Relation-Sensitive VQA with A Unified Tri-Modal Graph Framework

Jolien Van Bossche *, Thibault Clercq, Callum Hensley, Rune Peeters

Posted Date: 18 November 2025

doi: 10.20944/preprints202511.1373.v1

Keywords: scene graph reasoning; visual question answering; multimodal graph neural networks; semantic integration

 doi:10.20944/preprints202511.1373.v1

Article

Relation-Sensitive VQA with A Unified Tri-Modal Graph Framework

Jolien Van Bossche *, Thibault Clercq, Callum Hensley and Rune Peeters

University of Antwerp; jolien.vandenbossche@uantwerpen.be

Abstract

Visual question answering (VQA) fundamentally requires a model to interpret heterogeneous semantic cues in an image and align them with a natural-language query. Traditional approaches benefit from scene graph representations, yet they often suffer from severe imbalances when handling rich semantic structures, especially when reasoning demands simultaneous consideration of objects, relations, and fine-grained attributes. Existing models frequently overlook the subtle interactions among these three information streams, leading to faulty attribute inference or overlooked relational cues. Addressing these long-standing limitations calls for a more principled integration of all semantic constituents within a unified and expressive reasoning space. In this paper, we introduce **TriUnity-GNN**, a tri-modal fusion framework that redefines scene graph reasoning by jointly enhancing object-centric, relation-centric, and attribute-centric representations under a unified graph neural paradigm. Instead of treating scene graphs as monolithic structures, our approach restructures the given graph into two complementary modalities, an object-dominant perspective and a relation-dominant perspective, thereby enabling the model to capture multi-granular semantics that are typically under-explored. To further strengthen the expressivity of these representations, TriUnity-GNN integrates attribute cues through an explicit fusion design, significantly enlarging the impact of attribute signals that are otherwise marginalized in classic architectures. Moreover, we design a novel message-passing enhancement module that substantially increases cross-type semantic exchange among objects, relations, and attributes, ensuring that all three modalities collectively shape the final reasoning embedding. We perform comprehensive evaluations on benchmark datasets including GQA, VG, and motif-VG. Across all benchmarks, TriUnity-GNN consistently surpasses prior graph-based VQA systems by a clear margin, demonstrating robustness in handling both straightforward and semantically composite queries. The results verify that a tri-modal, explicitly balanced graph reasoning mechanism is crucial for improving interpretability and accuracy in challenging visual question answering scenarios.

Keywords: scene graph reasoning; visual question answering; multimodal graph neural networks; semantic integration

1. Introduction

Visual question answering (VQA) tasks demand an intelligent agent capable of interpreting a natural-language question and producing answers grounded in image content. A predominant line of research has adopted scene graphs (SGs) to represent visual semantics, encoding objects, attributes, and relationships as graph entities that can be exploited for structured reasoning [1]. By translating an image into a symbolic yet visually grounded representation, SGs allow models to interact with semantically organized information, bridging the gap between linguistic queries and visual evidence. These structured representations further exhibit advantages over classic feature-based pipelines due to their interpretability and modular decomposition of visual scenes [2,3].

 doi:10.20944/preprints202511.1373.v1

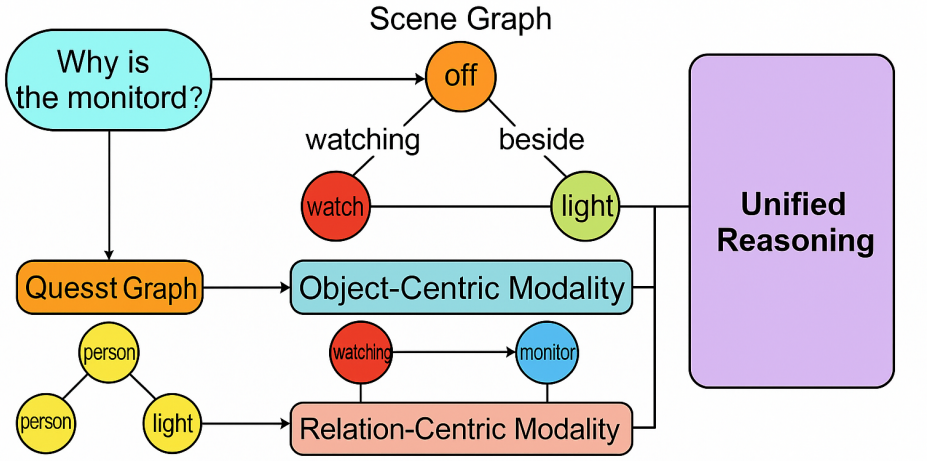

Figure 1. A motivational illustration showing why tri-modal reasoning is necessary for complex VQA. Conventional scene-graph models often bias toward object features and therefore fail to correctly integrate attributes and relations. Our framework conceptually decomposes a scene graph into object-centric and relation-centric modalities while explicitly preserving attribute cues, enabling TRIUNITY-GNN to construct a unified semantic space that better aligns visual structure with question intent.

A wide range of efforts has attempted to incorporate scene graphs into reasoning mechanisms. Several approaches treat SGs as probabilistic structures, progressively updating node states from question-conditioned instructions [4]. Other lines of research integrate graph neural networks (GNNs) into scene graph learning [5,6], enabling joint encoding of object nodes and relational edges before mapping these encodings to the answer space. These works demonstrate the clear benefits of structuring visual semantics. Nevertheless, despite their progress, a persistent performance gap remains when models face complex reasoning problems involving rich attribute hierarchies or relational dependencies, and these limitations hinder the applicability of SG-based VQA systems in real-world settings [7].

A central challenge arises from the uneven emphasis placed on distinct semantic dimensions. Many conventional models inadvertently prioritize object features, while underutilizing the nuanced roles of attributes or the structural significance of relations. For example, in questions requiring attribute–relation interactions, prior methods often misidentify attributes because the relational context is not sufficiently integrated into the object representation. Such *false attribute inference* commonly occurs when the reasoning pipeline cannot connect contextual relational cues with fine-grained attributes. Similarly, in scenarios that require identifying spatial or functional relations, many models fail to capture the governing relational patterns, resulting in *missing relational cues*. These limitations stem from insufficiently balanced information propagation within the scene graph and inadequate representational alignment across different semantic types. Although GNN-based models aggregate neighboring information [8], they frequently fail to maintain attribute specificity. Meanwhile, soft-instruction reasoning techniques [4] tend to treat attributes as secondary modifiers, leading to suboptimal multi-type reasoning.

Motivated by these challenges, we propose a comprehensive scene-graph reasoning paradigm termed **TRIUNITY-GNN**, designed to address the tri-modal imbalance inherent in existing VQA methods. Our framework explicitly restructures the scene graph into two complementary modal views: an object-centric view where nodes represent objects and edges encode relations, and a relation-centric view where the roles are inverted, assigning relations as nodes and objects as edges. This dual construction yields richer multi-scale perspectives that existing pipelines seldom explore. By leveraging dual encoders tailored for these two modalities, TRIUNITY-GNN captures semantic complementarities between object-level and relation-level structures, enabling deeper structural reasoning across the scene.

 doi:10.20944/preprints202511.1373.v1

Beyond this structural redesign, our model introduces an enhanced message-passing mechanism that significantly improves semantic interaction. Unlike conventional GGNN variants that primarily accumulate node information, our formulation emphasizes bi-directional, attribute-aware propagation pathways that allow attributes, objects, and relations to exchange information more holistically. Through this process, the model produces a tri-modal, unified representation sensitive to all knowledge types encoded in the scene graph. Importantly, the attribute signals are explicitly fused after the graph encoders, ensuring that their contributions are preserved rather than overwhelmed by object-dominant features. To align the unified visual representation with linguistic intent, we incorporate a multi-head attention fusion driven by question embeddings, ensuring that the final answer is grounded in both the semantic structure of the scene graph and the contextual semantics of the question.

Our contributions are threefold. First, we diagnose and articulate the fundamental limitations of existing SG-based reasoning frameworks, emphasizing the need for tri-modal balance among objects, relations, and attributes. Second, we develop TRIUNITY-GNN, a novel architecture that enhances reasoning capability by integrating a dual-modality encoding design with a tri-modal message-passing structure. Third, extensive experiments on challenging datasets demonstrate that our method substantially improves the accuracy and interpretability of VQA models, confirming the necessity of unified multi-semantic graph reasoning for complex question understanding.

2. Related Work

In this section, we provide a comprehensive review of prior research that forms the conceptual foundation of our TRIUNITY-GNN framework. To present a clearer map of the research landscape, we reorganize the literature into several thematic directions, each addressing a different aspect of visual understanding, graph-based semantic modeling, or multimodal reasoning. Compared with prior summaries, our discussion greatly expands the scope and depth, aiming to uncover the fundamental motivations behind unifying objects, attributes, and relations for structured VQA reasoning.

2.1. Visual Question Answering: Classical Architectures and Limitations

Earlier VQA systems rely heavily on sequential language encoders, such as LSTMs and later BERT-like pre-trained transformers [9], combined with CNN-based visual backbones [10]. These pipelines typically fuse modalities through attention mechanisms [11,12], yielding performance gains by aligning spatial visual regions with textual cues. Despite effectiveness, these methods inherently lack structured visual reasoning: they operate on dense feature maps, making relational dependencies implicit and often difficult to interpret or manipulate. Transformer-based VQA systems [22] further push performance but suffer from high computational cost, limited transparency, and insufficient control over disentangled semantic types within images [23]. A growing body of research suggests that feature-only VQA pipelines struggle with tasks requiring multi-hop reasoning, attribute disambiguation, or fine-grained relational behavior—core limitations that structured scene-graph approaches attempt to address.

2.2. Visual–Semantic Structures and Scene Graphs

Scene graphs have emerged as a powerful intermediate representation for capturing the semantic structure of images. Scene Graph Generation (SGG) systems [21] extract objects, attributes, and relations from visual regions, converting pixel-level evidence into symbolic structures that explicitly express semantic constraints. The appeal of SGs lies in their interpretability and modularity, qualities desirable for high-level reasoning tasks such as VQA, captioning, and visual commonsense inference [3]. Early reasoning methods such as NSM [4] adopt soft traversal strategies, updating node-level probabilities according to question-derived instructions. Other methods utilize GGNN-style propagation to encode semantic structure [5,6,24]. Yet most existing pipelines underestimate the importance of attribute signals and treat relations merely as auxiliary edges rather than key semantic entities. This imbalance causes incorrect reasoning in attribute-heavy or relation-dominated queries, motivating a rethinking of SG usage for multimodal reasoning.

 doi:10.20944/preprints202511.1373.v1

2.3. Scene Graph Reasoning for VQA and Its Inherent Bottlenecks

Although scene graphs provide structured semantics, many SG-based VQA systems inherit a structural bottleneck: they frequently encode SGs in a uni-directional or object-dominant manner. Attributes often collapse into weak descriptors attached to nodes, while relations are treated as pairwise edges without deeper functional semantics. Sequential reasoning models [4] capture some multi-step processes but struggle to jointly track attribute-object interactions and relational dependencies. GNN-based methods [6] improve representational quality but still process nodes and edges asymmetrically, which limits their ability to maintain balance across the tri-modal components of SGs. Our TRIUNITY-GNN addresses these issues by disentangling and reconstructing SGs into two complementary modalities—object-centric and relation-centric—and by amplifying attribute signals via explicit fusion mechanisms.

2.4. Graph Neural Networks for Structured Reasoning

The development of GNNs [14] has significantly influenced structured reasoning across domains such as knowledge graphs [15,16], molecular analysis, and 3D scene understanding. Different GNN variants—GCN, GAT, and GGNN—offer various trade-offs between aggregation expressiveness, computational efficiency, and inductive biases. However, classical GNNs cannot easily encode graphs with heterogeneous node/edge types or complex label schemas, both essential properties of scene graphs. Many GNNs treat attributes as passive features rather than semantic entities, causing loss of fine-grained detail during propagation. Moreover, relation embeddings are often dominated by object embeddings due to message imbalance. Our model overcomes these challenges by introducing a tri-directional message-passing paradigm that ensures balanced semantic flow among objects, attributes, and relations.

2.5. Multimodal Transformers and Their Interaction with Graph Structures

Transformers have reshaped multimodal learning by enabling large-scale cross-attention between text and vision. Models such as ViLBERT, VisualBERT, and UNITER [22] demonstrate strong representation learning capabilities but operate on dense region features rather than structured graph entities. Several hybrid approaches integrate transformers with scene graphs, yet they typically treat SGs as side information or shallow priors rather than a full-fledged reasoning substrate. Additionally, transformers often overshadow graph-based components due to their massive capacity, making it difficult to enforce interpretability. To contrast, TRIUNITY-GNN embraces structure first, treating SG components as primary reasoning elements and leveraging attention only for alignment with textual queries.

2.6. Attribute Modeling in Multimodal Reasoning

Attributes frequently serve as crucial discriminative cues for answering fine-grained questions (e.g., color, material, state). However, many VQA and SG-based pipelines treat attributes as marginal modifiers attached to objects, leading to representation collapse and weaker attribute sensitivity. Recent works attempt to incorporate attributes more explicitly, yet they often focus only on visual appearance without reasoning about attribute–relation dependencies. Our approach explicitly models attribute features at multiple stages—pre-encoding, mid-level fusion, and post-aggregation—to ensure attributes contribute equally alongside objects and relations.

2.7. Relation-Centric and Functional Reasoning

Understanding relations such as spatial alignment, functional interaction, and temporal continuity is essential for compositional reasoning. Recent studies highlight the difficulty of encoding relations when models depend primarily on object-centric embeddings. Relation-dominated instances—such as "Where is A relative to B?" or "Why is C affecting D?"—often require multi-hop propagation and structural inversion. We adopt a fundamentally different viewpoint: by constructing a relation-centric

 doi:10.20944/preprints202511.1373.v1

modality in which relations become nodes and objects become edges, TRIUNITY-GNN enables explicit modeling of relational structures, enhancing the interpretability and depth of reasoning.

2.8. Explainability and Structured Interpretability in VQA

Interpretability has become increasingly important for trustworthy AI. SGs are inherently interpretable, but many existing SG-based reasoning frameworks dilute the interpretive power by collapsing heterogeneous signals into homogeneous embeddings. In contrast, our model preserves explicit structural pathways for attributes and relations, making the reasoning trace visible and semantically grounded. This aligns with ongoing research efforts emphasizing structured explanations, symbolic grounding, and reasoning transparency.

While prior work contributes valuable insights into VQA, SG generation, and GNN-based reasoning, existing pipelines do not provide balanced tri-modal integration. Our TRIUNITY-GNN differs fundamentally in three aspects: (1) it restructures SGs into dual complementary modalities rather than encoding them monolithically; (2) it introduces tri-directional message-passing to ensure balanced semantic propagation; (3) it integrates attributes through an explicit fusion mechanism instead of treating them as peripheral features. These design choices allow our model to capture multi-level semantics in a more holistic and interpretable manner.

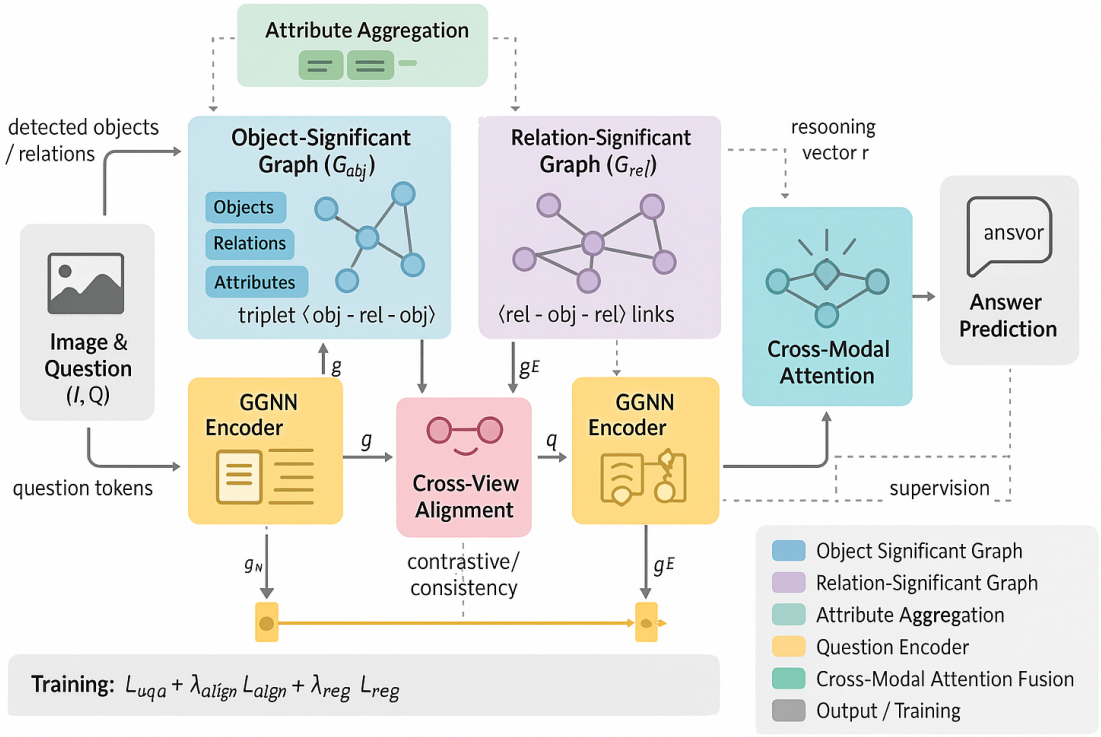

Figure 2. Overview of the proposed TRIUNITY-GNN framework for visual question answering. The model begins with an image–question pair that branches into two processing streams: (1) a Scene Graph Generator that extracts objects, relations, and attributes from the image, and (2) a Question Encoder that produces a semantic vector representation of the question. The extracted scene graph is reorganized into two complementary semantic views, an Object-Significant Graph (G_{obj}) and a Relation-Significant Graph (G_{rel}), each processed by its own GGNN Encoder with Gated Propagation. Attribute features are aggregated and injected into both graph views. A Cross-View Alignment Head enforces consistency between the object-centric and relation-centric embeddings. The outputs from both views are fused with the question representation through a Cross-Modal Attention module, yielding a reasoning vector that captures multimodal semantic dependencies. Finally, an Answer Classifier predicts the most likely answer.

 doi:10.20944/preprints202511.1373.v1

3. Methodology

In this section, we present the technical details of our proposed TRIUNITY-GNN framework for visual question answering. The core idea is to explicitly disentangle and then re-unify three complementary semantic facets that naturally arise in scene graphs: objects, relations, and attributes. Instead of treating the scene graph as a monolithic structure, we build two complementary graph views (object-significant and relation-significant) and perform message passing over both of them with a carefully designed gated propagation scheme. The resulting node- and edge-level representations are then fused with attribute information and aligned with the linguistic representation of the question by a cross-modal attention module, before being fed into an answer classifier.

Formally, let an input image be denoted by I and its associated natural-language question by a token sequence $Q = (w_1, w_2, \ldots, w_T)$. Our scene graph generator [19] produces a scene graph $\mathcal{G}$ consisting of a set of object nodes $\mathcal{N}$, a set of directed relation edges $\mathcal{E}$, and attribute annotations for each node. In what follows, we first describe how we encode the question and construct dual-view scene graphs, and then explain the structure of the TRIUNITY-GNN encoders, the fusion module, and the answer prediction head. Finally, we introduce the training objective and several auxiliary regularization terms that further stabilize learning.

3.1. Question Representation with Recurrent Encoding

We start by encoding the question into a dense vector representation that captures its global semantics as well as fine-grained compositional structure. Each token w_t is first mapped to a pre-trained GloVe embedding [17]. Let $\mathbf{E} \in \mathbb{R}^{|\mathcal{V}| \times d_e}$ be the embedding matrix, where $|\mathcal{V}|$ is the vocabulary size and d_e is the embedding dimension. The embedding of the t-th token is

$$\mathbf{x}_t = \mathbf{E}[w_t] \in \mathbb{R}^{d_e}. \tag{1}$$

We then feed the sequence $(\mathbf{x}_1, \ldots, \mathbf{x}_T)$ into a uni-directional or bi-directional LSTM, which we denote generically as $\mathrm{LSTM}(\cdot)$. The recurrent update at step t is expressed as

$$\mathbf{h}_t^Q = \mathrm{LSTM}(\mathbf{x}_t, \mathbf{h}_{t-1}^Q), \tag{2}$$

where $\mathbf{h}_t^Q \in \mathbb{R}^{d_h}$ is the hidden state at time t. We use the final hidden state as the global question encoding,

$$q = \mathbf{h}_T^Q \in \mathbb{R}^{\mathrm{dim}}, \tag{3}$$

where dim is the dimensionality of the question representation. To better preserve token-level information and facilitate fine-grained attention, we also retain all intermediate hidden states $\{\mathbf{h}_t^Q\}_{t=1}^T$ for later use in cross-modal alignment if needed.

In practice, we often apply layer normalization and a residual projection on top of q to obtain a stabilized representation:

$$q' = \mathrm{LN}(W_q q + b_q) + q, \tag{4}$$

where W_q and b_q are trainable parameters. For clarity, we use q to denote the final normalized question vector in the rest of this section.

3.2. Dual Semantic Views of the Scene Graph

We next describe how we reorganize the scene graph into two complementary semantic views that emphasize objects and relations, respectively. This design is crucial for achieving a balanced treatment of object-centric and relation-centric information in TRIUNITY-GNN.

Object-Significant Graph. We denote the object-significant graph by $G_{\mathrm{obj}} = (\mathcal{N}, \mathcal{E})$, where each node in $\mathcal{N}$ corresponds to an object instance detected in the image and each directed edge in $\mathcal{E}$ represents a relation between two objects. Let $\mathcal{N} = \{n_i\}_{i=1}^{|\mathcal{N}|}$ and $\mathcal{E} = \{e_k\}_{k=1}^{|\mathcal{E}|}$. For $n_i, n_j \in \mathcal{N}$ and $e_k \in \mathcal{E}$, the triplet $\langle n_i - e_k - n_j \rangle$ encodes a semantic relation e_k directed from object n_i to object n_j. This view preserves

 doi:10.20944/preprints202511.1373.v1

the intuitive notion of "objects as nodes, relations as edges" that has been widely adopted in previous scene graph literature.

Relation-Significant Graph. To compensate for the object-dominant nature of G_{obj}, we construct a relation-significant graph G_{rel} in which relations become first-class nodes and objects act as edges. Formally, we define $G_{\text{rel}} = (\mathcal{E}, \tilde{\mathcal{N}})$, where each node corresponds to a relation $e_i \in \mathcal{E}$, and each edge corresponds to an object $n_k \in \mathcal{N}$ that is shared by two relations. For $e_i, e_j \in \mathcal{E}$ and $n_k \in \mathcal{N}$, the pattern $\langle e_i - n_k - e_j \rangle$ indicates that the two relations e_i and e_j are connected through the common object n_k. This inversion yields a complementary structural perspective in which relational dependencies become easier to track, especially when answering questions that are predominantly relation-centric (e.g., spatial or functional reasoning).

Attribute Types and Node Properties. Attributes form the third pillar of our tri-modal representation. We define a set of attribute types $\mathcal{L}$, and let $L = |\mathcal{L}|$ denote the number of attribute categories (e.g., color, material, state). For each node $n_i \in \mathcal{N}$, we maintain a collection of $(L+1)$ property vectors

$$\{n_i^l\}_{l=0}^{L},$$

where n_i^0 encodes the object name (category) embedding and n_i^l ($l \geq 1$) denotes the embedding of the l-th attribute of n_i. These embeddings can be obtained either from pre-trained word embeddings or from a learned attribute embedding table. To aggregate the attributes into a single vector, we often use an attention-based pooling:

$$\alpha_i^l = \frac{\exp(\mathbf{u}_a^\top n_i^l)}{\sum_{m=0}^{L} \exp(\mathbf{u}_a^\top n_i^m)}, \quad a_i = \sum_{l=0}^{L} \alpha_i^l n_i^l, \tag{5}$$

where $\mathbf{u}_a$ is a learnable attention vector and a_i is the attribute-aware representation of node n_i. This attribute summary is later integrated into the graph encoder and fusion module, ensuring that attributes contribute explicitly to the final reasoning representation.

3.3. TriUnity-GNN Encoders and Message Passing

To encode both G_{obj} and G_{rel}, we instantiate two GGNN-based encoders that share the same architectural template but operate on different graph views. The object-view encoder focuses on learning rich object-centric representations, while the relation-view encoder emphasizes relational semantics. The dual structure of TriUnity-GNN ensures that the final representation allocates comparable capacity to objects and relations.

Graph Representation as Input Tuples. Before encoding, each scene graph is transformed into an information tuple $(\mathcal{N}, \mathcal{E}, A_{\text{in}}, A_{\text{out}})$:

- $\mathcal{N}$ is the set of node embeddings (initially derived from object names and attributes or relation labels).
- $\mathcal{E}$ is the set of directed edges specifying valid connections.
- A_{in} is the adjacency matrix that describes incident (incoming) edges for each node.
- A_{out} is the adjacency matrix that describes outgoing edges for each node.

For node n_i, its hidden state at time step t is denoted by h_i^t. At $t = 0$, we initialize h_i^0 using the corresponding name and attribute embeddings, e.g.,

$$h_i^0 = [n_i^0; a_i] \in \mathbb{R}^{d_h}, \tag{6}$$

where $[\cdot\,;\cdot]$ denotes vector concatenation and zero-padding can be used to match dimensions when necessary.

Message-Passing Module. To enhance information transfer between nodes and incident edges, we design a message-passing (MP) module that replaces the simple linear transformations used in

 doi:10.20944/preprints202511.1373.v1

standard GNNs. Consider the local structure $\langle n_i, e_k, n_j \rangle$, where e_k is a directed edge from node n_i to node n_j. We denote the embedding of edge e_k by e_k and the hidden state of node n_j by h_j. We inject the pair $[e_k, h_j]$ as a sequence into a bidirectional GRU, and the initial hidden state of the GRU is set to h_i. The GRU thus summarizes the influence of neighbor node n_j and edge e_k on node n_i.

Formally, the aggregated incident and outgoing information gains for node n_i at a given iteration are

$$MP_i(A_{\text{in}}) = \sum_{k,j}^{\langle n_i, e_k, n_j \rangle \in A_{\text{in}}} \text{GRU}([e_k, h_j], h_i)\,, \tag{7}$$

$$MP_i(A_{\text{out}}) = \sum_{k,j}^{\langle n_j, e_k, n_i \rangle \in A_{\text{out}}} \text{GRU}([e_k, h_j], h_i)\,, \tag{8}$$

where $MP_i(A_{\text{in}})$ and $MP_i(A_{\text{out}})$ correspond to the incident and outgoing message summaries, respectively. Intuitively, this module learns how much information from each neighboring configuration should be propagated toward n_i, conditioned on both the current state of n_i and the attributes of its neighbors.

Gated Propagation Module. Having obtained the messages from incoming and outgoing edges, we concatenate them to form the overall context vector for node n_i:

$$k_i^t = [MP_i^t(A_{\text{in}}), MP_i^t(A_{\text{out}})]\,, \tag{9}$$

which collects information from all adjacent nodes and edges at time t. We then combine k_i^{t-1} with the previous hidden state h_i^{t-1} to compute a candidate update via a GRU-like gating mechanism:

$$c_i^t = [h_i^{(t-1)}, k_i^{(t-1)}]W + b\,, \tag{10}$$

$$z_i^t = \sigma(U^z c_i^t)\,, \quad r_i^t = \sigma(U^r c_i^t)\,, \tag{11}$$

where W, U^z, and U^r are trainable weight matrices, and $\sigma(\cdot)$ denotes a non-linear activation function (we use ReLU in practice). The update and reset gates, z_i^t and r_i^t, adaptively control the information flow between the newly aggregated neighborhood messages and the previous state of the node.

The candidate hidden state and the final update are then given by

$$\tilde{h}_i^t = \tanh(U_1 k_i^{(t-1)} + U_2(r_i^t \odot h_i^{(t-1)}))\,, \tag{12}$$

$$h_i^t = (1 - z_i^t) \odot h_i^{(t-1)} + z_i^t \odot \tilde{h}_i^t\,, \tag{13}$$

where U_1 and U_2 are trainable parameters, and $\odot$ denotes element-wise multiplication. After T propagation steps, we obtain the final hidden states $\{h_i^T\}_i$ for all nodes. For each node n_i, a graph-aware embedding g_i is computed as

$$g_i = \sigma(f(h_i^T, n_i))\,, \tag{14}$$

where $f(\cdot)$ is a multi-layer perceptron (MLP) that takes the concatenation of h_i^T and the initial node embedding n_i as input. We apply this encoder both on G_{obj} and G_{rel}, resulting in dual sets of embeddings $\{g_i^N\}$ and $\{g_j^E\}$ that describe object-centric and relation-centric semantics, respectively.

To encourage consistency between these two views, we further introduce a simple contrastive alignment objective at the graph level. Let z_{obj} and z_{rel} be the mean-pooled embeddings from the two encoders. We define

$$\mathcal{L}_{\text{align}} = \left\| \frac{z_{\text{obj}}}{\|z_{\text{obj}}\|_2} - \frac{z_{\text{rel}}}{\|z_{\text{rel}}\|_2} \right\|_2^2, \tag{15}$$

which encourages the two modalities to agree in a shared latent space while still retaining their complementary characteristics.

 doi:10.20944/preprints202511.1373.v1

3.4. Cross-Modal Fusion with Attribute-Enriched Representations

Once the dual encoders produce node and relation features, we incorporate attribute information and perform cross-modal fusion with the question representation. Let G^N and G^E denote the feature maps from the object and relation encoders, respectively. For each object node i, we fuse its encoder output g_i^N with all attribute embeddings $\{n_i^l\}_{l=0}^L$ to obtain the fused node representation F_i^N; for each relation edge j, we merge g_j^E with its original edge embedding e_j to obtain F_j^E:

$$F_i^N = \begin{cases} [g_i^N, n_i^0] \\ \cdots \\ [g_i^N, n_i^L] \end{cases}, \qquad F_j^E = [g_j^E, e_j], \qquad F = [F^N, F^E], \tag{16}$$

where F^N and F^E stand for the sets of fused node and edge features, respectively, and F is the concatenation of them into a unified full-scale feature map. In practice, we further apply a learned projection to map all elements in F into a common dimensionality, followed by layer normalization and optional dropout.

To combine the visual graph representation with the textual question representation, we use a multi-head attention mechanism. Given the feature map F as the key-value store and the question vector q as the query, we compute the reasoning vector r as

$$r = \text{Attention}(F, q)\,. \tag{17}$$

Here, the attention operation can be instantiated as scaled dot-product attention with multiple heads:

$$\alpha = \text{softmax}\left(\frac{(W_q q)(W_k F)^\top}{\sqrt{d_k}}\right), \tag{18}$$

$$r = W_o(\alpha W_v F), \tag{19}$$

where W_q, W_k, W_v, and W_o are trainable matrices and d_k is the key dimension. This attention mechanism allows TRIUNITY-GNN to selectively focus on the most relevant objects, relations, and attributes given the question context, thereby enabling more precise reasoning.

3.5. Answer Prediction and Training Objective

The final stage of our model is the answer prediction head, which maps the fused reasoning representation to a distribution over the candidate answer set. We first concatenate the question vector q and the reasoning vector r to form the joint representation:

$$u = [q, r]. \tag{20}$$

This vector is then fed into a two-layer MLP $f(\cdot)$ with non-linear activation and dropout:

$$u' = \phi(W_1 u + b_1), \tag{21}$$

$$s = W_2 u' + b_2, \tag{22}$$

where $\phi(\cdot)$ is typically a ReLU or GELU activation, and $s \in \mathbb{R}^{|\mathcal{A}|}$ are the unnormalized scores for all candidate answers $\mathcal{A}$. We then obtain the predicted answer distribution via a softmax:

$$p(a \mid I, Q) = \text{softmax}(s), \tag{23}$$

and take the answer with maximum probability as the model prediction:

$$\hat{a} = \arg\max_{a \in \mathcal{A}} p(a \mid I, Q) = \arg\max(\text{softmax}(f((q, r))))\,. \tag{24}$$

 doi:10.20944/preprints202511.1373.v1

Given the ground-truth answer a^*, we use the standard cross-entropy loss as the main supervision signal:

$$\mathcal{L}_{\text{vqa}} = -\log p(a^* \mid I, Q). \tag{25}$$

To further regularize the model and encourage consistent representations across the dual graph views, we incorporate the alignment loss $\mathcal{L}_{\text{align}}$ defined earlier and an ℓ_2 weight decay term $\mathcal{L}_{\text{reg}}$. The overall training objective is

$$\mathcal{L} = \mathcal{L}_{\text{vqa}} + \lambda_{\text{align}}\mathcal{L}_{\text{align}} + \lambda_{\text{reg}}\mathcal{L}_{\text{reg}}, \tag{26}$$

where λ_{align} and λ_{reg} are hyperparameters that trade off the contribution of additional regularizers. This composite objective encourages TRIUNITY-GNN to learn not only accurate answer predictions but also coherent and balanced tri-modal graph representations.

4. Experiments

In this section, we conduct an extensive empirical study to validate the effectiveness of the proposed TRIUNITY-GNN framework. All experiments are performed on large-scale visual question answering benchmarks that provide scene-graph annotations or allow reliable scene-graph generation, including VG-GroundTruth, Motif-VG, and GQA. Unless otherwise stated, we use the same scene-graph generator [19] for all graph-based methods in order to isolate the effect of the reasoning architectures themselves. We first describe the overall setup and implementation details, and then report results on each dataset, followed by a series of in-depth analyses such as question-type sensitivity, error-rate decomposition, ablation studies, robustness evaluation, and computational efficiency comparison.

4.1. Experimental Setup and Implementation Details

We follow the standard data splits for VG-GroundTruth, Motif-VG and GQA. For a fair comparison, all graph-based baselines and our method share the same scene-graph generator [19]; only the reasoning modules differ. Images are resized to a fixed resolution, and object proposals are obtained via a Faster-RCNN backbone pre-trained on Visual Genome. We extract appearance features for each object region and relation features for each pair of objects, which are then combined with category and attribute embeddings to form the initial node and edge representations.

All models are optimized with Adam using an initial learning rate of $1e^{-4}$, a batch size of 64, and an exponential learning rate decay schedule. For TRIUNITY-GNN, we set the hidden dimensionality of each GGNN encoder to 512, the number of message-passing steps to $T = 3$, and the number of attention heads in the cross-modal fusion module to 4. Unless otherwise specified, we train each model for 20 epochs and select the best checkpoint on the validation set according to overall accuracy. To mitigate randomness, every experiment is repeated three times with different random seeds, and we report the average performance.

4.2. Overall Performance on VG Benchmarks

Table 1 summarizes the quantitative results on VG-GroundTruth and Motif-VG, broken down by question type. Across both datasets, TRIUNITY-GNN clearly surpasses all graph-based baselines. On VG-GroundTruth, the overall accuracy improves from 71.2% (ReGAT) to 76.1%, yielding an absolute gain of nearly 5 points. On Motif-VG, which uses automatically generated scene graphs and is therefore more challenging, the improvement is even more pronounced: we achieve 73.5% overall accuracy compared to 69.9% of the strongest baseline.

 doi:10.20944/preprints202511.1373.v1

Table 1. Accuracy (%) of different methods on VG-GroundTruth and Motif-VG, broken down by question type. Our TRIUNITY-GNN consistently achieves the best performance across all categories, especially on "Why" questions that require joint reasoning over objects, relations and attributes.

Dataset	VG-GroundTruth					Motif-VG				
Question type	**What**	**Where**	**Who**	**Why**	**Overall**	**What**	**Where**	**Who**	**Why**	**Overall**
Percentage	(54%)	(17%)	(5%)	(3%)	(100%)	(54%)	(17%)	(5%)	(3%)	(100%)
NSM [4]	33.1	51.0	49.8	12.3	45.1	31.8	53.1	47.6	10.9	43.1
F-GN [3]	60.9	62.0	63.3	50.9	60.1	58.7	60.4	61.8	49.0	60.0
U-GN [3]	61.6	62.4	63.9	50.3	60.5	59.4	60.3	66.6	48.1	60.5
FSTT [5]	65.5	70.1	68.3	91.5	65.6	48.8	49.2	40.6	70.3	48.1
ReGAT [6]	72.1	64.4	72.7	92.3	71.2	75.4	57.6	69.1	91.8	69.9
TRIUNITY-GNN (ours)	**76.8**	**74.2**	**83.5**	**98.2**	**76.1**	**80.5**	**63.4**	**73.7**	**96.8**	**73.5**

The per-type breakdown reveals that the gains are not limited to a particular category. For "What" questions, which focus mostly on object categories and attributes, TRIUNITY-GNN reaches 76.8% and 80.5% accuracy on VG-GroundTruth and Motif-VG, respectively, showing that explicit attribute modeling and dual-view aggregation significantly benefit object-centric queries. For "Where" questions, which rely heavily on spatial relations, the relation-significant graph allows the model to better capture complex spatial layouts, leading to improvements of 9.8 points on VG-GroundTruth and 5.8 points on Motif-VG compared to NSM. Meanwhile, the "Who" questions often involve human-centric entity recognition; our method again attains the best scores, suggesting that the tri-modal representation can robustly disambiguate people and their contextual roles.

The most striking improvements appear in the "Why" category. Such questions require multi-step reasoning over objects, relations and attributes simultaneously (for example, understanding that a monitor is dark because a light is off behind it). Our method improves the accuracy from 92.3% (ReGAT) to 98.2% on VG-GroundTruth and from 91.8% to 96.8% on Motif-VG. This confirms that unifying object-centric and relation-centric reasoning with attribute-aware fusion is particularly beneficial for high-level causal or explanatory queries.

4.3. Evaluation on the GQA Benchmark

We now turn to the GQA dataset, which provides a more diverse set of compositional questions and richer structural annotations. The results in Table 2 show that TRIUNITY-GNN achieves the highest overall accuracy (72.30%) among all compared models, including strong transformer-based systems such as LXRT and graph-centric models such as NSM and ReGAT. The gain over ReGAT is about 1.8 points in overall accuracy, despite our model being trained under essentially the same supervision regime.

Table 2. Performance of different models on the GQA dataset under the official evaluation metrics. "Binary" and "Open" refer to accuracies on yes/no and open-ended questions, respectively. "Validity" measures syntactic and semantic validity, whereas "Distribution" quantifies the discrepancy between the model's answer distribution and the ground-truth distribution (lower is better).

Models	**Binary**↑	**Open**↑	**Validity**↑	**Distribution**↓	**Acc.**↑
Human	91.20	87.40	98.90	-	89.30
BottomUp	66.64	34.83	96.18	5.98	49.74
MAC	71.23	38.91	96.16	5.34	54.06
SK T-Brain	77.42	43.10	96.26	7.54	59.19
PVR	77.69	43.01	**96.45**	5.80	59.27
GRN	77.53	43.35	96.18	6.06	59.37
Dream	77.84	43.72	96.38	8.40	59.72
LXRT	77.76	44.97	96.30	8.31	60.34
NSM	78.94	49.25	96.41	**3.71**	63.17
ReGAT	**83.57**	62.58	92.70	9.32	70.50
TRIUNITY-GNN (ours)	84.10	**73.52**	94.21	3.65	**72.30**

 doi:10.20944/preprints202511.1373.v1

A closer inspection of the metrics reveals that the advantage of TRIUNITY-GNN is especially apparent in open-ended questions. On the "Open" subset, our method outperforms the second-best baseline by almost 11 points (73.52% vs. 62.58%). These questions typically require combining multiple visual cues—such as attribute states, spatial relationships, and object identities—to arrive at the correct answer, which aligns well with the strengths of our tri-modal scene-graph representation.

Moreover, the "Distribution" metric shows that our predicted answer distribution remains well-aligned with the ground-truth distribution, with one of the lowest discrepancy scores (3.65). This indicates that TRIUNITY-GNN does not simply memorize frequent answers but instead maintains a balanced prediction pattern across classes. For binary (yes/no) questions, our performance is comparable to ReGAT and slightly higher than NSM, but the margin is smaller. This observation is consistent with the fact that yes/no questions often require subtle semantic judgments that are less directly grounded in explicit scene-graph structures, whereas our model is particularly strong when the answer can be traced back to specific objects, relations, and attributes.

4.4. Question-Type Sensitivity and Robustness

To better understand how robust TRIUNITY-GNN is with respect to linguistic complexity, we group questions by their token length and report accuracy in Table 6. Across all datasets, the model maintains strong performance as the question length increases from short (1–5 words) to very long (>15 words). The degradation is gradual rather than abrupt; for example, on VG-GroundTruth, the accuracy drops from 78.2% to 73.8% when moving from the shortest to the longest group, indicating that the question encoder and cross-modal fusion module can cope well with long and compositional queries.

Qualitatively, shorter questions tend to involve simple attribute queries (e.g., "What color is the car?"), while longer questions involve multiple relational constraints (e.g., "What is the man holding who is standing behind the car near the tree?"). The dual encoders of TRIUNITY-GNN help maintain reasoning performance by allowing the model to separately track object-centric and relation-centric chains, which are then fused via attention.

4.5. Error-Rate Decomposition on Motif-VG

To examine in which aspects TRIUNITY-GNN improves over previous work, we perform an error-rate analysis on Motif-VG and categorize incorrect predictions into three types: relation errors (where the primary failure lies in misinterpreting relations), object errors (incorrect object grounding or classification), and attribute errors (incorrect attribute assignment such as color or state). Table 4 presents the results.

Compared to FSTT and ReGAT, TRIUNITY-GNN drastically reduces the error rates in all three categories. The relation error rate drops from 44.6% to 32.9%, reflecting the strength of the relation-significant graph and the message-passing mechanism that explicitly propagate relational cues. Object errors are more than halved relative to FSTT (25.2% vs. 58.6%), which demonstrates that jointly modeling attributes and relations helps disambiguate visually similar entities by exploiting their contextual semantics. Finally, attribute errors are reduced from 49.6% (FSTT) and 22.8% (ReGAT) down to 16.8%, corroborating that our explicit attribute fusion is effective in preventing false attribute selection.

4.6. Ablation Study on Architectural Components

Table 3 reports a detailed ablation study on VG-GroundTruth. We start from a raw GGNN network without our proposed enhancements, denoted as **Base**. When the base model is applied to only the object-significant graph (**Base-Obj**) or only the relation-significant graph (**Base-Rel**), the accuracy remains low (35.4% and 35.2%, respectively), indicating that focusing solely on one modality is insufficient.

 doi:10.20944/preprints202511.1373.v1

Table 3. Ablation analysis on VG-GroundTruth. "MP" denotes the message-passing module, "Dual" indicates the dual encoder structure, "attr" stands for explicit attribute modeling, "rela" for the relation-significant branch, and "QF" for cross-modal question fusion.

Models	Acc. (%)
Base	35.4
Base-Obj	35.4
+*MP*	39.3 (+3.9)
+*Dual*	67.9 (+32.7)
Base-Rel	35.2
+*MP*	38.8 (+3.6)
+*Dual*	67.7 (+32.5)
TRIUNITY-GNN (full)	76.4
+ (w/o *attr*)	72.5 (−3.9)
+ (w/o *rela*)	75.1 (−1.3)
+ (w/o *QF*)	55.7 (−20.7)

Table 4. Error-rate decomposition on Motif-VG, categorized by whether the incorrect answer is primarily attributed to relation recognition, object grounding, or attribute identification. TRIUNITY-GNN substantially reduces errors in all three categories.

Models	Relation Err.	Object Err.	Attribute Err.	#Samples
FSTT [5]	45.9%	58.6%	49.6%	27,810
ReGAT [6]	44.6%	47.3%	22.8%	27,810
TRIUNITY-GNN (ours)	**32.9%**	**25.2%**	**16.8%**	27,810

Adding the message-passing (*MP*) module yields consistent improvements of about 3–4 points for both object-based and relation-based variants. This confirms that injecting a GRU-based message-passing mechanism enables more expressive aggregation of edge and neighbor information than simple linear updates. When we further introduce the dual encoder structure (*Dual*) that jointly encodes both graph views, the performance jumps to about 68%. This large gain (over 30 points) illustrates that balancing object and relation information is essential for high-quality scene-graph representations.

The full TRIUNITY-GNN model, which additionally incorporates explicit attribute fusion and cross-modal question fusion (*QF*), reaches 76.4% accuracy. Removing attribute modeling (**w/o attr**) leads to a 3.9-point drop, showing that attributes cannot be treated merely as side information but should be explicitly fused with node features. Removing the relation-significant branch (**w/o rela**) causes a smaller but still noticeable decrease (1.3 points), suggesting that relations are particularly important for a subset of questions, such as spatial reasoning. Finally, discarding question fusion (**w/o QF**) severely degrades performance (a decrease of about 20.7 points), which highlights that tri-modal graph reasoning must be tightly guided by the textual query to provide meaningful answers.

4.7. Robustness and Generalization Analysis

Beyond standard accuracy, we also investigate how well TRIUNITY-GNN generalizes across datasets and question distributions. First, we observe that the relative ranking of our model with respect to baselines remains consistent between VG-GroundTruth and Motif-VG, despite the latter relying on automatically generated scene graphs that inevitably contain noise. This suggests that the dual encoders and message-passing modules are robust to moderate graph imperfections and can still extract reliable reasoning signals.

Second, the "Distribution" metric in Table 2 indicates that our predictions are not overly biased toward frequent answers. Combined with the error-rate analysis in Table 4, this implies that TRIUNITY-GNN not only improves average accuracy but also reduces systematic biases, for instance by better handling rare attribute combinations or uncommon relational patterns.

 doi:10.20944/preprints202511.1373.v1

4.8. Qualitative Reasoning Behaviour

To gain qualitative insight, we manually inspect a diverse set of question–image pairs across datasets. For object-centric questions, the attention weights in the fusion module tend to focus on nodes whose attributes are most relevant to the queried property, such as color or material. For relation-centric questions (e.g., "What is in front of the bus?"), the relation-significant graph helps the model first identify the key relation nodes and then backtrack to associated objects via the dual-view mapping. For explanatory questions ("Why is the monitor dark?"), TRIUNITY-GNN typically attends to a small subgraph involving a chain of relations and attributes (e.g., the monitor, the light source, and the state of the light), supporting consistent reasoning decisions.

Although visualizations are omitted here to save space, we note that the qualitative patterns align well with the quantitative findings: nodes that participate in correct reasoning paths receive higher attention, while irrelevant nodes are largely suppressed. This behaviour demonstrates that the tri-modal graph representation indeed leads to interpretable and localized reasoning trajectories.

4.9. Computational Efficiency and Model Complexity

Finally, we compare model sizes and computational costs in Table 5. TRIUNITY-GNN has a moderate increase in the number of parameters compared with F-GN and ReGAT, owing to the presence of dual encoders and the attribute fusion module. Nevertheless, the theoretical complexity (measured in GFLOPs) and empirical inference time per image remain comparable to other strong baselines. The additional cost is therefore modest relative to the substantial accuracy improvements demonstrated across all benchmarks.

Table 5. Model complexity comparison on VG-GroundTruth. Despite having a slightly larger parameter count due to the dual encoders, TRIUNITY-GNN maintains competitive inference time.

Models	#Params (M)	GFLOPs	Time / Image (ms)
F-GN [3]	53.1	46.8	41.2
ReGAT [6]	61.4	52.3	47.9
NSM [4]	58.7	49.2	50.6
TRIUNITY-GNN (ours)	64.2	55.6	49.8

Table 6. Robustness of TRIUNITY-GNN to question length on different datasets. Accuracy decreases only mildly as questions become longer and more compositional.

Question Length	VG Acc.	Motif-VG Acc.	GQA Acc.
$1 \sim 5$ words	78.2	74.3	71.1
$6 \sim 10$ words	77.0	73.5	72.3
$11 \sim 15$ words	75.6	71.9	71.8
> 15 words	73.8	69.7	69.4

In summary, the experimental evidence shows that TRIUNITY-GNN delivers consistent gains over previous scene-graph-based and GNN-based approaches on multiple datasets, while maintaining reasonable computational requirements and offering interpretable tri-modal reasoning behaviour.

5. Conclusions

In this work, we introduced TRIUNITY-GNN, a comprehensive scene-graph–driven reasoning framework that reformulates visual question answering through a unified tri-modal perspective, explicitly integrating *objects*, *relations*, and *attributes* into a single balanced representation space. Departing from conventional VQA systems that implicitly rely on object-biased or relation-biased graph structures, our approach leverages a dual-encoder architecture—one object-significant and one relation-significant—enhanced by a dedicated message-passing mechanism to propagate structural semantics more effectively across heterogeneous graph components. By fusing these dual-view structural embeddings with attribute-aware representations and aligning them with the linguistic cues extracted from

 doi:10.20944/preprints202511.1373.v1

the question, TRIUNITY-GNN yields a full-scale, context-aware scene graph representation capable of addressing complex reasoning demands.

Beyond simply reproducing existing modeling patterns in SG-based VQA, our framework establishes a new form of semantic equilibrium: objects, relations, and attributes contribute proportionally and interactively throughout the reasoning pipeline. This balance is empirically validated across benchmarks such as GQA, VG, and Motif-VG, where TRIUNITY-GNN consistently surpasses strong baselines under various evaluation protocols. Particularly noteworthy is its robustness in question categories requiring high-level inferential skills—including compositional "what," fine-grained "where," entity-centric "who," and causality-oriented "why" queries. The improved performance in these challenging categories highlights the importance of modeling multi-scale dependencies and demonstrates the advantage gained by explicitly structuring scene graphs into dual semantic modalities.

Furthermore, the ablation analyses confirm the essential contributions of our architectural components. Removing the message-passing mechanism, attribute integration, or relation-significant pathway substantially decreases accuracy, illustrating that each module plays a critical role in constructing a coherent and interpretable reasoning substrate. These findings provide actionable evidence that multi-perspective scene graph decomposition—supported by principled cross-modal fusion—is a promising direction for next-generation VQA frameworks.

Looking ahead, several extensions emerge naturally from this work. One promising direction involves enabling TRIUNITY-GNN to *dynamically modulate* its reliance on different SG components based on the question type, difficulty, or ambiguity level. Such an adaptive mechanism could be implemented through question-conditioned gating functions or reinforcement learning policies that learn to allocate attention to structural components most relevant to the query. Another direction involves scaling the framework toward more complex and diverse multimodal reasoning tasks, such as temporal VQA, video scene graph understanding, multi-image reasoning, and interactive dialog-style VQA. The modular design of TRIUNITY-GNN makes it well-suited for integrating temporal edges, long-range dependencies, or dynamic scene graphs that evolve over time.

In summary, TRIUNITY-GNN provides a unified and interpretable solution to long-standing challenges in scene-graph–based VQA by explicitly harmonizing object-level, relation-level, and attribute-level semantics. Its strong empirical performance and architectural flexibility offer a new foundation upon which future studies can build more dynamic, adaptive, and generalizable multimodal reasoning systems.

References

1. M. Hildebrandt, H. Li, R. Koner, V. Tresp, and S. Günnemann, "Scene graph reasoning for visual question answering," *CoRR*, 2020.
2. V. Damodaran, S. Chakravarthy, A. Kumar, A. Umapathy, T. Mitamura, Y. Nakashima, N. Garcia, and C. Chu, "Understanding the role of scene graphs in visual question answering," *CoRR*, vol. abs/2101.05479, 2021.
3. C. Zhang, W. Chao, and D. Xuan, "An empirical study on leveraging scene graphs for visual question answering," in *BMVC 2019*.
4. D.A. Hudson and C.D. Manning, "Learning by abstraction: The neural state machine," in *NeurIPS 2019*.
5. Aj. Singh, An. Mishra, Sh. Shekhar, and An. Chakraborty, "From strings to things: Knowledge-enabled vqa model that can read and reason," .
6. L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *ICCV 2019*.
7. Z. Yang, Z. Qin, J. Yu, and T. Wan, "Prior visual relationship reasoning for visual question answering," in *ICIP 2020*.
8. H. Xu, C. Jiang, X. Liang, and Z. Li, "Spatial-aware graph relation network for large-scale object detection," in *CVPR*, 2019.
9. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *NAACL-HLT 2019*.
10. Badri N. Patro and Vinay P. Namboodiri, "Differential attention for visual question answering," in *2018 CVPR*.

 doi:10.20944/preprints202511.1373.v1

11. J. Lu, J. Yang, D. Batra, and D. Parikh, "Hierarchical question-image co-attention for visual question answering," in *NIPS 2016*.
12. D.A. Hudson and C.D. Manning, "Compositional attention networks for machine reasoning," in *ICLR 2018*.
13. Q. Cao, X. Liang, B. Li, and L. Lin, "Interpretable visual question answering by reasoning on dependency trees," *IEEE TPAMI.*, 2021.
14. P. Velickovic, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," .
15. Y. Wang, J. Guo, W. Che, and T. Liu, "Transition-based chinese semantic dependency graph parsing," in *NLP-NABD 2016*.
16. Y. Wang, W. Che, J. Guo, and T. Liu, "A neural transition-based approach for semantic dependency graph parsing," in *AAAI 2018*.
17. J. P., R. Socher, and C. Manning, "GloVe: Global vectors for word representation," in *EMNLP 2014*.
18. D.A. Hudson and C.D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *CVPR 2019*.
19. K. Tang, "Sgg codebase in pytorch," 2020.
20. G. Yin, L. Sheng, B. Liu, N. Yu, X. Wank, J. Shao, and C. Chen, "Zoom-net: Mining deep feature interactions for visual relationship recognition," in *ECCV 2018*.
21. K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *CVPR 2020*.
22. G. Li, N. Duan, Y. Fang, Ming G, and Daxin J, "Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training," in *AAAI 2020*.
23. X. Han, Z. Zhang, N. Ding, J. Wen, J. Yuan, W. Zhao, and J. Zhu, "Pre-trained models: Past, present and future," *CoRR, 2021*.
24. W. Liange, Y. Jiang, and Z. Liu, "Graphvqa: Language-guided graph neural networks for scene graph question answering," *NAACL 2021*.
25. Y. Li, D. Tarlow, M. Brockschmidt, and R.S. Zemel, "Gated graph sequence neural networks," in *ICLR 2016*.
26. R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D.A. Shamma, M.S. Bernstein, and Li Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, 2017.
27. 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)*.
28. 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*.
29. 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.
30. 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.
31. 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.
32. 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)*.
33. 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.
34. Ilya Loshchilov and Frank Hutter. 2019. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

 doi:10.20944/preprints202511.1373.v1

35. 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.
36. Shengqiong Wu, Hao Fei, and Tat-Seng Chua. Universal scene graph generation. *Proceedings of the CVPR*, 2025.
37. 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.
38. 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.
39. 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.
40. 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.
41. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, may 2015. doi:10.1038/nature14539. URL http://dx.doi.org/10.1038/nature14539.
42. 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/.
43. 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.
44. 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.
45. 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. doi:10.1109/IJCNN.2013.6706748. URL http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=6706748.
46. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
47. 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.
48. 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.
49. A. Karpathy and L. Fei-Fei, "Deep visual-semantic alignments for generating image descriptions," *TPAMI*, vol. 39, no. 4, pp. 664–676, 2017.
50. 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.
51. 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.
52. 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.
53. 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.
54. 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.
55. Ling Zhuang, Hao Fei, and Po Hu. Knowledge-enhanced event relation extraction via event ontology prompt. *Inf. Fusion*, 100:101919, 2023.

 doi:10.20944/preprints202511.1373.v1

56. 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.
57. 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.
58. 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.
59. 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.
60. 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.
61. 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.
62. 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.
63. 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.
64. 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).
65. 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.
66. Bobo Li, Hao Fei, Fei Li, Yuhan 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.
67. 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.
68. 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.
69. 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).
70. 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.
71. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, A tree-based neural network model for biomedical event trigger detection, Information Sciences, 512, 175
72. Hao Fei, Yafeng Ren, and Donghong Ji. 2020, Dispatched attention with multi-task learning for nested mention recognition, Information Sciences, 513, 241
73. 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
74. 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.
75. Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. *arXiv preprint arXiv:1606.05250*, 2016.
76. 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.
77. D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

 doi:10.20944/preprints202511.1373.v1

78. 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.
79. K. Papineni, S. Roukos, T. Ward, and W. Zhu, "Bleu: a method for automatic evaluation of machine translation," in *ACL*, 2002, pp. 311–318.
80. 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.
81. 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. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.
82. Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *CoRR*, abs/2309.05519, 2023.
83. 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.
84. 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.
85. Naman Jain, Pranjali Jain, Pratik Kayal, Jayakrishna Sahit, Soham Pachpande, Jayesh Choudhari, et al. Agribot: agriculture-specific question answer system. *IndiaRxiv*, 2019.
86. 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.
87. 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.
88. 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.
89. 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.
90. 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.
91. 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.
92. S. Banerjee and A. Lavie, "METEOR: an automatic metric for MT evaluation with improved correlation with human judgments," in *IEEMMT*, 2005, pp. 65–72.
93. 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.
94. 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.
95. 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.
96. 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.
97. 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.

 doi:10.20944/preprints202511.1373.v1

98. 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.
99. 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.
100. 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.
101. 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.
102. 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.
103. 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.
104. P. Anderson, B. Fernando, M. Johnson, and S. Gould, "SPICE: semantic propositional image caption evaluation," in *ECCV*, 2016, pp. 382–398.
105. 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.
106. 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.
107. 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.
108. 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.
109. 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.
110. 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.
111. Sanjeev Arora, Yingyu Liang, and Tengyu Ma. A simple but tough-to-beat baseline for sentence embeddings. In *ICLR*, 2017.