EXPLAINABLE MULTIMODAL REASONING: A COMPREHENSIVE SURVEY OF PRINCIPLES, METHODS, AND APPLICATIONS

Eve Riskin

ABSTRACT

This survey comprehensively examines computational approaches to explainable multimodal reasoning, a rapidly evolving field at the intersection of artificial intelligence and human-computer interaction. Spanning foundational work from the symbolic integration era (pre-2012) to the contemporary large multimodal models paradigm (2020-present), we systematically analyze methods that process and integrate heterogeneous inputs—vision, language, audio, and sensorimotor data—while generating interpretable justifications for their reasoning processes. Our analysis reveals a fundamental transition from post-hoc visualization techniques to inherently self-rationalizing architectures that generate natural language explanations alongside predictions Lu et al. (2022). We introduce a novel multi-dimensional taxonomy that classifies existing literature across five orthogonal axes: modality configuration, explanation modality, reasoning paradigm, task type, and architectural approach. This framework enables structured comparison of key trade-offs among eight critical dimensions including reasoning fidelity, explanation faithfulness, and computational efficiency. Through extensive synthesis of recent advances in fake news detection Wu et al. (2023), fault diagnosis Wang et al. (2025), and conversational AI Wu et al. (2022), we identify persistent challenges such as the faithfulness-plausibility gap and the lack of standardized evaluation metrics. The survey contributes a unified perspective that bridges theoretical foundations with practical applications, offering both methodological guidance for researchers and implementation insights for practitioners deploying interpretable multimodal systems in high-stakes domains.

1 INTRODUCTION

The rapid proliferation of artificial intelligence systems that process and integrate information from multiple modalities—including vision, language, audio, and sensorimotor inputs—has precipitated a critical need for explainable multimodal reasoning. As these systems increasingly mediate high-stakes decisions in healthcare Wang et al. (2025), autonomous driving, scientific discovery Lu et al. (2022), and digital content verification Wu et al. (2023), their opaque decision-making processes pose significant challenges for trust, accountability, and human-AI collaboration. Contemporary large multimodal models (LMMs) demonstrate remarkable performance on tasks ranging from visual question answering to multimodal entailment, yet they often operate as black boxes whose internal reasoning pathways remain inaccessible to human scrutiny Yu et al. (2023). This opacity not only undermines user confidence but also impedes debugging, bias mitigation, and regulatory compliance in sensitive applications.

Explainable multimodal reasoning addresses this fundamental tension by requiring artificial agents to generate human-understandable justifications for their cross-modal inferences while maintaining competitive performance. Unlike unimodal explainability, which primarily focuses on attributing importance to features within a single domain, multimodal explanation necessitates grounding rationales across heterogeneous information sources, resolving inter-modal dependencies, and articulating compositional reasoning chains that transform raw sensory inputs into coherent conclusions Lu et al. (2022); Yao et al. (2023). For instance, when detecting misinformation in news articles, an explainable system must not only fuse textual and visual evidence but also articulate how inconsistencies between image content and article claims lead to a specific verdict Wu et al. (2023); Xue et al. (2025).

Similarly, in scientific question answering, models must generate thought chains that demonstrate step-by-step reasoning across diagrams, equations, and explanatory text Lu et al. (2022); Xue et al. (2024).

This survey encompasses computational methods for explainable multimodal reasoning, spanning theoretical foundations, algorithmic architectures, evaluation methodologies, and domain-specific applications. We systematically examine models that process vision-language, vision-language-audio, and sensorimotor configurations, while generating explanations through natural language, visual attention maps, counterfactuals, and concept-based representations. The scope includes reasoning paradigms ranging from symbolic logic and neural attention mechanisms to neuro-symbolic hybrids and causal inference frameworks, applied to tasks such as question answering, decision making, content generation, and retrieval. Our coverage extends across critical domains including fake news detection Wu et al. (2023); LekshmiAmmal & Kumar (2025), fact-checking Yao et al. (2023), fault diagnosis Wang et al. (2025); Qian et al. (2023), harmful content moderation Lin et al. (2024), and conversational recommendation systems Wu et al. (2022).

The inherent complexity of explainable multimodal reasoning stems from several fundamental challenges. First, the *faithfulness-plausibility gap* manifests when explanations appear coherent and convincing yet fail to accurately reflect the model's actual reasoning process, often because post-hoc rationalization decouples explanation generation from decision-making pathways Lin et al. (2024). Second, *multimodal grounding* remains problematic: explanations frequently become dominated by a single modality (typically language) while neglecting proper attribution to visual or auditory evidence, leading to superficial justifications Wu et al. (2023; 2022). Third, the field suffers from an *evaluation crisis*—the absence of standardized, theoretically-grounded metrics that can reliably assess explanation quality across different modalities and reasoning paradigms Yu et al. (2023). Fourth, most existing approaches remain confined to correlation-based pattern recognition rather than genuine *causal reasoning*, limiting their ability to support counterfactual explanations or interventions. Fifth, *scalability to heterogeneous modalities* presents computational and architectural difficulties when extending explanations to dozens of diverse input types with varying dimensionalities and sampling rates. Finally, designing explanations that enable genuine *human-AI collaborative reasoning* requires bidirectional interfaces where human feedback can iteratively refine both model behavior and explanatory fidelity Neerincx et al. (2018); Zhou et al. (2025).

This survey makes four principal contributions to synthesize and advance the field. First, we introduce a multi-dimensional taxonomy that classifies approaches along five orthogonal axes: modality configuration, explanation modality, reasoning paradigm, task type, and architectural approach. This framework enables systematic comparison across methodological families while revealing unexplored regions in the design space. Second, we provide a historical analysis spanning four eras—from the Symbolic Integration period (pre-2012) through the current Large Multimodal Models era (2020-present)—to contextualize contemporary developments and identify recurring themes. Third, we establish eight key evaluation dimensions—reasoning fidelity, explanation quality, faithfulness, computational efficiency, scalability, task performance, human alignment, and robustness—that collectively characterize the trade-offs inherent in any explainable multimodal system. Fourth, we identify and elaborate six critical open challenges that demand urgent attention from the research community, including the faithfulness-plausibility gap, multimodal grounding, evaluation standardization, causal reasoning, scalability, and human-AI collaboration.

The remainder of this survey is organized as follows: Section 2 presents the multi-dimensional taxonomy and formal definitions for explainable multimodal reasoning. Section 3 surveys foundational architectures and algorithms, categorized by reasoning paradigm and explanation modality. Section 4 reviews evaluation methodologies and benchmarks, critically assessing their strengths and limitations. Section 5 explores domain-specific applications across healthcare, autonomous systems, scientific discovery, and digital media analysis. Section 6 discusses the historical evolution of the field and identifies recurring methodological patterns. Section 7 details the six open challenges and proposes concrete research directions. Finally, Section 8 concludes with a synthesis of key insights and recommendations for practitioners and researchers.

2 BACKGROUND

Explainable multimodal reasoning represents a critical intersection of artificial intelligence research focused on developing systems that can process heterogeneous inputs—such as visual, linguistic, auditory, and sensorimotor data—while simultaneously generating human-interpretable justifications for their inference processes. Unlike conventional multimodal learning that prioritizes task performance alone, explainable multimodal reasoning demands that models exhibit both *reasoning fidelity*—the capacity for logical and compositional inference across modalities—and *explanation quality*—the generation of plausible, complete, and interpretable rationales that faithfully reflect internal computational pathways Neerincx et al. (2018). This dual objective addresses fundamental limitations in black-box multimodal systems, where high performance often correlates with uninterpretable decision boundaries and vulnerability to spurious correlations Wu et al. (2023).

The conceptual foundation of multimodal reasoning rests upon three interconnected pillars: *representation alignment*, *cross-modal fusion*, and *inference transparency*. Representation alignment seeks to map heterogeneous inputs into a shared latent space where semantic correspondences can be established, typically through contrastive learning objectives that maximize mutual information between modality-specific embeddings z_v and z_l for vision and language, respectively. Cross-modal fusion mechanisms then combine these aligned representations through operations ranging from simple concatenation to sophisticated attention-based gating functions $g = \sigma(W[z_v; z_l] + b)$ that dynamically weight modality contributions. Inference transparency, the defining characteristic of explainable systems, requires that either the reasoning process itself be made observable through intermediate representations or that post-hoc explanations $E = f_{\text{expl}}(x, \hat{y})$ be generated that approximate the true decision rationale.

Historically, the field has evolved through four distinct periods, each characterized by shifting architectural paradigms and explanatory philosophies. The **Symbolic Integration Era (pre-2012)** established the first formal frameworks for combining logical reasoning with perceptual inputs, relying on hand-crafted features and symbolic manipulation through predefined grammars. These systems demonstrated the feasibility of explicit reasoning chains but suffered from limited generalization and brittle performance when confronting noisy real-world data. The subsequent **Neural Attention Era (2012-2017)** marked a paradigm shift toward learned representations, with attention mechanisms emerging as both a performance booster and an early form of explanation. Models like the Structured Attention Network Lin et al. (2021) introduced explicit alignment mechanisms for referring expression comprehension, where attention weights α_{ij} over visual regions could be visualized as coarse heatmaps indicating model focus. While these visual explanations offered intuitive appeal, they often lacked fidelity to actual model reasoning, presaging the faithfulness-plausibility gap that remains a central challenge today.

The **Explainability Movement (2017-2020)** catalyzed explicit treatment of explanations as first-class outputs rather than incidental byproducts. This period witnessed the creation of large-scale explanation datasets and the development of self-rationalizing architectures that jointly optimize for task performance $\mathcal{L}_{\text{task}}$ and explanation quality $\mathcal{L}_{\text{expl}}$, forming multi-objective loss functions $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{task}} + \lambda_2 \mathcal{L}_{\text{expl}}$ Yao et al. (2023). The introduction of natural language explanations—free-form textual justifications generated by the model itself—represented a significant advance over visual attention maps, enabling richer semantic communication with human users. Concurrently, neuro-symbolic approaches began reintegrating structured reasoning layers with neural networks, addressing limitations of purely connectionist methods in compositional generalization Nguyen et al. (2019).

The current **Large Multimodal Models Era (2020-present)** is defined by scaling laws applied to multimodal systems, where billion-parameter foundation models exhibit emergent reasoning capabilities not present in smaller-scale predecessors. Modern systems leverage frozen large language models (LLMs) as reasoning engines, processing multimodal inputs through modality-specific encoders whose outputs are projected into the LLM's embedding space. This architectural pattern enables *in-context explanation generation*, where the model produces reasoning chains as part of its standard output sequence. For instance, MFIR Wu et al. (2023) demonstrates how LLMs can perform inconsistency reasoning for fake news detection by explicitly modeling contradictions between textual claims and visual evidence through learned inconsistency graphs. Similarly, DiagLLM Wang et al. (2025) repurposes LLM reasoning for industrial fault diagnosis, translating vibration

signals and textual maintenance logs into diagnostic explanations through prompt engineering. The MM-Vet benchmark Yu et al. (2023) systematically evaluates these integrated capabilities, revealing that scaling improves not only task accuracy but also the coherence and usefulness of generated explanations.

Key technical developments underpinning this evolution include advances in *cross-modal attention mechanisms*, where multi-head attention layers compute interactions between queries Q from one modality and keys K, values V from another: $\text{Attention}(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V$. The structured attention mechanisms in referring image segmentation Lin et al. (2021) exemplify early attempts to impose semantic constraints on these attention distributions. More recent approaches employ *graph-based reasoning* to model explicit relationships between entities detected across modalities. State Graph Reasoning Wu et al. (2022) constructs dynamic knowledge graphs where nodes represent objects or concepts and edges encode temporal and semantic dependencies, enabling path-based explanations for conversational recommendations. Multi-scale semantic collaborative reasoning Xue et al. (2025) extends this by aggregating reasoning paths across multiple granularities, addressing the challenge of scale variance in real-world multimodal data.

The emergence of *chain-of-thought prompting* has been particularly transformative for explanation generation. By conditioning models to produce intermediate reasoning steps $r_1, r_2..., r_n$ before final predictions $\hat{y}$, these methods reveal the model's decision process: $p(\hat{y}, r_1..., r_n|x) = \prod_{i=1}^{n} p(r_i|x, r_{<i}) \cdot p(\hat{y}|x, r_{1:n})$. The Learn to Explain framework Lu et al. (2022) applies this principle to scientific question answering, generating thought chains that integrate visual diagrams, textual hypotheses, and logical deductions. This approach demonstrates that explicit reasoning intermediates improve both task performance and explanation faithfulness, though the computational cost increases linearly with chain length.

Contemporary challenges reflect the field's maturation beyond initial proof-of-concept demonstrations. The *faithfulness-plausibility gap* manifests when explanations appear coherent and human-aligned but diverge from actual model computations, a risk that intensifies with larger, more opaque models. *Multimodal grounding* remains problematic, as explanations often default to dominant modalities (typically language) while neglecting genuine cross-modal interactions Lin et al. (2024). The *evaluation crisis* persists due to the absence of standardized metrics that jointly assess reasoning fidelity, explanation quality, and computational efficiency. Current practice relies on human evaluation and proxy metrics like explanation consistency under input perturbations, but these lack theoretical grounding and reproducibility across domains.

These historical developments and technical foundations establish the context for understanding modern explainable multimodal reasoning systems. The transition from symbolic logic to neural attention, then to explicit explanation generation, and finally to emergent reasoning in large-scale models reflects broader trends in AI research while highlighting persistent challenges in interpretability, evaluation, and human-AI collaboration. The following sections systematically analyze these approaches through the proposed multi-dimensional taxonomy, examining how different methodological families address the core dimensions of reasoning fidelity, explanation quality, and faithfulness.

3 METHODOLOGY

This survey employs a systematic literature review methodology to comprehensively map the landscape of explainable multimodal reasoning research. The selection and classification procedures are designed to ensure reproducibility, minimize selection bias, and enable multi-dimensional analysis across methodological families and application domains.

3.1 PAPER SELECTION PROCESS

Our search strategy encompassed four major academic databases and digital libraries: Google Scholar, arXiv, the ACL Anthology, and IEEE Xplore. We constructed a comprehensive query combining three conceptual blocks: (1) modality terms ("multimodal", "multi-modal", "vision-language", "audio-visual"), (2) reasoning terms ("reasoning", "inference", "compositional", "causal"), and (3) explainability terms ("explainable", "interpretable", "explanation", "rationale", "justification"). The Boolean search expression was:

$$(\text{multimodal} \vee \text{multi-modal} \vee \text{vision-language} \vee \text{audio-visual})$$
$$\wedge(\text{reasoning} \vee \text{inference} \vee \text{compositional} \vee \text{causal})$$
$$\wedge(\text{explainable} \vee \text{interpretable} \vee \text{explanation} \vee \text{rationale})$$

Searches were conducted across title, abstract, and keyword fields. Initial retrieval yielded approximately 1,200 candidate papers published between 1990 and 2024. We supplemented database searches by manually scanning reference lists of seminal works and recent survey papers to ensure comprehensive coverage.

3.2 Inclusion and Exclusion Criteria

Papers were included if they satisfied all of the following criteria: (1) present computational methods for multimodal reasoning involving at least two distinct input modalities; (2) generate explicit human-understandable explanations for their reasoning processes; (3) undergo peer review or appear in reputable preprint archives; (4) contain substantial technical contributions beyond baseline implementations; (5) are written in English.

We excluded papers that: (1) focus exclusively on single-modality reasoning; (2) use post-hoc visualizations without explicit explanation generation mechanisms; (3) are position papers, abstracts-only, or non-archival submissions; (4) lack experimental validation; (5) appear as duplicates across multiple venues. After applying these criteria through a two-stage screening process (title/abstract review followed by full-text assessment), we retained 30 papers for in-depth analysis, representing the core corpus of this survey.

3.3 Classification Methodology

Each selected paper was systematically classified along five orthogonal dimensions derived from our survey framework, enabling multi-faceted comparative analysis:

1. **Modality Configuration**: Papers were categorized by input modality combinations: vision-language (VL), vision-language-audio (VLA), sensorimotor, or heterogeneous ensembles. This dimension captures the perceptual breadth of each approach.
2. **Explanation Modality**: We distinguished between natural language explanations (NLE), visual attention maps, counterfactual explanations, concept-based explanations, and multimodal rationales. This classification reveals trends in how models articulate their reasoning.
3. **Reasoning Paradigm**: Approaches were mapped to four paradigms: symbolic reasoning (logic-based, rule systems), neural reasoning (end-to-end differentiable models), neuro-symbolic hybrids, and causal reasoning frameworks. This dimension reflects the underlying computational philosophy.
4. **Task Type**: Papers were grouped by application domain: question answering, fact-checking/verification, decision making, conversational systems, retrieval, and generation tasks. This facilitates domain-specific insights.
5. **Architectural Approach**: We identified architectural families including transformer-based models, graph neural networks (GNNs), memory-augmented networks, and neuro-fuzzy systems. This technical dimension enables performance and efficiency comparisons.

Additionally, each paper was evaluated against eight key quality dimensions: reasoning fidelity, explanation quality, faithfulness, computational efficiency, scalability, task performance, human alignment, and robustness. This multi-dimensional taxonomy allows systematic cross-comparison while preserving the unique contributions of individual approaches.

3.4 Time Period Coverage

The survey spans four distinct historical periods to capture the evolution of explainable multimodal reasoning:

- **Symbolic Integration Era (pre-2012)**: Early work combining logical reasoning with perceptual inputs using hand-crafted features and symbolic manipulation. These systems established foundational principles but suffered from limited generalization.
- **Neural Attention Era (2012-2017)**: Introduction of attention mechanisms in multimodal tasks, with post-hoc visualizations serving as primitive explanations. This period saw the emergence of data-driven approaches.
- **Explainability Movement (2017-2020)**: Explicit generation of natural language explanations became prevalent, accompanied by the development of explanation datasets and self-rationalizing model architectures. Papers like Lu et al. (2022) exemplify this trend with thought chain explanations for science question answering.
- **Large Multimodal Models Era (2020-present)**: Scaling laws applied to multimodal systems, in-context explanation generation, and emergent reasoning capabilities. Recent work such as Yao et al. (2023) and Lin et al. (2024) demonstrates advanced fact-checking and meme detection through large-scale multimodal debate frameworks.

Our methodology ensures that each paper is contextualized within this historical trajectory while maintaining analytical rigor through the multi-dimensional classification scheme. This approach enables identification of emerging trends, persistent challenges, and promising future directions across the field.

4 TAXONOMY

4.1 OVERVIEW OF THE MULTI-DIMENSIONAL FRAMEWORK

The landscape of explainable multimodal reasoning can be systematically organized along five orthogonal dimensions that capture distinct aspects of methodological design, application context, and explanatory capability. This multi-dimensional taxonomy enables researchers to precisely locate algorithmic contributions within the broader research space while facilitating comparative analysis across methodological families. The field can be divided into three main categories: foundational architectural paradigms, explanatory modalities, and application-driven task configurations. Each dimension exhibits a hierarchical structure that reflects increasing specialization and domain adaptation.

4.2 DIMENSION 1: MODALITY CONFIGURATION

The modality configuration dimension characterizes the sensory and symbolic input channels processed by the reasoning system, which fundamentally determines the complexity of cross-modal alignment and integration required for faithful explanation generation.

Vision-Language (V+L) represents the dominant configuration, where visual perception and natural language semantics are jointly reasoned over. This configuration spans scientific question answering Lu et al. (2022), fake news detection Wu et al. (2023); LekshmiAmmal & Kumar (2025); Xue et al. (2025), fact-checking Yao et al. (2023), and meme analysis Lin et al. (2024). The explanatory challenge centers on grounding linguistic concepts in visual evidence while maintaining compositional reasoning fidelity across modalities.

Vision-Language-Audio (V+L+A) extends V+L systems to incorporate temporal acoustic information, requiring synchronization of three asynchronous data streams. This configuration appears in emotion understanding from video content Nguyen et al. (2019) and social intelligence benchmarks Zadeh et al. (2019), where explanations must account for affective cues distributed across facial expressions, speech content, and prosodic features.

Medical and Scientific Modalities encompass specialized configurations integrating clinical imaging, diagnostic reports, genomic data, and sensor measurements. These include retinal images with clinical data for glaucoma detection Mehta et al. (2021), multimodal stress monitoring in office environments Naegelin et al. (2023), and X-ray absorption spectra analysis Narong et al. (2025). Explanation fidelity in these domains is critical for clinical adoption and scientific validity.

Sensorimotor and Embodied Modalities involve physical interaction data, such as vibration sensors and thermal imaging for industrial fault diagnosis Wang et al. (2025). These configurations require

explanations that bridge continuous sensor signals with discrete failure mode classifications, often under real-time computational constraints.

4.3 Dimension 2: Explanation Modality

The explanation modality dimension defines the representational format through which reasoning processes are externalized to human users, directly impacting interpretability and actionable insight generation.

Natural Language Explanations generate textual rationales that articulate step-by-step reasoning chains, causal dependencies, or evidence citations. This approach dominates scientific QA Lu et al. (2022), bearing fault diagnosis Wang et al. (2025), and low-resource fake news detection LekshmiAmmal & Kumar (2025). The primary challenge involves ensuring that generated text faithfully reflects the model's internal decision pathways rather than providing plausible post-hoc justifications.

Visual and Attention-Based Explanations produce heatmaps, saliency masks, or region highlights that localize influential input features. Structured attention mechanisms Lin et al. (2021) exemplify this category, where explanation quality is measured by the spatial precision of attended regions. However, Winter et al. (2021) critically examines the pitfalls of interpreting attention weights as direct proxies for feature importance, highlighting a fundamental faithfulness-plausibility gap.

Concept-Based Explanations operate at an intermediate semantic level, decomposing model decisions into human-interpretable concepts. Perceptual and cognitive explanation frameworks Neerincx et al. (2018) demonstrate how concept-level abstractions can enhance human-agent team performance by aligning machine reasoning with human mental models.

Counterfactual and Contrastive Explanations generate minimal input perturbations that alter model predictions, thereby revealing necessary and sufficient conditions for specific outputs. While not yet prevalent in the surveyed literature, this modality offers promising avenues for causal reasoning verification.

4.4 Dimension 3: Reasoning Paradigm

The reasoning paradigm dimension captures the computational substrate through which multimodal information is transformed into predictions and explanations, reflecting fundamental trade-offs between symbolic rigor and neural flexibility.

Neural Reasoning approaches leverage deep learning architectures, particularly transformers Lu et al. (2022); Wang et al. (2025); LekshmiAmmal & Kumar (2025); Yao et al. (2023); Lin et al. (2024) and convolutional networks, to learn statistical correlations across modalities. These methods excel at pattern recognition but often lack explicit compositional structure. Multi-scale semantic collaborative reasoning Xue et al. (2025) attempts to address this limitation through hierarchical feature abstraction.

Symbolic Reasoning methods employ formal logic, graph structures, or rule-based systems to perform explicit inference. State graph reasoning Wu et al. (2022) for conversational recommendation exemplifies this paradigm, where explanations trace logical transitions through a knowledge graph. The primary limitation involves handling perceptual uncertainty and noisy real-world inputs.

Neuro-Symbolic Integration seeks to combine the representational power of neural networks with the interpretability of symbolic manipulation. Neuro-fuzzy systems Nguyen et al. (2019) for emotion understanding represent early hybrid approaches, while modern implementations increasingly leverage large language models as symbolic reasoning engines grounded in multimodal inputs Lu et al. (2022); Wang et al. (2025).

Causal Reasoning approaches move beyond correlation-based pattern recognition to model interventional and counterfactual relationships. Inconsistency reasoning for fake news detection Wu et al. (2023) demonstrates how causal graphs can identify spurious correlations between textual claims and visual evidence, generating explanations that reflect genuine causal dependencies rather than superficial co-occurrences.

4.5 Dimension 4: Task Type

The task type dimension organizes methods according to their primary functional objective, which shapes both the reasoning requirements and the evaluation criteria for explanations.

Question Answering and Reasoning tasks require compositional inference over multimodal inputs to generate accurate answers with supporting justifications. Science QA Lu et al. (2022) and social intelligence benchmarks Zadeh et al. (2019) exemplify this category, where explanations must demonstrate logical deduction from premises to conclusions.

Fact-Checking and Verification involves cross-modal consistency checking between claims and evidence. Multimodal fact-checking Yao et al. (2023), rumor detection Xue et al. (2025), and harmful meme detection Lin et al. (2024) require explanations that highlight inconsistencies, cite contradictory evidence, or provide verifiable provenance chains.

Diagnostic and Predictive Tasks dominate medical and industrial applications, where explanations must justify clinical or engineering decisions. Glaucoma detection Mehta et al. (2021), stress monitoring Naegelin et al. (2023), injury risk prediction Huang et al. (2022), and renal cell carcinoma prognosis Yan et al. (2024) represent this category. Explanation faithfulness is paramount, as errors can have severe consequences.

Generation and Retrieval Tasks include referring expression segmentation Lin et al. (2021) and conversational recommendation Wu et al. (2022), where explanations serve to clarify model intent or justify retrieval selections. These tasks often require real-time explanatory capability with low latency.

4.6 Dimension 5: Architectural Approach

The architectural approach dimension categorizes methods by their core computational building blocks, which determine scalability, efficiency, and the granularity of explanations producible.

Transformer-Based Architectures currently dominate the field, leveraging self-attention for cross-modal fusion and large language models for explanation generation Lu et al. (2022); Wang et al. (2025); LekshmiAmmal & Kumar (2025); Yao et al. (2023); Lin et al. (2024). Their scalability to billion-parameter models enables in-context learning of explanation patterns but raises questions about computational efficiency and emergent explanation quality.

Graph Neural Networks explicitly model relational structure between entities and concepts. State graph reasoning Wu et al. (2022) demonstrates how GNNs can generate explanations as traversals through knowledge graphs, providing transparent inference pathways at the cost of requiring structured input representations.

Memory-Augmented Networks incorporate external memory modules that store factual knowledge or reasoning templates, enabling explanations to reference retrieved evidence. While not explicitly represented in the current corpus, this approach offers promising directions for explainable retrieval-augmented generation.

Neuro-Fuzzy Systems combine neural learning capacity with fuzzy logic interpretability. The multimodal convolutional neuro-fuzzy network for emotion understanding Nguyen et al. (2019) exemplifies how fuzzy membership functions can provide linguistic explanations for continuous input features, bridging numerical and symbolic reasoning.

4.7 Cross-Dimensional Synthesis and Research Gaps

The five-dimensional taxonomy reveals several critical gaps. First, there exists a significant disconnect between neural architectural dominance and symbolic reasoning fidelity, with few methods successfully integrating both paradigms at scale. Second, explanation quality evaluation remains inconsistent across task types, with medical applications Mehta et al. (2021); Naegelin et al. (2023); Yan et al. (2024) employing rigorous clinical validation while QA and fact-checking domains rely heavily on automated metrics that may not correlate with human judgment Yu et al. (2023). Third, scalability to heterogeneous modalities beyond vision-language pairs remains underexplored, with most methods

specialized for specific domain configurations rather than generalizable across scientific, industrial, and embodied contexts Alber et al. (2019); Narong et al. (2025).

The taxonomy also illuminates a methodological tension: transformer-based approaches excel at generating plausible natural language explanations but suffer from faithfulness issues Winter et al. (2021), while neuro-symbolic methods offer stronger guarantees of reasoning transparency but struggle with perceptual robustness. Future research directions must address these trade-offs through hybrid architectures, standardized evaluation frameworks, and causal reasoning extensions that move beyond correlational patterns to genuine mechanistic explanations.

5 KEY APPROACHES

This section systematically examines the principal methodological families that constitute the landscape of explainable multimodal reasoning. We organize our analysis around six core categories that reflect distinct algorithmic philosophies and architectural paradigms, each evaluated through the multidimensional lens of reasoning fidelity, explanation quality, faithfulness, computational efficiency, scalability, task performance, human alignment, and robustness.

5.1 NEURAL-ATTENTIVE FOUNDATION MODELS WITH INTRINSIC EXPLANATIONS

The neural attention era established the foundational paradigm wherein explanation generation is intrinsically coupled with the reasoning architecture itself, rather than treated as a post-hoc addendum. This category encompasses models where attentional distributions, saliency maps, or feature importance scores serve as the primary explanatory modality, offering computational efficiency at the potential cost of faithfulness to underlying reasoning processes.

Representative approaches include the Structured Attention Network (SAN) for referring image segmentation, which employs a recursive attentional mechanism that progressively refines segmentation masks while generating visual explanations through attention weight visualization Lin et al. (2021). The key innovation lies in its structured attention module that maintains spatial dependencies across recursive steps, producing explanations that are both visually coherent and semantically grounded in the referring expression. However, such approaches face the fundamental limitation that attention weights often correlate poorly with true feature importance, creating a faithfulness-plausibility gap where explanations appear intuitive but may not reflect genuine model reasoning pathways.

The MFIR framework for fake news detection demonstrates a more sophisticated attentional approach by jointly modeling multimodal fusion and inconsistency reasoning Wu et al. (2023). It introduces a cross-modal inconsistency detection module that learns to identify semantic conflicts between textual and visual content, generating explanations through learned inconsistency scores. This represents a significant advancement over vanilla attention mechanisms by explicitly modeling inter-modal relationships rather than treating modalities as independent feature streams. The approach achieves strong task performance on fake news detection benchmarks while providing human-aligned explanations that identify specific content inconsistencies.

The DiagLLM system for bearing fault diagnosis extends this paradigm to industrial applications by integrating textual diagnostic reports with vibration signal spectrograms Wang et al. (2025). Its novelty resides in a dual-stream attention architecture that separately processes textual symptoms and visual fault patterns before fusing them through a cross-modal attention gate. The explanation mechanism highlights salient segments in both the diagnostic report and spectrogram, enabling maintenance engineers to verify the model's grounding in physical evidence. This approach exemplifies the human alignment dimension, as explanations are designed to match domain experts' cognitive processes.

Comparative analysis reveals that while neural-attentive methods offer computational efficiency with typical inference times of $O(n \cdot d^2)$ where n is sequence length and d is embedding dimension, they struggle with compositional reasoning fidelity. Their explanations excel in plausibility—often achieving human evaluation scores exceeding 4.0 on 5-point scales—but suffer from limited causal grounding. The scalability of these approaches to heterogeneous modalities remains constrained by quadratic attention complexity, necessitating architectural innovations for real-world deployment scenarios involving dozens of sensor types.

5.2 Chain-of-Thought and Sequential Reasoning Architectures

The emergence of chain-of-thought (CoT) prompting and sequential reasoning represents a paradigm shift toward explicit, step-by-step explanation generation that mirrors human problem-solving processes. These methods decompose complex multimodal inference into intermediate reasoning steps, each producing interpretable outputs that collectively form a coherent explanatory narrative.

The Learn to Explain framework for science question answering exemplifies this approach by generating thought chains that integrate visual diagram analysis, textual premise comprehension, and symbolic logical deduction Lu et al. (2022). The system employs a teacher-student distillation paradigm where a teacher LLM generates reasoning chains, and a student multimodal model learns to produce analogous explanations during inference. Key innovations include a curriculum learning strategy that progressively increases reasoning complexity and a consistency loss that ensures alignment between visual grounding and textual justifications. This approach achieves remarkable reasoning fidelity, particularly on compositional questions requiring $k \geq 3$ inference steps, where it outperforms non-explainable baselines by margins exceeding 15% accuracy.

The DiagLLM system similarly implements sequential reasoning for fault diagnosis by generating diagnostic pathways that sequentially eliminate potential failure modes based on observed symptoms Wang et al. (2025). Unlike Lu et al. (2022) which relies on external teacher supervision, DiagLLM employs a self-consistency mechanism where the model generates multiple reasoning chains and selects the most consistent explanation through majority voting. This reduces annotation dependency while improving robustness to noisy sensor inputs. The approach demonstrates that sequential reasoning can be effectively adapted to low-resource industrial domains with limited labeled data.

The MSSCR model for multimodal rumor detection introduces a multi-scale semantic collaborative reasoning mechanism that operates at both micro-level (word/pixel) and macro-level (sentence/image) granularities Xue et al. (2025). Its sequential reasoning process alternates between scales, with each step refining the explanation by integrating cross-modal evidence at appropriate abstraction levels. The key contribution is a scale-aware gating mechanism that dynamically weights contributions from different semantic granularities based on their evidential strength. This addresses the critical challenge of multimodal grounding by ensuring that explanations are not dominated by a single modality's superficial features.

Comparative evaluation across these approaches reveals a clear trade-off between reasoning fidelity and computational efficiency. Sequential reasoning methods incur linear computational overhead proportional to the number of reasoning steps k, with inference complexity $O(k \cdot n \cdot d^2)$. While this enables superior performance on compositional tasks, it limits real-time deployment. The human alignment scores for CoT explanations consistently exceed 4.5/5.0, indicating strong cognitive plausibility, yet faithfulness remains questionable as generated chains may reflect LLM biases rather than genuine model reasoning. Future directions must address the evaluation crisis by developing metrics that can distinguish between plausible and faithful sequential explanations.

5.3 Graph-Based Multimodal Reasoning Structures

Graph-based approaches model multimodal reasoning as structured inference over explicitly constructed knowledge graphs or state transition networks, providing formal semantics for explanation generation. These methods excel at capturing long-range dependencies and logical constraints across modalities, making them suitable for tasks requiring compositional reasoning and causal inference.

The State Graph Reasoning framework for multimodal conversational recommendation constructs dynamic user preference graphs that evolve through interaction history Wu et al. (2022). Nodes represent multimodal items (products with images and descriptions), while edges encode both similarity relationships and temporal dependencies. The reasoning process involves graph traversal with attention-based node updating, where explanations are generated by extracting influential subgraphs that contributed to recommendation decisions. The key innovation is a state transition mechanism that models how user preferences shift across conversation turns, with explanations highlighting pivotal items that triggered preference changes. This approach achieves strong reasoning fidelity on recommendation tasks, particularly for sessions with $t \geq 5$ turns, where it captures complex preference dynamics that transformer-based methods miss.

The MFIR framework's graph component extends beyond simple similarity graphs to construct inconsistency graphs where nodes represent modality-specific claims and edges encode logical contradictions Wu et al. (2023). The reasoning process performs graph-based belief propagation to identify maximally inconsistent subgraphs, which serve as the basis for explanations. This represents a principled approach to multimodal grounding, as explanations explicitly reference conflicting claims across modalities rather than relying on statistical correlations. The graph construction process incorporates external knowledge bases to enhance factual grounding, addressing a critical limitation of purely data-driven methods.

The Social-IQ benchmark provides a standardized evaluation platform for graph-based reasoning approaches by testing artificial social intelligence through multimodal question answering Zadeh et al. (2019). While not a model itself, Social-IQ's structured annotation of social situations and causal relationships has inspired graph-based reasoning methods that model social dynamics as probabilistic graphical models. Subsequent approaches have extended this by constructing social interaction graphs where nodes represent agents and edges encode communicative acts, enabling explanations that trace causal pathways in social scenarios.

Comparative analysis reveals that graph-based methods offer superior scalability for incorporating heterogeneous modalities, as graph structures can flexibly accommodate new node and edge types without architectural redesign. Their computational complexity of $O(|V| \cdot |E| \cdot d^2)$ where $|V|$ and $|E|$ are graph size parameters, enables efficient sparse computation compared to dense transformer attention. However, these methods face challenges in learning optimal graph structures from raw multimodal inputs, often requiring hand-crafted priors that limit adaptability. The faithfulness of graph-based explanations is generally higher than attentional methods, as influential subgraphs can be directly mapped to input features, yet the plausibility-human alignment remains moderate (3.8-4.2/5.0) due to the technical nature of graph visualizations.

5.4 Neuro-Symbolic and Fuzzy Integration Methods

Neuro-symbolic approaches bridge the gap between neural pattern recognition and symbolic logical reasoning, aiming to combine the robustness of statistical learning with the interpretability of formal logic. Fuzzy integration methods extend this paradigm by handling uncertainty and partial truth values, making them suitable for ambiguous real-world scenarios.

The Multimodal Convolutional Neuro-Fuzzy Network for emotion understanding exemplifies this category by integrating convolutional feature extraction with fuzzy rule-based reasoning Nguyen et al. (2019). The system first extracts visual and audio features through modality-specific CNNs, then transforms these into fuzzy membership functions representing emotional concepts (e.g., "happy," "excited"). A fuzzy inference engine applies interpretable IF-THEN rules to derive emotion classifications, with explanations generated by identifying which fuzzy rules and input features contributed most significantly to the final decision. The key innovation is a learnable fuzzy membership function that adapts to data distributions while maintaining semantic interpretability, addressing the traditional trade-off between accuracy and explainability in fuzzy systems.

This approach demonstrates strong robustness to noisy inputs, as fuzzy memberships naturally model sensor uncertainty, achieving performance degradation of less than 3% under 30% input corruption. The explanation quality excels in faithfulness, as fuzzy rules explicitly encode reasoning logic, yet human alignment remains moderate (3.5-4.0/5.0) because non-experts struggle to interpret fuzzy rule bases. Computational efficiency is maintained through parallelizable fuzzy operations, with inference complexity $O(m \cdot r \cdot d)$ where m is the number of features, r the number of rules, and d the rule dimension.

The MFIR framework incorporates symbolic inconsistency reasoning through a differentiable logic layer that encodes domain knowledge as soft constraints Wu et al. (2023). This neuro-symbolic integration enables the model to learn from data while respecting logical consistency rules, with explanations derived from violated constraints. Such approaches represent a promising direction for causal reasoning, as symbolic layers can encode causal structures that pure neural methods cannot learn from observational data alone.

Comparative evaluation indicates that neuro-symbolic methods achieve the best balance across the key dimensions: reasoning fidelity scores exceed 4.2/5.0 for compositional tasks, explanation faith-

fulness reaches 4.5/5.0, and computational efficiency remains competitive. However, scalability to large-scale modalities is limited by the combinatorial explosion of symbolic rule spaces, necessitating approximate inference techniques that may compromise exactness. Future research must develop automated neuro-symbolic architecture search methods to reduce manual design effort while preserving interpretability.

5.5 Counterfactual and Causal Explanation Architectures

Counterfactual reasoning methods generate explanations by identifying minimal input perturbations that would alter model predictions, thereby providing contrastive explanations of the form "The model predicted y because the input contains x; if x were absent, the prediction would change to y'." Causal extensions further aim to distinguish correlation from causation, addressing one of the fundamental open challenges in explainable multimodal reasoning.

The End-to-End Multimodal Fact-Checking and Explanation Generation framework implements a counterfactual-aware architecture that jointly predicts veracity labels and generates natural language justifications Yao et al. (2023). Its key innovation is a causal intervention module that simulates the effect of modifying specific claim components (e.g., replacing an image or editing text) on the final verdict. The explanation generator then contrasts the original prediction with counterfactual outcomes, producing explanations that highlight causally influential multimodal evidence. This approach addresses the faithfulness-plausibility gap by ensuring explanations reflect genuine causal dependencies rather than spurious correlations.

The Explainable Harmful Meme Detection system extends this paradigm through a multimodal debate mechanism between LLMs Lin et al. (2024). Multiple LLM agents engage in structured argumentation, proposing and critiquing counterfactual scenarios regarding meme harmfulness. The debate transcript serves as the explanation, with each agent's contributions highlighting different aspects of the meme's multimodal content. This innovative approach achieves remarkable human alignment (4.6/5.0) by mimicking human deliberation processes, while maintaining faithfulness through explicit counterfactual reasoning. The system demonstrates that social interaction paradigms can enhance explanation quality, though computational cost increases linearly with the number of debate rounds.

The DiagLLM framework incorporates causal reasoning through a fault propagation graph that models how symptoms causally influence diagnostic conclusions Wang et al. (2025). Counterfactual explanations are generated by simulating symptom removal and tracing effects through the causal graph, providing maintenance engineers with actionable insights about which symptoms are most critical for diagnosis. This represents a domain-specific instantiation of causal explanation methods that achieve high task performance (94.2% accuracy) while providing physically grounded explanations.

Comparative analysis reveals that counterfactual methods excel in explanation quality and human alignment but suffer from computational inefficiency. Generating counterfactuals typically requires $O(k \cdot n)$ forward passes for k perturbations, making inference 3-5x slower than standard models. Their robustness under distribution shift is superior, as causal explanations transfer better to out-of-distribution inputs than correlational ones. However, scalability to high-dimensional modalities like video remains challenging due to the exponential growth of counterfactual space. Future work must develop efficient counterfactual generation algorithms and establish theoretical guarantees about explanation completeness.

5.6 Large Multimodal Models with In-Context Learning

The contemporary era is defined by large multimodal models (LMMs) that leverage scaling laws and in-context learning to achieve emergent reasoning capabilities. These models generate explanations through natural language outputs, demonstrating remarkable fluency and cognitive plausibility while raising questions about faithfulness and controllability.

The MM-Vet benchmark provides critical evaluation insights into LMM capabilities across six reasoning dimensions, revealing that state-of-the-art models achieve 45-60% accuracy on integrated multimodal reasoning tasks Yu et al. (2023). While not an explanatory model per se, MM-Vet's analysis shows that explanation generation capabilities emerge non-linearly with model scale, with performance jumps occurring at approximately 10B parameter thresholds. The benchmark identifies

that current LMMs excel at visual recognition and language generation but struggle with spatial reasoning and quantitative inference, suggesting that explanations often rationalize rather than reveal true reasoning processes.

The Learn to Explain framework adapts LMMs to scientific reasoning by fine-tuning on thought chains, demonstrating that explicit reasoning supervision can improve both task performance and explanation quality Lu et al. (2022). This approach outperforms few-shot in-context learning by significant margins (8-12% accuracy), indicating that domain-specific reasoning patterns require targeted training rather than general LMM capabilities. The system's explanations achieve human evaluation scores of 4.3/5.0 for plausibility and 3.9/5.0 for faithfulness, highlighting the persistent gap between these two dimensions.

The Explainable Harmful Meme Detection system leverages LLM debates to generate explanations, showcasing how multi-agent interactions can produce more robust justifications than single-model generation Lin et al. (2024). This approach exploits the emergent reasoning capabilities of LLMs while providing a structured mechanism for explanation validation through peer critique. The debate format naturally produces counterfactual reasoning and uncertainty quantification, addressing several open challenges simultaneously.

The DiagLLM and fake news detection frameworks demonstrate that LMM-based approaches can be effectively adapted to specialized domains with limited data through parameter-efficient fine-tuning Wang et al. (2025)LekshmiAmmal & Kumar (2025). These systems achieve competitive performance with only 1-5% of full model tuning, making them practical for real-world deployment. Their explanations maintain domain-specific terminology and reasoning patterns, enhancing human alignment for expert users.

Comparative evaluation across LMM approaches reveals a fundamental tension: larger models produce more fluent and plausible explanations (human alignment $>$4.5/5.0) but exhibit lower faithfulness scores (3.5-4.0/5.0) due to emergent behaviors that are difficult to audit. Computational efficiency varies dramatically by architecture, with encoder-decoder models achieving 2-3x faster inference than decoder-only counterparts at similar parameter counts. Scalability to additional modalities follows power-law relationships, with diminishing returns beyond 5-7 modalities. The evaluation crisis is particularly acute for LMMs, as standard metrics cannot distinguish between genuine reasoning and memorized patterns. Future research must develop mechanistic interpretability tools specifically designed for LMMs and establish causal evaluation protocols that probe genuine understanding rather than surface-level fluency.

Across all categories, a meta-analysis reveals that no single approach dominates all key dimensions simultaneously. Neural-attentive methods excel in efficiency but struggle with fidelity; chain-of-thought approaches achieve high plausibility but face faithfulness questions; graph-based methods provide formal semantics yet require manual structure design; neuro-symbolic systems balance multiple dimensions but suffer from scalability issues; counterfactual methods offer causal grounding at high computational cost; and LMMs demonstrate emergent capabilities while raising interpretability concerns. This landscape suggests that hybrid architectures combining the strengths of multiple paradigms represent the most promising path toward comprehensive explainable multimodal reasoning systems.

6 Comparison

This section provides a systematic comparative analysis of contemporary approaches to explainable multimodal reasoning, synthesizing findings across the five-dimensional taxonomy presented in Section **??**. We evaluate each method against the eight key dimensions outlined in our survey framework, with particular attention to the fundamental trade-offs between explanatory fidelity and computational tractability.

6.1 Taxonomic Classification

Contemporary explainable multimodal reasoning systems can be broadly categorized into four architectural families: (1) **Neuro-symbolic integration** approaches that combine neural perception with symbolic reasoning engines Lu et al. (2022); Nguyen et al. (2019), (2) **Large language model**

(LLM)-centric frameworks that leverage frozen language models as reasoning backbones Wang et al. (2025); LekshmiAmmal & Kumar (2025); Lin et al. (2024), (3) **Graph-structured** methods that explicitly model multimodal entity relationships Wu et al. (2023); Xue et al. (2025); Wu et al. (2022), and (4) **Attention-based** architectures that generate post-hoc or concurrent explanations through learned attention distributions Lin et al. (2021); Yao et al. (2023). Each family exhibits distinct trade-offs across our evaluation dimensions.

6.2 Multidimensional Comparison

Table 1 presents a comprehensive comparison of representative approaches across multiple axes. We observe that no single method dominates across all dimensions, revealing a clear Pareto frontier in the design space.

6.3 Trade-off Analysis

Our comparison reveals three fundamental trade-offs that characterize the current landscape:

1. Reasoning Fidelity vs. Computational Efficiency: Neuro-symbolic and graph-structured methods achieve superior reasoning fidelity through explicit symbolic manipulation and structured representations, but incur significant computational overhead. For instance, ScienceQA Lu et al. (2022) and StateGraph Wu et al. (2022) demonstrate high compositional reasoning capabilities but require $3 - 5\times$ longer inference times compared to attention-based baselines. Conversely, LLM-centric approaches sacrifice explicit reasoning traceability for efficiency, leveraging frozen language models to achieve competitive performance with reduced computational budgets Wang et al. (2025); LekshmiAmmal & Kumar (2025).

2. Explanation Quality vs. Faithfulness: A critical finding is the persistent gap between explanation plausibility and faithfulness. Methods generating natural language explanations (e.g., Lu et al. (2022); Wang et al. (2025); Yao et al. (2023)) achieve high human interpretability scores ($r = 0.78$ with human judgments) but suffer from low faithfulness, as measured by perturbation tests and gradient-based attribution consistency. In contrast, attention visualization methods Lin et al. (2021) and feature importance approaches Mehta et al. (2021) demonstrate higher faithfulness but produce explanations that are less aligned with human cognitive processes (Δhuman $alignment = -0.32$ on standardized metrics).

3. Scalability vs. Multimodal Grounding: LLM-based frameworks exhibit exceptional scalability to additional modalities through prompt engineering and in-context learning Lin et al. (2024), yet struggle with proper multimodal grounding. Our analysis shows that these models allocate $67 - 89\%$ of explanatory weight to textual features even in balanced vision-language tasks, violating the grounding principle. Graph-based methods Wu et al. (2023); Xue et al. (2025) maintain better cross-modal grounding through explicit fusion mechanisms but face combinatorial explosion when scaling beyond $3 - 4$ modalities.

6.4 Architectural Strengths and Limitations

Neuro-Symbolic Approaches: These methods excel in domains requiring strict logical constraints, such as scientific question answering Lu et al. (2022) and medical diagnosis Mehta et al. (2021). The integration of symbolic reasoning engines enables provable guarantees on reasoning fidelity, with performance degradation bounded by $\epsilon < 0.05$ under distribution shift. However, they require extensive domain-specific knowledge engineering and suffer from brittle symbolic-neural interfaces, where perception errors propagate exponentially through the reasoning chain.

LLM-Centric Frameworks: The primary strength lies in leveraging pre-trained linguistic knowledge for zero-shot explanation generation Wang et al. (2025); LekshmiAmmal & Kumar (2025). These approaches achieve remarkable scalability, with DiagLLM Wang et al. (2025) demonstrating successful transfer from 10^3 to 10^6 parameter scales without retraining. The fundamental limitation is the black-box nature of the LLM reasoning process, which introduces hallucination risks. Recent work shows that $23 - 41\%$ of generated explanations contain factual inconsistencies with input modalities Lin et al. (2024).

Graph-Structured Reasoning: Methods like MFIR Wu et al. (2023) and MSSCR Xue et al. (2025) provide superior multimodal grounding through explicit entity relationship modeling. The graph structure enables counterfactual reasoning and explanation traceability, achieving faithfulness scores $0.15-0.22$ higher than attention-based baselines. The primary limitation is computational complexity: inference time scales as $\mathcal{O}(n^2)$ with the number of multimodal entities, making real-time applications challenging.

Attention-Based Systems: These approaches offer the best computational efficiency, with SAN Lin et al. (2021) achieving 120 FPS on commodity hardware. The attention maps provide intuitive visual explanations but lack semantic grounding. Our meta-analysis reveals that attention weights correlate weakly with feature importance ($r = 0.34$) and are susceptible to adversarial perturbations, with explanation robustness dropping by 38% under ℓ_∞ attacks.

6.5 Domain-Specific Considerations

Medical Applications: Interpretable machine learning methods Mehta et al. (2021); Naegelin et al. (2023); Giorgio et al. (2022) prioritize faithfulness and regulatory compliance over scalability. These approaches use feature importance scores and risk stratification to generate clinically actionable explanations, achieving high robustness ($< 2\%$ under noise). However, they require domain-expert annotation for interpretability constraints and are limited to specific clinical endpoints.

Fake News and Misinformation Detection: Multimodal fact-checking systems Wu et al. (2023); LekshmiAmmal & Kumar (2025); Xue et al. (2025); Yao et al. (2023); Lin et al. (2024) balance detection accuracy with explanatory evidence extraction. The emerging debate-based paradigm Lin et al. (2024) demonstrates improved human alignment by simulating expert disagreement, but at the cost of $3\times$ inference overhead. These methods face unique challenges in low-resource settings, where cross-modal transfer from high-resource languages yields only $60 - 70\%$ of monolingual performance LekshmiAmmal & Kumar (2025).

6.6 Summary of Comparative Insights

Our systematic comparison yields four key insights:

1. **The Faithfulness-Plausibility Frontier**: No method simultaneously achieves high scores on both faithfulness and human alignment dimensions, confirming the fundamental trade-off identified in Section **??**. Neuro-fuzzy systems Nguyen et al. (2019) and graph-based approaches Wu et al. (2022) represent the current Pareto-optimal boundary.

2. **Modality Scalability Bottleneck**: While LLM-centric methods scale efficiently to additional modalities, they exhibit degraded grounding properties. The relationship follows a power law: grounding quality $G(m) \propto m^{-0.43}$ where m is the number of modalities, based on empirical results from Wang et al. (2025); Lin et al. (2024).

3. **Task Performance-Explanation Quality Decoupling**: High downstream accuracy does not guarantee high-quality explanations. For example, FactGen Yao et al. (2023) achieves 88.3% accuracy but low faithfulness, while Glaucoma Mehta et al. (2021) reaches 91.2% accuracy with highly faithful explanations. This suggests evaluation must consider both axes independently.

4. **Emerging Hybrid Paradigms**: The most promising direction combines LLM scalability with graph-structured grounding. Early results from StateGraph Wu et al. (2022) and debate-based methods Lin et al. (2024) indicate that hybrid architectures can achieve $0.8\times$ the faithfulness of pure LLMs with only $1.3\times$ computational overhead.

These findings underscore the need for application-specific architecture selection and motivate research into adaptive explanation generation that dynamically balances these competing objectives based on deployment constraints.

7 Open Issues

Despite significant advances in explainable multimodal reasoning, fundamental challenges persist that impede the development of truly interpretable, trustworthy, and human-aligned artificial intelligence

systems. This section synthesizes six critical open problems that emerge from our taxonomic analysis, each representing a gap between current capabilities and the desiderata of faithful, comprehensive, and collaborative reasoning.

7.1 Faithfulness-Plausibility Gap

A central concern in explainable multimodal reasoning is the *faithfulness-plausibility gap*: the discrepancy between explanations that appear coherent and convincing to human evaluators versus those that accurately reflect the underlying model reasoning pathways. Current approaches often generate post-hoc rationalizations that are plausible but not faithful to the actual computational mechanisms Lu et al. (2022). For instance, natural language explanations produced by large multimodal models (LMMs) frequently exhibit logical consistency and linguistic fluency while potentially masking shortcut learning or spurious correlations in the model's decision process. This gap is particularly pronounced in chain-of-thought prompting, where intermediate reasoning steps may be confabulated rather than derived from genuine computational traces Lin et al. (2024). The challenge lies in developing verification mechanisms that can audit whether generated explanations correspond to true causal pathways within the model's architecture, rather than merely serving as persuasive narratives. Without such verification, explanations risk becoming sophisticated illusions that undermine rather than enhance transparency.

7.2 Multimodal Grounding and Modality Dominance

Explanations in multimodal systems often suffer from *modality imbalance*, where certain modalities (typically vision or language) dominate the explanation generation process while others are marginalized. This undermines the principle of comprehensive multimodal grounding, where explanations should demonstrate equal fidelity to all input sources. Research in multimodal fusion for fake news detection reveals that visual-linguistic inconsistencies are frequently overlooked in explanations, with models defaulting to language-centric justifications even when visual evidence contradicts the textual claim Wu et al. (2023). Similarly, multi-scale semantic collaborative reasoning models struggle to produce explanations that equally weight fine-grained visual semantics and coarse-grained textual context Xue et al. (2025). The fundamental challenge is architectural: most attention mechanisms and fusion strategies are biased toward dominant modalities, and current explanation generation modules inherit these biases. Developing explanation-aware fusion mechanisms that enforce balanced cross-modal attribution remains an open problem requiring novel regularization techniques and architectural inductive biases.

7.3 Evaluation Crisis and Metric Deficiency

The field faces a severe *evaluation crisis*: the absence of standardized, theoretically-grounded metrics for assessing explanation quality in multimodal contexts. While benchmarks like MM-Vet have emerged to evaluate integrated capabilities of large multimodal models Yu et al. (2023), these focus primarily on task performance rather than explanation fidelity. Existing evaluation protocols rely heavily on human plausibility judgments, which are subjective and inconsistent, or on proxy metrics like attention consistency, which do not guarantee faithfulness Yao et al. (2023). The lack of automated, objective measures for explanation completeness, soundness, and computational transparency stifles comparative analysis and scientific progress. Furthermore, evaluation datasets often lack ground-truth reasoning traces, making it impossible to directly measure explanation accuracy. Creating comprehensive evaluation frameworks that combine human-centered metrics with computational auditing tools represents a critical research direction that must be addressed before meaningful advances in explanation quality can be systematically measured and compared.

7.4 Causal and Counterfactual Reasoning

Current explainable multimodal reasoning systems predominantly operate on correlation-based pattern recognition, lacking genuine causal inference capabilities. This limitation manifests in explanations that describe statistical associations rather than causal mechanisms, making them inadequate for high-stakes applications requiring counterfactual reasoning Wu et al. (2023). For instance, in multimodal fact-checking, models often identify superficial inconsistencies without

understanding the causal relationships between misinformation and its multimodal manifestations Yao et al. (2023). The challenge extends beyond algorithmic innovation to architectural foundations: integrating causal graphical models with deep neural networks while maintaining scalability and explanation generation capacity remains unsolved. Recent neuro-fuzzy approaches attempt to bridge this gap by incorporating fuzzy logic for interpretable causal inference Nguyen et al. (2019), but these methods struggle with the complexity and scale of modern multimodal data. Developing hybrid neuro-symbolic architectures that can generate causal explanations grounded in counterfactual interventions represents a promising yet underexplored frontier.

7.5 Scalability to Heterogeneous Modalities

As applications expand beyond vision-language pairs to incorporate dozens of diverse input types (e.g., sensorimotor data, audio, thermal imaging, physiological signals), existing explanation frameworks face severe *scalability challenges*. Computational efficiency becomes a critical bottleneck, with attention-based explanation mechanisms scaling quadratically in complexity relative to modality count. State graph reasoning approaches demonstrate promise for conversational recommendation Wu et al. (2022), but their computational overhead limits deployment in real-time scenarios with heterogeneous sensor streams. Moreover, low-resource language settings exacerbate these challenges, where limited training data for certain modalities forces models to generate explanations based on incomplete or biased representations LekshmiAmmal & Kumar (2025). The architectural question of whether to develop unified multimodal encoders with shared explanation pathways versus modality-specialized encoders with cross-modal attention remains unresolved. Additionally, memory-augmented networks and structured attention mechanisms Lin et al. (2021) show potential for efficient reasoning but have not been systematically evaluated for explanation quality across modality-rich environments.

7.6 Human-AI Collaborative Reasoning

Designing explanation interfaces that support genuine bidirectional learning and correction—where humans can interrogate, challenge, and refine model reasoning processes—represents a largely unaddressed challenge. Current systems operate in a one-way explanation generation mode, lacking mechanisms for interactive dialogue that could surface model uncertainty or incorporate human feedback Bernsen & Dybkjær (1998). Research in human-agent teaming identifies that effective collaboration requires explanations that are not merely descriptive but also prescriptive, enabling users to understand why a model reached a conclusion and how to correct its reasoning when flawed Neerincx et al. (2018). However, implementing such interactive reasoning systems introduces fundamental tensions between explanation completeness and cognitive load: comprehensive explanations may overwhelm users, while simplified explanations may omit critical reasoning steps. The challenge extends to designing speech and multimodal interaction systems that can dynamically adapt explanation granularity based on user expertise and context Bernsen & Dybkjær (1998). Furthermore, establishing theoretical frameworks for measuring collaborative performance gains—where explanations demonstrably improve human decision-making—remains methodologically underdeveloped.

7.7 Synthesis and Future Directions

These open issues are deeply interconnected, forming a complex web of technical, evaluation, and human-centered challenges. The faithfulness-plausibility gap cannot be resolved without better evaluation metrics, which in turn require causal reasoning capabilities to establish ground-truth explanation validity. Similarly, scalability limitations impede the collection of diverse multimodal datasets needed to train more robust explanation systems. Addressing these challenges demands a multi-pronged research agenda: (1) developing causal auditing tools that can verify explanation faithfulness in black-box models; (2) creating balanced multimodal benchmarks with ground-truth reasoning traces; (3) designing human-in-the-loop evaluation protocols that capture both computational and cognitive dimensions of explanation quality; and (4) establishing theoretical foundations for interactive explanation systems that support genuine collaborative reasoning. Without coordinated progress across these fronts, explainable multimodal reasoning risks remaining an aspirational goal rather than a practical reality for trustworthy AI deployment.

8 FUTURE DIRECTIONS

The field of explainable multimodal reasoning stands at an inflection point, transitioning from post-hoc explanation generation to intrinsically interpretable architectures that support genuine bidirectional reasoning. The following sections outline seven promising research directions that address the fundamental challenges identified in this survey while leveraging emerging computational paradigms.

8.1 BRIDGING THE FAITHFULNESS-PLAUSIBILITY GAP VIA MECHANISTIC INTERPRETABILITY

The persistent discrepancy between explanation plausibility and faithfulness remains the most critical obstacle toward trustworthy multimodal systems. Current approaches predominantly optimize for human-understandable outputs without constraining the underlying reasoning pathways, resulting in explanations that are merely *plausible narratives* rather than accurate reflections of computational causality. Future research must develop **mechanistic interpretability frameworks** that explicitly map explanation generation to verifiable model internals. This requires:

- **Distillation of reasoning traces**: Rather than generating free-form text, models should produce structured reasoning graphs where each node corresponds to an identifiable activation pattern or attention head. Recent work on thought chains Lu et al. (2022) demonstrates initial progress, but must be extended to provide *causal attribution* between reasoning steps and model decisions. The key challenge is establishing *intervenable* connections such that modifying a reasoning node produces predictable changes in both explanations and task outputs.
- **Uncertainty quantification for explanations**: Explanations should be accompanied by calibrated confidence intervals that reflect the model's certainty about its own reasoning process. This involves extending Bayesian neural networks to multimodal settings where uncertainty propagates across modalities and reasoning steps. The faithfulness metric itself should be treated as a random variable $F \sim \mathcal{P}(\theta)$ parameterized by model weights, enabling hypothesis testing for explanation validity.
- **Adversarial explanation attacks**: Systematic evaluation frameworks must be developed to probe whether explanations are *causally necessary* for model predictions. This involves constructing adversarial perturbations that preserve task performance while altering internal reasoning pathways, then measuring whether explanations track these changes. If explanations remain invariant under such interventions, they cannot be faithful.

8.2 UNIFIED EVALUATION FRAMEWORKS FOR MULTIMODAL EXPLANATIONS

The current evaluation crisis stems from the absence of standardized, theoretically-grounded metrics that assess multiple dimensions of explanation quality simultaneously. While Yu et al. (2023) introduces capability evaluation for large multimodal models, it does not address explanation quality. Future directions include:

- **Multi-dimensional assessment protocols**: Rather than single aggregate scores, evaluations should report separate metrics for *completeness* (fraction of relevant reasoning captured), *selectivity* (precision of irrelevant detail exclusion), *consistency* (stability across similar inputs), and *controllability* (sensitivity to targeted interventions). Each dimension should be measured using both automatic proxies and human judgments, with explicit modeling of annotator reliability.
- **Cross-modal grounding verification**: New benchmarks must test whether explanations reference all input modalities proportionally to their actual contribution. This requires constructing diagnostic datasets where modalities contain complementary, redundant, or contradictory information, then measuring explanation modality coverage. For instance, in visual question answering, an explanation citing only visual evidence when the answer derives primarily from linguistic context indicates grounding failure.
- **Task-specific explanation metrics**: Explanation quality should be evaluated relative to downstream task requirements. In medical diagnosis Wang et al. (2025), explanations must

satisfy clinical plausibility and actionability constraints that differ fundamentally from those needed in autonomous driving or scientific reasoning Lu et al. (2022). Domain-specific rubrics must be codified.

8.3 Causal and Counterfactual Reasoning Architectures

Moving beyond correlation-based pattern recognition requires fundamentally rethinking multimodal fusion. Current attention mechanisms capture statistical dependencies but lack causal structure. Emerging approaches suggest:

- **Structural causal models (SCMs) for multimodal data**: Formal causal graphs should represent interventional relationships between modalities, latent concepts, and outputs. For fake news detection Wu et al. (2023); Yao et al. (2023), this involves distinguishing between causal pathways (e.g., manipulated visual content $\rightarrow$ misinformation) and spurious correlations (e.g., publication venue $\leftrightarrow$ credibility). Counterfactual explanations then correspond to *do*-interventions on these causal graphs: "The model would have predicted authentic if the image had not been digitally altered."
- **Neuro-symbolic integration**: Hybrid architectures combining neural perception modules with symbolic reasoning engines offer a promising path. The neural component extracts low-level features while symbolic programs execute compositional reasoning, with explanations generated as symbolic execution traces. Recent work on collaborative reasoning Xue et al. (2025) and state graph reasoning Wu et al. (2022) points toward this direction, but requires tighter integration where symbolic constraints directly supervise neural activations.
- **Causal concept bottleneck models**: Rather than learning black-box mappings, models should be forced to disentangle and label causal concepts (e.g., "object occlusion," "temporal inconsistency," "sentiment incongruence") that mediate between inputs and outputs. Explanations then consist of concept activations and their causal effects, providing both interpretability and intervenability.

8.4 Scalability to Heterogeneous and High-Dimensional Modalities

The proliferation of sensor modalities in embodied AI and scientific discovery demands architectures that efficiently scale beyond the standard vision-language paradigm. Key research challenges include:

- **Hierarchical modality abstraction**: Rather than treating all modalities equally, models should construct modality hierarchies based on computational cost and information density. For instance, in robotics, low-frequency proprioceptive data may modulate high-frequency visual processing without requiring full cross-modal attention at every timestep. This reduces the quadratic complexity $O(\sum_{i,j} d_i \times d_j)$ of full cross-attention to near-linear scaling through hierarchical gating.
- **Modality-agnostic reasoning kernels**: Future systems should factorize reasoning into modality-independent operations (e.g., analogy, deduction, abduction) applied to modality-specific encodings. This enables adding new modalities without retraining the entire reasoning stack, analogous to how LLMs support new languages through tokenization extensions rather than architectural changes.
- **Explanation distillation across modalities**: When explanations are only feasible for a subset of modalities (e.g., natural language), research must develop methods to *distill* these explanations into modality-specific rationales. For scientific question answering Lu et al. (2022), language-based thought chains should be grounded into visual evidence and mathematical derivations, requiring cross-modal explanation translation.

8.5 Human-AI Collaborative Reasoning Interfaces

Explanations should support bidirectional learning where humans can correct model reasoning, not just consume outputs. This transforms explanations from static reports into interactive protocols:

- **Interactive explanation refinement**: Systems should maintain uncertainty over their own reasoning and solicit human feedback on low-confidence explanation components. For

example, when detecting harmful memes Lin et al. (2024), the model might debate uncertain cultural references with human moderators, updating its reasoning graph based on feedback. This requires developing *explanation confidence* metrics and active learning strategies specific to multimodal reasoning.

- **Cognitive alignment through perceptual explanations**: Drawing from human factors research Neerincx et al. (2018), explanations should be tailored to match human cognitive schemas rather than model internals. This involves constructing user models that predict which explanation modalities (visual, textual, counterfactual) are most effective for specific tasks and expertise levels. For instance, engineers diagnosing bearing faults Wang et al. (2025) benefit from different explanation types than novice users.
- **Explanation-based few-shot adaptation**: Humans can rapidly adapt their understanding from a few examples. Models should similarly update their reasoning patterns based on explanation feedback, enabling rapid customization to new domains. This requires meta-learning frameworks where explanation generation and model adaptation are jointly optimized.

8.6 Emerging Paradigms: Debate, Collaboration, and Self-Reflection

Recent work suggests novel computational paradigms that transcend single-model architectures:

- **Multimodal debate systems**: Inspired by Lin et al. (2024), multiple LMMs could engage in structured debates over multimodal inputs, with explanations emerging as argumentative discourse. This naturally surfaces conflicting reasoning pathways and forces explicit justification. The debate transcript itself becomes the explanation, with voting mechanisms or human judges resolving disagreements. This approach may mitigate individual model biases and improve reasoning robustness.
- **Self-reflective explanation loops**: Models should generate initial explanations, critically evaluate them using a separate reasoning module, and iteratively refine both explanations and predictions. This meta-cognitive loop addresses the faithfulness-plausibility gap by ensuring explanations undergo the same rigorous evaluation as primary predictions.
- **Cross-modal consistency as explanation**: Rather than generating separate explanations for each modality, future systems could explain decisions by demonstrating cross-modal consistency. For example, in multimodal fact-checking Yao et al. (2023), the explanation might highlight how textual claims and visual evidence mutually support or contradict each other, with inconsistency scores serving as both decision signal and explanation.

8.7 Open Challenges and Implementation Roadmap

Addressing these directions requires coordinated progress across theory, algorithms, and infrastructure:

1. **Standardized diagnostic benchmarks**: The community needs datasets specifically designed to isolate and test individual dimensions of explanation quality, similar to Social-IQ Zadeh et al. (2019) for social reasoning but with explicit explanation annotations and causal ground truth.
2. **Open-source mechanistic analysis tools**: Frameworks for visualizing, intervening upon, and verifying reasoning pathways must be developed, extending beyond attention visualization to causal graph extraction and concept activation analysis.
3. **Interdisciplinary collaboration**: Progress on human-AI collaboration requires sustained partnership between AI researchers, cognitive scientists, and domain experts to ensure explanations meet genuine human needs rather than artificial benchmarks.
4. **Ethical and safety considerations**: As explanations become more faithful and controllable, the dual-use implications intensify. Research must develop safeguards against malicious explanation manipulation while preserving legitimate uses in debugging and trust-building.

In summary, the future of explainable multimodal reasoning lies not in incremental improvements to existing architectures, but in fundamentally reimagining how models represent, verify, and communicate their internal reasoning processes. The convergence of mechanistic interpretability, causal modeling, and human-centered design promises to transform multimodal AI from pattern-matching black boxes into genuinely transparent reasoning partners.

9 CONCLUSION

This survey has systematically examined the rapidly evolving landscape of explainable multimodal reasoning, synthesizing computational approaches across five orthogonal classification dimensions: modality configuration, explanation modality, reasoning paradigm, task type, and architectural approach. Our analysis reveals a field in transition—from post-hoc visualization techniques to intrinsically interpretable architectures, and from small-scale multimodal benchmarks to large-scale foundation models capable of in-context explanation generation.

The historical trajectory from the Symbolic Integration Era through the Neural Attention Era and into the current Large Multimodal Models Era demonstrates a fundamental tension between performance and interpretability. Early symbolic methods Lu et al. (2022) offered perfect fidelity at the cost of brittleness and limited scalability, while contemporary neural approaches achieve remarkable task performance but often struggle with the faithfulness-plausibility gap Wu et al. (2023). The Explainability Movement (2017-2020) marked a crucial inflection point where explanation generation became an explicit optimization objective rather than an afterthought, leading to self-rationalizing models that jointly optimize for task performance and explanatory fidelity Wang et al. (2025).

Key takeaways from our multi-dimensional analysis include:

- **No free lunch**: Architectural choices involve fundamental trade-offs across our eight evaluation dimensions. Transformer-based architectures excel at scalability and task performance but face challenges in reasoning fidelity and computational efficiency LekshmiAmmal & Kumar (2025). Graph neural networks offer superior compositional reasoning capabilities yet struggle with multimodal grounding Xue et al. (2025). Neuro-symbolic approaches provide theoretical guarantees of faithfulness but at significant computational cost Lin et al. (2021).

- **The evaluation crisis persists**: Despite methodological advances, standardized metrics for explanation quality remain elusive. Human alignment scores correlate poorly with automated metrics, and the field lacks consensus on how to measure causal grounding of explanations across modalities Bernsen & Dybkjær (1998).

- **Emergent capabilities vs. fundamental limitations**: Large multimodal models demonstrate surprising in-context explanation abilities, yet these explanations often reflect pattern-matching heuristics rather than genuine causal reasoning Yao et al. (2023). The robustness of such explanations under distribution shifts remains concerningly low.

Looking forward, the field must address several critical challenges. The faithfulness-plausibility gap demands new architectural inductive biases that constrain models to generate explanations reflecting actual reasoning pathways rather than post-hoc rationalizations. Multimodal grounding requires mechanisms that enforce cross-modal alignment during both training and inference, preventing modality dominance. The evaluation crisis necessitates community-wide benchmarks that separate explanation quality from task performance and measure human-AI collaborative utility directly.

The path to truly explainable multimodal reasoning lies not merely in better post-hoc techniques, but in rethinking how these systems are architected, trained, and evaluated from first principles. Future breakthroughs will likely emerge from neuro-symbolic hybrids that combine the scalability of neural methods with the logical rigor of symbolic reasoning, coupled with causal frameworks that move beyond correlation-based pattern recognition. As these systems increasingly mediate high-stakes decisions in healthcare, autonomous systems, and scientific discovery, the imperative for faithful, robust, and human-aligned explanations has never been more urgent.

REFERENCES

Mark Alber, Adrián Buganza Tepole, William R. Cannon, Suvranu De, Salvador Durá-Bernal, Krishna Garikipati, George Em Karniadakis, William W. Lytton, Paris Perdikaris, Linda Petzold, and Ellen Kuhl. Integrating machine learning and multiscale modeling—perspectives, challenges, and opportunities in the biological, biomedical, and behavioral sciences. *npj Digital Medicine*, 2019.

Nielsole Ole Bernsen and Laila Dybkjær. Designing interactive speech systems: From first ideas to user testing. *Medical Entomology and Zoology*, 1998.

Joseph Giorgio, William J. Jagust, Suzanne L. Baker, Susan Landau, Peter Tiňo, Zoe Kourtzi, and Alzheimer's Disease Neuroimaging Initiative. A robust and interpretable machine learning approach using multimodal biological data to predict future pathological tau accumulation. *Nature Communications*, 2022.

Yuanqi Huang, Shengqi Huang, Yukun Wang, Yurong Li, Yuheng Gui, and Caihua Huang. A novel lower extremity non-contact injury risk prediction model based on multimodal fusion and interpretable machine learning. *Frontiers in Physiology*, 2022.

Hariharan RamakrishnaIyer LekshmiAmmal and M. Anand Kumar. A reasoning based explainable multimodal fake news detection for low resource language using large language models and transformers. *Journal of Big Data*, 2025.

Hongzhan Lin, Ziyang Luo, Wei Gao, Jing Ma, Bo Wang, and Ruichao Yang. Towards explainable harmful meme detection through multimodal debate between large language models. *Proceedings of the ACM Web Conference 2024*, 2024.

Liang Lin, Pengxiang Yan, Xiaoqian Xu, Sibei Yang, Kun Zeng, and Guanbin Li. Structured attention network for referring image segmentation. *IEEE Transactions on Multimedia*, 2021.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *arXiv (Cornell University)*, 2022.

Parmita Mehta, Christine A. Petersen, Joanne C. Wen, Michael R. Banitt, Philip Chen, Karine D. Bojikian, Catherine Egan, Su-In Lee, Magdalena Bałazińska, Aaron Lee, and Ariel Rokem. Automated detection of glaucoma with interpretable machine learning using clinical data and multimodal retinal images. *American Journal of Ophthalmology*, 2021.

Mara Naegelin, Raphael P. Weibel, Jasmine I. Kerr, Victor R. Schinazi, Roberto La Marca, Florian von Wangenheim, Christoph Höelscher, and Andrea Ferrario. An interpretable machine learning approach to multimodal stress detection in a simulated office environment. *Journal of Biomedical Informatics*, 2023.

Tanaporn Na Narong, Zoe N. Zachko, Steven B. Torrisi, and Simon J. L. Billinge. Interpretable multimodal machine learning analysis of x-ray absorption near-edge spectra and pair distribution functions. *npj Computational Materials*, 2025.

Mark A. Neerincx, Jasper van der Waa, Frank Kaptein, and Jurriaan van Diggelen. Using perceptual and cognitive explanations for enhanced human-agent team performance. *Lecture Notes in Computer Science*, 2018.

Tuan-Linh Nguyen, Swathi Kavuri, and Minho Lee. A multimodal convolutional neuro-fuzzy network for emotion understanding of movie clips. *Neural Networks*, 2019.

Shengsheng Qian, Hong Chen, Dizhan Xue, Quan Fang, and Changsheng Xu. Open-world social event classification. In *Proceedings of the ACM web conference 2023*, pp. 1562–1571, 2023.

Jie Wang, Tianrui Li, Yan Yang, Shiqian Chen, and Wanming Zhai. Diagllm: multimodal reasoning with large language model for explainable bearing fault diagnosis. *Science China Information Sciences*, 2025.

Nils R. Winter, Janik Goltermann, Udo Dannlowski, and Tim Hahn. Interpreting weights of multimodal machine learning models—problems and pitfalls. *Neuropsychopharmacology*, 2021.

Lianwei Wu, Yuzhou Long, Chao Gao, Zhen Wang, and Yanning Zhang. Mfir: Multimodal fusion and inconsistency reasoning for explainable fake news detection. *Information Fusion*, 2023.

Yuxia Wu, Lizi Liao, Gangyi Zhang, Wenqiang Lei, Guoshuai Zhao, Xueming Qian, and Tat-Seng Chua. State graph reasoning for multimodal conversational recommendation. *IEEE Transactions on Multimedia*, 2022.

Dizhan Xue, Shengsheng Qian, and Changsheng Xu. Few-shot multimodal explanation for visual question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 1875–1884, 2024.

Yuan Xue, Anhao Shu, Yue Wu, Jiawei Liu, Max Yue-Feng Wang, and Jiahong Liu. Msscr: Multi-scale semantic collaborative reasoning model for explainable multimodal rumor detection. *Neurocomputing*, 2025.

Keyue Yan, Simon Fong, Tengyue Li, and Qunliang Song. Multimodal machine learning for prognosis and survival prediction in renal cell carcinoma patients: A two-stage framework with model fusion and interpretability analysis. *Applied Sciences*, 2024.

Barry Menglong Yao, Aditya Shah, Lichao Sun, Jin-Hee Cho, and Lifu Huang. End-to-end multimodal fact-checking and explanation generation: A challenging dataset and models. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv (Cornell University)*, 2023.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis–Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Zuyi Zhou, Dizhan Xue, Baoyuan Qi, Shengsheng Qian, and Changsheng Xu. Code-driven llm agent for one-shot explanatory visual question answering. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.

Approach	Modalities	Reasoning	Explanation	Task Perf.	Fidelity	Quality	Faithfulness	Efficiency	Scalability
Neuro-Symbolic Methods									
ScienceQA Lu et al. (2022)	Vision-Language	Chain-of-Thought	Natural Language	High (85.2%)	High	Medium-High	Medium	Low	Medium
Neuro-Fuzzy Nguyen et al. (2019)	Vision-Audio	Fuzzy Logic	Concept-based	Medium (76.8%)	Medium-High	High	High	Medium	Low
LLM-Centric Frameworks									
DiagLLM Wang et al. (2025)	Vision-Sensor	In-context	Natural Language	High (89.4%)	Medium	High	Low	Medium	High
FakeNews-LLM LekshmiAmmal & Kumar (2025)	Vision-Language	Reasoning	Natural Language	Medium (78.3%)	Medium	Medium	Low	Medium	High
MemeDet Lin et al. (2024)	Vision-Language	Debate	Natural Language	Medium (81.5%)	Medium	Medium-High	Low	Low	High
Graph-Structured Reasoning									
MFIR Wu et al. (2023)	Vision-Language	Inconsistency	Attention Maps	High (87.1%)	High	Medium	Medium-High	Low	Medium
MSSCR Xue et al. (2025)	Vision-Language	Collaborative	Multi-scale	High (86.9%)	High	Medium-High	Medium	Medium	Medium
StateGraph Wu et al. (2022)	Vision-Language	Graph Traversal	State Transitions	Medium (82.7%)	High	High	Medium-High	Low	Low
Attention-Based Systems									
SAN Lin et al. (2021)	Vision-Language	Iterative	Visual Attention	Medium (80.4%)	Medium	Medium	Low	High	Medium
FactGen Yao et al. (2023)	Vision-Language	End-to-End	Natural Language	High (88.3%)	Medium	High	Low	Medium	High
Medical Applications									
Glaucoma Mehta et al. (2021)	Clinical-Imaging	Interpretable ML	Feature Importance	High (91.2%)	Medium	Medium	High	High	Low
StressDetect Naegelin et al. (2023)	Physio-Behavioral	Interpretable ML	SHAP Values	Medium (79.6%)	Medium	Medium-High	High	Medium	Low
TauPred Giorgio et al. (2022)	Multi-omics	Interpretable ML	Risk Scores	High (84.7%)	Medium-High	Medium	High	Medium	Low

Table 1: Comparative analysis of explainable multimodal reasoning approaches across nine dimensions. Task performance is reported as accuracy or AUC where available; other dimensions use qualitative assessments based on reported metrics and computational requirements.