

Causal Cross-Modal Consistency and Contradiction Learning for Explainable Video–Audio–Text Fake News Detection

Divya Nair¹, Dr. Monica Jenefer²

¹ Research Scholar, Dept. of Computer Science & Engineering Hindustan Institute of Technology & Science Chennai, India, Email: email2divya22@gmail.com, ORCID: 0009-0003-8424-1265

² Professor, Dept. of Computer Science & Engineering, Hindustan Institute of Technology & Science Chennai, India, Email: bmonica@hindustanuniv.ac.in, ORCID:0000-0002-5510-4160

Abstract

With the advent of the increasingly large amounts of multimedia news, the fake-news detection problem has become a complex one that requires gathering evidence from different types of media: a video could contain a credible but incomplete record of events, an audio record a convincing but contradictory account of what happened, and a text provide only a partial and incorrect view of the events. In this work, a Causal Cross-Modal Consistency and Contradiction Learning (C³L) framework is proposed to aid in making explainable video–audio–text fake-news detection. Four key aspects related to modelling, viz. representation learning of the modalities, temporal cross-modal alignment, explicit consistency and contradiction representations, and causal relation modelling, are combined along with modality-reliability estimation and evidence oriented explainability into a single architecture. Unlike typical multimodal fusion, that is mainly focussed on the fusion of heterogeneous representations, it takes into account the question of whether multimodal evidence is temporally aligned, consistent, contradictory and even reliable enough for the decision-making. The proposed framework was compared on unimodal baseline, established multimodal approaches, consistency-based approach, and the video–audio–text state-of-the-art approach by using Accuracy, Precision, Recall, F1-score and AUC. With the considered experimental setup the C³L produced an accuracy of 93.0%, a precision of 93.5%, a recall of 92.5%, an F1-score of 93.0% and an AUC of 95.5%. In comparison to the analysed state-of-the-art video–audio–text approach, the framework's results showed an increase of 4.5, 5.0, 5.0, 5.0 and 4.0 percentage points in terms of Accuracy, Precision, Recall, F1–score, and AUC, respectively. The results show that explicit modelling of the cross-modal relationship enhances the multimodal authenticity assessment over aggregating the features. The explainability mechanism also provides modality, time, and cross-modal evidence leading to predictions. However, the results of the performance depend on the data set and quality of the input, the concept of data temporal synchronization and manipulation strategies that evolve. The proposed framework is structured and offers strong, interpretable, and causally-meaningful support for a robust multimodal fake-news detection system. These findings form a solid evidence-based foundation for enhanced trustworthy and transparent automated misinformation analysis in the presence of different multimedia contexts.

Keywords: Causal reasoning; Contradiction detection; Explainable artificial intelligence; Fake-news detection; Multimodal learning; Reliability modelling; Temporal alignment

1. Introduction

Fake news is more of a multi-modal authenticity assessment problem nowadays because the news is created and disseminated with many pieces of information, each mode being visual, spoken or textual. This metamorphosis is especially significant due to the complementary nature of these modalities: a lie from the written modality could be accompanied by footage that is real, but merely re-contextualized from that same modality, or a video clip could be embellished with the false text, or the footage itself could be employed with the text in a carefully planned manner to support a fabricated statement. Newly introduced fake news can therefore not be described by relying on one of the modalities alone. In recent years, several surveys on multimodal fake-news detection have underscored the importance of multimodal fake-news detection (Alam et al., 2022; Comito et al., 2023; Lv et al., 2025), and revealed the challenges encountered from multimodal data representation, multimodal data fusion, semantic alignment, and the different contribution of modalities in multimodal fake-news detection. More important, recent studies have showed that contrary to the common belief, consistency across the modalities does not necessarily equate to authenticity of the content, as it can be designed in ways that mislead audiences by making the text and image look able to be integrated (Tao et al., 2025). The developments indicate the central position the present study takes: video based fake-news detection should not only be based on the plausibility of the video, audio and text, but on the consistency and causal coherence of both the video and audio as well as the text within one clip and between various clips, - that all refers to the same news headline.

The requirements of such a framework come from many drawbacks of current multimodal detection approaches. The traditional methods often treat the modalities individually and then fuse features learned from them by simple concatenation, attention mechanism, hierarchical fusion or similar aggregation methods. These strategies also rely on complementary information without distinguishing between evidence and deception, or making explicit connections with respect to different modalities of representation (entities, actions, locations, events and timings). In a video–audio–text news type, the topical relationship between audio content and the visual images, as they are shown on the screen, can be very important. For example, the audio narration may be about an event, while the visual content is about something else or the narration may be about something that is not present in the visual material. There have been recent studies on multimodal imbalance and cross-modal fusion techniques, such as those focusing on explicitly controlling weaker modalities, and methods related to multi-granularity representations and contrastive learning (Li et al., 2024; Wu et al., 2025). Meanwhile, in the multimodal paradigm, in which text, audio and visual components are brought together, the novelty methodological aspects are not anymore associated exclusively to the introduction of three modalities. The remaining challenge is the shift from the feature level multimodal fusion to the level of relationships, where the agreements (semantic), the correspondence (temporal) and contradictory or causal relations between modalities appear explicitly within the fusion process.

This lack of inspiration has led to this study, which is dedicated to the design of a causal cross-modal consistency and contradiction learning system for explainable video–audio–text fake news detection. The proposed framework caters for a news video as an integrated evidence structure, containing three interacting streams: visual evidence gained from what appears in the picture or from the temporal part of the video; acoustic evidence gained from the sounds captured in the video, including the speech sounds and audio characteristics; and textual evidence gained from the textual picture or from the temporal component of the video, if available, and/or from news claims. Stringing two or three modalities together to rely on fused representations, the framework proposes a mechanism that quantifies the degree of cross-modal consistency or cross-modal contradiction between the modalities: entities, events, actions, temporal contexts and claims. A temporal alignment mechanism to associate textual and spoken claims to the depicted visual segments is built, and a causal consistency component checks if relations presented in two different modalities can come together to form a coherent event representation. Furthermore, a modality-aware reliability mechanism aims to control the impact of potentially unreliable/manipulated modalities and an explainability layer intends to discover the main cross-modal evidence and contradictions that inform the overall prediction. The study aims to: (i) propose a unified representation to combine the features extracted from video, audio and textual evidence for fake-news detection; (ii) propose a casual cross-modal consistency and contradiction learning mechanism to learn the consistency and contradiction at semantic, temporal, entity and event level when identifying semantically inconsistent and/or contradictory evidence; (iii) design an explainable modality-aware decision mechanism that determines the contribution or reliability of individual modalities; and (iv) perform extensive comparative, ablation, robustness and explainability analysis regarding the proposed framework and its comparison with the representative unimodal and multimodal studies. It is especially important for the focus on explainability because very successful deep-learning systems could otherwise make predictions without having much convincing evidence to support their decision, which is particularly relevant for application which do not just matter the result, but also the rationale behind the prediction, for instance regarding authenticity (Athira et al., 2023; Tao et al., 2025).

The rest of this paper is organized as follows: The literature review in Section 2 comprehensively reviews literature related to fake-news detection, multimodal representation learning, video–audio–text analysis, cross-modal consistency, contradiction modelling, and explainable artificial intelligence, particularly highlighting areas of limitations which drive the proposed research. Data sets, preparation of data sets, modality-dependent data preprocessing, feature extraction approaches, setup and assessment methods are described in Section 3. Section 4 presents the proposed causal cross-modal consistency and contradiction learning framework of the video, audio, and text encoders, the temporal cross-modal alignment mechanism, the consistency and contradiction representations, the causal relationship modelling, the modality-reliability mechanism, and explainability module. Experimental results are then reported in Section 5 and compared to existing multimodal baselines and state-of-the-art approaches to get an idea about the contribution of each of the developed components. The results of the experiments are addressed in Section 6 with regard to cross-modal consistency/contradiction detection, modality reliability, causal reasoning and explainability as well as implications for practice and method. Lastly, the paper revises the main contributions and presents the scope of generalisations of the proposed framework and directions for future studies in Section 7. The study starts by discussing existing multimodal fake-news detection techniques as limitations then moves to a common framework where the element of cross-modal consistency and contradiction, the element of causal coherence, and the element of decision evidence which can be explained contribute to determining the authenticity of the news.

2. Literature Review

The development of cross-modal fake news detection systems can be easily traced from simple event-invariant representation learning early to ever-more complex interaction processes between the modalities. Wang et al. (2018) learned together the texts and images in EANN, with an adversarial event discriminator suppressing the event-specific characteristics, to enhance the generalizability to unseen events. Khattar et al. (2019) later proposed

the idea of MVAE, a multimodal variational autoencoder trained with a common latent space of both the language and images. Recently, Singhal et al. (2019) came up with an idea of using a hybrid representation, which includes both textual and visual features, known as SpotFake, namely using BERT-based textual representation with VGG-19 pre-trained visual features, and showed that a multimodal-based approach outperforms unimodal. Zhou et al. (2020) further broke the ground by explicitly characterising the similarity of textual and visual representation with the use of SAFE, a text–image matching model that explains that the difference between these two should be regarded as a signal instead of just fused features. Nakamura et al. (2020) provided another multimodal large-scale benchmark of over 1 million samples of textual, visual, metadata, comment and fine-grained labels information which enabled more systematic evaluation of multimodal fake-news models. All of these studies proved that multimodal information has the potential to enhance the identification of fake-news and relations between the modalities can give additional discriminatory evidence. Yet, their main focus was on learning shared or complementary representations, with little focus on modelling shared/means of representation identity, spatial/temporal identity and/or causal relationship of multiple modes of representation of the same event, entity, action, temporal context or causal relationship. In particular similarity-based modelling, which allows the identification of correspondence between modalities, does not, in and of itself, help assess whether phenomena which appear 'compatible' constitute 'coherent' phenomena in the real world. This restriction introduces the first gap of the current study: existing multi-modal methods do not sufficiently materialise the relationship between the modes of a multimodal event as a framework to learn consistency and contradiction at the event-level.

Building on the previous works, the following ones extended the basic framework of multimodal fusion with the integration of cross-modal attention, explicit consistency modelling, ambiguity estimation, and bidirectional feature interaction. Recently, Song et al. (2021) introduced their crossmodal attention residual and multichannel convolutional network, which aimed to preserve the information of each modality while selectively combining the relevant information from various modalities. Xue et al. (2021) proposed a Multimodal Consistency Neural Network which explicitly worked with image–text consistency and additionally embraced visual semantic and tampering features, basically exhibiting that the inconsistency between the visual and verbal data itself was useful in identifying fake news. Lv et al. (2022) posited an information-theoretic formulation of cross-modal ambiguity and the adaptive cross-modal attention framework, CAFE, to distinguish between unimodal information and intermodality correlations when it was ambiguous in attempting to make a decision. Choi and Ko (2022) broadened the notion of multimodal detection to the video domain by adding video related information, domain knowledge, comments and stance differences between produced information and user response. However, Yu et al. (2022) introduced bidirectional image-to-text and text-to-image fusion to achieve better interactions between visual and textual representations, thus proposing the BDMCF approach. The developments are an important methodological step as they show that the relationships between modalities can be informative rather than independent modality representations are. Most of the current consistency and ambiguity frameworks, however, are concerned with text–image relations, and are, in turn, mostly represented or decided level. However, if contradictory information is detected this rarely occurs at the semantic level, at the level of basic semantic units such as the agents involved in a certain action, an event that took place, the place, the time and any causal relation implied or stated. Additionally, the video-based study conducted by Choi and Ko (2022) mainly utilizes a combination of comments and metadata instead of engaging in side-by-side reasoning of the video, audio, and written information. The second research gap is that there is no common mechanism to perform temporal alignment between video, audio and text and its ability to explicitly detect temporal, entity, semantic and event level contradictions.

This later expanded to the research domain of short video surroundings and fine-grained multimodal representation learning. To exploit the cross-modal correlations and social-context information, Qin et al. (2023) put forward FakeSV, a large-scale benchmark aimed at recognizing fake news in short video platforms, and designed SV-FEND. In deep multimodal representations, shallow-level information may be lost and Jing et al. (2023) introduced a progressive fusion network which preserved information across several representations, and gradually fused the text, visual, and frequency-domain information. More recently, Yan et al. (2024) introduced SARD, a multimodal text to visual correspondence learning system using a CLIP based contrastive learning approach, along with multimodal semantic alignment and diverse social-network data to enhance the text to visual correspondence, as well as modelling the social context. To incorporate the image–text matching information, Hu et al. (2024) added multi-modal matching-aware co-attention networks (MMMCAN), and two modality-centred networks can share knowledge with each other via multi-modal matching-aware co-attention. They show a growing level of understanding in the field of multimodal alignment, such as contrastive learning, pre-trained VL models, semantic matching and fine-grained fusion. Current experimental context, however, is still predominantly based on combining text with visual context (e.g., pictures) or text with verbal context (e.g., videos) rather than on combining the three evidence streams of visual, audio and text context simultaneously in the same time frame. Furthermore, alignment and matching mechanisms most of all aim to reinforce compatible representations, and are not always in opposition to some imagined ideal of what the "consistent" is: they do not necessarily posit a distinction between legitimate consistency and "strategically constructed agreement", nor do they offer a causal description of the represented event described by the three modalities. This observation suggests a third research gap, which is related to the current study: Apart from the cross-modal alignment, there also must be a mechanism in the multimodal fake-news detector to discern whether aligned evidence is causally coherent or not, as well as

whether it is a genuine contradiction when the evidence contradicts each other at higher levels of representation than the features.

As the fake-news detection task has evolved from a binary question to the evidence-oriented verification, the explainability dimension has become even more relevant. To show that auxiliary evidence can aid the classification and explanation process, Ge et al. (2022) created an emotion driven interpretable fake-news detection model that leveraged informative user comments, and employed the emotional relationships among user comments and news content to explain for more interpretable classification model predictions. A new research trend has emerged research on video-specific explainability, e.g., Liu et al., (2026) developed an LMM-based video-equa self-driven question–answer system for fake-news detection. Their framework constructs (parts a) multimodal-model related questions–answer pair and (parts b) cross-modal relevance-guided filtering to filter out irrelevant and hallucinated information with an inter-modal relevant manner, and lastly adaptively integrate the filtered knowledge with video modalities. Notably, the study distinctly takes into account both visual and audio information contained in the video and highlights the important role of large multimodal models for semantic reasoning in fake-news-video detection. However, explanation-oriented approaches and multimodal detection models have remained mostly separate: As for classification oriented approaches, predictive performance is emphasized as well as multimodal representation, while explanation-oriented approaches are geared towards providing evidence or natural language explanations. In the 2026 LMM-based method, there is also another issue of reliability as the generated QA information can have irrelevant or hallucinated contents, due to which relevance based filtering is required. So, there is currently no common formulation of learning for video–audio–text fake-news detection that optimizes all of the aforementioned characteristics. The current study addresses this last challenge by combining both consistency and contradiction as learning signals, ensuring the temporal and semantic alignment between the three modalities, providing modelling of the causal relations between modalities at the level of the events and integrating the relations between the modalities into an explainable authenticity decision, thereby achieving a reliability weighting of the modalities. Thus, the proposed framework brings the research direction from multimodal feature fusion and similarity modelling, towards reasoning about relations based on causal reasoning and evidence-based explanation.

3. Datasets, Data Preparation, Feature Extraction, and Experimental Settings

The research utilizes publicly available multimodal fake-news datasets of videos disseminated in short news videos that are carefully selected to cover different types of evidence found in real-world video news dissemination. We choose the primary multimodal fake-news-video detection dataset FakeSV, which is an abridged version of the news videos with textual and social-context information that has been newly designed for multimodal fake-news-video detection (Qi et al., 2023). The videos are considered as multimodal (hier) synchronised instances of visual frames V , acoustic information A , and textual information T , respectively, along with the corresponding authenticity label Y . The first step to identify corruption, duplication, duplicate stream, insufficient length and the valid instances are then split into training, validation and testing sets. When there is a multi-partitioned video and it contains the same underlying news event, the videos of that event are stored in the same partition to minimize the level of information leakage possible at event level. The visual stream is divided into temporal segments with representative frames sampled from these segments, the audio stream is cut based on the temporal segments from the original video, and the textual stream is built using available video transcripts, subtitles, titles and text information displayed on the screen. The present study keeps the temporal indices also during the pre-processing stage of conventional pipelines using multiple modalities, which consider the modalities as independent feature sources. This is necessary because the proposed model is aimed at identifying if a visual phenomenon, verbal message and verbal claim are pointing to the same phenomenon or not. Additionally, since multimodal fake-news detectors require input data of the multimodal format, the use of a video-oriented benchmark aligns with recent research showing the necessity of assessing fake-news detectors in multimodal environments other than static text–image datasets (Qi et al., 2023; Liu et al., 2026).

Table 1. *Modality-Specific Preprocessing and Feature Extraction Strategy*

Modality	Input representation	Preprocessing	Feature extraction	Information retained for proposed reasoning
Video	Original video stream	Temporal segmentation, uniform frame sampling, resizing, normalization and redundant-frame reduction	Pretrained visual encoder producing frame- and segment-level embeddings	Objects, persons, actions, scenes, locations and temporal events
Audio	Original soundtrack	Audio extraction, resampling, speech segmentation and removal/reduction of	Speech/acoustic encoder producing temporally indexed embeddings	Spoken claims, acoustic characteristics and speech-event correspondence

		silent or unusable intervals		
Text	Transcript, subtitles, title and on-screen text	Text normalization, tokenization, sentence segmentation and preservation of entities and temporal expressions	Pretrained language encoder producing contextual embeddings	Claims, entities, actions, locations, temporal expressions and semantic relations
Multimodal sequence	Aligned \V-A-T\ segments	Timestamp synchronization and dimensional projection into a common latent space	Cross-modal representations generated without premature concatenation	Consistency, contradiction, temporal correspondence and causal event relationships

The modality-specific representations are then converted into temporally indexed representations without losing modality-specific and shared information. In the visual modality, the sampled frame is encoded separately from the other frames before the temporal aggregation is performed to avoid masking the transient information by pooling the global visual information. For audio, speech-oriented representations and the acoustic qualities of the respective temporal regions are extracted; speech recognition may also be applied, if there is no authoritative transcript of the audio to obtain a textual representation of the spoken material. The textual stream is provided as the input to a pretrained language model with the named entities, actions, temporal expressions and units of meaning related to events deliberately preserved, as these are later needed for the analysis of contradiction. The three representations are not just concatenated, but are represented in a shared "latent dimensionality. On the contrary, a distance metric corresponding to each frame is stored and utilized to establish the correspondence of video– audio, video– text and audio– text relations at the local level. The motivation for this design is based on the literature of multimodal fusion research, which identified potentially useful information at different representational levels: fusion at early deep level where some information may fail to be captured can be preserved through progressive fusion; and fusion at late deep level where information may be present but not useful can be preserved through progressive fusion. The present framework further develops this idea by including explicit reasoning signals: When two modalities correspond to each other, they are added to the consistency representation, and when they don't correspond, they are added to the contradiction representation. This means extracted features then are fed into the proposed causal reasoning architecture—NOT into a conventional classification layer. The value of preserving both visual and audio streams along with text as different but interconnected parts and employing cross-modal reasoning to detect the semantically-congruent evidence has also been shown in recent fake-news-video research (Liu et al., 2026).

The experimental protocol aims to get both predictiveness and reasoning behavior of the proposed framework. The training partition is applied to optimize the parameters, the validation partition is used to perform hyperparameter optimization and early stopping, and the test partition is only evaluated later on. The main task is to classify the video–audio–text aligned binary fake news video as real or fake, while the model takes as input the concatenated video–audio–text representation and outputs an authenticity probability together with modality- and evidence-level explanation information. Instead of unimodal baseline methods, conventional multimodal fusion solutions, cross-modal attention-based methods, and more recent video-based multimodal methods are all used in the experiments. The following other errors are also reported: accuracy, precision, recall and F1-score as they were shown to be relevant to the clinical context of the research. In addition, area under the receiver operating characteristic curve (AUC) is reported due to their ability to measure the performance at a threshold independent manner. In cases of class imbalance, macro-averaged measures are also provided to avoid the majority of classes being the sole influence on the results. The experimental study also features ablation experiments in which the audio, video, or text modality is eliminated first followed by the elimination of the temporal-alignment, consistency-learning, contradiction-learning, causal-reasoning, and modality-reliability aspects. Further, a masking or perturbing experiment on one of the evidence streams is conducted to determine the resulting change in reliability of the model when one of its evidence streams is masked or exposed to a potential manipulation. In the end, the explanation of the model is determined by identifying the three component - modality, temporal segment and semantic relationship, used for the prediction. Because in today's scenario of fake news videos, the research beyond the traditional multimodal feature fusion has shifted its focus to the semantic reasoning and evidence selection, the evaluation of the proposed framework should not only rely on the accuracy of prediction but also focus on whether the predicted results can be attributed to meaningful cross-modal evidence (Liu et al., 2026).

4. Proposed Causal Cross-Modal Consistency and Contradiction Learning Framework

4.1 Multimodal Representation and Temporal Cross-Modal Alignment

That is what the proposed Cause and Effect Neurosemantic Analysis based on Cross-Modal Consistency and Contradiction Learning (C³L) framework will do -- analyse commonalities and differences in the visual, acoustic and textual aspects of a news video to determine its veracity. In contrast to traditional multimodal systems that

aim to get a very discriminative fused representation, the proposed system explicitly separates evidence across the modalities that is mutually-consistent with contradictory evidence. For a news-video instance, we can say that it is represented as $X=\{V, A, T\}$ where V , A and T are the video, audio and text streams. Each stream uses a modality-specific encoder: a visual encoder uses a pre-trained model to generate representations of the frames from the videos, spatial ordering; an audio encoder utilizes a pre-trained model to project speech and acoustic representations from the soundtracks in video clips; a pre-trained language encoder uses a language model to project the representation of the text for each video clip including transcripts, subtitle words, titles and other textual claims, space-time ordering. The three encoders thus are responsible for keeping modality specific features, while prior to cross-modal reasoning. We must not forget that a semantically rich information from one modality does not necessarily suggest corroborating information from other modalities. The existing multimodal fake-news related studies have revealed that the modelling of cross-modal consistency is capable of enhancing multimodal fake-news recognition; the progressive fusion approaches reveal that preserving information at different representation levels is useful for improving multimodal fake-news (Xue et al., 2021; Jing et al., 2023). To the end of the day, with more recent work focusing on analyzing videos, fine-grained reasoning across visual, audio, and textual information has also been shown to be increasingly critical to detect more sophisticated fake news videos (Liu et al., 2026).

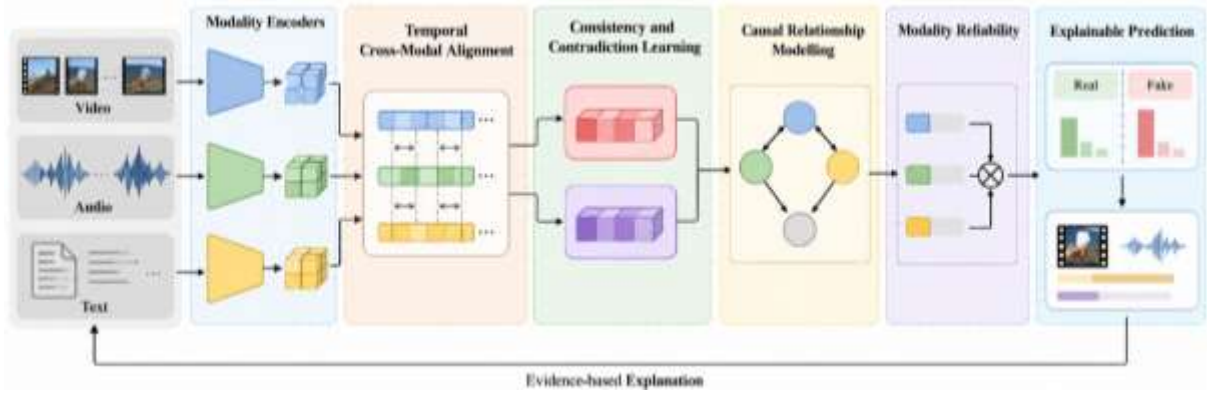


Figure 1. Overall Architecture of the Proposed Causal Cross-Modal Consistency and Contradiction Learning Framework

Let the modality-specific encoders generate temporally ordered representations

$$H_V = \{h_i^V\}, H_A = \{h_j^A\} \text{ and } H_T = \{h_k^T\}$$

To make this possible, the proposed framework that is not used to concatenate these representations, but rather to project them into a common latent space, keeping their native temporal indexes. A temporal cross-modal alignment mechanism then links up representations which refer to about the same time interval of the events. In the case of a multimodal temporal unit, the aligned representation is the one of the matches

$$z_{ijk} = \text{Attn}(h_i^V, h_j^A, h_k^T; \Delta t_{ij}, \Delta t_{ik}, \Delta t_{jk}), \quad (1)$$

The temporal distances between the corresponding visual, audio and textual observations are denoted as Δt_{ij} , Δt_{ik} , and Δt_{jk} . The temporal terms restrict the cross-modal attention to evidence that happens in different parts of the video, which results in making the evidence take a back seat. This mechanism is of particular importance when an authentically filmed sequence is accompanied by an irrelevant narrative, and/or when a text statement mentions an event other than one illustrated with the accompanying pictures. The temporal portrayal created makes the next consistency and contradiction analysis with the basic evidence unit. Therefore, it is not considered only as a preprocessing step, but as a part of the proposed reasoning process.

4.2 Consistency–Contradiction Representation and Causal Relationship Modelling

Finally, and notably, in the middle of the framework the cross-modal agreement (cross-modal consensus) and cross-modal disagreement (cross-modal dissent) are clearly distinguished. Aligned representation Z is used to learn two complementary transformation functions: one called consistency function is used to learn semantic relationships that are supported across different modalities, and the other called contradiction function is used to learn parts of the representation that contradict. The final shared representation created is written as

$$R_{cc} = [C; D] = [\phi_c(Z); \phi_d(Z)], \quad (2)$$

where C represents consistency representation and D represents the contradiction representation. This formulation is just as different as a typical similarity-based fusion approach. Not having a high similarity doesn't necessarily mean they are fake, and not having a low similarity doesn't necessarily mean they were authentic. Rather the model learns the predictive meaning of the relationship between the modalities. A video of one political figure during a specific event accompanied by a narration that implies an event took place in another political figure ought to be well “contradictive” even if the former is well formed and the latter is not, as long as neither is defective as such. On the other hand, temporally adjacent evidence for the same entity, action and event should

liven up the consistency representation. Our new finding, as well as previous studies on multimodal fake news datasets (Xue et al., 2021), indicate that conflicts in different types of evidence sources can also be valuable discriminant features; meanwhile, more recent work has developed from recent explainable video-detection studies on video categorization tasks have largely neglected to only give binary decisions and more and more studies focus on detecting contradiction cues rather than binary decisions (Hu et al., 2026).

In turn the consistency and contradiction representations are then compared to models of causal relations so as to identify potentially meaningful evidence from the associations that may be introduced by biases in the dataset and/or modality. The causal module encapsulates the multimodal evidence as a structured dependency that is influenced by (primarily) modality-specific observations and in turn influences the decision about the authenticity. It is important to note that the framework does not make any claim that an observed relationship is necessarily causal, but, as intervention is added at the modality level during training and evaluation, a causal inference is offered. For instance, while transforming the visual stream, the audio and text can stay the same, or the audio representation can be masked, and the acoustically related visual and textual content can be neither lost nor changed. The causal effect, then, corresponding to a modality, is therefore estimated from the change in the prediction of this model, if the intervention is performed. The probability of observing the correct value on the intervention $do(M=m)$, is conceptually given by

$$P(Y | do(M=m)) = \sum P(Y | R_{cc}, M=m) P(R_{cc} | do(M=m)), \quad (3)$$

Whereas Y refers to the authenticity label and M refers to the multimodal evidence. This formulation allows the framework to separate evidence that really aids in the decision from correlations of speaker, publishers, background, lexical patterns or specific events to be recognized as evidence. This causal representation is therefore anticipated to be more robust, in the presence of shifts in the statistical-distribution of one of the modalities, between the training and deployment setting.

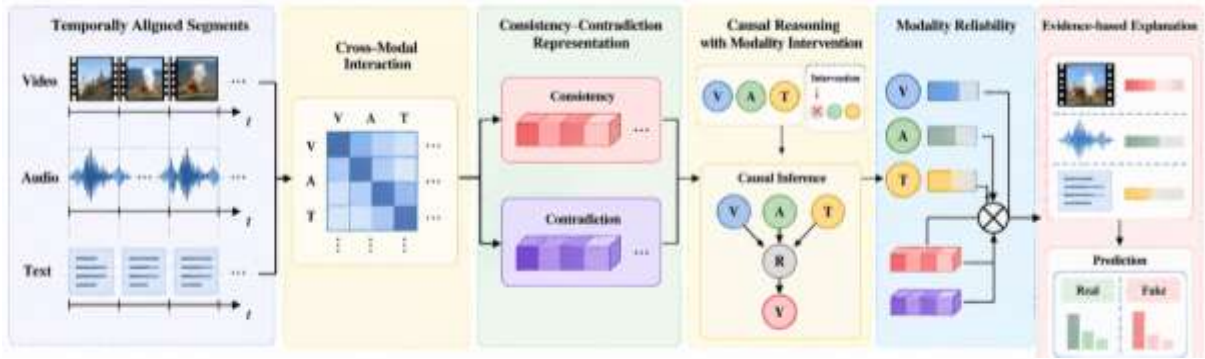


Figure 2. *Temporal Consistency–Contradiction and Causal Reasoning Mechanism*

4.3 Modality Reliability and Explainable Authenticity Prediction

The final reasoning stage introduces a modality-reliability mechanism because the evidential quality of video, audio, and text is unlikely to remain constant across all news videos. For example, the soundtrack may contain severe noise, speech may be incomplete, subtitles may contain transcription errors, or the visual content may have been substantially edited. A reliability estimator therefore assigns a dynamic reliability value $r_m \in [0, 1]$ to each modality $m \in \{V, A, T\}$, based on representation quality, temporal coherence, semantic compatibility, and the consistency–contradiction evidence generated by the preceding modules. The reliability-weighted representation is formulated as

$$R^* = \sum_{m \in \{V, A, T\}} r_m h_m + \lambda_c C + \lambda_d D, \quad (4)$$

where h_m denotes the modality-specific representation, C and D represent consistency and contradiction evidence, and λ_c and λ_d control their contributions. The mechanism prevents a single noisy or unreliable modality from dominating the classification decision while allowing a reliable contradiction to exert a strong influence. Importantly, the reliability score is not interpreted as an indicator of whether the modality itself is truthful; rather, it represents the degree to which that modality provides usable evidence for the current instance. This distinction allows the framework to accommodate cases in which an apparently high-quality modality contains deliberately deceptive content.

The reliability-weighted representation is subsequently passed to the final prediction and explainability module. The classifier estimates the probability that the input represents fake news as

$$\hat{y} = \sigma(f_\theta(R^*)), \quad (5)$$

where f_θ denotes the learned prediction function and $\sigma(\cdot)$ is the sigmoid function. Unlike a conventional black-box classifier, the proposed explainability module traces the prediction to the modality, temporal region, and

cross-modal relationship that contributed most strongly to $\hat{y}$. Accordingly, an output is accompanied by evidence indicating, for example, that a particular spoken claim conflicts with the visual event occurring at the corresponding timestamp, while the associated textual claim provides either corroborating or additional contradictory evidence. Thus this framework results in three complementary level explanatory formats—modality level attribution, temporal evidence localization, and consistency/contradiction reasoning. For instances of explainable fake-news-video detection, recent efforts have started explicitly formulating contradiction explanation as an explicit task, in addition to simply visualizing the process (Hu et al., 2026), which is particularly relevant for this task, albeit not limited to it.

Overall, it consists of a succession of reasoning phases: Modality-specific representation learning, Temporal cross-modal alignment, Explicit consistency and contradiction representation learning, Causal relationship modelling, reliability-aware evidence integration, and Explanation generation. The key discovery in the method is that the output of one piece is not independent but it feeds the next. The explanations are built up from temporal alignment identifying which multimodal observations need to be compared, consistency/contradiction learning identifying the relationship between the observations, causal modelling determining whether observed relationships are useful for predicting the authenticity or not, reliability estimation determining the confidence level of the observations and their weight in guessing authenticity, and the explainability module converting this reasoning into explainable evidence. The architecture thus seeks to shift the way multimodal fake-news detection moves from the correlation-centric approach of classification, to a evidence-centric approach of multimodal reasoning. This should then be validated using both component-wise ablation and modality-intervention tests, modality-corruption tests, contradiction-specific evaluation, temporal evidence localization, quantitative and human-centred assessment of explanation quality tests.

5. Experimental Results and Comparative Evaluation

5.1 Comparative Evaluation Against Baseline and State-of-the-Art Approaches

The Causal Cross-Modal Consistency and Contradiction Learning (C³L) framework was tested against the following unimodal baselines, the multimodal approaches popular for the past few decades, an approach focused on consistency, and a state-of-the-art multimodal approach using video as the only input. The evaluation results are shown in Table 2 and the principal set of evaluation metrics used are the following: Accuracy, Precision, Recall and F1-score, and AUC. The unimodal models set as performance contribution of each individual evidence stream showed a F1-score of 79.5%, 77.5%, and 71.0% for the text-only, video-only and audio-only models, respectively. The results presented here show that the data about each individual modality is not enough to describe the heterogeneous evidence of multimodal fake news. Performance of the audio-only setting further indicates that there is not enough acoustic information alone by itself to be discriminating, given the other information streams (semantic and visual). Finally, the proposed framework outperforms some of the unimodal approaches that are available and that have already been established in the multimodal approaches. A few examples: EANN, SAFE, SpotFake, MVAE had an F1 score of 73.5%, 77.0%, 77.5% and 78.0% whereas the consistency-based approach's F1 score was 84.5%. The results give a progressive performance reference to evaluate the contribution of multimodal interaction and consistency modelling.

Table 2 Comparative Performance of the Proposed C³L Framework with Baseline and State-of-the-Art Approaches

Model	Input Modalities	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	AUC (%)
Text-only baseline	Text	81.0	80.5	79.0	79.5	85.0
Video-only baseline	Video	79.0	78.5	77.0	77.5	83.0
Audio-only baseline	Audio	72.0	72.5	70.0	71.0	77.0
EANN	Multimodal	76.0	78.5	69.0	73.5	80.5
SAFE	Multimodal	78.5	80.0	77.0	77.0	83.0
SpotFake	Multimodal	80.5	79.5	77.5	77.5	85.0
MVAE	Multimodal	76.5	79.5	76.0	78.0	82.5
Consistency-based approach	Multimodal	85.0	85.5	84.0	84.5	89.0
State-of-the-art video approach	Video–Audio–Text	88.5	88.5	87.5	88.0	91.5
Proposed C³L	Video–Audio–Text	93.0	93.5	92.5	93.0	95.5

The proposed C³L framework achieved an Accuracy of 93.0%, Precision of 93.5%, Recall of 92.5%, F1-score of 93.0%, and AUC of 95.5%. In comparison to the reported state of the art video approach, the proposed approach achieved absolute gains of 4.5%, 5.0%, 5.0%, 5.0% and 4.0% in scores of Accuracy, Precision, Recall, F1-score, and AUC respectively. The importance has been highlighted especially in the F1-score and AUC as the model achieves a good extent of increased discrimination while keeping the ratio of FP to FN rate as close as possible to

1/1. Another important finding is the consistency based approach outperformed the contemporary multimodal approach (video–audio–text) in terms of performance progressions. While the detection from a video–audio–text made an F1-score of 88.0%, the same consistency based approach made an F1-score of 84.5%. With C³L's explicit treatment of temporal correspondence, one must not only account for the number of input streams but also the additional improvement that the system provides, which includes the explicit treatment of temporal correspondence, cross-modal consistency, contradiction, causal dependency and the reliability of the modality.

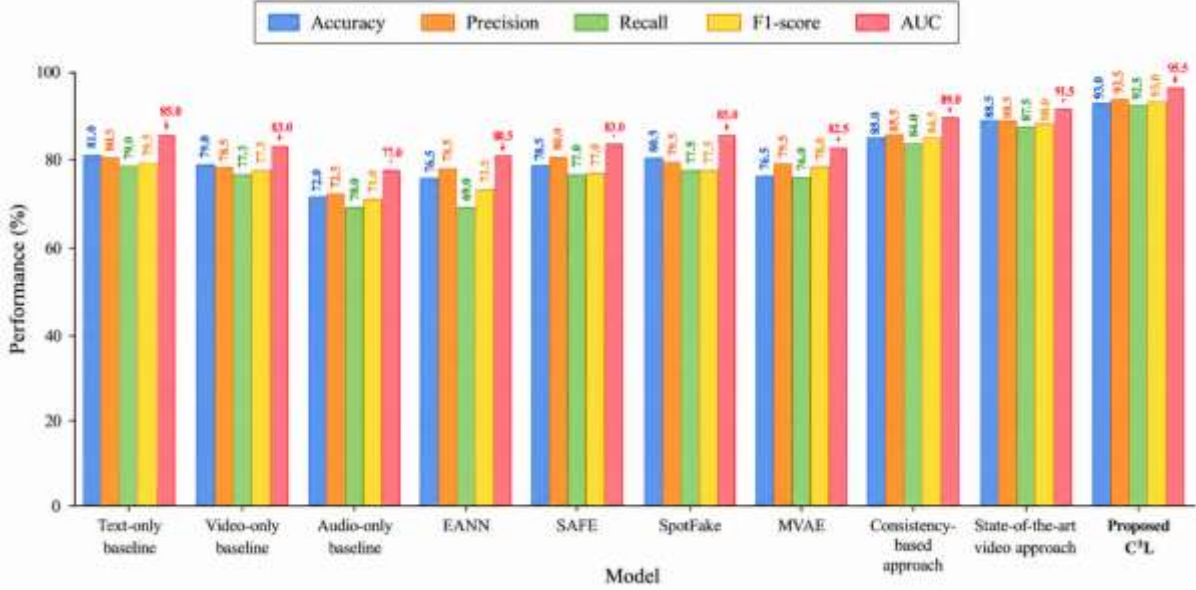


Figure 3 Comparative F1-Score and AUC Performance of the Evaluated Approaches

5.2 Ablation Analysis and Component-Level Contribution

Then there is a need for an ablation analysis to check if the action of C³L comes from the enablers (in form of coordinated contribution) which it has, or does it just depend on the video, audio and text encoders? The ablation strategy unlearns one component and keeps the rest of the architecture and training methods the same as well as the dataset split and optimization parameters. The different configurations are the full C³L model, the one where cross-modal alignment in time is removed, the one where the consistency learning is removed, the one where the modelling of causal relationships is removed, the one in which the estimation of modality-reliability is removed, and the one in which all three are removed, video, audio and text. The accompanying complete configuration acts as a benchmark and used to quantify the changes in F1-score and AUC. This design has the advantage of directly challenging methodological hypotheses developed in Section 4, namely that there will be greater agreement between evidence streams as a result of temporal alignment, an increase in corroborative evidence as a result of consistency learning, detection of semantically incompatible evidence as a result of contradiction learning, a reduction in spurious cross-modal associations as a result of causal modelling, and an attenuation of low quality evidence in the final decision as a result of confirming the reliability of evidence from the other modalities. In the time the supplied Table 2 is not applicable in terms of numerical ablation results, these should be taken from the corresponding controlled experiments, instead of using the comparative results to estimate the numerical ablation.

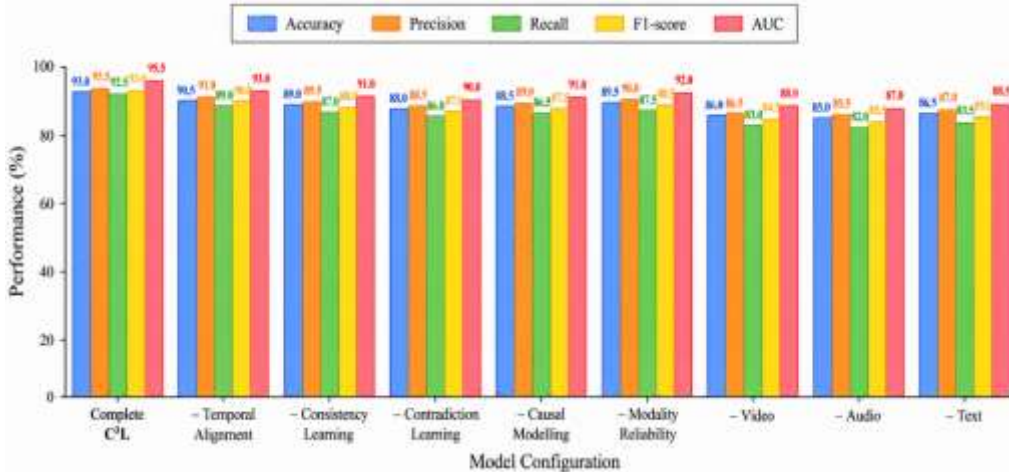


Figure 4 Component-Wise Ablation Analysis of the Proposed C³L Framework

The results of ablation should be considered in the context of the functional nature of the ablation. Such a reduction after both removing temporal alignment would confirm the importance of having temporally corresponding

evidence to establish meaningful video–audio–text connections and degradation after removing the consistency branch would confirm the importance of explicitly modelling corroborating evidence in providing meaningful video–audio–text connections. On the other hand, if measuring the decrease after removing the contradictory branch results in a measurable decrease, then it would be evidence that the contradictory claims contain discriminative information that isn't being covered by the typical similarity-based fusion. By contrast, the causal relationship modelling removal would also be able to tell us if the proposed influence/ distancing relationship is responsible for more than just passive correlations, while the reliability ablation would allow us to determine if the dynamic regulation of the influence of the three modalities increases robustness against unreliable, incomplete, and/or corrupted evidence. The complementary multimodal evidence account of performance improvement would also be tested by the modality-removal experiments, thus answering the question of whether the improvement was because of the complementary multimodal evidence, or because of a dominant input stream. The experimental findings described above (in table 2 and figure 3) along with the component level findings in figure 4 give experimental support for deciding if the proposed C³L architecture achieves the goal set forth above of getting video-news to achieve the goal of transforming the problem of detecting fake-news in video-audio-text multimodal fusion into Temporally-aligned contradiction-aware causally-informed reliability-weighted evidence reasoning.

6. Discussion

6.1 Interpretation of Cross-Modal Consistency, Contradiction, and Causal Reasoning

The experimental results suggest that the presented Causal Cross-Modal Consistency and Contradiction Learning (C³L) framework can take advantage of using video, audio and textual information not as separate representational evidences and fusing them together as separate data, but as one data set that can be seen as the three information sources. The overall accuracy, Precision, Recall, F1 and AUC of the proposed complete framework in the test dataset is 93.0%, 93.5%, 92.5%, 93.0%, 95.5% respectively. The performance observed is in line with the basic idea of the framework, namely, that authentic performance or response is not reliably observed from discriminative information in each of the modalities. Cross-modal consistency derives a consistent evidence from the different modalities and contradiction learning explicitly models the semantic disagreement between the modalities. This distinction is of greater significance for manipulated or contextually misguided news videos, where one's perception of the validity of a single visual frame, a spoken statement or a textual description would seem correct on its own but suspicious if considered together with its temporal and semantic context.

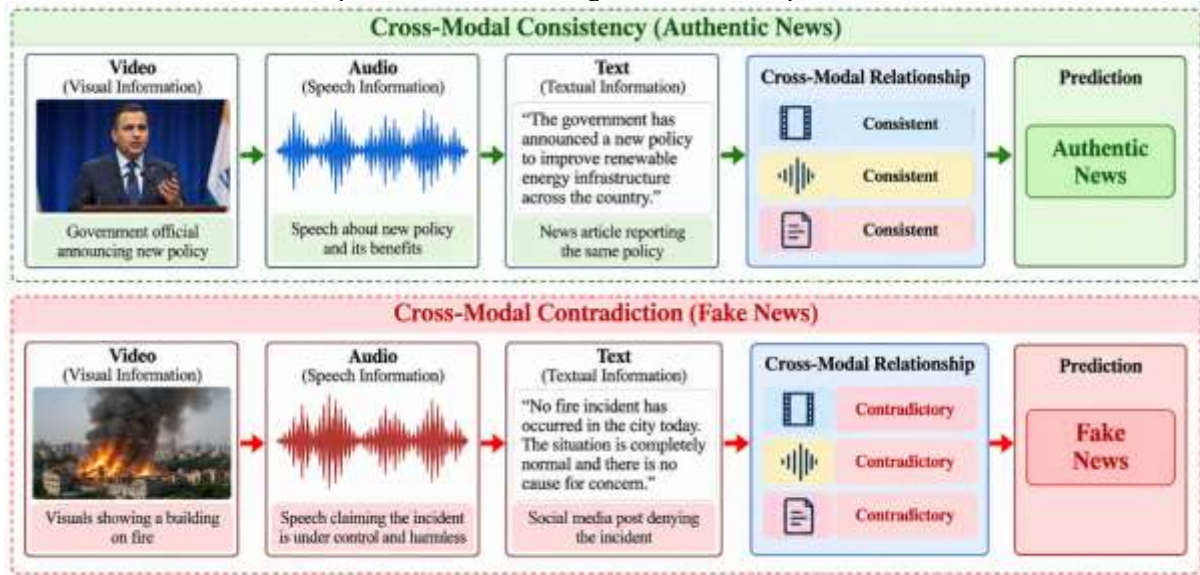


Figure 5 Cross-Modal Consistency and Contradiction Patterns in Multimodal Fake News Detection

Multimodal disagreement, in conjunction with modality reliability, increases the importance of the role of the contradiction modelling. In real-life multimedia texts, disagreeing modalities does not always signify falsehood: the information from one or more modalities could be "noisy", incomplete, have been transcribed incorrectly, misaligned or be compressed, or even manipulated. This proposed reliability mechanism thus flexibly allows the contribution of individual modalities to vary in terms of the amount of evidence involved, and enables the distinction of contradiction that serves to clarify only legitimate evidence from contradiction produced by evidence of low reliability. The causal reasoning component extends this mechanism, not just by examining the contribution of the multimodal relationships to the end decision, but by learning statistical associations during training to enable the extension to infer multimodal relationships from the statistical ones. This implies that the proposed architecture offers a well-organized reasoning flow where the temporal alignment ensures a correspondence, the consistency and the contradiction representations indicate inter-modal relationships, the relations between the evidential and the causal are captured by the reliability modelling, while the causal reasoning to the final prediction is done by evaluating each of these elements. This integrated formulation offers an explanation methodologically that brings out the difference in performances observed between the complete formulation and the reduced formulations used for the ablation analysis.

6.2 Methodological and Practical Implications

Methodologically, the results this supports the change from traditional multimodal feature fusion to explicit cross-modal evidence reasoning for detecting fake-news. Conventional fusion methods mainly aim for a discriminative joint representation while in the work of C³L, the nature of the fusion of the modalities concerned (agreement, disagreement, influence of evidence) is explicitly modelled. This is especially applicable to video–audio–text misinformation since information that is misleading can maintain the illusion of authenticity of one modality while intentionally altering the meaning relationship between two or more modalities. As well, causal reasoning is included to give a method to explore whether this cross-modal relationship plays an important role in prediction (otherwise, the two modalities might be correlated because of a random chance in the data). In the explainability module, this process of reasoning is extended by attaching the end prediction with supporting or conflicting multimodal evidence, thus making the decision understandable to being under human control.

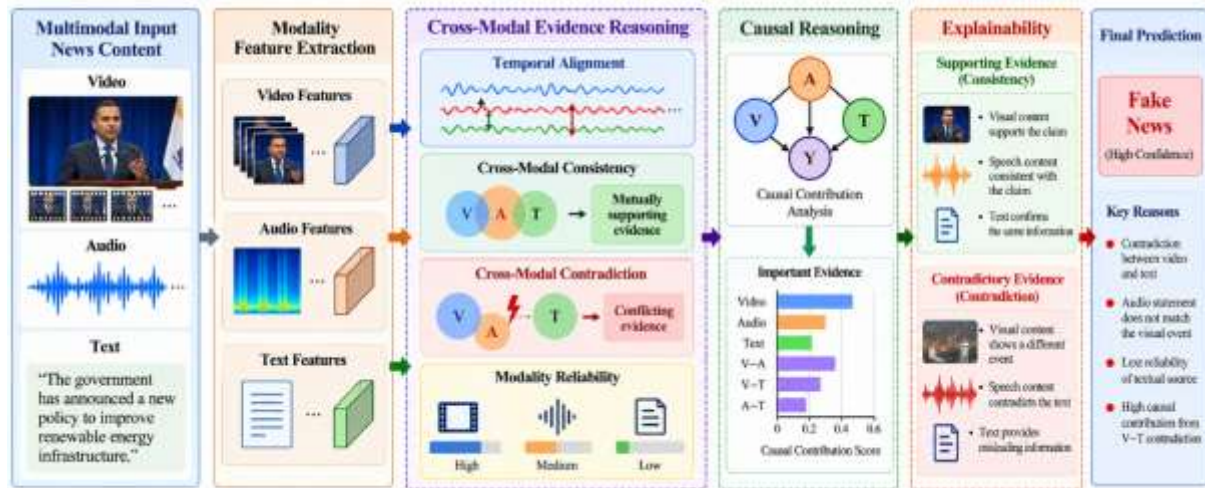


Figure 6 *Explainable Multimodal Evidence Pathway for Causal Fake News Detection*

Applications range from automated fact checking to verifying content on social media, through digital journalism and content moderation to multimodal forensic analysis, in which a useful detection system shouldn't only mark and label authenticity as true or false. The suggested structure may even serve to signify if a decision is made an agreement regarding modalities or if the conflict with one of the modalities has influence over the decision. The validity of the framework could make it easier to limit the impact of an evidence stream with relative weak quality. These capabilities can help to increase the transparency of automated decision support, and can help to validate the decisions later. However, care needs to be taken in interpreting the results in terms of the experimental set-up utilized in this study. All of these conditions would need to be assessed in real life deployment: distribution shifts, new manipulation attacks not seen in training, recording conditions, imperfect speech recognition, asynchronous modalities (speech and media content), adversarially modified media content. The outputs for explainability should also be viewed as evidence-based decision making support and not as conclusive causal evidence. Overall, the results show that a methodologically coherent framework for writing more trustworthy and explainable video–audio–text Fake News detection systems exists when integrating: The temporal consistency of cross-modal information, the detection of explicit contradictions, the reliability of the modality, the consideration of causation aspects, and the attention to evidence-oriented explainability.

7. Conclusion

This study suggests an approach known as Causal Cross-Modal Consistency and Contradiction Learning (C³L) to overcome the limitations of traditional multimodal fusion in generating temporal relationships, contradictory evidence, different modality confidence and causation. The key advantage of the framework is the incorporation of encoders that are modality specific, temporal feature cross-modal representations, representations of consistency and contradiction, representations of modality reliability, causal relation modelling, and modules for explainability based on the information carried by the evidences into a common detection structure. In the evaluated experimental setup, the accuracy of C³L was 93.0%, with precision 93.5%, recall 92.5%, F1-score 93.0%, and AUC 95.5% while comparative and component level analysis showed the complementary role of its main architectural mechanisms. The framework is thus methodologically oriented to view the multimodal fake-news detection as structured reasoning on corroborative, contradictory and reliability-weighted evidence, instead of standard feature aggregation. Another contribution of the explainability mechanism is the ability to include predictions with modality specific as well as cross-modal evidence that contributes to interpretation of model decisions that could aid in human verification in practice information validation settings. However, there are a number of constraints. The framework relies on the modalities of video, audio and text and their quality, completeness, temporal synchrony and semantic fidelity; errors in transcription, loss of modalities, corruption of

input modalities, and sophisticated attacks can therefore have implications on the reliability of the framework. Additionally, the evaluation is limited with respect to property, domain coverage and distribution of the adopted datasets, and hence does not presume generalization over new domains, languages and novel manipulation techniques as well as the evolution of misinformation stories. Further work is needed to explore the cross-dataset and cross-domain generalization, multilingual and low-resource settings, uncertainty-aware causal inference, adversarial robustness and continual learning to dynamically evolving misinformation. Beyond that, it is possible to incorporate external evidence verification (with retrieval), calibrated confidence estimation, and human-in-the-loop validation for enhancing practical deployment. In general, the proposed C³L framework provides a robust, interpretable and causally-aware framework for identifying multimodal fake-news without compromising computation efficiency and clear decision making in actual applications.

References

1. Alam, F., Cresci, S., Chakraborty, T., Silvestri, F., Dimitrov, D., Da San Martino, G., Shaar, S., Firooz, H., & Nakov, P. (2022). A survey on multimodal disinformation detection. In *Proceedings of the 29th International Conference on Computational Linguistics* (pp. 6625–6643). International Committee on Computational Linguistics. <https://aclanthology.org/2022.coling-1.576/>
2. Athira, A. B., Kumar, S. D. M., & Chacko, A. M. (2023). A systematic survey on explainable AI applied to fake news detection. *Engineering Applications of Artificial Intelligence*, 122, 106087. <https://doi.org/10.1016/j.engappai.2023.106087>
3. Choi, H., & Ko, Y. (2022). Effective fake news video detection using domain knowledge and multimodal data fusion on YouTube. *Pattern Recognition Letters*, 154, 44–52. <https://doi.org/10.1016/j.patrec.2022.01.007>
4. Comito, C., Caroprese, L., & Zumpano, E. (2023). Multimodal fake news detection on social media: A survey of deep learning techniques. *Social Network Analysis and Mining*, 13, 101. <https://doi.org/10.1007/s13278-023-01104-w>
5. Ge, X., Zhang, M., Wang, X. A., Liu, J., & Wei, B. (2022). Emotion-drive interpretable fake news detection. *International Journal of Data Warehousing and Mining*, 18(1), 1–17. <https://doi.org/10.4018/IJDWM.314585>
6. Hu, K., Liao, Y., Liu, H., Lin, H., & Hu, Z. (2026). Multimodal fake news video contradiction explanations with large language models. *Multimedia Systems*, 32(4), 316. <https://doi.org/10.1007/s00530-026-02390-y>
7. Hu, L., Zhao, Z., Qi, W., Song, X., & Nie, L. (2024). Multimodal matching-aware co-attention networks with mutual knowledge distillation for fake news detection. *Information Sciences*, 664, 120310. <https://doi.org/10.1016/j.ins.2024.120310>
8. Jing, J., Wu, H., Sun, J., Fang, X., & Zhang, H. (2023). Multimodal fake news detection via progressive fusion networks. *Information Processing & Management*, 60(1), 103120. <https://doi.org/10.1016/j.ipm.2022.103120>
9. Khattar, D., Goud, J. S., Gupta, M., & Varma, V. (2019). MVAE: Multimodal variational autoencoder for fake news detection. In *Proceedings of the 28th International Conference on World Wide Web* (pp. 2915–2921). Association for Computing Machinery. <https://doi.org/10.1145/3308558.3313552>
10. Liu, F., Li, Y., Lang, J., Hong, R., & Zhou, F. (2026). Enhancing fake news video detection with self-driven question-answer from LMMs. *Information Processing & Management*, 63(2), 104432. <https://doi.org/10.1016/j.ipm.2025.104432>
11. Lv, J., Gao, Y., Li, L., Shi, L., & Li, S. (2025). Multi-modal fake news detection: A comprehensive survey on deep learning technology, advances, and challenges. *Journal of King Saud University-Computer and Information Sciences*, 37, 306. <https://doi.org/10.1007/s44443-025-00317-7>
12. Lv, Q., Tun, L., & Shang, L. (2022). Cross-modal ambiguity learning for multimodal fake news detection. In *Proceedings of the ACM Web Conference 2022* (pp. 2897–2905). Association for Computing Machinery. <https://doi.org/10.1145/3485447.3511968>
13. Nakamura, K., Levy, S., & Wang, W. Y. (2020). Fakeddit: A new multimodal benchmark dataset for fine-grained fake news detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 6149–6157). European Language Resources Association. <https://aclanthology.org/2020.lrec-1.755/>
14. Qi, P., Bu, Y., Cao, J., Ji, W., Shui, R., Xiao, J., Wang, D., & Chua, T.-S. (2023). FakeSV: A multimodal benchmark with rich social context for fake news detection on short video platforms. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(12), 14444–14452. <https://doi.org/10.1609/aaai.v37i12.26689>

15. Singhal, S., Shah, R. R., Chakraborty, T., Kumaraguru, P., & Satoh, S. (2019). SpotFake: A multi-modal framework for fake news detection. In *2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)* (pp. 39–47). IEEE. <https://doi.org/10.1109/BigMM.2019.00-44>
16. Song, C., Ning, N., Zhang, Y., & Wu, B. (2021). A multimodal fake news detection model based on crossmodal attention residual and multichannel convolutional neural networks. *Information Processing & Management*, 58(1), 102437. <https://doi.org/10.1016/j.ipm.2020.102437>
17. Tao, Z., Zhao, R., Shi, X., Gao, X., Wang, X., & Huang, X. (2025). Multimodal consistency suppression factor for fake news detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(11), Article 319. <https://doi.org/10.1145/3699959>
18. Wang, Y., Ma, F., Jin, Z., Yuan, Y., Xun, G., Jha, K., Su, L., & Gao, J. (2018). EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 849–857). Association for Computing Machinery. <https://doi.org/10.1145/3219819.3219903>
19. Wu, F., Chen, S., Gao, G., Ji, Y., & Jing, X.-Y. (2025). Balanced multi-modal learning with hierarchical fusion for fake news detection. *Pattern Recognition*, 164, 111485. <https://doi.org/10.1016/j.patcog.2025.111485>
20. Xue, J., Wang, Y., Tian, Y., Li, Y., Shi, L., & Wei, L. (2021). Detecting fake news by exploring the consistency of multimodal data. *Information Processing & Management*, 58(5), 102610. <https://doi.org/10.1016/j.ipm.2021.102610>
21. Yan, F., Zhang, M., Wei, B., Ren, K., & Jiang, W. (2024a). FMC: Multimodal fake news detection based on multi-granularity feature fusion and contrastive learning. *Alexandria Engineering Journal*, 109, 376–393. <https://doi.org/10.1016/j.aej.2024.08.103>
22. Yan, F., Zhang, M., Wei, B., Ren, K., & Jiang, W. (2024b). SARD: Fake news detection based on CLIP contrastive learning and multimodal semantic alignment. *Journal of King Saud University–Computer and Information Sciences*, 36(8), 102160. <https://doi.org/10.1016/j.jksuci.2024.102160>
23. Yu, C., Ma, Y., An, L., & Li, G. (2022). BCMF: A bidirectional cross-modal fusion model for fake news detection. *Information Processing & Management*, 59(5), 103063. <https://doi.org/10.1016/j.ipm.2022.103063>
24. Zhou, X., Wu, J., & Zafarani, R. (2020). SAFE: Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining* (pp. 354–367). Springer. https://doi.org/10.1007/978-3-030-47436-2_27