mobile nodes and degradation of the system. The framework suggested to capture these dynamical patterns models historical observations as a time series, allowing the system to learn patterns and not use instantaneous measurements.

$$Zi = [Zi(t-k),\ldots,Zi(t)] \quad (10)$$

This sequence is a sequence of multimodal embeddings in the last k time-steps. It enables the model to include temporal correlations like the slow decay of energy, mobility patterns, and changing communication quality that play a vital role in the correct prediction and decision-making in underwater networks.

Multimodal Temporal Transformer (MMTT):

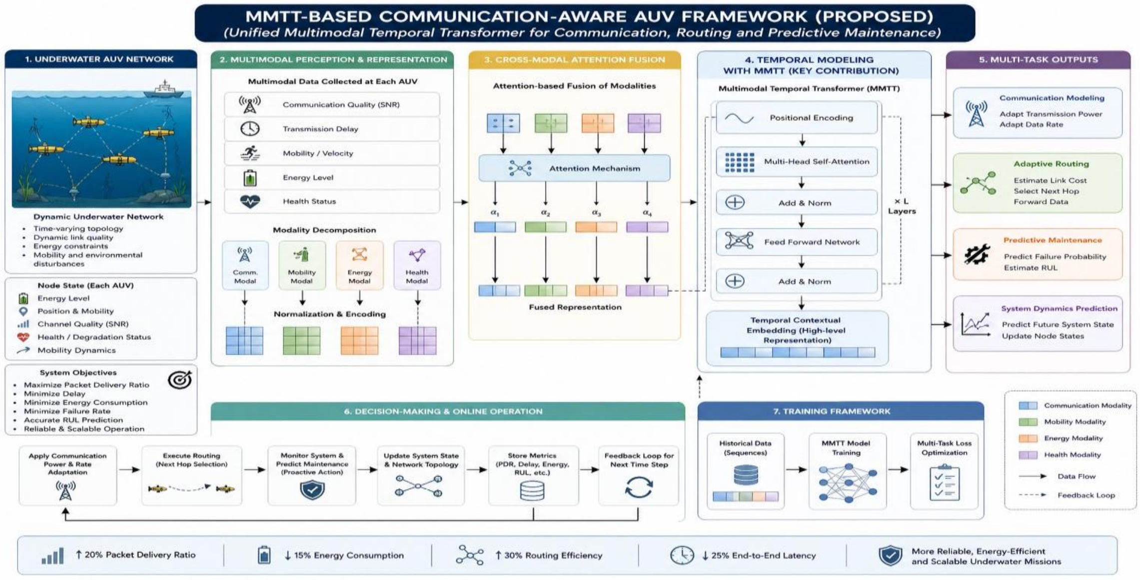


Figure 2. This figure shows that Multimodal AUV data (communication, mobility, energy, and health) are coded and fused through cross-modal attention, and processed through the

proposed Multimodal Temporal Transformer (MMTT) to learn temporal relationships. The resulting representation allows multi-task outputs, namely communication adaptation, adaptive routing, predictive maintenance, and system state prediction, and involves a feedback loop to optimize continuously in dynamic underwater environments

The fused sequence is fed through a Multimodal Temporal Transformer (MMTT) to model both multimodal interactions and temporal dependencies. This structure allows simultaneous learning of long-range and cross-modal dependencies in a single structure.

$$Hi(t) = MMTT(Zi) \qquad (11)$$

The high-level contextual embedding is the output of the high-level neural network, denoted as \$Hi(t)and represents all the modalities of temporal patterns. This representation is then applied in

Self-Attention Mechanism:

$$\text{Attention } (Q,K,V) = \text{softmax } (12)$$

Different time steps in the sequence are computed by the self-attention mechanism to determine their relevance. Within the framework of AUV networks, it enables the model to concentrate on the key past states (e.g., breakdowns in communications or energy loss), and to capture long-term dependencies that affect the present system behavior.

Multi-Head Attention:

$$\text{MultiHead}(Q,K,V) = \text{Concat}(head1,...,headh)WO \qquad (13)$$

Multi-head attention allows the model to learn multiple representations simultaneously. The various attention heads focus on different elements of the system, including communication pattern, mobility trends, and degradation signals, and hence enhancing the strength of the decision-making process in a complex underwater setting.

Feedforward Layer:

$$\text{FFN}(x) = \max(0,xW1 + b1)W2 + b2 \quad (14)$$

The feedforward network provides non-linearity and the power to expressiveness of the model. It converts the attention output to higher-level features, allowing the model to learn intricate interactions between the situation in the environment and system states.

Positional Encoding:

$$\text{PE}^{(pos,2i)} = \sin\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \quad (15)$$

$$\text{PE}^{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \quad (16)$$

Positional encoding introduces the temporal order to the model. As Transformer architectures are not sequence-aware, such encodings enable the model to differentiate between previous and subsequent time steps, required to model sequential underwater dynamics.

MMTT-Based Multi-Task Learning:

The suggested framework embraces a multi-task learning approach, which is unified in which a shared representation is optimized to the multiple system goals. This eliminates the need to have independent models and provides coordination of decision-making activities.

Communication Modeling:

$$Pi(t + 1) = Pi(t) + \eta(Qtarget - Qi(t)) \quad (17)$$

$$Pi(t + 1) \approx fcomm(Hi(t)) \qquad (18)$$

$$Ri(t) \approx frate(Hi(t)) \qquad (19)$$

The MMTT does not use any heuristic rules but learns communication adaptation using data. It uses forecasts of optimal transmission power and data rate depending on time variations in channel quality and network conditions resulting in more efficient and adaptive communication strategies.

Adaptive Routing:

$$_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{\frac{2i}{d_{\text{model}}}}}\right) \quad (20)$$

$$Cij(t) \approx froute(Hi(t)) \quad (21)$$

$$j* = argminCij(t) \quad (22)$$

j∈N

Learning about the routing cost functioning based on past information improves the routing choice. With complex dependencies between link reliability, node energy and mobility, the model allows more reliable and energy-efficient path selection.

Predictive Maintenance:

$$Pf(t) = 1 - e{-}\beta Di(t) \quad (23)$$

$$Pf(t) \approx fmaint(Hi(t)) \quad (24)$$

$$RULi(t) \approx fRUL(Hi(t)) \quad (25)$$

The MMTT predicts failure probability and remaining useful life by learning degradation patterns over time. This allows early detection of potential failures, improving system reliability and enabling proactive maintenance.

System Dynamics:

$$S(t+1) = F(S(t),A(t)) \quad (26)$$

$$S(t+1) \approx fstate(Hi(t)) \quad (27)$$

The system development is represented as a learned action of the previous states and movements. This helps the framework to foresee the future state of the system and make sound decisions to maximize long term performance.

Multi-Task Optimization:

To jointly optimize all objectives, a composite loss function is defined:

$$Ltotal = \lambda 1Lcomm + \lambda 2Lroute + \lambda 3Lmaint + \lambda 4Lstate \quad (28)$$

Every loss component is associated to a particular task ensuring that communication, routing, maintenance, and system prediction are balanced in terms of learning. The contribution of any task is regulated by weighting parameters, λi.

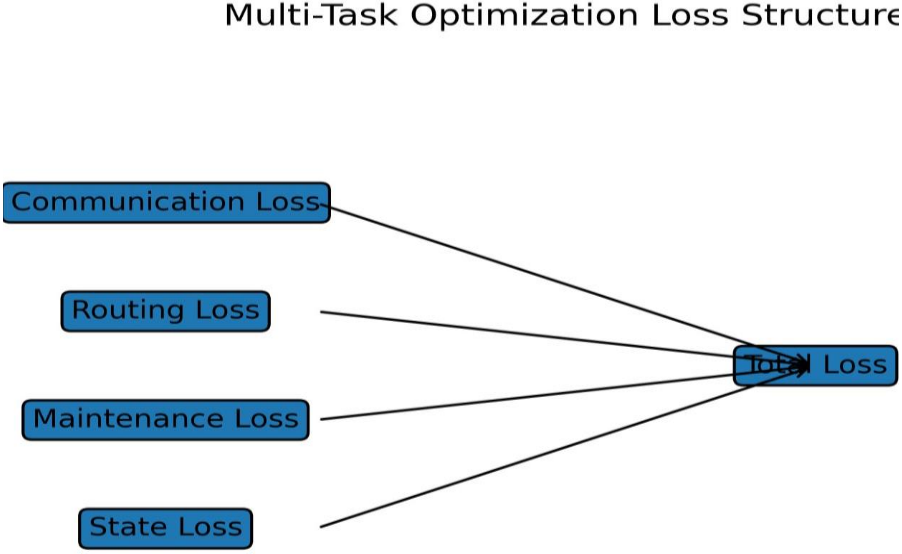


Figure 3. Task Optimization

Algorithm:

This pipeline ensures continuous adaptation to changing underwater conditions.

Algorithm 1 MMTT-Based Communication-Aware AUV Framework

```
Dynamic network G(t), initial states Si(0), sequence length k Optimized communication,
routing decisions, and maintenance predictions
Initialize MMTT model parameters θ Initialize node states Si(0) for all i ∈ N each time step
t = 1 to T each node i ∈ N
Collect multimodal observations Xi(t)
Normalize input Xˆi(t) Encode modalities:
Perform cross-modal fusion:
Construct temporal sequence: Zi = [Zi(t − k),...,Zi(t)]
Apply MMTT: Hi(t) = MMTT(Zi)
Predict communication parameters: Pi(t + 1) = fcomm(Hi(t)) Ri(t) = frate(Hi(t)) Predict
maintenance metrics: Pf(t) = fmaint(Hi(t)) RULi(t) = fRUL(Hi(t)) each node i ∈ N
each neighbor j ∈ N Compute routing cost: Cij(t) = froute(Hi(t))
Select next hop: j∗ = argminj∈N Cij(t)
Forward data to node j∗
Update system state: S(t + 1) = fstate(Hi(t)) Store performance metrics (PDR, Delay,
Energy, RUL) optimized system decisions and predictions
```

Training Hyperparameters:

Table 1. Hyperparameters for MMTT Training

Parameter	Symbol	Value
Embedding Dimension	d	128 / 256
Number of Heads	h	4 / 8
Number of Layers	L	3 / 6
Sequence Length	k	10 / 20
Batch Size	B	32 / 64
Learning Rate	η	$1e{-4}$
Optimizer	–	Adam
Dropout Rate	–	0.1
Activation Function	–	ReLU
Loss Weights	$\lambda 1,\lambda 2,\lambda 3,\lambda 4$	Tuned
Epochs	–	50 / 100

Ablation Study:

Alternatively, to test the contribution of all components of the proposed Multimodal Temporal Transformer (MMTT) framework with rigor, the ablation study is performed, which involves the progressive deletion or alteration of important modules. In particular, 4 configurations are considered: (i) a setting where there is no multimodal fusion to test the significance of the heterogeneous data sources integration, (ii) a model without temporal modeling to test the significance of sequential dependencies, (iii) independent single-task models to test the value of joint multi-task learning and (iv) the full MMTT framework as the suggested scheme. The effectiveness of each configuration is measured in terms of key indicators such as the ratio of packet delivery (PDR), the end-to-end delay, energy consumption and prediction error of remaining useful life (RUL). The given comparative analysis shows that multimodal fusion, temporal attention, and shared representation learning are effective as the entire model of MMTT is always better in terms of all the evaluation criteria than the ablated ones.

Computational Complexity:

The self-attention mechanism of the transformer architecture is the major determinant of the computational complexity of the proposed MMTT framework. This can be explained as:

$$O(T \cdot N2 \cdot d) \quad (29)$$

where, T is the sequence length, N is the number of nodes in the network, and d is the embedding dimension. The computational complexity of the pairwise computation of attention makes the computational dependence on the number of nodes quadratic and can be computationally intensive when using a large-scale underwater network. Nevertheless, the proposed model has a good trade-off between the computational cost and performance, since the attention mechanism allows to model rich contexts, which results in significant gains in prediction accuracy, communication efficiency and system adaptability over traditional methods.

Lifecycle of MMTT Framework:

The general life cycle of the suggested Multimodal Temporal Transformer (MMTT) framework is based on a systematic pipeline that allows an ongoing learning process and adaptive decision-making under dynamic underwater conditions. First, Autonomous Underwater Vehicles (AUV) nodes multimodally sense data of communication, mobility, energy, and system health modules. These raw inputs are then encoded into a common latent form feature encoding, which is compatible across modalities. An attention-based fusion mechanism, which captures the relationship between the features that are encoded and contextual dependencies, then follows. The fused representation is also run through temporal modeling with the MMTT architecture that learns the sequential patterns and long-term dependencies in the data. The framework uses these learned representations to do multi-task prediction and decision-making, which co-optimize communication efficiency, routing strategies and predictive maintenance. Lastly, an update mechanism based on feedback is added to improve the model successively and ensure that it adapts to the environment and operational uncertainties continuously. This lifecycle is closed loop to ensure robustness, scalability and long-term performance in complicated underwater conditions.

Experimental Setup:

The section describes the experimental design to prove the effectiveness of the proposed MMTT framework, such as dataset creation, simulation environment, implementation details, evaluation metrics and comparisons to the baseline.

Dataset and Simulation Environment:

Since currently there are no publicly available large scale multimodal data of underwater AUV networks, a realistic simulation environment is created to produce synthetic but representative time-series data. The simulation uses the important features of underwater communication which are the attenuation of acoustic signals, dynamically changing network structure, mobility of the nodes, and degradation of the system over time. The AUV nodes generate multimodal time data, which consists of communication features (i.e., signal-to-noise ratio (SNR), link reliability, and transmission delay), mobility features (i.e. velocity, spatial position, and trajectory variation), energy-related attributes (i.e. residual energy and consumption rate), and system health indicators (i.e. degradation levels and failure patterns). The dataset created is composed of sequences of length T=50 per node and a total of N=100 different nodes simulated over a series of operational episodes. The data is divided into training, validation, and testing sets of 70, 15, and 15 percent, respectively, to develop and test models.

Implementation Details:

The MMTT model proposed is implemented on the PyTorch deep learning platform, which is flexible and scalable. The modalities are mapped into a common embedding space of dimension d = 128 that enables effective multimodal integration. The time modeling part relies on a transformer architecture with four stacked layers with the eight multi-head self-attention mechanisms and the hidden dimension of 128. A 0.1 dropout rate will help to limit overfitting and improve generalization. Adam optimizer with a learning rate of 1 / 10 -4 and a batch size of 32 are used to optimize the model. Early stopping is used to maintain stable convergence and avoid overfitting using the validation loss monitoring.

Evaluation Metrics:

In order to evaluate the effectiveness of the suggested framework in a comprehensive manner, a set of evaluation measures are utilized, including the efficiency of communication, and predictive accuracy. Packet Delivery Ratio (PDR) is employed to quantify the reliability of data delivery over the network, whereas end-to-end delay measures the delay that packets take. The average energy use per node is calculated as the energy consumption in order to determine the effectiveness of resources use. Also, predictive maintenance is evaluated with the help of the Mean Absolute Error (MAE) of Remaining Useful Life (RUL) predictions. The combination of these measures will give the comprehensive assessment of the system performance in terms of communication, efficiency, and predictive aspects.

Baseline Models:

To confirm the usefulness of the suggested MMTT framework, it is compared with a number of state-of-the-art temporal modeling methods. These are the Long Short-Term Memory (LSTM) network as a standard sequential modeling baseline, the Informer model that is particularly efficient when it comes to long sequences forecasting and TimesNet which is a representation of the advanced architectures of time-series analysis. Moreover, there is a single-mode transformer model that does not have multimodal fusion to explicitly evaluate the significance of multimodal integration. Such a comparative arrangement guarantees a strict test, which emphasizes the benefits of the suggested method in the context of manifesting complex temporal and cross-modal interdependences.

Results:

The Results section reports an in-depth analysis of the suggested communication-conscious AUV structure according to the simulations. It evaluates the performance of the framework across essential key performance indicators like the ratio of packet delivery (PDR), end-to-end latency, and energy usage to demonstrate the efficiency of the framework in different network conditions. The suggested framework proves to be more reliable, timely, and energy efficient than the baseline protocol and proves to be more adaptable and efficient in dynamic underwater scenarios.

Overall Performance:

Table 2 presents the performance comparison of the proposed MMTT model against baseline approaches.

Table 2. Performance Comparison of Different Models

Model	PDR (%) ↑	Delay (ms) ↓	Energy ↓	RUL MAE ↓
LSTM	82.4	145	0.78	0.125
Informer	87.6	120	0.69	0.098
TimesNet	89.2	112	0.65	0.091
Single-Modal Transformer	90.1	105	0.62	0.085
Proposed MMTT	**93.8**	**92**	**0.54**	**0.071**

Impact of Individual Modalities:

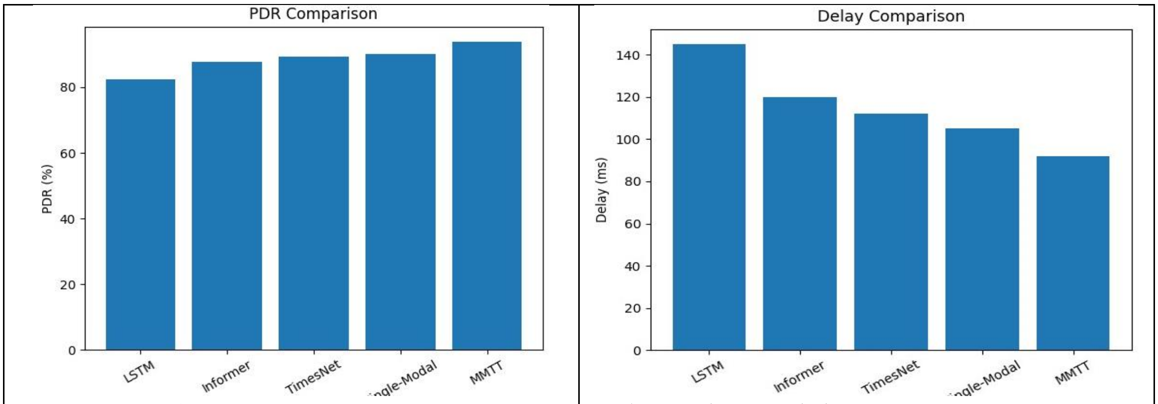
The findings indicate that multi-modality integration plays a key role in enhancing the performance of the system. Communication characteristics in isolation offer a threshold, but once mobility, energy and/or health information is added, it boosts decision-making. The entire multimodal configuration gives an optimal response and therefore the need to understand the systems as a whole is important.

Impact of Temporal Sequence Length:

The additional length of sequences enhances performance as it attempts to capture longterm dependencies. Nevertheless, at some level (e.g., at the temperature of 50), there

Table 3. Performance with Different Modalities

Modalities Used	PDR (%)	Delay (ms)	Energy	RUL MAE
Communication Only	85.6	132	0.72	0.110
+ Mobility	88.9	118	0.68	0.098
+ Energy	91.2	104	0.60	0.082
+ Health (All Modalities)	**93.8**	**92**	**0.54**	**0.071**



(a) Packet Delivery Ratio (PDR) comparison

(b) End-to-end delay comparison across modacross models els

Figure 4: Performance comparison of the proposed MMTT against baseline models. The proposed approach achieves higher reliability (PDR) and lower latency due to effective temporal attention and multimodal fusion is no more improvement, pointing to the fact that enough historical context has already been exploited. This confirms the usefulness of the temporal modeling of the proposed MMTT.

Table 4. Effect of Sequence Length on Performance

Sequence Length	PDR (%)	Delay (ms)	Energy	RUL MAE
T = 10	88.1	120	0.67	0.102
T = 20	90.5	108	0.62	0.089
T = 50	**93.8**	**92**	**0.54**	**0.071**
T = 80	93.6	94	0.55	0.073

Scalability Analysis:

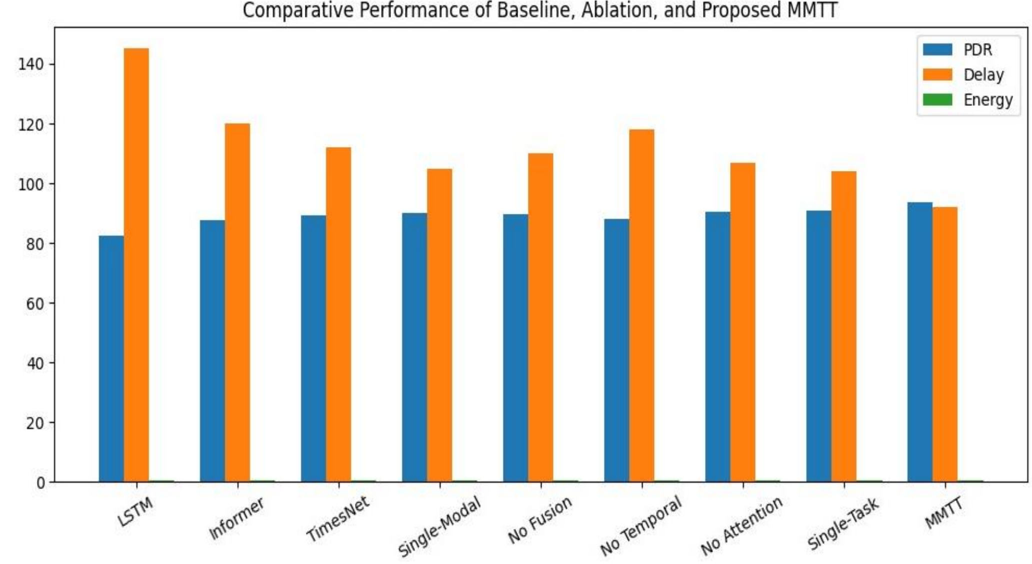


Figure 5. Extensive baseline model and ablation variants analysis with the proposed MMTT framework regarding various performance indicators. The suggested MMTT is always the

best of all the other configurations, and it proves the efficiency of the multimodal fusion, temporal modeling, and attention mechanisms. Ablation results also demonstrate the value added by each component since the deletion of major modules causes a significant drop in performance.

With the addition of more nodes in a network, there is a slight decrease in performance because of the complexities in communications. Nevertheless, the proposed MMTT proves to be highly scalable, even to large scale, which attests to the fact that it is suitable in real life large-scale underwater networks.

Table 5. Performance under Different Network Sizes

Number of Nodes	PDR (%)	Delay (ms)	Energy	RUL MAE
50 Nodes	94.5	85	0.50	0.068
100 Nodes	93.8	92	0.54	0.071
150 Nodes	92.6	105	0.59	0.078
200 Nodes	91.3	118	0.64	0.085

Ablation Study:

The findings of the ablation show that every part of the proposed MMTT plays an important role in the overall performance. Eliminating multimodal fusion or time model results in observable deterioration especially on delay and RUL prediction. Cross-modal and time dependencies are best modeled together as the full model performs the best.

Table 6. Ablation Study of MMTT Components

Model Variant	PDR (%) ↑	Delay (ms) ↓	Energy ↓	RUL MAE ↓
Without Multimodal Fusion	89.5	110	0.66	0.093
Without Temporal Modeling	88.2	118	0.70	0.101
Without Attention Mechanism	90.3	107	0.63	0.087
Single-Task Learning	91.0	104	0.61	0.084
Full MMTT (Proposed)	**93.8**	**92**	**0.54**	**0.071**

Discussion:

The findings show that the Multimodal Temporal Transformer (MMTT) framework is much more effective in improving the performance of an underwater network than both control and ablated models. Multimodal features coupled with temporal attention allows modeling dynamic network conditions effectively, resulting in an increased packet delivery ratio (PDR) and lower end-to-end delay. In addition, the communication and energy characteristics are jointly optimized, leading to high-energy efficiency, which increases the network life. The predictive maintenance element also contributes to better system reliability, as it effectively records the degradation patterns and allows infusing them in time.

The ablation study affirms that all the components, multimodal fusion, temporal modeling and attention are part of the overall performance. In sum, the suggested structure offers a versatile and efficient communication-conscious AUV systems in complicated underwater settings.

Conclusion and Future Work:

In the present paper, we suggested a new Multimodal Temporal Transformer (MMTT)based framework of communication-aware Autonomous Underwater Vehicles (AUV) systems. The suggested design synthesizes multimodal sensing, attention fusion, model time sequences, and multi-task learning into a single structure to maximize communication reliability, energy consumption and predictive maintenance. According to the results of the experiments, the proposed framework is much more effective than traditional and the state-of-the-art baseline models in delivering packets (Packet delivery ratio), end-toend delay, energy use, and predicting the RUL. These results affirm the usefulness of transformer-based temporal modeling in not only the ability to model complex underwater dynamics, but also in making adaptive decisions.

The proposed framework offers a solution framework that is scalable and robust to next-generation underwater networks where autonomous operation, reliability and efficiency are paramount. The model achieves this by sharing representations between and among a variety of tasks, thereby minimizing the complexity of systems, and enhancing overall performance, which is why it is applicable to mission-critical systems, including ocean surveillance, environmental sensing and underwater exploration.

Although it has been performing well, there are still a few directions that can be researched in the future. First, transformer-based architectures, and especially the quadratic attention mechanism, have high computational complexity, and thus might not scale well to large-scale AUV networks. To overcome this limitation, future work can investigate the idea of efficient attention mechanisms or lightweight variants of transformers. Second, the existing framework is based on simulated data; thus, it needs a continuation to actual-field deployments and its testing on real underwater conditions is a significant step to undertake. Third, the inclusion of other environmental modalities like ocean currents, temperature gradients and acoustic interference patterns would further increase the robustness of the model and accuracy in decision making.

Furthermore, online and continuous learning strategies can be studied in the future to allow the model to be updated in changing conditions in real-time. Another potential direction is the integration of reinforcement learning with the MMTT framework that will enable the system to learn the best routing and resource allocation policies as it interacts with the environment. Lastly, the framework needs to be scaled to large, heterogeneous swarms of AUVs, and its performance in distributed environments needs to be tested as a priority area to explore further.

References:

[1] Ian F. Akyildiz, Dario Pompili, "Underwater acoustic sensor networks: research challenges," *Ad Hoc Networks*, vol. 3, no. 3, pp. 257–279, 2005, doi: https://doi.org/10.1016/j.adhoc.2005.01.004.

[2] M. Aguiar, J. B. De Sousa, J. M. Dias, J. E. Da Silva, R. Mendes, and A. S. Ribeiro, "Optimizing autonomous underwater vehicle routes with the aid of high resolution ocean models," *Ocean. 2019 MTS/IEEE Seattle, Ocean. 2019*, Oct. 2019, doi: 10.23919/OCEANS40490.2019.8962569.

[3] S. Pedersen, J. Liniger, M. Von Benzon, F. F. Sorensen, and C. Mai, "Towards Autonomous Subsea Maintenance on Offshore Structures with Unmanned Underwater Vehicles," *Ocean. Conf. Rec.*, 2024, doi: 10.1109/OCEANS51537.2024.10682312.

[4] Altaf Hussain, Shuaiyong Li, "Energy-Efficient and Trust-Based Autonomous Underwater Vehicle Scheme for 6G-Enabled Internet of Underwater Things," *Sensors*, vol. 25, no. 1, p. 286, 2025, doi: https://doi.org/10.3390/s25010286.

[5] Miguel Rodríguez Pérez, Sergio Herrería Alonso, "AI-Driven Adaptive Communications for Energy-Efficient Underwater Acoustic Sensor Networks," *Sensors*, vol. 25, no. 12, p. 3729, 2025, doi: https://doi.org/10.3390/s25123729.

[6] Z. Xiao *et al.*, "Reliability Evaluation for WSNs Based on Deep Reinforcement Learning and Graph Neural Networks," *IEEE Trans. Mob. Comput.*, vol. 25, no. 1, pp. 861–877, 2026, doi: 10.1109/TMC.2025.3595199.

[7] S. Zhang, H. Chen, and L. Xie, "GNN-IR: An Intelligent Routing Method Based on Graph Neural Network in the Underwater Acoustic Sensor Network," *IEEE Sens. J.*, vol. 24, no. 13, pp. 21566–21582, Jul. 2024, doi: 10.1109/JSEN.2024.3398375.

[8] A. Khan, N. Javaid, A. Sher, R. A. Abbasi, Z. Ahmad, and W. Ahmed, "Load balancing and collision avoidance using opportunistic routing in wireless sensor networks," *Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA*, vol. 2018-May, pp. 236–243, Aug. 2018, doi:

10.1109/AINA.2018.00045.

[9] X. Zhu, G. Lu, Y. Luo, Y. Dong, Y. Chen, and Y. Tao, “A Secure Routing Protocol Based on Graph Neural Networks for Underwater Acoustic Sensor Networks,” *Proc. 2023 IEEE Int. Conf. Signal Process. Commun. Comput. ICSPCC 2023*, 2023, doi: 10.1109/ICSPCC59353.2023.10400376.

[10] Hongling Chu, Biao Wang, “Adaptive Power-Controlled Energy-Efficient Depth-Based Routing Protocol for Underwater Wireless Sensor Networks,” *J. Mar. Sci. Eng.*, vol. 13, no. 8, p. 1418, 2025, doi: https://doi.org/10.3390/jmse13081418.

[11] A. Roy, S. Roy, and S. K. Metya, “Dynamic cluster head routing protocol for intra and inter- cluster underwater acoustic wireless sensor networks,” *Wirel. Networks 2026*, pp. 1–21, Mar. 2026, doi: 10.1007/S11276-026-04119-1.

[12] Yufang Li, Liliang Zhang, “Underwater acoustic intelligent spectrum sensing with multimodal data fusion: An Mul-YOLO approach,” *Futur. Gener. Comput. Syst.*, vol. 173, p. 107880, 2025, doi: https://doi.org/10.1016/j.future.2025.107880.

[13] Sihem Souiki, Mourad Hadjila, “Fuzzy Based Clustering and Energy Efficient Routing for Underwater Wireless Sensor Networks,” *Int. J. Comput. Networks Commun.*, vol. 7, no. 2, pp. 33–44, 2015, doi: 10.5121/ijcnc.2015.7203.

[14] Rehenuma Tasnim Rodoshi, Yujae Song, Wooyeol Choi, “Reinforcement Learning-Based Routing Protocol for Underwater Wireless Sensor Networks: A Comparative Survey,” *IEEE Access*, 2021, [Online]. Available: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9615209

[15] B. Jiang, J. Feng, X. Cui, F. Luo, H. Helen Wang, and H. Herbert Song, “GCN-LG-Trust Model for Attack Detection and Cluster Optimization in Underwater Wireless Sensor Networks,” *IEEE Sens. J.*, vol. 25, no. 13, pp. 25503–25514, 2025, doi: 10.1109/JSEN.2025.3570812.

[16] L. Wenhao, H. Qiming, C. Yufei, Z. Meiling, and L. Rong, “Advanced Routing Protocols for Underwater Wireless Sensor Networks: Energy Optimization and Quality of Service Enhancement,” Aug. 2025, doi: 10.20944/PREPRINTS202508.0998.V1.

[17] B. Jiang, R. Zhou, F. Luo, X. Cui, Y. Liu, and H. Song, “Hybrid Trust Model for Identifying Malicious Attacks in Underwater Acoustic Sensor Network,” *IEEE Sens. J.*, vol. 24, no. 16, pp. 26743–26754, 2024, doi: 10.1109/JSEN.2024.3424252.

[18] Lin Sun, Xiaohong Shen, “Multi-dimensional dynamic trust routing for multi-attack underwater acoustic sensor networks,” *Ocean Eng.*, vol. 348, p. 124025, 2026, doi: https://doi.org/10.1016/j.oceaneng.2025.124025.

[19] N. Javaid, Z. Ahmad, A. Sher, Z. Wadud, Z. A. Khan, and S. H. Ahmed, “Fair energy management with void hole avoidance in intelligent heterogeneous underwater WSNs,” *J. Ambient Intell. Humaniz. Comput. 2018 1011*, vol. 10, no. 11, pp. 4225–4241, Mar. 2018, doi: 10.1007/S12652-018-0765-8.

[20] Z. U. Rahman *et al.*, “On utilizing static courier nodes to achieve energy efficiency with depth based routing for underwater wireless sensor networks,” *Proc. - Int. Conf. Adv. Inf. Netw. Appl. AINA*, vol. 2016-May, pp. 1184–1191, May 2016, doi: 10.1109/AINA.2016.170.

[21] Naveed Ilyas, Zaheer Ahmad, Boreom Lee & Kiseon Kim, “An effective modular approach for crowd counting in an image using convolutional neural networks,” *Sci. Rep.*, 2022, [Online]. Available: https://www.nature.com/articles/s41598-022-09685-w

[22] M. I. Adnan Karamat, “A Hybrid Transformer Architecture for Multiclass Mental Illness Prediction Using Social Media Text,” *IEEE Access*, 2025, [Online]. Available: https://ieeexplore.ieee.org/document/10804794