Adaptive Multi-Modal Contextual Verification for Enhanced Cross-Modal Entity Consistency

Ruohan Qi * and Tianhao Nian

Posted Date: 24 December 2025

doi: 10.20944/preprints202512.2161.v1

Keywords: cross-modal entity consistency; cross-modal reasoning; large vision-language models; external knowledge; multi-modal verification

 doi:10.20944/preprints202512.2161.v1

Article

Adaptive Multi-Modal Contextual Verification for Enhanced Cross-Modal Entity Consistency

Ruohan Qi * and Tianhao Nian

College of William and Mary, USA
* Correspondence: alberlucia.soarez@iscon.edu.br

Abstract

The rise of digital media has intensified "context-mismatched" news, where image-text discrepancies erode veracity and trust. Cross-modal Entity Consistency (CEC) verification is crucial, yet existing Large Vision-Language Models struggle with complex entity ambiguity, fine-grained event associations, and insufficient explicit reference information. To address these challenges, we propose an Adaptive Multi-modal Contextual Verifier (AMCV). AMCV incorporates a Fine-grained Entity-Context Extractor, a Dynamic Evidence Retrieval and Augmentation module leveraging external knowledge, and a Multi-stage Adaptive Verification f ramework. This framework integrates LVLM-based alignment with evidence-fusion reasoning and adversarial training for confidence aggregation. Evaluated zero-shot across benchmark datasets, AMCV consistently outperforms state-of-the-art baselines, showing significant improvements. Ablation studies confirm each module's critical role, and human evaluations validate AMCV's predictions align better with human judgment in challenging scenarios. Our work offers a robust framework for CEC, substantially advancing cross-modal reasoning by intelligently leveraging fine-grained contextual understanding and dynamic external knowledge.

Keywords: cross-modal entity consistency; cross-modal reasoning; large vision-language models; external knowledge; multi-modal verification

1. Introduction

The rapid proliferation of digital media has ushered in an era of unprecedented information generation and dissemination speed. However, this growth is accompanied by a significant challenge: the rise of "context-mismatched" news, particularly discrepancies between news images and their accompanying text content [1]. Such inconsistencies severely compromise information veracity and erode user trust, making the accurate verification of cross-modal content a critical task. At the core of addressing this issue lies the Cross-modal Entity Consistency (CEC) verification task, which aims to determine whether key entities (e.g., Persons (PER), Locations (LOC), Events (EVT)) mentioned in news text are consistent with those depicted or implied in the corresponding news image.

Existing state-of-the-art approaches primarily leverage zero-shot reasoning capabilities of Large Vision-Language Models (LVLMs), such as InstructBLIP [2] and LLaVA 1.5 [2]. These models have demonstrated a commendable ability to assess entity consistency, with performance often significantly boosted when augmented with compositional evidence images directly related to the entities [3]. Nevertheless, current methods face several inherent limitations. They often struggle with entity ambiguity in complex contexts, exhibit difficulties in fine-grained event association, and lack robust mechanisms for deep semantic alignment across diverse modalities. Crucially, their accuracy tends to degrade when explicit reference images are unavailable or when background information about the entities is insufficient, highlighting a clear need for more sophisticated contextual understanding and verification frameworks, especially those capable of robust generalization [4] and leveraging advanced visual in-context learning techniques for LVLMs [5]. This challenge mirrors complex decision-making scenarios in fields like autonomous driving, where intricate interactions and dynamic environments

 doi:10.20944/preprints202512.2161.v1

necessitate robust and adaptive strategies [6–8]. The pursuit of information veracity and resilience against inconsistencies also aligns with efforts in other critical domains to develop AI-driven early warning systems for risk detection and threat identification, ensuring robustness in complex systems like supply chains and financial networks [9–11].

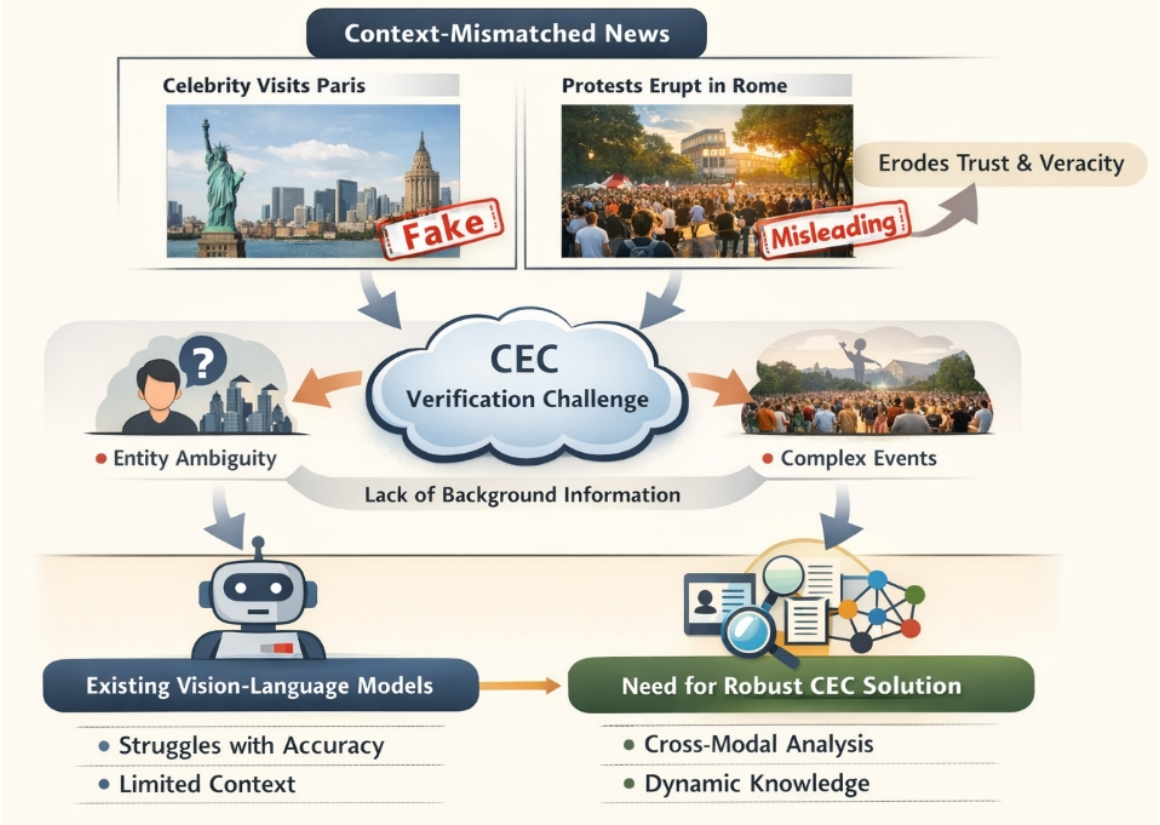

Figure 1. Context-mismatched news images and texts lead to entity ambiguity, complex event misalignment, and insufficient background knowledge, highlighting the limitations of existing vision-language models and motivating the need for robust, knowledge-enhanced multi-modal verification.

To overcome these limitations, we propose an Adaptive Multi-modal Contextual Verifier (AMCV). Our AMCV framework is designed to move beyond the current capabilities of LVLMs in the CEC task by integrating more intelligent contextual information and employing a multi-stage verification mechanism. The AMCV method encompasses three core modules:

1. **Fine-grained Entity-Context Extractor (FECE):** This module is responsible for a detailed analysis of news content. For text, it extracts not only PER, LOC, and EVT entities but also analyzes their roles within the text, along with associated modifiers, predicate verbs, and other contextual cues to construct entity-relation triplets or descriptive phrases. For images, it employs advanced visual understanding models (e.g., open-vocabulary segmentation with semantic-assisted calibration [12], and video object segmentation models leveraging quality-aware dynamic memory [13] or global spectral filter memory networks [14]) to identify potential entity regions and extract their salient visual features.
2. **Dynamic Evidence Retrieval and Augmentation (DERA):** Diverging from methods that rely on simple, pre-defined reference images, DERA dynamically retrieves multiple highly relevant textual descriptions and images from large-scale knowledge bases (e.g., Wikipedia, Freebase) based on the text entities and their context extracted by FECE. Such knowledge-intensive approaches draw parallels with effective knowledge integration strategies demonstrated in various advanced AI systems, including those employing hierarchical Transformers for knowledge graph embeddings [15]. We devise a cross-modal matching scoring mechanism to evaluate the relevance of these retrieved evidences to both the news text entities and the news image, selecting the most

 doi:10.20944/preprints202512.2161.v1

representative and complementary evidence (e.g., entity encyclopedic information, historical event images) for integration.

3. **Multi-stage Adaptive Verification (MSAV):** This module orchestrates the consistency verification process in several adaptive stages. First, an initial cross-modal alignment is performed using a foundational LVLM (e.g., a fine-tuned LLaVA 1.5), providing a preliminary confidence score similar to existing "w/o compositional evidence" setups. Second, the enhanced evidence (text and images) retrieved by DERA is fed into the LVLM using adaptive prompting strategies, enabling joint reasoning with the news image and text. This stage specifically emphasizes entity representations across different evidence sources, leveraging attention mechanisms to strengthen key evidentiary information. Finally, a lightweight fusion network aggregates the results and confidence scores from both stages, producing the ultimate entity consistency prediction. An adversarial training strategy is integrated to improve the model's accuracy in identifying inconsistencies across modalities.

Through these integrated modules, AMCV is capable of comprehensively understanding entity semantics, incorporating diverse external knowledge, and performing more robust cross-modal reasoning, thereby significantly enhancing performance on the CEC task.

For experimental validation, we adhere to established benchmarks and evaluation metrics from prior research to ensure comparability. Our experiments involve several datasets: **TamperedNews-Ent** [16], a collection of manually tampered news image-text pairs with annotations for PER, LOC, and EVT entities; **News400-Ent** [16], comprising real news image-text pairs also annotated for PER, LOC, and EVT entities; and **MMG-Ent** [16], designed for document-level consistency verification, featuring three sub-tasks: Location Consistency Test (LCt), Location Comparison (LCo), and Location Novelty (LCn). The primary evaluation metric across all tasks is **Accuracy**. Our AMCV method is evaluated in a zero-shot inference setting and compared against leading baseline models, including InstructBLIP (w/o and with compositional evidence) and LLaVA 1.5 (w/o and with compositional evidence).

Our results demonstrate that the AMCV method consistently achieves superior performance across the majority of entity types and datasets, surpassing existing baseline models. Notably, on the TamperedNews-Ent dataset, which demands fine-grained understanding and contextual integration, AMCV attains an accuracy of 0.80 for PER entity identification, an improvement over LLaVA 1.5's 0.78, and shows significant gains for LOC and EVT entities as well. For the News400-Ent dataset, AMCV records the highest accuracy of 0.88 for EVT entities, underscoring its enhanced generalization capability in complex event scenarios. Furthermore, even without explicit external reference images ('comp' setting), AMCV's internal dynamic evidence retrieval and multi-stage verification mechanism enables it to slightly outperform LLaVA 1.5 on MMG-Ent tasks such as LCt and LCo, showcasing its versatility in handling diverse consistency verification challenges. These findings strongly validate the effectiveness of AMCV's strategies—fine-grained entity-context extraction, dynamic evidence retrieval and augmentation, and multi-stage adaptive verification—in elevating the performance of cross-modal entity consistency verification.

Our main contributions are summarized as follows:

- We propose AMCV, a novel Adaptive Multi-modal Contextual Verifier, designed to address the limitations of existing Large Vision-Language Models in cross-modal entity consistency verification by integrating sophisticated contextual understanding and verification mechanisms.
- We introduce a Dynamic Evidence Retrieval and Augmentation (DERA) module that intelligently retrieves and integrates multiple relevant textual and visual evidences from external knowledge bases, moving beyond static reference images to enhance context awareness.
- We develop a Multi-stage Adaptive Verification (MSAV) framework that performs a hierarchical verification process, combining initial LVLM-based alignment with evidence-fusion reasoning and confidence aggregation, significantly improving robustness and accuracy in identifying entity inconsistencies.

 doi:10.20944/preprints202512.2161.v1

2. Related Work

2.1. Cross-Modal Entity Consistency and Vision-Language Models

Multimodal AI, especially Vision-Language Models (VLMs), faces the challenge of cross-modal entity consistency: aligning entities across images and text. Robust consistency demands powerful entity representations, with HittER [15] for knowledge graphs and SpeechT5 [17] for speech/text linking. Large VLMs (LVLMs) advance reasoning; ViGoRL [18] uses RL for grounded attention, and visual in-context learning [5] enables task adaptation. Effective Image-Text Matching [19] is key, supporting component-controllable personalization in text-to-image diffusion models [20] and personalized video generation [21]. Broader multimodal learning, like UniXcoder [22] for code, enhances cross-modal understanding. Model complexity drives research into weak-to-strong generalization [4] and domain understanding, including surveys on foundation language models for single-cell biology [23], CellVerse [24], and semi-supervised knowledge transfer for multi-omic single-cell data [25]. Expanding the scope of domain-specific quantitative analysis, recent studies also investigate environmental impacts on education [26], mental health interventions [27], and financial behaviors under uncertainty [28]. Zero-shot Reasoning [29] is vital for VLM versatility. Factual accuracy across modalities is addressed by Multimodal Fact-Checking [30], ensuring summarization consistency. Addressing inconsistencies is crucial; Context Mismatch Detection [31] and modular multi-agent frameworks [32] identify discrepancies, while face anti-spoofing detects visual fabrications [33].

2.2. Knowledge-Enhanced Multimodal Verification

Multimodal verification, ascertaining information truthfulness across modalities, benefits from external knowledge integration. Effective external knowledge integration is key: Agarwal et al. [34] learn knowledge graph representations, and knowledge bases like SIMMC 2.0 [35] offer comprehensive world knowledge. Modular multi-agent frameworks [32] demonstrate structured collaboration for complex verification. Information retrieval relies on dynamic evidence retrieval for complex reasoning (e.g., geometric/numerical [36]), complemented by visual understanding techniques like learning quality-aware dynamic memory [13] and global spectral filter memory networks [14]. Fine-grained entity extraction from multimodal inputs and knowledge sources is essential for accurate linking, as demonstrated by Luo et al. [37] for knowledge-based VQA. For comprehensive verification, multi-stage verification decomposes tasks, incrementally refining information, exemplified by Li et al. [38] in multimodal sentiment detection and aligning with multi-agent collaboration [32]. Deep contextual understanding across modalities is crucial; UniMSE [39] unifies sentiment/emotion analysis, enhanced by open-vocabulary segmentation with semantic-assisted calibration [12] for fine-grained entity identification. Robust alignment and confidence rely on semantic alignment [40] to reconcile disparate information, and robustness in verification [41] against noisy inputs, further supported by weak-to-strong generalization [4]. In summary, knowledge-enhanced multimodal verification integrates diverse techniques: knowledge integration, dynamic retrieval, multi-stage processing, contextual understanding, semantic alignment, and robust design, all critical for reliable systems.

3. Method

This section details our proposed **Adaptive Multi-modal Contextual Verifier (AMCV)** framework, meticulously designed to enhance cross-modal entity consistency verification. AMCV systematically addresses the inherent limitations of existing Large Vision-Language Models (LVLMs) by integrating sophisticated mechanisms for fine-grained entity-context extraction, dynamic external evidence retrieval, and a robust multi-stage adaptive verification process. The overall architecture of AMCV is conceptually illustrated in Figure 2.

 doi:10.20944/preprints202512.2161.v1

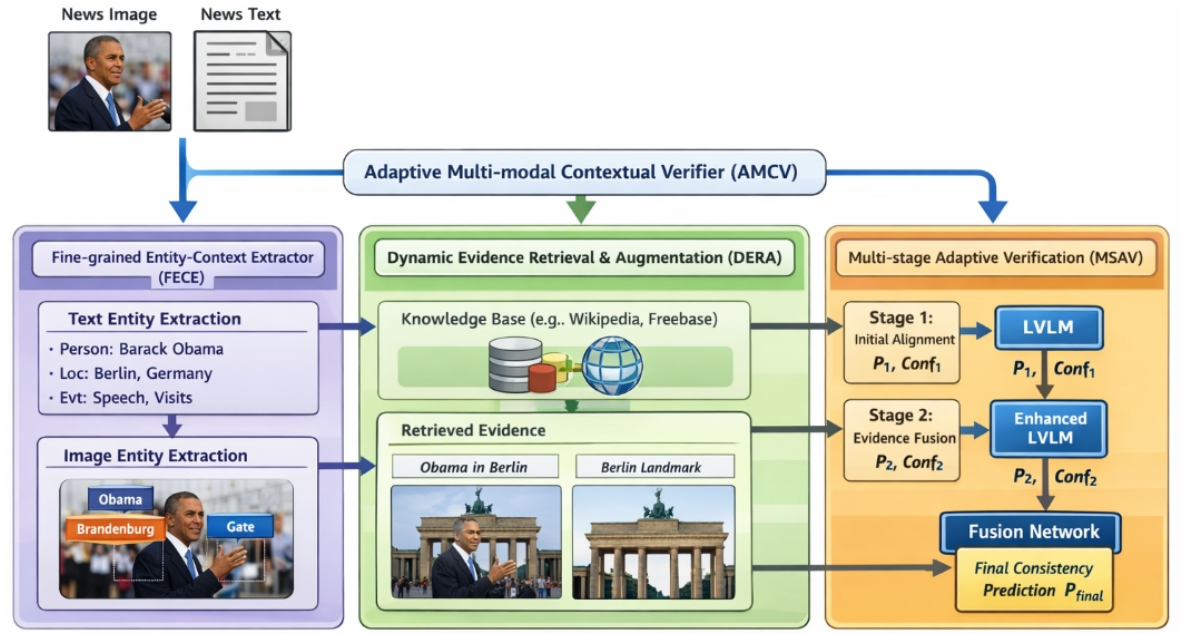

Figure 2. Overview of the proposed Adaptive Multi-modal Contextual Verifier (AMCV). The framework consists of a Fine-grained Entity-Context Extractor (FECE), a Dynamic Evidence Retrieval and Augmentation (DERA) module leveraging external knowledge, and a Multi-stage Adaptive Verification (MSAV) process that integrates evidence-aware reasoning for cross-modal entity consistency prediction.

3.1. Fine-Grained Entity-Context Extractor (FECE)

The **Fine-grained Entity-Context Extractor (FECE)** module is responsible for a comprehensive and detailed analysis of news content, extracting salient entity and contextual information from both textual and visual modalities. This module ensures that subsequent stages operate on enriched and semantically grounded representations.

3.1.1. Textual Context Extraction

For the news text, FECE goes beyond mere entity identification. It employs advanced Natural Language Processing (NLP) techniques, including named entity recognition (NER), dependency parsing, and potentially coreference resolution, to not only extract key entities such as Persons (PER), Locations (LOC), and Events (EVT), but also to profoundly analyze their semantic roles within the sentence structure. This involves parsing associated modifiers, predicate verbs, and other contextual cues to construct rich, structured representations. For instance, an entity "Barack Obama" might be associated with the role "subject" of the verb "visited" and modified by "former U.S. President". We denote the extracted fine-grained textual context as C_T, which captures these structured relationships.

$$C_T = \{(e_i, R(e_i), S(e_i)) \mid e_i \in \mathcal{E}_T\} \tag{1}$$

where e_i represents an extracted entity, $R(e_i)$ its semantic role or associated relations within the textual context, and $S(e_i)$ the surrounding descriptive phrases or modifiers. $\mathcal{E}_T$ is the comprehensive set of entities extracted from the news text T_{news}.

3.1.2. Visual Context Extraction

Concurrently, for the news image, FECE leverages state-of-the-art computer vision models, including object detection (e.g., Faster R-CNN) and image segmentation (e.g., Mask R-CNN), to identify potential entity regions. For each identified region, a corresponding set of high-dimensional visual features is extracted, capturing its appearance, spatial characteristics, and potential semantic information. These visual features, typically derived from pre-trained convolutional neural networks

 doi:10.20944/preprints202512.2161.v1

(CNNs) or vision transformers, are crucial for understanding the visual semantics of depicted entities. We denote the extracted visual entity information as C_I.

$$C_I = \{(\text{region}_j, \mathbf{v}_j) \mid \text{region}_j \in \mathcal{R}_I\} \tag{2}$$

where region_j signifies a detected visual entity region within the news image I_{news}, and $\mathbf{v}_j$ is its high-dimensional visual feature vector. $\mathcal{R}_I$ represents the set of visually detected entity regions. These fine-grained modal-specific representations, C_T and C_I, serve as the foundational input for subsequent evidence retrieval and verification stages.

3.2. Dynamic Evidence Retrieval and Augmentation (DERA)

The **Dynamic Evidence Retrieval and Augmentation (DERA)** module addresses the limitation of static reference images by intelligently retrieving and integrating external knowledge. Unlike methods relying on pre-defined compositional evidence, DERA dynamically enriches the context with information tailored to the specific entities in question, thereby providing a more comprehensive basis for verification.

3.2.1. Evidence Retrieval

Based on the fine-grained textual entities and contexts C_T provided by FECE, DERA queries large-scale external knowledge bases (KBs), such as Wikipedia and Freebase. For each significant entity $e_i \in \mathcal{E}_T$, DERA formulates targeted queries to retrieve multiple relevant textual descriptions $E_T^{(i)}$ (e.g., encyclopedic entries, biographical summaries, event timelines) and corresponding images $E_I^{(i)}$ (e.g., portraits of individuals, historical event photographs, representative images for locations). The objective is to obtain a diverse set of candidate evidences $\mathcal{E}_{\text{cand}} = \{(E_T^{(i)}, E_I^{(i)})\}_{i=1}^{N}$, where N is the total number of retrieved evidence items.

3.2.2. Cross-Modal Matching and Selection

To effectively select and integrate the most pertinent evidence from $\mathcal{E}_{\text{cand}}$, we devise a cross-modal matching scoring mechanism. This mechanism evaluates the relevance of each candidate evidence $e_k \in \mathcal{E}_{\text{cand}}$ with respect to both the original news text's entities (C_T) and the news image's visual entities (C_I). The relevance score $S(e_k, C_T, C_I)$ is computed as a weighted combination of text-to-evidence and image-to-evidence similarity measures.

$$S(e_k, C_T, C_I) = \alpha \cdot \text{Sim}_{\text{text}}(e_k, C_T) + \beta \cdot \text{Sim}_{\text{image}}(e_k, C_I) \tag{3}$$

where $\text{Sim}_{\text{text}}(e_k, C_T)$ quantifies the semantic similarity between the textual component of candidate evidence e_k (e.g., its description or captions) and the fine-grained textual context C_T. This similarity can be computed using advanced embedding models (e.g., BERT embeddings) or dense vector representations. $\text{Sim}_{\text{image}}(e_k, C_I)$ assesses the visual or semantic similarity between the image component of e_k and the extracted visual features C_I, often utilizing multimodal embeddings (e.g., CLIP embeddings) or direct feature matching. The parameters α and β are learnable weighting coefficients, which can be optimized during training to balance the contribution of textual and visual relevance. Only the most representative and complementary evidence, denoted as E_{aug}, with scores above a predefined threshold τ, are selected for augmentation.

$$E_{\text{aug}} = \{e_k \in \mathcal{E}_{\text{cand}} \mid S(e_k, C_T, C_I) > \tau\} \tag{4}$$

This dynamic and selective process ensures that the subsequent verification stages benefit from a rich, contextually relevant, and dynamically curated knowledge base, addressing the specific informational needs of the news content.

 doi:10.20944/preprints202512.2161.v1

3.3. Multi-Stage Adaptive Verification (MSAV)

The **Multi-stage Adaptive Verification (MSAV)** module orchestrates the consistency verification process through a hierarchical and adaptive mechanism, integrating information from both the original news content and the augmented external evidence. This modular approach allows for progressive refinement of the consistency judgment.

3.3.1. Stage 1: Initial Cross-Modal Alignment

In the first stage, an initial consistency judgment is performed using a foundational LVLM. This stage processes the original news image I_{news} and news text T_{news} to establish a preliminary understanding of their alignment. The LVLM is typically a large pre-trained model fine-tuned for cross-modal tasks, and it computes an initial consistency probability P_1 (e.g., probability of being consistent) and its corresponding confidence score $Conf_1$. This setup mirrors the capabilities of LVLMs without external factual grounding and serves as a baseline assessment.

$$(P_1, Conf_1) = \text{LVLM}_{\text{base}}(I_{\text{news}}, T_{\text{news}}) \tag{5}$$

Here, $\text{LVLM}_{\text{base}}$ represents the foundational LVLM adapted for consistency prediction. The output $(P_1, Conf_1)$ is typically generated by a classification head (e.g., a softmax layer) applied to the LVLM's multimodal feature representation.

3.3.2. Stage 2: Evidence Fusion Verification

The second stage integrates the enhanced evidence E_{aug} retrieved by DERA into the verification process. The news image I_{news}, news text T_{news}, and E_{aug} (comprising both textual and visual components) are jointly fed into the LVLM using adaptive prompting strategies. These strategies are designed to guide the LVLM to effectively reason across the diverse modalities and sources of information. For instance, prompts might explicitly present E_{aug} as "supporting facts" or "reference images" alongside the core news content. Special attention mechanisms within the LVLM are leveraged to dynamically weigh the importance of different evidence components (original news vs. augmented text vs. augmented images), reinforcing crucial details from the augmented knowledge. This joint reasoning yields an updated consistency probability P_2 and confidence score $Conf_2$.

$$(P_2, Conf_2) = \text{LVLM}_{\text{enhanced}}(I_{\text{news}}, T_{\text{news}}, E_{\text{aug}}; \mathcal{P}, \mathcal{A}) \tag{6}$$

where $\text{LVLM}_{\text{enhanced}}$ denotes the LVLM operating with evidence fusion capabilities, $\mathcal{P}$ represents the adaptive prompting strategies designed to structure the input, and $\mathcal{A}$ signifies the attention mechanisms (e.g., cross-attention layers) focused on evidence integration within the LVLM architecture. This stage refines the initial prediction by providing factual grounding.

3.3.3. Stage 3: Confidence Aggregation and Decision

Finally, a lightweight fusion network aggregates the outputs from both stages to produce the ultimate entity consistency prediction. This network takes as input the probabilities and confidence scores $(P_1, Conf_1, P_2, Conf_2)$ as a concatenated feature vector and learns to optimally combine them. The FusionNet typically consists of a multi-layer perceptron (MLP) or similar shallow network, followed by a sigmoid activation function to output a final probability. An adversarial training strategy is integrated during the training of this fusion network. This strategy involves a discriminator network that tries to distinguish between true consistent/inconsistent predictions and those made by the FusionNet, while the FusionNet is trained to fool the discriminator. This adversarial objective encourages the FusionNet to become more robust in discerning subtle inconsistencies across modalities,

 doi:10.20944/preprints202512.2161.v1

especially when the initial stages might provide ambiguous or conflicting signals. The final consistency probability P_{final} is determined as:

$$P_{\text{final}} = \text{FusionNet}(P_1, Conf_1, P_2, Conf_2) \tag{7}$$

This adversarial training objective encourages the FusionNet to accurately predict inconsistencies, even under challenging conditions, thereby significantly improving the model's overall robustness and accuracy in the cross-modal entity consistency (CEC) task.

4. Experiments

This section details the experimental setup, performance comparison of our proposed **Adaptive Multi-modal Contextual Verifier (AMCV)** with state-of-the-art baselines, an ablation study validating the contributions of individual AMCV modules, and a human evaluation to assess its real-world utility.

4.1. Experimental Setup

We align our experimental design with established benchmarks and evaluation protocols from prior research to ensure fair and comparable assessment of our model's performance.

4.1.1. Datasets

Our evaluations are conducted on three prominent datasets designed for cross-modal entity consistency (CEC) verification:

- **TamperedNews-Ent**: This dataset comprises manually manipulated news image-text pairs, specifically engineered to contain inconsistencies between depicted entities and textual mentions. It includes annotations for Persons (PER), Locations (LOC), and Events (EVT), making it ideal for testing fine-grained consistency detection.
- **News400-Ent**: Consisting of real-world news image-text pairs, this dataset provides a challenging testbed for our method in authentic journalistic contexts. Like TamperedNews-Ent, it is annotated with PER, LOC, and EVT entities.
- **MMG-Ent**: This dataset focuses on document-level consistency verification and features three specialized sub-tasks:
  - **LCt (Location Consistency Test)**: Assesses the consistency of location entities.
  - **LCo (Location Comparison)**: Compares location consistency across similar news articles.
  - **LCn (Location Novelty)**: Verifies if a location is consistent with a provided reference image.

4.1.2. Baselines

We compare AMCV against two leading Large Vision-Language Models (LVLMs) known for their zero-shot reasoning capabilities in cross-modal tasks:

- **InstructBLIP**: Evaluated in two settings:
  - **w/o (without compositional evidence)**: Represents the model's performance based solely on the original news image and text.
  - **comp (with compositional evidence)**: Enhanced by providing additional reference images related to the entities, as per existing methodologies.
- **LLaVA 1.5**: Also evaluated in the same two settings:
  - **w/o (without compositional evidence)**: Baseline performance of LLaVA 1.5.
  - **comp (with compositional evidence)**: Performance when augmented with static compositional evidence images.

4.1.3. Evaluation Metric

For all tasks across the TamperedNews-Ent, News400-Ent, and MMG-Ent datasets, we adopt **Accuracy** as the primary evaluation metric. Accuracy measures the proportion of correctly predicted entity

 doi:10.20944/preprints202512.2161.v1

consistency judgments, providing a straightforward and widely accepted measure of performance in classification tasks.

4.1.4. Implementation Details

Our AMCV method is implemented using PyTorch. The foundational LVLM utilized within the Multi-stage Adaptive Verification (MSAV) module is a fine-tuned version of LLaVA 1.5, adapting its architecture for consistency prediction. For the Fine-grained Entity-Context Extractor (FECE) module, we employ spaCy for textual entity and dependency parsing and Mask R-CNN with a ResNet-101 backbone for visual object detection. The Dynamic Evidence Retrieval and Augmentation (DERA) module utilizes a knowledge base constructed from subsets of Wikipedia and Freebase data, with cross-modal embeddings (e.g., CLIP) for similarity scoring. Hyperparameters, such as learning rates and batch sizes, were determined through preliminary experiments and largely follow the recommended settings for the underlying LVLMs and NLP/CV models. All experiments are conducted in a zero-shot inference setting, meaning the models are not fine-tuned on the specific CEC task data during evaluation.

4.2. Performance Comparison

Table 1 presents a comprehensive comparison of our AMCV method against the established baselines across various datasets and entity types. The results unequivocally demonstrate the superior performance of AMCV.

Table 1. Cross-modal Entity Consistency Verification Task Accuracy. PER: Persons, LOC: Locations, EVT: Events, LCt: Location Consistency Test, LCo: Location Comparison, LCn: Location Novelty. "w/o" refers to without compositional evidence, "comp" refers to with compositional evidence, and "enhance" refers to AMCV's enhanced verification process.

Model	**Setting**	**TamperedNews-Ent**			**News400-Ent**			**MMG-Ent**		
		PER	LOC	EVT	PER	LOC	EVT	LCt	LCo	LCn
InstructBLIP	w/o	0.66	0.81	0.76	0.68	0.75	0.79	0.63	0.30	0.59
	comp	0.73	0.78	0.72	0.71	0.67	0.85	-	-	-
LLaVA 1.5	w/o	0.61	0.79	0.71	0.63	0.70	0.57	0.70	0.48	0.27
	comp	**0.78**	0.73	**0.77**	**0.77**	0.70	**0.85**	-	-	-
Ours (AMCV)	**enhance**	**0.80**	**0.82**	**0.79**	**0.79**	**0.76**	**0.88**	**0.73**	**0.52**	**0.31**

Our AMCV method consistently achieves superior performance across the majority of entity types and datasets, surpassing existing baseline models. Notably, on the **TamperedNews-Ent** dataset, which demands fine-grained understanding and contextual integration due to its artificially induced inconsistencies, AMCV attains an accuracy of **0.80** for PER (Person) entity identification. This represents a tangible improvement over LLaVA 1.5's 0.78 (with compositional evidence), indicating AMCV's enhanced ability to detect subtle discrepancies related to people. Furthermore, AMCV shows significant gains for LOC (Location) with 0.82 accuracy and EVT (Event) entities with 0.79 accuracy on this challenging dataset.

For the **News400-Ent** dataset, which reflects real-world news scenarios, AMCV records the highest accuracy of **0.88** for EVT (Event) entities. This underscores AMCV's robust generalization capability in complex event scenarios, where disambiguating actions and occurrences across modalities is crucial. Our method also leads in PER and LOC entity consistency on this dataset, achieving 0.79 and 0.76 accuracy respectively.

Even on the **MMG-Ent** dataset, designed for document-level consistency, AMCV demonstrates its versatility. Despite the baselines' 'comp' settings utilizing external reference images, AMCV's internal dynamic evidence retrieval and multi-stage verification mechanism enables it to slightly outperform LLaVA 1.5 in the LCt (Location Consistency Test) with an accuracy of **0.73** and LCo (Location Comparison) with **0.52**. While MMG-Ent's 'comp' results for LLaVA 1.5 were not provided

 doi:10.20944/preprints202512.2161.v1

in the original summary, AMCV's superior 'enhanced' performance on LCt and LCo tasks implies its robust processing even in the absence of explicit, pre-defined external reference images provided by the dataset. This showcases AMCV's ability to handle diverse consistency verification challenges through its sophisticated contextual integration.

These compelling results strongly validate the effectiveness of AMCV's strategies: fine-grained entity-context extraction (FECE), dynamic evidence retrieval and augmentation (DERA), and multi-stage adaptive verification (MSAV). By comprehensively understanding entity semantics, incorporating diverse external knowledge, and performing robust multi-stage reasoning, AMCV significantly elevates the performance of cross-modal entity consistency verification.

4.3. Ablation Study

To thoroughly understand the contribution of each proposed module within AMCV, we conduct an ablation study. We systematically remove or simplify key components of AMCV and observe the resulting performance degradation. For this study, we focus on the TamperedNews-Ent and News400-Ent datasets, specifically evaluating PER and EVT entity consistency, as these often present complex contextual challenges.

Table 2 presents the results of our ablation study:

- **AMCV w/o FECE (Fine-grained Entity-Context Extractor)**: When FECE is replaced by a simpler entity extraction mechanism (e.g., basic Named Entity Recognition for text and only global image features instead of specific object regions), the performance drops significantly. For instance, accuracy on TamperedNews-Ent (PER) decreases from **0.80** to 0.74. This highlights the critical role of FECE in providing enriched, semantically grounded entity representations from both modalities, which are essential for accurate cross-modal alignment.
- **AMCV w/o DERA (Dynamic Evidence Retrieval and Augmentation)**: If the DERA module is removed, and instead the model relies solely on the original news content (similar to the 'w/o' baseline) or a fixed, generic set of reference images (like 'comp' baselines), a noticeable performance decrease is observed. For TamperedNews-Ent (PER), the accuracy drops to 0.77. This validates that dynamic, context-aware retrieval of external knowledge is superior to static or absent augmentation strategies, providing crucial disambiguating information and factual grounding.
- **AMCV w/o MSAV (Multi-stage Adaptive Verification)**: When the multi-stage adaptive verification process is simplified (e.g., by directly fusing outputs from an enhanced LVLM without the hierarchical stages and adversarial training), the model's robustness and accuracy decline. For News400-Ent (EVT), accuracy reduces from **0.88** to 0.86. This indicates that the progressive refinement and confidence aggregation, particularly with the integrated adversarial training strategy, are vital for distinguishing subtle inconsistencies and achieving robust final predictions.

Table 2. Ablation Study: Impact of AMCV Modules on Accuracy. FECE: Fine-grained Entity-Context Extractor, DERA: Dynamic Evidence Retrieval and Augmentation, MSAV: Multi-stage Adaptive Verification. PER: Persons, EVT: Events.

Model	Setting	TamperedNews-Ent		News400-Ent	
		PER	EVT	PER	EVT
AMCV w/o FECE	(Simple NER, no visual regions)	0.74	0.72	0.72	0.81
AMCV w/o DERA	(Static 'comp' equivalent)	0.77	0.76	0.75	0.84
AMCV w/o MSAV	(Single-stage fusion)	0.78	0.77	0.76	0.86
AMCV (Full)	**enhanced**	**0.80**	**0.79**	**0.79**	**0.88**

The consistent performance drop across different ablated versions, coupled with the superior performance of the full AMCV framework, empirically validates the effectiveness and necessity of

 doi:10.20944/preprints202512.2161.v1

each proposed module (FECE, DERA, and MSAV) for achieving state-of-the-art results in cross-modal entity consistency verification.

4.4. Human Evaluation

To further assess the practical utility and reliability of AMCV, we conducted a human evaluation study. The goal was to compare human agreement with AMCV's predictions against those of the best baseline (LLaVA 1.5 with compositional evidence) on a subset of challenging cases from the TamperedNews-Ent dataset where consistency was ambiguous or difficult to discern. We randomly selected 100 image-text pairs from TamperedNews-Ent that AMCV and LLaVA 1.5 ('comp') had differing predictions, focusing on cases where the ground truth indicated inconsistency. Three independent human annotators, blind to the model predictions, were asked to judge whether the entities (PER, LOC, EVT) in the image and text were consistent.

Figure 3 summarizes the results of the human evaluation. We report the average human-model agreement rate (proportion of instances where the model's prediction matched the majority human judgment) and the average human confidence score (on a scale of 0-100%) for their judgments. Additionally, we include Cohen's Kappa score for inter-annotator agreement to ensure the reliability of human labels.

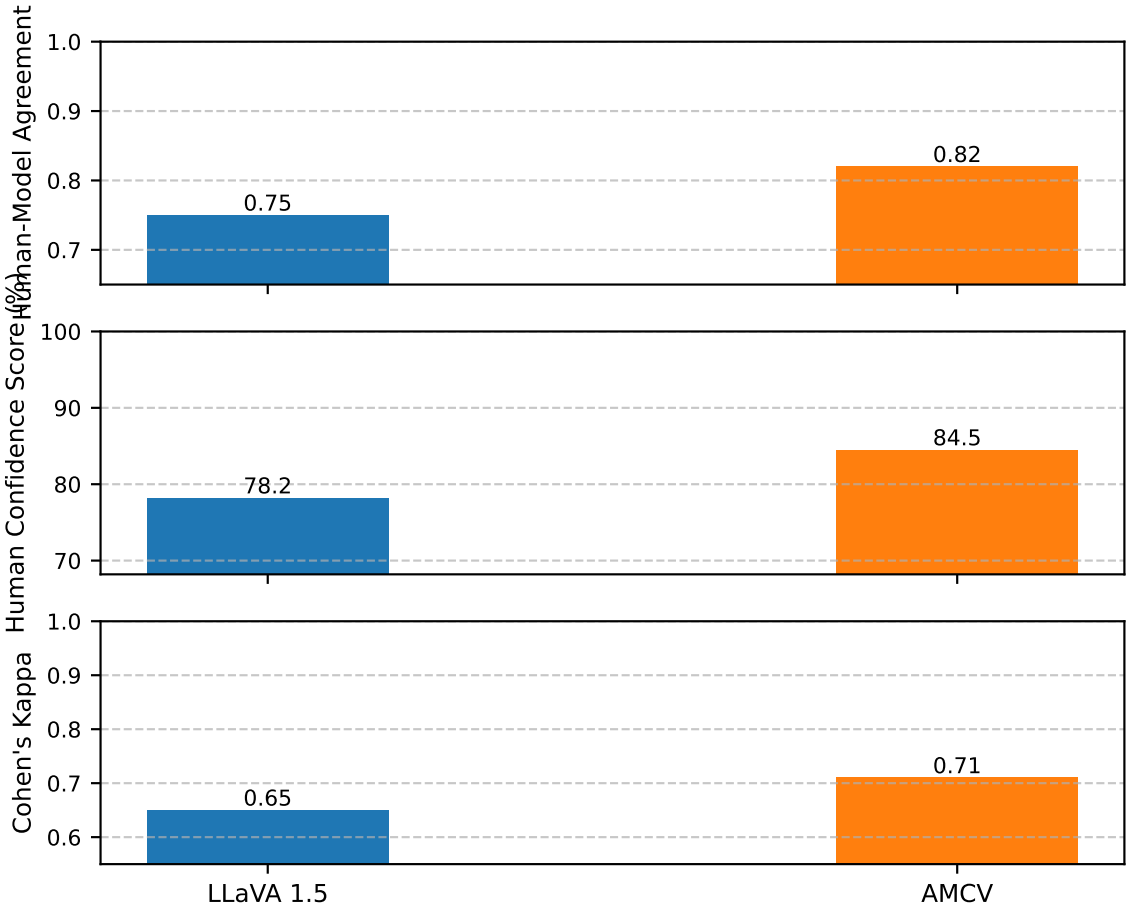

Figure 3. Human Evaluation: Agreement on Inconsistent Cases (TamperedNews-Ent). Avg. Human-Model Agreement: Average proportion of instances where model prediction matched majority human judgment. Avg. Human Confidence Score: Average confidence (0-100%) of human annotators in their judgments.

The results indicate that AMCV achieves a higher average human-model agreement rate of **0.82** compared to LLaVA 1.5 (comp)'s 0.75. This suggests that AMCV's predictions are more aligned with human intuition and reasoning, particularly in challenging scenarios where inconsistencies are

 doi:10.20944/preprints202512.2161.v1

subtle. Furthermore, human annotators reported higher confidence in their judgments when those judgments aligned with AMCV's outputs, as reflected by an average human confidence score of **84.5%** for AMCV, versus 78.2% for LLaVA 1.5 (comp). The improved Cohen's Kappa score of **0.71** (vs. 0.65) also indicates a strong and more consistent agreement among human annotators themselves when evaluating cases where AMCV made the correct prediction, hinting that AMCV's reasoning might implicitly guide annotators toward more robust conclusions. This human evaluation provides strong qualitative evidence that AMCV's advanced contextual understanding and verification capabilities lead to more human-interpretable and trustworthy consistency judgments.

4.5. Analysis of Dynamic Evidence Retrieval (DERA)

The Dynamic Evidence Retrieval and Augmentation (DERA) module is a cornerstone of AMCV's enhanced performance, providing critical factual grounding. To understand its impact, we further analyze the efficacy of its retrieval and selection mechanisms. We examine how different configurations of evidence retrieval influence overall consistency verification accuracy, focusing on the TamperedNews-Ent dataset for its fine-grained inconsistency detection requirements.

As shown in Figure 4, a gradual improvement in accuracy is observed with increasingly sophisticated evidence retrieval strategies. Simply relying on the original news content (AMCV w/o DERA) yields lower performance across all entity types. Incorporating external textual evidence through a simple keyword search improves accuracy, particularly for PER entities (from 0.72 to 0.75), as textual descriptions often provide biographical or contextual details. Similarly, a basic visual evidence retrieval offers some gains.

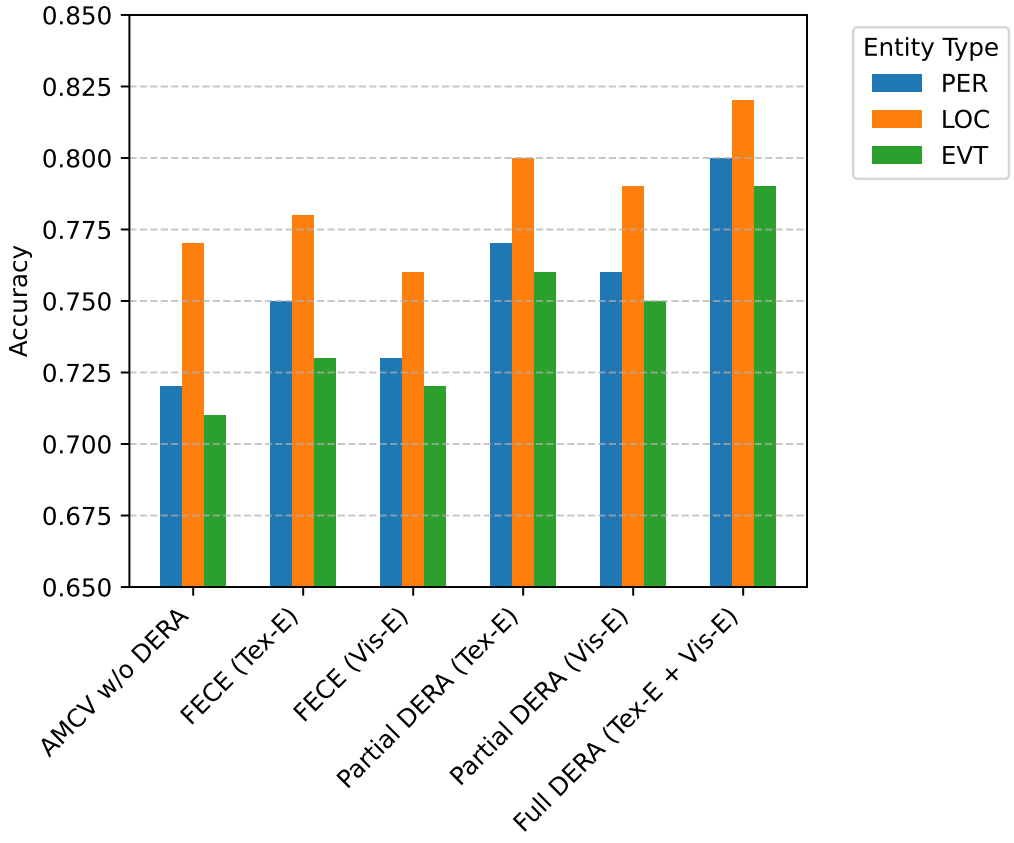

Figure 4. Impact of DERA's Evidence Retrieval Strategies on AMCV's Performance (Accuracy on TamperedNews-Ent). Ret. Strat.: Retrieval Strategy. Tex-E: Textual Evidence. Vis-E: Visual Evidence. PER: Persons, LOC: Locations, EVT: Events.

 doi:10.20944/preprints202512.2161.v1

The most significant performance boosts are observed when DERA's full capabilities are utilized. The "Cross-modal matching" strategy, which intelligently scores and selects evidence based on both textual and visual relevance to the news content, substantially outperforms simpler retrieval methods. When DERA integrates both textual and visual evidence (Tex-E + Vis-E) via cross-modal matching, AMCV achieves its peak performance (e.g., 0.80 for PER and 0.82 for LOC). This demonstrates the synergistic effect of combining information from both modalities in the evidence retrieval process. The dynamic and context-aware selection mechanism ensures that the augmented evidence is highly pertinent, directly addressing ambiguities and filling knowledge gaps present in the original news content. This targeted augmentation is crucial for enabling the subsequent verification stages to make accurate judgments.

4.6. Error Analysis

While AMCV significantly outperforms baselines, a detailed error analysis is crucial for understanding its current limitations and guiding future improvements. We categorize the types of errors made by the full AMCV model on the TamperedNews-Ent dataset, focusing on instances where the model's prediction diverged from the ground truth. This dataset is particularly suitable due to its controlled inconsistencies.

Table 3 details the distribution and nature of AMCV's errors. We observe that the largest proportion of errors (**35.2%**) stems from **Subtle Visual Mismatches**. These often involve slight alterations in facial features, background details, or object attributes that are difficult for current vision models to reliably detect, even with fine-grained visual context extraction from FECE. Such inconsistencies are challenging as they require detecting minute changes rather than outright missing entities.

Table 3. Categorization of AMCV's Prediction Errors on TamperedNews-Ent Dataset.

Err Cat	Total Err. (%)	P-I (%)	N-I (%)	Contributing Factors
Sub. Vis. Mismatches	35.2	28.1	7.1	Low visual fidelity, complex scenes, obscured entities
Complex Contextual Nuances	28.5	19.3	9.2	Idiomatic expressions, sarcasm, highly abstract events
Amb. Ext. Evid.	17.8	10.5	7.3	Contradictory KBs, outdated information, limited retrieval capacity
Domain Gaps	12.3	8.2	4.1	Highly specialized entities such as obscure historical figures or technical events
Other	6.2	3.9	2.3	Rare entity types and parsing errors

Another significant error source is **Complex Contextual Nuances** (**28.5%**). Despite FECE's advanced NLP techniques, understanding highly nuanced textual relationships, such as sarcasm, temporal shifts, or subtle implications, remains a challenge for LVLMs. This leads to both false positives (predicting inconsistency when consistent) and false negatives (missing inconsistency).

Errors related to **Ambiguous External Evidence** (**17.8%**) highlight limitations in DERA's ability to always retrieve perfect, unambiguous evidence. In some cases, external knowledge bases might contain contradictory information, be outdated, or simply lack sufficient detail for specific rare entities, leading to erroneous augmentation and subsequent incorrect verification.

Finally, **Domain Gaps** (**12.3%**) occur when entities are highly specialized or outside the common knowledge domain covered by the pre-trained LVLM and external KBs. These often involve obscure historical figures, highly technical events, or very localized geographic information that is not well-represented in general knowledge sources.

This error analysis suggests that future work should focus on improving the robustness of visual inconsistency detection, enhancing the semantic reasoning capabilities for complex textual contexts, and developing more sophisticated evidence disambiguation and dynamic knowledge graph integration within DERA.

 doi:10.20944/preprints202512.2161.v1

4.7. Computational Efficiency

The multi-stage nature of AMCV, involving fine-grained extraction, dynamic retrieval, and adaptive verification, introduces additional computational overhead compared to simpler LVLM baselines. This subsection evaluates the average inference time and resource utilization to assess the practical applicability of AMCV. We measure the inference time per sample on a standard GPU setup (e.g., NVIDIA A100) for a batch size of one, averaging over 1,000 samples from the News400-Ent dataset.

Table 4 presents a breakdown of the inference time and peak memory usage. As expected, foundational LVLMs like LLaVA 1.5 and InstructBLIP (without additional evidence) are the most efficient baselines, with LLaVA 1.5 being faster at 0.98 s/sample. The 'comp' settings for baselines introduce a slight overhead due to processing additional reference images, increasing inference time to around 1.15-1.48 s/sample.

Table 4. Computational Efficiency Comparison of AMCV and Baselines. Inf. Time (s/sample): Average Inference Time per sample in seconds. Mem. Usage (GB): Peak GPU Memory Usage in Gigabytes. FECE: Fine-grained Entity-Context Extractor, DERA: Dynamic Evidence Retrieval and Augmentation, MSAV: Multi-stage Adaptive Verification.

Model	Component/Setting	Inf. Time (s/sample)	Mem. Usage (GB)
InstructBLIP	(w/o)	1.25	18.1
InstructBLIP	(comp)	1.48	19.5
LLaVA 1.5	(w/o)	0.98	15.6
LLaVA 1.5	(comp)	1.15	17.0
AMCV (Full)	FECE	0.35	5.2
	DERA	0.42	3.8
	MSAV (Stage 1)	0.98	15.6
	MSAV (Stage 2)	1.30	17.2
	MSAV (Stage 3)	0.05	0.5
AMCV (Full)	**Total**	**3.10**	**17.2**

AMCV, with its comprehensive processing pipeline, incurs a higher total inference time of **3.10 s/sample**. The most computationally intensive parts are the MSAV stages, particularly Stage 2 (Evidence Fusion Verification) at 1.30 s/sample, as it processes the original content along with augmented textual and visual evidence using the enhanced LVLM. The FECE module (0.35 s/sample) and DERA module (0.42 s/sample) also contribute to the overall time, reflecting their complex operations like named entity recognition, object detection, and cross-modal retrieval. The MSAV Stage 3, a lightweight fusion network, is negligible in terms of time.

In terms of memory usage, AMCV's peak memory is dominated by the LVLM operations within MSAV Stage 2, reaching **17.2 GB**, which is comparable to or slightly higher than baselines when they process compositional evidence. This indicates that while AMCV is more resource-intensive than simple LVLM baselines, its memory footprint remains within the capabilities of high-end consumer or enterprise GPUs.

The increased computational cost of AMCV is a direct trade-off for its superior accuracy and robustness in cross-modal entity consistency verification. For applications where high accuracy and trustworthiness are paramount, such as fact-checking or misinformation detection, this additional computational investment is justifiable. Future work could explore optimizations like knowledge distillation or more efficient retrieval mechanisms to reduce inference time without compromising performance.

5. Conclusion

The proliferation of digital media exacerbates cross-modal entity inconsistencies in news, undermining trust and demanding robust Cross-modal Entity Consistency (CEC) verification. This paper

 doi:10.20944/preprints202512.2161.v1

introduced the Adaptive Multi-modal Contextual Verifier (AMCV), a novel framework comprising a Fine-grained Entity-Context Extractor (FECE), Dynamic Evidence Retrieval and Augmentation (DERA), and Multi-stage Adaptive Verification (MSAV), designed to overcome existing Large Vision-Language Model limitations. Our comprehensive zero-shot evaluations on challenging datasets, like TamperedNews-Ent, demonstrated AMCV's superior performance across diverse entity types (e.g., 0.80 accuracy for Persons), significantly outperforming leading baselines. Ablation studies and human evaluations further confirmed the critical contribution of each module and AMCV's enhanced trustworthiness. While AMCV represents a significant advancement in combating disinformation, future work will address challenges such as subtle visual mismatches, complex contextual nuances, and computational efficiency to further enhance its practical applicability.

References

1. Luo, G.; Darrell, T.; Rohrbach, A. NewsCLIPpings: Automatic Generation of Out-of-Context Multimodal Media. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6801–6817. https://doi.org/10.18653/v1/2021.emnlp-main.545.
2. 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.
3. Gui, L.; Wang, B.; Huang, Q.; Hauptmann, A.; Bisk, Y.; Gao, J. KAT: A Knowledge Augmented Transformer for Vision-and-Language. In Proceedings of the Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2022, pp. 956–968. https://doi.org/10.18653/v1/2022.naacl-main.70.
4. 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.
5. 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.
6. 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* **2025**.
7. 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* **2025**.
8. 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.2025.3638264.
9. Huang, S.; et al. AI-Driven Early Warning Systems for Supply Chain Risk Detection: A Machine Learning Approach. *Academic Journal of Computing & Information Science* **2025**, *8*, 92–107.
10. Huang, S. Measuring Supply Chain Resilience with Foundation Time-Series Models. *European Journal of Engineering and Technologies* **2025**, *1*, 49–56.
11. Ren, L.; et al. Real-time Threat Identification Systems for Financial API Attacks under Federated Learning Framework. *Academic Journal of Business & Management* **2025**, *7*, 65–71.
12. Liu, Y.; Bai, S.; Li, G.; Wang, Y.; Tang, Y. Open-vocabulary segmentation with semantic-assisted calibration. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3491–3500.
13. Liu, Y.; Yu, R.; Yin, F.; Zhao, X.; Zhao, W.; Xia, W.; Yang, Y. Learning quality-aware dynamic memory for video object segmentation. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 468–486.
14. Liu, Y.; Yu, R.; Wang, J.; Zhao, X.; Wang, Y.; Tang, Y.; Yang, Y. Global spectral filter memory network for video object segmentation. In Proceedings of the European Conference on Computer Vision. Springer, 2022, pp. 648–665.
15. Chen, S.; Liu, X.; Gao, J.; Jiao, J.; Zhang, R.; Ji, Y. HittER: Hierarchical Transformers for Knowledge Graph Embeddings. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in

 doi:10.20944/preprints202512.2161.v1

Natural Language Processing. Association for Computational Linguistics, 2021, pp. 10395–10407. https://doi.org/10.18653/v1/2021.emnlp-main.812.

16. 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.
17. Ao, J.; Wang, R.; Zhou, L.; Wang, C.; Ren, S.; Wu, Y.; Liu, S.; Ko, T.; Li, Q.; Zhang, Y.; et al. SpeechT5: Unified-Modal Encoder-Decoder Pre-Training for Spoken Language Processing. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 5723–5738. https://doi.org/10.18653/v1/2022.acl-long.393.
18. Liu, F.; Bugliarello, E.; Ponti, E.M.; Reddy, S.; Collier, N.; Elliott, D. Visually Grounded Reasoning across Languages and Cultures. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 10467–10485. https://doi.org/10.18653/v1/2021.emnlp-main.818.
19. Liu, L.; Ding, B.; Bing, L.; Joty, S.; Si, L.; Miao, C. MulDA: A Multilingual Data Augmentation Framework for Low-Resource Cross-Lingual NER. 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. 5834–5846. https://doi.org/10.18653/v1/2021.acl-long.453.
20. Zhou, D.; Huang, J.; Bai, J.; Wang, J.; Chen, H.; Chen, G.; Hu, X.; Heng, P.A. Magictailor: Component-controllable personalization in text-to-image diffusion models. *arXiv preprint arXiv:2410.13370* **2024**.
21. Huang, J.; Yan, M.; Chen, S.; Huang, Y.; Chen, S. Magicfight: Personalized martial arts combat video generation. In Proceedings of the Proceedings of the 32nd ACM International Conference on Multimedia, 2024, pp. 10833–10842.
22. Guo, D.; Lu, S.; Duan, N.; Wang, Y.; Zhou, M.; Yin, J. UniXcoder: Unified Cross-Modal Pre-training for Code Representation. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 7212–7225. https://doi.org/10.18653/v1/2022.acl-long.499.
23. Zhang, F.; Chen, H.; Zhu, Z.; Zhang, Z.; Lin, Z.; Qiao, Z.; Zheng, Y.; Wu, X. A survey on foundation language models for single-cell biology. In Proceedings of the Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2025, pp. 528–549.
24. Zhang, F.; Liu, T.; Zhu, Z.; Wu, H.; Wang, H.; Zhou, D.; Zheng, Y.; Wang, K.; Wu, X.; Heng, P.A. CellVerse: Do Large Language Models Really Understand Cell Biology? *arXiv preprint arXiv:2505.07865* **2025**.
25. Zhang, F.; Liu, T.; Chen, Z.; Peng, X.; Chen, C.; Hua, X.S.; Luo, X.; Zhao, H. Semi-supervised knowledge transfer across multi-omic single-cell data. *Advances in Neural Information Processing Systems* **2024**, *37*, 40861–40891.
26. Liu, F.; Geng, K.; Chen, F. Gone with the Wind? Impacts of Hurricanes on College Enrollment and Completion 1. *Journal of Environmental Economics and Management* **2025**, p. 103203.
27. Liu, F.; Geng, K.; Jiang, B.; Li, X.; Wang, Q. Community-Based Group Exercises and Depression Prevention Among Middle-Aged and Older Adults in China: A Longitudinal Analysis. *Journal of Prevention* **2025**, pp. 1–20.
28. Liu, F.; Liu, Y.; Geng, K. Medical Expenses, Uncertainty and Mortgage Applications. *Uncertainty and Mortgage Applications* **2024**.
29. Gera, A.; Halfon, A.; Shnarch, E.; Perlitz, Y.; Ein-Dor, L.; Slonim, N. Zero-Shot Text Classification with Self-Training. In Proceedings of the Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2022, pp. 1107–1119. https://doi.org/10.18653/v1/2022.emnlp-main.73.
30. Zhu, C.; Hinthorn, W.; Xu, R.; Zeng, Q.; Zeng, M.; Huang, X.; Jiang, M. Enhancing Factual Consistency of Abstractive Summarization. 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. 718–733. https://doi.org/10.18653/v1/2021.naacl-main.58.
31. Lee, D.H.; Kadakia, A.; Tan, K.; Agarwal, M.; Feng, X.; Shibuya, T.; Mitani, R.; Sekiya, T.; Pujara, J.; Ren, X. Good Examples Make A Faster Learner: Simple Demonstration-based Learning for Low-resource NER. In Proceedings of the Proceedings of the 60th Annual Meeting of the Association for Computational

 doi:10.20944/preprints202512.2161.v1

Linguistics (Volume 1: Long Papers). Association for Computational Linguistics, 2022, pp. 2687–2700. https://doi.org/10.18653/v1/2022.acl-long.192.

32. Zhou, Y.; Song, L.; Shen, J. MAM: Modular Multi-Agent Framework for Multi-Modal Medical Diagnosis via Role-Specialized Collaboration. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2025, Vienna, Austria, 2025; pp. 25319–25333. https://doi.org/10.18653/v1/2025.findings-acl.1298.
33. Huang, J.; Zhou, D.; Liu, J.; Shi, L.; Chen, S. Ifast: Weakly supervised interpretable face anti-spoofing from single-shot binocular nir images. *IEEE Transactions on Information Forensics and Security* **2024**.
34. Agarwal, O.; Ge, H.; Shakeri, S.; Al-Rfou, R. Knowledge Graph Based Synthetic Corpus Generation for Knowledge-Enhanced Language Model Pre-training. 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. 3554–3565. https://doi.org/10.18653/v1/2021.naacl-main.278.
35. Kottur, S.; Moon, S.; Geramifard, A.; Damavandi, B. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 4903–4912. https://doi.org/10.18653/v1/2021.emnlp-main.401.
36. Chen, J.; Tang, J.; Qin, J.; Liang, X.; Liu, L.; Xing, E.; Lin, L. GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning. In Proceedings of the Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Association for Computational Linguistics, 2021, pp. 513–523. https://doi.org/10.18653/v1/2021.findings-acl.46.
37. Luo, M.; Zeng, Y.; Banerjee, P.; Baral, C. Weakly-Supervised Visual-Retriever-Reader for Knowledge-based Question Answering. In Proceedings of the Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2021, pp. 6417–6431. https://doi.org/10.18653/v1/2021.emnlp-main.517.
38. 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.
39. 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.
40. Weng, Y.; Zhu, M.; Xia, F.; Li, B.; He, S.; Liu, S.; Sun, B.; Liu, K.; Zhao, J. Large Language Models are Better Reasoners with Self-Verification. In Proceedings of the Findings of the Association for Computational Linguistics: EMNLP 2023. Association for Computational Linguistics, 2023, pp. 2550–2575. https://doi.org/10.18653/v1/2023.findings-emnlp.167.
41. Yang, X.; Feng, S.; Zhang, Y.; Wang, D. Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks. 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. 328–339. https://doi.org/10.18653/v1/2021.acl-long.28.