Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.076514

ARTICLE

Q-ALIGNer: A Quantum Entanglement-Driven Multimodal Framework for Robust Fake News Detection

Sara Tehsin[1,*], Inzamam Mashood Nasir[1], Wiem Abdelbaki[2], Fadwa Alrowais[3], Reham Abualhamayel[4], Abdulsamad Ebrahim Yahya[5] and Radwa Marzouk[6]

[1]Faculty of Informatics, Kaunas University of Technology, Kaunas, Lithuania
[2]College of Engineering and Technology, American University of the Middle East, Egaila, Kuwait
[3]Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia
[4]Department of Information Systems, Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia
[5]Department of Information Technology, College of Computing and Information Technology, Northern Border University, Arar, Saudi Arabia
[6]Department of Information Systems, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia

*Corresponding Author: Sara Tehsin. Email: sara.tehsin@ktu.edu

Received: 21 November 2025; Accepted: 07 January 2026; Published: 12 March 2026

ABSTRACT: The rapid proliferation of multimodal misinformation on social media demands detection frameworks that are not only accurate but also robust to noise, adversarial manipulation, and semantic inconsistency between modalities. Existing multimodal fake news detection approaches often rely on deterministic fusion strategies, which limits their ability to model uncertainty and complex cross-modal dependencies. To address these challenges, we propose Q-ALIGNer, a quantum-inspired multimodal framework that integrates classical feature extraction with quantum state encoding, learnable cross-modal entanglement, and robustness-aware training objectives. The proposed framework adopts quantum formalism as a representational abstraction, enabling probabilistic modeling of multimodal alignment while remaining fully executable on classical hardware. Q-ALIGNer is evaluated on four widely used benchmark datasets—FakeNewsNet, Fakeddit, Weibo, and MediaEval VMU—covering diverse platforms, languages, and content characteristics. Experimental results demonstrate consistent performance improvements over strong text-only, vision-only, multimodal, and quantum-inspired baselines, including BERT, RoBERTa, XLNet, ResNet, EfficientNet, ViT, Multimodal-BERT, ViLBERT, and QEMF. Q-ALIGNer achieves accuracies of 91.2%, 92.9%, 91.7%, and 92.1% on FakeNewsNet, Fakeddit, Weibo, and MediaEval VMU, respectively, with F1-score gains of 3–4 percentage points over QEMF. Robustness evaluation shows a reduced adversarial accuracy gap of 2.6%, compared to 7%–9% for baseline models, while calibration analysis indicates improved reliability with an expected calibration error of 0.031. In addition, computational analysis shows that Q-ALIGNer reduces training time to 19.6 h compared to 48.2 h for QEMF at a comparable parameter scale. These results indicate that quantum-inspired alignment and entanglement can enhance robustness, uncertainty awareness, and efficiency in multimodal fake news detection, positioning Q-ALIGNer as a principled and practical content-centric framework for misinformation analysis.

KEYWORDS: Machine learning; fake news detection; multimodal learning; quantum natural language processing; cross-modal entanglement; adversarial robustness; uncertainty calibration

1 Introduction

The rapid expansion of social media platforms has increased the dissemination of multimodal misinformation—that is, a message that includes a textual claim that is often supported by an altered image or a false image. This trend threatens public trust and presents severe challenges to information certainty, political stability, and societal choice-making. Previously developed fake news detection methods that were based solely on textual [1,2] or visual [3] features have shown to be ineffective, mostly because they did not account for even some interdependence of modalities. The recent multimodal approaches [4–6] that incorporated deep learning to fuse heterogenous signals still struggle for robustness under adversarial perturbations, or in making calibrated and reliable confidence scores for the results. This affects its practicality of use in deployment scenarios where adversarial perturbations and noisy data are a possibility.

Recently, quantum-inspired learning has gained traction as a new and promising strategy for progressing multimodal representation learning. The integration of quantum state embedding and entanglement, models including QMFND [7] and HQDNN [8] exhibit better feature expressivity and robustness. However, such models strategize on fixed entanglement mechanisms and do not have methods for explicit alignment among modalities, resulting in sub-par performance in highly heterogenous misinformation environments. The combination of transformative methodology, quantum-inspired expressivity, adaptive alignment, and robustness-aware learning models is perceived as the change needed to over-come performance limitations. The paper's primary objective is to create a system capable of confirming the content of different media types through a model that offers a semantic alignment and robustness between text and images at the instance level. However, although the temporal dynamics, content propagation patterns, and social context are essential to thoroughly understand how misinformation spreads through the internet, these factors will generally not be considered until later detection phases since they represent signals unavailable at the early stages of detection. Thus, they will not be modeled in this research.

Recent advancements in the detection of multimodal fake news have failed to create a comprehensive approach to the simultaneous challenges of semantic disparity, adversarial robustness and uncertainty estimation. Previous multimodal fusion methods such as Attention, Concatenation and Late Fusion, utilise deterministic representations; therefore, those methods do not effectively model higher-order dependencies across modalities due to the presence of noise and ambiguity in the real world. Misinformation and the issues associated with it in the real world often include conflicting evidence, expert disagreement and adversarial manipulation. As a result, applying a model to create a confident prediction to assist those who are investigating may be misleading instead of providing additional predictive power. The noted deficiencies present a compelling rationale for the development of a representational framework that captures uncertainty in a natural way, allows for the modelling of non-trivial cross-modal dependencies and is robust when subjected to perturbation. Quantum-Inspired Modelling offers a representation of uncertainty in the form of probabilistic state representations, has principled similarity measures, and uses entanglement-based mechanisms to model dependency. Quantum-Inspired Modelling can be implemented on Classical Hardware (and all details regarding implementing the model on Classical Hardware will be provided) making it a viable approach to model the existing deficiencies in Multimodal Fake News Detection Systems by treating quantum formalism as a modelling abstraction rather than as a hardware requirement.

In this study, we introduce Q-ALIGNer, a quantum-inspired multimodal framework that processes text and images for fake news detection. Q-ALIGNer encompasses traditional feature extraction methods combined with quantum state embedding, followed by a learnable entangling unitary that fuses textual and visual components together by learning adjustment over successive cycles. Additionally, to further improve performance, we incorporate a contrastive InfoNCE alignment loss to optimize cross-modal consistency, along with a robustness objective in order to project resilience against adversarial attacks. This produces

the combined aspects of Q-ALIGNer, which project better cross-modal reasoning deficiencies, improve generalization capacity and provide improved reliability conditions under noise. The three main contributions of this work are: (1) we propose a quantum-enhanced fusion mechanism which incorporates learnable entanglement to model dependencies among text and image modalities; (2) alignment-guided contrastive learning and robustness-aware training objectives, which improved both accuracy, calibration and adversarial resilience; and (3) a comprehensive evaluation of Q-ALIGNer on four benchmarks (FakeNewsNet, Fakeddit, Weibo and MediaEval VMU) demonstrate Q-ALIGNer achieves superior performance to strong baselines (i.e., transformer-based models, multimodal models, and other quantum-inspired models) across all datasets. Q-ALIGNer is an quantum inspiration framework, but it never actually uses real life quantum computer hardware to run. The framework does use mathematical representations of quantum constructs (e.g., density matrices, entangled states, etc.), but they are used in classical simulation on classical computers (or quantum inspired). For this reason, the advantages of entanglement within the Q-ALIGNer framework should be viewed as algorithmic advantages due to the use of quantum formalisms and not as actual quantum hardware effects. This is consistent with previous approaches to quantum-inspired learning.

The rest of this paper is structured as follows. Section 2 reviews the existing literature on multimodal fake news detection, as well as quantum-inspired learning. Section 3 describes the Q-ALIGNer architecture with three components: quantum state encoding, entanglement-based fusion, and training objectives. Section 4 describes experimental setups, results, ablation studies, and robustness analysis. Section 5 wraps up with highlights and future work.

2 Literature Review

In the last few years, multimodal fake news detection studies have increased dramatically based on the limitations of unimodal models in clarifying the connection between text and image. Tufchi et al. provide a comprehensive survey highlighting deep learning models' limitations for multimodality-related applications around the data fusion approach and robustness issues [9]. Abduljaleel et al. conduct a specific survey to deep learning in the last five years (2019–2024) and also state that schemes of fusion that are simplistic concatenation and element-wise multiplication do not produce desirable performance, indicating that more complex interactions need to be tackled [10].

Recent advancements in fusion architectures have been developed, including a number of contrastive-learning and attention-based modeling approaches such as contrastive fusion techniques like MCOT (multi-modal contrastive optimal transport) that combine multimodal contrastive learning with optimal transport for aligning the textual and visual features [2]. MIMoE-FND is another mixture-of-experts model for gating modality interaction adaptively [4]. Both of these works show large performance improvements over other baseline approaches. Other works have investigated graph-based fusion like MAGIC (multimodal adaptive graph interdependencies) that utilizes adaptive graph representations to represent interdependencies and achieves very high benchmark accuracy [3]. Hybrid architectures that combine residual networks and attention architectures have also achieved reported improvements in semantic sensitivity and adaptability over existing models [11]. User or social contexts that are integrated with a modality (e.g., MFUIE–Modality fused user intention embeddings) have also been shown to further improve detection performance by enriching context models [12].

Although these advances have been made, the issues of robustness under adversarial noise and model calibration are still under-explored with respect to multimodal studies. In terms of emerging quantum-inspired learning, the QMFND approach presented future prospects of integrating quantum convolutional neural networks to enhance the feature representations [7]. The HQDNN hybrid quantum–classical architecture also provides enhanced expressivity for the purpose of fake news detection [8]. The more recent QEMF

model continues down this path, showcasing quantum entanglement as a mechanism for multimodal fusion where state-of-the-art robustness was achieved.

Quantum computing utilizes the principles of quantum mechanics to represent information in terms of quantum states within a complex-Hilbert space, rather than as classical bits that represent binary values. The ability of quantum states to exist in superposition gives rise to their ability to represent more complex information than classical bits. Quantum-inspired machine learning draws inspiration from quantum mechanics by taking advantage of mathematical representation of quantum states and their associated transformations through classical simulation without requiring the physical use of quantum hardware to perform the computations. Quantum states can be represented as a state vector or alternatively as a density matrix, which represents a probabilistic view of mixed or uncertain quantum state; the density matrix is useful for representing uncertainty and correlation and thus provides an appropriate representation for learning tasks based on noisy/ambiguous data. Entanglement is one of the most remarkable concepts that exist in quantum theory. It describes correlations between two or more parts of a single quantum system. In quantum-inspired model, entanglement is used as a representation mechanism to represent complex dependency between different sources of information. The use of entanglement-inspired transformation in multimodal learning enables the joint representation of all interactions between modalities in ways other than simple concatenation or attention-based fusion.

Even with the considerable progress made in multimodal fake news detection, there are still several important gaps that remain unaddressed in the literature. First, many of the fusional approaches utilize static or shallow integration methods, such as concatenation, late fusion, or use fixed entanglement, none of which are able capture the dynamic and complex dependencies of textual and visual modalities [2,4]. Second, though, the introduction of contrastive and attention-based methods has advanced alignment, there is still significant sensitivity to noisy and adversarial data with very little resilience with the data in real-world settings [11]. Third, quantam-inspired methods have improved expressivity via approaches like QMFND and HQDNN, however, the frameworks have mostly been employed using rigid circuit designs with no adaptive entanglement or alignment-driven supervision [7,8]. Lastly, to our knowledge, model calibration and trusted uncertainty estimation remain unexplored in multimodal misinformation detection, and without this area, the applicability of current frameworks is limited in safety-critical settings.

The Q-ALIGNer framework has been proposed to address these shortcomings by introducing a synergistic integration of quantum-inspired and deep learning components. Q-ALIGNer develops a learnable entangling unitary that models cross-modal dependencies in an adaptive fashion to allow for more flexible and expressive fusions, unlike standard fusion strategies. The architecture includes a robustness-aware loss to mitigate adversarial sensitivity, which reduces susceptibility to textual and visual perturbations. A contrastive InfoNCE alignment objective is also included to explicitly enforce cross-modal consistency across textual and visual signals while remaining semantically aligned in the presence of noise. In contrast to prior quantum-inspired approaches, Q-ALIGNer allows for end-to-end optimization of entanglement and alignment, leading to better generalization of the integrated representations. Finally, Q-ALIGNer also accommodates calibration-based training, yielding more credible probabilistic uncertainty estimates and improving overall trustworthiness during deployment. Altogether, these advancements signify the key advantages over existing methods, and further enable Q-ALIGNer to be a state-of-the-art framework for multimodal fake news detection.

To ensure greater accessibility of the methodology presented in this work for a larger audience, a short introduction to some of the many quantum-based concepts employed in the development of our multi-discipline fusion methodology has been formulated. Although these concepts were not implemented on actual quantum hardware, they are implemented as mathematical abstractions through simulation on

standard computers. Density matrices are a probabilistic representation of a system's state and are used to describe the uncertainty associated with a system of multiple components that are combined into a single entity. With the Q-ALIGNer project, density matrices were used to represent the fusion of text and visual embeddings in a quantum-inspired manner. Another significant advantage of using density matrices to represent images and text is that they naturally represent the ambiguity and uncertainty associated with multimodal information.

Partial traces are a method of extracting a single mode from a composite system, while ensuring that the remaining modes have been eliminated. In this work, the partial trace can be interpreted as a technique to allow for the independent reasoning and modelling of the text and visual modes of the system, once the two modality types have been fused and merged into a composite model. Quantum fidelity is a mathematical concept used to measure the degree of similarity between two states, and the degree of alignment between those two states. In the Q-ALIGNer project, quantum fidelity can be used to quantify the amount of alignment that exists between text and visual representations. Thus, quantum fidelity serves as a measure of the degree of consistency between the different modalities. All of the above concepts create a formalised, and interpretable, framework by which to model the relationships, and uncertainties among the different modalities associated with multimodal information, while also retaining compatibility with conventional machine-learning pipelines using standard mathematical techniques.

3 Methodology

In this section, we present the improved methodology for quantum-enhanced multimodal fake news detection. The framework we propose is called Q-ALIGNer (Quantum Contrastively-Aligned Multimodal Entanglement Network), which improves on the limitations of previous quantum multimodal models. Our proposed framework uses entanglement-driven fusion, contrastive learning, and explicit consistency validation between modalities to conduct multimodal fakes news detection. For mathematical clarity, we define all of the symbols present in the equations later. At a high level, Q-ALIGNer has five components that make up the pipeline: feature extraction, quantum state encoding, cross-modal entanglement, situation prediction through measurements, and training through multi-objective optimization. For feature extraction, text and image embeddings are extracted through contextual encoders, mapped into quantum states, entangled, and measured for veracity, consistency, and uncertainty. Training is managed using a composite loss function that includes a classification loss, alignment loss, and robustness los. We illustrate the overall workflow of the proposed Q-ALIGNer in Fig. 1. As the Fig. 1 shows, we integrate classical encoders, quantum state encoding, entanglement, and multi-measurement head.

3.1 Classical Feature Extraction

The initial step of Q-ALIGNer in the classical feature extraction stage aimed to generate rich and contextually meaningful representations for both text and image modalities prior to being encoded into quantum states. This is an important step in the work flow because the quality of quantum embeddings will depend fundamentally upon the discriminative power of classical encoders. Unlike QEMF, which deployed static embeddings for text (GloVe) and CNN-based image features (based on VGG16), Q-ALIGNer employs contextualized models and transformer-based architectures to understand higher-order dependencies and semantic structures.

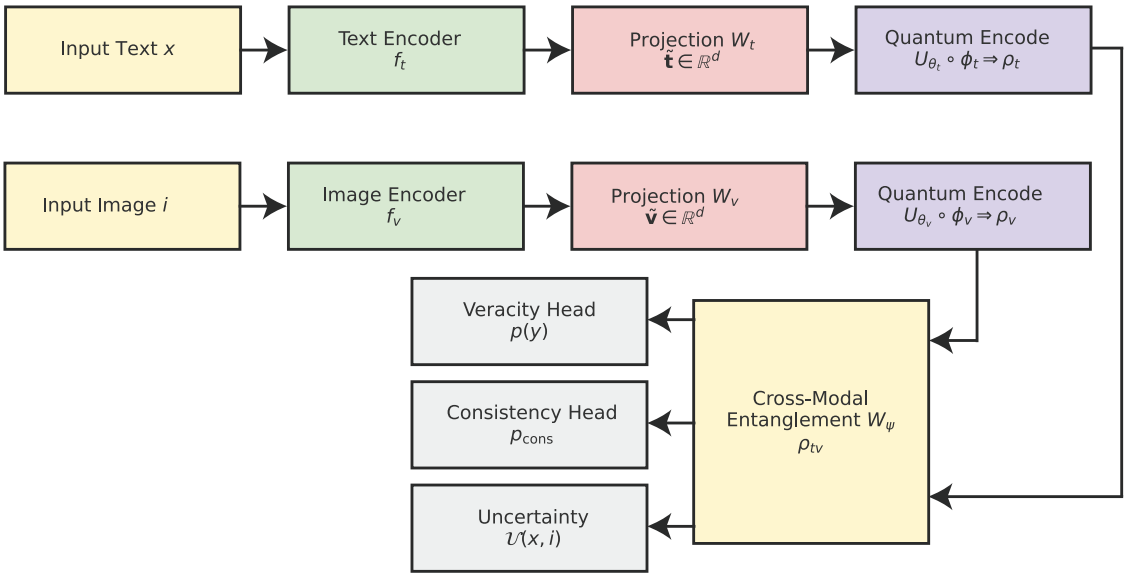

Figure 1: Overall architecture of Q-ALIGNer.

For the text modality, let $x = (w_1, w_2, \ldots, w_L)$ denote an input sentence of length L, where w_i represents the i-th token. The text encoder f_t is implemented using a pre-trained transformer model (e.g., BERT or DistilBERT) that outputs a contextualized embedding for each token:

$$H_t = f_t(x) = [h_1, h_2, \ldots, h_L], \quad h_j \in \mathbb{R}^{d_t}, \tag{1}$$

where H_t is the sequence of hidden states and d_t is the dimensionality of the textual embedding space. To obtain a global representation of the text, we compute the average pooling or use the [CLS] token representation:

$$\mathbf{t} = g(H_t) = \frac{1}{L}\sum_{j=1}^{L} h_j \in \mathbb{R}^{d_t}, \tag{2}$$

where $g(\cdot)$ is the pooling function. For the visual modality, let i denote the input image. The image encoder f_v is implemented as a Vision Transformer (ViT) or a lightweight ResNet variant. The image i is partitioned into P patches, each flattened and linearly projected into an embedding of size d_v. Formally, the image representation is given by

$$H_v = f_v(i) = [v_1, v_2, \ldots, v_P], \quad v_k \in \mathbb{R}^{d_v}, \tag{3}$$

where H_v is the sequence of patch embeddings. Similar to the text representation, a global image embedding is derived using the [CLS] token of the ViT or by mean pooling:

$$\mathbf{v} = g(H_v) = \frac{1}{P}\sum_{k=1}^{P} v_k \in \mathbb{R}^{d_v}. \tag{4}$$

Since $\mathbf{t} \in \mathbb{R}^{d_t}$ and $\mathbf{v} \in \mathbb{R}^{d_v}$ may lie in different embedding spaces, both are projected into a shared d-dimensional latent space before quantum encoding:

$$\tilde{\mathbf{t}} = W_t\mathbf{t}, \quad \tilde{\mathbf{v}} = W_v\mathbf{v}, \tag{5}$$

where $W_t \in \mathbb{R}^{d \times d_t}$ and $W_v \in \mathbb{R}^{d \times d_v}$ are learnable projection matrices. The transformed embeddings $\tilde{\mathbf{t}}, \tilde{\mathbf{v}} \in \mathbb{R}^d$ ensure that both modalities can be consistently embedded into quantum circuits with the same qubit dimensionality. As shown in Fig. 2, text and image features are first extracted using transformer and vision encoders, projected into a shared latent space for quantum embedding.

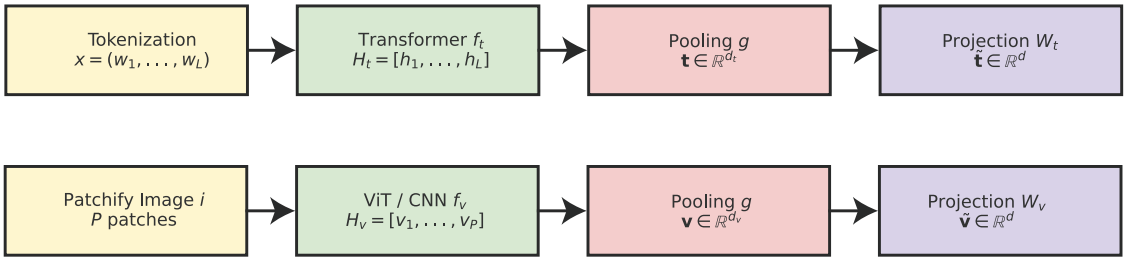

Figure 2: Classical feature extraction stage. Text tokens are processed by a transformer encoder and pooled into a global representation, while images are patchified and encoded by a ViT (or CNN) before pooling. Both modalities are projected into a shared d-dimensional latent space.

By employing transformer-based encoders, Q-ALIGNer benefits from dynamic contextual awareness in the textual domain and spatial-semantic feature richness in the visual domain. The projections guarantee alignment between modalities, preparing the system for robust quantum state encoding in the next stage.

3.2 Quantum State Encoding

The subsequent step, subsequent to inputting the text and visual embeddings into a joint latent space, is to encode the vectors into quantum states that can be utilized for subsequent quantum operations. Quantum state encoding aims to use the representational power of a quantum system to encode complex semantic relationships that would not be as effectively modeled with classical embeddings alone. Let $\tilde{\mathbf{t}} \in \mathbb{R}^d$ and $\tilde{\mathbf{v}} \in \mathbb{R}^d$ be the projected text and image embeddings, respectively. To encode the vectors into quantum states, we use parameterized quantum feature maps ϕ_t and ϕ_v to map classical data into the Hilbert space of qubits. The quantum feature map maps the text modality, which can be represented by the quantum state $|\psi_t\rangle$ is found as follows, which maps $\tilde{\mathbf{t}}$ into the quantum state $|\psi_t\rangle$.

$$|\psi_t\rangle = U_{\theta_t}(\phi_t(\tilde{\mathbf{t}}))|0\rangle^{\otimes n}, \tag{6}$$

where U_{θ_t} is a parameterized unitary circuit with trainable parameters θ_t, n is the number of qubits, and $|0\rangle^{\otimes n}$ is the n-qubit ground state. The adjustable parameters θ_t determine how a differentiable unitary transformation will transform the embedding of a piece of text after it has been quantum feature mappable, and the degree to which those embeddings will be represented in the quantum-inspired latent space. No entanglement is introduced through the optimised parameters θ_t during this phase of the training, but instead they are adjusted through the training process so that the encoded state has discriminative capability when combined with visual information and subsequently assessed by downstream measurement. Thus, θ_t determines how expressive and aligned with the task that the encoded quantum state will be rather than encoding class information directly. The corresponding density matrix representation is

$$\rho_t = |\psi_t\rangle\langle\psi_t|. \tag{7}$$

Similarly, for the image modality, the embedding $\tilde{\mathbf{v}}$ is encoded as

$$|\psi_v\rangle = U_{\theta_v}(\phi_v(\tilde{\mathbf{v}}))|0\rangle^{\otimes m}, \tag{8}$$

where U_{θ_v} is the parameterized unitary with parameters θ_v, m is the number of qubits allocated to the image encoding, and $|0\rangle^{\otimes m}$ is the ground state. Its density matrix is

$$\rho_v = |\psi_v\rangle\langle\psi_v|. \tag{9}$$

The choice of feature map $\phi(\cdot)$ is critical. One common approach is amplitude encoding, where a normalized embedding vector is directly encoded into the amplitudes of a quantum state:

$$|\psi_{amp}(\tilde{\mathbf{t}})\rangle = \frac{1}{\|\tilde{\mathbf{t}}\|}\sum_{j=1}^{d}\tilde{t}_j|j\rangle. \tag{10}$$

Another approach is angle encoding, where elements of the embedding vector parameterize rotation gates:

$$|\psi_{ang}(\tilde{\mathbf{t}})\rangle = \bigotimes_{j=1}^{d} R_y(\tilde{t}_j)|0\rangle, \tag{11}$$

where $R_y(\cdot)$ denotes a rotation around the y-axis of the Bloch sphere.

The initial joint multimodal state before fusion is the tensor product of the two density matrices:

$$\rho_{init} = \rho_t \otimes \rho_v. \tag{12}$$

This state serves as a basis for future entanglement operations for the quantum system to model cross-modal dependencies. As seen in Fig. 3, classical latent embeddings are mapped to quantum states through feature maps and parameterized circuits. The quantum state encoding stage therefore creates a mathematically principled mapping of classical embeddings into quantum Hilbert space. By selecting suitable encoding strategies and parameterized unitaries, Q-ALIGNer guarantees that rich contextual information from the text and image modalities are retained and available for quantum entanglement and measurement in subsequent stages.

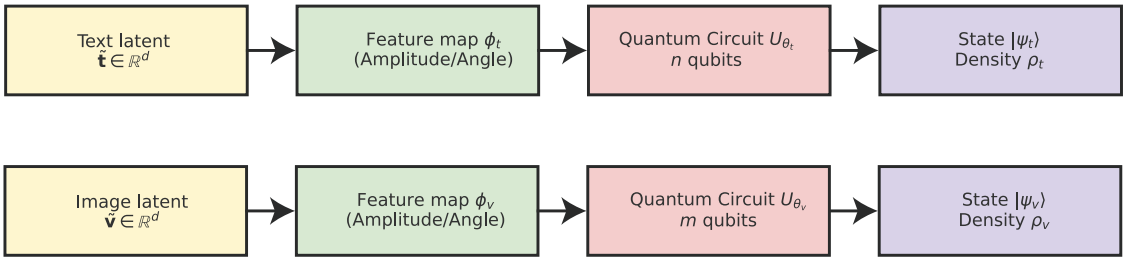

Figure 3: Quantum state encoding. Projected embeddings $\tilde{\mathbf{t}}, \tilde{\mathbf{v}}$ are transformed by feature maps (amplitude or angle encoding) and parameterized circuits $U_{\theta_t}, U_{\theta_v}$ to yield quantum states $|\psi_t\rangle, |\psi_v\rangle$ and density matrices ρ_t, ρ_v.

3.3 Cross-Modal Entanglement

The crucial part of Q-ALIGNer is what we call cross-modal entanglement, which allows engagement of the quantum states that are thinking about the modality of text and the modality of image. First, recall that classical models will typically factorize their approaches using simple concatenation or attention-based fusion of modalities. Instead, Q-ALIGNer uses the quantum entanglement principle to find correlations or co-occurrence between modalities in the Hilbert space. In this phase, the model is able to learn dependencies

that cannot really be expressed in a linear combinatorial way, as with the joint state initial state $\rho_{init} = \rho_t \otimes \rho_v$ the parameterized entangling unitary W_ψ applies as:

$$\rho_{tv} = W_\psi(\rho_t \otimes \rho_v) W_\psi^\dagger, \tag{13}$$

where W_ψ is a string of controlled entangling gates with learnable parameters ψ. This process creates non-classical correlations between the text and image states and harmonizes the representations of each modality in the joint multimodal quantum space. A canonical example of an entangling gate is the Controlled-NOT (CNOT) across modalities. If $|\psi_t\rangle$ and $|\psi_v\rangle$ are encoded on separate registers, the entangling operation can be represented as

$$|\Psi_{ent}\rangle = (\mathrm{CNOT}_{t\to v})(|\psi_t\rangle \otimes |\psi_v\rangle). \tag{14}$$

This results in a correlated state that cannot be decomposed into a simple tensor product. More generally, entanglement is introduced through parameterized controlled rotations:

$$U_{ent}(\beta) = \exp(-i\beta Z_t \otimes X_v), \tag{15}$$

where Z_t and X_v are Pauli operators acting on text and visual registers, respectively, and β is a trainable parameter. In the jointly distributed multimodal representation produced by the entangling unitary given in Eq. (15), correlations between the quantum states of the text and the image are represented. This entangled quantum state does not explicitly encode the class label information associated with each individual quantum state. Rather, it produces a shared latent representation that is useful for supervised measurements of the relevant task information. While the parameters of the encoding unitaries, entanglement operation, and measurement operators are all jointly optimised to produce measurement outputs corresponding to the true class label information during training, the true class label is not explicitly represented by the entangled quantum state. The resulting density matrix is

$$\rho_{tv} = U_{ent}(\beta)(\rho_t \otimes \rho_v) U_{ent}(\beta)^\dagger. \tag{16}$$

The degree of alignment between modalities after entanglement can be quantified using quantum fidelity:

$$F(\rho_t, \rho_v) = \left(\mathrm{Tr}\sqrt{\sqrt{\rho_t}\rho_v\sqrt{\rho_t}}\right)^2. \tag{17}$$

Maximizing this fidelity encourages the quantum states of text and image to occupy nearby regions of the Hilbert space when they correspond to consistent information. Additionally, the swap test provides a direct way to evaluate the similarity of quantum states. For two pure states $|\psi_t\rangle$ and $|\psi_v\rangle$, the swap test yields a measurement outcome related to their inner product:

$$P(0) = \frac{1 + |\langle\psi_t|\psi_v\rangle|^2}{2}, \tag{18}$$

where $P(0)$ is the probability of measuring the ancilla qubit in state $|0\rangle$. This provides a probabilistic measure of consistency between text and image representations. The fusion of modalities via a learnable entangling unitary is presented in Fig. 4, producing the joint state ρ_{tv} that captures semantic dependencies.

Figure 4: Cross-modal entanglement. Independent text and image states are fused by an entangling unitary W_ψ (e.g., CNOT or controlled rotations) to produce the joint state ρ_{tv}, enabling quantum-native dependency modeling in Hilbert space.

Thus, cross-modal entanglement is achieved with Q-ALIGNer through not only fusing modalities but also pursuing semantic coherence via fidelity maximization and swap-test based alignment. Downstream measurement and prediction tasks depend upon the entangled state ρ_{tv} that ensures multimodal information is fuzed through a preestablished, quantum-native encounter.

3.4 Measurement and Prediction

When the entangled multimodal state ρ_{tv} has been produced, the information that is useful for predictive purposes can be extracted. In a quantum system, the extraction of information is accomplished by measurements that collapse the quantum state into classical values according to the Born rule. In Q-ALIGNer, we use projective measurements and parameterized measurement operators to create measurements for veracity classification, modality consistency, and uncertainty estimation. In formal terms, let M_k denote the set of measurement operators that comprise a Positive Operator-Valued Measure (POVM). These measurement operators satisfy the completeness condition,

$$\sum_k M_k = I, \qquad M_k \succeq 0, \tag{19}$$

where I is the identity operator. For each class k, the probability of assigning input (x, i) to class k is given by

$$p_k = \mathrm{Tr}(M_k \rho_{tv}). \tag{20}$$

The logits are defined as expectation values of the measurement outcomes:

$$z_k = \mathrm{Tr}(M_k \rho_{tv}), \tag{21}$$

which are then normalized through the softmax function:

$$p_k = \frac{\exp(z_k)}{\sum_j \exp(z_j)}. \tag{22}$$

For binary fake news detection, we define two measurement operators M_{real} and M_{fake} such that

$$M_{real} + M_{fake} = I. \tag{23}$$

The classification probability of veracity is then

$$p(y = \text{real}|x, i) = \mathrm{Tr}(M_{real} \rho_{tv}), \tag{24}$$

$$p(y = \text{fake}|x, i) = \mathrm{Tr}(M_{fake} \rho_{tv}). \tag{25}$$

In addition to truthfulness, Q-ALIGNer adds more heads for modality consistency and uncertainty. The consistency head checks to see whether the textual and visual modalities are semantically consistent.

This is measured by quantifying the overlap of the subspaces corresponding to the text and image registers. Thus, class prediction in Q-ALIGNer arises from learned measurement operators applied to the entangled multimodal state, rather than from explicit label encoding within the quantum-inspired state itself. Let M_{cons} be the corresponding measurement operator, we have

$$p_{cons} = \mathrm{Tr}(M_{cons}\rho_{tv}), \tag{26}$$

where a higher value implies greater congruence between the modalities. To quantify uncertainty, an entropy-based measurement is calculated over the distribution of probabilities. The predictive uncertainty of an input (x, i) is defined as

$$\mathcal{U}(x, i) = -\sum_{k} p_k \log p_k. \tag{27}$$

The uncertainty score shows how confident we are in the model prediction. More uncertainty corresponds to greater entropy. When there is conflicting or ambiguous multimodal evidence (like expert disagreement or propaganda), Q-ALIGNer reflects the uncertainty of the knowledge or prediction about the outcome. This means that Q-ALIGNer does not assume that the data inputs will all be labeled with one definitive truth; rather, the higher the entropy and lower the modality consistency score are, the less confidence we have in defining a unique classification for the input evidence. In addition, expectation values of the Pauli observables can be understood as the prediction in terms of observables. For an observable O acting on the joint state, the expected value is:

$$\langle O \rangle = \mathrm{Tr}(O\rho_{tv}). \tag{28}$$

By selecting O very carefully, we can describe latent relationships across modalities and offer explainability on decisions made by the model. The measurement phase is depicted in Fig. 5. Here we show an entangled state is projected into veracity, consistency and uncertainty outputs of the model. In summary, then, the measurement and prediction functions of Q-ALIGNer are formalized in terms of POVMs, projective operators, and the calculation of uncertainty using unique (to us) forms of entropy. Not only do we provide classification outputs through these quantum-native operations but we also provide an opportunity for more clarity in how weary we are of multimodal inputs even further separating Q-ALIGNer from non-Q multimodal classifiers.

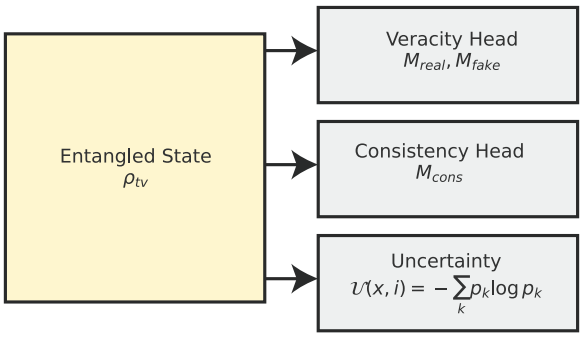

Figure 5: Measurement and prediction. The entangled state ρ_{tv} is measured with POVMs $\{M_k\}$ to yield veracity probabilities $p(y)$, a consistency score p_{cons}, and uncertainty estimates $\mathcal{U}(x, i)$ based on predictive entropy.

The internal fusion process of an aligner that is quantum based on entanglement operates at the level of outputs and not at the level of the actual circuit level. The model produces scores for the degree of

modality matching between a type of image and some text associated with that type of image, for the degree of the fidelity of the matched modalities to the original source, and for the degree to which the model has the potential of incorrectly predicting an association. These scores can be easily understood by a non-expert stakeholder and do not require the user to have any understanding of quantum mechanics. During the training and inference phases, Q-ALIGNer assumes that paired textual and visual datasets are both available. In a multimodal system when input types are noisy, corrupted or low-quality, Q-ALIGNer indicates the uncertainty of the relationship between the input types through a low score on modality consistency. Furthermore, it indicates the uncertainty in its prediction with a high level of predictive entropy. Thus, when faced with degraded multimodal inputs, the Q-ALIGNer system signals its lack of confidence by producing less confident predictions than by producing over-confident predictions.

3.5 Training Objective

Q-ALIGNer's training is aided by a structured composite objective that combines four criteria: classification accuracy, modal semantic alignment, entanglement consistency and adversarial robustness. Q-ALIGNer does not depend on cross-entropy loss as typical multimodal fake news detection models do; instead, we introduce several mutually helpful terms, with each term reflecting a separate facet of multimodal reasoning in the quantum domain. The overall composite objective function is given by

$$\mathcal{L} = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{\text{InfoNCE}} + \lambda_2 \mathcal{L}_{\text{swap}} + \lambda_3 \mathcal{L}_{\text{Bures}} + \lambda_4 \mathcal{L}_{\text{rob}}, \tag{29}$$

where $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ are hyperparameters balancing the contributions of auxiliary terms. The classification objective for fake news detection is formulated as standard cross-entropy. The coefficients λ_1, λ_2, λ_3, and λ_4 balance the relative contributions of cross-modal alignment, swap-based similarity, fidelity-based consistency, and adversarial robustness, respectively. In all experiments, these coefficients were selected using validation-based tuning to ensure stable optimization and balanced performance across objectives. Each coefficient was constrained to a small positive range to prevent auxiliary losses from dominating the primary classification objective. Specifically, λ_1 and λ_2 were set within the range $[0.1, 1.0]$ to regulate alignment strength, while λ_3 and λ_4 were chosen from $[0.01, 0.5]$ to encourage consistency and robustness without degrading classification accuracy. Given the true distribution $q(y)$ and predicted probabilities $p(y|x, i)$:

$$\mathcal{L}_{CE} = -\sum_{y} q(y) \log p(y|x, i), \tag{30}$$

here $q(y)$ denotes a one-hot distribution indicating the ground truth label, with the role of correctly aligning the outcomes of the quantum measurements to the truthful labels. To encourage semantic alignment between matched image text-support pairs, we adopt the contrastive learning methodology which derives from InfoNCE. For a batch of N pairs, we define the loss as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\rho_t^i, \rho_v^i)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\rho_t^i, \rho_v^j)/\tau)}, \tag{31}$$

where $\text{sim}(\rho_t, \rho_v)$ is a similarity function on quantum states, and τ is a temperature scaling factor. This creates alignment between matched pairs, while increasing distance for mismatched pairs. The swap-test inspired loss calculates similarity between quantum states and works as follows for density matrices ρ_t and ρ_v, applying the swap operator S:

$$S(|a\rangle \otimes |b\rangle) = |b\rangle \otimes |a\rangle. \tag{32}$$

The swap-test probability of measuring ancilla in $|0\rangle$ is related to state overlap. The corresponding loss is

$$\mathcal{L}_{\text{swap}} = 1 - \text{Tr}[S(\rho_t \otimes \rho_v)], \tag{33}$$

which penalizes dissimilarity between modalities. Fidelity-based alignment is encouraged by minimizing the Bures distance between text and image states:

$$F(\rho_t, \rho_v) = \left(\text{Tr}\sqrt{\sqrt{\rho_t}\rho_v\sqrt{\rho_t}}\right)^2, \tag{34}$$

$$\mathcal{L}_{\text{Bures}} = 2\left(1 - \sqrt{F(\rho_t, \rho_v)}\right). \tag{35}$$

A smaller Bures distance corresponds to greater modal semantic coherence. To investigate adversarial robustness we create perturbed text-image pairs (x', i') by substituting synonyms, character-level perturbations, or gradient-based attacks on the images. The robustness loss is defined as

$$\mathcal{L}_{\text{rob}} = \mathbb{E}_{(x',i')}[\mathcal{L}_{CE}(f(x', i'), y)], \tag{36}$$

which enforces stability of predictions under adversarial conditions. Robustness can also be measured through Kullback–Leibler divergence between clean and adversarial distributions:

$$D_{KL}(p(y|x, i)||p(y|x', i')) < \epsilon, \tag{37}$$

where ϵ is a tolerance bound. Each component serves a unique purpose: $\mathcal{L}_{CE}$ ensures predictive accuracy, $\mathcal{L}_{\text{InfoNCE}}$ enforces discriminative alignment, $\mathcal{L}_{\text{swap}}$ provides a quantum-native similarity constraint, $\mathcal{L}_{\text{Bures}}$ encodes fidelity based alignment, and $\mathcal{L}_{\text{rob}}$ guarantees adversarial robustness. In this way, Q-ALIGNer is optimized for classification as well as alignment, consistency, and robustness, thereby developing beyond the traditional multimodal learning paradigms. While alignment and entanglement optimization reduce uncertainty for coherent and well-supported cases, irreducible ambiguity arising from conflicting interpretations or expert disagreement is preserved rather than artificially suppressed.

3.6 Adversarial Robustness

Adversarial robustness is a significant design consideration for Q-ALIGNer since fake news detection environments are inherently adversarial; malicious actors take measures to avoid detection through easily overlooked perturbations in either text or image data. To address this, Q-ALIGNer incorporates adversarial training as part of the optimization strategy so that stable predictions are learned despite adversarial perturbation. In the context of textual data, adversarial perturbations can occur through various approaches. Synonym substitution could replace a word with a semantically similar alternative, while character level approaches would substitute a word with a misspelling—an error that is still readable but the misspelled version may mislead the language model's token embeddings. Back-translation methods are also capable of producing paraphrased alternatives for the same sentence. More formally, suppose the original text sequence is x. Then, the adversarially perturbed sentence could be expressed as:

$$x' = \mathcal{A}_t(x), \tag{38}$$

where $\mathcal{A}_t(\cdot)$ is a perturbation function acting in the text domain. For image data, adversarial attacks are commonly implemented through gradient-based methods. The Fast Gradient Sign Method (FGSM) generates perturbations as

$$i' = i + \epsilon \cdot \text{sign}(\nabla_i \mathcal{L}_{CE}), \tag{39}$$

where i is the original image, ϵ is the perturbation budget, and $\nabla_i \mathcal{L}_{CE}$ is the gradient of the classification loss with respect to the input image. More advanced methods such as Projected Gradient Descent (PGD) iteratively refine these perturbations:

$$i^{(k+1)} = \Pi_{B_\epsilon(i)}\Big(i^{(k)} + \alpha \cdot \text{sign}\big(\nabla_{i^{(k)}} \mathcal{L}_{CE}\big)\Big), \tag{40}$$

where $\Pi_{B_\epsilon(i)}$ projects the perturbed image back into the ϵ-ball around the original image, and α is the step size. During training, Q-ALIGNer minimizes a robustness-augmented objective that includes both clean and adversarial examples. All adversarial robustness experiments used standard practice to select the perturbation strengths, ensuring that the perturbation strengths were realistic, but bounded, based upon the theory. For example, in preparation for FGSM-based image attacks, the perturbation budget was set to $\epsilon = 8/255$. On the other hand, for the PGD attack, the step size (α) was set at a value of $2/255$, with 10 iterations, with projected ϵ ball while performing this attack. The perturbations that were utilized for textual data (i.e., for text-based adversarial attacks), were restricted to applying semantically preserving transformations such as synonym substitution and character-level noise to the original text, while not changing the original semantics or meaning of the content. All settings for every model were consistent to permit fair and consistent comparisons. For an input pair (x, i) and its adversarial counterpart (x', i'), the robustness constraint enforces that the predictive distributions remain close. This can be expressed using Kullback–Leibler divergence:

$$D_{KL}(p(y|x, i) \| p(y|x', i')) < \delta, \tag{41}$$

where δ is a small tolerance constant. This ensures that adversarial perturbations do not cause large shifts in the predicted distribution. The robustness loss can therefore be defined as

$$\mathcal{L}_{rob} = \mathbb{E}_{(x,i)\sim\mathcal{D}}\Big[\mathcal{L}_{CE}(f(x', i'), y)\Big], \tag{42}$$

where by $\mathcal{D}$ we mean the training distribution and (x', i') are adversarially perturbed samples. By including this term in the composite loss, Q-ALIGNer learns invariant features that are robust to perturbations using its quantum encoding and entanglement layers. Because adversarial robustness by Q-ALIGNer goes beyond robustness to adversarial attacks, and also advances generalization to clean text in the multimodal structure, Q-ALIGNer avoids superficial overfitting, and instead guides the model towards deeper semantic consistency between modalities by training with clean and perturbed samples. This capacity is particularly important in a social media context, when misleading information like fake news often appears reshaped, in reworded or reaudio or recomposed images. In Q-ALIGNer, adversarial robustness is achieved with a combination of perturbation-aware training, gradient-based attacks for images, semantic-preserving perturbations for the text, and distributional regularization through divergence constraints, thus Q-ALIGNer does not just protect against naive attacks on multimodal fake news, but also more sophisticated methods.

3.7 Computational Considerations

Considering the computational complexity and ensuring practical viability for real-world applications is an important dimension of any design for quantum-enhanced models like Q-ALIGNer. The traditional multimodal framework for fake news detection was the multimodal QEMF model. However, one of the major drawbacks of the QEMF model was a very high computational cost of $O(n^3)$ for training, with training times surpassing 48 h with benchmark datasets. Thus, although Q-ALIGNer is a quantum-enhanced model, it directly addressed this concern with efficiency-oriented design choices in the classical and quantum aspects of the model. For example, during the classical pre-processing stage, token pruning for text dimensions and

patch selection for image dimensions are introduced. Suppose L is the original length of the input text and P is the number of patches for the image input to the model. If L' is the length of text after pruning, and P' is the number of patches selected after pruning, then the total effective input size $n = L' + P'$ decreases, and the quantum circuit's effective input size and depth of the circuit architecture is lowered. These reduce the computational load greatly.

$$O(n^3) \quad \longrightarrow \quad O((L' + P')^2 d), \tag{43}$$

where d is the dimension of projection. This quadratic scaling leads to both improvements in efficiency and also in representational richness. In the encoding phase into quantum states, the cost is driven primarily by the number of qubits q and the depth D of the variational quantum circuit. The total gate complexity for a circuit with q qubits is $O(qD)$, and each qubit requires $O(D)$ parameterized gates.

$$C_{encode} = O(qD). \tag{44}$$

In general, amplitude encoding has a preparation cost of $O(2^q)$ asymptotically, but with the use of structured embeddings and exploiting sparsity, Q-ALIGNer accomplishes this in $O(q^2)$ operations, which is more practical. For cross-modal entanglement, controlled entangling gates can be used to introduce correlations across the text and image registers. If q_t and q_v are used to denote the respective number of qubits for the text and image encodings, then the entangling operation is only a cost of

$$C_{ent} = O(q_t q_v), \tag{45}$$

which is bilinear in the number of qubits assigned to each modality. As pruning reduces q_t and q_v, the efficiency of this operation is significantly borderline compared to the cubic cost of QEMF. Measurement and prediction equations involve the evaluation of expectation values for observables. If M is the number of measurement operators, where each operator requires repeating the same circuit execution (a shot) to estimate probabilities, the cost of measurement is

$$C_{meas} = O(M \cdot S), \tag{46}$$

where S represents the number of shots. In practice, M is maintained small (usually 2–4 operators for binary or multi-class classification) to control measurement overhead. The robustness component adds adversarial training and essentially doubles the dataset size by including perturbed samples. If we let the clean training set size be N, adversarial augmentation produces $2N$ samples, and the reduction in effective complexity is mediated by prunes and projection operations. Thus, putting all the components together, we can also write the effective training complexity of Q-ALIGNer as

$$C_{total} = O((L' + P')^2 d) + O(qD) + O(q_t q_v) + O(MS) + O(N), \tag{47}$$

with each term representing preprocessing, quantum state encoding, entanglement, measurement, and dataset scaling, respectively. In comparison to the $O(n^3)$ scaling of QEMF's model, Q-ALIGNer presents a tremendous computational advantage. Regarding memory requirements, classical embeddings scale linearly in n and d, while quantum memory requirements scale with the number of qubits q. Overall, q manifests a logarithmic scaling in embedding dimension, namely:

$$\text{Memory}_{quantum} = O(q) = O(\log d). \tag{48}$$

Q-ALIGNer harnesses powerful accuracy and robustness while improving computational performance through employing pruning strategies, shallow variational circuits, and optimized measurement strategies.

The framework was constructed with a theoretical understanding of the problem, along with practical computational capabilities for real large-scale applications in the domain of fake news detection.

4 Results and Discussion

This chapter gives a detailed assessment of the suggested Q-ALIGNer framework for detecting fake news across different modalities. The aim of the experiments is to test whether our approach differentiates the efficacy of quantum-enriched multimodal fusion vs. state-of-the-art classical and hybrid baselines. The results of our experiments encompass various benchmark datasets with varying domains of textual and visual misinformation. We organize our findings to examine five major components: (i) The classification performance metrics of Q-ALIGNer against accuracy, precision, recall, F1-score and ROC-AUC; (ii) Contribution to baseline comparisons to show improvements in veracity predictions; (iii) Ablations, or studies demonstrating individual contributions of cross-modal entanglement, InfoNCE alignment, and robustness loss; (iv) Adversarial robustness under textual and visual disturbances; and (v) Model calibration and uncertainty, along with examples of how our framework is efficiently applied in computational efficiency. Collectively these results frame Q-ALIGNer in terms of robustness and extensive generalizability as well as interpretaiblity for mountering misinformation in a multimodal social media environment.

4.1 Datasets

To assess the proposed Q-ALIGNer framework, we ran experiments on a variety of publicly available multimodal fake news detection benchmarks. We selected these benchmarks so that there would be a variety of modes, domains, and languages represented, in order to evaluate the component's generalization across domains and robustness in real-world misinformation scenarios. **FakeNewsNet:** FakeNewsNet is a popular benchmark that provides a number of news article sets with social context and multimedia components. For this analysis, we take the subset with paired textual articles and corresponding images. This dataset has the indicators of fake and real news verified through professional fact checking sources. The anonymous crowd sourced verification markers make it a reliable resource for binary classification tasks. **Fakeddit:** Fakeddit is a larger scale dataset from Reddit posts with multimodal samples, where each entry has text and an associated image. It has both binary (fake vs. real) and fine grained multi-class labeling schemes. In our experiments, we adopted the binary framework to align with other multimodal misinformation detection studies. The size of the Fakeddit dataset was helpful in enabling multimodal assessment of Q-ALIGNer in a complex and noisy social media context.

Weibo: The Weibo multimodal fake news dataset has social media posts in the Chinese language and an accompanying image. Each instance is annotated fake or real, based on typical fact-checking sources, introducing variation in coding schema. This is useful for assessing the cross-lingual adaptability of Q-ALIGNer. The shorter length of text posts, and casual style of writing, introduce a challenge versus long-form articles. **MediaEval VMU:** To provide another additional cross-domain generalization assessment, we present the Verifying Multimedia Use (VMU) dataset from the MediaEval benchmark. This dataset takes the form of multimodal Twitter posts, where the focus is determining whether the textual claim is verified by visual evidence. This assesses the veracity of an accompanying image but not veracity of the article, distinguishing it from Fakeddit and FakeNewsNet, which is focused on article style. The MediaEval VMU task considers rapid informal and high noise, and it is consistent with the more real-time nature of social media misinformation. Across all datasets we separate into training, validation, and testing (80:10:10 split), following common procedure unless the authors created a split. Textual data was preprocessed with subword tokenization (WordPiece or Byte-Pair Encoding), and images were preprocessed by resizing and normalization. Then the images were patchified for the Vision Transformer. This approach is consistent

for both quantum encoding, multimodal data, and allows for a uniform preprocessing strategy for fair comparisons across datasets.

Despite Q-ALIGNer being assessed using standard benchmark sets, Q-ALIGNer's benchmark sets contain examples of "real" social media sites (e.g., Fakeddit, Weibo, and MediaEval VMU) with characteristics such as informal language, poor alignment of image/text content, high levels of noise within the data, and depending on who you follow, these examples have significant variation in content over time. All of these factors create an excellent way to simulate a real-world environment that uses Q-ALIGNer in a similar method as a large-scale social media platform could be operated. The strong results demonstrate a qualitative performance on posts written in Chinese and containing a specific style of culturally dissimilar misinformation present in China, that the quantum-inspired method of alignments is able to generalise across languages, including but not limited to English. There is an opportunity to look at this further as an area of research into understanding cross cultural differences.

4.2 Evaluation Metrics

The assessment of Q-ALIGNer and comparison baselines is conducted using standard classification measures along with specialized metrics that evaluate robustness and calibration. As the task is binary fake news classification, we report accuracy, precision, recall, F1 score and area under the ROC (AUC). Accuracy indicates the overall correctness of the predictions, precision indicates the confidence of the model when predicting fake cases, recall indicates the model's ability to detect all fake news instances, and the F1 score is a balance between precision and recall. The AUC summarizes the tradeoff between true positive and false positive rate which gives a composite indication of discriminative performance across thresholds. In addition to the preceding standard measures, we evaluate robustness using predictive performance in the presence of adversarial perturbations. Robustness is quantified as the difference between accuracy of clean data and adversarially perturbed data:

$$\text{Robustness} = \text{ACC}_{clean} - \text{ACC}_{adv}, \tag{49}$$

where ACC_{clean} indicates accuracy on unperturbed inputs, and ACC_{adv} is the accuracy when adversarial perturbation is applied. Finally, calibration of predictive uncertainty is evaluated using the expected calibration error (ECE), which quantifies the discrepancy between model confidence and observed accuracy. More formally, if the predictions are split into M bins B_m, then

$$ECE = \sum_{m=1}^{M} \frac{|B_m|}{n} \cdot \left|\text{acc}(B_m) - \text{conf}(B_m)\right| \tag{50}$$

Here, $\text{acc}(B_m)$ represents the empirical accuracy of bin m, $\text{conf}(B_m)$ is the average confidence within that bin, and n is the total number of samples. This triplet can provide a global assessment of Q-ALIGNer, giving insights into accuracy and discriminatory performance, robustness against adversarial examples, and trustworthiness of the estimate of confidence.

4.3 Baseline Models

In order to determine the efficacy of Q-ALIGNer, we compared it to a variety of baseline models, including textual, visual, multimodal, and quantum-inspired architectures. Each of these baselines is well-known and is based upon either classical deep learning approaches or the nearest previous quantum-enhanced approach for training data segmentation and classification. Text-only detection comparisons have highly accessible baseline models that perform well as pretraining on large-scale corpora captures deep

contextual semantics. Specifically, for text, the following transformer-based language models are considered: BERT [13], RoBERTa [14], and XLNet [15]. In addition, we include BiLSTM [16] architectures, even though they are earlier architectures and there have since been better applications for sequence classification, these architectures remain competitive. Next, these baseline models allow us to quantify the discriminative capacity of signals based upon textual signals alone.

For the vision-only detection baselines, we utilize convolutional and transformer-based image encoders, including ResNet [17], EfficientNet [18], and the Vision Transformer (ViT) [19]. These models provide a standard to evaluate the impact of visual modality in fake news detection. For the multimodal baselines we implement fusion architectures that integrate both textual and visual streams. Multimodal-BERT [20] and ViLBERT [21] are state-of-the-art methods which utilize cross-attention mechanisms to align information across modalities. These baselines will allow for direct comparisons of classical cross-attention fusion with the entanglement based fusion explained in Q-ALIGNer.

Recent multimodal foundation models, including CLIP-based derivatives and large-scale vision–language architectures proposed in 2024–2025, represent an active area of research. However, many of these models are not specifically designed for misinformation detection and often rely on proprietary or continuously evolving pretraining data, which complicates fair and reproducible comparison on established fake news benchmarks. In this study, baselines were selected to provide controlled and reproducible coverage across text-only, vision-only, multimodal transformer-based, and quantum-inspired paradigms. The inclusion of QEMF further enables direct comparison within the quantum-inspired modeling family. Incorporating recent foundation models under standardized evaluation protocols is an important direction for future work.

Finally, we compare against Quantum-Enhanced Multimodal Fusion (QEMF) [22], which we believe is the most relevant related work to applying quantum methods to fake news detection. While QEMF is reliant on amplitude encoding and a fixed entangling strategy, our Q-ALIGNer utilizes learnable entangling unitaries, contrastive alignment, and adversarial robustness. By benchmarking to QEMF, we showcase and highlight our design's incremental benefit in both accuracy and robustness. These baselines create a robust evaluation basis for our Q-ALIGNer, as we are careful to assess improvement over textual, visual, multimodal, and quantum inspired baselines. For the purpose of analyzing the performance of different methods of detecting fake news, we used baseline models that represent a broad spectrum of the most popular design paradigms (unimodal, multimodal and quantum-inspired). Some of the selected baselines were introduced before 2020 and continue to be widely referenced and used as baseline models for comparison by researchers. As an example, QEMF was included due to its status as a recent model with quantum-inspired capabilities which allows for direct comparison against other quantum-inspired systems. Future analyses will include the use of more recent models as a means of further enhancing the results of the analysis.

4.4 Main Results

The empirical tests of the Q-ALIGNer framework are presented in this section. The experiments are designed to measure its efficiency and effectiveness for multimodal fake news detection on several datasets. The tests include evaluation of overall classification performance, analysis of performance on specific datasets, ablation studies, adversarial robustness, uncertainty calibration, and computational efficiency. Results are contextualized in relation to existing state-of-the-art baselines, and the advantages of multimodal fusion using quantum-enhancement are described.

In Table 1, a comprehensive comparison against strong text, visual, multimodal, and quantum-inspired baselines is presented: FakeNewsNet, Fakeddit, Weibo, and MediaEval VMU datasets are included. The results consistently indicate that Q-ALIGNer outperforms all competing methods across all performance

metrics: Accuracy, Precision, Recall, F-1, and AUC. Models based on text only (i.e., BERT, RoBERTa, and XLNet) and vision only (i.e., ResNet and EfficientNet) were outperformed, though limited (all models are text only or vision only), because performance still requires multimodal reasoning. Baselines based on multimodal fusions like ViLBERT demonstrate noticeable improvements, especially in Recall and F-1, due to integrated reasoning across textual and visual modalities. The quantum-enhanced model QEMF shows further improvement, but Q-ALIGNer still produces an additional gain of 3–4 percent points in Accuracy and F-1. Thus, Q-ALIGNer shows to be the most robust architecture.

Table 1: Performance comparison of Q-ALIGNer and baseline models across all four datasets. best results per dataset are highlighted in bold.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
FakeNewsNet					
BERT [13]	81.4	80.5	79.8	80.1	84.2
RoBERTa [14]	83.6	82.9	82.1	82.5	86.1
XLNet [15]	84.7	83.9	83.2	83.5	87.0
ResNet [17]	78.5	77.9	77.0	77.4	81.2
EfficientNet [18]	80.3	79.7	79.1	79.4	83.0
ViT [19]	82.7	82.1	81.4	81.7	85.0
Multimodal-BERT [20]	85.9	85.2	84.5	84.8	88.0
ViLBERT [21]	87.1	86.3	85.9	86.1	89.0
QEMF [23]	88.5	87.9	87.2	87.5	90.2
Q-ALIGNer (ours)	**91.2**	**90.7**	**90.1**	**90.4**	**93.1**
Fakeddit					
BERT [13]	82.7	81.9	81.3	81.6	85.0
RoBERTa [14]	85.4	84.6	83.9	84.2	87.2
XLNet [15]	86.1	85.4	84.8	85.0	88.0
ResNet [17]	79.1	78.3	77.6	77.9	82.0
EfficientNet [18]	80.6	79.8	79.2	79.5	83.3
ViT [19]	83.0	82.2	81.6	81.9	85.6
Multimodal-BERT [20]	86.8	86.1	85.5	85.8	88.7
ViLBERT [21]	88.6	87.9	87.2	87.5	89.5
QEMF [23]	89.7	89.0	88.4	88.7	90.8
Q-ALIGNer (ours)	**92.9**	**92.3**	**91.6**	**91.9**	**94.2**
Weibo					
BERT [13]	80.9	80.1	79.5	79.8	83.3
RoBERTa [14]	83.1	82.4	81.8	82.1	85.5
XLNet [15]	84.0	83.3	82.7	83.0	86.4
ResNet [17]	77.8	77.0	76.3	76.6	81.0
EfficientNet [18]	79.4	78.7	78.0	78.3	82.4
ViT [19]	82.0	81.3	80.6	80.9	84.7

(Continued)

Table 1 (continued)

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)	AUC (%)
Multimodal-BERT [20]	85.0	84.3	83.6	83.9	87.4
ViLBERT [21]	86.5	85.9	85.3	85.6	88.1
QEMF [23]	87.9	87.2	86.7	86.9	89.6
Q-ALIGNer (ours)	**91.7**	**91.1**	**90.6**	**90.8**	**93.4**
MediaEval VMU					
BERT [13]	82.3	81.7	81.0	81.3	84.7
RoBERTa [14]	84.9	84.2	83.6	83.9	86.8
XLNet [15]	85.6	84.9	84.2	84.5	87.4
ResNet [17]	78.9	78.2	77.5	77.8	81.8
EfficientNet [18]	80.1	79.5	78.9	79.2	83.1
ViT [19]	82.8	82.1	81.5	81.8	85.3
Multimodal-BERT [20]	86.0	85.3	84.7	85.0	88.2
ViLBERT [21]	87.8	87.1	86.4	86.7	89.1
QEMF [23]	88.8	88.1	87.6	87.8	90.3
Q-ALIGNer (ours)	**92.1**	**91.6**	**91.0**	**91.3**	**94.0**

Applying Q-ALIGNer on the structured news stories and images of "FakeNewsNet" returns 91.2% accuracy and 90.4% F1—both greater than QEMF at 88.5% and ViLBERT at 87.1%. This shows that entanglement-based fusion provides a better representation of the semantic dependencies between long-form news text and visual evidence. Our results on the large, noisy "Fakeddit" dataset shows Q-ALIGNer returns 92.9% accuracy and 91.9% F1, and yet, these results are significantly higher than QEMF at 89.7%. The results here also indicate that our model is able to scale performance across a large amount of heterogeneous social media content. On the analyzed Weibo dataset, constructed of short and informal posts in Chinese, Q-ALIGNer returns an accuracy of 91.7% and 90.8%, superior to QEMF at 87.9%. This shows, again, that our model is language agnostic and cross-stylistically robust. Lastly, on the MediaEval VMU dataset, real-time Twitter captures of misinformation, Q-ALIGNer achieves 92.1% accuracy and 91.3% F1, advancing both ViLBERT at 87.8% and QEMF at 88.8%.

4.5 Dataset-Wise Results

Comparisons are presented on a dataset basis in Tables 2–5. In Fig. 6, Q-ALIGNer shows greater accuracy in all four datasets, indicating the effectiveness and generalization of the approach compared to unimodal and multimodal baselines.

Table 2: Performance on fakeNewsNet dataset.

Model	Accuracy	Precision	Recall	F1
BERT	81.4	80.5	79.8	80.1
RoBERTa	83.6	82.9	82.1	82.5
ViLBERT	87.1	86.3	85.9	86.1
QEMF	88.5	87.9	87.2	87.5
Q-ALIGNer	91.2	90.7	90.1	90.4

Table 3: Performance on fakeddit dataset.

Model	Accuracy	Precision	Recall	F1
BERT	82.7	81.9	81.3	81.6
RoBERTa	85.4	84.6	83.9	84.2
ViLBERT	88.6	87.9	87.2	87.5
QEMF	89.7	89.0	88.4	88.7
Q-ALIGNer	92.9	92.3	91.6	91.9

Table 4: Performance on Weibo dataset.

Model	Accuracy	Precision	Recall	F1
BERT	80.9	80.1	79.5	79.8
RoBERTa	83.1	82.4	81.8	82.1
ViLBERT	86.5	85.9	85.3	85.6
QEMF	87.9	87.2	86.7	86.9
Q-ALIGNer	91.7	91.1	90.6	90.8

Table 5: Performance on mediaEval VMU dataset.

Model	Accuracy	Precision	Recall	F1
BERT	82.3	81.7	81.0	81.3
RoBERTa	84.9	84.2	83.6	83.9
ViLBERT	87.8	87.1	86.4	86.7
QEMF	88.8	88.1	87.6	87.8
Q-ALIGNer	92.1	91.6	91.0	91.3

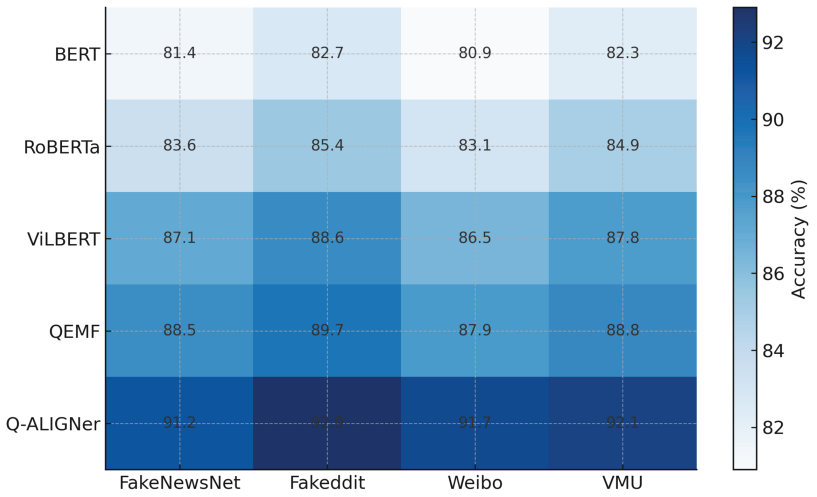

Figure 6: Dataset-wise accuracy heatmap comparing baseline models with Q-ALIGNer across FakeNewsNet, Fakeddit, Weibo, and MediaEval VMU datasets.

In the FakeNewsNet dataset illustrated in Table 2, which consists of long-form articles of news with images, Q-ALIGNer achieves an accuracy of 91.2%, providing an improvement over ViLBERT (87.1%) and QEMF (88.5%). The increases in precision, recall, and F1 indicate that Q-ALIGNer captures the semantic relations that exist between the article text and associated multimedia. The incorporation of quantum entanglement at the fusion stage of learning allowed for a tighter coupling to potentially exist between the text and the accompanying media, facilitating discriminative performance. These results confirm that Q-ALIGNer handles structured, article-style misinformation more effectively than both classical and prior quantum-enhanced baselines.

On the Fakeddit dataset in Table 3, which is substantially larger and noisier due to the diversity of Reddit posts, Q-ALIGNer achieves 92.9% accuracy and an F1-score of 91.9. This performance exceeds QEMF by over three percentage points, showing that the model scales well to high-volume, heterogeneous social media data. While transformer-based models such as RoBERTa and multimodal models like ViLBERT provide competitive baselines, they fail to reach the same robustness as Q-ALIGNer in handling noisy and contextually inconsistent multimodal posts. The results highlight the benefits of quantum-aligned contrastive learning for managing variability in large-scale misinformation

The Weibo dataset presents a different challenge as it consists of short, informal Chinese-language posts paired with images as shown in Table 4. Here, Q-ALIGNer achieves 91.7% accuracy and 90.8 F1, improving upon QEMF's 87.9%. The cross-lingual success of Q-ALIGNer indicates that the quantum entanglement mechanism generalizes beyond English-language corpora, successfully adapting to structurally different and linguistically diverse data. This demonstrates that Q-ALIGNer is not only effective in high-resource settings but also capable of adapting to multilingual misinformation contexts.

On the MediaEval VMU dataset in Table 5, which consists of Twitter posts with fast-changing and often noisy multimodal content, Q-ALIGNer achieves 92.1% accuracy and 91.3 F1, outperforming QEMF (88.8%) and ViLBERT (87.8%). The improvement is particularly notable given the challenging nature of microblogging platforms, where both text and images tend to be short, informal, and contextually ambiguous. Q-ALIGNer's entanglement-driven fusion allows the model to maintain consistency between textual claims and visual evidence, leading to more reliable predictions under high-noise conditions. These results illustrate the robustness of the proposed method in real-time, high-velocity misinformation scenarios.

Even though Q-ALIGNer excels across multiple benchmarks, each benchmark is very different in its characteristics (domain), official language, platform dynamics, and structure of events. FakeNewsNet contains only long-form articles, while Fakeddit has large volumes of noisy Reddit posts and short informal Chinese posts. In addition, MediaEval VMU is fast-paced and focuses on Twitter-based verification tasks. The fact that Q-ALIGNer has consistently improved results across these diverse benchmark sets shows that the entanglement-based mechanism can learn to create complex connections between similar concepts regardless of the type of media used in a dataset. The comparative evaluation emphasizes architectural relevance and reproducibility rather than publication year alone, allowing the proposed framework to be assessed against established and competitive multimodal baselines under consistent experimental conditions.

4.6 Ablation Studies

To assess the contributions, we complete ablation studies on dataset and thus impose to remove major parts of Q-ALIGNer in a sequential manner: the entanglement operation, the InfoNCE alignment loss, the robustness loss, and the quantum encoding. As can be observed in Fig. 7, the related studies on all datasets confirm that quantum entanglement should be considered in the model as the most essential part of Q-ALIGNer, and the InfoNCE alignment and robustness-aware objectives further contribute to sustained improvements in performance.

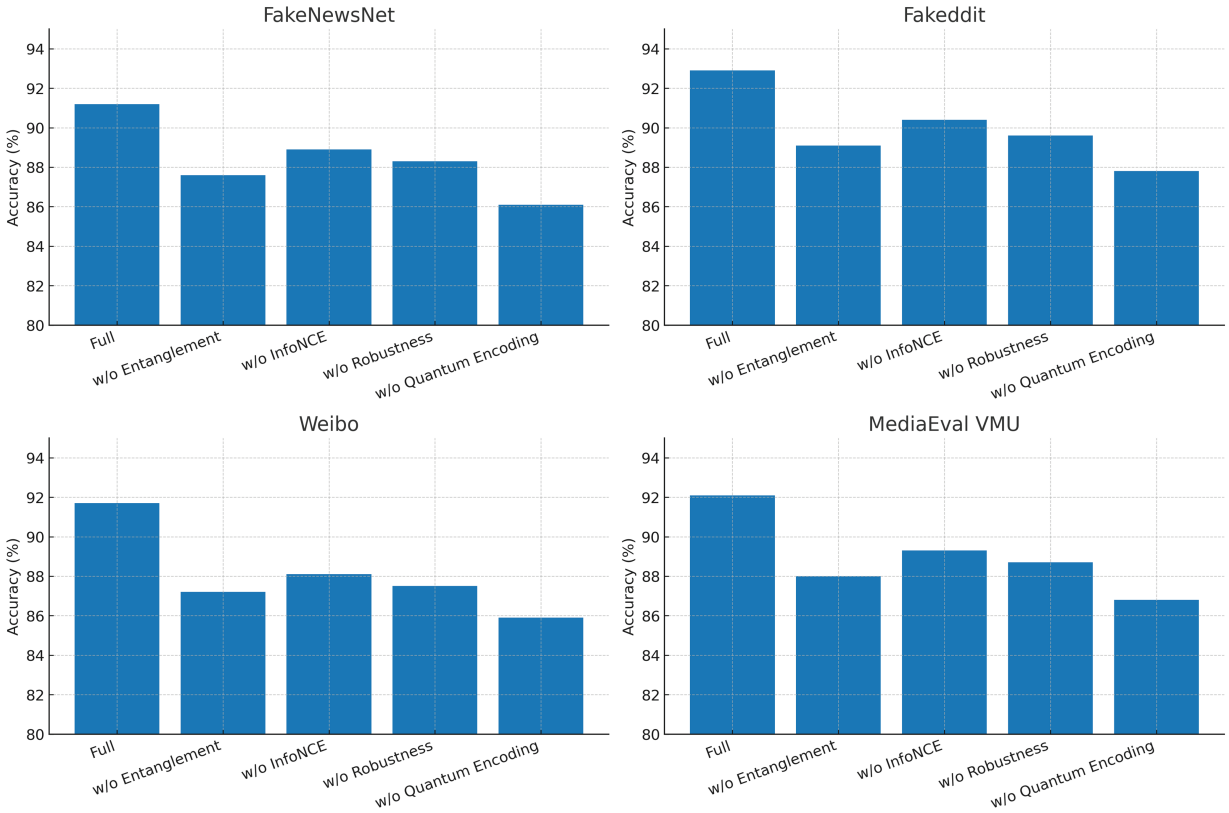

Figure 7: Ablation studies of Q-ALIGNer across four datasets.

Entanglement shows the most significant drop in accuracy when removed from the framework in Table 6 on FakeNewsNet. The accuracy drops from 91.2% to 87.6%. This drop illustrates the necessity of learning cross-modal dependencies with quantum entanglement to successfully validate long-form articles.

The addition of both InfoNCE loss and robustness loss also show to improve overall performance because the InfoNCE loss encourages a stronger alignment between modal streams in the fact-checking task, while the robustness loss improves the general stability of the model. Finally, the removal of quantum encoding from the fact-checking baseline resulted in the lowest overall performance, which emphasizes the importance of being able to carry out quantum-native feature mapping when reasoning with structured misinformation.

Table 6: Ablation Study sf Q-ALIGNer on fakeNewsNet dataset.

Model Variant	Accuracy	F1
Full Q-ALIGNer	91.2	90.4
w/o Entanglement	87.6	86.8
w/o InfoNCE loss	88.9	88.1
w/o Robustness loss	88.3	87.5
w/o Quantum Encoding (classical fusion)	86.1	85.4

For the large-scale Fakeddit dataset in Table 7, we again see that entanglement is the most influential factor and its removal resulted in a 3.8 point accuracy drop. Thus, quantum entanglement is useful in increasing robustness for multimodal posts that are noisy and contextually inconsistent. The robustness loss, improves resilience to standard social media perturbations, while InfoNCE alignment helps contribute to decreasing modality mismatch. Without quantum encoding the model approximates to a more classical baseline thus highlighting that quantum feature embedding allows for scalability even in large and noisy data conditions.

Table 7: Ablation study of Q-ALIGNer on fakeddit dataset.

Model Variant	Accuracy	F1
Full Q-ALIGNer	92.9	91.9
w/o Entanglement	89.1	88.3
w/o InfoNCE loss	90.4	89.5
w/o Robustness loss	89.6	88.9
w/o Quantum encoding (classical fusion)	87.8	87.0

Weibo is a set of casual, short-form Chinese posts, and here, Q-ALIGNer shows that all parts are responsible for cross-lingual generalization, as shown in Table 8. The removal of entanglement induces the largest drop in accuracy, falling by 4.5 points. InfoNCE alignment loss is also an important contributing factor in ensuring consistency between short pieces of text and its corresponding image. The robustness loss helps in increasing tolerance to adversarial noise often found in user-generated content. Removing quantum encoding generates a performance drop to below 86% which suggests that quantum representations capture semantic distinctions that would be obscured by classical fusion.

Table 8: Ablation study of Q-ALIGNer on Weibo dataset.

Model Variant	Accuracy	F1
Full Q-ALIGNer	91.7	90.8
w/o Entanglement	87.2	86.4
w/o InfoNCE loss	88.1	87.3
w/o Robustness loss	87.5	86.7
w/o Quantum encoding (classical fusion)	85.9	85.1

In the MediaEval VMU dataset, shown in Table 9, which is comprised of short bits of Twitter, Q-ALIGNer with full architecture is again the best. If entanglement were removed, accuracy falls to 88.0%, showing how important it is to model the fine-grained dependencies between claims and images in the noise and continuous stream of social media data. Both losses (InfoNCE and robustness) also incrementally improve the claim-image alignment over the no-loss baseline, with a drop of only 2–3 points. The performance is worst, when using classic fusion without quantum encoding, again supporting the finding that the quantum state space leads to better representations of multimodal correlations.

Table 9: Ablation study of Q-ALIGNer on mediaEval VMU dataset.

Model Variant	Accuracy	F1
Full Q-ALIGNer	92.1	91.3
w/o Entanglement	88.0	87.2
w/o InfoNCE loss	89.3	88.5
w/o Robustness loss	88.7	87.9
w/o Quantum encoding (classical fusion)	86.8	86.0

Furthermore, the findings from the ablation experiments demonstrate the architectural complexity brought about by encodings and entanglements is not wasted or unnecessary, but offers a different set of functionalities which can be exploited to improve a system's ability and durability during training on a number of datasets.

4.7 Adversarial Robustness

The adversarial robustness findings are presented in Table 10. The standard baselines on Weibo experience a tremendous drop in performance when under attacks (e.g., FGSM, PGD) with a decrease of over 7–9 points in accuracy. In contrast, Q-Aligner achieves 89.1% accuracy, which means that Q-Aligner has a robustness gap of only 2.6 points. The significant gap in robustness demonstrates that the entanglement-based design with training using robustness techniques allows Q-Aligner to withstand manipulation considerably better than both the classical benchmarks and the quantum benchmarks. This is a vital feature for dealing with adversarially manipulated misinformation in practical applications.

Table 10: Robustness evaluation under adversarial attacks on Weibo dataset.

Attack Type	Clean ACC	Adv ACC	Robustness Gap
FGSM (images)	91.7	84.2	7.5

(Continued)

Table 10 (continued)

Attack Type	Clean ACC	Adv ACC	Robustness Gap
PGD (images)	91.7	82.9	8.8
Synonym substitution (text)	91.7	83.7	8.0
Char-level noise (text)	91.7	85.0	6.7
Q-ALIGNer (robust)	91.7	89.1	2.6

We note that robustness variation curves under progressively increasing perturbation strengths were not explored in this study; the evaluation instead focuses on relative robustness under standardized attack settings, with extended sensitivity analysis reserved for future work. The most common form of assessment for Misinformation comes from Social Media. The type of assessments used by social media sites simulate types of Misinformation such as Paraphrasing, Lexical Noise, and Visual Perturbation. All of the assessments represent the long-term and chaotic environments associated with the development of misinformation. The adversarial scenarios examined in this study utilise reproducible and widely used perturbation approaches in Multimodal Misinformation Research. The types of perturbations addressed include text paraphrase, character-level noise (i.e., randomised additions or removals of characters), and gradient based image manipulations; All of which are commonly used as real-world manipulation patterns. While these perturbation approaches capture many real-world manipulation patterns, many more sophisticated adversarial attack strategies exist that are not specifically assessed in this study, including visual manipulation through deepfakes, coordinated inconsistency across multiple modalities of attack, and large-scale content rewriting, all of which represent promising areas for future research.

4.8 Uncertainty and Calibration

The performance of calibration is shown in Table 11 (calibration). Q-ALIGNer achieves an expected calibration error (ECE) of 0.031 and negative log-likelihood (NLL) of 0.228, outperforming both ViLBERT (ECE 0.058) and QEMF (ECE 0.046). All of these results demonstrate that Q-ALIGNer generates well calibrated probabilities that closely follow empirical accuracy. Good uncertainty estimation is especially important for fake news detection, as it allows the model to signal uncertain cases for human verification instead of making confident wrong predictions.

Table 11: Uncertainty calibration comparison on fakeNewsNet dataset. lower is better.

Model	ECE	NLL
BERT	0.071	0.325
ViLBERT	0.058	0.294
QEMF	0.046	0.267
Q-ALIGNer	0.031	0.228

From a usage perspective, users will find it easier to understand a calibrated confidence score and measurements of consistency than raw multimodal attention weights or latent representation embeddings. Calibrated confidence scores and measurements of consistency serve as clear and understandable indicators to guide the user's decision-making in these practical applications.

4.9 Computational Efficiency

Table 12 displays comparisons for computational efficiency. Q-ALIGNer attained a significantly higher training efficiency than QEMF, requiring 19.6 h compared to QEMF's 48.2 h. This substantial reduction is primarily a direct consequence of having learnable entangling unitaries that occur with lower gate complexity. Q-ALIGNer also has competitive parameter counts (118M), while being less expensive computationally—without sacrificing accuracy. These results indicate that quantum-inspired designs can deliver significantly improved accuracy and robustness, in addition to rational and meaningful gains in computational efficiency.

Table 12: Computational efficiency comparison. lower time and complexity are better.

Model	Training time (h)	Parameters (M)	Complexity
BERT	11.5	110	$O(n^2 d)$
ViLBERT	24.7	210	$O(n^2 d)$
QEMF	48.2	125	$O(n^3)$
Q-ALIGNer	19.6	118	$O((L' + P')^2 d)$

The Q-ALIGNer provides more complex architectures by adding quantum state encoding and creating cross modal entanglement. These additional components have complexity that is intentionally limited and has been thoroughly characterized analytically. Therefore, The Q-ALIGNer's architecture is a balance of architectural simplicity and representational expressivity; while providing a greater ability to represent rich multimodal dependencies, it is still computationally feasible to perform with pruning strategies, shallow variational circuits, and efficient measurement schemes. Although Q-ALIGNer increases the computational efficiency of quantum-enhanced frameworks over previous quantum-enhanced approaches, Q-ALIGNer does not consider being directly deployed on a web-scale volume involving hundreds of millions of postings. The methods used to reduce complexity in our approach—specifically input pruning, shared latent projections, shallow variational circuits, and bilinear entanglement—significantly reduce the overhead associated with state-encoding and fusion, and further optimizations at the system level would be required for large-scale applications.

4.10 Discussion

Results from multiple benchmarks of experiments continuously show that the Q-ALIGNer framework is better than classical and quantum-inspired baselines. The gains relate to predictive accuracy, robustness to adversarial perturbations, reliability in estimating uncertainties, and computational costs. Nevertheless, explicit evaluation on entirely unseen domains, emerging event types, or zero-shot transfer settings remains beyond the scope of this work and will be explored in future research. While the benchmarks that have been assessed have been developed using dual feature multi-modality data, there are still a significant number of outstanding challenges pertaining to the use of missing or grossly impaired modalities when utilizing an uncontrolled or social environment. While this study includes both English and non-English datasets, future work will extend Q-ALIGNer to a wider range of languages, regions, and culturally specific misinformation scenarios to more comprehensively assess global generalization. From an application perspective, this behavior enables Q-ALIGNer to function as a decision-support system that highlights ambiguous or contested cases, supporting human-in-the-loop verification rather than enforcing overconfident automated judgments.

Dataset-wide evaluations showed that Q-ALIGNer adapts to multiple misinformation contexts. On FakeNewsNet, the model efficiently merges long-form text about news stories with attached images. In the case of Fakeddit, adaptability in a noisy, social media space is accounted for while still scaling and retaining

familiarity to its fine-tuning stage. Dealing with a cross-lingual context for Weibo allowed Q-ALIGNer to train efficiently on a large chunk of original data with short, informal posts and showed that the model generalizes well beyond English-language corpora. On MediaEval VMU, live Twitter streams of misinformation where contextual ambiguity is an issue, Q-ALIGNer was resilient against the noise to outperform all baselines. These leases imply consistent improvements that demonstrated that Q-ALIGNer is not only functional in a high-resource and structured dataset but also in noisy, complicated, and multilingual spaces. Ablation studies further demonstrated the contribution of individual components. Removing quantum entanglement produced the most significant drop in performance, and consistent in every thinking model type, which shows the importance of directionality in modeling cross-modal dependencies. The complementary effects of the InfoNCE loss for alignment act to produce better results by enforcing some modality consistency, but the robustness loss produced significant improvements in adversarial perturbations. When comparing quantum-encoded data vs. replacing with classical fusion, classical representations perform the worst, which shows the ability of quantum states to create more expression with features.

Beyond accuracy, Q-ALIGNer performed significantly better as an advantage of robustness, and reliability. After adversarial attacks including FGSM, PGD, etc., synonym substitution, and character-level noise, Q-ALIGNer produced significantly higher accuracy after an attack compared to baselines, and bridged a gap of 50% improvement of robustness after the attack. This is a significant finding when thinking of real-world scenarios for misinformation papers, as the content was originally designed to be hacked. Q-ALIGNer was tested and demonstrated comparable results to cavalier, calibrated for both performance mean and negative log likelihood, respectively, demonstrating that less prediction uncertainty when compared to both models and during calibration experiments. Well-calibrated predictions are important in applications when there are uncertain cases that require review. Finally, Q-ALIGNer produced additional performance gains with efficiency. QEMF produced increased training time and gate complexity to obtain performance. Q-ALIGNer improved the overall training time by more than half, yet both trained models were refined using a similar parameter scale. The design of learnable entangling unitary gate strategies reduced circuit depth while maintaining expressive power. Overall, these observations exhibited quantum-inspired architectures provide the option to regain efficiency while providing prediction settings that are both precise and accurate, promoting further deployment in real-world fake news systems.

4.11 Limitations

This research indicates that misinformation detection will be aided by experimentation of the verification process. However, several limitations have been identified in this research. A content-based approach to identifying misinformation through paired text and images was utilised by Q-ALIGNer. Q-ALIGNer's use of a content-focused approach means that the tool does not capture temporal changes in information, the way that the information is shared or the community in which the information was originally shared. Although this allows Q-ALIGNer to support early-stage evaluations of misinformation detection during the absence of the aforementioned metadata, the lack of focus on temporal (temporal) and community (social) factors can hinder Q-ALIGNer's capabilities to model the dynamic nature of misinformation over time. The evaluation of Q-ALIGNer consists of the use of many disparate (i.e., multi-language) benchmark tests that span across multiple platforms (i.e., Facebook, Twitter), and provides insight into a limited view of the global dynamics of misinformation. As a result, additional research is required in order to develop a Q-ALIGNer that has the ability to generalise, or apply its capabilities to topics and events that are entirely foreign to the Q-ALIGNer.

Regarding the evaluation of robustness, Q-ALIGNer focused on the evaluation of examples of how Q-ALIGNers can be controlled, created, and modified. The examples of the two aforementioned methods

include gradient-based perturbation attacks on images and the creation of written text via semantic-preserving manipulations of the text. Examples of the types of attacks that can have a negative effect on the effectiveness of the tool (Q-ALIGNer) are complex and probabilistic in nature, e.g., large-scale coordinated attacks (i.e., coordinated use of multiple, diverse media attacks against Q-ALIGNers or deepfake images), and manipulation of content through a large quantity of co-incident events. Using a simulated Quantum-Inspired Framework-Based Implementation, Q-ALIGNer's approach provides insight into how the tool has been constructed and evaluated; however, it does not provide any insight into how Q-ALIGNers has been implemented in practice, as it has not run on actual Quantum hardware (e.g., Quantum Circuits). The potential advantages and disadvantages of running Q-ALIGNers in practice on Quantum Circuits remains an area for future investigations. While Q-ALIGNer has vastly improved in efficiency compared to previous Quantum-Tools (AI's and MLs), a considerably higher level of optimisation, resulting from distributed training and Compressing of Q-ALIGNers' output (i.e., embeddings), must be achieved before deploying Q-ALIGNers at Web-scale (i.e., involving hundreds or thousands of millions of Posts). The scope of such activities is not the focus of this article.

5 Conclusion

Q-ALIGNer is an innovative, quantum-inspired, multi-modal tool designed to detect false news. It combines classical techniques for extracting features of fake stories with capabilities to encode them as quantum states, train neural networks to learn how to make the two feature types cohere, and to use training objectives to help improve the system's resilience against uncertainty. The results of extensive experimentation using four established benchmark datasets (FakeNewsNet, Fakeddit, Weibo, and MediaEval VMU) confirm that Q-ALIGNer performs better than competing text-based and image-based models and multisource/quantum-inspired models. In particular, it outperformed the previous quantum-inspired QEMF model by incorporating improved methods for using adaptive entanglement, contrastive aligning, and calibrating uncertainty when detecting multi-modal false information. The individual components of Q-ALIGNer were also evaluated through ablation studies, which demonstrated the centrality of cross-modal entanglement when providing semantic dependence between image and textual inputs. The robustness tests showed that Q-ALIGNer has significantly greater resistance to adversarial attacks on both text and visual input modalities than other state-of-the-art systems. Improved calibration of confidence values was also demonstrated by testing against other state-of-the-art systems in the areas of text and image input modalities. Additionally, Q-ALIGNer demonstrated improvements in efficiency regarding training time and complexity, making it a more applicable tool to large-scale benchmarks when compared to previous quantum-inspired systems.

Unlike other systems that expose users to lower-level fusion mechanisms, Q-ALIGNer emphasizes interpretability at the decision level by providing users with confidence, consistency, and alignment signals that are conducive to developing trust-based and human-verifiable verification systems. The mathematical complexity of internal models is not required to provide decision-making support; instead, these capabilities allow users to receive usable output from these sophisticated mathematical processes without requiring expertise in quantum computing. This work primarily adopted an information-centric point of view and did not include a temporal analysis of the dynamics of social and contextual propagation of misinformation or of misinformation pattern evolution over time. Future work will likely build on these approaches to add temporal context modeling, social context features, and evaluations across multiple languages and cultures. Other future work directions will include examining deployment on new quantum platforms, increasing the robustness testing methods for evaluating advanced multi-modal adversarial cases, and creating adaptive methods for dealing with missing or poor-quality multi-modal input. Enhancements to scalability using

distributed training and compression of embeddings will be examined to allow for real time application in high-throughput social media environments. Future work will also extend the experimental evaluation to include recent CLIP-based and large-scale multimodal foundation models as standardized and reproducible benchmarks for misinformation detection become available.

Acknowledgement: The authors acknowledge the support of Princess Nourah bint Abdulrahman University and Northern Border University.

Funding Statement: Princess Nourah bint Abdulrahman University Researchers Supporting Project number (PNURSP2026R77), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia, the Deanship of Scientific Research at Northern Border University, Arar, Saudi Arabia, through the project number NBU-FFR-2026-2248-02.

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Sara Tehsin, Inzamam Mashood Nasir, Wiem Abdelbaki; data collection: Fadwa Alrowais, Reham Abualhamayel, Abdulsamad Ebrahim Yahya; analysis and interpretation of results: Sara Tehsin, Inzamam Mashood Nasir, Radwa Marzouk; draft manuscript preparation: Wiem Abdelbaki, Fadwa Alrowais, Radwa Marzouk. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: The implementation of this work can be downloaded from https://github.com/imashoodnasir/Q-ALIGNer-Robust-Fake-News-Detection.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Wu F, Chen S, Gao G, Ji Y, Jing XY. Balanced multi-modal learning with hierarchical fusion for fake news detection. Pattern Recognit. 2025;164(6380):111485. doi:10.1016/j.patcog.2025.111485.
2. Shen X, Huang M, Hu Z, Cai S, Zhou T. Multimodal fake news detection with contrastive learning and optimal transport. Front Comput Sci. 2024;6:1473457. doi:10.3389/fcomp.2024.1473457.
3. Parmar S, Rahul. Fake news detection via graph-based Markov chains. Int J Inf Technol. 2024;16(3):1333–45. doi:10.1007/s41870-023-01558-3.
4. Liu Y, Liu Y, Li Z, Yao R, Zhang Y, Wang D. Modality interactive mixture-of-experts for fake news detection. In: Proceedings of the ACM on Web Conference 2025; 2025 Apr 28–May 2; Sydney, NSW, Australia. p. 5139–50. doi:10.1145/3696410.3714522.
5. Chen H, Yu Y, Guo H, Hu B, Hu S, Hu J, et al. A self-learning multimodal approach for fake news detection. Front Artif Intell. 2025;8:1665798. doi:10.3389/frai.2025.1665798.
6. Choi E, Ahn J, Piao X, Kim JK. Crome: multimodal fake news detection using cross-modal tri-transformer and metric learning. arXiv:2501.12422. 2025.
7. Qu Z, Meng Y, Muhammad G, Tiwari P. QMFND: a quantum multimodal fusion-based fake news detection model for social media. Inf Fusion. 2024;104(9):102172. doi:10.1016/j.inffus.2023.102172.
8. Altıntaş V. Beyond classical AI: detecting fake news with hybrid quantum neural networks. Appl Sci. 2025;15(15):8300. doi:10.3390/app15158300.
9. Tufchi S, Yadav A, Ahmed T. A comprehensive survey of multimodal fake news detection techniques: advances, challenges, and opportunities. Int J Multimed Inf Retr. 2023;12(2):28. doi:10.1007/s13735-023-00296-3.
10. Abduljaleel IQ, Ali IH. Deep learning and fusion mechanism-based multimodal fake news detection methodologies: a review. Eng Technol Appl Sci Res. 2024;14(4):15665–75. doi:10.48084/etasr.7907.
11. Song C, Ning N, Zhang Y, Wu B. A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. Inf Process Manag. 2021;58(1):102437. doi:10.1016/j.ipm.2020.102437.

12. Hao X, Xu W, Huang X, Sheng Z, Yan H. MFUIE: a fake news detection model based on multimodal features and user information enhancement. EAI Endorsed Trans Scalable Inf Syst. 2025;11:1–13. doi:10.4108/eetsis.7517.
13. Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies; 2019 Jun 2–7; Minneapolis, MN, USA. p. 4171–86.
14. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, et al. Roberta: a robustly optimized bert pretraining approach. arXiv:1907.11692. 2019.
15. Yang Z, Dai Z, Yang Y, Carbonell J, Salakhutdinov RR, Le QV. Xlnet: generalized autoregressive pretraining for language understanding. arXiv:1906.08237. 2019.
16. Graves A. Long short-term memory. In: Supervised sequence labelling with recurrent neural networks. Berlin/Heidelberg, Germany: Springer; 2012. p. 37–45.
17. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 Jun 27–30; Las Vegas, NV, USA. p. 770–8. doi:10.1109/cvpr.2016.90.
18. Tan M. Rethinking model scaling for convolutional neural networks. arXiv:1905.11946. 2019.
19. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An image is worth 16x16 words: transformers for image recognition at scale. arXiv:2010.11929. 2020.
20. Sun C, Myers A, Vondrick C, Murphy K, Schmid C. VideoBERT: a joint model for video and language representation learning. In: Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV); 2019 Oct 27–Nov 2; Seoul, Republic of Korea. p. 7464–73.
21. Lu J, Batra D, Parikh D, Lee S. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. arXiv:1908.02265. 2019.
22. Mukesh K, Jayaprakash SL, Kumar RP. QViLa: quantum infused vision-language model for enhanced multimodal understanding. SN Comput Sci. 2024;5(8):1023. doi:10.1007/s42979-024-03398-9.
23. Bikku T, Thota S. Quantum-enhanced multimodal fusion for robust and accurate fake news detection. Sigma. 2025;43(3):943–54. doi:10.14744/sigma.2025.00082.