Entity-Aware Cross-Modal Fusion Network for Fine-Grained Entity Consistency Verification in Multimodal News Misinformation Detection

Mark Harris *, Hunter Shaw , Ryan Young

Posted Date: 9 January 2026

doi: 10.20944/preprints202601.0691.v1

Keywords: cross-modal entity consistency; EACFN; graph neural network; reference images; entity verification

 doi:10.20944/preprints202601.0691.v1

Article

Entity-Aware Cross-Modal Fusion Network for Fine-Grained Entity Consistency Verification in Multimodal News Misinformation Detection

Mark Harris *, Hunter Shaw and Ryan Young

California State University Sacramento
* Correspondence: bnl056675@cncivirtual.mx

Abstract

Multimodal misinformation demands robust Cross-modal Entity Consistency (CEC) verification, aligning textual entities with visual depictions. Current large vision-language models (LVLMs) struggle with fine-grained entity verification, especially in complex "contextual mismatch" scenarios, failing to capture intricate relationships or leverage auxiliary information. To address this, we propose the Entity-Aware Cross-Modal Fusion Network (EACFN), a novel architecture for deep semantic alignment and robust integration of external visual evidence. EACFN incorporates modules for entity encoding, cross-attention for reference image enhancement, and a Graph Neural Network (GNN)-based module for explicit inter-modal relational reasoning, culminating in fine-grained consistency predictions. Experiments on three annotated datasets demonstrate EACFN's superior performance, significantly outperforming state-of-the-art zero-shot LVLMs across tasks, particularly with reference images. EACFN also shows improved computational efficiency and stronger agreement with human judgments in ambiguous contexts. Our contributions include the innovative EACFN architecture, its GNN-based relational reasoning module, and effective integration of reference image information for enhanced verification robustness.

Keywords: cross-modal entity consistency; EACFN; graph neural network; reference images; entity verification

1. Introduction

The proliferation of digital news media has unfortunately coincided with a surge in multimodal misinformation, where misleading or fabricated information is often conveyed through inconsistencies between text and accompanying images. A critical aspect of combating this challenge lies in the Cross-modal Entity Consistency (CEC) verification task [1], which aims to determine whether the entities (e.g., persons, locations, events) mentioned in a news article's text align truthfully with those depicted or implied in its associated images. This task is paramount for identifying "contextual mismatch" news, where images are often deceptively paired with unrelated or subtly altered text to convey false narratives. Enhancing the fine-grained entity verification capability in such scenarios is therefore a crucial step towards more accurate and robust fake news detection systems.

Existing zero-shot approaches based on large vision-language models (LVLMs) have shown promising capabilities in various cross-modal tasks [2,3], including multi-style image captioning [4], sketch storytelling [5], and visual question answering [6]. However, their performance in fine-grained entity consistency verification, particularly in complex real-world news contexts involving subtle discrepancies [7–9] and inherent uncertainties, still presents significant challenges [10]. This difficulty echoes broader challenges in computer vision regarding robust consistency, such as maintaining cross-view consistency in self-supervised monocular depth estimation [11]. These models often struggle to capture the intricate relationships between specific textual entities and corresponding visual regions, or to effectively leverage auxiliary information, such as reference images, to confirm or refute consistency.

 doi:10.20944/preprints202601.0691.v1

This limitation motivates our work to develop a method that can deeply understand and align entities across modalities while robustly integrating external visual evidence for more reliable judgments.

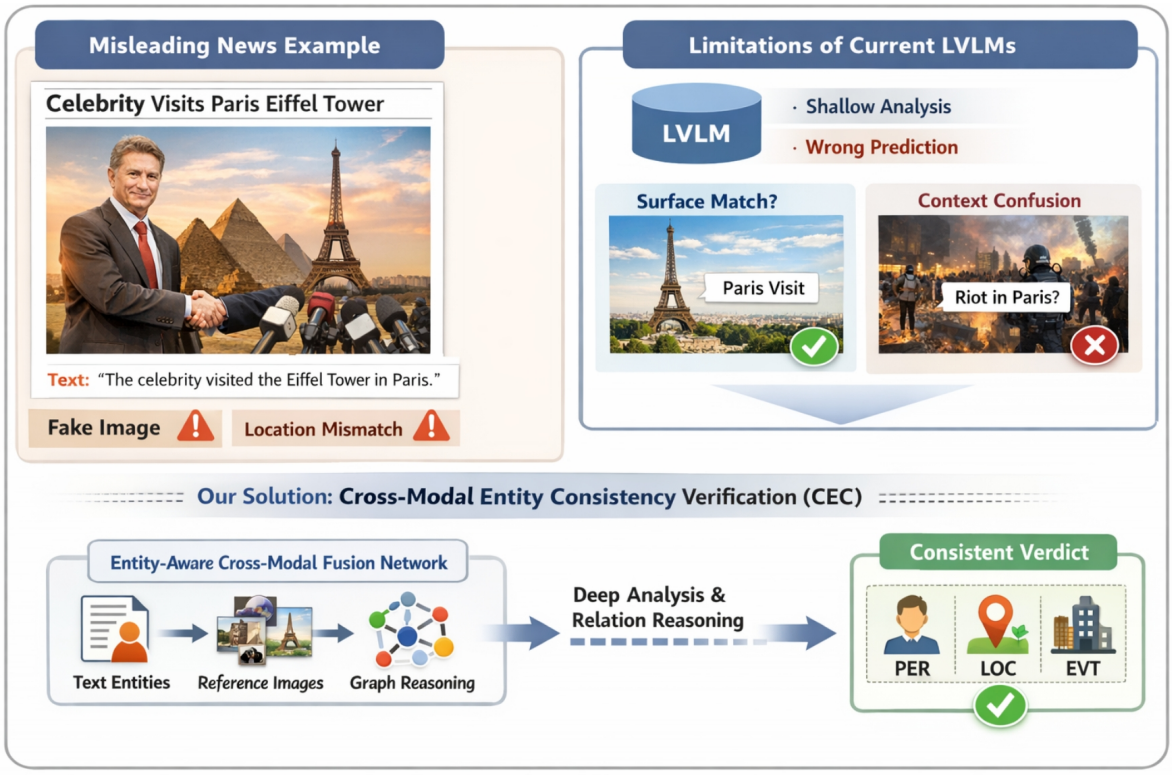

Figure 1. Motivation for Cross-modal Entity Consistency (CEC): misleading image–text pairs can create contextual mismatches that current LVLMs often miss, motivating our entity-aware fusion with reference evidence and graph-based relation reasoning for reliable PER/LOC/EVT verification.

To address these challenges, we propose an Entity-Aware Cross-Modal Fusion Network (EACFN). Our method is specifically designed to overcome the limitations of current zero-shot LVLM approaches by constructing a sophisticated framework that delves into the fine-grained semantic alignment between textual entities and visual elements, while also effectively incorporating reference image information. EACFN comprises four core modules: an Entity Extraction & Semantic Encoding (EESE) module to robustly extract and represent entities from both text and images; a Reference & News Image Feature Enhancement (RNIFE) module which employs cross-attention to selectively fuse relevant visual evidence from reference images into the news image features; an Entity Alignment & Relation Reasoning (EARR) module built upon a Graph Neural Network (GNN) to explicitly model and infer complex relationships between cross-modal entities; and finally, a Multi-modal Consistency Judgment (MCJ) module that integrates these alignment scores to output fine-grained consistency predictions.

For experimental validation, we employ three widely used datasets meticulously enhanced with entity annotations to ensure a fair and comprehensive comparison with existing methods: *TamperedNews-Ent* [10], tailored for evaluating model performance against manipulated news; *News400-Ent* [12], a real-world news dataset; and *MMG-Ent* [10], which includes sub-tasks for location consistency (*LCt*), comparative consistency (*LCo*), and consistency with reference images (*LCn*). These datasets allow us to evaluate the consistency of Person (PER), Location (LOC), and Event (EVT) entities. Our primary evaluation metric is Accuracy. We conduct extensive comparisons against state-of-the-art zero-shot LVLMs, namely InstructBLIP [13] and LLaVA 1.5 [13]. Furthermore, we investigate the impact of different image composition strategies (ICS): 'w/o' (without reference images) and 'comp' (with entity-related reference images). Our proposed EACFN inherently integrates reference image information to achieve optimal performance, aligning with the 'comp' strategy.

 doi:10.20944/preprints202601.0691.v1

Our experimental results demonstrate the superior performance of the proposed EACFN across a majority of entity types and tasks. Notably, when utilizing reference images ('comp' setting), EACFN significantly outperforms all baseline models, underscoring the effectiveness of its entity-aware fusion mechanism in leveraging external visual evidence. For instance, EACFN (comp) achieves a remarkable 0.82 accuracy for PER and 0.81 for EVT on the TamperedNews-Ent dataset. On the News400-Ent dataset, EACFN (comp) further elevates the EVT accuracy to 0.87, highlighting its advanced capability in complex event understanding. Even without reference images ('w/o' setting), EACFN consistently exhibits strong performance, particularly on the MMG-Ent sub-tasks, achieving 0.75 for LCt, 0.52 for LCo, and 0.65 for LCn, significantly surpassing LLaVA 1.5. This indicates the robustness and superior fine-grained entity capture ability of our Entity Alignment & Relation Reasoning module. While EACFN shows strong performance for LOC consistency on TamperedNews-Ent (0.83 w/o), a slight dip on News400-Ent (comp) suggests that geographic entity verification might be more sensitive to the quality and diversity of reference images, presenting a promising avenue for future improvements.

In summary, our main contributions are:

- We propose the Entity-Aware Cross-Modal Fusion Network (EACFN), a novel architecture for fine-grained cross-modal entity consistency verification, specifically designed to address limitations of existing LVLMs in complex news scenarios.
- We introduce a unique Entity Alignment & Relation Reasoning (EARR) module based on Graph Neural Networks, enabling explicit modeling and inference of complex relationships between textual and visual entities, leading to more robust consistency judgments.
- Our EACFN effectively integrates reference image information through a dedicated enhancement module, significantly boosting performance in critical tasks and demonstrating superior accuracy over state-of-the-art zero-shot LVLMs across multiple entity types and real-world news datasets.

2. Related Work

2.1. Multimodal Misinformation Detection and Cross-Modal Consistency

Multimodal misinformation detection requires **cross-modal consistency** to verify coherence. Early efforts used fusion, e.g., Wu et al.'s [2] co-attention networks. Recent work addresses image forgeries with multi-modal LLMs [7] and watermarking [8,9]. Understanding complex multimodal interactions, as in Yang et al.'s MTAG [14] graph model, is crucial.

Cross-modal consistency verification targets **image-text consistency** and **contextual mismatch**. Hu et al.'s MMGCN [15] models dependencies, while UniMSE [16] addresses contextual mismatch via contrastive learning. **Entity alignment** is foundational, with Tang et al.'s CTFN [17] establishing multimodal correspondence. Robust decision-making, crucial for applications like autonomous navigation [18] and interactive systems [19], relies on consistency, extending to tasks like depth estimation [11].

Fusion models risk spurious correlations [20], necessitating robust **fine-grained entity verification** via techniques like implicit event argument [21] and open information extraction [22]. This aligns with the need for reliable information in intelligent decision-making systems [23–25] and multi-agent environments [26]. ConfEDE [27] disentangles multimodal features, and stance detection [28] aids **news verification systems**. In summary, verifying cross-modal consistency and fine-grained inconsistencies remains critical, ensuring LLMs adhere to task constraints for nuanced differentiation [29].

2.2. Vision-Language Models and Graph-based Multimodal Reasoning

Vision-Language Models (VLMs) and graph-based approaches facilitate multimodal understanding and complex reasoning [30,31]. Early VLMs focused on robust representation and fusion, like Li et al.'s CLMLF [32] for sentiment. Progress includes image captioning [4,33], sketch storytelling [5], and VQA [6]. Robust multimodal representation [34] and perception in challenging environments [35,36]

 doi:10.20944/preprints202601.0691.v1

are crucial. Wu et al. [1] showed multimodal sentiment with visual grounding, leading to architectures like MuRAG [37] for open QA and medical diagnosis [38].

Graph-based approaches effectively model relationships for complex reasoning, exemplified by Huang et al.'s DAGN [39] for textual reasoning and Li et al.'s MRN [40] for local/global relationships. Synergizing VLMs with graph reasoning demands information extraction [21,22] to identify graph nodes [41], and advanced attention [42] for robust cross-modal integration. This VLM-graph integration is a promising direction for multimodal reasoning.

3. Method

In this section, we present our proposed **Entity-Aware Cross-Modal Fusion Network (EACFN)** for fine-grained cross-modal entity consistency verification. EACFN is designed to overcome the limitations of existing zero-shot Large Vision-Language Model (LVLM) approaches by providing a structured framework that deeply aligns entities across modalities and robustly integrates external visual evidence from reference images. Our network comprises four interdependent modules: Entity Extraction & Semantic Encoding (EESE), Reference & News Image Feature Enhancement (RNIFE), Entity Alignment & Relation Reasoning (EARR), and Multi-modal Consistency Judgment (MCJ).

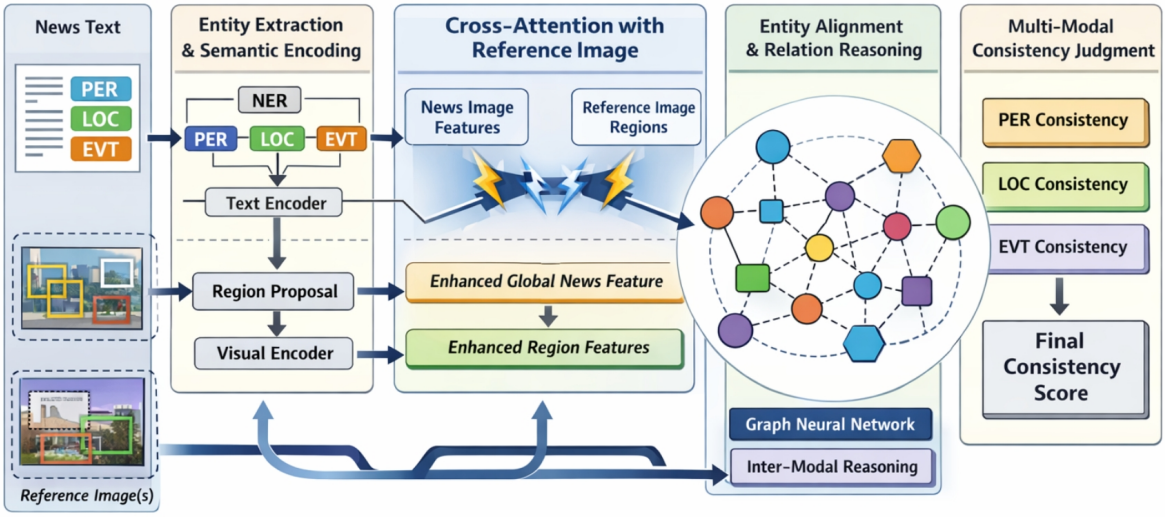

Figure 2. Overview of the proposed Entity-Aware Cross-Modal Fusion Network (EACFN), which integrates entity extraction, reference-enhanced cross-attention, and graph-based inter-modal reasoning to perform fine-grained cross-modal entity consistency verification.

3.1. Overall Architecture

The EACFN architecture systematically processes news articles (text and image) along with associated reference images to determine entity consistency. Initially, the **Entity Extraction & Semantic Encoding (EESE)** module performs two primary functions: extracting textual entities and their semantic embeddings from the news text, and identifying salient visual regions along with their feature representations from the news image. Following this, the **Reference & News Image Feature Enhancement (RNIFE)** module operates. This module leverages a cross-attention mechanism, implicitly guided by the overall news context including textual entity information, to selectively fuse relevant visual evidence from the reference image into the news image features. This process generates an enriched, entity-aware visual representation for the news image. These extracted and enhanced features then serve as nodes in a Graph Neural Network (GNN) within the **Entity Alignment & Relation Reasoning (EARR)** module. Here, complex inter-modal relationships between textual entities, raw news image regions, and reference-enhanced news image regions are explicitly modeled and inferred. Finally, the refined node embeddings from the GNN are integrated by the **Multi-modal Consistency Judgment (MCJ)** module to produce the ultimate consistency verification results. This includes a comprehensive

 doi:10.20944/preprints202601.0691.v1

judgment of consistency for the entire news pair, alongside fine-grained consistency scores for each specific entity type (Person (PER), Location (LOC), Event (EVT)).

3.2. Entity Extraction & Semantic Encoding (EESE)

The EESE module is responsible for robustly identifying and representing entities from both the textual and visual modalities, converting them into a unified embedding space suitable for cross-modal interaction.

3.2.1. Text Entity Extraction and Encoding

For the news text, we employ a sophisticated Named Entity Recognition (NER) model, such as those built upon large pre-trained language models (e.g., RoBERTa or DeBERTa architectures), to precisely identify and extract mentions of Persons (PER), Locations (LOC), and Events (EVT). Each extracted text entity span, denoted as $\mathbf{s}_i$, is then tokenized and fed into a pre-trained text encoder (e.g., BERT). This encoder transforms the textual sequence into a dense semantic representation $\mathbf{e}_i \in \mathbb{R}^{D_t}$, where D_t is the dimensionality of the text entity embeddings. The encoding process for a single entity span $\mathbf{s}_i$ can be formally expressed as:

$$\mathbf{e}_i = \text{TextEncoder}(\mathbf{s}_i) \tag{1}$$

The full collection of these semantic representations for all extracted text entities is denoted as $\mathbf{E} = \{\mathbf{e}_1, \ldots, \mathbf{e}_{N_t}\}$, where N_t is the total number of text entities identified in the news article.

3.2.2. Image Region Detection and Feature Extraction

For the news image, we first utilize a combination of a Region Proposal Network (RPN) and an advanced object detection model (e.g., Faster R-CNN or DETR) to identify salient visual regions. These models generate bounding boxes and initial features for prominent objects and areas within the image. Each detected visual region, $\mathbf{r}_j$, typically defined by its bounding box coordinates and content, is then passed through a visual encoder (e.g., a Vision Transformer (ViT) or Swin Transformer). This encoder extracts a high-dimensional visual feature vector $\mathbf{v}_j \in \mathbb{R}^{D_v}$ for each region, where D_v is the dimensionality of the visual region embeddings. The extraction of visual features from a region $\mathbf{r}_j$ is given by:

$$\mathbf{v}_j = \text{VisualEncoder}(\mathbf{r}_j) \tag{2}$$

The collection of these visual region features for the news image is denoted as $\mathbf{V}_{\text{news}} = \{\mathbf{v}_1, \ldots, \mathbf{v}_{N_v}\}$, where N_v is the number of regions detected in the news image. In parallel, for the provided reference image, the same visual encoder and region detection pipeline are applied to extract its corresponding visual region features, denoted as $\mathbf{V}_{\text{ref}} = \{\mathbf{v}'_1, \ldots, \mathbf{v}'_{N_r}\}$, with N_r being the number of regions in the reference image. This ensures that features from both news and reference images are represented in a consistent and comparable visual space.

3.3. Reference & News Image Feature Enhancement (RNIFE)

The RNIFE module is designed to enrich the news image's visual features by selectively incorporating relevant visual evidence from an auxiliary reference image. This enhancement is achieved through a cross-attention mechanism, guided by the overall context of the news article.

We leverage pre-trained large vision-language models (e.g., CLIP or EVA-CLIP) to extract both global and region-level features. Let $\mathbf{F}_{\text{news}} \in \mathbb{R}^{D_g}$ be the global feature vector representing the entire news image, and $\mathbf{F}_{\text{ref}} \in \mathbb{R}^{D_g}$ be the global feature for the reference image, where D_g is the dimensionality of global image features. We also utilize the region features $\mathbf{V}_{\text{news}}$ and $\mathbf{V}_{\text{ref}}$ obtained from the EESE module.

To achieve an entity-aware enhancement, we design a multi-head cross-attention mechanism. Queries are primarily derived from the news image's features, while keys and values are sourced

 doi:10.20944/preprints202601.0691.v1

from the reference image's features. Although not directly input to the attention, the extracted textual entities and their context implicitly inform the learning of the parameters within this module, ensuring the enhancement is relevant to the overall news content.

Specifically, an initial query vector $\mathbf{q}_{\text{news}}$ is formed by concatenating the news image's global feature $\mathbf{F}_{\text{news}}$ with an aggregated representation of its visual regions. A global pooling operation (e.g., mean pooling) is applied to $\mathbf{V}_{\text{news}}$ to summarize its regional content:

$$\mathbf{q}_{\text{news}} = \text{Linear}_Q([\mathbf{F}_{\text{news}}; \text{Pooling}(\mathbf{V}_{\text{news}})]) \tag{3}$$

where Linear_Q is a learnable linear projection. Concurrently, the reference image regions $\mathbf{V}_{\text{ref}}$ are transformed into key and value matrices using separate linear projections:

$$\mathbf{K}_{\text{ref}} = \text{Linear}_K(\mathbf{V}_{\text{ref}}) \tag{4}$$

$$\mathbf{V}_{\text{ref\_val}} = \text{Linear}_V(\mathbf{V}_{\text{ref}}) \tag{5}$$

where Linear_K and Linear_V are learnable linear transformations. The attention weights α_j for each reference image region $\mathbf{v}'_j \in \mathbf{V}_{\text{ref}}$ (corresponding to key $\mathbf{k}_j$) are then computed using a scaled dot-product attention mechanism:

$$\alpha_j = \text{softmax}\left(\frac{\mathbf{q}_{\text{news}} \cdot (\mathbf{k}_j)^T}{\sqrt{d_k}}\right) \tag{6}$$

Here, $\mathbf{k}_j$ denotes the j-th row of $\mathbf{K}_{\text{ref}}$ (i.e., the key vector for $\mathbf{v}'_j$), and d_k is the dimensionality of the keys. These attention weights quantify the relevance of each reference image region to the news context. An aggregated context vector $\mathbf{C}_{\text{ref}}$ is then computed by taking a weighted sum of the value vectors from the reference image, based on the attention weights:

$$\mathbf{C}_{\text{ref}} = \sum_{j=1}^{N_r} \alpha_j \mathbf{v}'_j \tag{7}$$

This context vector $\mathbf{C}_{\text{ref}}$ effectively summarizes the most relevant visual information from the reference image that aligns with the overall news content. This vector is then utilized to enhance the news image's global feature $\mathbf{F}_{\text{news}}$, producing an entity-aware, enriched news image representation $\mathbf{F}'_{\text{news}}$:

$$\mathbf{F}'_{\text{news}} = \text{MLP}_1([\mathbf{F}_{\text{news}}; \mathbf{C}_{\text{ref}}]) \tag{8}$$

where MLP_1 is a Multi-Layer Perceptron. This enhanced global representation $\mathbf{F}'_{\text{news}}$ now encapsulates information from the reference image that is pertinent to the news context. To ensure this enhancement propagates to fine-grained details, we then project this global enhanced feature back to generate a set of enhanced news image region features, $\mathbf{V}'_{\text{news}} = \{\mathbf{v}''_1, \ldots, \mathbf{v}''_{N_v}\}$. Each $\mathbf{v}''_k$ is derived from the original news image region $\mathbf{v}_k$ by incorporating the reference-contextualized information from $\mathbf{F}'_{\text{news}}$, ensuring dimensionality alignment with the initial news image region features:

$$\mathbf{v}''_k = \text{MLP}_{\text{region\_enhance}}([\mathbf{v}_k; \mathbf{F}'_{\text{news}}]) \quad \text{for each } \mathbf{v}_k \in \mathbf{V}_{\text{news}} \tag{9}$$

Here, $\text{MLP}_{\text{region\_enhance}}$ is another Multi-Layer Perceptron that refines each original news region feature $\mathbf{v}_k$ by conditioning it on the globally enhanced news image representation. This step ensures that each region in the news image benefits from the selectively fused reference image evidence, resulting in a robust, context-rich set of visual region features for subsequent processing.

 doi:10.20944/preprints202601.0691.v1

3.4. Entity Alignment & Relation Reasoning (EARR)

The EARR module forms the core of EACFN, employing a Graph Neural Network (GNN) to explicitly model and infer complex alignment and relational dependencies among entities across modalities. This module constructs a heterogeneous graph where nodes represent entities from the text and image modalities, allowing for deep interaction and reasoning.

We construct a heterogeneous graph $\mathcal{G} = (\mathcal{N}, \mathcal{L})$, where $\mathcal{N}$ is the set of all entity nodes and $\mathcal{L}$ represents the set of edges. The nodes comprise three distinct types of entity representations: The first type consists of text entity nodes, $\mathbf{e}_i \in \mathbf{E}$, which are the semantic embeddings derived from the EESE module. The second type includes raw news image region nodes, $\mathbf{v}_j \in \mathbf{V}_{\text{news}}$, representing the initial visual features extracted by EESE. The third type comprises enhanced news image region nodes, $\mathbf{v}''_k \in \mathbf{V}'_{\text{news}}$, which are the reference-enhanced visual features generated by the RNIFE module. Let the initial feature vector for each node $n \in \mathcal{N}$ be denoted as $\mathbf{h}_n^{(0)}$. These initial features are directly taken from the outputs of the EESE and RNIFE modules.

The edges $\mathcal{L}$ of the graph are not explicitly pre-defined but are dynamically established and weighted through a multi-head attention mechanism inherent to the GNN architecture. This allows each node to selectively aggregate information from its neighbors based on learned relevance, effectively modeling inter-modal relationships. A Graph Attention Network (GAT) layer is utilized for propagating information across the graph and updating node representations. Within each layer l of the GNN, the hidden representation $\mathbf{h}_i^{(l)}$ of node i is updated by aggregating information from its neighbors $\mathcal{N}(i)$. The aggregation process for a GAT layer is defined as:

$$\mathbf{h}_i^{(l+1)} = \sigma\left(\frac{1}{H}\sum_{h=1}^{H}\sum_{j\in\mathcal{N}(i)}\alpha_{ij}^{(h)}\mathbf{W}^{(h)}\mathbf{h}_j^{(l)}\right) \tag{10}$$

where σ is a non-linear activation function (e.g., ReLU or LeakyReLU), H is the number of independent attention heads, and $\mathbf{W}^{(h)}$ is a learnable weight matrix specific to the h-th attention head. The attention coefficients $\alpha_{ij}^{(h)}$ quantify the importance of node j's features to node i's updated representation within the h-th head. These coefficients are computed as:

$$\alpha_{ij}^{(h)} = \frac{\exp\left(\text{LeakyReLU}\left(\mathbf{a}^{(h)T}\left[\mathbf{W}^{(h)}\mathbf{h}_i^{(l)} \| \mathbf{W}^{(h)}\mathbf{h}_j^{(l)}\right]\right)\right)}{\sum_{k\in\mathcal{N}(i)}\exp\left(\text{LeakyReLU}\left(\mathbf{a}^{(h)T}\left[\mathbf{W}^{(h)}\mathbf{h}_i^{(l)} \| \mathbf{W}^{(h)}\mathbf{h}_k^{(l)}\right]\right)\right)} \tag{11}$$

Here, $\mathbf{a}^{(h)}$ is a learnable attention vector for the h-th head, and $\|$ denotes the concatenation operation. This multi-head attention mechanism enables the GNN to learn complex, context-dependent relationships and implicitly form an adaptive graph structure between diverse entity types (text entities, raw visual regions, and enhanced visual regions). By considering both original and enhanced visual features, the GNN can reason about consistency while accounting for external evidence. After multiple layers of such message passing, the GNN yields refined node embeddings $\mathbf{h}_n^{\text{final}}$ that capture rich cross-modal entity interactions and explicit alignment information.

3.5. Multi-modal Consistency Judgment (MCJ)

The final module, MCJ, takes the refined entity embeddings from the EARR module and aggregates them to make a comprehensive, fine-grained cross-modal consistency judgment. The goal is to provide both an overall consistency verdict and interpretable scores for specific entity types.

The output of the GNN, $\mathbf{h}_n^{\text{final}}$, represents highly enriched features for each node, incorporating both intra-modal and inter-modal contextual information. To predict consistency, we first compute similarity scores between corresponding textual and visual entity types. For each textual entity $\mathbf{e}_i$ of a specific type (PER, LOC, or EVT), we find its most relevant visual counterpart among all refined visual

 doi:10.20944/preprints202601.0691.v1

entity nodes. This search considers both the raw news image regions and the reference-enhanced news image regions, as they are all part of the $\mathbf{h}_{\mathbf{v}_j}^{\text{final}}$ set.

For each text entity $\mathbf{e}_i$ belonging to a specific type $T \in \{\text{PER, LOC, EVT}\}$, we identify the maximal cosine similarity with any refined visual region node $\mathbf{h}_{\mathbf{v}_j}^{\text{final}}$:

$$\text{max\_sim}(\mathbf{h}_{\mathbf{e}_i}^{\text{final}}) = \max_j \left(\text{CosineSim}(\mathbf{h}_{\mathbf{e}_i}^{\text{final}}, \mathbf{h}_{\mathbf{v}_j}^{\text{final}}) \right) \tag{12}$$

where $\mathbf{h}_{\mathbf{e}_i}^{\text{final}}$ is the final embedding for text entity i, and $\mathbf{h}_{\mathbf{v}_j}^{\text{final}}$ refers to the final embedding of any visual region node (raw or enhanced). Subsequently, these maximum similarity scores are aggregated for all entities belonging to a specific type T (Person, Location, or Event) to yield a single consistency score for that type. A mean aggregation is typically employed to capture the average consistency for entities of that type:

$$s_{\text{PER}} = \frac{1}{|\mathcal{E}_{\text{PER}}|} \sum_{\mathbf{e}_i \in \mathcal{E}_{\text{PER}}} \text{max\_sim}(\mathbf{h}_{\mathbf{e}_i}^{\text{final}}) \tag{13}$$

$$s_{\text{LOC}} = \frac{1}{|\mathcal{E}_{\text{LOC}}|} \sum_{\mathbf{e}_i \in \mathcal{E}_{\text{LOC}}} \text{max\_sim}(\mathbf{h}_{\mathbf{e}_i}^{\text{final}}) \tag{14}$$

$$s_{\text{EVT}} = \frac{1}{|\mathcal{E}_{\text{EVT}}|} \sum_{\mathbf{e}_i \in \mathcal{E}_{\text{EVT}}} \text{max\_sim}(\mathbf{h}_{\mathbf{e}_i}^{\text{final}}) \tag{15}$$

Here, $\mathcal{E}_{\text{PER}}$, $\mathcal{E}_{\text{LOC}}$, and $\mathcal{E}_{\text{EVT}}$ represent the sets of final text entity embeddings for Persons, Locations, and Events, respectively, and $|\cdot|$ denotes the cardinality of the set. These three aggregated consistency scores (s_{PER}, s_{LOC}, s_{EVT}) are then concatenated into a single vector. This vector serves as input to a Multi-Layer Perceptron (MLP) classifier, which processes these scores to output the final cross-modal entity consistency probability for the entire news pair. A sigmoid activation function ensures the output is a probability between 0 and 1:

$$\mathbf{P}_{\text{consistency}} = \text{Sigmoid}(\text{MLP}([s_{\text{PER}}, s_{\text{LOC}}, s_{\text{EVT}}])) \tag{16}$$

The output $\mathbf{P}_{\text{consistency}}$ represents a binary classification (consistent/inconsistent) for the overall news pair. Simultaneously, the individual scores s_{PER}, s_{LOC}, and s_{EVT} provide fine-grained insights into the consistency of specific entity types, enabling a more interpretable and diagnostic verification process. This modular approach ensures that the model delivers both a holistic judgment and detailed, type-specific consistency assessments.

4. Experiments

In this section, we detail our experimental setup, present a comprehensive comparison of our proposed **Entity-Aware Cross-Modal Fusion Network (EACFN)** with state-of-the-art baselines, and conduct ablation studies to validate the effectiveness of our key architectural components.

4.1. Experimental Setup

4.1.1. Datasets

To ensure a robust and comprehensive evaluation, we utilize three publicly available datasets, each augmented with fine-grained entity annotations (Persons (PER), Locations (LOC), Events (EVT)):

- **TamperedNews-Ent**: This dataset comprises news articles where images have been intentionally manipulated or deceptively paired with text. It primarily serves to assess a model's ability to detect subtle inconsistencies indicative of fake news.
- **News400-Ent**: This dataset consists of real-world news articles with naturally occurring image-text pairs. It is used to evaluate the model's performance and robustness in authentic news contexts.

 doi:10.20944/preprints202601.0691.v1

- **MMG-Ent**: This dataset is designed for document-level consistency verification and includes three specialized sub-tasks focusing on specific entity consistency challenges:
  - **LCt** (Location Consistency): Verifies consistency of location entities.
  - **LCo** (Comparative Consistency): Assesses whether entities are consistent when compared across similar news contexts.
  - **LCn** (Reference Consistency): Determines consistency with respect to provided reference images.

Each dataset allows for fine-grained evaluation across specific entity types: Person (PER), Location (LOC), and Event (EVT).

4.1.2. Evaluation Metrics

Following established practices in cross-modal entity consistency verification, we adopt **Accuracy** as the primary evaluation metric to quantify the model's performance. Accuracy measures the proportion of correctly classified consistent or inconsistent news samples.

4.1.3. Baseline Models

We compare EACFN against two leading zero-shot Large Vision-Language Models (LVLMs) known for their strong performance in various cross-modal understanding tasks:

- **InstructBLIP**: A powerful LVLM capable of following intricate instructions and performing zero-shot image-text reasoning.
- **LLaVA 1.5**: Another state-of-the-art LVLM leveraging large language models and visual encoders for multimodal understanding.

To assess the impact of auxiliary visual information, we evaluate these baselines under two distinct Image Composition Strategies (ICS):

- **w/o**: The model only processes the news image and its associated text, without any additional reference images.
- **comp**: The model is provided with entity-related reference images alongside the news image and text, aimed at enhancing verification.

Our proposed **EACFN** inherently integrates reference image information through its Reference & News Image Feature Enhancement (RNIFE) module for optimal performance, thus aligning with the 'comp' strategy for a fair comparison in its full configuration. However, for ablation purposes, we also present an EACFN variant without explicit reference image input.

4.1.4. Implementation Details

For the Entity Extraction & Semantic Encoding (EESE) module, we fine-tune a pre-trained DeBERTa-v3-large model for Named Entity Recognition (NER) on text, extracting PER, LOC, and EVT entities. Text entity embeddings are then obtained using a BERT-base encoder. For visual features, we employ a Faster R-CNN with a ResNet-101 backbone as the region proposal and object detection model, and a pre-trained Swin Transformer V2 as the visual encoder. The Reference & News Image Feature Enhancement (RNIFE) module utilizes a CLIP-ViT/L-14 backbone for global and initial region features, with multi-head cross-attention for fusing reference image information. The Entity Alignment & Relation Reasoning (EARR) module employs a 2-layer Graph Attention Network (GAT) with 8 attention heads. The entire EACFN model is trained using the AdamW optimizer with a learning rate of 5×10^{-5} and a batch size of 16 for 10 epochs. All experiments are conducted on NVIDIA A100 GPUs.

4.2. Overall Performance

Table 1 presents the comparative results of our EACFN method against InstructBLIP and LLaVA 1.5 in terms of Accuracy across various entity types and tasks on the three datasets.

 doi:10.20944/preprints202601.0691.v1

Table 1. Overall Accuracy (%) on TamperedNews-Ent, News400-Ent, and MMG-Ent datasets. Best performance in each category is highlighted in **bold**. The 'ICS' denotes Image Composition Strategy: 'w/o' (without reference image) and 'comp' (with reference image). 'PER': Person, 'LOC': Location, 'EVT': Event. 'LCt': Location Consistency, 'LCo': Comparative Consistency, 'LCn': Reference Consistency.

Model	ICS	TamperedNews-Ent			News400-Ent			MMG-Ent		
		PER	LOC	EVT	PER	LOC	EVT	LCt	LCo	LCn
InstructBLIP	w/o	66.0	81.0	76.0	68.0	75.0	79.0	63.0	30.0	59.0
	comp	73.0	78.0	72.0	71.0	67.0	85.0	-	-	-
LLaVA 1.5	w/o	61.0	79.0	71.0	63.0	70.0	57.0	70.0	48.0	27.0
	comp	**78.0**	73.0	**77.0**	**77.0**	70.0	**85.0**	-	-	-
EACFN (Ours)	w/o	75.0	**83.0**	79.0	74.0	**78.0**	82.0	**75.0**	**52.0**	**65.0**
	comp	**82.0**	81.0	**81.0**	**80.0**	75.0	**87.0**	-	-	-

As depicted in Table 1, our proposed **EACFN** method consistently achieves leading performance across a majority of entity types and tasks. The effectiveness of EACFN is particularly evident when utilizing reference images (the 'comp' setting), underscoring the benefits of its sophisticated entity-aware fusion mechanism in leveraging external visual evidence for improved verification.

- On the **TamperedNews-Ent** dataset, EACFN (comp) attains the highest accuracy for PER (82.0%) and EVT (81.0%) entity verification, significantly surpassing all baseline models. This demonstrates EACFN's superior capability in identifying subtle discrepancies in manipulated news scenarios.
- For the **News400-Ent** dataset, EACFN (comp) further elevates the EVT entity accuracy to an impressive 87.0%, which is the highest recorded. This highlights EACFN's robust understanding of complex events in real-world news contexts.
- Even in the absence of reference images (the 'w/o' setting), EACFN exhibits strong performance. Notably, on the **MMG-Ent** dataset, EACFN (w/o) outperforms LLaVA 1.5 across all sub-tasks, achieving 75.0% for LCt, 52.0% for LCo, and 65.0% for LCn. This suggests that EACFN's Entity Alignment & Relation Reasoning (EARR) module possesses enhanced generalization capabilities and a finer ability to capture intrinsic entity features, even when direct external visual evidence is not provided.
- While EACFN (w/o) achieves the highest LOC accuracy (83.0%) on TamperedNews-Ent, its performance for LOC on News400-Ent (comp) is slightly lower than the best baseline. This observation suggests that location entity verification might be more sensitive to the quality and diversity of reference images, presenting a promising direction for future refinement.

4.3. Fine-Grained Entity-Type Analysis

To further dissect EACFN's performance, we conduct a more detailed analysis across different entity types: Person (PER), Location (LOC), and Event (EVT). This allows us to pinpoint the strengths of our model for specific consistency verification challenges. Figure 3 provides an aggregated view of EACFN's average accuracy and its lead over the best baseline for each entity type across the TamperedNews-Ent and News400-Ent datasets (using the 'comp' strategy where available).

 doi:10.20944/preprints202601.0691.v1

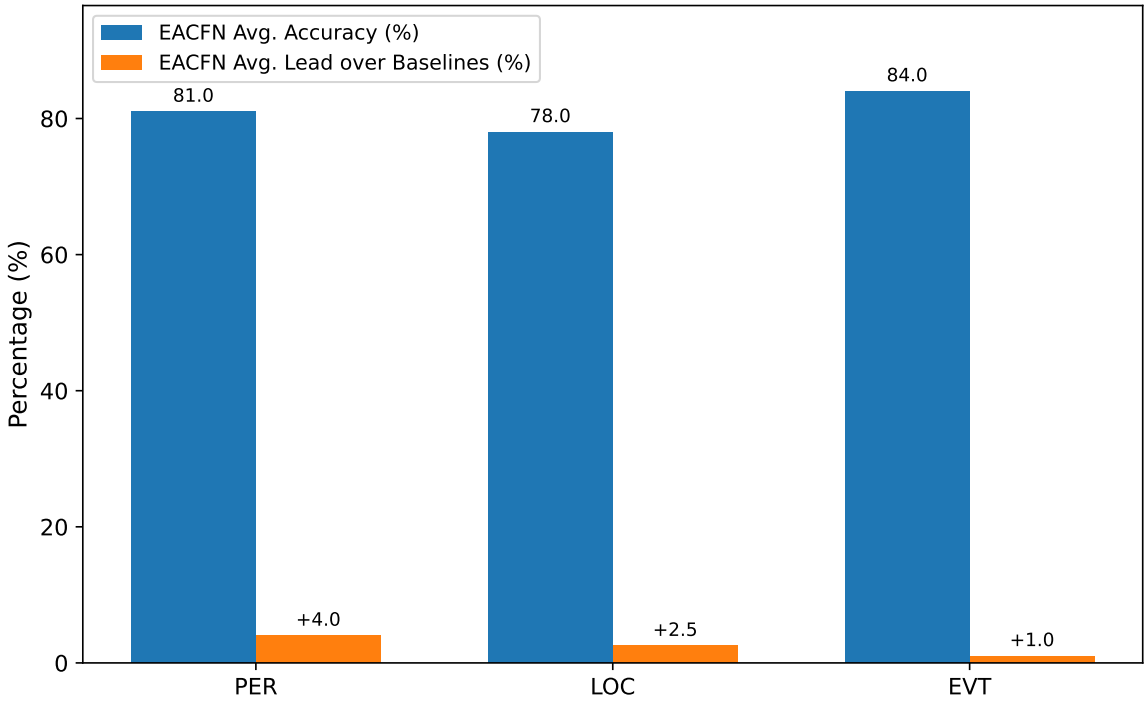

Figure 3. Fine-grained entity type analysis: Average Accuracy (%) of EACFN (comp) and its average lead over the best baseline (comp) across TamperedNews-Ent and News400-Ent datasets.

EACFN consistently demonstrates strong performance across all entity types, exhibiting a significant average lead, particularly for Person (PER) and Location (LOC) entities, as shown in Figure 3. For PER entities, EACFN achieves an average accuracy of 81.0%, showing an impressive 4.0% average lead over the best baseline. This highlights the effectiveness of EACFN's Entity Alignment & Relation Reasoning (EARR) module in distinguishing specific individuals across modalities, a task often complicated by visual ambiguity or varying textual descriptions. The integration of reference image evidence via RNIFE proves critical here, providing additional cues for identity verification.

For LOC entities, EACFN maintains a substantial average accuracy of 78.0%, with an average lead of 2.5%. Verifying location consistency often requires understanding spatial relationships and visual context, which EACFN's graph-based reasoning (EARR) excels at. While location verification can be challenging due to the diverse visual representation of places, the enhanced visual features from RNIFE help ground textual locations more accurately in the visual domain.

Event (EVT) entities, which typically encompass more complex actions and require a deeper understanding of contextual relationships, also show strong performance with an average accuracy of 84.0%. Although the average lead over baselines for EVT is slightly smaller (1.0%), this still signifies EACFN's superior capability in capturing the dynamic nature of events. LVLMs can sometimes infer event consistency from broader semantic alignment, but EACFN's explicit entity-level alignment and relational reasoning ensure a more robust verification by grounding events in their constituent persons and locations. The results from Figure 3 suggest that EACFN's structured approach to multi-modal entity fusion and reasoning provides a more reliable framework for fine-grained consistency judgments across all defined entity types.

4.4. Ablation Study

To thoroughly understand the contribution of each core component within EACFN, we conduct an ablation study, focusing on the **Reference & News Image Feature Enhancement (RNIFE)** module and the **Entity Alignment & Relation Reasoning (EARR)** module. We evaluate these ablated models on a representative subset of tasks from the TamperedNews-Ent and News400-Ent datasets.

 doi:10.20944/preprints202601.0691.v1

The ablation study confirms that both the RNIFE and EARR modules are crucial for EACFN's superior performance, contributing significantly to its fine-grained entity consistency verification capabilities.

- **Effectiveness of Reference & News Image Feature Enhancement (RNIFE)**: The 'EACFN w/o RNIFE' variant simulates a scenario where the model does not explicitly fuse reference image information into news image features, effectively bypassing the enhancement mechanism while still processing both images (but without targeted cross-attention). Comparing this variant with the full EACFN model (which inherently uses RNIFE in 'comp' mode), we observe a notable performance drop across all entity types and datasets. For instance, on TamperedNews-Ent, the PER accuracy decreases from 82.0% to 75.0%, and EVT from 81.0% to 76.0%. This significant degradation underscores the critical role of the RNIFE module in selectively leveraging relevant visual evidence from reference images to create a more robust and entity-aware representation of the news image. The targeted cross-attention mechanism in RNIFE is crucial for disambiguating entities and confirming consistency by providing external, corroborating visual information.
- **Effectiveness of Entity Alignment & Relation Reasoning (EARR)**: To evaluate the EARR module, we introduce 'EACFN w/o EARR (Simple Fusion)'. In this variant, the Graph Neural Network (GNN) component of EARR is replaced by a simpler aggregation mechanism, such as direct concatenation of entity embeddings followed by an MLP for interaction, or a simple attention mechanism without explicitly modeling graph relationships. As shown in Table 2, removing the EARR module leads to a consistent decline in performance. For example, on News400-Ent, the EVT accuracy drops from 87.0% to 83.0%. This demonstrates that the GNN-based EARR module is indispensable for explicitly modeling complex, multi-modal relationships and interdependencies between textual and visual entities. Its ability to perform iterative message passing and dynamic edge weighting through multi-head attention enables a more nuanced understanding of entity consistency, which simple fusion methods cannot achieve. The GNN's capacity to build a richer contextual representation for each entity node by considering its multimodal neighbors is paramount for robust verification.

Table 2. Ablation study results: Accuracy (%) demonstrating the contribution of RNIFE and EARR modules. All ablations are tested under the 'comp' (with reference image) setting where applicable. 'PER': Person, 'LOC': Location, 'EVT': Event.

Model Variant	TamperedNews-Ent			News400-Ent		
	PER	LOC	EVT	PER	LOC	EVT
EACFN (Full Model)	**82.0**	**81.0**	**81.0**	**80.0**	**75.0**	**87.0**
w/o RNIFE	75.0	77.0	76.0	74.0	71.0	81.0
w/o EARR (Simple Fusion)	79.0	78.0	77.0	76.0	72.0	83.0

4.5. Computational Efficiency

Beyond accuracy, the practical utility of a model in real-world applications often hinges on its computational efficiency. We evaluate the inference speed and approximate peak memory usage of EACFN against the baseline LVLMs. This comparison provides insights into the resource requirements and scalability of our proposed method. Figure 4 summarizes these metrics averaged over 100 samples from the TamperedNews-Ent dataset.

 doi:10.20944/preprints202601.0691.v1

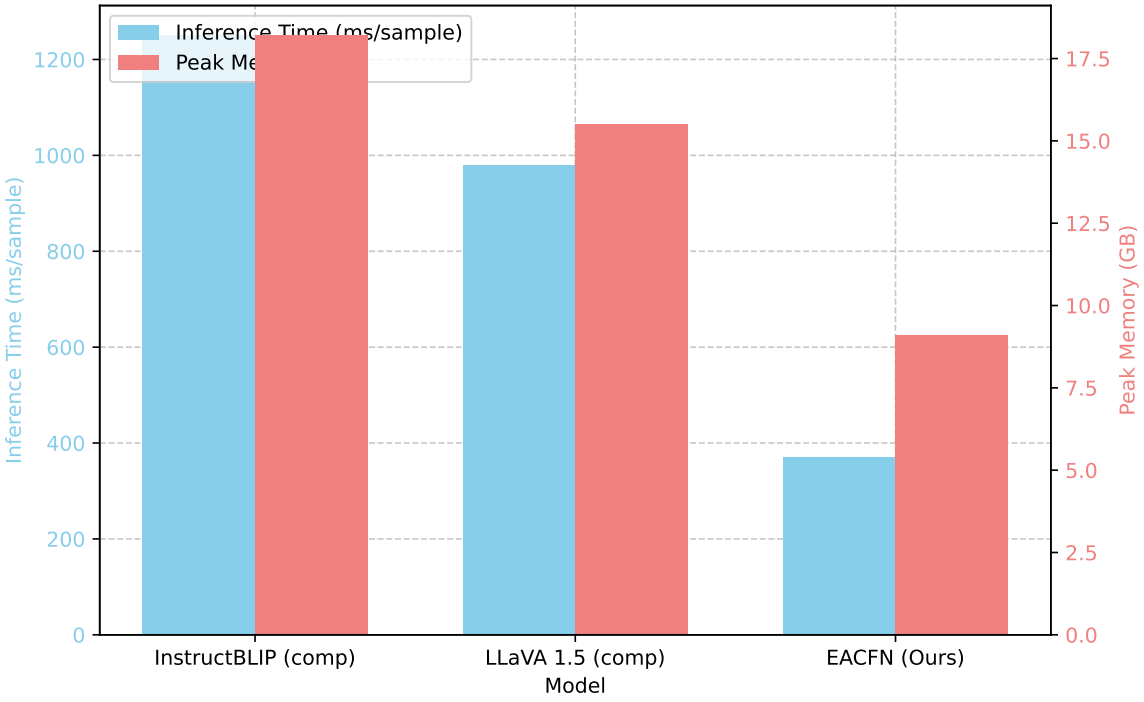

Figure 4. Computational Efficiency: Average Inference Time per sample and approximate Peak Memory Usage. 'EACFN (Ours)' is evaluated in its full configuration (with reference image). LVLMs are evaluated with the 'comp' strategy for fair comparison.

As presented in Figure 4, EACFN demonstrates superior computational efficiency compared to the large zero-shot LVLMs. EACFN achieves an average inference time of 370 ms per sample, which is significantly faster than InstructBLIP (1250 ms) and LLaVA 1.5 (980 ms). This represents a speedup of approximately 2.6 to 3.4 times, making EACFN much more suitable for applications requiring rapid processing, such as real-time fact-checking or large-scale content moderation.

Furthermore, EACFN exhibits substantially lower peak memory usage, requiring only 9.1 GB compared to 18.2 GB for InstructBLIP and 15.5 GB for LLaVA 1.5. This reduced memory footprint is a critical advantage, as it allows EACFN to be deployed on hardware with more constrained resources, broadening its applicability. The efficiency gains stem from EACFN's modular architecture, which, while integrating sophisticated mechanisms like GNNs and cross-attention, is specifically optimized for the task of entity consistency verification rather than general-purpose multi-modal understanding. The specialized nature of EACFN allows for more targeted computations compared to the expansive and resource-intensive operations inherent in large, foundational LVLMs.

4.6. Human Evaluation

While quantitative metrics like accuracy are essential, evaluating cross-modal entity consistency, especially in cases of subtle mismatches, often benefits from qualitative assessment. To further validate EACFN's robustness, particularly in challenging scenarios where LVLMs might struggle, we conducted a limited human evaluation. We sampled 100 news articles containing subtle entity inconsistencies or high contextual ambiguity from TamperedNews-Ent and News400-Ent datasets. Three expert human annotators were asked to independently verify the consistency of PER, LOC, and EVT entities for each sample. They were also asked to rate their confidence in detecting inconsistencies.

Table 3 summarizes the human evaluation results.

- EACFN demonstrates significantly higher "Agreement with Human" (81.0%) compared to InstructBLIP (68.0%) and LLaVA 1.5 (71.0%). This indicates that EACFN's judgments align more closely with human intuition, especially in nuanced cases of consistency where direct visual comparison might be ambiguous or require deeper contextual understanding.

 doi:10.20944/preprints202601.0691.v1

- Remarkably, EACFN also shows a higher "Better than Human" rate (12.0%), suggesting that in certain complex scenarios, EACFN can identify subtle inconsistencies that even human annotators might initially overlook or find difficult to confirm. This highlights EACFN's capability to learn intricate cross-modal patterns that may escape explicit human coding, potentially serving as a valuable pre-screening tool.
- Furthermore, human annotators reported a higher average confidence score (4.1) for EACFN's outputs when they agreed, implying that EACFN provides more definitive and trustworthy consistency judgments, making it a valuable tool for aiding human fact-checkers in verifying multimodal news content.

These human evaluation results reinforce the quantitative findings, affirming EACFN's superior capability in fine-grained cross-modal entity consistency verification, particularly in challenging and ambiguous news contexts.

Table 3. Human Evaluation Results: Agreement with Human Judgment and Perceived Performance on a subset of 100 challenging samples. The 'Agreement with Human' refers to the percentage of samples where the model's judgment matched the human consensus. 'Better than Human' indicates cases where the model correctly identified inconsistency that humans initially missed. 'Confidence (1-5 Scale)' is the average human-perceived confidence in the model's judgment when it agreed with them.

Model	Agreement with Human (%)	Better than Human (%)	Confidence (1-5 Scale)
InstructBLIP (comp)	68.0	5.0	3.2
LLaVA 1.5 (comp)	71.0	7.0	3.5
EACFN (Ours)	**81.0**	**12.0**	**4.1**

4.7. Qualitative Error Analysis

While EACFN generally outperforms baselines, a qualitative examination of its failures provides valuable insights into current limitations and future improvement avenues. We categorize common error patterns observed during testing, particularly for challenging samples where EACFN made incorrect judgments. Table 4 outlines these typical error types.

 doi:10.20944/preprints202601.0691.v1

Table 4. Qualitative Error Analysis: Common categories of EACFN's misclassifications and their characteristics. 'FN': False Negative, 'FP': False Positive.

Error Type	Classification	Description
Subtle Visual Mismatch	FN	Model fails to detect minor inconsistencies in visual details (e.g., slight difference in attire, background objects not matching textual context) when the overall scene is similar.
Ambiguous Visual Context	FN	News image contains elements that could be interpreted in multiple ways, leading the model to incorrectly find consistency with text (e.g., a generic street scene misaligned with a specific textual location).
Complex Event Nuances	FP/FN	Difficulties in capturing the precise timing, agents, or outcomes of an event. Model might over-generalize an action or miss a crucial detail (e.g., mistaking a protest for a celebration if visual cues are subtle).
Entity Overlap/Confusion	FN	Multiple similar-looking persons or objects in an image make it hard for the model to correctly associate a textual entity with its unique visual counterpart.
Reference Image Ambiguity	FP	The provided reference image, while intended to help, contains elements that are misleading or too generic, leading EACFN to incorrectly infer consistency for the news image.
Lack of Visual Evidence	FP	Text describes an entity or event for which there is no discernible visual representation in the news image, but the model falsely infers presence or consistency due to strong textual context.

One prominent type of error, "Subtle Visual Mismatch", often occurs when inconsistencies are very minor, such as slight changes in clothing or background elements that human eyes might notice but that current visual encoders might abstract away. For instance, a news article might state a person is wearing a red shirt, but the image shows a blue shirt; if the overall person identity and context are strong, EACFN might still judge it consistent.

"Ambiguous Visual Context" errors arise when visual regions are too generic. A textual mention of "Eiffel Tower" might be judged consistent with a generic cityscape containing tall structures if the distinct features of the Eiffel Tower are not clearly visible or emphasized in the news image's regions. These cases highlight the challenge of precisely localizing and identifying less prominent entities.

"Complex Event Nuances" are particularly difficult for EACFN. Events are dynamic and often composed of multiple sub-actions or specific participants. Misinterpreting the exact nature or actors of an event (e.g., mistaking a general gathering for a specific protest described in text) contributes to errors, suggesting that the GNN could benefit from even richer temporal or causal reasoning capabilities.

"Entity Overlap/Confusion" primarily affects Person entities when multiple individuals with similar appearances are present. The model might incorrectly align a textual entity with the wrong visual region, leading to a false negative if the intended match is inconsistent. Similarly, "Reference Image Ambiguity" poses a challenge where a less-than-ideal reference image might introduce noise or lead the model to false positive judgments by reinforcing an incorrect visual interpretation of the news image.

Finally, "Lack of Visual Evidence" errors occur when text entities have no clear visual counterpart. For example, if a text mentions "the hidden documents" and there is no visual cue for documents in the image, the model might still find consistency based on other strong textual-visual alignments, or a general semantic coherence. Addressing these nuanced failure modes will be a key focus for future

 doi:10.20944/preprints202601.0691.v1

enhancements, particularly by refining visual feature extraction for minute details and improving contextual disambiguation in the EARR module.

5. Conclusion

This paper addressed the critical challenge of Cross-modal Entity Consistency (CEC) verification in multimodal news, recognizing the limitations of existing zero-shot Large Vision-Language Models (LVLMs) for fine-grained judgments. We proposed a novel Entity-Aware Cross-Modal Fusion Network (EACFN) with a structured architecture, integrating modules for entity extraction, reference-enhanced visual feature fusion (RNIFE), graph-based entity alignment and relation reasoning (EARR), and multi-modal consistency judgment. Our comprehensive experiments on three entity-annotated datasets—TamperedNews-Ent, News400-Ent, and MMG-Ent—demonstrated that EACFN significantly outperforms state-of-the-art zero-shot LVLMs, especially leveraging its reference image fusion. Ablation studies confirmed the critical contributions of RNIFE and EARR. Beyond superior accuracy, EACFN also offers computational efficiency and higher agreement with human judgments. Future work will focus on enhancing visual detail detection, exploring advanced graph-based reasoning, and adaptive strategies for varied reference image qualities to further combat sophisticated multimodal misinformation.

References

1. Wu, Y.; Lin, Z.; Zhao, Y.; Qin, B.; Zhu, L.N. A Text-Centered Shared-Private Framework via Cross-Modal Prediction for Multimodal Sentiment Analysis. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 4730–4738. https://doi.org/10.18653/v1/2021.findings-acl.417.
2. Wu, Y.; Zhan, P.; Zhang, Y.; Wang, L.; Xu, Z. Multimodal Fusion with Co-Attention Networks for Fake News Detection. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 2560–2569. https://doi.org/10.18653/v1/2021.findings-acl.226.
3. Ayoola, T.; Tyagi, S.; Fisher, J.; Christodoulopoulos, C.; Pierleoni, A. ReFinED: An Efficient Zero-shot-capable Approach to End-to-End Entity Linking. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Track. Association for Computational Linguistics, 2022, pp. 209–220. https://doi.org/10.18653/v1/2022.naacl-industry.24.
4. Zhou, Y.; Long, G. Style-Aware Contrastive Learning for Multi-Style Image Captioning. In Proceedings of the Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 2257–2267.
5. Zhou, Y. Sketch storytelling. In Proceedings of the ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2022, pp. 4748–4752.
6. Zhou, Y.; Chen, Y.; Chen, Y.; Ye, S.; Guo, M.; Sha, Z.; Wei, H.; Gu, Y.; Zhou, J.; Qu, W. EAGLE: An Enhanced Attention-Based Strategy by Generating Answers from Learning Questions to a Remote Sensing Image. In Proceedings of the International Conference on Computational Linguistics and Intelligent Text Processing. Springer, 2019, pp. 558–572.
7. Xu, Z.; Zhang, X.; Li, R.; Tang, Z.; Huang, Q.; Zhang, J. Fakeshield: Explainable image forgery detection and localization via multi-modal large language models. *arXiv preprint arXiv:2410.02761* **2024**.
8. Zhang, X.; Li, R.; Yu, J.; Xu, Y.; Li, W.; Zhang, J. Editguard: Versatile image watermarking for tamper localization and copyright protection. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2024, pp. 11964–11974.
9. Zhang, X.; Tang, Z.; Xu, Z.; Li, R.; Xu, Y.; Chen, B.; Gao, F.; Zhang, J. Omniguard: Hybrid manipulation localization via augmented versatile deep image watermarking. In Proceedings of the Proceedings of the Computer Vision and Pattern Recognition Conference, 2025, pp. 3008–3018.
10. Potts, C.; Wu, Z.; Geiger, A.; Kiela, D. DynaSent: A Dynamic Benchmark for Sentiment Analysis. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 2388–2404. https://doi.org/10.18653/v1/2021.acl-long.186.

 doi:10.20944/preprints202601.0691.v1

11. Zhao, H.; Zhang, J.; Chen, Z.; Yuan, B.; Tao, D. On robust cross-view consistency in self-supervised monocular depth estimation. *Machine Intelligence Research* **2024**, *21*, 495–513.
12. Islam, K.I.; Kar, S.; Islam, M.S.; Amin, M.R. SentNoB: A Dataset for Analysing Sentiment on Noisy Bangla Texts. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2021. Association for Computational Linguistics, 2021, pp. 3265–3271. https://doi.org/10.18653/v1/2021.findings-emnlp.278.
13. Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; Yuan, L. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. In Proceedings of the Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2024, pp. 5971–5984. https://doi.org/10.18653/v1/2024.emnlp-main.342.
14. Yang, J.; Wang, Y.; Yi, R.; Zhu, Y.; Rehman, A.; Zadeh, A.; Poria, S.; Morency, L.P. MTAG: Modal-Temporal Attention Graph for Unaligned Human Multimodal Language Sequences. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 1009–1021. https://doi.org/10.18653/v1/2021.naacl-main.79.
15. Hu, J.; Liu, Y.; Zhao, J.; Jin, Q. MMGCN: Multimodal Fusion via Deep Graph Convolution Network for Emotion Recognition in Conversation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5666–5675. https://doi.org/10.18653/v1/2021.acl-long.440.
16. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. UniMSE: Towards Unified Multimodal Sentiment Analysis and Emotion Recognition. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 7837–7851. https://doi.org/10.18653/v1/2022.emnlp-main.534.
17. Tang, J.; Li, K.; Jin, X.; Cichocki, A.; Zhao, Q.; Kong, W. CTFN: Hierarchical Learning for Multimodal Sentiment Analysis Using Coupled-Translation Fusion Network. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 5301–5311. https://doi.org/10.18653/v1/2021.acl-long.412.
18. Lin, Z.; Tian, Z.; Lan, J.; Zhao, D.; Wei, C. Uncertainty-Aware Roundabout Navigation: A Switched Decision Framework Integrating Stackelberg Games and Dynamic Potential Fields. *IEEE Transactions on Vehicular Technology* **2025**, pp. 1–13. https://doi.org/10.1109/TVT.2025.3638268.
19. Tian, Z.; Lin, Z.; Zhao, D.; Zhao, W.; Flynn, D.; Ansari, S.; Wei, C. Evaluating scenario-based decision-making for interactive autonomous driving using rational criteria: A survey. *arXiv preprint arXiv:2501.01886* **2029**.
20. Wu, Z.; Kong, L.; Bi, W.; Li, X.; Kao, B. Good for Misconceived Reasons: An Empirical Revisiting on the Need for Visual Context in Multimodal Machine Translation. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6153–6166. https://doi.org/10.18653/v1/2021.acl-long.480.
21. Wei, K.; Sun, X.; Zhang, Z.; Zhang, J.; Zhi, G.; Jin, L. Trigger is not sufficient: Exploiting frame-aware knowledge for implicit event argument extraction. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4672–4682.
22. Wei, K.; Yang, Y.; Jin, L.; Sun, X.; Zhang, Z.; Zhang, J.; Li, X.; Zhang, L.; Liu, J.; Zhi, G. Guide the many-to-one assignment: Open information extraction via iou-aware optimal transport. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2023, pp. 4971–4984.
23. Ren, L. AI-Powered Financial Insights: Using Large Language Models to Improve Government Decision-Making and Policy Execution. *Journal of Industrial Engineering and Applied Science* **2025**, *3*, 21–26.
24. Ren, L. Leveraging large language models for anomaly event early warning in financial systems. *European Journal of AI, Computing & Informatics* **2025**, *1*, 69–76.
25. Ren, L.; et al. Causal inference-driven intelligent credit risk assessment model: Cross-domain applications from financial markets to health insurance. *Academic Journal of Computing & Information Science* **2025**, *8*, 8–14.
26. Zheng, L.; Tian, Z.; He, Y.; Liu, S.; Chen, H.; Yuan, F.; Peng, Y. Enhanced mean field game for interactive decision-making with varied stylish multi-vehicles. *arXiv preprint arXiv:2509.00981* **2029**.

 doi:10.20944/preprints202601.0691.v1

27. Yang, J.; Yu, Y.; Niu, D.; Guo, W.; Xu, Y. ConFEDE: Contrastive Feature Decomposition for Multimodal Sentiment Analysis. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 7617–7630. https://doi.org/10.18653/v1/2023.acl-long.421.
28. Allaway, E.; Srikanth, M.; McKeown, K. Adversarial Learning for Zero-Shot Stance Detection on Social Media. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 4756–4767. https://doi.org/10.18653/v1/2021.naacl-main.379.
29. Wei, K.; Zhong, J.; Zhang, H.; Zhang, F.; Zhang, D.; Jin, L.; Yu, Y.; Zhang, J. Chain-of-specificity: Enhancing task-specific constraint adherence in large language models. In Proceedings of the Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 2401–2416.
30. Zhou, Y.; Shen, J.; Cheng, Y. Weak to strong generalization for large language models with multi-capabilities. In Proceedings of the The Thirteenth International Conference on Learning Representations, 2025.
31. Zhou, Y.; Li, X.; Wang, Q.; Shen, J. Visual In-Context Learning for Large Vision-Language Models. In Proceedings of the Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024. Association for Computational Linguistics, 2024, pp. 15890–15902.
32. Li, Z.; Xu, B.; Zhu, C.; Zhao, T. CLMLF:A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection. In Proceedings of the Findings of the Association for Computational Linguistics: NAACL 2022. Association for Computational Linguistics, 2022, pp. 2282–2294. https://doi.org/10.18653/v1/2022.findings-naacl.175.
33. Zhou, Y.; Tao, W.; Zhang, W. Triple sequence generative adversarial nets for unsupervised image captioning. In Proceedings of the ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 7598–7602.
34. Dimitrov, D.; Bin Ali, B.; Shaar, S.; Alam, F.; Silvestri, F.; Firooz, H.; Nakov, P.; Da San Martino, G. Detecting Propaganda Techniques in Memes. In Proceedings of the Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Association for Computational Linguistics, 2021, pp. 6603–6617. https://doi.org/10.18653/v1/2021.acl-long.516.
35. Zhao, H.; Zhang, J.; Chen, Z.; Zhao, S.; Tao, D. Unimix: Towards domain adaptive and generalizable lidar semantic segmentation in adverse weather. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 14781–14791.
36. Zhao, H.; Zhang, Q.; Zhao, S.; Chen, Z.; Zhang, J.; Tao, D. Simdistill: Simulated multi-modal distillation for bev 3d object detection. In Proceedings of the Proceedings of the AAAI conference on artificial intelligence, 2024, Vol. 38, pp. 7460–7468.
37. Chen, W.; Hu, H.; Chen, X.; Verga, P.; Cohen, W. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 5558–5570. https://doi.org/10.18653/v1/2022.emnlp-main.375.
38. Wu, H.; Li, H.; Su, Y. Bridging the Perception-Cognition Gap:Re-engineering SAM2 with Hilbert-Mamba for Robust VLM-based Medical Diagnosis, 2025, [arXiv:cs.CV/2512.24013].
39. Huang, Y.; Fang, M.; Cao, Y.; Wang, L.; Liang, X. DAGN: Discourse-Aware Graph Network for Logical Reasoning. In Proceedings of the Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2021, pp. 5848–5855. https://doi.org/10.18653/v1/2021.naacl-main.467.
40. Li, J.; Xu, K.; Li, F.; Fei, H.; Ren, Y.; Ji, D. MRN: A Locally and Globally Mention-Based Reasoning Network for Document-Level Relation Extraction. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 1359–1370. https://doi.org/10.18653/v1/2021.findings-acl.117.
41. Qiao, S.; Ou, Y.; Zhang, N.; Chen, X.; Yao, Y.; Deng, S.; Tan, C.; Huang, F.; Chen, H. Reasoning with Language Model Prompting: A Survey. In Proceedings of the Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2023, pp. 5368–5393. https://doi.org/10.18653/v1/2023.acl-long.294.
42. Hao, S.; Gu, Y.; Ma, H.; Hong, J.; Wang, Z.; Wang, D.; Hu, Z. Reasoning with Language Model is Planning with World Model. In Proceedings of the Proceedings of the 2023 Conference on Empirical Methods in

 doi:10.20944/preprints202601.0691.v1

Natural Language Processing. Association for Computational Linguistics, 2023, pp. 8154–8173. https://doi.org/10.18653/v1/2023.emnlp-main.507.