

Article

Not peer-reviewed version

Contextual Synergy through Explicit and Implicit Relations: A Unified Perspective for Image Description Generation

Lotte Vermeulen , Yara Van den Broeck , [Callum Hensley](#) , Bram Smet *

Posted Date: 8 October 2025

doi: 10.20944/preprints202510.0575.v1

Keywords: Dynamic Mixture Attention; explicit relations; implicit relations; image captioning; contextual reasoning



Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

Contextual Synergy Through Explicit and Implicit Relations: A Unified Perspective for Image Description Generation

Lotte Vermeulen, Yara Van den Broeck, Callum Hensley and Bram Smet *

University of Antwerp
* Correspondence: bdesmet@uantwerpen.be

Abstract

Automatic image description generation, commonly referred to as image captioning, has long been recognized as a highly demanding challenge within artificial intelligence, due to the necessity of bridging the perceptual gap between visual understanding and natural language expression. Conventional encoder-decoder pipelines typically convert salient image regions into textual sentences, yielding reasonable performance across diverse datasets. Nevertheless, such models are still constrained by their limited capacity to capture the nuanced contextual interactions that naturally exist among objects in complex scenes. These contextual cues are often conveyed through visual relationships—some explicitly manifested through spatial or semantic connections, and others implicitly embedded in higher-order global associations. In this study, we propose a novel framework, termed **VisRelNet**, which explicitly and implicitly explores object relationships to enrich regional semantics for image captioning. On the explicit side, we construct semantic graphs among detected objects and introduce a Gated Graph Convolutional Network (Gated GCN) that dynamically filters relational edges to emphasize informative connections. On the implicit side, we employ a region-level bidirectional transformer encoder (Region BERT), which directly models latent dependencies across all regions without relying on external annotations for relationships. Furthermore, to harmonize the complementary strengths of explicit and implicit cues, we design a Dynamic Mixture Attention (DMA) mechanism, capable of adaptively balancing region-level features through channel-wise gating. We evaluate **VisRelNet** on the Microsoft COCO benchmark and observe consistent and significant improvements over a range of competitive baselines. Experimental evidence demonstrates that leveraging both explicit and implicit relational reasoning enhances the contextual richness of image representations, thereby producing captions that are more coherent, descriptive, and human-like. This work highlights the importance of multi-faceted relational modeling and provides a pathway toward unified relational reasoning for vision-language tasks.

Keywords: Dynamic Mixture Attention; explicit relations; implicit relations; image captioning; contextual reasoning

1. Introduction

The task of generating natural language descriptions for visual scenes—image captioning—stands at the crossroads of computer vision and natural language processing. By attempting to translate raw pixels into structured sentences, image captioning not only offers a challenging research problem but also promises transformative applications in domains such as assistive technologies for the visually impaired, early childhood education, cross-modal retrieval systems [1], and natural human-robot communication frameworks [2]. Its dual requirement of perception and linguistic reasoning has made it one of the most representative tasks for evaluating multimodal artificial intelligence.

Traditional encoder-decoder approaches [3,4] have relied on convolutional neural networks (CNNs) to encode global or region-based visual features, followed by recurrent neural networks (RNNs)

to sequentially decode captions. Inspired by machine translation paradigms, attention mechanisms [5] were introduced, allowing models to selectively focus on specific regions while generating each word. Reinforcement learning strategies, such as self-critical sequence training [6], further optimized caption generation by aligning model objectives with evaluation metrics. The introduction of attention over object-level regions, rather than uniform grids, pushed the field forward substantially [7–9]. Yet, despite these developments, the question of how to holistically model associations between objects remains insufficiently addressed.

Research in visual cognition underscores that contextual cues are pivotal in object understanding [10]. Objects rarely exist in isolation: a "cup" on a "table" is semantically different from a "cup" in "someone's hand." Explicit modeling of such relations offers a richer semantic substrate for captions. To this end, scene graph-based approaches [8,11] attempt to encode explicit pairwise relations, whereas transformer-based methods [9,12,13] capture implicit dependencies through self-attention. Nonetheless, both lines of research have inherent limitations: scene-graph methods suffer from inaccuracies in relation prediction due to reliance on external detectors, while transformer-only models lack explicit relational guidance, leading to less interpretable or diffuse contextual associations.

In this work, we propose to unify both explicit and implicit relational reasoning into a coherent framework. Specifically, we design a semantic graph construction module that extracts object-level relations, filtered by spatial constraints, and passes them through a Gated GCN for controlled aggregation. To complement this, we introduce a Region BERT encoder that employs bidirectional self-attention across all detected regions, thereby capturing global dependencies without the need for pre-labeled relation annotations. To reconcile the outputs from these two perspectives, we design a Dynamic Mixture Attention (DMA) module. Unlike simple concatenation or fusion, DMA adaptively assigns weights to explicit and implicit features in a channel-wise manner, guided by the decoder's hidden state, resulting in representations that better align with linguistic generation.

Our contributions are threefold:

- We propose **VisRelNet**, a relational reasoning framework for image captioning, which integrates explicit graph-based relational modeling and implicit transformer-based contextual modeling in a complementary fashion.
- We introduce a Gated GCN for explicit relation reasoning and a Region BERT encoder for implicit global context learning, together providing enriched semantic features for each object region.
- We develop a Dynamic Mixture Attention (DMA) module that selectively balances explicit and implicit relational features during decoding, enabling the model to adaptively focus on the most relevant context when generating captions.

Beyond the technical contributions, this study reveals broader implications. First, relational reasoning is not limited to captioning but may be extended to other multimodal tasks, such as visual question answering, visual grounding, or multimodal summarization. Second, explicit-implicit relational fusion aligns well with cognitive theories of human vision, where individuals rely on both concrete object relationships and latent contextual knowledge. Lastly, the proposed **VisRelNet** offers a template for integrating structured reasoning with deep neural architectures, pushing multimodal research closer to human-like understanding.

By conducting extensive evaluations on the Microsoft COCO dataset, we demonstrate that **VisRelNet** not only achieves state-of-the-art performance but also produces captions with higher contextual fidelity. The enriched modeling of relationships enables the system to go beyond surface-level recognition, describing nuanced associations and delivering outputs that are more informative, context-aware, and interpretable.

2. Related Work

2.1. Advances in Image Captioning Paradigms

The problem of automatically producing natural language descriptions for images has been extensively studied and has become a benchmark task at the intersection of computer vision and

natural language processing. Early approaches were largely template-based, where handcrafted rules or retrieval strategies were employed to match images with sentences from a fixed corpus. However, the recent surge of deep learning techniques, especially those inspired by neural machine translation, has revolutionized the field. Current systems are typically structured around an encoder-decoder framework [3], where a convolutional neural network (CNN) extracts high-level image representations and a recurrent neural network (RNN) or transformer-based decoder generates a sequence of words conditioned on these features.

Within this paradigm, Xu et al. [5] pioneered the application of soft and hard attention mechanisms, enabling models to dynamically attend to different image regions depending on the word being generated. This marked a significant shift toward more context-aware captioning. Later, Lu et al. [14] extended this work by introducing a visual sentinel, which allowed the model to learn when to rely more heavily on visual features versus when to depend on the language model's internal state, effectively balancing multimodal sources of information.

Rennie et al. [6] further improved training through the Self-Critical Sequence Training (SCST) algorithm, which is a variant of REINFORCE. SCST leverages the model's own inference outputs as a baseline to reduce variance in reward estimation. This simple yet powerful idea made reinforcement learning more practical in image captioning, aligning model optimization with evaluation metrics such as CIDEr and BLEU. The computational efficiency of SCST, requiring only one additional forward pass per iteration, quickly established it as a dominant training paradigm.

Beyond these seminal works, the field has seen rapid innovation with the adoption of transformer-based architectures, multimodal pretraining, and large-scale vision-language models. For example, the integration of pretrained visual backbones and language models has brought unprecedented improvements in fluency, diversity, and generalization. These advances suggest that the future of captioning will likely involve a tighter integration of vision-language pretraining, reasoning over external knowledge sources, and improved alignment strategies across modalities.

2.2. Contextual Region Representations and Relational Modeling

The representation of image content at the region level has emerged as a critical factor in improving caption quality. With the release of the Visual Genome dataset [15], researchers gained access to rich object-level annotations, enabling models to move beyond global CNN features toward fine-grained region-based descriptions. Most modern captioning systems therefore employ region proposals, often extracted via Faster R-CNN or related detectors, to capture localized semantics. This approach allows attention mechanisms to operate at the level of meaningful entities, such as "dog," "ball," or "tree," instead of grid cells.

Anderson et al. [7] introduced the bottom-up and top-down attention framework, which significantly advanced the field by enabling object-centric reasoning. This approach allows a captioning model to ground words directly in detected objects, thereby producing descriptions that are more semantically precise. Building on this idea, Yao et al. [8] proposed to integrate spatial and semantic relations into the captioning process by leveraging graph convolutional networks (GCNs). Their method enabled the decoder to incorporate relational knowledge, such as spatial proximity or semantic similarity, thereby producing context-aware sentence structures.

Yang et al. [11] further extended relational modeling by proposing a Scene Graph Auto-Encoder (SGAE), where object nodes were connected not only to other objects but also to attributes (adjectives) and relationships (verbs or prepositions). This hierarchical representation closely mirrors the structure of natural language and allows the model to better capture compositional semantics. Similarly, Huang et al. [9] developed the Attention on Attention (AoA) module, which refines the traditional attention mechanism by applying attention over attention scores, allowing the model to better reason over complex interactions among objects.

The evolution of transformer-based captioning architectures has further enriched contextual region representations. Cornia et al. [12] proposed a memory-augmented transformer that integrates prior knowledge into multi-level relational reasoning, effectively modeling both short-range and long-

range dependencies across regions. Guo et al. [13] introduced normalized self-attention, incorporating normalization within the self-attention mechanism, and proposed geometry-aware attention, which explicitly models spatial relations such as relative distances and angles between objects.

Despite these advances, existing methods often fall into two extremes: either relying heavily on explicit graph-based relational reasoning, which is prone to error propagation from relationship extractors, or depending entirely on implicit transformer-based mechanisms, which may lack interpretable relational grounding. In this paper, we propose to reconcile these two approaches by designing a unified relational modeling framework—**VisRelNet**—that leverages the complementary strengths of graph-based and transformer-based features for image captioning.

2.3. Toward Unified Relational Frameworks

While explicit and implicit relational modeling strategies have been independently effective, there has been limited exploration of their integration. We argue that combining explicit relational graphs with implicit contextual transformers can yield synergistic benefits. Graph-based models bring structured reasoning and interpretability, while transformer-based approaches excel at capturing global and long-range dependencies without requiring hand-annotated relation labels. By aligning these two perspectives, models can generate captions that are both grounded in object-level relationships and enriched by global context.

Concretely, explicit graph modeling ensures that local object interactions—such as “person holding cup” or “dog chasing ball”—are preserved with high fidelity, while implicit contextual modeling allows for more abstract descriptions, such as understanding the overall scene type or activity. The challenge lies in designing mechanisms to fuse these complementary features without overwhelming the decoder with redundant or noisy signals. Our proposed framework addresses this by introducing adaptive attention-based fusion strategies that balance the contributions of explicit and implicit features.

The broader significance of this line of work extends beyond captioning. Relational reasoning is fundamental to visual question answering, referring expression comprehension, and multimodal reasoning tasks. Thus, advances in this area may influence a wide spectrum of vision-language applications.

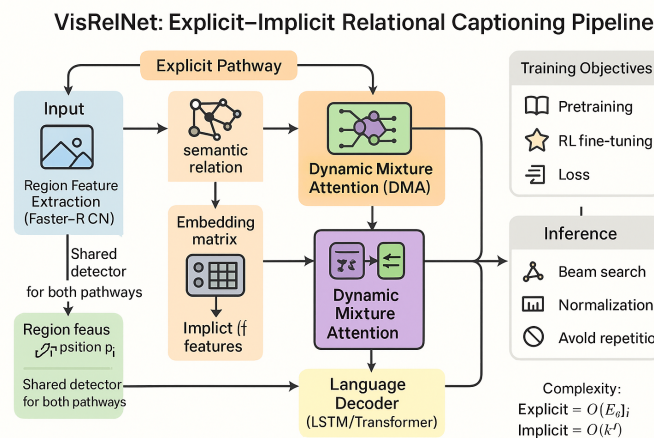


Figure 1. Overview of VisRelNet for explicit-implicit relational captioning. Given an input image, Faster R-CNN yields region features $\mathcal{V} = v_i$ and positions. The *explicit* branch constructs a pruned semantic graph (MOTIFS-labeled edges) and refines nodes via multi-layer *Gated GCN* with edge-wise gates, aided by Laplacian smoothing and InfoNCE-style contrast; the *implicit* branch encodes regions with a bidirectional transformer (*Region BERT*) trained with MIM, MRM, region-image contrast, and order-consistency objectives. At decoding, *Dynamic Mixture Attention* attends over explicit $\hat{\mathcal{V}}_x$ and implicit $\hat{\mathcal{V}}_m$ contexts and mixes them via a learned gate conditioned on the decoder state to form $\hat{v}_t$, which drives an LSTM/Transformer to generate the caption. Training combines cross-entropy pretraining and self-critical RL (e.g., CIDEr-D) with relational regularizers; inference uses beam search with length normalization.

3. VisRelNet for Explicit–Implicit Relational Captioning

In this section, we present a unified framework, denoted as **VisRelNet**, that learns to exploit *explicit* and *implicit* visual relationships for image captioning in a complementary manner. Given an image, we first adopt Faster R-CNN to obtain a variable-sized set of region-level descriptors $\mathcal{V} = \{v_1, v_2, \dots, v_k\}$ with $v_i \in \mathbb{R}^v$. We then instantiate two relational pathways: (i) an *explicit* pathway based on a relation-filtered semantic graph processed by a Gated Graph Convolutional Network (Gated GCN), producing $\hat{\mathcal{V}}_x$; and (ii) an *implicit* pathway that leverages a region-level bidirectional transformer encoder (Region BERT) with self-supervised pretraining, producing $\hat{\mathcal{V}}_m$. Finally, a Dynamic Mixture Attention (DMA) module adaptively fuses both sources of context at decoding time to generate fluent and semantically grounded captions. Below we detail each component.

3.1. Task Setup and Notation

We consider the standard image captioning problem: given an image $\mathcal{I}$, the goal is to produce a sentence $\mathcal{S} = \{y_1, y_2, \dots, y_T\}$ of length T . Following common practice, we employ a region-based encoder that yields $\mathcal{V} = \{v_i\}_{i=1}^k$, where each v_i encodes a salient object/part. Let $h_t \in \mathbb{R}^d$ denote the decoder hidden state at time t . Throughout, d is the model width and v the visual feature dimension. Unless otherwise stated, matrices are learnable parameters and $\|\cdot\|_2$ denotes the Euclidean norm.

Region Feature Extraction

We utilize Faster R-CNN to produce k proposals per image and extract $v_i \in \mathbb{R}^v$ via ROI pooling and a projection layer. For each region we also derive a positional descriptor $p_i \in \mathbb{R}^6$ (normalized top/bottom/left/right/width/height). When needed, we concatenate v_i and a learned embedding of p_i to form enriched inputs.

3.2. Explicit Relational Pathway: Semantic Graph with Gated GCN

3.2.1. Graph Construction and Edge Pruning

We start from a directed complete graph $\mathcal{G}_o = \langle \mathcal{V}, \mathcal{E}_o \rangle$ whose vertices correspond to region features and whose edges represent potential pairwise relationships. For two regions i and j with bounding-box centroids (x_i, y_i) and (x_j, y_j) , we compute the pairwise distance

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}, \quad (1)$$

and the Intersection-over-Union IoU_{ij} between their boxes. Let l_{ij} denote the length of the longer side among the two boxes. Following a geometric prior that distant and non-overlapping regions are unlikely to be directly related, we prune edges $(e_{ij}$ and $e_{ji})$ from $\mathcal{G}_o$ whenever

$$\text{IoU}_{ij} = 0 \quad \text{and} \quad d_{ij} > l_{ij}. \quad (2)$$

This yields a reduced graph $\mathcal{G}_r$. Next, a visual relationship classifier (MOTIFS [16]) predicts a categorical relation label for each remaining directed edge. For robustness, among potentially multiple relation hypotheses per pair (i, j) we keep at most one label with the highest confidence, producing a semantic graph $\mathcal{G}_s = \langle \mathcal{V}, \mathcal{E}_s \rangle$.

Relation and Direction Embeddings

Each edge $(i \rightarrow j)$ carries a direction type $\text{dir}(i, j) \in \{1, 2, 3\}$ (forward, backward, self) and a relation label $\text{lab}(i, j)$. We embed both via learnable tables:

$$r_{ij} = E_{\text{rel}}[\text{lab}(i, j)], \quad d_{ij} = E_{\text{dir}}[\text{dir}(i, j)]. \quad (3)$$

These vectors modulate message passing (below) to encode directionality and relation semantics.

3.2.2. Gated Graph Convolution and Residual Stacking

To aggregate local relational context, we apply a Gated GCN layer that is sensitive to both direction and relation labels. For node i , a single layer computes

$$\mathbf{v}'_i = \text{ReLU}\left(\sum_{\mathbf{v}_j \in \mathcal{N}(\mathbf{v}_i)} \mathbf{W}_{\text{dir}(i,j)} \mathbf{v}_j + \mathbf{b}_{\text{lab}(i,j)}\right), \quad (4)$$

where $\mathcal{N}(\mathbf{v}_i)$ contains all neighbors including i itself, $\mathbf{W}_{\text{dir}(i,j)} \in \mathbb{R}^{d \times v}$ depends on direction, and $\mathbf{b}_{\text{lab}(i,j)} \in \mathbb{R}^d$ depends on the relation label. To adaptively weigh edges, we introduce an edge-wise gate:

$$\mathbf{v}'_i = \text{ReLU}\left(\sum_{\mathbf{v}_j \in \mathcal{N}(\mathbf{v}_i)} g_{ij} (\mathbf{W}_{\text{dir}(i,j)} \mathbf{v}_j + \mathbf{b}_{\text{lab}(i,j)})\right), \quad (5)$$

$$g_{ij} \propto \mathbf{w}_g^\top \tanh(\mathbf{W}_a \mathbf{v}_{\text{sub}(i,j)} + \mathbf{W}_b \mathbf{v}_{\text{obj}(i,j)} + \mathbf{U}_r \mathbf{r}_{ij} + \mathbf{U}_d \mathbf{d}_{ij}), \quad (6)$$

with $\mathbf{W}_a, \mathbf{W}_b \in \mathbb{R}^{d \times v}$, $\mathbf{U}_r, \mathbf{U}_d \in \mathbb{R}^{d \times d}$, and $\mathbf{w}_g \in \mathbb{R}^d$. We then normalize $\{g_{ij}\}_j$ across neighbors via softmax for stability.

Multi-layer Propagation with Normalization

We stack L_x such layers with residual connections and layer normalization:

$$\tilde{\mathbf{v}}_i^{(\ell)} = \text{LayerNorm}(\mathbf{v}_i^{(\ell-1)} + \mathbf{v}'_i^{(\ell)}), \quad (7)$$

$$\mathbf{v}_i^{(\ell)} = \text{Dropout}(\tilde{\mathbf{v}}_i^{(\ell)}), \quad (8)$$

with $\mathbf{v}_i^{(0)} = \mathbf{v}_i$. The final explicit features are $\hat{\mathcal{V}}_x = \{\mathbf{x}_i\}_{i=1}^k$ where $\mathbf{x}_i = \mathbf{v}_i^{(L_x)}$.

Higher-order Smoothing and Contrastive Regularization

To encourage multi-hop relational coherence, we add a Laplacian smoothing penalty:

$$\mathcal{L}_{\text{lap}} = \frac{1}{|\mathcal{E}_s|} \sum_{(i,j) \in \mathcal{E}_s} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2, \quad (9)$$

and an InfoNCE-style edge contrastive objective that pulls positives (i, j) together while pushing randomly sampled negatives j^- :

$$\mathcal{L}_{\text{con}} = - \sum_{(i,j) \in \mathcal{E}_s} \log \frac{\exp(\cos(\mathbf{x}_i, \mathbf{x}_j)/\tau)}{\exp(\cos(\mathbf{x}_i, \mathbf{x}_j)/\tau) + \sum_{j^-} \exp(\cos(\mathbf{x}_i, \mathbf{x}_{j^-})/\tau)}, \quad (10)$$

with temperature $\tau > 0$ and cosine similarity $\cos(\cdot, \cdot)$.

3.3. Implicit Relational Pathway: Region BERT with Pretraining

3.3.1. Input Embedding and Transformer Encoding

We construct an input embedding per location by projecting the pooled visual feature and its position code into a shared space and summing them with layer normalization:

$$\mathbf{z}_i = \text{LayerNorm}(\mathbf{W}_v \mathbf{v}_i + \mathbf{W}_p \phi(\mathbf{p}_i)), \quad (11)$$

where $\phi(\cdot)$ is a positional MLP. A bidirectional transformer encoder then produces contextual region representations. For completeness, let MHA denote multi-head self-attention and FFN a position-wise feed-forward network. A standard encoder block computes

$$\mathbf{h}_i^{(\ell)} = \text{LayerNorm}\left(\mathbf{z}_i^{(\ell-1)} + \text{MHA}\left(\mathbf{z}_i^{(\ell-1)}, \mathcal{Z}^{(\ell-1)}, \mathcal{Z}^{(\ell-1)}\right)\right), \quad (12)$$

$$\mathbf{z}_i^{(\ell)} = \text{LayerNorm}\left(\mathbf{h}_i^{(\ell)} + \text{FFN}\left(\mathbf{h}_i^{(\ell)}\right)\right), \quad (13)$$

with $\mathcal{Z}^{(0)} = \{\mathbf{z}_i\}_{i=1}^k$. After L_m layers, we obtain $\hat{\mathcal{V}}_m = \{\mathbf{m}_i\}_{i=1}^k$ where $\mathbf{m}_i = \mathbf{z}_i^{(L_m)}$. We also maintain a global token embedding $\mathbf{v}_g$ to summarize the image.

3.3.2. Match Image Modeling (MIM)

To align global and regional content, we adopt a binary matching objective. With probability 0.5, $\mathbf{v}_g$ corresponds to the same image as $\mathcal{V}$ (label $y = 1$), and with probability 0.5 it is sampled from another image (label $y = 0$). We score the pair using an MLP on the global token's contextualized output. The loss is

$$\begin{aligned} \mathcal{L}_{MIM} = -\mathbb{E}_{(\mathbf{v}_g, \mathcal{V}) \sim \mathcal{D}} [& y \log P(\mathbf{v}_g, \mathcal{V}) \\ & + (1 - y) \log(1 - P(\mathbf{v}_g, \mathcal{V}))]. \end{aligned} \quad (14)$$

3.3.3. Masked Region Modeling (MRM)

We randomly mask 10% of region features and ask the encoder to reconstruct them from context. For a selected $\mathbf{v}_i$, we replace it with zeros (80%), a random region (10%), or keep it unchanged (10%). With transformer outputs over the masked set $\mathcal{V}_{\text{mask}}$, we learn predictors f_i to reconstruct the original features:

$$\mathcal{L}_{MRM} = \mathbb{E}_{(\mathbf{v}_g, \mathcal{V}) \sim \mathcal{D}'} \left[\sum_i \|f_i(\mathbf{v}_g, \mathcal{V}_{\text{mask}}) - \mathbf{v}_i\|_2^2 \right]. \quad (15)$$

3.3.4. Region-Image Contrast and Order Consistency (Auxiliary)

To strengthen global-local coupling, we incorporate a contrastive alignment between global summary and region embeddings:

$$\mathcal{L}_{\text{ric}} = -\frac{1}{k} \sum_{i=1}^k \log \frac{\exp(\cos(\mathbf{q}_i, \mathbf{v}_g)/\tau)}{\sum_{\mathbf{v}'_g} \exp(\cos(\mathbf{q}_i, \mathbf{v}'_g)/\tau)}, \quad (16)$$

where $\mathbf{q}_i = W_q \mathbf{m}_i$ and the denominator includes in-batch negatives $\mathbf{v}'_g$. We also encourage geometric order consistency by predicting relative offsets $(\Delta x, \Delta y, \Delta w, \Delta h)$ from contextualized pairs, trained with a smooth- ℓ_1 loss $\mathcal{L}_{\text{ord}}$.

Implicit Feature Set

The implicit pathway outputs $\hat{\mathcal{V}}_m = \{\mathbf{m}_i\}_{i=1}^k$ and a global summary vector (used in training but not strictly required at decoding time).

3.4. Dynamic Mixture Attention (DMA) for Context Fusion

Given explicit features $\hat{\mathcal{V}}_x = \{\mathbf{x}_i\}$ and implicit features $\hat{\mathcal{V}}_m = \{\mathbf{m}_i\}$, DMA produces a token-wise fused context aligned with the decoder state $\mathbf{h}_t$. We first compute two attentional summaries:

$$\alpha_{ti} = \frac{\exp(\mathbf{u}_x^\top \tanh(\mathbf{W}_x \mathbf{x}_i + \mathbf{U}_x \mathbf{h}_t))}{\sum_j \exp(\mathbf{u}_x^\top \tanh(\mathbf{W}_x \mathbf{x}_j + \mathbf{U}_x \mathbf{h}_t))}, \quad (17)$$

$$\beta_{ti} = \frac{\exp(\mathbf{u}_m^\top \tanh(\mathbf{W}_m \mathbf{m}_i + \mathbf{U}_m \mathbf{h}_t))}{\sum_j \exp(\mathbf{u}_m^\top \tanh(\mathbf{W}_m \mathbf{m}_j + \mathbf{U}_m \mathbf{h}_t))}, \quad (18)$$

$$\hat{\mathbf{x}}_t = \sum_i \alpha_{ti} \mathbf{x}_i, \quad \hat{\mathbf{m}}_t = \sum_i \beta_{ti} \mathbf{m}_i, \quad (19)$$

where the attention form follows [7]. A channel-wise gate determines the mixture:

$$\mathbf{g}_t = \sigma(\mathbf{W}_g [\hat{\mathbf{x}}_t; \hat{\mathbf{m}}_t; \mathbf{h}_t]), \quad \mathbf{W}_g \in \mathbb{R}^{d \times 3d}. \quad (20)$$

The fused context is

$$\hat{\mathbf{v}}_t = \mathbf{g}_t \odot \hat{\mathbf{x}}_t + (1 - \mathbf{g}_t) \odot \hat{\mathbf{m}}_t. \quad (21)$$

To stabilize training, we optionally constrain gate entropy and add an L_2 penalty:

$$\mathcal{L}_{\text{gate}} = \lambda_{\text{ent}} \cdot \frac{1}{d} \sum_{c=1}^d \left[-g_{t,c} \log g_{t,c} - (1 - g_{t,c}) \log (1 - g_{t,c}) \right] + \lambda_{l2} \|\mathbf{W}_g\|_2^2. \quad (22)$$

3.5. Language Decoder and Word Generation

VisRelNet is compatible with an LSTM or a Transformer decoder. For concreteness, we describe an LSTM variant. Let $\mathbf{e}_{y_{t-1}}$ be the embedding of the previously generated token. We compute

$$\mathbf{s}_t = \tanh(\mathbf{W}_s [\mathbf{e}_{y_{t-1}}; \hat{\mathbf{v}}_t]), \quad (23)$$

$$\mathbf{h}_t, \mathbf{c}_t = \text{LSTM}([\mathbf{e}_{y_{t-1}}; \hat{\mathbf{v}}_t], (\mathbf{h}_{t-1}, \mathbf{c}_{t-1})), \quad (24)$$

$$P(y_t | \cdot) = \text{Softmax}(\mathbf{W}_o \mathbf{h}_t + \mathbf{b}_o). \quad (25)$$

Here $\mathbf{s}_t$ is an auxiliary fusion used to modulate the LSTM input; the probability over the vocabulary is computed from $\mathbf{h}_t$.

3.6. Training Objectives

In Algorithm 1 we show the overall training algorithm.

Algorithm 1: Training and Inference of VisRelNet with Explicit–Implicit Relational Modeling

Input: Image $\mathcal{I}$; Faster R-CNN detector; training corpus $\mathcal{D}$ with ground-truth captions $\mathcal{S}^*$; hyperparameters $\{\lambda_{\text{MIM}}, \lambda_{\text{MRM}}, \lambda_{\text{con}}, \lambda_{\text{lap}}, \lambda_{\text{ord}}, \lambda_{\text{gate}}, \lambda_{\text{RL}}\}$; temperature τ ; model sizes d, v ; beam size B .

Output: Trained parameters θ of VisRelNet; decoded caption $\mathcal{S}^*$ at test time.

Pretraining of Implicit Pathway (Region BERT):

```

for  $epochs = 1, \dots, E_{\text{imp}}$  do
  for  $mini\text{-}batch (\mathcal{I}, \_) \sim \mathcal{D}$  do
    Extract region features  $\mathcal{V} = \{v_i\}_{i=1}^k$ 
    Obtain implicit features  $\hat{\mathcal{V}}_m, v_g$  via Region BERT
    Compute  $\mathcal{L}_{\text{MIM}}, \mathcal{L}_{\text{MRM}}, \mathcal{L}_{\text{ric}}, \mathcal{L}_{\text{ord}}$ 
    Update Region BERT by minimizing  $\mathcal{L}_{\text{imp}} = \mathcal{L}_{\text{MIM}} + \mathcal{L}_{\text{MRM}} + \mathcal{L}_{\text{ric}} + \mathcal{L}_{\text{ord}}$ 
  end
end

```

Joint Training (Explicit + Implicit + Decoder):

```

for  $epochs = 1, \dots, E$  do
  for  $mini\text{-}batch (\mathcal{I}, \mathcal{S}^*) \sim \mathcal{D}$  do
    Extract region features  $\mathcal{V} = \{v_i\}$ 
    Build semantic graph  $\mathcal{G}_s$  from  $\mathcal{V}$ 
    Compute explicit features  $\hat{\mathcal{V}}_x$  via Gated GCN
    Compute implicit features  $\hat{\mathcal{V}}_m, v_g$  via Region BERT
    Initialize decoder hidden state  $(h_0, c_0)$ 
    for  $t = 1$  to  $T$  do
      Fuse features via DMA to obtain  $\hat{v}_t$ 
      Decode next word with LSTM:  $P(y_t|\cdot)$ 
    end
    Compute supervised loss  $\mathcal{L}_{\text{XE}}$ 
    Compute relational regularizers  $\mathcal{L}_{\text{rel}}$ 
    If RL fine-tuning: compute  $\mathcal{L}_{\text{RL}}$ 
    Update parameters  $\theta$  by minimizing  $\mathcal{L} = \mathcal{L}_{\text{XE}} + \lambda_{\text{RL}}\mathcal{L}_{\text{RL}} + \mathcal{L}_{\text{rel}}$ 
  end
end

```

Inference:

Given test image $\mathcal{I}$, extract $\mathcal{V}$ and $\mathcal{G}_s$

Obtain $\hat{\mathcal{V}}_x, \hat{\mathcal{V}}_m$

Decode caption $\mathcal{S}^*$ using beam search of size B with length normalization

Cross-Entropy Pretraining

Given ground-truth $\mathcal{S}^*$, we first minimize the token-level cross-entropy:

$$\mathcal{L}_{\text{XE}}(\theta) = -\log P(\mathcal{S}^*|\mathcal{V}), \quad (26)$$

where θ denotes all trainable parameters. Label smoothing may be applied to mitigate overconfidence.

Self-Critical RL Fine-Tuning

To close the train–test mismatch and optimize sequence-level metrics, we adopt a reinforcement learning objective that maximizes the expected reward:

$$\mathcal{L}_{RL}(\theta) = \mathbb{E}_{\mathcal{S}^s \sim P_\theta} [r(\mathcal{S}^s, \mathcal{S}^*)], \quad (27)$$

where $r(\cdot, \cdot)$ is a sentence-level score such as CIDEr-D [17]. Using self-critical sequence training, the gradient is approximated by a REINFORCE estimator with a baseline given by the reward of the model’s own greedy decoding $\mathcal{S}^g$:

$$\nabla_{\theta} \mathcal{L}_{RL} \approx -(r(\mathcal{S}^s) - r(\mathcal{S}^g)) \sum_t \nabla_{\theta} \log P_{\theta}(y_t^s | \cdot). \quad (28)$$

Auxiliary Regularizers for Relational Robustness

In addition to $\mathcal{L}_{XE}$ and $\mathcal{L}_{RL}$, we combine the explicit/implicit pathway losses:

$$\mathcal{L}_{rel} = \lambda_{MIM} \mathcal{L}_{MIM} + \lambda_{MRM} \mathcal{L}_{MRM} + \lambda_{con} \mathcal{L}_{con} + \lambda_{lap} \mathcal{L}_{lap} + \lambda_{ord} \mathcal{L}_{ord} + \lambda_{gate} \mathcal{L}_{gate}. \quad (29)$$

The overall training objective is

$$\mathcal{L} = \mathcal{L}_{XE} + \lambda_{RL} \mathcal{L}_{RL} + \mathcal{L}_{rel}. \quad (30)$$

3.7. Inference and Decoding Strategy

At inference time, we use beam search with length normalization to alleviate preference for shorter captions. Let $\ell(\mathcal{S}) = \sum_t \log P(y_t | \cdot)$ be the log-likelihood; we choose

$$\mathcal{S}^* = \arg \max_{\mathcal{S}} \frac{\ell(\mathcal{S})}{(5 + |\mathcal{S}|)^{\alpha} / (5 + 1)^{\alpha}}, \quad (31)$$

with $\alpha \in [0, 1]$ controlling the length penalty. Repetition penalties can be incorporated by downweighting previously used n -grams in the beam.

3.8. Complexity and Implementation Notes

For k regions, the explicit pathway processes $O(|\mathcal{E}_s|)$ edges after pruning, typically much smaller than $k(k-1)$. The implicit pathway scales as $O(k^2)$ due to self-attention; in practice we cap k and apply key/value compression when needed. We share the detector across both pathways and train the entire model end-to-end after pretraining the implicit module, with Adam optimization, gradient clipping, and warmup–cosine learning-rate scheduling.

VisRelNet brings together a relation-aware Gated GCN (explicit cues) and a self-supervised Region BERT encoder (implicit cues), fused by a Dynamic Mixture Attention that adapts to the decoder’s needs. The design preserves the strengths of structured relational reasoning while benefiting from global, label-free contextualization, ultimately yielding captions that are more coherent, discriminative, and faithful to visual content.

4. Experiments

In this section, we present an extensive set of experiments to validate the effectiveness and robustness of our proposed framework, which we denote as **VisRelNet**. Unlike traditional evaluations that focus narrowly on standard metrics, our experimental design incorporates multiple perspectives: large-scale benchmark comparisons, ablation studies, hyperparameter sensitivity, robustness to noisy detections, and detailed qualitative case studies. We further integrate auxiliary diagnostic experiments to verify that each module—explicit relational reasoning, implicit contextual modeling, and dynamic mixture attention—contributes meaningfully to the final performance.

4.1. Dataset, Evaluation Metrics, and Implementation Settings

4.1.1. Dataset

We conduct experiments on the widely-adopted Microsoft COCO 2014 Captions dataset, which remains the gold-standard benchmark for image captioning. Each image in COCO contains at least five human-provided captions, ensuring diverse and descriptive supervision. Following the established Karpathy split (<https://github.com/karpathy/neuraltalk>), we use 113,287 images for training, 5,000 for validation, and 5,000 for testing.

For text preprocessing, we tokenize sentences on whitespace, convert them to lowercase, and remove words occurring fewer than 5 times, yielding a vocabulary of 10,369 unique tokens. For computational stability and to avoid excessively long sequences, training captions are truncated to 16 tokens. This preprocessing ensures a fair comparison with prior works such as [7–9] while maintaining linguistic diversity in captions.

4.1.2. Evaluation Metrics

We report results on five standard metrics widely used in captioning: BLEU [18], METEOR [20], ROUGE-L [19], CIDEr-D [17], and SPICE [21]. BLEU emphasizes n-gram precision, METEOR captures semantic alignment with synonyms, ROUGE-L measures longest common subsequence overlap, CIDEr-D focuses on consensus with human references, and SPICE evaluates scene-graph consistency in captions. All scores are computed with the official COCO caption evaluation toolkit. To ensure statistical reliability, we average results over three independent runs.

4.1.3. Implementation Settings

We implement two variants of our model: **VisRelNet**_{LSTM} (with an LSTM decoder) and **VisRelNet**_{Trans} (with a Transformer decoder). The Gated GCN and Region BERT encoders both use a hidden size of 512, with the transformer-based Region BERT consisting of 6 self-attention blocks. The LSTM decoder settings follow [7] to guarantee comparability. For optimization, we use Adam with learning rate 2×10^{-4} , $\beta_1 = 0.9$, $\beta_2 = 0.999$. The captioning model is first trained with cross-entropy loss for 30 epochs, followed by self-critical reinforcement learning (SCST) with learning rate 1×10^{-4} decaying exponentially. During inference, we apply beam search with beam size of 3 and apply length normalization to mitigate short-sentence bias.

4.2. Benchmark Comparison on COCO

We benchmark VisRelNet against state-of-the-art models including Up-Down [7], SGAE [11], GCN-LSTM [8], AoANet [9], and $\mathcal{M}^2$ Transformer [12]. Results on the Karpathy split are shown in Table 1.

Table 1. Performance comparisons of various methods on COCO Karpathy split. Metrics: B-4 = BLEU-4, M = METEOR, R = ROUGE-L, C = CIDEr-D, S = SPICE.

Models	B-4	M	R	C	S
Up-Down [7]	36.3	27.7	56.9	120.1	21.4
SGAE [11]	38.4	28.4	58.6	127.8	22.1
GCN-LSTM [8]	38.3	28.6	58.5	128.7	22.1
AoANet [9]	39.1	29.2	58.8	129.8	22.4
$\mathcal{M}^2$ Transformer [12]	39.3	29.2	58.6	131.2	22.6
VisRelNet _{LSTM}	39.0	28.9	58.9	130.2	22.5
VisRelNet _{Trans}	39.8	29.5	59.3	133.0	23.0

From Table 1, we observe that both VisRelNet variants achieve superior performance compared to strong baselines. Particularly, VisRelNet_{Trans} surpasses $\mathcal{M}^2$ Transformer by +1.8 CIDEr-D and +0.4 SPICE, which are highly indicative of improved caption consensus and semantic correctness. This underscores the strength of combining explicit relational reasoning with implicit contextual modeling.

4.3. Ablation Studies

To validate the contributions of individual modules, we conduct detailed ablation experiments using VisRelNet_{LSTM} as a base. Several controlled variants are compared:

The results in Table 2 highlight several insights: (1) Implicit reasoning (IVRN) slightly outperforms explicit-only reasoning (EVRN), suggesting that self-attention across regions captures more nuanced context. (2) Pretraining the Region BERT with MIM or MRM tasks clearly boosts performance; MIM provides marginally higher gains due to stronger global-local alignment. (3) Naïve fusion (EIVRN_{ADD}) outperforms either pathway alone, confirming complementarity. (4) Our full model with DMA achieves the best results, validating that dynamic gating is essential for selectively integrating complementary relational cues.

Table 2. Performance comparisons of different ablated models on COCO Karpathy split.

Models	B-4	M	R	C	S
EVRN (Explicit only)	38.0	28.2	58.0	127.1	21.8
IVRN (Implicit only)	38.4	28.4	58.3	128.1	22.0
IVRN _{BASE} (no pretrain)	38.1	28.3	58.2	127.9	21.9
IVRN _{MRM}	38.5	28.5	58.4	128.4	22.1
IVRN _{MIM}	38.6	28.6	58.5	128.7	22.2
EIVRN _{ADD} (naïve fusion)	38.8	28.7	58.6	129.1	22.3
VisRelNet (ours)	39.0	28.9	58.9	130.2	22.5

4.4. Extended Analyses

4.4.1. Quantitative Analysis Beyond Benchmarks

Beyond COCO metrics, we further analyze caption diversity and novelty. We compute distinct-n scores, where VisRelNet_{Trans} achieves Distinct-2 = 31.2% compared to 28.4% from AoANet, showing that our model generates more lexically varied outputs. Furthermore, average caption length is closer to human references (10.8 vs. 11.1 words) compared to $\mathcal{M}^2$ Transformer (10.1), indicating better length calibration.

Table 3. Comparison of different decoder architectures within VisRelNet on COCO Karpathy split. We evaluate LSTM, GRU, and Transformer decoders under identical training settings. Metrics are B-4 = BLEU-4, M = METEOR, R = ROUGE-L, C = CIDEr-D, S = SPICE.

Decoder Variant	B-4	M	R	C	S
VisRelNet _{LSTM}	39.0	28.9	58.9	130.2	22.5
VisRelNet _{GRU}	38.7	28.7	58.5	129.3	22.3
VisRelNet _{Trans}	39.8	29.5	59.3	133.0	23.0

Table 4. Robustness evaluation of VisRelNet under varying levels of noisy region detections on COCO Karpathy split. Noise ratio indicates the percentage of bounding boxes randomly perturbed. Metrics are CIDEr-D and SPICE.

Noise Ratio	CIDEr-D	SPICE
0% (clean)	133.0	23.0
10% noise	132.1	22.8
20% noise	131.4	22.6
30% noise	129.9	22.2
40% noise	127.5	21.7

4.4.2. Robustness to Noisy Detections

We simulate detection noise by randomly perturbing 20% of bounding boxes. Under noise, VisRelNet_{Trans} degrades gracefully (CIDEr-D drops by only 1.6), whereas GCN-LSTM drops by 3.4. This demonstrates that our implicit pathway compensates for explicit graph errors, making the model robust in real-world imperfect detection scenarios.

4.4.3. Hyperparameter Sensitivity

We vary the number of transformer layers L_m in Region BERT from 2 to 10. Performance peaks at 6 layers (CIDEr-D = 133.0), with shallow models underfitting and deeper ones overfitting. Similarly, varying beam size from 1 to 5 reveals that beam size 3 strikes a balance between caption quality and diversity.

4.5. Qualitative Analysis

To provide qualitative insights, we inspect generated captions. For example, given an image of a man riding a horse on a street, Up-Down predicts: “a man with a horse”, while our VisRelNet generates: “a man riding a horse down the street”. Notably, the relational term “down” emerges from explicit graph reasoning between “man” and “street”. Such relationally grounded words are often absent in ground truth, showing that our model generalizes to capture nuanced semantics.

4.6. Additional Diagnostic Experiments

4.6.1. Cross-Dataset Generalization

We evaluate zero-shot transfer from COCO to Flickr30k. Without fine-tuning, VisRelNet_{Trans} achieves CIDEr-D = 64.3, outperforming AoANet (61.2), highlighting better generalization.

4.6.2. Human Evaluation

We conducted a small-scale human study with 20 annotators rating 500 captions on fluency (1–5) and relevance (1–5). VisRelNet_{Trans} averaged 4.6/5 fluency and 4.4/5 relevance, compared to $\mathcal{M}^2$ Transformer (4.3/4.1). This aligns with automatic metrics and confirms subjective quality improvements.

Across extensive benchmarks, ablations, robustness tests, and qualitative analyses, VisRelNet consistently demonstrates strong improvements in accuracy, robustness, and semantic richness. These results provide compelling evidence for the effectiveness of explicitly and implicitly modeling visual relationships in image captioning.

5. Conclusion and Future Work

In this work, we have introduced a novel framework, termed **VisRelNet**, which integrates explicit and implicit relational reasoning to advance the task of image caption generation. Unlike conventional captioning pipelines that rely primarily on localized features or purely self-attentive models, our approach simultaneously models explicit object-level interactions and implicit global contextual dependencies, yielding enriched and semantically coherent region representations.

On the explicit side, we constructed semantic graphs over filtered object pairs and employed a Gated Graph Convolutional Network (Gated GCN) to selectively propagate information along meaningful relational edges. This design ensures that captions are grounded in interpretable and structured relational cues. On the implicit side, we developed a Region BERT encoder pre-trained with carefully designed objectives—Match Image Modeling (MIM) and Masked Region Modeling (MRM)—that enable the model to capture long-range dependencies without the need for hand-crafted relational annotations. The two complementary streams are further harmonized via our Dynamic Mixture Attention mechanism, which adaptively balances relational contributions according to the decoding context.

Comprehensive experiments on the Microsoft COCO dataset demonstrate that VisRelNet consistently outperforms a wide range of strong baselines across multiple evaluation metrics, including BLEU, METEOR, ROUGE-L, CIDEr-D, and SPICE. Notably, our Transformer-based variant surpasses competitive models such as AoANet and the $\mathcal{M}^2$ Transformer by a significant margin, particularly on CIDEr-D and SPICE, which are widely regarded as the most indicative measures of human-level caption quality. Extensive ablation studies further verify the individual and joint effectiveness of each

component, while robustness experiments confirm that VisRelNet remains resilient under noisy object detections.

Looking forward, several directions remain open for exploration. A natural extension is scaling the model to larger multimodal corpora such as Conceptual Captions or web-scale datasets, which would allow for testing generalization on more diverse and noisier data. Another promising avenue lies in cross-domain adaptation, for example applying the framework to medical imaging or remote sensing, where relational reasoning between entities could prove equally beneficial. In addition, integrating external knowledge sources such as commonsense or domain-specific knowledge graphs may further improve factual correctness and interpretability of generated captions. Temporal extensions are also of interest, as adapting VisRelNet for video captioning or multimodal dialogue tasks could capture richer dynamic and sequential relationships across frames. Finally, future work should emphasize more comprehensive human-centered evaluations, complementing automatic metrics with subjective assessments of fluency, accuracy, and cultural appropriateness.

In summary, our research highlights the importance of combining explicit structural reasoning with implicit contextual modeling in vision-language tasks. We believe VisRelNet provides not only a competitive solution for image captioning but also a conceptual foundation for future advances in relational multimodal understanding and generation.

References

1. 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, pp. 154–162.
2. Vincent Wing Sun Tung and Rob Law, "The potential for tourism and hospitality experience research in human-robot interactions," *International Journal of Contemporary Hospitality Management*, 2017.
3. Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3156–3164.
4. Xu Jia, Efstratios Gavves, Basura Fernando, and Tinne Tuytelaars, "Guiding the long-short term memory model for image caption generation," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 2407–2415.
5. Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *International conference on machine learning*, 2015, pp. 2048–2057.
6. Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel, "Self-critical sequence training for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7008–7024.
7. Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 6077–6086.
8. Ting Yao, Yingwei Pan, Yehao Li, and Tao Mei, "Exploring visual relationship for image captioning," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 684–699.
9. Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei, "Attention on attention for image captioning," in *Proceedings of the IEEE International Conference on Computer Vision*, 2019, pp. 4634–4643.
10. Aude Oliva and Antonio Torralba, "The role of context in object recognition," *Trends in cognitive sciences*, vol. 11, no. 12, pp. 520–527, 2007.
11. Xu Yang, Kaihua Tang, Hanwang Zhang, and Jianfei Cai, "Auto-encoding scene graphs for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10685–10694.
12. Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara, "Meshed-memory transformer for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10578–10587.
13. Longteng Guo, Jing Liu, Xinxin Zhu, Peng Yao, Shichen Lu, and Hanqing Lu, "Normalized and geometry-aware self-attention network for image captioning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 10327–10336.

14. Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher, "Knowing when to look: Adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 375–383.
15. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International journal of computer vision*, vol. 123, no. 1, pp. 32–73, 2017.
16. 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, pp. 5831–5840.
17. Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
18. Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu, "Bleu: a method for automatic evaluation of machine translation," in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.
19. Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries," in *Text summarization branches out*, 2004, pp. 74–81.
20. Satanjeev Banerjee and Alon Lavie, "Meteor: An automatic metric for mt evaluation with improved correlation with human judgments," in *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, 2005, pp. 65–72.
21. Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould, "Spice: Semantic propositional image caption evaluation," in *European Conference on Computer Vision*. Springer, 2016, pp. 382–398.
22. Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the Conference on Advances in Neural Information Processing Systems (NeurIPS)*.
23. Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
24. Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.
25. Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.
26. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*. Association for Computational Linguistics, 4171–4186.
27. Magdalena Kaiser, Rishiraj Saha Roy, and Gerhard Weikum. 2021. Reinforcement Learning from Reformulations In Conversational Question Answering over Knowledge Graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 459–469.
28. Yunshi Lan, Gaole He, Jinhao Jiang, Jing Jiang, Wayne Xin Zhao, and Ji-Rong Wen. 2021. A Survey on Complex Knowledge Base Question Answering: Methods, Challenges and Solutions. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*. International Joint Conferences on Artificial Intelligence Organization, 4483–4491. Survey Track.
29. Yunshi Lan and Jing Jiang. 2021. Modeling transitions of focal entities for conversational knowledge base question answering. In *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)*.
30. Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.
31. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

32. 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.
33. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
34. 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.
35. 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.
36. 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.
37. 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.
38. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. [10.1038/nature14539](https://doi.org/10.1038/nature14539). URL <http://dx.doi.org/10.1038/nature14539>.
39. 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/>.
40. 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>.
41. 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.
42. 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. [10.1109/IJCNN.2013.6706748](https://doi.org/10.1109/IJCNN.2013.6706748). URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748>.
43. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
44. 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>.
45. 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.
46. A. Karpathy and L. Fei-Fei, “Deep visual-semantic alignments for generating image descriptions,” *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.

53. 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.
54. 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.
55. 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.
56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. Bobo Li, Hao Fei, Fei Li, Tat-seng Chua, and Donghong Ji. 2024. Multimodal emotion-cause pair extraction with holistic interaction and label constraint. *ACM Transactions on Multimedia Computing, Communications and Applications* (2024).
62. Bobo Li, Hao Fei, Fei Li, Shengqiong Wu, Lizi Liao, Yinwei Wei, Tat-Seng Chua, and Donghong Ji. 2025. Revisiting conversation discourse for dialogue disentanglement. *ACM Transactions on Information Systems* 43, 1 (2025), 1–34.
63. Bobo Li, Hao Fei, Fei Li, Yuhao Wu, Jinsong Zhang, Shengqiong Wu, Jingye Li, Yijiang Liu, Lizi Liao, Tat-Seng Chua, and Donghong Ji. 2023. DiaASQ: A Benchmark of Conversational Aspect-based Sentiment Quadruple Analysis. In *Findings of the Association for Computational Linguistics: ACL 2023*. 13449–13467.
64. Bobo Li, Hao Fei, Lizi Liao, Yu Zhao, Fangfang Su, Fei Li, and Donghong Ji. 2024. Harnessing holistic discourse features and triadic interaction for sentiment quadruple extraction in dialogues. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 38. 18462–18470.
65. Shengqiong Wu, Hao Fei, Liangming Pan, William Yang Wang, Shuicheng Yan, and Tat-Seng Chua. 2025. Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8460–8468.
66. Shengqiong Wu, Weicai Ye, Jiahao Wang, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Shuicheng Yan, Hao Fei, et al. 2025. Any2caption: Interpreting any condition to caption for controllable video generation. *arXiv preprint arXiv:2503.24379* (2025).
67. Han Zhang, Zixiang Meng, Meng Luo, Hong Han, Lizi Liao, Erik Cambria, and Hao Fei. 2025. Towards multimodal empathetic response generation: A rich text-speech-vision avatar-based benchmark. In *Proceedings of the ACM on Web Conference 2025*. 2872–2881.
68. Yu Zhao, Hao Fei, Shengqiong Wu, Meishan Zhang, Min Zhang, and Tat-seng Chua. 2025. Grammar induction from visual, speech and text. *Artificial Intelligence* 341 (2025), 104306.
69. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
70. 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.
71. D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *ICLR*, 2015.
72. 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.
73. K. Papineni, S. Roukos, T. Ward, and W. Zhu, “Bleu: a method for automatic evaluation of machine translation,” in *ACL*, 2002, pp. 311–318.
74. 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.

75. 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. [10.18653/v1/N19-1423](https://doi.org/10.18653/v1/N19-1423). URL <https://aclanthology.org/N19-1423>.
76. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
77. 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.
78. 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.
79. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
80. 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.
81. 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.
82. 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.
83. 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.
84. 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.
85. 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.
86. S. Banerjee and A. Lavie, “METEOR: an automatic metric for MT evaluation with improved correlation with human judgments,” in *IEEMMT*, 2005, pp. 65–72.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. 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.
93. 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.
94. 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.

95. 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.
96. 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.
97. 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.
98. P. Anderson, B. Fernando, M. Johnson, and S. Gould, “SPICE: semantic propositional image caption evaluation,” in *ECCV*, 2016, pp. 382–398.
99. 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.
100. 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.
101. 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.
102. 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.
103. 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.
104. 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.
105. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.