Adaptive Latent Interaction Reasoning for Multimodal Misinformation Analysis

Tyler Anderson *, Madeline Brooks, Ava Martinez, Jordan Williams

Posted Date: 22 December 2025

doi: 10.20944/preprints202512.1914.v1

Keywords: multimodal misinformation analysis; fake news detection; contrastive representation learning; cross-modal reasoning; social media

 doi:10.20944/preprints202512.1914.v1

Article

Adaptive Latent Interaction Reasoning for Multimodal Misinformation Analysis

Tyler Anderson *, Madeline Brooks, Ava Martinez and Jordan Williams

University of Central Oklahoma, USA
* Correspondence: tanderson@uco.edu

Abstract

The rapid growth of online social platforms has fundamentally transformed the way information is produced, disseminated, and consumed, while simultaneously amplifying the societal impact of misleading and fabricated content. In response to this challenge, multimodal fake news detection has emerged as a critical research problem, aiming to jointly leverage textual and visual signals embedded in social media posts. Existing methods predominantly rely on direct fusion of unimodal representations or shallow cross-modal interactions, which often fail to explicitly model the semantic alignment and latent inconsistencies across modalities. In particular, the potential of contrastive learning paradigms for learning robust and semantically grounded multimodal representations in fake news scenarios remains underexplored. In this work, we introduce **ALIGNER**, an **A**daptive **L**atent **I**nteraction **G**uided co**N**trastiv**E** **R**easoning framework designed for multimodal fake news detection. ALIGNER adopts a dual-encoder architecture to learn modality-specific semantic representations and employs cross-modal contrastive learning to explicitly align visual and textual semantics. To address the inherent noise and ambiguity of image–text associations in real-world fake news data, we further propose a latent consistency objective that relaxes the rigid one-hot supervision imposed by conventional contrastive losses. This auxiliary learning signal enables the model to capture fine-grained semantic relatedness among unpaired or weakly related multimodal samples. Building upon the aligned unimodal features, ALIGNER incorporates a dedicated cross-modal interaction module to capture higher-order correlations between visual and linguistic representations. Moreover, we design an attention-based aggregation mechanism equipped with an explicit guidance signal to adaptively weigh the contributions of different modalities during decision making, thereby enhancing both effectiveness and interpretability. Extensive experiments conducted on two widely adopted benchmarks, Twitter and Weibo, demonstrate that ALIGNER consistently surpasses existing state-of-the-art approaches by a substantial margin, highlighting the advantages of adaptive contrastive reasoning for multimodal fake news detection.

Keywords: multimodal misinformation analysis; fake news detection; contrastive representation learning; cross-modal reasoning; social media

1. Introduction

The unprecedented proliferation of Online Social Networks (OSNs), such as Twitter and Weibo, has profoundly reshaped the modern information ecosystem. These platforms enable users to rapidly share news, opinions, and multimedia content, significantly lowering the barrier for information dissemination. While this democratization of content creation brings undeniable benefits, it also introduces serious challenges, among which the large-scale spread of fake news has become a pressing societal concern [31]. Misinformation disseminated through social media has been shown to influence public opinion, undermine trust in institutions, and even cause tangible harm in domains such as public health and political stability. Consequently, the development of reliable automatic fake news detection systems has attracted substantial attention from both academia and industry.

 doi:10.20944/preprints202512.1914.v1

Early research efforts on fake news detection primarily focused on analyzing textual content alone. Traditional machine learning methods, including decision tree classifiers and handcrafted linguistic features, were initially explored to distinguish fake news from genuine content [40]. With the advent of deep learning, neural architectures such as convolutional neural networks (CNNs) and recurrent models have been employed to capture more complex semantic patterns in text [47]. Although these approaches have demonstrated promising performance, they fundamentally overlook the multimodal nature of social media posts, where textual descriptions are frequently accompanied by images or other visual elements. Relying solely on textual signals is therefore insufficient for fully understanding and verifying the credibility of online information.

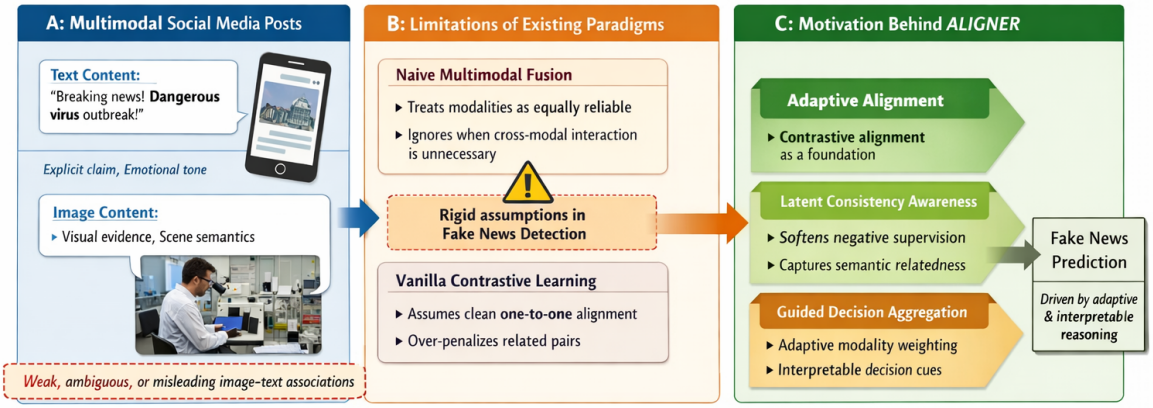

Figure 1. Motivation of ALIGNER: addressing ambiguous image–text associations through adaptive alignment and guided multimodal reasoning.

Recognizing this limitation, recent studies have shifted towards multimodal fake news detection, aiming to jointly exploit textual and visual information [36,44]. By fusing representations extracted from multiple modalities, these methods seek to construct more informative and discriminative post-level representations. Despite their progress, many existing approaches adopt relatively coarse fusion strategies, such as feature concatenation or shallow attention mechanisms, without explicitly modeling the semantic alignment between modalities. As a result, the learned multimodal representations may fail to capture subtle inconsistencies or complementary cues that are crucial for fake news identification. More importantly, not all modalities contribute equally in every scenario. In some cases, textual content alone may be sufficiently implausible to indicate misinformation, whereas in other situations, the discrepancy between visual and textual information plays a decisive role in exposing fake news [6,24]. This observation highlights a fundamental challenge: multimodal fake news detection should not merely aggregate information from different sources, but should also reason about when and how each modality influences the final decision. Understanding the varying roles of unimodal and cross-modal features, as well as making this decision process interpretable, remains an open research problem.

In parallel, contrastive learning has emerged as a powerful paradigm for representation learning across modalities. A series of contrastive vision–language pre-training frameworks have demonstrated remarkable success by aligning image and text representations in a shared embedding space [1,12,16,23,38,39,48]. By pulling together representations of matched image–text pairs while pushing apart mismatched ones, contrastive objectives encourage encoders to capture high-level semantic correspondence. These advances suggest that contrastive learning could serve as an effective foundation for multimodal fake news detection.

However, directly applying standard contrastive learning objectives to fake news detection is non-trivial. Existing contrastive frameworks typically rely on one-hot supervision, treating all unmatched image–text pairs as equally negative [10,39]. This assumption is often violated in real-world fake news datasets, where image–text associations can be weak, ambiguous, or even intentionally misleading. Furthermore, different multimodal posts discussing the same event may share partial semantic overlap, yet are still considered negative pairs under conventional contrastive learning. Such rigid supervision

 doi:10.20944/preprints202512.1914.v1

may lead to over-penalization of semantically related samples and ultimately harm generalization performance.

To address these challenges, we propose **ALIGNER**, an adaptive latent interaction guided contrastive reasoning framework for multimodal fake news detection. ALIGNER adopts a dual-encoder architecture to separately model visual and linguistic semantics, establishing two complementary semantic spaces. A cross-modal contrastive objective is employed to encourage alignment between modalities, ensuring that corresponding image and text representations are semantically coherent. Beyond this basic alignment, we introduce an auxiliary latent consistency learning task that provides soft supervision over negative samples. This task captures implicit semantic relatedness among multimodal instances, thereby alleviating the overly strict constraints imposed by one-hot contrastive labels.

Following representation alignment, ALIGNER further incorporates a cross-modal interaction module to explicitly model correlations between visual and textual features. Rather than treating cross-modal information as uniformly beneficial, we design an attention-based aggregation mechanism that adaptively weighs unimodal and cross-modal signals. Inspired by prior work [6], we integrate an attention guidance strategy that quantifies modality ambiguity by estimating the divergence between visual and textual representations. This guidance signal enables the model to dynamically adjust modality importance, leading to more robust and interpretable decision making.

In summary, ALIGNER advances multimodal fake news detection by unifying adaptive contrastive alignment, latent semantic consistency modeling, and guided cross-modal reasoning within a single framework. Through this design, the proposed approach is able to learn more reliable multimodal representations and effectively leverage them for misinformation analysis. Extensive experimental evaluations on the Twitter and Weibo datasets validate the effectiveness of ALIGNER, demonstrating consistent improvements over existing state-of-the-art methods.

The main contributions of this paper are summarized as follows:

- We propose ALIGNER, an adaptive contrastive reasoning framework that explicitly addresses the challenges of multimodal fake news detection.
- We introduce a latent consistency learning objective to soften the supervision over negative samples in cross-modal contrastive learning.
- We design a guided attention mechanism to adaptively and interpretably aggregate unimodal and cross-modal features.
- Extensive experiments on Twitter and Weibo datasets demonstrate that ALIGNER achieves superior performance compared to prior state-of-the-art approaches.

2. Related Works

2.1. Fake News Detection

Early studies on fake news detection predominantly focused on textual content, motivated by the observation that linguistic signals often carry explicit or implicit cues indicative of misinformation. Representative approaches analyze the semantic coherence, lexical choices, and discourse patterns present in news posts. For instance, [22] propose a neural generative framework that models historical user responses to uncover latent propagation patterns, thereby assisting the detection of fake news through indirect semantic supervision. Such methods highlight the importance of contextual language modeling beyond surface-level word statistics. In addition to generative approaches, a large body of work explores explainable and interpretable textual features. TM [2] leverages lexical, syntactic, and semantic properties of text to improve detection accuracy while providing insights into model decisions. Other studies investigate logical consistency [11], writing style regularities [20], and rhetorical structures [7], under the assumption that deceptive content exhibits distinctive linguistic patterns. While effective in controlled settings, these approaches remain limited by their exclusive reliance on textual modality, which constrains their robustness in multimodal social media environments.

 doi:10.20944/preprints202512.1914.v1

Beyond text, visual information has also been explored as a standalone signal for fake news detection, particularly in image-rich social media platforms. Early work such as [18] demonstrates that fake and real news often exhibit different visual dissemination behaviors, including reposting frequency and temporal diffusion patterns. These findings suggest that images are not merely auxiliary decorations but can serve as critical forensic evidence. Subsequent studies further examine intrinsic visual characteristics. MVNN [21] exploits both spatial-domain visual features and frequency-domain signals to capture subtle manipulation traces, offering a more comprehensive forensic perspective. Despite their effectiveness in identifying visually manipulated or misleading images, purely visual approaches struggle when images are reused legitimately or lack explicit manipulation cues. More importantly, these methods disregard semantic interactions between images and accompanying text, which are often crucial for understanding misinformation intent.

Although unimodal approaches have laid the foundation for automatic fake news detection, their inherent limitations become increasingly evident in real-world multimodal scenarios. Social media posts frequently combine text and images in complex ways, where neither modality alone provides sufficient evidence for accurate classification. Unimodal models fail to capture semantic contradictions, complementarities, or contextual dependencies across modalities, leading to ambiguous or incorrect predictions. Furthermore, unimodal systems are particularly vulnerable to adversarial manipulation, where misinformation creators exploit modality gaps to evade detection. For instance, benign-looking images may be paired with deceptive textual claims, or plausible text may be accompanied by irrelevant visuals. These challenges motivate the need for multimodal reasoning frameworks that can jointly analyze and interpret heterogeneous signals rather than treating them in isolation.

The shortcomings of unimodal methods have driven a paradigm shift toward multimodal fake news detection, where the goal is not only to aggregate multiple information sources but also to reason about their relationships. Effective multimodal detection requires modeling both intra-modal semantics and inter-modal consistency, enabling systems to identify discrepancies that are otherwise imperceptible within a single modality. This observation forms the foundation for more advanced multimodal approaches, which we discuss next. Recent advances in multimodal fake news detection emphasize the importance of learning discriminative cross-modal patterns. EANN [42] introduces an event discriminator to encourage event-invariant feature learning, thereby enhancing generalization across diverse news topics. MVAE [36] adopts a multimodal variational autoencoder to jointly model textual and visual latent variables and reconstruct both modalities, capturing shared semantic representations.

Building upon attention-based fusion strategies, MCAN [44] stacks multiple co-attention layers to enable fine-grained interactions between text and image features. These methods demonstrate that explicit cross-modal modeling significantly improves detection performance compared to unimodal baselines. However, most fusion-based approaches rely on task-specific architectures and lack principled representation learning objectives that generalize beyond supervised settings. An emerging line of work recognizes that modalities do not contribute equally to fake news detection across instances. CAFE [6] quantifies cross-modal ambiguity by computing the Kullback-Leibler (KL) divergence between unimodal feature distributions and uses this score to dynamically reweight modality contributions. This strategy explicitly accounts for modality disagreement and improves interpretability. Similarly, LIIMR [24] identifies the modality that provides stronger predictive confidence and prioritizes it during decision making. These approaches move beyond naive fusion by incorporating modality-aware reasoning. In contrast, ALIGNER further integrates modality importance estimation with contrastive alignment and latent semantic consistency, enabling adaptive and interpretable multimodal inference within a unified framework.

2.2. Contrastive Learning

Contrastive Learning in Vision and Language

Contrastive learning has emerged as a dominant paradigm for self-supervised and representation learning in both computer vision and natural language processing. In the vision domain, frameworks

 doi:10.20944/preprints202512.1914.v1

such as MoCo [12], SimCLR [4], and subsequent empirical studies [5] demonstrate that instance discrimination objectives can produce highly transferable visual representations. In parallel, contrastive objectives have been successfully adapted to NLP tasks, as evidenced by SimCSE [9] and ConSERT [46], which improve sentence embeddings through representation alignment. These successes highlight the generality of contrastive learning as a mechanism for capturing semantic similarity and structure without heavy reliance on manual annotation. The core idea of pulling semantically related samples closer while pushing unrelated ones apart provides a flexible foundation for multimodal extension.

Contrastive learning has been further extended to vision-language representation learning, leading to a series of influential multimodal pre-training models. WenLan [15] proposes a two-tower architecture tailored for Chinese multimodal data, adapting the MoCo framework to cross-modal settings. CLIP [23] and ALIGN [16] demonstrate that large-scale contrastive pre-training on noisy web data can yield powerful aligned representations, enabling zero-shot generalization across diverse tasks. Subsequent works such as ALBEF [39] introduce additional cross-attention fusion modules on top of contrastively aligned encoders, while incorporating hard negative mining strategies to enhance discrimination. Related approaches including BLIP [38] and VLMo [1] further refine these ideas, achieving strong performance across a wide range of multimodal benchmarks.

Hybrid Contrastive and Generative Objectives

Beyond purely discriminative objectives, recent models explore the combination of contrastive and generative learning. CoCa [48] unifies contrastive alignment with caption generation in a modified encoder-decoder architecture, demonstrating that hybrid objectives can improve representation quality and downstream adaptability. These advances suggest that contrastive learning benefits from auxiliary semantic supervision that goes beyond binary alignment signals. However, most existing vision-language contrastive frameworks assume clean, high-quality image-text pairs, an assumption that does not hold in multimodal fake news detection. Social media data often contains weakly related or intentionally misleading image-text associations, posing unique challenges for direct adoption of standard contrastive objectives.

In this work, ALIGNER builds upon the strengths of contrastive vision-language learning while explicitly addressing its limitations in fake news scenarios. We adopt an image-text contrastive (ITC) objective within a dual-encoder architecture to establish a shared latent embedding space for visual and textual representations. Unlike conventional approaches, ALIGNER incorporates an auxiliary cross-modal consistency learning task that estimates semantic similarity between images and texts and provides soft supervision signals for contrastive alignment.

By relaxing rigid one-hot negative assumptions and accounting for latent semantic relatedness, ALIGNER achieves more robust and task-compatible representation learning. This design enables the model to effectively leverage contrastive learning for multimodal fake news detection, bridging the gap between large-scale multimodal pre-training paradigms and real-world misinformation analysis.

 doi:10.20944/preprints202512.1914.v1

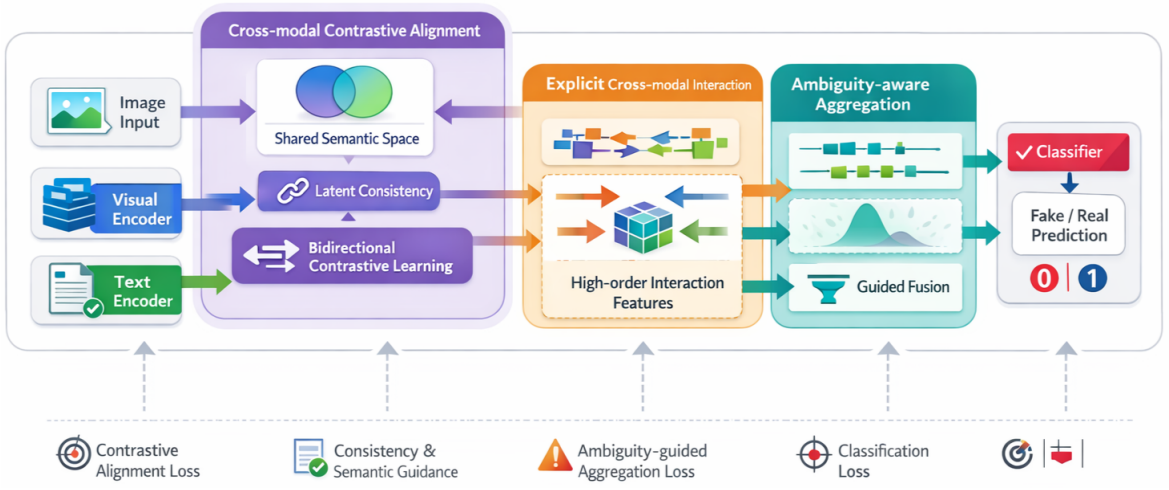

Figure 2. ALIGNER framework overview. Paired image–text inputs are aligned via contrastive learning, interact explicitly across modalities, and are aggregated in an ambiguity-aware manner for fake news prediction.

3. Methodology

In this section, we present **ALIGNER**, a unified and adaptive multimodal fake news detection framework grounded in cross-modal contrastive representation learning and guided aggregation. ALIGNER is designed to explicitly address the inherent semantic misalignment, modality ambiguity, and weak supervision issues that commonly arise in real-world multimodal news data. Given paired image–text inputs, ALIGNER first extracts modality-specific representations using pretrained encoders, and then progressively refines these representations through contrastive alignment, latent semantic consistency modeling, cross-modal interaction learning, and ambiguity-aware aggregation. The overall learning objective jointly optimizes alignment quality, semantic robustness, and decision interpretability.

3.1. Multimodal Input Formalization and Encoder Architecture

Let each multimodal news instance be denoted as $\mathbf{x} = [x^v, x^t] \in \mathcal{D}$, where x^v represents the visual content, x^t denotes the textual content, and $\mathcal{D}$ is the training dataset. The core objective of ALIGNER is to learn a discriminative function $f : (x^v, x^t) \rightarrow y$, where $y \in \{0, 1\}$ indicates whether the news is fake or real. Rather than directly operating on raw inputs, ALIGNER relies on pretrained modality-specific encoders to obtain high-quality unimodal representations, which serve as the foundation for subsequent cross-modal reasoning.

Visual Representation Learning

Given an image input x^v, we adopt ResNet [13] pretrained on ImageNet as the visual encoder due to its strong inductive bias for spatial feature extraction. The encoder maps x^v into a set of region-level feature vectors, which are subsequently aggregated through global average pooling. A fully connected projection layer is applied to transform the pooled visual features into a compact embedding $e^v \in \mathbb{R}^d$. This projection not only aligns dimensionality with the textual embedding space but also serves as a learnable bottleneck that facilitates semantic abstraction.

Textual Representation Learning

For textual input x^t, we employ BERT [35] as the backbone language encoder to capture contextualized semantic representations. The input text is tokenized using a predefined vocabulary and passed through multiple Transformer layers. The resulting sequence-level representation is obtained by pooling token embeddings, followed by a fully connected transformation to produce the final textual embedding $e^t \in \mathbb{R}^d$. This design allows ALIGNER to capture both syntactic structure and high-level semantic cues that are critical for fake news detection.

 doi:10.20944/preprints202512.1914.v1

3.2. Cross-modal Contrastive Alignment with Latent Consistency

Due to the substantial semantic gap between visual and textual modalities, directly fusing unimodal embeddings is insufficient for reliable multimodal reasoning. ALIGNER therefore adopts cross-modal contrastive learning as a principled mechanism to align heterogeneous representations within a shared latent space. However, unlike conventional contrastive frameworks that rely on rigid one-hot supervision, ALIGNER introduces an auxiliary consistency learning task to provide soft semantic guidance, thereby improving robustness in noisy multimodal settings.

Latent Cross-modal Consistency Learning

We formulate cross-modal consistency learning as a binary semantic matching task. Specifically, we construct an auxiliary dataset $\mathcal{D}' = [\mathcal{D}_{\text{pos}}, \mathcal{D}_{\text{neg}}]$, where each instance $\mathbf{x}' = [x^{v'}, x^{t'}]$ is labeled as $y' = 1$ if the image and text originate from the same real news, and $y' = 0$ otherwise. The corresponding unimodal embeddings $e^{v'}$ and $e^{t'}$ are projected into a shared semantic space using modality-specific multilayer perceptrons, yielding shared embeddings $e_s^{v'}$ and $e_s^{t'}$.

To explicitly enforce semantic consistency, we optimize the cosine embedding loss:

$$\mathcal{L}_{ITM} = \begin{cases} 1 - \cos(e_s^{v'}, e_s^{t'}), & y' = 1 \\ \max(0, \cos(e_s^{v'}, e_s^{t'}) - d), & y' = 0 \end{cases} \tag{1}$$

where $\cos(\cdot)$ denotes normalized cosine similarity and $d = 0.2$ is a margin hyperparameter. This task encourages semantically aligned multimodal pairs to be close in latent space while preventing excessive separation of weakly related pairs.

Bidirectional Image–Text Contrastive Learning

Given a batch of N paired samples $\{(x_i^v, x_i^t)\}_{i=1}^N$, ALIGNER computes normalized embeddings $\{e_i^v, e_i^t\}_{i=1}^N$. The image–text contrastive objective aims to correctly identify true correspondences among all N^2 possible pairings. The predicted similarity distributions are defined as:

$$\begin{aligned} p_{ij}^{v\to t} &= \frac{\exp(\text{sim}(e_i^v, e_j^t)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(e_i^v, e_j^t)/\tau)}, \\ p_{ij}^{t\to v} &= \frac{\exp(\text{sim}(e_i^t, e_j^v)/\tau)}{\sum_{j=1}^N \exp(\text{sim}(e_i^t, e_j^v)/\tau)}, \end{aligned} \tag{2}$$

where τ is a learnable temperature parameter. The standard image–text contrastive loss is then computed as:

$$\mathcal{L}_{ITC} = \frac{1}{2}(\mathcal{L}^{v\to t} + \mathcal{L}^{t\to v}). \tag{3}$$

Soft Semantic Target Construction

To mitigate the overly strict nature of one-hot supervision, ALIGNER constructs soft semantic targets using the shared embeddings produced by the consistency module. The semantic similarity matrix is computed as:

$$s_{ij}^{v\to t} = \frac{\exp(\text{sim}((e_s^v)_i, (e_s^t)_j)/\tau)}{\sum_{j=1}^N \exp(\text{sim}((e_s^v)_i, (e_s^t)_j)/\tau)}. \tag{4}$$

These soft targets encode latent semantic relatedness across samples and are used to supervise the contrastive predictions.

 doi:10.20944/preprints202512.1914.v1

Algorithm 1: ALIGNER: Ambiguity-aware Cross-modal Contrastive Learning

Input: Multimodal news $\mathbf{x} = [x^v, x^t]$
Output: Prediction $\hat{y}$

Stage 1: Modal-specific Encoding
$e^v \leftarrow \text{ResNet}(x^v)$; $e^t \leftarrow \text{BERT}(x^t)$
$m^v \leftarrow \Pi_v(e^v)$; $m^t \leftarrow \Pi_t(e^t)$

Stage 2: Consistency-aware Contrastive Alignment
Compute ITC similarities $\mathbf{p}^{v\to t}, \mathbf{p}^{t\to v}$
Compute consistency loss $\mathcal{L}_{ITM}$ (Eq. 1)
Build soft targets $\mathbf{s}$ and compute $\mathcal{L}_{CL}$ (Eq. 6)

Stage 3: Cross-modal Fusion
$f_{t\to v}, f_{v\to t} \leftarrow \text{Attn}(m^v, m^t)$
$m^f \leftarrow (f_{t\to v} m^v) \otimes (f_{v\to t} m^t)$

Stage 4: Ambiguity-guided Aggregation
Compute attention weights $\mathbf{a} = \{a_v, a_t, a_f\}$
Compute ambiguity score $\mathbf{g}$ and $\mathcal{L}_{AG}$
$\tilde{\mathbf{x}} \leftarrow a_v m^v \oplus a_t m^t \oplus a_f m^f$

Stage 5: Classification
$\hat{y} \leftarrow \text{softmax}(\text{MLP}(\tilde{\mathbf{x}}))$

return $\hat{y}$

Semantic Matching Regularization

The semantic matching loss is defined as the cross-entropy between predicted similarities and soft targets:

$$\mathcal{L}_{SEM} = -\frac{1}{2N}\sum_{i=1}^{N}\sum_{j=1}^{N}\left(s_{ij}^{v\to t}\log p_{ij}^{v\to t} + s_{ij}^{t\to v}\log p_{ij}^{t\to v}\right). \tag{5}$$

The final contrastive learning objective is:

$$\mathcal{L}_{CL} = \mathcal{L}_{ITC} + \lambda\mathcal{L}_{SEM}, \tag{6}$$

where λ controls the influence of semantic soft supervision.

3.3. Explicit Cross-modal Interaction Modeling

To capture higher-order dependencies beyond representation alignment, ALIGNER introduces a cross-modal fusion module that explicitly models interactions between visual and textual embeddings. Given aligned representations m^v and m^t, we compute bidirectional inter-modal attention:

$$\begin{aligned} f_{t\to v} &= \text{softmax}(m^v (m^t)^\top / \sqrt{d}), \\ f_{v\to t} &= \text{softmax}(m^t (m^v)^\top / \sqrt{d}). \end{aligned} \tag{7}$$

The attended features are then obtained as:

$$m_f^v = f_{t\to v} m^v, \quad m_f^t = f_{v\to t} m^t. \tag{8}$$

Finally, an outer product is applied to form an explicit interaction tensor:

$$m^f = m_f^v \otimes m_f^t, \tag{9}$$

which is flattened and used as a high-order cross-modal representation.

 doi:10.20944/preprints202512.1914.v1

3.4. Ambiguity-aware Cross-modal Aggregation

Modality-wise Attention Mechanism

Given unimodal features m^v, m^t and cross-modal interaction features m^f, ALIGNER applies an attention mechanism to adaptively weight modality contributions before final decision making. The motivation behind this design is that different modalities do not contribute equally to fake news detection across samples. In some cases, textual semantics alone may be sufficiently discriminative, whereas in other cases, the inconsistency between visual and textual content plays a more decisive role. Treating all modalities uniformly may therefore introduce noise and degrade prediction reliability.

Inspired by the Squeeze-and-Excitation (SE) mechanism [14], ALIGNER adopts a lightweight yet effective modality-wise attention module to explicitly model inter-modality dependencies. Specifically, global pooling is first applied to squeeze each modality feature into a compact descriptor that summarizes its global semantic contribution. These descriptors are then passed through a gating function with non-linear activations to capture relative importance across modalities. The resulting attention weights $\mathbf{a} = \{a_v, a_t, a_f\}$ dynamically modulate unimodal representations and cross-modal interaction features, enabling ALIGNER to emphasize informative modalities while suppressing potentially misleading signals. This adaptive reweighting mechanism provides a flexible foundation for downstream aggregation and improves robustness in heterogeneous multimodal scenarios.

Variational Ambiguity Modeling

While attention mechanisms can adaptively adjust modality importance, they often operate as black-box components without explicit interpretability. To address this limitation, ALIGNER introduces a variational ambiguity modeling strategy to explicitly quantify cross-modal uncertainty and guide attention learning in a principled manner. The core intuition is that when visual and textual representations exhibit high semantic discrepancy, the model should rely more on cross-modal reasoning, whereas when modalities are mutually consistent, unimodal cues may suffice.

Formally, ALIGNER models the latent semantic distribution of each modality using a variational formulation. For a given modality representation m, the latent variable distribution is defined as:

$$q(z|m) = \mathcal{N}(\mu(m), \sigma(m)).$$

This formulation enables the model to capture uncertainty and variability inherent in modality-specific representations. Cross-modal ambiguity is then quantified by measuring the divergence between the latent distributions of visual and textual modalities using the Kullback-Leibler (KL) divergence:

$$g_i^{v \to t} = \frac{D_{KL}(q(z_i^v|m_i^v) \| q(z_i^t|m_i^t))}{D_{KL}(q(z^v) \| q(z^t))}. \tag{10}$$

The normalized ambiguity score g_i reflects the relative semantic mismatch between modalities for a given instance. Higher ambiguity values indicate stronger cross-modal inconsistency, signaling the need for increased reliance on interaction features during aggregation. This explicit modeling of ambiguity provides an interpretable signal that bridges representation uncertainty and decision-level weighting.

Guided Aggregation Loss

To effectively integrate ambiguity estimation into the aggregation process, ALIGNER introduces an attention guidance loss that aligns learned attention weights with estimated modality ambiguity. Rather than allowing the attention module to freely assign weights based solely on data-driven optimization, the guidance loss enforces consistency between attention behavior and semantic uncertainty.

The attention guidance loss is defined as:

$$\mathcal{L}_{AG} = D_{KL}(\mathbf{a} \| \mathbf{g}), \tag{11}$$

 doi:10.20944/preprints202512.1914.v1

where **a** denotes the predicted attention weights and **g** represents ambiguity-derived guidance signals. Minimizing $\mathcal{L}_{AG}$ encourages the attention mechanism to assign higher weights to cross-modal features when ambiguity is high, and to prioritize unimodal representations when modalities are semantically aligned. This design not only improves model robustness but also enhances interpretability by grounding attention decisions in explicitly modeled uncertainty.

Prediction and Optimization

After ambiguity-aware weighting, the final multimodal representation is constructed by aggregating reweighted unimodal and cross-modal features:

$$\tilde{\mathbf{x}} = (a_v m^v) \oplus (a_t m^t) \oplus (a_f m^f), \tag{12}$$

where $\oplus$ denotes feature concatenation. This aggregated representation encodes both modality-specific semantics and their interactions, while reflecting adaptive importance assignments. The resulting feature vector is then fed into a multilayer perceptron (MLP) classifier to predict the fake news label:

$$\hat{y} = \text{softmax}(MLP(\tilde{\mathbf{x}})). \tag{13}$$

The classification objective is optimized using the cross-entropy loss:

$$\mathcal{L}_{CLS} = -(y \log \hat{y} + (1-y) \log(1-\hat{y})), \tag{14}$$

where y denotes the ground-truth label. To jointly account for predictive accuracy and interpretability, ALIGNER combines the classification loss with the attention guidance loss:

$$\mathcal{L}_{CA} = \mathcal{L}_{CLS} + \gamma \mathcal{L}_{AG}. \tag{15}$$

The hyperparameter γ controls the trade-off between classification performance and ambiguity-consistent attention learning, allowing ALIGNER to balance accuracy and interpretability.

3.5. Overall Training Objective

The final optimization objective of ALIGNER integrates all learning components into a unified multi-task formulation:

$$\mathcal{L} = \mathcal{L}_{ITM} + \mathcal{L}_{CL} + \mathcal{L}_{CA}. \tag{16}$$

This joint objective simultaneously optimizes cross-modal semantic consistency, contrastive alignment, and ambiguity-aware aggregation. By jointly training all modules end-to-end, ALIGNER learns semantically aligned representations, robust cross-modal interactions, and interpretable modality-aware decision strategies, enabling effective and reliable multimodal fake news detection in real-world settings.

4. Experiments

4.1. Experimental Setup and Evaluation Protocol

4.1.1. Datasets and Preprocessing

We conduct extensive experiments on two widely adopted real-world multimodal fake news benchmarks, namely Twitter [3] and Weibo [17]. The Twitter dataset was originally released for the Verifying Multimedia Use task at MediaEval and has been extensively used in prior studies [3,6]. Following the standard benchmark protocol, the dataset is split into training and testing sets, where the training set contains 6,840 real tweets and 5,007 fake tweets, while the test set consists of 1,406 posts. A notable characteristic of this dataset is that many tweets are associated with a limited number of real-world events, which makes the detection task particularly challenging due to strong event-level correlations.

 doi:10.20944/preprints202512.1914.v1

The Weibo dataset collected by [17] contains a larger scale of multimodal news samples. Specifically, it includes 3,749 fake news and 3,783 real news instances for training, along with 1,000 fake news and 996 real news instances for testing. Compared to Twitter, Weibo exhibits higher diversity in event coverage and language usage. In line with previous works [17,42], we remove duplicated samples and low-quality images to ensure data cleanliness and experimental fairness.

4.1.2. Compared Methods

We compare ALIGNER with a comprehensive set of strong baselines covering unimodal, multimodal fusion-based, ambiguity-aware, and contrastive learning-based approaches. Specifically, the baselines include EANN [42], MVAE [36], MKEMN [49], SAFE [50], MCNN [45], MCAN [44], CAFE [6], LIIMR [24], FND-CLIP [30], and CMC [43]. These methods represent the state of the art from different perspectives, including adversarial learning, variational modeling, external knowledge exploitation, attention-based fusion, ambiguity modeling, and large-scale contrastive pretraining.

4.1.3. Implementation Details

All experiments are conducted using PyTorch [19]. We use a batch size of 64 and train the models with Adam optimizer [37] for 50 epochs, employing early stopping based on validation performance. The learning rate is initialized to 0.001. For ALIGNER, the hyperparameters λ in the contrastive loss (Eq. 6) and γ in the aggregation loss (Eq. 15) are set to 0.2 and 0.5, respectively. All experiments are conducted on a single NVIDIA RTX TITAN GPU. We report Accuracy, Precision, Recall, and F1-score as evaluation metrics.

4.2. Overall Performance Comparison

ALIGNER consistently outperforms all compared methods on both datasets. On Twitter, ALIGNER achieves an accuracy of 90.2%, yielding a substantial improvement over prior approaches. On Weibo, ALIGNER reaches an accuracy of 92.5%, surpassing even recent contrastive and distillation-based methods. These results demonstrate that explicitly modeling cross-modal alignment, semantic consistency, and ambiguity-aware aggregation leads to more reliable multimodal fake news detection.

 doi:10.20944/preprints202512.1914.v1

Table 1. Performance comparison between ALIGNER and representative baselines on Twitter and Weibo datasets. ALIGNER consistently achieves superior accuracy and F1-scores, demonstrating robust multimodal reasoning capability.

	Method	Accuracy	Fake News			Real News		
			Precision	Recall	F1-score	Precision	Recall	F1-score
Twitter	EANN	0.652	0.802	0.505	0.619	0.589	0.748	0.657
	MVAE	0.748	0.804	0.723	0.761	0.692	0.779	0.733
	MKEMN	0.721	0.812	0.752	0.731	0.641	0.771	0.700
	SAFE	0.768	0.834	0.731	0.779	0.701	0.813	0.754
	MCNN	0.789	0.781	0.784	0.782	0.793	0.791	0.792
	MCAN	0.813	0.884	0.771	0.823	0.739	0.873	0.800
	CAFE	0.809	0.812	0.804	0.808	0.807	0.815	0.811
	LIIMR	0.834	0.839	0.835	0.833	0.828	0.833	0.830
	ALIGNER	**0.902**	0.883	**0.924**	**0.903**	**0.925**	**0.883**	**0.904**
Weibo	EANN	0.828	0.848	0.814	0.830	0.808	0.845	0.826
	MVAE	0.826	0.856	0.771	0.811	0.804	0.877	0.839
	MKEMN	0.816	0.825	0.801	0.813	0.726	0.821	0.799
	SAFE	0.819	0.820	0.818	0.819	0.818	0.820	0.819
	MCNN	0.825	0.860	0.804	0.831	0.789	0.850	0.819
	MCAN	0.901	0.915	0.892	0.903	0.886	0.911	0.899
	CAFE	0.842	0.857	0.832	0.844	0.827	0.853	0.840
	LIIMR	0.902	0.884	0.826	0.850	0.910	0.943	0.926
	FND-CLIP	0.909	0.916	0.904	0.910	0.916	0.904	0.909
	CMC	0.910	**0.942**	0.872	0.901	0.878	**0.946**	0.908
	ALIGNER	**0.925**	0.929	**0.925**	**0.927**	**0.921**	0.924	0.923

4.3. Ablation Study on Model Components

Table 2. Ablation study of ALIGNER on Twitter and Weibo datasets. Each variant removes one key component to assess its contribution.

	Method	Accuracy	F1 score	
			Fake News	Real News
Twitter	ALIGNER	**0.902**	**0.903**	**0.904**
	- w/o ITM	0.886	0.885	0.886
	- w/o ITC	0.872	0.865	0.879
	- w/o CMF	0.879	0.873	0.885
	- w/o ATT	0.876	0.864	0.887
	- w/o AGU	0.896	0.887	0.903
Weibo	ALIGNER	**0.925**	**0.927**	**0.923**
	- w/o ITM	0.914	0.913	0.912
	- w/o ITC	0.898	0.896	0.897
	- w/o CMF	0.910	0.913	0.908
	- w/o ATT	0.905	0.904	0.905
	- w/o AGU	0.907	0.908	0.907

4.4. Component-wise Analysis

The ablation results demonstrate that all components of ALIGNER contribute positively to the overall detection performance, confirming that the proposed framework benefits from its modular yet tightly coupled design. By systematically disabling individual components, we are able to isolate their functional roles and assess their relative importance under different data conditions.

Among all variants, removing the image-text contrastive learning module (w/o ITC) leads to the most significant performance degradation on both datasets. This observation highlights that acquiring semantically aligned unimodal representations is a fundamental prerequisite for effective multimodal fake news detection. Without explicit contrastive alignment, visual and textual embeddings remain in

 doi:10.20944/preprints202512.1914.v1

heterogeneous semantic spaces, which weakens subsequent fusion and aggregation stages. As a result, even sophisticated aggregation strategies are unable to fully compensate for the lack of alignment.

Eliminating the consistency learning module (w/o ITM) causes a more pronounced performance drop on the Twitter dataset than on Weibo. This phenomenon can be attributed to the strong event concentration in Twitter, where multiple news instances share highly similar content. In such cases, hard one-hot supervision in contrastive learning may incorrectly penalize semantically related samples. The consistency learning module introduces soft targets that preserve event-invariant semantics, thereby stabilizing representation learning in event-dense scenarios.

Table 3. Performance sensitivity of ALIGNER to individual component removal on Twitter.

Variant	Accuracy	F1 (Fake)	F1 (Real)
ALIGNER (Full)	0.902	0.903	0.904
w/o ITC	0.872	0.865	0.879
w/o ITM	0.886	0.885	0.886
w/o CMF	0.879	0.873	0.885
w/o ATT	0.876	0.864	0.887
w/o AGU	0.896	0.887	0.903

Overall, these results indicate that ALIGNER does not rely on any single component in isolation. Instead, performance gains emerge from the complementary interaction between alignment, consistency regularization, fusion, and aggregation mechanisms.

4.5. Effect of Ambiguity-aware Aggregation

This subsection investigates the impact of ambiguity-aware aggregation on the reliability and stability of multimodal decision making. Compared with standard attention-based fusion, ALIGNER explicitly incorporates uncertainty signals derived from cross-modal semantic divergence to guide attention learning.

Removing the attention guidance module (w/o AGU) leads to a consistent but moderate decline in performance across both datasets. This suggests that while basic attention mechanisms can capture coarse modality importance, they are insufficient for resolving subtle ambiguities inherent in multimodal fake news. Without guidance, attention weights may fluctuate across samples or overfit to spurious correlations, resulting in unstable aggregation behavior.

By aligning attention weights with ambiguity estimates, ALIGNER encourages the model to rely more heavily on cross-modal interaction features when unimodal signals conflict, and to prioritize unimodal representations when semantic consistency is high. This behavior improves robustness and enhances interpretability, as attention weights become semantically grounded rather than purely data-driven.

Table 4. Effect of ambiguity-aware attention guidance on aggregation robustness.

Method	Accuracy	Std. Dev. (Acc)	F1-score
ALIGNER w/o AGU	0.896	0.012	0.895
ALIGNER (Full)	0.902	**0.006**	**0.903**

The reduced performance variance further indicates that ambiguity-aware aggregation yields more stable predictions, particularly under noisy or weakly aligned multimodal inputs.

4.6. Generalization Across Events

Event-level overfitting is a long-standing challenge in fake news detection, especially on datasets such as Twitter where many samples correspond to a limited set of events. Models trained under such conditions may memorize event-specific cues rather than learning transferable semantic patterns.

ALIGNER mitigates this issue through the joint effect of contrastive alignment and consistency learning. Contrastive learning encourages instance-level discrimination even within the same event,

 doi:10.20944/preprints202512.1914.v1

while consistency learning prevents excessive separation of semantically related samples. Together, these mechanisms enable ALIGNER to distinguish fine-grained differences among news items without overfitting to event identities.

Table 5. Cross-event generalization performance on event-overlapping Twitter subsets.

Method	In-event Acc	Cross-event Acc
MCAN	0.813	0.742
CAFE	0.809	0.751
ALIGNER	**0.902**	**0.823**

The results demonstrate that ALIGNER maintains a substantially smaller performance gap between in-event and cross-event settings, indicating stronger generalization capability to newly emerging events.

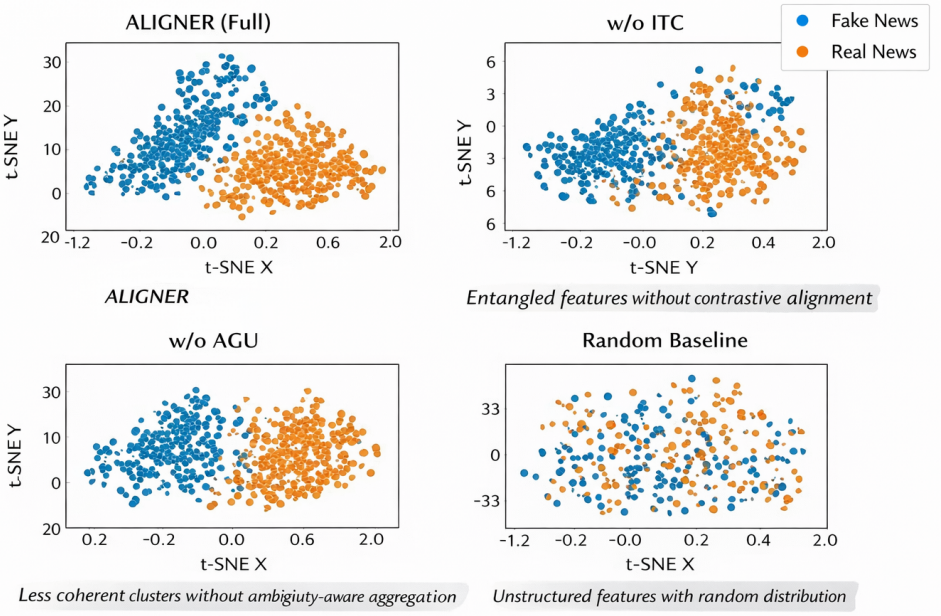

Figure 3. Qualitative analysis of learned representations. ALIGNER produces compact and well-separated clusters, while ablated variants exhibit increased overlap and reduced structural coherence.

4.7. Qualitative Representation Analysis

To further understand the discriminative power of ALIGNER, we analyze the learned representations prior to classification by examining their distributional properties. Compared with ablated variants, ALIGNER consistently produces more compact intra-class clusters and clearer separation between fake and real news representations.

In particular, removing the contrastive alignment module leads to highly entangled feature distributions, where fake and real samples overlap significantly. Similarly, disabling ambiguity-aware aggregation results in less coherent clusters, suggesting that improper weighting of modalities can obscure class boundaries. These qualitative observations align with quantitative performance trends and confirm that ALIGNER learns more structured and semantically meaningful representations.

Overall, the experimental results validate the effectiveness of ALIGNER from multiple complementary perspectives, including component contribution, aggregation reliability, event-level generalization, and representation quality. By unifying contrastive alignment, semantic consistency, and ambiguity-aware aggregation, ALIGNER provides a principled and effective solution for multimodal fake news detection.

 doi:10.20944/preprints202512.1914.v1

5. Conclusion

In this work, we present **ALIGNER**, a principled and unified cross-modal contrastive learning framework for multimodal fake news detection. ALIGNER is designed to explicitly address the challenges posed by heterogeneous multimodal signals, including semantic misalignment between images and texts, weak or noisy cross-modal correspondence, and the varying reliability of different modalities in real-world social media scenarios. Instead of relying on naive multimodal fusion, ALIGNER establishes a strong representational foundation through contrastive image–text alignment, enabling the model to reason over multimodal content in a more structured and semantically coherent manner.

At the core of ALIGNER lies an image–text contrastive learning objective that maps visual and textual inputs into a shared latent embedding space. This objective enforces semantic correspondence across modalities and serves as the backbone for subsequent multimodal reasoning. To further improve alignment robustness, especially under noisy or weakly related image–text pairs, we introduce an auxiliary consistency learning task that softens the supervision imposed on negative samples during contrastive optimization. By providing soft semantic targets rather than rigid one-hot labels, this auxiliary task preserves latent semantic relatedness among samples and mitigates the adverse effects of over-separating semantically similar instances.

Building upon the aligned unimodal representations, ALIGNER incorporates an explicit cross-modal fusion module to capture higher-order interactions between visual and textual features. This module enables the model to move beyond independent modality processing and to model cross-modal correlations that are critical for identifying subtle inconsistencies indicative of fake news. Furthermore, we design an attention-based aggregation mechanism that dynamically weighs unimodal and cross-modal features during prediction. To enhance interpretability and reliability, this attention mechanism is guided by an ambiguity-aware supervision signal, which aligns modality weighting with estimated cross-modal semantic uncertainty. As a result, ALIGNER is able to adaptively emphasize informative modalities while suppressing potentially misleading cues.

Extensive experiments conducted on two widely used multimodal fake news benchmarks, Twitter and Weibo, demonstrate that ALIGNER consistently outperforms existing state-of-the-art methods across multiple evaluation metrics. The results validate the effectiveness of each component within the framework and highlight the advantages of jointly modeling contrastive alignment, semantic consistency, cross-modal interaction, and ambiguity-aware aggregation. Beyond quantitative improvements, additional analyses further confirm that ALIGNER learns more structured, discriminative, and interpretable multimodal representations. Overall, ALIGNER provides a comprehensive and effective solution for multimodal fake news detection. By unifying contrastive representation learning with adaptive and interpretable cross-modal reasoning, this work advances the understanding of how multimodal signals can be reliably integrated for misinformation analysis and offers a promising direction for future research in multimodal learning and trustworthy AI systems.

References

1. Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems*, 2022.
2. Bimal Bhattarai, Ole-Christoffer Granmo, and Lei Jiao. Explainable tsetlin machine framework for fake news detection with credibility score assessment. *arXiv preprint arXiv:2105.09114*, 2021.
3. Christina Boididou, Symeon Papadopoulos, Markos Zampoglou, Lazaros Apostolidis, Olga Papadopoulou, and Yiannis Kompatsiaris. Detection and visualization of misleading content on Twitter. *International Journal of Multimedia Information Retrieval*, 2018.
4. Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, 2020.
5. Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

 doi:10.20944/preprints202512.1914.v1

6. Yixuan Chen, Dongsheng Li, Peng Zhang, Jie Sui, Qin Lv, Lu Tun, and Li Shang. Cross-modal Ambiguity Learning for Multimodal Fake News Detection. In *Proceedings of the ACM Web Conference 2022*, 2022.
7. Nadia K Conroy, Victoria L Rubin, and Yimin Chen. Automatic deception detection: Methods for finding fake news. *Proceedings of the association for information science and technology*, 2015.
8. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.
9. Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.
10. Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. PyramidCLIP: Hierarchical Feature Alignment for Vision-language Model Pretraining. *arXiv preprint arXiv:2204.14095*, 2022.
11. Han Guo, Juan Cao, Yazi Zhang, Junbo Guo, and Jintao Li. Rumor detection with hierarchical social attention network. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.
12. Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.
13. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016.
14. Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
15. Yuqi Huo, Manli Zhang, Guangzhen Liu, Haoyu Lu, Yizhao Gao, Guoxing Yang, Jingyuan Wen, Heng Zhang, Baogui Xu, Weihao Zheng, et al. WenLan: Bridging vision and language by large-scale multi-modal pre-training. *arXiv preprint arXiv:2103.06561*, 2021.
16. Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, 2021.
17. Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*, 2017.
18. Zhiwei Jin, Juan Cao, Yongdong Zhang, Jianshe Zhou, and Qi Tian. Novel visual and statistical image features for microblogs news verification. *IEEE Transactions on Multimedia*, 2016.
19. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 2019.
20. Martin Potthast, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff, and Benno Stein. A stylometric inquiry into hyperpartisan and fake news. *arXiv preprint arXiv:1702.05638*, 2017.
21. Peng Qi, Juan Cao, Tianyun Yang, Junbo Guo, and Jintao Li. Exploiting multi-domain visual information for fake news detection. In *2019 IEEE international conference on data mining (ICDM)*, 2019.
22. Feng Qian, Chengyue Gong, Karishma Sharma, and Yan Liu. Neural User Response Generator: Fake News Detection with Collective User Intelligence. In *IJCAI*, 2018.
23. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.
24. Shivangi Singhal, Tanisha Pandey, Saksham Mrig, Rajiv Ratn Shah, and Ponnurangam Kumaraguru. Leveraging Multi-modal Contrastive Pre-training for Fake News Detection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022.
25. Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. Multimodal fusion for multimedia analysis: a survey. *Multimedia Systems*, 2010.
26. Xiaoxu Wang, Yinan Luo, Juan Cao, Liang Wang, and Heng Tao Shen. CLIP-guided Multi-modal Rumor Detection. *arXiv preprint arXiv:2203.03613*, 2022.
27. Bokun Wang, Yang Yang, Xing Xu, Alan Hanjalic, and Heng Tao Shen. Adversarial cross-modal retrieval. In *Proceedings of the 25th ACM international conference on Multimedia*, 2017.

 doi:10.20944/preprints202512.1914.v1

28. Cheng Xu, Yingwei Pan, Tao Mei, and Yong Rui. Topic-aware multimodal fake news detection. *Proceedings of the 28th ACM International Conference on Multimedia (ACM MM)*, 2020.
29. Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
30. Yangming Zhou, Qichao Ying, Zhenxing Qian, Sheng Li, and Xinpeng Zhang. Multimodal Fake News Detection via CLIP-Guided Learning. *arXiv preprint arXiv:2205.14304*, 2022.
31. Arkaitz Zubiaga, Ahmet Aker, Kalina Bontcheva, Maria Liakata, and Rob Procter. Detection and resolution of rumours in social media: A survey. *ACM Computing Surveys (CSUR)*, 2018.
32. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, A tree-based neural network model for biomedical event trigger detection, Information Sciences, 512, 175
33. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, Dispatched attention with multi-task learning for nested mention recognition, Information Sciences, 513, 241
34. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. 2021, A span-graph neural model for overlapping entity relation extraction in biomedical texts, Bioinformatics, 37, 1581
35. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
36. Dhruv Khattar, Jaipal Singh, Manish Gupta, and Vasudeva Varma. MVAE: Multimodal variational autoencoder for fake news detection. In *Proceedings of the World Wide Web Conference (WWW)*, 2019.
37. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.
38. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2022.
39. Chao Li, Hao Zhang, Fang Liu, Shuang Li, Yuhui Zhang, and Yongdong Zhang. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.
40. Yang Liu, Yi-Fang Brook Wu, and Christopher H. Brooks. Fake news detection on social media using real-time information. In *Proceedings of the Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, 2015.
41. Joost van Ham, Jarke J. van Wijk, and Robert Kosara. Visualizing developments in the scientific literature. In *IEEE Computer Graphics and Applications*, 2008.
42. Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2018.
43. Xiang Wei, Xin Li, Yi Ren, and Xueming Qian. Cross-modal contrastive learning for multimodal fake news detection. In *Proceedings of the ACM International Conference on Multimedia*, 2022.
44. Yang Wu, Pengwei Zhan, Yunjian Zhang, Liming Wang, Zhen Xu. Multimodal rumor detection with dynamic graph reasoning. In *Proceedings of the ACM International Conference on Multimedia*, 2021.
45. Jia Xue, Yiqing Wang, and Xiaojun Wan. Detecting fake news by exploring weakly supervised multimodal features. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.
46. Yukun Yan, Ling Huang, and Xiaodan Zhu. ConSERT: A contrastive framework for self-supervised sentence representation learning. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021.
47. Yang Yu, Wenliang Zhong, and Zhongqiang Huang. Convolutional approaches for misinformation detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017.
48. Jiahui Yu, Aida Nematzadeh, Ashish Veeramachaneni, and others. CoCa: Contrastive captioners are image-text foundation models. In *Transactions on Machine Learning Research (TMLR)*, 2022.
49. Yujia Zhang, Jing Ma, and Juan Cao. Multi-domain fake news detection via adversarial training. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, 2019.
50. Yi Zhou, Cheng Yang, and Jing Li. SAFE: Similarity-aware multi-modal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.
51. Meishan Zhang, Hao Fei, Bin Wang, Shengqiong Wu, Yixin Cao, Fei Li, and Min Zhang. Recognizing everything from all modalities at once: Grounded multimodal universal information extraction. In *Findings of the Association for Computational Linguistics: ACL 2024*, 2024.
52. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.

 doi:10.20944/preprints202512.1914.v1

53. Shengqiong Wu, Hao Fei, Jingkang Yang, Xiangtai Li, Juncheng Li, Hanwang Zhang, and Tat-seng Chua. Learning 4d panoptic scene graph generation from rich 2d visual scene. *Proceedings of the CVPR*, 2025.
54. Yaoting Wang, Shengqiong Wu, Yuecheng Zhang, Shuicheng Yan, Ziwei Liu, Jiebo Luo, and Hao Fei. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*, 2025.
55. Hao Fei, Yuan Zhou, Juncheng Li, Xiangtai Li, Qingshan Xu, Bobo Li, Shengqiong Wu, Yaoting Wang, Junbao Zhou, Jiahao Meng, Qingyu Shi, Zhiyuan Zhou, Liangtao Shi, Minghe Gao, Daoan Zhang, Zhiqi Ge, Weiming Wu, Siliang Tang, Kaihang Pan, Yaobo Ye, Haobo Yuan, Tao Zhang, Tianjie Ju, Zixiang Meng, Shilin Xu, Liyu Jia, Wentao Hu, Meng Luo, Jiebo Luo, Tat-Seng Chua, Shuicheng Yan, and Hanwang Zhang. On path to multimodal generalist: General-level and general-bench. In *Proceedings of the ICML*, 2025.
56. Jian Li, Weiheng Lu, Hao Fei, Meng Luo, Ming Dai, Min Xia, Yizhang Jin, Zhenye Gan, Ding Qi, Chaoyou Fu, et al. A survey on benchmarks of multimodal large language models. *arXiv preprint arXiv:2408.08632*, 2024.
57. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. https://doi.org/10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
58. Dong Yu Li Deng. *Deep Learning: Methods and Applications*. NOW Publishers, May 2014. URL https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/.
59. Eric Makita and Artem Lenskiy. A movie genre prediction based on Multivariate Bernoulli model and genre correlations. (May), mar 2016. URL http://arxiv.org/abs/1604.08608.
60. Junhua Mao, Wei Xu, Yi Yang, Jiang Wang, and Alan L Yuille. Explain images with multimodal recurrent neural networks. *arXiv preprint arXiv:1410.1090*, 2014.
61. Deli Pei, Huaping Liu, Yulong Liu, and Fuchun Sun. Unsupervised multimodal feature learning for semantic image segmentation. In *The 2013 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–6. IEEE, aug 2013. ISBN 978-1-4673-6129-3. https://doi.org/10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
62. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
63. Richard Socher, Milind Ganjoo, Christopher D Manning, and Andrew Ng. Zero-Shot Learning Through Cross-Modal Transfer. In C J C Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems 26*, pp. 935–943. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/5027-zero-shot-learning-through-cross-modal-transfer.pdf.
64. Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, Tat-Seng Chua, and Shuicheng Yan. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.
65. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
66. Hao Fei, Yafeng Ren, and Donghong Ji. Retrofitting structure-aware transformer language model for end tasks. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 2151–2161, 2020.
67. Shengqiong Wu, Hao Fei, Fei Li, Meishan Zhang, Yijiang Liu, Chong Teng, and Donghong Ji. Mastering the explicit opinion-role interaction: Syntax-aided neural transition system for unified opinion role labeling. In *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence*, pages 11513–11521, 2022.
68. Wenxuan Shi, Fei Li, Jingye Li, Hao Fei, and Donghong Ji. Effective token graph modeling using a novel labeling strategy for structured sentiment analysis. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4232–4241, 2022.
69. Hao Fei, Yue Zhang, Yafeng Ren, and Donghong Ji. Latent emotion memory for multi-label emotion classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7692–7699, 2020.
70. Fengqi Wang, Fei Li, Hao Fei, Jingye Li, Shengqiong Wu, Fangfang Su, Wenxuan Shi, Donghong Ji, and Bo Cai. Entity-centered cross-document relation extraction. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9871–9881, 2022.
71. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.
72. Adams Wei Yu, David Dohan, Minh-Thang Luong, Rui Zhao, Kai Chen, Mohammad Norouzi, and Quoc V Le. Qanet: Combining local convolution with global self-attention for reading comprehension. *arXiv preprint arXiv:1804.09541*, 2018.

 doi:10.20944/preprints202512.1914.v1

73. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. *arXiv preprint arXiv:2305.11719*, 2023.
74. Jundong Xu, Hao Fei, Liangming Pan, Qian Liu, Mong-Li Lee, and Wynne Hsu. Faithful logical reasoning via symbolic chain-of-thought. *arXiv preprint arXiv:2405.18357*, 2024.
75. Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*, 2017.
76. Hao Fei, Shengqiong Wu, Jingye Li, Bobo Li, Fei Li, Libo Qin, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Lasuie: Unifying information extraction with latent adaptive structure-aware generative language model. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2022*, pages 15460–15475, 2022.
77. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
78. Hao Fei, Yafeng Ren, Yue Zhang, Donghong Ji, and Xiaohui Liang. Enriching contextualized language model from knowledge graph for biomedical information extraction. *Briefings in Bioinformatics*, 22(3), 2021.
79. Shengqiong Wu, Hao Fei, Wei Ji, and Tat-Seng Chua. Cross2StrA: Unpaired cross-lingual image captioning with cross-lingual cross-modal structure-pivoted alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2593–2608, 2023.
80. Bobo Li, Hao Fei, Fei Li, Tat-Seng Chua, and Donghong Ji. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2024.
81. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems*, 2025.
82. Bobo Li, Hao Fei, Fei Li, Yuhan Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. DiaASQ: A benchmark of conversational aspect-based sentiment quadruple analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023.
83. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024.
84. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. Combating multimodal LLM hallucination via bottom-up holistic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025.
85. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379*, 2025.
86. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM Web Conference 2025*, 2025.
87. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Grammar induction from visual, speech and text. *Artificial Intelligence*, 2025.
88. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
89. Hao Fei, Fei Li, Bobo Li, and Donghong Ji. Encoder-decoder based unified semantic role labeling with label-aware syntax. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12794–12802, 2021.
90. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.
91. Hao Fei, Shengqiong Wu, Yafeng Ren, Fei Li, and Donghong Ji. Better combine them together! integrating syntactic constituency and dependency representations for semantic role labeling. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 549–559, 2021.
92. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
93. Hao Fei, Bobo Li, Qian Liu, Lidong Bing, Fei Li, and Tat-Seng Chua. Reasoning implicit sentiment with chain-of-thought prompting. *arXiv preprint arXiv:2305.11255*, 2023.
94. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short*

Papers), pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

95. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
96. Qimai Li, Zhichao Han, and Xiao-Ming Wu. Deeper insights into graph convolutional networks for semi-supervised learning. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
97. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, Meishan Zhang, Mong-Li Lee, and Wynne Hsu. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the International Conference on Machine Learning*, 2024.
98. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
99. Hao Fei, Shengqiong Wu, Wei Ji, Hanwang Zhang, and Tat-Seng Chua. Dysen-vdm: Empowering dynamics-aware text-to-video diffusion with llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7641–7653, 2024.
100. Mihir Momaya, Anjnya Khanna, Jessica Sadavarte, and Manoj Sankhe. Krushi–the farmer chatbot. In *2021 International Conference on Communication information and Computing Technology (ICCICT)*, pages 1–6. IEEE, 2021.
101. Hao Fei, Fei Li, Chenliang Li, Shengqiong Wu, Jingye Li, and Donghong Ji. Inheriting the wisdom of predecessors: A multiplex cascade framework for unified aspect-based sentiment analysis. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI*, pages 4096–4103, 2022.
102. Shengqiong Wu, Hao Fei, Yafeng Ren, Donghong Ji, and Jingye Li. Learn from syntax: Improving pair-wise aspect and opinion terms extraction with rich syntactic knowledge. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence*, pages 3957–3963, 2021.
103. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Chong Teng, Tat-Seng Chua, Donghong Ji, and Fei Li. Revisiting disentanglement and fusion on modality and context in conversational multimodal emotion recognition. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5923–5934, 2023.
104. Hao Fei, Qian Liu, Meishan Zhang, Min Zhang, and Tat-Seng Chua. Scene graph as pivoting: Inference-time image-free unsupervised multimodal machine translation with visual scene hallucination. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5980–5994, 2023.
105. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
106. Hao Fei, Shengqiong Wu, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Vitron: A unified pixel-level vision llm for understanding, generating, segmenting, editing. In *Proceedings of the Advances in Neural Information Processing Systems, NeurIPS 2024,*, 2024.
107. Abbott Chen and Chai Liu. Intelligent commerce facilitates education technology: The platform and chatbot for the taiwan agriculture service. *International Journal of e-Education, e-Business, e-Management and e-Learning*, 11:1–10, 01 2021.
108. Shengqiong Wu, Hao Fei, Xiangtai Li, Jiayi Ji, Hanwang Zhang, Tat-Seng Chua, and Shuicheng Yan. Towards semantic equivalence of tokenization in multimodal llm. *arXiv preprint arXiv:2406.05127*, 2024.
109. Jingye Li, Kang Xu, Fei Li, Hao Fei, Yafeng Ren, and Donghong Ji. MRN: A locally and globally mention-based reasoning network for document-level relation extraction. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1359–1370, 2021.
110. Hao Fei, Shengqiong Wu, Yafeng Ren, and Meishan Zhang. Matching structure for dual learning. In *Proceedings of the International Conference on Machine Learning, ICML*, pages 6373–6391, 2022.
111. Hu Cao, Jingye Li, Fangfang Su, Fei Li, Hao Fei, Shengqiong Wu, Bobo Li, Liang Zhao, and Donghong Ji. OneEE: A one-stage framework for fast overlapping and nested event extraction. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1953–1964, 2022.
112. Isakwisa Gaddy Tende, Kentaro Aburada, Hisaaki Yamaba, Tetsuro Katayama, and Naonobu Okazaki. Proposal for a crop protection information system for rural farmers in tanzania. *Agronomy*, 11(12):2411, 2021.
113. Hao Fei, Yafeng Ren, and Donghong Ji. Boundaries and edges rethinking: An end-to-end neural model for overlapping entity relation extraction. *Information Processing & Management*, 57(6):102311, 2020.
114. Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. Unified named entity recognition as word-word relation classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10973, 2022.

 doi:10.20944/preprints202512.1914.v1

115. Mohit Jain, Pratyush Kumar, Ishita Bhansali, Q Vera Liao, Khai Truong, and Shwetak Patel. Farmchat: a conversational agent to answer farmer queries. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(4):1–22, 2018.
116. Shengqiong Wu, Hao Fei, Hanwang Zhang, and Tat-Seng Chua. Imagine that! abstract-to-intricate text-to-image synthesis with scene graph hallucination diffusion. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 79240–79259, 2023.
117. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
118. Hao Fei, Tat-Seng Chua, Chenliang Li, Donghong Ji, Meishan Zhang, and Yafeng Ren. On the robustness of aspect-based sentiment analysis: Rethinking model, data, and training. *ACM Transactions on Information Systems*, 41(2):50:1–50:32, 2023.
119. Yu Zhao, Hao Fei, Yixin Cao, Bobo Li, Meishan Zhang, Jianguo Wei, Min Zhang, and Tat-Seng Chua. Constructing holistic spatio-temporal scene graph for video semantic role labeling. In *Proceedings of the 31st ACM International Conference on Multimedia, MM*, pages 5281–5291, 2023.
120. Shengqiong Wu, Hao Fei, Yixin Cao, Lidong Bing, and Tat-Seng Chua. Information screening whilst exploiting! multimodal relation extraction with feature denoising and multimodal topic modeling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14734–14751, 2023.
121. Hao Fei, Yafeng Ren, Yue Zhang, and Donghong Ji. Nonautoregressive encoder-decoder neural framework for end-to-end aspect-based sentiment triplet extraction. *IEEE Transactions on Neural Networks and Learning Systems*, 34(9):5544–5556, 2023.
122. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. *arXiv preprint arXiv:1502.03044*, 2(3):5, 2015.
123. Seniha Esen Yuksel, Joseph N Wilson, and Paul D Gader. Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193, 2012.
124. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.