

Explainable Multimodal Transformer Architecture for Context-Aware Decision Intelligence in Smart Applications

Prof. Weisong Shi

Alumni Distinguished Professor & Chair, Computer & Information Sciences, University of Delaware, USA

Email: weisong@udel.edu

Article Received 11-07-2026, Revised 17-08-2026, Accepted 04-09-2026

Author(s) Retains the Copyrights of This Article

Abstract

Smart applications increasingly depend on information obtained from multiple sources, including images, text, speech, sensor measurements, video streams, and contextual information. Although combining these modalities can provide a more comprehensive representation of a user's environment, conventional multimodal learning systems frequently provide limited insight into how individual information sources influence their final decisions. This lack of transparency can restrict the deployment of intelligent systems in applications where decisions need to be understood, verified, or reviewed by human users. This research proposes an explainable multimodal Transformer architecture for context-aware decision intelligence in smart applications. The proposed architecture employs separate modality encoders to transform heterogeneous inputs into compatible representations, followed by cross-modal Transformer layers that learn relationships among visual, textual, auditory, and contextual features. A context integration mechanism is introduced to dynamically adjust the contribution of each modality according to the current application state and available evidence. Instead of generating a prediction alone, the architecture produces an associated explanation identifying the principal modalities, features, and contextual elements contributing to the decision. A hierarchical explanation mechanism provides both modality-level and feature-level interpretations, allowing users to examine why a particular decision was generated and which information sources had the greatest influence. The framework is designed to handle incomplete or unevenly informative modalities by dynamically modifying their contribution during inference. Its performance can be evaluated using predictive accuracy together with explanation fidelity, modality contribution consistency, robustness, context sensitivity, and computational efficiency. The proposed architecture provides a unified approach for combining multimodal perception with interpretable decision intelligence and can support smart healthcare, intelligent transportation, smart homes, human-computer interaction, industrial monitoring, and context-aware assistance. The study establishes a framework in which multimodal Transformer reasoning is accompanied by evidence-oriented explanations suitable for human inspection.

Keywords: Explainable Artificial Intelligence, Multimodal Transformer, Context-Aware Intelligence, Multimodal Fusion, Decision Intelligence, Cross-Modal Attention, Interpretable Deep Learning, Smart Applications, Modality Attribution, Contextual Reasoning, Multimodal Representation Learning, Explainable Decision-Making.

1. Introduction

The rapid development of connected sensors, intelligent cameras, wearable devices, mobile platforms, conversational interfaces, and edge computing has resulted in smart applications that operate on information originating from multiple heterogeneous sources. A modern intelligent application may simultaneously receive an image from a camera, a textual description from a user, speech from a microphone, measurements from sensors, and contextual information describing time, location, environmental conditions, or application state. Processing these information sources independently can provide only partial knowledge of the situation, whereas multimodal learning attempts to combine

complementary evidence to obtain a richer representation. The Transformer architecture introduced by Vaswani et al. demonstrated that attention-based processing can capture relationships among information elements without relying on conventional recurrent structures, creating an important foundation for subsequent multimodal architectures [1].

Language-oriented Transformer models further demonstrated the ability of attention mechanisms to capture contextual relationships within large-scale representations. Devlin et al. introduced BERT, showing that bidirectional Transformer representations can support a range of language understanding tasks [2]. The success of Transformer-

based language models subsequently encouraged researchers to investigate methods for connecting language representations with visual information. Lu et al. developed ViLBERT using separate visual and linguistic streams connected through co-attentional Transformer layers, demonstrating that information from different modalities can be jointly processed while retaining modality-specific representations [3]. Tan and Bansal proposed LXMERT, which explicitly modeled intra-modal and cross-modal relationships between language and visual representations [4].

Other multimodal architectures further demonstrated the potential of joint representation learning. VisualBERT incorporated visual regions into Transformer processing to establish relationships between visual and textual information [5]. UNITER investigated universal image-text representations using large-scale multimodal pretraining and word-region alignment [6]. Radford et al. demonstrated the potential of large-scale image-text representation learning through CLIP, which established relationships between natural-language descriptions and visual representations and supported transfer to different visual tasks [7]. Alayrac et al. subsequently presented Flamingo, demonstrating few-shot multimodal learning through an architecture capable of combining visual information with language generation [8].

Despite these advances, prediction accuracy alone is insufficient for many smart applications. In healthcare, industrial monitoring, transportation, and other decision-sensitive environments, users may need to understand which evidence contributed to a system's output. Park et al. demonstrated the value of multimodal explanations by combining visual evidence with textual rationales to justify decisions [10]. Lundberg and Lee introduced SHAP as a unified framework for feature attribution, providing a basis for examining the contribution of individual input features to model outputs [11]. These developments indicate that multimodal intelligence should not be limited to combining heterogeneous information; it should also provide interpretable evidence associated with its decisions.

The problem becomes more complex when the reliability and usefulness of each modality vary according to context. A camera may provide strong evidence in a well-lit environment but unreliable information in poor visibility. Speech may be highly informative when the audio channel is clear but less useful under substantial background noise. Sensor measurements may provide reliable contextual information while remaining insufficient for identifying visual events. Consequently, a fixed fusion

mechanism may not be appropriate for all operating conditions. A context-aware architecture should be capable of adjusting modality contributions according to the available evidence and the current application state.

Multimodal learning has developed rapidly, but the explainability of multimodal Transformer systems remains an active research problem. Xu, Zhu, and Clifton surveyed Transformer-based multimodal learning and identified challenges involving multimodal representation, fusion, pretraining, alignment, and deployment [9]. More recent work has also examined methods for evaluating modality-level contributions and interpreting Transformer-based multimodal models [17], [19], [20]. These developments motivate the design of a framework in which multimodal representation, context-aware fusion, decision generation, and explanation generation are considered as interconnected components.

The present research proposes an explainable multimodal Transformer architecture designed for context-aware decision intelligence. The framework receives heterogeneous information through modality-specific encoders, maps the resulting representations into a common latent space, and applies cross-modal Transformer processing to learn relationships between modalities. A context integration mechanism estimates the relevance of each modality under the current operating condition. The resulting representations are then used for decision generation and hierarchical explanation. The explanation mechanism identifies the dominant modalities and influential features associated with the generated decision.

The research addresses four related objectives. First, it develops a multimodal representation mechanism capable of integrating heterogeneous visual, textual, auditory, sensor, and contextual information. Second, it introduces context-guided fusion to dynamically adjust modality contributions according to application conditions. Third, it incorporates evidence attribution to connect the generated decision with the information that influenced it. Fourth, it proposes an evaluation methodology that measures predictive performance, explanation fidelity, modality contribution consistency, robustness, context sensitivity, and computational efficiency.

The proposed framework is intended for smart applications in which heterogeneous information must be combined while maintaining a degree of transparency. Rather than treating explainability as an independent visualization added after prediction, the proposed architecture considers explanation as an

integrated component of the decision process. This design provides a basis for developing intelligent systems that can produce both a decision and an evidence-oriented interpretation of that decision.

2. Related Work

Transformer-based architectures have become an important foundation for multimodal representation learning. Vaswani et al. introduced the Transformer through an attention-based architecture capable of modeling relationships between sequence elements without recurrent computation [1]. The self-attention mechanism allows a representation to consider information from multiple positions, making it suitable for learning long-range relationships. Devlin et al. subsequently demonstrated the effectiveness of bidirectional Transformer representations in language understanding through BERT [2]. These developments established the architectural principles later adopted by multimodal systems in which information from different sources must be aligned and jointly processed.

Lu et al. extended Transformer-based representation learning to vision-language processing through ViLBERT. Their architecture maintained separate streams for visual and textual information while allowing interaction through co-attentional Transformer layers [3]. This design is particularly relevant to multimodal decision intelligence because it demonstrates that heterogeneous representations can retain modality-specific characteristics while still participating in cross-modal reasoning. Tan and Bansal proposed LXMERT, which explicitly modeled relationships within visual and language modalities as well as relationships between the modalities [4]. The architecture demonstrated that multimodal reasoning benefits from distinguishing intra-modal processing from cross-modal interaction.

VisualBERT provided another approach to vision-language representation by integrating visual regions into Transformer-based processing and learning relationships between visual and linguistic elements [5]. Chen et al. developed UNITER, which focused on universal image-text representation learning and explicit alignment between textual words and image regions [6]. These approaches established the importance of representation alignment when information originating from different modalities is incorporated into a common Transformer framework. Radford et al. introduced CLIP, demonstrating that large-scale image-text learning can produce transferable visual representations using natural-language supervision [7]. The significance of this approach extends beyond image classification because

it demonstrated that semantic relationships between modalities can be learned at scale. Alayrac et al. subsequently introduced Flamingo, which combined visual information with language modeling and demonstrated few-shot multimodal learning capabilities [8]. These models illustrate the increasing capacity of multimodal systems to process heterogeneous information, although their primary objectives do not necessarily address detailed explanations of how individual modalities contribute to a particular decision.

Xu, Zhu, and Clifton provided a broad survey of multimodal learning with Transformers, examining multimodal pretraining, representation learning, fusion, applications, and challenges [9]. Their work indicates that multimodal Transformers have progressed from relatively task-specific architectures toward general-purpose representation systems. However, the survey also highlights unresolved challenges related to multimodal alignment, missing modalities, computational requirements, and effective fusion.

Explainability introduces another dimension to multimodal learning. Park et al. investigated multimodal explanations that combine visual evidence with textual rationales to justify decisions [10]. Their work is important because it treats evidence identification as part of the explanation process rather than merely presenting a numerical confidence score. Lundberg and Lee developed SHAP as a general framework for attributing predictions to input features [11]. SHAP has subsequently influenced numerous explainability approaches, although direct application to complex multimodal Transformer architectures can be challenging because representations may be highly distributed and interactions may occur across modalities.

Sundararajan, Taly, and Yan introduced Integrated Gradients as an axiomatic attribution method for deep networks [13]. Selvaraju et al. proposed Grad-CAM for generating visual explanations from deep neural networks by identifying image regions associated with model predictions [14]. Ribeiro, Singh, and Guestrin introduced LIME, which explains predictions through locally interpretable approximations [16]. These approaches provide useful principles for feature attribution but were not originally designed specifically for context-aware multimodal Transformer architectures.

Gilpin et al. reviewed interpretability and emphasized the distinction between different forms of model explanation and transparency [15]. Parcalabescu and Frank developed MM-SHAP to address modality contributions in vision-language models, making their

work especially relevant to multimodal systems in which explanations must distinguish between information sources [17]. Fantozzi and Naldi examined Transformer explainability and categorized approaches according to the Transformer components used to produce explanations [19]. Sun et al. reviewed multimodal explainable artificial intelligence and identified challenges associated with multimodal contribution analysis, evaluation, and reliable explanation generation [20].

The literature therefore demonstrates substantial progress in Transformer-based multimodal learning and explainability, but the two research directions are not always integrated into a single context-aware decision framework. Existing multimodal architectures have demonstrated effective cross-modal representation learning, while conventional XAI methods have provided feature-attribution mechanisms. However, smart applications frequently require an additional capability: determining how the relevance of each modality changes according to context and communicating that variation to human users. The proposed research addresses this gap by integrating modality-specific representation, cross-modal Transformer reasoning, context-guided fusion, and hierarchical evidence attribution into one architecture.

3. Multimodal Decision Intelligence and Context Representation

The proposed framework considers a smart application as a multimodal decision environment in which several information sources provide complementary and sometimes conflicting evidence. The input layer can contain images, textual information, speech or other audio signals, sensor observations, and contextual variables. These inputs are not assumed to have identical reliability. Instead, each modality is represented independently before being integrated through the proposed multimodal reasoning mechanism.

Visual information can be obtained from cameras, medical images, surveillance systems, industrial inspection equipment, or mobile devices. A visual encoder transforms an input image or video representation into a sequence of visual embeddings. These embeddings contain spatial or semantic information that can subsequently interact with representations from other modalities. Textual information may originate from user queries, reports,

descriptions, system logs, or natural-language instructions. A Transformer-based language encoder converts the textual input into contextual token representations.

Audio information can include spoken instructions, environmental sounds, machine sounds, or other acoustic observations. An audio encoder transforms the waveform or spectral representation into an embedding sequence. Sensor information provides another source of evidence. Temperature, pressure, motion, heart rate, location, acceleration, environmental measurements, and other sensor variables can be normalized and encoded into contextual representations.

The context representation acts as an additional information layer rather than simply another independent modality. It describes the state under which the decision is being generated. Context can include time, application state, environmental conditions, task requirements, device status, user interaction state, and modality availability. The context representation allows the model to distinguish between identical observations occurring under different conditions.

For example, a visual observation of an object may have different implications during daytime and nighttime operation. Similarly, a spoken command received in a quiet environment may be more reliable than the same command received under severe background noise. The context encoder provides information that can help the model estimate these differences.

The encoded modalities are projected into a common dimensional space so that their representations can participate in subsequent Transformer operations. Positional or modality embeddings can be incorporated to preserve information about the origin and order of representations. A modality identifier enables the Transformer to distinguish visual, textual, audio, sensor, and contextual tokens.

The representation layer therefore establishes a common computational language across heterogeneous inputs while retaining modality identity. This is important because multimodal fusion should not eliminate the distinctions between information sources. Instead, the system should preserve sufficient information to determine which modality provided particular evidence for the final decision.

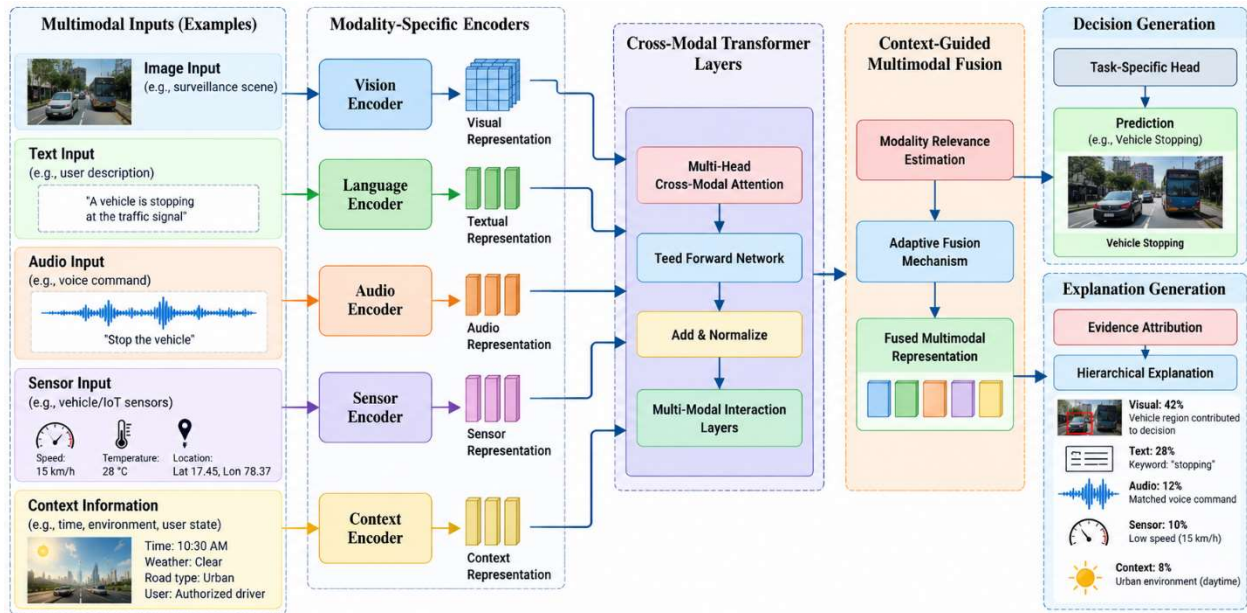


Figure 1. Proposed Explainable Multimodal Transformer Architecture

4. Explainable Multimodal Transformer and Context-Guided Fusion

The central component of the proposed framework is a cross-modal Transformer that learns relationships between representations generated by the individual modality encoders. Instead of concatenating all representations directly, the Transformer uses attention mechanisms to determine which elements should interact during decision formation. This enables visual features to interact with textual descriptions, speech information to interact with contextual observations, and sensor information to influence the interpretation of other modalities.

Let the representations of the visual, textual, audio, sensor, and contextual modalities be represented by H_v , H_t , H_a , H_s , and H_c , respectively. Each representation is projected into a common feature space. The combined representation can be expressed conceptually as

$$H = [H_v; H_t; H_a; H_s; H_c].$$

The Transformer then computes attention relationships among these representations. For a standard attention operation, the representation can be expressed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where Q , K , and V represent query, key, and value projections and d denotes the dimensionality used for scaling.

In the proposed architecture, cross-modal attention is used to identify relationships that are not necessarily

available within individual modalities. For example, a textual phrase describing an object can interact with image regions representing that object. A sensor measurement can alter the interpretation of a visual observation, while an audio signal can provide additional evidence for a textual command.

Context-guided fusion is introduced after or within the cross-modal representation process. Let z_i denote the representation associated with modality i , and let c denote the contextual representation. A modality relevance coefficient can be represented conceptually as

$$a_i = \text{softmax}(f(z_i, c)),$$

where f represents a learned compatibility function. The fused representation can then be expressed as

$$zf = \sum_{i=1}^M (a_i z_i)$$

where M represents the number of available modalities.

The coefficients a_i provide an interpretable representation of relative modality contribution, although they should not automatically be treated as complete causal explanations. Their reliability should be evaluated against explanation fidelity and other attribution measures. Additional attribution mechanisms such as Integrated Gradients, SHAP, or

Grad-CAM can be applied to obtain feature-level evidence [11], [13], [14].

The explanation mechanism operates at two levels. At the first level, the system identifies which modalities contributed most strongly to the decision. At the second level, it identifies specific features or regions within those modalities. For an image, this can correspond to important image regions. For text, it can correspond to influential words or phrases. For audio, it can correspond to important temporal segments or

acoustic patterns. For sensor data, it can identify influential measurements or temporal intervals.

The explanation is then aligned with the generated decision. If the system predicts a particular application state, the explanation should identify the evidence that supports that state rather than producing a generic description of the input. This distinction is essential because an explanation that is visually plausible but unrelated to the actual decision process may provide a misleading interpretation.

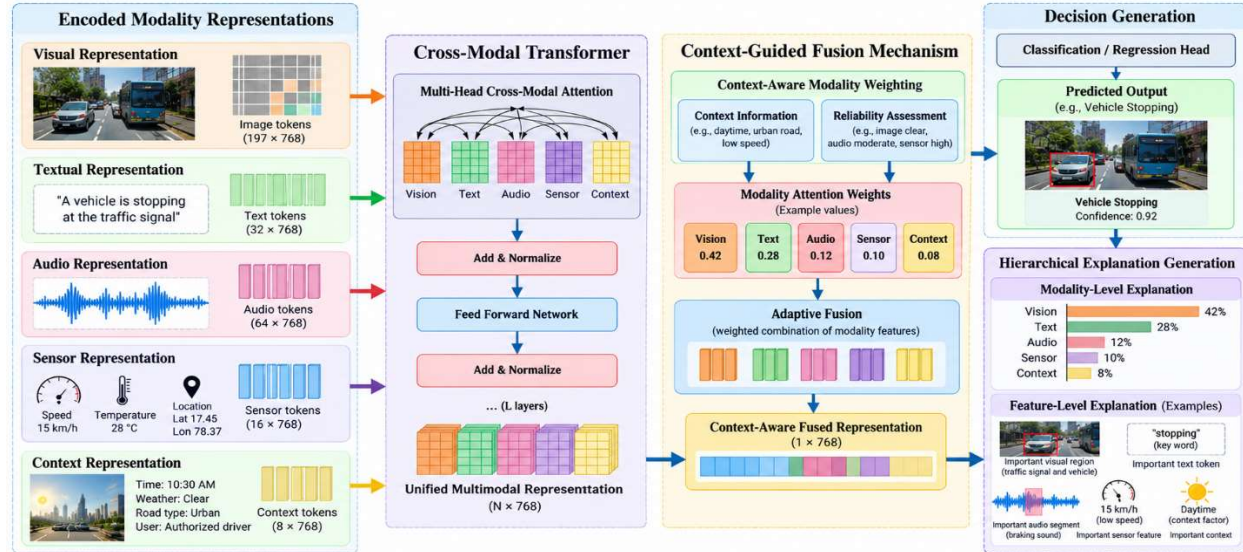


Figure 2. Context-Guided Multimodal Fusion and Explanation Generation

5. Experimental Configuration, Evaluation and Smart-Application Scenarios

The experimental evaluation should determine whether the proposed architecture can maintain predictive performance while providing meaningful explanations. Since the framework is intended for general smart applications, the evaluation can be organized around multiple representative scenarios rather than relying exclusively on a single application domain. Smart healthcare can involve medical images, clinical text, physiological sensor information, and patient context. Intelligent transportation can combine camera images, traffic descriptions, audio signals, vehicle sensors, and environmental conditions. Smart-home applications can integrate camera observations, speech commands, environmental sensors, and user context.

A multimodal experimental dataset should contain aligned or approximately synchronized observations from the participating modalities. Each sample should contain a target decision and sufficient contextual information to evaluate whether the model adapts appropriately to changes in operating conditions.

Where a public multimodal dataset is used, the training, validation, and testing partitions should be kept separate to prevent information leakage.

The proposed model should be compared with appropriate unimodal and multimodal baselines. A visual-only model can establish the contribution of visual information, while a text-only model can measure the performance of language information. A simple concatenation model can be used to determine whether the proposed attention-based fusion provides benefits beyond direct feature concatenation. A conventional multimodal Transformer can serve as a stronger baseline for evaluating the contribution of context-guided fusion.

Predictive performance should be measured using accuracy, precision, recall, and F1-score for classification tasks. If the application involves continuous outputs, suitable regression metrics can be substituted. Explainability should be evaluated separately. Explanation fidelity should measure whether the identified evidence corresponds to the model's actual behavior. Modality contribution consistency should examine whether the explanation

remains stable under small perturbations that should not substantially alter the decision. Robustness should be examined by systematically removing or degrading individual modalities. For example, visual quality can be reduced through image corruption, audio can be contaminated with background noise, textual information can be partially masked, and sensor measurements can contain missing values. The model should then be evaluated to

determine whether it can redistribute its reliance toward the remaining informative modalities. Context sensitivity should be evaluated by modifying contextual variables while maintaining the underlying content as much as possible. The objective is to determine whether the model appropriately changes its decision or modality weighting when context genuinely changes the interpretation of the input.

Table 1. Predictive and Explainability Evaluation Framework

Evaluation aspect	Metric	Purpose
Predictive performance	Accuracy, Precision, Recall, F1-score	Measures decision quality
Explanation fidelity	Fidelity score	Determines whether explanations reflect model behavior
Modality contribution	Contribution consistency	Measures stability of modality attribution
Context sensitivity	Context-response variation	Examines adaptation to changing context
Missing-modality robustness	Performance degradation	Measures resilience to unavailable inputs
Computational efficiency	Inference time and parameter count	Evaluates practical deployment requirements

The table should contain actual measured values after experimentation rather than predetermined values.



Figure 3: Multimodal Contribution and Context Sensitivity Analysis

6. Robustness, Ablation, Discussion and Practical Implications

Ablation analysis is necessary to determine whether the individual components of the proposed architecture contribute meaningfully to its performance. The complete architecture can be compared with versions in which context integration, cross-modal attention, adaptive modality weighting, or hierarchical explanation is removed. Such experiments provide evidence concerning the

functional contribution of the architecture rather than relying solely on overall performance.

The context integration component can be evaluated by comparing the complete model with a model that treats all modalities using fixed fusion weights. A difference between the two configurations would provide information about the value of context-dependent modality weighting. Similarly, removing cross-modal attention would indicate whether direct multimodal combination is sufficient or whether

learned relationships between modalities provide additional value.

The explanation mechanism can also be evaluated independently. Removing the explanation module should not necessarily produce a large change in prediction accuracy because its principal purpose is interpretability. Instead, the relevant comparison should concern explanation fidelity, evidence localization, modality attribution consistency, and computational overhead.

Robustness analysis should investigate situations in which one or more modalities are unavailable. Real-world smart applications cannot assume continuous availability of every sensor. Cameras may be blocked,

microphones may experience interference, sensors may fail, and textual descriptions may be incomplete. A practical multimodal architecture should therefore degrade gracefully rather than producing unstable decisions whenever one information source becomes unavailable.

Conflicting modalities represent another important robustness condition. A camera may suggest one interpretation while a textual description suggests another. Rather than forcing an arbitrary decision, the architecture should identify the disagreement and use contextual information to determine which evidence is more appropriate. Such behavior is particularly important in safety-sensitive smart applications.

Table 2. Ablation and Robustness Evaluation

Model configuration	Predictive performance	Explanation fidelity	Missing-modality robustness	Context sensitivity	Computational cost
Without context integration	Actual result	Actual result	Actual result	Actual result	Actual result
Without cross-modal attention	Actual result	Actual result	Actual result	Actual result	Actual result
Without adaptive modality weighting	Actual result	Actual result	Actual result	Actual result	Actual result
Without hierarchical explanation	Actual result	Actual result	Actual result	Actual result	Actual result
Proposed complete architecture	Actual result	Actual result	Actual result	Actual result	Actual result

The practical value of explainability is particularly relevant in environments where human users must review intelligent decisions. In smart healthcare, clinicians may require information concerning which image regions, clinical descriptions, physiological measurements, or contextual factors influenced a recommendation. In industrial monitoring, operators may need to understand whether a warning originated primarily from visual inspection, machine telemetry, acoustic evidence, or operational context. In intelligent transportation, an explanation can help indicate whether a decision was influenced by a detected road object, a traffic instruction, vehicle sensor data, or environmental conditions.

However, explainability should not be treated as a guarantee that a model is correct. An explanation describes aspects of the model's decision process or identifies features associated with the output. It does not by itself establish causal correctness. Therefore, explanation quality should be evaluated independently from predictive performance.

The computational requirements of multimodal Transformers are another practical consideration. Processing several modalities simultaneously can

increase memory consumption and inference time. Efficient projection layers, modality-specific compression, lightweight attention mechanisms, model distillation, and edge-aware inference strategies can potentially reduce deployment costs.

Privacy is also important because multimodal systems can process highly sensitive information. Images, speech, location data, sensor observations, and textual records can collectively provide substantially more information about individuals than any single modality. Practical deployments should therefore incorporate appropriate data protection, access control, encryption, and privacy-preserving learning mechanisms.

7. Limitations, Future Research and Conclusion

The proposed framework has several limitations. First, multimodal Transformer architectures can require considerable computational resources, particularly when several high-dimensional modalities are processed simultaneously. This can make direct deployment on resource-constrained edge devices difficult. Model compression and efficient inference

are therefore important areas for subsequent investigation.

Second, the quality of context representation strongly influences the behavior of the proposed fusion mechanism. If the context information is incomplete, outdated, or incorrectly measured, the model may assign inappropriate importance to one or more modalities. Reliable context acquisition is therefore a fundamental requirement for effective context-aware decision intelligence.

Third, modality contribution does not necessarily represent a causal explanation. Attention weights or attribution values can identify information associated with a prediction without proving that the identified information caused the decision. For this reason, multiple explanation methods should be compared when evaluating the interpretability of the architecture.

Fourth, explanation evaluation itself remains challenging. A visually intuitive explanation may not necessarily be faithful to the underlying model. Human evaluations can also be subjective. A robust evaluation methodology should therefore combine quantitative fidelity measures with controlled perturbation experiments and, where appropriate, expert assessment.

Future research can investigate lightweight multimodal Transformers designed specifically for edge deployment. Knowledge distillation and parameter-efficient adaptation could reduce computational requirements while retaining cross-modal reasoning capabilities. Another promising direction is federated multimodal learning, where multiple organizations or edge devices contribute to model training without directly exchanging sensitive raw data.

Continual learning could allow the model to adapt to new application contexts, emerging sensor configurations, and changing user behavior without complete retraining. Multimodal foundation models also provide an opportunity to extend the proposed architecture to a larger range of input types and application domains.

Further research should investigate explanation alignment under modality conflict. A model should not merely identify the strongest modality; it should determine whether multiple modalities provide mutually consistent evidence. Confidence estimation and uncertainty modeling could therefore be incorporated into future versions of the framework.

Human-in-the-loop decision intelligence is another important direction. Rather than treating the human user as a passive recipient of an explanation, future systems could allow users to inspect, challenge, or

modify the evidence used by the model. This could lead to interactive intelligent systems in which human feedback becomes part of the decision process.

In conclusion, this research proposes an explainable multimodal Transformer architecture for context-aware decision intelligence in smart applications. The architecture combines modality-specific representation learning, cross-modal Transformer reasoning, context-guided fusion, adaptive modality contribution, decision generation, and hierarchical evidence attribution. The central principle is that an intelligent system should not only combine heterogeneous information to produce a decision but should also provide evidence that enables a human user to understand the basis of that decision.

The proposed framework is applicable to smart healthcare, intelligent transportation, industrial monitoring, smart homes, human-computer interaction, and other environments where multimodal information and contextual reasoning are required. Its evaluation should consider both conventional predictive measures and explainability-oriented measures, including explanation fidelity, modality contribution consistency, context sensitivity, robustness to missing information, and computational efficiency. The resulting framework provides a foundation for developing multimodal intelligent systems in which perception, contextual reasoning, decision generation, and explanation are treated as interconnected components rather than independent processing stages.

References

- [1]. A. Vaswani *et al.*, "Attention Is All You Need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [2]. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *Proceedings of NAACL-HLT*, pp. 4171–4186, 2019.
- [3]. J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks," *Advances in Neural Information Processing Systems*, vol. 32, 2019. ViLBERT uses separate visual and textual streams connected through co-attentional Transformer layers.
- [4]. H. Tan and M. Bansal, "LXMERT: Learning Cross-Modality Encoder Representations from Transformers," *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th*

- International Joint Conference on Natural Language Processing*, pp. 5100–5111, 2019. LXMERT explicitly models relationships within and across vision and language modalities.
- [5]. L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, “VisualBERT: A Simple and Performant Baseline for Vision and Language,” *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1001–1015, 2020. VisualBERT employs Transformer self-attention to align textual elements with visual regions.
- [6]. Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, “UNITER: UNiversal Image-TExt Representation Learning,” *European Conference on Computer Vision*, pp. 104–120, 2020. UNITER develops joint image-text representations through large-scale multimodal pretraining and explicit word-region alignment.
- [7]. A. Radford *et al.*, “Learning Transferable Visual Models From Natural Language Supervision,” *Proceedings of the 38th International Conference on Machine Learning*, vol. 139, pp. 8748–8763, 2021. CLIP demonstrated large-scale image-text representation learning and zero-shot transfer across visual tasks.
- [8]. J.-B. Alayrac *et al.*, “Flamingo: A Visual Language Model for Few-Shot Learning,” *Advances in Neural Information Processing Systems*, vol. 35, pp. 23716–23736, 2022.
- [9]. P. Xu, X. Zhu, and D. A. Clifton, “Multimodal Learning With Transformers: A Survey,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 10, pp. 12113–12132, 2023. This survey reviews multimodal Transformer architectures, multimodal pretraining, applications, challenges, and open research directions.
- [10]. D. H. Park, L. A. Hendricks, Z. Akata, A. Rohrbach, B. Schiele, T. Darrell, and M. Rohrbach, “Multimodal Explanations: Justifying Decisions and Pointing to the Evidence,” *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 8779–8788, 2018. The study combines visual evidence with textual rationales to explain multimodal decisions.
- [11]. S. Lundberg and S.-I. Lee, “A Unified Approach to Interpreting Model Predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017. The paper introduced SHAP as a unified framework for attributing model predictions to input features.
- [12]. M. B. K. V. K. Bharadwaj Parasaram, S. K. Nalluri and V. T. Bathini, “Attention-Guided Multi-Modal Deep Intelligence Framework for Early Cancer Detection Using Imaging and Genomic Signatures,” *2026 International Conference on Data Science for Cyber-Physical Systems Resilience using Advanced Applications (ICDCA)*, BANGALORE, India, 2026, pp. 1–6, doi: 10.1109/ICDCA69396.2026.11620255.
- [13]. M. Sundararajan, A. Taly, and Q. Yan, “Axiomatic Attribution for Deep Networks,” *Proceedings of the 34th International Conference on Machine Learning*, pp. 3319–3328, 2017. The work introduced Integrated Gradients for attribution of predictions to input features.
- [14]. R. R. Selvaraju *et al.*, “Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization,” *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.
- [15]. L. H. Gilpin *et al.*, “Explaining Explanations: An Overview of Interpretability of Machine Learning,” *2018 IEEE 5th International Conference on Data Science and Advanced Analytics*, pp. 80–89, 2018.
- [16]. M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1135–1144, 2016.
- [17]. L. Parcalabescu and A. Frank, “MM-SHAP: A Performance-Agnostic Metric for Measuring Multimodal Contributions in Vision and Language Models & Tasks,” *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pp. 4032–4059, 2023. MM-SHAP specifically addresses the contribution of different modalities in vision-

- language models, making it particularly relevant to the proposed architecture.
- [18]. Koeneni, C., S. Lokula, and L. Panjiyar. "A Cloud-Native Approach to Statistical Anomaly Detection and Automated Data Quality Validation." *International Journal of Data Science and IoT Management System* 4.4 (2025): 533-543
- [19]. P. Fantozzi and M. Naldi, "The Explainability of Transformers: Current Status and Directions," *Computers*, vol. 13, no. 4, Art. 92, 2024. The survey categorizes Transformer explainability methods according to the Transformer components used to produce explanations.
- [20]. S. Sun, W. An, F. Tian, F. Nan, Q. Liu, J. Liu, N. Shah, and P. Chen, "A Review of Multimodal Explainable Artificial Intelligence: Past, Present and Future," 2024. The review examines multimodal XAI approaches, evaluation methods, datasets, and emerging challenges across different generations of multimodal models.
- [21]. V. K. Bharadwaj Parasaram, V. Teja Bathini and S. K. Nalluri, "An Intelligent Digital Framework for Clinical Trial Monitoring Using AI and Cloud-Based Software Platforms," *2026 2nd International Conference on Artificial Intelligence Systems (AIS)*, San Antonio, TX, USA, 2026, pp. 524-530, doi: 10.1109/AIS69919.2026.11621841.
- [22]. Vishram Singh, 2026. "Enterprise Resource Planning for Financial Applications: Engineering, Architecture, and Design Principles," *International Journal of Scientific Research in Science and Technology*, Technoscience Academy, vol. 13(2), pages 653-661, April.
- [23]. C. Koeneni and S. Lokula, "Silent Data Corruption in Distributed Genai-Augmented Data Pipelines: an Empirical Study of Hallucination-Induced Failure Propagation," *2026 IEEE 46th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, Seoul, Korea, Republic of, 2026, pp. 137-143,