

ORIGINAL RESEARCH

Open access

Following the Evidence Across Modalities: A Conceptual Architecture for Explainability in Text, Vision, Audio, Sensor, and Structured-Data Intelligent Systems

Amelia Clarke^{1*}, Carlos Mendoza²

Abstract

Multimodal intelligent systems combine evidence from text, vision, audio, sensors, and structured data, but their explanations often remain fragmented into separate saliency maps, feature scores, textual rationales, attention visualisations, or confidence indicators. Such outputs may describe isolated aspects of a model while obscuring how modality-specific evidence was represented, transformed, fused, and communicated. This article develops a conceptual multimodal explainability architecture organised around the principle of following evidence across modalities. The approach begins by distinguishing the evidential and representational characteristics of textual, visual, acoustic, temporal-sensor, and structured inputs. It then identifies attribution difficulties created by early, intermediate, late, and hybrid fusion, including modality dominance, missing evidence, conflicting inputs, representation alignment, and uncertainty about joint contributions. Evidence traceability is treated as an end-to-end requirement connecting input provenance, modality-specific representation, fusion operations, attribution processes, consistency assessment, uncertainty communication, and decision explanation. The architecture further incorporates human interpretation because technically detailed multimodal explanations may remain cognitively inaccessible, visually persuasive, incomplete, or poorly calibrated to user expertise and decision context. Its central contribution is a layered framework linking modality inventory, evidence classification, representation mapping, fusion attribution, provenance, consistency review, human-centred presentation, evaluation, and feedback updating. The architecture is intended to structure research and design reasoning rather than prescribe a particular model, explanation technique, evaluation metric, or deployment procedure. It does not establish that an explanation is faithful, that users understand it, that reliance is appropriate, or that the underlying intelligent system is trustworthy or ready for operational use.

Keywords Explainable AI, Multimodal AI, Multimodal explainability, Cross-modal attribution, Evidence traceability, Explanation consistency

*Correspondence:

Amelia Clarke
amelia.clarke@unimelb.edu.au

¹ Department of Multimodal AI and Explainability, University of Melbourne, Melbourne, Australia

² Department of Systems Engineering, Universidad de los Andes, Bogota, Colombia

Introduction

Multimodal intelligence extends machine learning beyond a single informational channel by combining linguistic, visual, acoustic, temporal, sensor-generated, and structured evidence. The resulting systems may represent, align, translate, and fuse heterogeneous inputs before producing a classification, recommendation, generated response, or decision-

support output. Explainability in this setting cannot be reduced to displaying the most salient image region or the most influential textual token because the operative evidence may be distributed across several modalities and transformed through multiple representational stages. Moreover, an explanation is not only a technical account of model behaviour; it is also a selective communication received and interpreted by a person within a particular task and social context [1-3].

The design problem is therefore broader than selecting an attribution method. Explanation objectives differ across developers, domain specialists, operators, decision makers, and people affected by system outputs. One stakeholder may need to diagnose model behaviour, another may need to determine whether material evidence is missing, and another may need an accessible account of why a recommendation was produced. A multidisciplinary design approach consequently requires the explanation purpose, recipient, interaction context, and evaluation criteria to be specified before explanatory outputs are selected or combined [4].

Multimodal systems intensify familiar explainability problems while introducing additional ones. Evidence from different channels may be redundant, complementary, contradictory, differently scaled, or unequally reliable. A system may appear multimodal while relying predominantly on one source, or it may compensate silently for an absent modality through imputation or correlations learned from other inputs. Even when separate explanations are available for text, images, audio, sensor streams, and tabular variables, those explanations may not reveal how cross-modal interaction altered the final output. A collection of modality-specific explanations is therefore not necessarily an explanation of multimodal reasoning.

This article proposes a conceptual architecture for following evidence across multimodal intelligent systems. Its purpose is to organise explainability around nine connected concerns: use-case definition, stakeholder needs, modality inventory, modality-specific representation, provenance and data quality, cross-modal fusion, fusion attribution, evidence traceability and consistency, and human interpretation and evaluation. The proposed contribution is conceptual rather than operational. It structures questions that explanation designers and researchers may investigate, but it does not validate a particular explanation method, establish accountability, demonstrate human understanding, certify trustworthy behaviour, or determine deployment readiness.

Multimodal Intelligence and the Explanation Problem

Multimodal intelligent systems commonly use distinct processing pathways to encode heterogeneous inputs before combining them through shared representations or decision mechanisms. Depending on the architecture, modalities may be fused near the input layer, after partial representation learning, at a high-level decision stage, or through several connected stages. These design choices affect what evidence remains identifiable and what forms of attribution are technically plausible. An explanation architecture must therefore document where each modality enters the system, how it is represented, and when it interacts with other evidence [5].

The explanation problem also depends on the quality and availability of each source. Multimodal biomedical systems, for example, may integrate imaging, text, molecular measurements, and structured clinical information, but these sources can differ in completeness, resolution, timing, and reliability. Similar challenges arise in other decision-support and socio-technical settings where sensor failure, document incompleteness, image artifacts, temporal misalignment, or inconsistent structured records can affect model behaviour. Adding modalities can broaden the evidential base, but it simultaneously enlarges the provenance, missingness, alignment, bias, and validation burden [6].

A further difficulty concerns the relationship between an explanation and the model process it purports to describe. Post-hoc explanations may be intuitively convincing while remaining detached from the operative decision mechanism. In consequential contexts, readable rationales or visually clear importance maps should not be assumed to reveal the model's actual reasoning merely because they are plausible to a human observer. The multimodal setting amplifies this concern because an explanation may represent one modality accurately while ignoring decisive interactions occurring elsewhere in the system [7].

Technical explanation methods can nevertheless support structured investigation when their scope is stated carefully. Attribution, relevance decomposition, sensitivity analysis, saliency, example-based explanation, and representation analysis can expose selected relationships between inputs, internal states, and outputs. Their results remain dependent on the model, explanation algorithm, parameterisation, and evaluation procedure. Accordingly, multimodal explainability requires validation of both modality-specific outputs and their relationship to fusion behaviour; no individual attribution artefact should be treated as a complete account of the system [8].

Modality-Specific Evidence and Representation

Textual evidence may enter a multimodal system as short utterances, long documents, retrieved passages, labels, metadata, or conversational history. Its representation can involve tokens, sentences, contextual embeddings, document-level features, or generated summaries. Token attribution identifies local associations between an output and selected textual elements, whereas a natural-language rationale presents a human-readable account that may be extracted from source evidence or generated after the decision. These explanation forms answer different questions, and a fluent rationale should not be treated as faithful merely because it is coherent or persuasive [9].

Visual evidence includes images, video frames, objects, regions, spatial relationships, motion, and learned visual features. Saliency and relevance maps may indicate where a model is locally sensitive, while object- or region-based explanations may connect an output to more recognisable visual entities. However, highlighted regions do not independently establish that the represented object was necessary, sufficient, or causally responsible for the output. Visual explanation therefore requires tracing the relationship between source imagery, preprocessing, latent visual representations, model layers, and the final decision [10].

Audio evidence is temporally structured and may include speech content, speaker characteristics, environmental sounds, rhythm, frequency patterns, or learned acoustic representations. Explanations may be presented as time-aligned segments, spectrogram regions, acoustic-feature importance, or links between a transcript and its corresponding sound. Methods transferred from image explainability can be useful, but their interpretation must reflect the distinct semantics of waveforms, temporal dependencies, and time-frequency transformations. Audio explanations consequently require domain-sensitive validation rather than the uncritical reuse of visual explanation assumptions [11].

Sensor and structured-data modalities introduce further representational differences. Sensor streams and time-series signals preserve ordering, duration, sampling frequency, missing intervals, and temporal patterns, whereas tabular data rely on schemas, numerical variables, categorical features, relational fields, units, and derived measures. Static feature rankings may erase important temporal structure, while structured-data attributions may overlook correlations, encodings, interactions, or proxy variables. Explanation design must therefore retain the semantic and temporal properties that make these modalities meaningful in their original context [12].

Across all modalities, provenance is essential. An explanation should distinguish primary evidence from transformed, retrieved, inferred, or imputed information and should identify the processing stage at which each representation was created. Modality-specific analysis remains necessary because text, images, audio, sensor streams, and structured variables support different forms of evidence and different failure modes. It is not sufficient, however, because the final output may depend on interaction and fusion processes that cannot be reconstructed by inspecting each modality independently.

Table 1 summarises modality-specific evidence and representation requirements for multimodal explainability across text, vision, audio, sensor, and structured-data intelligent systems.

Table 1. Modality-Specific Evidence and Representation Requirements for Multimodal Explainability: Text Evidence, Natural-Language Rationales, Token Attribution, Vision Evidence, Visual Attribution, Audio Evidence, Acoustic Features,

Sensor Streams, Time-Series Signals, Structured Data, Feature Importance, Data Provenance, and Modality-Specific Boundaries

Modality or evidence type	Core explanation question	Evidence or representation feature	Potential explanation form	Failure or overclaiming risk	Validation or traceability need	Human interpretation need	Architectural role
Text, natural language, and document evidence	Which textual source, passage, or linguistic element contributed to the output?	Tokens, sentences, documents, metadata, contextual embeddings, and retrieved passages	Passage citation, token highlighting, document provenance, or textual summary	A fluent explanation may not reflect the evidence actually used	Trace the source document, selected passage, preprocessing, embedding, and downstream use	Separate source evidence from generated interpretation	Text-evidence and provenance pathways
Text rationales	Is the rationale extracted from operative evidence or generated as a post-hoc account?	Extractive or generated explanatory language	Contrastive, selective, or question-responsive rationale	Coherence may be mistaken for fidelity or completeness	Compare the rationale with source evidence, attribution, and model behaviour	State uncertainty, omissions, and generation status	Human-facing language channels
Token attribution	Which tokens receive attributed relevance for a particular output?	Token-level importance, gradients, perturbations, or relevance estimates	Highlighted tokens or ranked spans	Scores may be unstable and may omit contextual interactions	Test sensitivity to tokenisation, context, perturbation, and method choice	Avoid presenting local relevance as causal proof	Local token attribution components
Vision, image, and video evidence	Which visual source, frame, object, or region contributed?	Pixels, patches, frames, motion, objects, regions, and visual embeddings	Saliency map, region attribution, key frames, or example-based comparison	Visually salient areas may appear more evidentially decisive than they are	Trace image or video provenance, preprocessing, representation, and temporal aggregation	Preserve enough visual context to prevent misleading interpretation	Visual evidence pathways
Object, region, and saliency evidence	Does the explanation identify sensitivity, semantic content, or	Object detections, region features, gradients, or relevance maps	Object-level account, bounding region, or heatmap	Detection errors, artifacts, and unstable maps may be hidden	Apply stability, sanity, perturbation, and representation-level checks	Explain what the visualisation can and cannot establish	Diagnostic visual attribution components

	causal contribution?						
Audio, speech, and sound evidence	Which utterance, acoustic event, or temporal segment influenced the output?	Waveforms, spectrograms, speech content, prosody, sound events, and audio embeddings	Time-aligned attribution, acoustic-feature account, or playable segment	Visualised spectra may be misread as direct semantic evidence	Trace waveform transformation, temporal alignment, and acoustic representation	Use domain-meaningful descriptors and accessible playback where appropriate	Audio evidence pathways
Sensor streams, time-series signals, and IoT evidence	Which device, period, trend, or subsequence contributed?	Timestamps, sampling intervals, temporal patterns, device metadata, and signal quality	Temporal attribution, event timeline, subsequence explanation, or quality indicator	Drift, gaps, noise, and aggregation may be concealed by static scores	Preserve device provenance, calibration status, missingness, and temporal context	Explain scale, duration, reliability, and missing intervals	Sensor temporal evidence pathways
Structured and tabular data	Which fields, values, categories, or interactions influenced the result?	Numerical, categorical, relational, derived, and encoded features	Local or global feature attribution, rule, or counterfactual account	Feature rankings may ignore dependencies, proxies, or implausible alternatives	Trace schema, units, encoding, scaling, derivation, and feature relationships	Use original domain labels and distinguish local from global scope	Structured data pathways
Data and modality provenance	Where did each evidential element originate and how was it transformed?	Source identity, acquisition context, lineage, quality, timestamps, and preprocessing	Provenance record, evidence ledger, or transformation chain	Derived or imputed evidence may appear primary or verified	Maintain source-to-representation lineage and version information	Communicate source quality and transformation limits	Cross-cutting evidence traceability inputs
No modality-specific sufficiency boundary	Can one modality-specific explanation account for the complete multimodal decision?	Comparison of modality-level, interaction-level, and fused evidence	Explicit scope and limitation statement	A clear explanation for one modality may be mistaken for a system-level explanation	Require fusion, interaction, missingness, and conflict review	Identify what remains unexplained across modalities	Mandatory architectural boundaries

Cross-Modal Fusion and Attribution Challenges

Cross-modal fusion changes the object that an explanation must address. Evidence may be integrated at the input level, through intermediate latent representations, through separate modality-specific decisions, or through hybrid combinations spanning several stages. In each case, joint behaviour may arise from complementarity, redundancy, suppression, or

conflict among modalities. Research on multimodal data integration and explainable information fusion indicates that the contribution of an individual source cannot always be reconstructed by listing independent feature scores, because the operative signal may emerge only through interaction and context-sensitive combination [13-15].

Existing multimodal explainability approaches vary according to the modalities involved, the stage at which explanation is produced, and whether the method attributes the output to inputs, representations, interactions, or generated explanatory content. This diversity is useful but creates comparability problems. An image heatmap paired with a textual rationale, for example, does not necessarily explain how visual and linguistic representations affected each other. A structured architecture must therefore specify whether an explanation concerns modality-specific evidence, the fusion mechanism, a cross-modal interaction, or the final communication presented to a user [16].

Early fusion can make modality-level attribution difficult because low-level features are combined before each modality develops a fully separable representation. Intermediate fusion shifts the challenge to latent states whose semantics may not correspond directly to human-recognisable evidence. Late fusion can appear more transparent because separate predictions or representation summaries are combined near the output, yet fusion weights or confidence values may still conceal calibration differences, dependency, and conditional interaction. Hybrid architectures compound these issues because evidence may be combined repeatedly at several locations.

Cross-attention and representation alignment provide additional points for analysis, especially in transformer-based and large multimodal systems, but neither should be treated as a complete explanation. Attention may identify which internal elements interact, while alignment measures may indicate proximity between learned representations. These operations do not independently establish that attended or aligned elements caused the output. Likewise, separate saliency maps for several modalities may remain incomplete because they do not show how fusion transformed, reinforced, or suppressed the attributed evidence.

Attribution must also address modality dominance, imbalance, noise, absence, and conflict. A nominally multimodal model may rely primarily on one channel, substitute correlated information when another is missing, or suppress evidence that contradicts a dominant representation. Different explanation methods may assign incompatible contributions, and such disagreement should be treated as information about explanatory uncertainty rather than removed through selective presentation. The conceptual architecture therefore requires explicit fusion mapping, modality-interaction review, missing-modality disclosure, conflict analysis, and qualified attribution. These processes may support better investigation, but they do not prove that attention is explanatory, that saliency is sufficient, that contribution estimates are uniquely correct, or that the system is ready for operational use.

Evidence Traceability and Explanation Consistency

Evidence traceability concerns whether an explanatory claim can be followed from its communicated form back through the relevant model operations to identifiable source evidence. In a multimodal system, this requires more than preserving input records. It requires documentation of modality provenance, preprocessing, representation formation, fusion location, attribution procedure, decision conditions, and explanation-generation processes. Explanation quality must therefore be assessed through specified properties such as fidelity, usefulness, comprehensibility, completeness, and consistency rather than through an undifferentiated claim that a system is explainable [17].

A multimodal evidence chain should connect each textual passage, image region, audio segment, sensor interval, or structured variable to its modality-specific representation and then identify how that representation entered the fusion process. The chain should further record relevant interaction analysis, attribution uncertainty, explanation-consistency checks, and the relationship between fused evidence and the communicated decision explanation. Evaluation should examine several properties because a traceable explanation can remain unstable, incomplete, unfaithful, or unusable [18]. **Figure 1** illustrates the multimodal evidence traceability pathway, showing how text, vision, audio, sensor, and

structured-data evidence can be connected to representation, fusion, attribution, explanation consistency, uncertainty, and decision explanation.

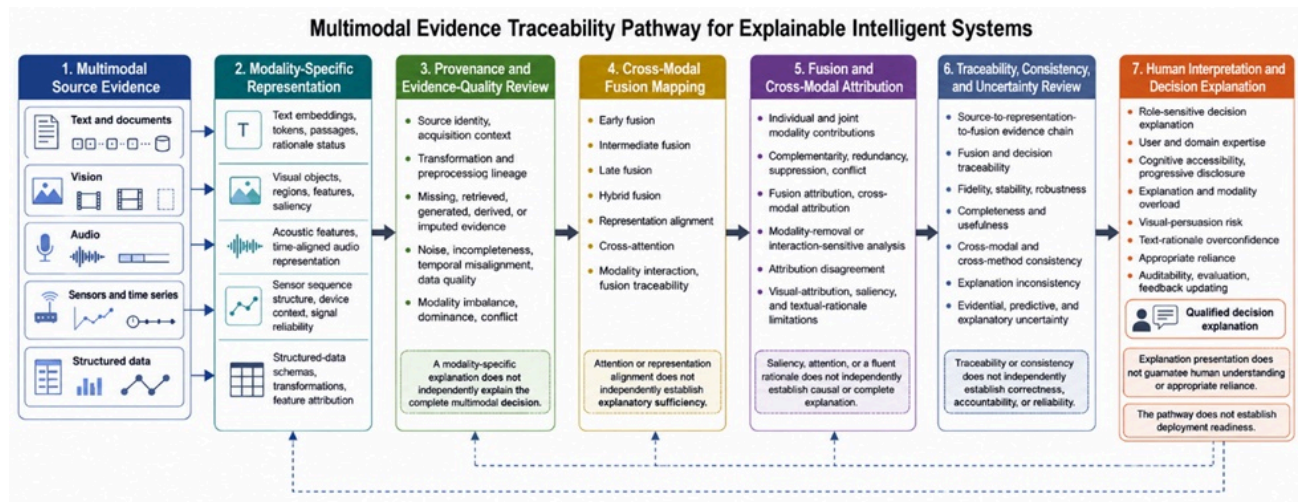


Figure 1. Multimodal Evidence Traceability Pathway for Explainable Intelligent Systems

Alt-text description:

A conceptual pathway showing text, vision, audio, sensor, and structured-data evidence moving through representation, provenance, fusion, attribution, consistency checking, uncertainty communication, and decision explanation.

Traceability is especially important when a model relies on unintended artifacts or shortcuts. An explanation can appear visually plausible while exposing evidence that is irrelevant to the intended task, and a confident output may rest on background features, acquisition artifacts, formatting patterns, or correlated proxies. Diagnostic explanation should therefore test whether the cited evidence reflects the intended phenomenon or merely a repeatable association learned from the data [19].

Uncertainty should also be attached to the evidence chain. Data-related uncertainty may arise from noisy images, ambiguous language, incomplete records, sensor variability, or conflicting modalities, whereas knowledge-related uncertainty may arise because the model has insufficient support for a particular case. These forms of uncertainty should not be collapsed automatically into one confidence value, and uncertainty in cross-modal attribution should be distinguished from uncertainty in the underlying prediction [20].

Explanation consistency concerns whether modality-level, fusion-level, and decision-level accounts can coexist without unresolved contradiction. Consistency may be examined across repeated inputs, small perturbations, explanation techniques, model versions, or different views of the same evidence. Nevertheless, agreement does not prove correctness: several explanations may consistently reproduce the same shortcut or bias. Traceability and consistency may support review, but neither independently establishes accountability, trustworthy operation, or justified reliance.

Human Interpretation of Multimodal Explanations

Human-centred evaluation must distinguish explanation availability from explanation success. Users may report that an explanation is clear or reassuring without demonstrating improved understanding, error detection, or task performance. Trust, perceived usefulness, acceptance, and actual comprehension are related but distinct outcomes, and understanding should be assessed in relation to the abilities needed for the relevant task rather than inferred from exposure to an explanatory display [21–23].

Static explanations are often insufficient for complex multimodal systems. A user may need to ask which modality dominated, inspect a conflicting source, compare an alternative explanation, reveal omitted evidence, or request a simpler account. Interactive explanation can support questioning and progressive disclosure, although interaction itself does not correct unfaithful attribution or guarantee that users will reach an appropriate interpretation [24].

Explanation presentation should reflect user expertise and decision context. A domain expert may require representation traces, source provenance, temporal detail, methodological uncertainty, and access to conflicting evidence. A lay user or affected person may instead require plain language, clear separation between source evidence and generated rationale, accessible uncertainty communication, and an opportunity to challenge inaccurate information. Operators and decision makers may need prioritised warnings about missing modalities, sensor quality, conflicting evidence, or unstable explanations.

Multimodal interfaces can create explanation overload. Simultaneous heatmaps, highlighted text, audio timelines, sensor plots, feature rankings, and narrative rationales may exceed the user’s capacity to integrate the information. Visual prominence may cause a heatmap to appear more authoritative than a qualified textual caveat, while a fluent narrative may conceal uncertainty or unsupported inference. Progressive disclosure, coordinated presentation, source-linked claims, and explicit scope statements can reduce these risks but require empirical evaluation.

Appropriate reliance should therefore be treated as an evaluation outcome rather than an assumed benefit of explanation. A useful architecture should help users recognise evidential strengths, gaps, conflicts, and uncertainty without implying that clearer explanations necessarily improve decisions. Human understanding and reliance require task-specific testing, and neither can be guaranteed by technical explainability alone. **Table 2** maps cross-modal fusion, attribution, evidence traceability, explanation consistency, and human-interpretation requirements for multimodal intelligent systems.

Table 2. Cross-Modal Fusion, Attribution, Evidence Traceability, Explanation Consistency, and Human Interpretation Requirements for Multimodal Explainability: Fusion Attribution, Modality Interaction, Modality Conflict, Missing Modality, Evidence Chains, Provenance, Consistency, Uncertainty, Cognitive Accessibility, Explanation Overload, and Interpretation Boundaries

Explainability requirement or challenge	Core explanation or interpretation question	Relevant modality or fusion condition	Explanation function	Potential output	Failure or misinterpretation risk	Evaluation or governance need
Fusion-stage identification	Where and how were modalities combined?	Early, intermediate, late, or hybrid fusion	Locates integration operations	Fusion map	Modality explanations may bypass decisive fusion logic	Architecture verification
Fusion and cross-modal attribution	Which individual and joint contributions affected the output?	Complementary, redundant, or interacting modalities	Examines conditional contributions	Qualified attribution account	Independent scores may omit interaction or imply false precision	Ablation, perturbation, and method comparison
Modality dominance and imbalance	Did one modality overwhelm or marginalise another?	Unequal quality, scale, availability, or informativeness	Tests meaningful multimodal use	Dominance or balance report	A nominally multimodal system may depend mainly on one source	Modality-removal and calibration testing

Missing modality	Which evidence was absent, imputed, or substituted?	Unavailable inputs or incomplete records	Discloses fallback behaviour	Availability and imputation statement	Users may believe all listed evidence was observed	Missing-modality stress testing
Modality noise and conflict	How were unreliable or contradictory sources handled?	Corrupted, drifting, or inconsistent inputs	Exposes disagreement and source quality	Conflict and quality report	The system may suppress contrary evidence silently	Conflict-case and robustness evaluation
Cross-attention and saliency boundaries	What do attention and saliency outputs establish?	Attention-based or gradient-based explanation	Describes internal association or sensitivity	Qualified visualisation	Visually compelling maps may be interpreted as causal proof	Sanity, intervention, and comparative testing
Evidence and modality provenance	Where did each evidential element originate?	All input modalities and derived representations	Preserves source identity and transformations	Provenance ledger	Derived, imputed, or stale data may appear primary	Lineage and version review
Representation and fusion traceability	How was evidence transformed before and during fusion?	Encoders, embeddings, latent states, and fusion modules	Links source evidence to internal processing	Evidence-chain graph	Hidden transformations may make explanations incomplete	Model-specific trace-completeness testing
Explanation fidelity, stability, and robustness	Does the explanation correspond to model behaviour and remain meaningful under variation?	All modalities and explanation techniques	Assesses technical dependability	Multi-property evaluation profile	Plausible explanations may be unfaithful or volatile	Perturbation, robustness, and method-specific evaluation
Cross-modal consistency	Do modality and fusion explanations agree for defensible reasons?	Multiple explanations of one output	Detects contradictions and unexplained divergence	Consistency matrix	Consistent explanations may reproduce a shared error	Consistency combined with fidelity testing
Explanation uncertainty and completeness	What is uncertain, omitted, or only partially represented?	Noisy, conflicting, or incompletely traceable evidence	Defines scope and confidence limits	Uncertainty and completeness statement	Point estimates may imply unsupported certainty	Communication and omission review
Cognitive accessibility	Can the intended user understand the	Lay, expert, operator, decision-maker,	Adapts detail, sequencing, and language	Layered or role-specific explanation	Complexity or oversimplification	Accessibility and

	explanation format and terminology?	or affected-person context			may impair understanding	comprehension testing
Explanation and modality overload	Are too many channels or details presented simultaneously?	Rich multimodal interfaces	Prioritises information and progressive disclosure	Coordinated explanation view	Users may ignore important caveats or conflicts	Usability and information-load evaluation
Visual-persuasion and rationale-overconfidence risks	Does presentation quality exceed evidential strength?	Heatmaps, diagrams, narratives, and confidence displays	Qualifies persuasive formats	Source-linked caveat and scope statement	Clear graphics or fluent language may induce overtrust	User-perception and fidelity evaluation
Appropriate reliance	Does use reflect demonstrated capability and limitations?	Decision-support and human-oversight settings	Supports calibrated interpretation	Reliance-oriented evaluation	More explanation may increase misplaced confidence	Behavioural and decision-outcome testing
Interpretation boundaries	Does explanation delivery establish understanding or beneficial reliance?	All users and use cases	Prevents unsupported human-outcome claims	Explicit boundary statement	Understanding and benefit may be assumed rather than tested	Task-specific human validation

Proposed Multimodal Explainability Architecture

The proposed architecture treats multimodal explainability as an evidence-following process rather than a collection of disconnected explanation techniques. Its first layer defines the use case, consequence level, modality context, stakeholder roles, and interpretation needs. A second layer records the modality inventory and classifies primary, derived, retrieved, generated, and imputed evidence. These layers establish what the system receives and what different users need to inspect before representation or attribution is considered [25].

The next layers document modality-specific representation and provenance. Separate pathways record how text, vision, audio, sensor, and structured-data evidence is preprocessed, encoded, and made available for explanation. The architecture does not require every modality to use the same explanatory format. Instead, it requires compatible records of source identity, transformation, quality, missingness, and representational stage. Explicit documentation of modality composition and fusion location is necessary because inconsistent descriptions can obscure materially different multimodal designs [26].

The fusion layer maps early, intermediate, late, and hybrid integration, including representation alignment, cross-attention, and other modality-interaction mechanisms where relevant. Cross-attention may help identify where internal interactions occur, but it is retained as a model operation rather than elevated automatically to an explanation. Fusion attribution is therefore paired with modality-removal tests, interaction analysis, disagreement review, and explicit uncertainty about contribution estimates [27].

Dedicated modality branches preserve modality-appropriate explanation requirements. Structured-data explanations should retain feature meanings, encodings, dependencies, units, and local or global scope [28]. Sensor and time-series explanations should preserve ordering, sampling, duration, subsequences, drift, and temporal stability rather than reducing sequential evidence to static importance rankings [29]. Parallel requirements apply to textual rationales, visual saliency, and audio attribution, each of which remains subject to its own fidelity and interpretation limits.

The final layers connect evidence traceability, explanation consistency, uncertainty communication, cognitive accessibility, auditability, evaluation, and feedback updating. Fusion-stage selection changes which evidence transformations can be reconstructed and which explanatory claims can be supported, so the architecture records both forward evidence flow and corrective feedback [30]. **Table 3** presents the proposed multimodal explainability architecture linking modality inventory, evidence representation, cross-modal fusion, attribution, traceability, consistency, uncertainty communication, human interpretation, evaluation, and feedback updating.

Table 3. Proposed Multimodal Explainability Architecture for Text, Vision, Audio, Sensor, and Structured-Data Intelligent Systems: Modality Inventory, Evidence Classification, Representation Mapping, Cross-Modal Fusion, Fusion Attribution, Modality Conflict Review, Evidence Traceability, Explanation Consistency, Uncertainty Communication, Human Interpretation, Auditability, Evaluation, Feedback Updating, and Deployment Boundaries

Architecture layer	Core purpose	Required input	Explainability function	Potential output	Validation or evaluation need	Failure risk	
Use-case and modality-context definition	Establish task, consequence, environment, and system scope	Intended use, domain, operating context, and system description	Defines relevant explanation questions	Context specification	Stakeholder and domain review	Generic explanations may be applied to unsuitable contexts	
Stakeholder and interpretation-need mapping	Identify explanation recipients and goals	Roles, expertise, decisions, and accessibility needs	Matches explanation depth and form to purpose	Stakeholder map	Human-centred task analysis	One interface may serve incompatible needs poorly	U
Modality inventory and evidence classification	Record all observed, derived, retrieved, generated, and imputed evidence	Modality list, source metadata, availability, and evidence status	Prevents hidden or conflated inputs	Evidence inventory	Completeness and provenance review	Important sources or transformations may remain undocumented	
Text representation and attribution	Trace documents, passages, tokens, embeddings, and rationales	Text sources and language-processing records	Links language evidence to model and explanation outputs	Text-evidence trace	Rationale-fidelity and token-attribution evaluation	Fluent rationales may be unfaithful	
Vision representation	Trace images, frames, objects,	Visual sources and	Supports region- and	Visual-evidence	Saliency sanity, stability, and	Heatmaps may be persuasive	n

and attribution	regions, and visual features	preprocessing records	representation-level analysis	trace	perturbation testing	but incomplete
Audio representation and attribution	Trace speech, sounds, acoustic features, and temporal segments	Waveform, transcript, acoustic representation, and timing data	Supports time-aligned explanation	Audio-evidence timeline	Audio-specific faithfulness and intelligibility testing	Spectrogram displays may be semantically misleading
Sensor and time-series representation	Preserve device, timing, calibration, and sequence context	Sensor identity, stream, timestamps, gaps, and quality information	Links attribution to temporal and device provenance	Sensor-evidence trace	Drift, corruption, missingness, and stability tests	Faulty or incomplete streams may appear authoritative
Structured-data representation and attribution	Preserve schema, feature meaning, transformation, and dependency	Numerical, categorical, relational, and derived variables	Supports local, global, and interaction-sensitive explanation	Structured-evidence profile	Correlation, proxy, encoding, and plausibility checks	Feature rankings may conceal dependencies
Provenance and data-quality review	Qualify every source before attribution	Lineage, acquisition context, transformations, and quality indicators	Attaches reliability context to evidence	Provenance and quality record	Source and lineage audit	Low-quality or imputed evidence may appear verified
Cross-modal fusion mapping	Record where and how evidence combines	Architecture graph, encoders, alignments, and fusion operators	Locates interaction stages	Fusion map	Architecture verification	Decisive fusion operations may be omitted
Fusion attribution and conflict review	Examine joint contribution, dominance, absence, and contradiction	Fusion states, modality outputs, availability, and quality records	Tests interaction-sensitive evidence use	Qualified fusion-attribution and conflict report	Ablation, perturbation, stress, and method-comparison testing	Contributions may be non-unique or method-dependent
Evidence traceability and consistency	Connect sources, representations, fusion, decision, and communicated explanation	Versioned evidence and model records	Supports reconstruction and contradiction detection	Evidence chain and consistency matrix	Fidelity, completeness, and independent validation	Consistent explanations may share an error

Uncertainty communication	Distinguish evidential, predictive, and explanatory uncertainty	Data-quality, model, and method-variation information	Qualifies explanatory claims	Layered uncertainty statement	Calibration and user-comprehension testing	Numeric confidence may imply false precision
Human interpretation and cognitive accessibility	Present evidence according to role, expertise, and task	Stakeholder requirements and explanation outputs	Supports progressive, interactive, and accessible interpretation	Role-sensitive explanation interface	Accessibility, comprehension, and reliance evaluation	Overload, simplification, or persuasive presentation may mislead
Auditability, evaluation, and feedback updating	Preserve reconstruction and revise the architecture over time	Version history, evaluations, incidents, and user feedback	Supports lifecycle review	Audit trail and updated requirements	Repeated technical, domain, and human assessment	Static explanation processes may become obsolete
Architecture boundary layer	Preserve the limited conceptual status of the proposal	Results from all preceding layers	Prevents sufficiency and readiness overclaims	Explicit boundary record	Independent empirical and contextual assessment	Architecture diagrams may be mistaken for validated controls

Evaluation and Research Implications

Evaluation should remain modality-specific while also examining system-level interaction. Text explanations require separation of source evidence, token attribution, and generated rationales [9]. Visual explanations require sanity, robustness, and representation-sensitive testing [10]. Audio explanations need temporal and acoustic validation [11], while sensor, time-series, and tabular explanations must preserve sequence structure, feature semantics, encoding, dependencies, and provenance [12]. A single universal metric would conceal these differences.

Cross-modal evaluation should test whether explanation methods identify complementarity, dominance, conflict, missingness, and fusion-stage behaviour. Research is needed on interaction-aware attribution, uncertainty in modality contribution, explanation disagreement, and the relationship between modality-specific outputs and fused decisions. Multimodal-XAI evaluation should therefore combine fidelity, robustness, consistency, completeness, and human usefulness rather than relying on visually persuasive examples [16].

Evidence traceability and human interpretation also require separate evaluation programmes. Technical studies can test reconstruction, perturbation sensitivity, representation mapping, and explanation stability [18]. Human-centred studies should assess comprehension, cognitive accessibility, misconception, error detection, task performance, and reliance without assuming that positive perceptions establish understanding [21]. Interactive explanation and feedback mechanisms require longitudinal investigation because user needs and system behaviour may change over time [24].

Emerging multimodal foundation models create additional research questions concerning cross-attention, retrieved evidence, generated language, latent representation, tool use, and provenance across long interaction chains [25]. Lifecycle research should examine how explanation quality changes after data shifts, sensor drift, model updates, or modality substitution [29]. These implications are conceptual research directions, not legal advice, compliance guidance, procurement or audit instruction, model- or platform-selection advice, data-fusion or modality-selection instruction, safety certification, or deployment guidance.

Conclusion

This article has proposed a conceptual multimodal explainability architecture for following evidence across text, vision, audio, sensor, and structured-data intelligent systems. The architecture links use-case and stakeholder definition with modality inventory, representation mapping, provenance, cross-modal fusion, fusion attribution, missingness and conflict review, evidence traceability, explanation consistency, uncertainty communication, human interpretation, evaluation, and feedback updating. Its central premise is that multimodal explainability should be assessed through complete and qualified evidence chains rather than isolated saliency maps, attention weights, textual rationales, feature rankings, or confidence values. Modality-specific explanations remain necessary, but they are insufficient without analysis of interaction, fusion, uncertainty, and omission. Similarly, traceability does not prove accountability, consistency does not prove correctness, and accessible presentation does not guarantee understanding or appropriate reliance. The architecture therefore supports disciplined conceptual development and future evaluation while remaining subject to empirical, domain-specific, human-centred, and lifecycle validation.

Acknowledgements

None

Conflict of interest

None

Financial support

None

Ethics statement

None

Received: 21 Jul 2025

Revised: 23 Aug 2025

Accepted: 15 Sep 2025

Published online: 15 March 2026

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Baltrušaitis T, Ahuja C, Morency LP. Multimodal machine learning: A survey and taxonomy. *IEEE Trans Pattern Anal Mach Intell*. 2019;41(2):423-43.
<https://doi.org/10.1109/TPAMI.2018.2798607>.
- Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J, Bennetot A, Tabik S, Barbado A, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Inf Fusion*. 2020;58:82-115.
<https://doi.org/10.1016/j.inffus.2019.12.012>.
- Miller T. Explanation in artificial intelligence: Insights from the social sciences. *Artif Intell*. 2019;267:1-38.
<https://doi.org/10.1016/j.artint.2018.07.007>.
- Mohseni S, Zarei N, Ragan ED. A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *ACM Trans Interact Intell Syst*. 2021;11(3-4):24:1-45.
<https://doi.org/10.1145/3387166>.
- Ramachandram D, Taylor GW. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Process Mag*. 2017;34(6):96-108.
<https://doi.org/10.1109/MSP.2017.2738401>.
- Acosta JN, Falcone GJ, Rajpurkar P, Topol EJ. Multimodal biomedical AI. *Nat Med*. 2022;28(9):1773-84.
<https://doi.org/10.1038/s41591-022-01981-2>.
- Rudin C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat Mach Intell*. 2019;1(5):206-15.
<https://doi.org/10.1038/s42256-019-0048-x>.
- Samek W, Montavon G, Lapuschkin S, Anders CJ, Müller KR. Explaining deep neural networks and beyond: A review of methods and applications. *Proc IEEE*. 2021;109(3):247-78.
<https://doi.org/10.1109/JPROC.2021.3060483>.
- Madsen A, Reddy S, Chandar S. Post-hoc interpretability for neural NLP: a survey. *ACM Comput Surv*. 2022;55(8):155:1-42.
<https://doi.org/10.1145/3546577>.
- Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. *Digit Signal Process*. 2018;73:1-15.
<https://doi.org/10.1016/j.dsp.2017.10.011>.
- Akman A, Schuller BW. Audio explainable artificial intelligence: A review. *Intell Comput*. 2024;3:0074.
<https://doi.org/10.34133/icomputing.0074>.
- Di Martino F, Delmastro F. Explainable AI for clinical and remote health applications: A survey on tabular and time series data. *Artif Intell Rev*. 2023;56(6):5261-315.
<https://doi.org/10.1007/s10462-022-10304-3>.
- Lipkova J, Chen RJ, Chen B, Lu MY, Barbieri M, Shao D, et al. Artificial intelligence for multimodal data integration in oncology. *Cancer Cell*. 2022;40(10):1095-110.
<https://doi.org/10.1016/j.ccell.2022.09.012>.
- Holzinger A, Dehmer M, Emmert-Streib F, Cucchiara R, Augenstein I, Del Ser J, et al. Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. *Inf Fusion*. 2022;79:263-78.
<https://doi.org/10.1016/j.inffus.2021.10.007>.

Yang G, Ye Q, Xia J. Unbox the black-box for the medical explainable AI via multi-modal and multi-centre data fusion: A mini-review, two showcases and beyond. *Inf Fusion*. 2022;77:29-52.
<https://doi.org/10.1016/j.inffus.2021.07.016>.

Rodis N, Sardianos C, Radoglou-Grammatikis P, Sarigiannidis P, Varlamis I, Papadopoulos GT. Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions. *IEEE Access*. 2024;12:159794-820.
<https://doi.org/10.1109/ACCESS.2024.3467062>.

Vilone G, Longo L. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Inf Fusion*. 2021;76:89-106.
<https://doi.org/10.1016/j.inffus.2021.05.009>.

Nauta M, Trienes J, Pathak S, Nguyen E, Peters M, Schmitt Y, et al. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI. *ACM Comput Surv*. 2023;55(13s):1-42.
<https://doi.org/10.1145/3583558>.

Lapuschkin S, Wäldchen S, Binder A, Montavon G, Samek W, Müller KR. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun*. 2019;10(1):1096.
<https://doi.org/10.1038/s41467-019-08987-4>.

Hüllermeier E, Waegeman W. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Mach Learn*. 2021;110(3):457-506.
<https://doi.org/10.1007/s10994-021-05946-3>.

Rong Y, Leemann T, Nguyen TT, Fiedler L, Qian P, Unhelkar V, et al. Towards human-centered explainable AI: A survey of user studies for model explanations. *IEEE Trans Pattern Anal Mach Intell*. 2024;46(4):2104-22.
<https://doi.org/10.1109/TPAMI.2023.3331846>.

Shin D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int J Hum Comput Stud*. 2021;146:102551.
<https://doi.org/10.1016/j.ijhcs.2020.102551>.

Speith T, Crook B, Mann S, Schomäcker A, Langer M. Conceptualizing understanding in explainable artificial intelligence (XAI): An abilities-based approach. *Ethics Inf Technol*. 2024;26(2):40.
<https://doi.org/10.1007/s10676-024-09769-3>.

Raees M, Meijerink I, Lykourantzou I, Khan VJ, Papangelis K. From explainable to interactive AI: A literature review on current trends in human-AI interaction. *Int J Hum Comput Stud*. 2024;189:103301.
<https://doi.org/10.1016/j.ijhcs.2024.103301>.

Liang PP, Zadeh A, Morency LP. Foundations and trends in multimodal machine learning: principles, challenges, and open questions. *ACM Comput Surv*. 2024;56(10):264:1-42.
<https://doi.org/10.1145/3656580>.

Sleeman WC 4th, Kapoor R, Ghosh P. Multimodal classification: current landscape, taxonomy and future directions. *ACM Comput Surv*. 2022;55(7):150:1-31.
<https://doi.org/10.1145/3543848>.

Xu P, Zhu X, Clifton DA. Multimodal learning with transformers: A survey. *IEEE Trans Pattern Anal Mach Intell*. 2023;45(10):12113-32.
<https://doi.org/10.1109/TPAMI.2023.3275156>.

Sahakyan M, Aung Z, Rahwan T. Explainable artificial intelligence for tabular data: A survey. *IEEE Access*. 2021;9:135392-422.
<https://doi.org/10.1109/ACCESS.2021.3116481>.

Theissler A, Spinnato F, Schlegel U, Guidotti R. Explainable AI for time series classification: A review, taxonomy and research directions. *IEEE Access*. 2022;10:100700-24.
<https://doi.org/10.1109/ACCESS.2022.3207765>.

Stahlschmidt SR, Ulfenborg B, Synnergren J. Multimodal deep learning for biomedical data fusion: a review. *Brief Bioinform*. 2022;23(2).