International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 782-791 www.ijaem.net ISSN: 2395-5252

VisualBERT for Credible Fake News Detection

Botsa Srinivasa Rao
Mtech Research scholar, Anil Neerukonda Institute of Technology & Sciences, Visakhapatnam

Venkata Gayathri Ganivada
Assistant Professor, Anil Neerukonda Institute of Technology & Sciences (ANITS), Visakhapatnam

Chandrasekhar Rao Pakki
Asst. Prof in CSE, N S Raju Institute of Engineering & Technology

Date of Submission: 13-05-2026 Date of Acceptance: 28-05-2026

Abstract
The increasing diffusion of multimodal misinformation has exposed important limitations in existing fake news detection research, which remains primarily focused on predictive classification, multimodal fusion, and post-hoc explainability. Although these approaches improve detection performance, they provide limited explanation regarding how textual and visual evidence becomes interpreted as credible. To understand this, wedevelop Cross-Modal Credibility Formation (CMCF) as a conceptual theory-building framework explaining how multimodal attention transforms heterogeneous evidence into interpretable credibility judgments. Integrating Information Processing Theory, Explainable Artificial Intelligence, and Socio-Technical Systems Theory, the framework conceptualizes credibility as a sequential interpretive process involving evidence congruence, attention coordination, interpretive visibility, evaluative confidence, and detection legitimacy. The study extends multimodal misinformation studies by repositioning VisualBERT as a credibility-oriented interpretive architecture rather than solely a predictive detection system.

Keywords: Multimodal misinformation; Fake news detection; Cross-Modal Credibility Formation; Explainable AI; VisualBERT; Credibility formation; Multimodal learning

I. Introduction

The contemporary information environment has undergone a significant transformation in which misinformation increasingly circulates through integrated combinations of textual and visual content rather than isolated linguistic narratives. Digital communication platforms have accelerated the diffusion of multimodal information formats where images, captions, contextual framing, and narrative sequencing collectively shape how individuals evaluate authenticity and credibility. Consequently, fake news has evolved from a predominantly textual phenomenon into a multimodal phenomenon in which credibility judgments emerge through interactions between visual and semantic signals.

This transformation reflects more than a change in communication format; it represents a shift in the mechanisms through which credibility is constructed within digital environments. Prior research demonstrates that visual information strongly influences credibility evaluation because individuals frequently perceive image-supported information as more authentic and trustworthy, even when visual evidence itself lacks factual verification (Fogg & Tseng, 1999; Metzger et al., 2010). Images reduce interpretive uncertainty, increase perceptual realism, and function as heuristic credibility cues that shape information acceptance. As a result, visual content does not merely supplement textual narratives but actively participates in the construction of believability.

The rise of multimodal misinformation therefore introduces challenges that extend beyond predictive identification and factual verification. Within contemporary digital environments, individuals increasingly evaluate authenticity through integrated assessments of visual plausibility, semantic consistency, and contextual coherence. Manipulated visual framing can reinforce misleading narratives despite weak evidentiary foundations, allowing misinformation to appear credible through coordinated multimodal presentation. Accordingly, multimodal misinformation should be understood not simply as false information but as a process through which heterogeneous signals jointly produce perceived authenticity. Existing misinformation researchlargely influenced by assumptions derived from earlier detection paradigms. Traditional

approaches primarily conceptualize misinformation as a classification problem centered on identifying linguistic inconsistencies and improving predictive accuracy. Although these approaches have significantly advanced detection capability, they remain limited in explaining how credibility judgments emerge through multimodal interpretation.

This limitation reveals an unresolved tension within misinformation studies. Advances in predictive performance have been accompanied by greater opacity in decision-making processes. Detection systems may accurately classify manipulated content while providing limited explanation regarding how textual and visual evidence jointly contributed to credibility judgments. This article conceptualizes this tension as the Credibility–Interpretability Paradox which is describes a condition in which improvements in predictive sophistication do not generate corresponding improvements in interpretability, transparency, or decision legitimacy. Systems may achieve increasingly accurate outcomes while remaining insufficiently capable of explaining how multimodal evidence becomes interpreted as believable or deceptive. Consequently, predictive capability alone cannot explain how credibility is formed within multimodal misinformation environments.

Existing studies addressing misinformation can broadly be organized into three major streams. The first stream focuses on fake news detection and conceptualizes misinformation primarily as a predictive classification problem aimed at improving identification capability (Vosoughi et al., 2018). Although this literature advances predictive effectiveness, credibility is generally treated as a final output rather than an interpretive process.The second stream examines multimodal learning and demonstrates that integrating textual and visual information improves representational quality and predictive performance (Baltrušaitis et al., 2018; Li et al., 2019; Lu et al., 2019). However, multimodal integration is predominantly operationalized as a technical fusion mechanism rather than a process through which meaning and credibility are constructed.The third stream emphasizes explainable artificial intelligence and argues that transparency improves accountability, interpretability, and trust in intelligent systems (Doshi-Velez & Kim, 2017; Adadi& Berrada, 2018; Miller, 2019; Arrieta et al., 2020). Despite these contributions, explainability is frequently treated as a post-hoc explanatory layer attached after prediction rather than an intrinsic component of credibility formation.

All, these streams explain prediction more effectively than credibility formation. Existing literature are limited in explaining how textual and visual evidence becomes jointly interpreted as believable, how multimodal reasoning shapes credibility judgments, and how detection outcomes become socially acceptable within digital environments.So, we introduceCross-Modal Credibility Formation which refers to the process through which multimodal attention transforms heterogeneous textual and visual evidence into interpretable credibility judgments. Unlike approaches that treat credibility as a direct outcome of prediction, CMCF conceptualizes credibility as an interpretive process emerging through multimodal coordination, attention alignment, and evaluative reasoning.

Therefore, weproposes that credibility should be understood not as an automatic consequence of predictive capability but as a sequential process of evidence interpretation. Within this process, multimodal attention facilitates evidence alignment, interpretation stabilizes meaning, and evaluative reasoning contributes to perceived authenticity and legitimacy.

This study has major contributions. It introduces Cross-Modal Credibility Formation as a novel conceptual construct explaining how credibility emerges through multimodal interpretation. It reconceptualizes VisualBERT as a credibility-oriented interpretive architecture rather than solely a multimodal detection system.And it develops a process framework explaining how multimodal attention mechanisms contribute to credibility attribution and interpretable misinformation detection.

This study is positioned as a conceptual theory-building article. Conceptual theorization becomes necessary when existing literature explains related phenomena but lacks an integrated mechanism capable of explaining emerging interpretive processes (Whetten, 1989; MacInnis, 2011). Although multimodal misinformation research has advanced predictive detection, multimodal learning, and explainable AI, existing studies remains limited in explaining how multimodal evidence becomes interpreted as credible within digital environments.Consistent with conceptual theory-building research (Whetten, 1989; MacInnis, 2011), the study contributes by introducing a new interpretive construct, clarifying its dimensions, and organizing these dimensions into a sequential explanatory framework. Accordingly, the article adopts a conceptual orientation focused on process explanation rather than predictive optimization or empirical model testing.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 782-791 www.ijaem.net ISSN: 2395-5252

II. Problematizing Existing Literature: Critical Literature review

Theexpansion of fake news research has generated substantial progress in automated misinformation identification; however, this progress has largely evolved within assumptions that prioritize prediction, integration, and explanation as sufficient conditions for effective misinformation control. Across computational and information systems literature, scholarly advancement is frequently evaluated according to improvements in detection accuracy, multimodal representation capability, and explainability outputs. While these developments have expanded analytical capability, they have simultaneously narrowed theoretical understanding of how credibility itself emerges during information evaluation.

This article argues that current studiesremain constrained by three dominant assumptions: fake news as classification logic, multimodal detection as fusion logic, and explainability as output logic. These assumptions have become increasingly institutionalized within the literature and collectively reinforce a view that misinformation control is fundamentally a technical optimization problem. Although such assumptions improve prediction, they remain insufficient for explaining how heterogeneous evidence becomes interpreted as believable. Consequently, existing studies explains detection outcomes but leaves credibility formation theoretically underdeveloped.

2.1 Fake News as Classification Logic: The Limits of Prediction and Benchmark Dependency

The dominant tradition in fake news research frames misinformation as a classification problem. Therefore, false information is assumed to possess detectable patterns that can be learned and separated from truthful content through increasingly sophisticated predictive systems. Progress is therefore measured through improvements in accuracy, precision, recall, F1 score, and comparative benchmark performance.This orientation has generated major advances in automated misinformation detection and significantly improved predictive effectiveness (Shu et al., 2017; Vosoughi et al., 2018). However, beneath this success lies an implicit assumption: better prediction produces better understanding.This article challenges that assumption.Prediction and explanation are not equivalent.

Existing studies increasingly exhibits what may be termed prediction obsession, the tendency to evaluate theoretical advancement through predictive improvement while overlooking the interpretive mechanisms through which information becomes believable. Under this logic, misinformation is reduced to an output variable rather than treated as a socially constructed credibility phenomenon.As a consequence, prediction becomes the destination of inquiry rather than the object of explanation.A related limitation emerges through benchmark dependency. Detection studies frequently rely on standardized datasets and leaderboard-based comparison as indicators of scientific contribution. Although benchmarking supports replicability and methodological rigor, it also encourages narrow forms of optimization detached from real-world interpretive complexity. And Benchmark superiority demonstrates that systems classify better.It does not demonstrate that systems explain credibility better.

An algorithm may correctly identify manipulated information while remaining incapable of explaining why users perceived that information as authentic.This distinction becomes increasingly important because misinformation succeeds not merely because it is present but because it appears credible.Accordingly, the classification paradigm produces a paradoxical outcome: greater predictive sophistication accompanied by limited understanding of credibility construction.

2.2 Multimodal Detection as Fusion Logic: When Integration Replaces Interpretation

Recognizing the limitations of text-centric detection, recent research has introduced multimodal approaches that integrate textual and visual information into unified representations (Baltrušaitis et al., 2018; Li et al., 2019; Lu et al., 2019). These approaches improve representational quality by aligning heterogeneous inputs through multimodal fusion mechanisms. However, existing studies largely assume that greater integration automatically produces greater understanding. This study conceptualizes this assumption as fusion logic. Although fusion enhances predictive capability, it provides limited explanation regarding how integrated evidence becomes interpreted as credible. Aligned textual and visual signals may reinforce, distort, or legitimize narratives independent of factual accuracy. Consequently, meaning emerges not from computational integration alone but through interpretive interaction across modalities. Existing multimodal research therefore explains connection more effectively than credibility formation.

2.3 Explainability as Output Logic: Why Post-Hoc Explanation Is Not Credibility

A third research stream emphasizes explainable artificial intelligence and interpretable decision-making (Doshi-Velez & Kim, 2017;

Adadi& Berrada, 2018; Miller, 2019; Arrieta et al., 2020). Although this literature improves transparency and accountability, it largely assumes that explanation after prediction produces understanding. This study conceptualizes this assumption as output logic. Within this perspective, explanation functions primarily as a post-hoc justification of outputs rather than an explanation of how credibility emerges. However, feature attribution and attention visualization may increase observability without explaining how multimodal evidence becomes believable. In multimodal misinformation environments, credibility develops during interpretation rather than after prediction.

2.4 The Missing Mechanism: From Multimodal Attention to Credibility

Existing literature explains prediction, integration, and transparency.Yet none explains credibility formation.Classification logic explains whether information can be detected.Fusion logic explains how information sources are combined.Output logic explains how outcomes become observable.However, no dominant theoretical perspective explains how multimodal attention becomes credibility.Specifically, existing studies lacks a mechanism explaining how heterogeneous visual and textual evidence is transformed into interpretable judgments of authenticity.This unresolved mechanism constitutes the central theoretical deficiency addressed in this article.Accordingly, the next section introduces Cross-Modal Credibility Formation as a new conceptual construct explaining how multimodal attention converts integrated evidence into credibility judgments and how credibility emerges as a process of interpretation rather than a consequence of prediction alone.

III. Theoretical Foundations

To explain how multimodal evidence becomes interpreted as credible, weintegrate Information Processing Theory, Explainable Artificial Intelligence and Socio-Technical Systems Theory into a sequential interpretive framework. Collectively, these perspectives explain how multimodal attention organizes heterogeneous evidence, how reasoning becomes interpretable, and how credibility judgments become socially acceptable within digital misinformation environments.

3.1 Information Processing Theory: From Inputs to Credibility Evaluation

The first theoretical anchor is Information Processing Theory (Galbraith, 1974; Daft & Lengel, 1986), which explains how heterogeneous information becomes transformed into meaningful judgment through interpretive processing. Rather than assuming that information possesses objective meaning, the theory argues that meaning emerges through processes that organize, prioritize, and contextualize incoming signals.

Within CMCF, this perspective explains how textual and visual evidence becomes jointly interpreted through cross-modal attention. In multimodal misinformation environments, credibility judgments rarely emerge from isolated textual claims; instead, users evaluate coordinated combinations of images, captions, semantic framing, and contextual cues. Accordingly, multimodal attention is conceptualized as an interpretive mechanism that aligns heterogeneous evidence into credibility evaluations.

3.2 Explainable Artificial Intelligence: From Interpretation to Interpretability

The second theoretical foundation is Explainable Artificial Intelligence (Doshi-Velez & Kim, 2017; Adadi& Berrada, 2018; Miller, 2019; Arrieta et al., 2020). Within CMCF, XAI provides the theoretical basis for understanding how multimodal reasoning becomes interpretable.

Rather than treating explainability as a post-hoc explanatory layer, the framework positions interpretability as an embedded component of credibility formation. Interpretability increases visibility into cross-modal reasoning, clarifies evidence attribution, and transforms attention-based interpretation into understandable credibility signals. Under this perspective, explanation is not external to credibility formation but part of the process through which credibility judgments become understandable.

3.3 Socio-Technical Systems Theory: From Interpretability to Legitimacy

The third theoretical anchor is Socio-Technical Systems Theory (Trist & Bamforth, 1951; Bostrom & Heinen, 1977; Pasmore et al., 1982), which explains how technical outputs become socially accepted judgments. The theory argues that system effectiveness depends not only on computational capability but also on alignment between technological processes and human interpretation.

Within CMCF, credibility emerges when users perceive coherent relationships among evidence, reasoning, and final judgment. Accordingly, legitimacy is treated as a socio-technical outcome shaped by interpretability, plausibility, and procedural coherence rather than prediction accuracy alone.Combinedly, these three perspectives explain how multimodal attention

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 782-791 www.ijaem.net ISSN: 2395-5252

contributes to interpretable and socially accepted credibility judgments. This integrated framework establishes the theoretical foundation of Cross-Modal Credibility Formation.

3.4 Why VisualBERT?

VisualBERT is theoretically relevant to CMCF because it enables unified interaction between textual and visual representations through shared transformer attention mechanisms (Li et al., 2019). Unlike earlier multimodal architectures that process modalities separately before fusion, VisualBERT embeds visual regions and textual tokens within a common representational space, allowing cross-modal relationships to emerge during contextual learning.

This architecture is particularly suitable for credibility analysis because multimodal misinformation operates through coordinated interpretation of images, captions, semantic framing, and contextual cues. Transformer attention mechanisms dynamically assign relational importance across heterogeneous inputs, enabling multimodal evidence to become contextually aligned (Vaswani et al., 2017).Compared with ViLBERT and LXMERT, which maintain relatively separate visual and linguistic streams (Lu et al., 2019; Tan & Bansal, 2019), VisualBERT's integrated representational structure more directly supports examination of unified interpretive processing. More recent systems such as CLIP, BLIP, and Flamingo emphasize scalability, generative reasoning, and zero-shot performance (Radford et al., 2021; Li et al., 2022; Alayrac et al., 2022). However, these architectures are less oriented toward interpretive transparency within credibility evaluation contexts.Accordingly, VisualBERT is positioned as a theoretically appropriate architecture for examining how multimodal attention organizes heterogeneous evidence into interpretable credibility judgments.

IV. Developing Cross-Modal Credibility Formation

Existing misinformation research has substantially improved predictive detection, multimodal integration, and explainability. However, these approaches remain limited in explaining how heterogeneous textual and visual evidence becomes interpreted as credible. Misinformation rarely succeeds through factual superiority alone; rather, it becomes persuasive when multimodal evidence appears coherent, believable, and legitimate. This limitation suggests the need for a framework capable of explaining credibility as an interpretive process rather than a predictive outcome. So, this study introduces Cross-Modal Credibility Formation. CMCF conceptualizes credibility as a sequential interpretive process through which multimodal attention transforms textual and visual evidence into believable judgment. Unlike approaches that treat credibility as a by-product of prediction or transparency, the framework explains how credibility emerges through coordination among evidence alignment, attention mechanisms, interpretability, and human evaluation.The conceptual necessity of CMCF emerges from the limitations of adjacent constructs. Classification explains whether misinformation can be identified, explainability clarifies whether decisions can be understood, and trust addresses whether outcomes will be accepted. However, these constructs do not explain how multimodal evidence becomes believable during interpretation. CMCF therefore occupies the interpretive space connecting multimodal reasoning with credibility attribution.

This study conceptualizes CMCF as a four-dimensional process construct.Evidence Congruence refers to the perceived consistency between textual narratives and accompanying visual information. Credibility formation begins when multimodal evidence appears mutually reinforcing rather than contradictory. Accordingly, coherence across modalities establishes the initial condition for credibility evaluation.

Attention Alignment refers to the interpretive coordination through which multimodal attention prioritizes and contextualizes heterogeneous evidence. Drawing conceptually from transformer attention mechanisms (Vaswani et al., 2017), this dimension explains how cross-modal relationships become organized into meaningful evaluative structures. Within this framework, VisualBERT is positioned as an interpretive architecture that enables evidence alignment across modalities.

Interpretive Visibility refers to the degree to which multimodal reasoning becomes understandable and cognitively accessible. Extending explainability research (Doshi-Velez & Kim, 2017; Miller, 2019), this dimension conceptualizes interpretability as an embedded mechanism of credibility formation rather than a post-hoc explanatory layer. Interpretive visibility clarifies how textual and visual evidence jointly contribute to credibility evaluation.

Credibility Attribution refers to the evaluative outcome through which integrated evidence becomes perceived as believable. Extending credibility literature (Fogg & Tseng, 1999; Metzger et al., 2010), this dimension conceptualizes credibility not as an objective

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 782-791 www.ijaem.net ISSN: 2395-5252

property of information but as a perceived judgment emerging through multimodal interpretation.
Collectively, these dimensions explain how multimodal attention transforms heterogeneous evidence into interpretable and socially acceptable credibility judgments.

4.4 Conceptual Distinctiveness of CMCF

CMCF remains conceptually distinct from adjacent constructs. Explainability focuses on making outputs understandable, whereas CMCF explains how multimodal evidence becomes believable. Trust reflects willingness to accept outcomes, while CMCF explains the interpretive conditions enabling acceptance. Similarly, classification predicts informational categories, whereas CMCF explains how credibility emerges through multimodal interpretation. Accordingly, CMCF is positioned as an interpretive construct linking multimodal reasoning with credibility judgments.

4.5 Operationalization of CMCF

To support future empirical validation, the dimensions of CMCF can be operationalized through observable indicators and measurement approaches. Table 3 outlines the conceptual definitions, illustrative indicators, and potential empirical measures associated with each construct dimension (MacInnis, 2011).

Table 3. Operational Definitions and Measurement Possibilities of CMCF

Construct	Operational Definition	Example Indicators	Possible Measurement Approaches
Evidence Congruence	Degree of perceived alignment between textual narratives and visual information	Semantic consistency, contextual fit, image–text coherence	Likert-scale agreement items, multimodal consistency assessment
Attention Coordination	Extent to which multimodal attention prioritizes and organizes relevant evidence across modalities	Salience alignment, evidence prioritization, cross-modal weighting	Attention visualization analysis, eye-tracking, attention-mapping metrics
Interpretive Visibility	Degree to which multimodal reasoning processes become understandable to evaluators	Explanation clarity, reasoning accessibility, transparency perception	Explainability perception scales, XAI visibility measures
Evaluative Confidence	Degree of confidence users develop toward credibility judgments generated through multimodal interpretation	Judgment certainty, perceived reliability, confidence in reasoning	User evaluation scales, cognitive confidence assessment
Detection Legitimacy	Extent to which detection outcomes are perceived as procedurally acceptable and justified	Perceived fairness, procedural coherence, acceptance of outcomes	Legitimacy perception scales, procedural trust measures
Misinformation Resistance	Degree to which users sustain confidence in corrective judgments and resist misleading information	Reduced susceptibility, resilience to misinformation, corrective reliance	Experimental misinformation tasks, resistance intention scales

The proposed operational structure serves two purposes. First, it clarifies boundaries among the dimensions of CMCF by distinguishing interpretive coordination, visibility, evaluative confidence, and legitimacy as related but analytically separate mechanisms. Second, it establishes a preliminary foundation for empirical validation through experimental, survey-based, and computational approaches.Future research can refine these indicators, develop psychometrically validated scales, and test the proposed sequential relationships across different multimodal misinformation environments.

V. Visual BERT as a Credibility Engine: A Process Theory of Cross-Modal Credibility Formation

The preceding sections established CMCF as a process through which multimodal evidence becomes transformed into credibility judgments. Based on this perspective, this section reframed VisualBERT not merely as a multimodal classification architecture but as a credibility-oriented interpretive system. Rather than treating attention mechanisms solely as computational tools for prediction optimization, the framework positions multimodal attention as a mechanism that organizes

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 782-791 www.ijaem.net ISSN: 2395-5252

evidence into understandable credibility structures (Li et al., 2019; Vaswani et al., 2017).Accordingly, credibility formation is conceptualized as a six-stage interpretive sequence through which multimodal evidence progresses from perceptual input to accepted judgment.

Stage 1: Visual Encoding — Establishing Perceptual Grounding

The process begins with visual encoding, where images provide initial perceptual cues that shape plausibility judgments. Prior research suggests that visual content strongly influences perceived authenticity because users frequently interpret images as immediate evidence of realism and credibility (Fogg & Tseng, 1999; Metzger et al., 2010). In multimodal misinformation environments, visual signals therefore function as interpretive anchors that orient subsequent evaluation.Outcome: Perceptual grounding.

Stage 2: Semantic Encoding — Structuring Narrative Context

The second stage involves semantic encoding, where textual information contextualizes visual evidence and guides interpretation. Rather than functioning as isolated informational content, language structures relationships among events, actors, and meanings. Through this process, textual framing shapes how visual evidence is understood and integrated into broader narratives.Outcome: Contextual plausibility.

Stage 3: Cross-Modal Attention — Coordinating Evidence

Cross-modal attention refers to the alignment of textual and visual evidence into coherent interpretive structures. Rather than functioning as independent inputs, multimodal signals become selectively coordinated to support credibility evaluation. Attention mechanisms therefore organize informational relevance by prioritizing relationships across modalities (Vaswani et al., 2017).Outcome**:** Interpretive coherence.

Stage 4: Meaning Formation — Stabilizing Interpretation

Once evidence becomes coordinated, information is transformed into stable interpretive meaning. At this stage, relationships among multimodal cues become sufficiently organized to reduce ambiguity and support evaluative understanding. Meaning formation therefore represents the transition from multimodal coordination to cognitively accessible interpretation.

Outcome: Interpretive stability.

Stage 5: Credibility Construction — Generating Believability

Credibility construction refers to the process through which stabilized interpretation becomes perceived as believable. Individuals evaluate whether evidence appears internally consistent, contextually plausible, and sufficiently supported across modalities. Credibility therefore emerges through evaluative assessment rather than prediction accuracy alone.

Outcome**:** Perceived authenticity.

Stage 6: Detection Legitimization — Producing Accepted Judgment

The final stage involves detection legitimization, where credibility judgments become accepted as procedurally valid outcomes. Detection decisions gain legitimacy when users perceive coherence between evidence, interpretation, and explanatory reasoning. Under these conditions, multimodal detection operates not only as prediction but as socially acceptable judgment formation.Outcome**:** Accepted detection.

Process Logic of the Model: Visual EncodingSemantic → EncodingCross-Modal AttentionMeaning → FormationCredibility ConstructionDetection Legitimization.

This framework positions VisualBERT as a credibility-oriented architecture that structures multimodal evidence into interpretable judgment sequences. The model suggests that credibility does not emerge directly from prediction or information fusion alone. Instead, credibility develops progressively through coordinated stages of perception, contextualization, attention alignment, interpretation, and evaluative acceptance.

VI. Process Model and Propositions

The previous sections established that credibility formation in multimodal misinformation environments cannot be sufficiently explained through prediction accuracy, information fusion, or post-hoc interpretability alone. CMCF was therefore introduced as a process construct explaining how heterogeneous evidence becomes transformed into accepted judgment, while VisualBERT was repositioned as the interpretive engine enabling this transformation.The remaining task is to specify the causal architecture connecting these components.This section develops a process model that explains credibility formation as a progressive sequence of interpretive transitions rather than independent effects. The model proposes that multimodal credibility develops through cumulative stages in which informational coherence enables coordinated attention, coordinated attention enables interpretive accessibility, interpretive accessibility supports evaluative confidence, evaluative confidence produces legitimacy, and legitimacy ultimately strengthens resistance to misinformation.

Unlike conventional detection models that assume

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 782-791 www.ijaem.net ISSN: 2395-5252

direct relationships between inputs and outcomes, the proposed framework conceptualizes credibility as a mediated process in which each stage becomes the enabling condition for the next.

The resulting process sequence is shown below:
Evidence → Congruence Attention Coordination Interpretive → Visibility Evaluative Confidence Detection Legitimacy → Misinformation Resistance.

Table 2. Process Logic of Cross-Modal Credibility Formation

Stage	Process Question	Mechanism	Immediate Outcome
Evidence Congruence	Does evidence fit together?	Cross-modal consistency	Interpretive readiness
Attention Coordination	What becomes cognitively prioritized?	Selective weighting	Structured interpretation
Interpretive Visibility	Can reasoning become understood?	Explanatory accessibility	Evaluative confidence
Detection Legitimacy	Does judgment become acceptable?	Procedural coherence	Accepted detection
Misinformation Resistance	Does acceptance translate into resilience?	Sustained confidence	Resistance capability

Proposition Development

Credibility formation begins when textual and visual evidence appears mutually reinforcing. Evidence congruence reduces interpretive uncertainty and enables multimodal attention to become selectively coordinated (Baltrušaitis et al., 2018; Li et al., 2019).

P1: *Higher evidence congruence positively strengthens attention coordination.*

Attention coordination structures heterogeneous evidence into meaningful interpretive pathways. As multimodal relationships become organized, reasoning becomes more cognitively accessible and interpretable (Vaswani et al., 2017).

P2: *Greater attention coordination positively increases interpretive visibility.*

Interpretive visibility refers to the extent to which multimodal reasoning becomes understandable to evaluators. Greater visibility reduces ambiguity and strengthens evaluative confidence in credibility judgments (Doshi-Velez & Kim, 2017; Miller, 2019; Arrieta et al., 2020).

P3: *Higher interpretive visibility positively strengthens evaluative confidence*.

Evaluative confidence contributes to detection legitimacy by increasing perceptions of procedural coherence and judgment acceptability. Accordingly, legitimacy emerges through confidence in interpretive reasoning rather than prediction accuracy alone (Fogg & Tseng, 1999; Metzger et al., 2010).

P4: *Greater evaluative confidence positively increases detection legitimacy.*

The model further proposes that Cross-Modal Credibility Formation functions as the central interpretive mechanism linking coordinated attention with accepted judgment outcomes.

P5: *Cross-Modal Credibility Formation mediates the relationship between attention coordination and detection legitimacy.*

Finally, legitimacy strengthens misinformation resistance by increasing willingness to rely on corrective judgments and reducing interpretive uncertainty (Vosoughi et al., 2018).

P6: *Higher detection legitimacy positively improves misinformation resistance.*

Collectively, the propositions explain misinformation resistance as a sequential interpretive process in which multimodal attention, interpretability, evaluative confidence, and legitimacy jointly contribute to accepted credibility judgments.

6.1 Boundary Conditions of Cross-Modal Credibility Formation

The proposed CMCF framework is context-sensitive rather than universally stable. Credibility formation may vary across high-involvement and emotionally charged misinformation contexts where users rely more heavily on visual plausibility and narrative coherence. Platform affordances such as algorithmic amplification, rapid sharing, and visual prioritization may further shape cross-modal interpretation differently across digital environments.

In addition, individual and cultural differences may influence credibility evaluation. Media literacy, political ideology, cognitive bias, and prior expertise can affect how users interpret multimodal evidence, while cultural variation may shape perceptions of authenticity, symbolism, and narrative meaning. Accordingly, credibility formation should be understood as contingent on broader technological, cognitive, and socio-cultural conditions.

International Journal of Advances in Engineering and Management (IJAEM)
Volume 8, Issue 5 May. 2026, pp: 782-791 www.ijaem.net ISSN: 2395-5252

VII. Discussion

This study develops CMCF as a process-oriented framework explaining how multimodal evidence becomes transformed into credibility judgments within misinformation environments. The proposed framework suggests that multimodal misinformation should be examined through interpretive rather than purely predictive mechanisms. In this perspective, credibility emerges through the coordination of visual evidence, semantic framing, attention alignment, and evaluative interpretation rather than through classification performance alone.

The study contributes to multimodal misinformation research by repositioning VisualBERT as a credibility-oriented architecture rather than solely a prediction system. This perspective extends existing multimodal learning research by emphasizing how attention mechanisms organize evidence into interpretable judgment structures that shape perceived authenticity.

The framework also contributes to explainable AI studies by treating interpretability as an embedded component of credibility formation rather than a post-hoc explanatory layer. Accordingly, interpretability becomes connected to evaluative understanding and procedural acceptance within multimodal environments. This study shifts analytical attention from benchmark optimization toward credibility generation. This perspective suggests that future misinformation research should examine not only whether systems detect false content accurately, but also how multimodal reasoning influences interpretation, acceptance, and legitimacy.

Future research can empirically validate the proposed dimensions of CMCF, develop measurement scales for interpretive visibility and evidence congruence, and examine whether credibility formation varies across platforms, cultural settings, and misinformation contexts.

VIII. Contributions of the Study

This study contributes to multimodal misinformation research in several ways. It introduces Cross-Modal Credibility Formation as a novel theoretical construct explaining how textual and visual evidence becomes transformed into credibility judgments through sequential interpretive processes. This study reframes VisualBERT as a credibility-oriented interpretive architecture rather than solely a multimodal classification system. This framework extends existing misinformation studies beyond prediction-centered approaches by emphasizing credibility generation as a process of multimodal interpretation. Finally, the study shifts analytical attention from prediction accuracy toward interpretable and legitimate credibility judgments, proposing that multimodal AI systems should be evaluated not only through detection performance but also through their capacity to support understandable and socially acceptable reasoning processes.

IX. Conclusion

This study addressed a major limitation in multimodal misinformation research. The lack of a theoretical explanation for how heterogeneous textual and visual evidence becomes interpreted as credible. So, we developed CMCF as a process-oriented framework explaining how multimodal attention, interpretation, and evaluation jointly shape credibility judgments within digital information environments.The proposed framework suggests that misinformation detection should be understood not only as a predictive task but also as an interpretive process through which evidence becomes meaningful, believable, and procedurally acceptable. By emphasizing credibility generation rather than prediction performance alone, the study offers a broader perspective for examining multimodal AI systems and interpretability in misinformation contexts.

Future research should empirically validate the proposed dimensions of CMCF, examine the framework across different digital platforms and cultural contexts, and investigate how multimodal credibility formation operates under varying levels of media literacy, political polarization, and visual manipulation.

References

[1]. Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access, 6*, 52138–52160.

[2]. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. *Advances in Neural Information Processing Systems*.

[3]. Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... & Herrera, F. (2020). Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion, 58*, 82–115.

[4]. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2018). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on*

Pattern Analysis and Machine Intelligence, 41(2), 423–443.

[5]. Bostrom, R. P., & Heinen, J. S. (1977). MIS problems and failures: A socio-technical perspective. *MIS Quarterly, 1*(3), 17–32.

[6]. Daft, R. L., & Lengel, R. H. (1986). Organizational information requirements, media richness and structural design. *Management Science, 32*(5), 554–571.

[7]. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*.

[8]. Fogg, B. J., & Tseng, H. (1999). The elements of computer credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 80–87).

[9]. Galbraith, J. R. (1974). Organization design: An information processing view. *Interfaces, 4*(3), 28–36.

[10]. Li, L. H., Yatskar, M., Yin, D., Hsieh, C. J., & Chang, K. W. (2019). VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*.

[11]. Li, J., Li, D., Savarese, S., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the International Conference on Machine Learning*.

[12]. Lu, J., Batra, D., Parikh, D., & Lee, S. (2019). ViLBERT: Pretraining task-agnostic Visio linguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems*.

[13]. MacInnis, D. J. (2011). A framework for conceptual contributions in marketing. *Journal of Marketing, 75*(4), 136–154.

[14]. Metzger, M. J., Flanagin, A. J., Eyal, K., Lemus, D. R., & McCann, R. M. (2010). Credibility for the 21st century: Integrating perspectives on source, message, and media credibility in the contemporary media environment. In *The Handbook of Mass Media Effects* (pp. 293–335).

[15]. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence, 267*, 1–38.

[16]. Pasmore, W. A., Francis, C. A., Haldeman, J., & Shani, A. B. (1982). Sociotechnical systems: A North American reflection on empirical studies of the seventies. *Human Relations, 35*(12), 1179–1204.

[17]. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... &Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *Proceedings of the International Conference on Machine Learning*.

[18]. Tan, H., & Bansal, M. (2019). LXMERT: Learning cross-modality encoder representations from transformers. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

[19]. Trist, E. L., & Bamforth, K. W. (1951). Some social and psychological consequences of the longwall method of coal-getting. *Human Relations, 4*(1), 3–38.

[20]. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... &Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*.

[21]. Vosoughi, S., Roy, D., & Aral, S. (2018). The spread of true and false news online. *Science, 359*(6380), 1146–1151.

[22]. Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review, 14*(4), 490–495.