IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 30 October 2025, accepted 12 December 2025, date of publication 17 December 2025, date of current version 8 January 2026.

Digital Object Identifier 10.1109/ACCESS.2025.3645220

RESEARCH ARTICLE

NeSyVQA: Neurosymbolic Visual Question Answering With Knowledge-Enriched Scene Graphs

M. JALEED KHAN[1,2], JOHN G. BRESLIN[1,3], (Senior Member, IEEE), AND EDWARD CURRY[1,3]

[1]Research Ireland Centre for Research Training in Artificial Intelligence, Data Science Institute, University of Galway, Galway, H91 TK33 Ireland
[2]Medical Sciences Division, University of Oxford, OX3 9DU Oxford, U.K.
[3]Insight Research Ireland Centre for Data Analytics, Data Science Institute, University of Galway, Galway, H91 TK33 Ireland

Corresponding author: M. Jaleed Khan (m.khan12@universityofgalway.ie)

This work was supported by the Research Ireland under Grant 18/CRT/6223 and Grant 12/RC/2289_P2.

ABSTRACT Incorporating multimodal features and heterogeneous common sense knowledge in scene representation and visual reasoning techniques is essential for accurate and intuitive Visual Question Answering (VQA). Knowledge-enhanced multimodal VQA offers transformative potential across computer vision applications, including accessibility, healthcare, security and education. Existing VQA techniques often neglect rich semantic and relational information about image content or rely on potentially biased knowledge sources with limited coverage. Our Neurosymbolic Visual Question Answering (NeSyVQA) framework addresses this by enriching scene graphs with rich background knowledge from a heterogeneous knowledge graph and employing the enriched scene graphs in an attention-based scene graph reasoning network. The framework employs a cascade of deep neural networks, including Convolutional Neural Networks (CNN) and Long Short Term Memory (LSTM) networks, for object detection and predicate classification within Scene Graph Generation (SGG). The initial scene graph is semantically enriched for improved expressiveness using background knowledge and related facts extracted from a heterogeneous knowledge graph. The enriched scene graphs are then employed for downstream VQA using an attention-based scene graph reasoning network. Promising evaluation results were obtained for SGG and VQA tasks on the General Question Answering (GQA) and Visual Genome (VG) benchmark datasets using their standard evaluation metrics. The proposed NeSyVQA framework outperformed the existing state-of-the-art SGG and VQA techniques with over 13% higher relationship recall rates in SGG and a 4% higher accuracy on open-ended questions in VQA, which underscores the efficacy of leveraging multimodal features and heterogeneous knowledge for complex visual reasoning. The source code is available at https://github.com/jaleedkhan/nesy-vqa

INDEX TERMS Knowledge enrichment, neurosymbolic integration, scene graph, scene understanding, visual question answering, visual reasoning.

I. INTRODUCTION

Visual Question Answering (VQA) is a fundamental task in visual reasoning that involves answering a variety of questions, posed in natural language, about a given image. The complexity and interdisciplinary nature of VQA have led to its recognition as an AI-complete task [1]. It incorporates image feature extraction from computer vision, question and answer feature extraction and natural language generation from natural language processing, and structured representation and semantic reasoning from knowledge representation and reasoning. Multimodal feature fusion is a crucial aspect of VQA, which involves creating a joint feature representation of the image-question pair for answer classification or generation. VQA is considered a type of Turing test for visual reasoning, as it evaluates the system's ability to perform semantic analysis of visual scenes at a level comparable to that of a human [2]. It has well-defined

The associate editor coordinating the review of this manuscript and approving it for publication was Liang-Bi Chen.

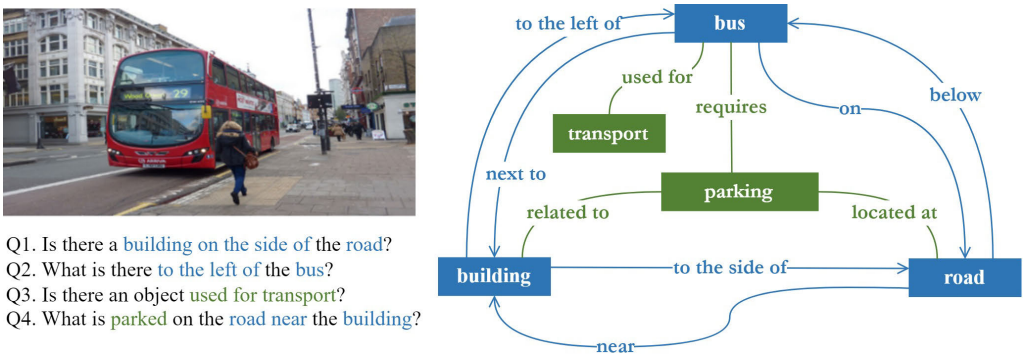


FIGURE 1. The scene graph (blue) of the image has sufficient information to answer Q1 and Q2. However, external background knowledge (green) about the visual concepts in the image is required to answer Q3 and Q4, indicating the need to incorporate heterogeneous common sense knowledge for scene graph reasoning.

evaluation benchmarks such as General Question Answering (GQA) [3] and potential applications across various domains. For instance, VQA can enable visually impaired people to navigate their surroundings and the internet with fewer visual barriers [4]. VQA can support automated medical diagnosis, assist healthcare staff with clinical decisions, and facilitate patient education [5]. VQA also holds promise for unattended surveillance systems, where it can help raise alarms in anomalous situations and assist human operators in making more informed and quicker decisions [6]. Other potential areas of application include education [7], art [8] and marketing [9], where it can enhance e-learning and student evaluation, replace audio guides in art galleries and museums, and evaluate human comprehension of advertising media, respectively. VQA stands at the intersection of multiple research domains and has the potential to revolutionise various aspects of our society.

VQA techniques often struggle with several challenges due to the complex nature of the task. These techniques typically require the execution of multiple computer vision sub-tasks for question answering and a significant volume of labelled images and question-answer pairs for generalisation [1]. However, collecting sufficient training data for all visual concepts can be tedious and often impractical [10]. Moreover, the conventional VQA models overlook the essential semantic and relational information about visual concepts in images, which is vital for visual reasoning. To this end, scene graph-based VQA emerged as a more effective solution, especially for questions that necessitate semantic reasoning [11]. The aim of neurosymbolic (NeSy) methodologies is to harness the extensive learning capacity and broad applicability of neural systems, complemented by the logical reasoning and interpretability provided by symbolic systems in AI [12]. Techniques within NeSy frameworks use neural representations to enhance symbolic reasoning, incorporate external knowledge into neural learning, or a synthesis of both, with the degree of integration between neural and symbolic elements varying from weak to moderate and then to a strong integration [13], [14]. Scene Graph Generation (SGG) employs a neurosymbolic approach that represents the image content as a structured graph, with objects as nodes and their relationships as edges. SGG usually involves a combination of deep learning-based multimodal techniques, mainly for object detection and relationship predicate classification. This graphical image representation proved beneficial for downstream visual reasoning techniques that require understanding the interactions between visual concepts [15], [16]. For instance, the scene graph in Figure 1 can be used to readily answer questions Q1 and Q2 about the image. However, answering complex questions requiring additional information and a deep semantic understanding of the scene and the question, such as Q3 and Q4, necessitates more than just objects and their relationships. Besides objects and relationships, background common sense knowledge about these visual concepts is required for a higher-level semantic understanding of the visual scene and enhanced reasoning capabilities.

Several knowledge-based methods have been introduced to address this challenge by leveraging statistical priors [17], [18], [19], [20], language priors [21], [22], and fact-based external knowledge [23], [24]. However, these methods have limitations due to their restricted knowledge and inherent biases. Knowledge Graphs (KGs) have emerged as a promising source of common sense knowledge. Some SGG methods [19], [24], [25], [26] incorporate knowledge from KGs, while others use graph-based message propagation [18], [27], [28], [29] to embed KG structural information into model representations. The Common Sense Knowledge Graph (CSKG) [30], a consolidated knowledge source with heterogenous common sense knowledge, was found to be beneficial for knowledge integration in SGG [31]. While several techniques have integrated knowledge from KGs for scene graph generation, their application to visual reasoning tasks like VQA remains limited. Some VQA techniques [28], [32] have used a limited subset [23] of DBpedia, ConceptNet, and WebChild, but these methods did not utilize scene graphs, thereby neglecting the structural features required for reasoning. There is a significant need to explore the use of rich and diverse common sense knowledge about visual concepts for VQA. This will enable VQA methods

to leverage the complementary structural features of scene graphs and broad coverage of common sense knowledge in heterogeneous KGs to answer complex questions accurately and intuitively.

To address the aforementioned challenges, we present NeSyVQA, a novel neurosymbolic visual reasoning framework for VQA based on the semantic enrichment of scene graphs and attention-based scene graph reasoning. The framework initiates SGG with CNN-based object detection, followed by LSTM-based classification of relationship predicates, which leverages visual-textual multimodal feature learning. The scene graphs are then semantically enriched using background knowledge and related facts extracted from a heterogeneous KG. Finally, the enriched scene graphs are used for downstream reasoning using an attention-based pipeline for VQA. NeSyVQA achieved state-of-the-art performance in SGG and VQA tasks on the GQA [3] and VG [33] datasets using the standard performance measures, which depicts the effectiveness of multimodal SGG and scene graph enrichment via heterogeneous common sense knowledge for visual reasoning. The NeSyVQA framework significantly and systematically extends the knowledge-based SGG method [31] with new methodological and experimental contributions, particularly on its application to VQA. The proposed framework addresses the expressiveness challenge in visual reasoning by bridging the gap between scene graphs and common sense knowledge and VQA. By enriching scene graphs with external knowledge, our framework enables deeper semantic understanding of scenes, which improved accuracy on open-ended questions and also enhanced performance on binary questions, consistency, validity and plausibility scores compared to existing state-of-the-art methods. Open-ended questions, in particular, benefit significantly from background common sense knowledge, as they often require reasoning that goes beyond visual features alone. The main contributions of this paper include:

1) We proposed a novel scene graph enrichment-based neurosymbolic visual reasoning framework for VQA (Figure 2), which consists of CNN- and LSTM-based multimodal scene graph generation (Algorithm 1), heterogenous KG-based scene graph enrichment (Algorithm 2) and attention-based scene graph reasoning for VQA (Algorithm 3).
2) In our SGG evaluation, the proposed approach achieved over 19% higher relationship recall rates in SGG, compared to the traditional data-centric approach without knowledge enrichment (Figure 3), and outperformed the existing state-of-the-art methods with over 13% higher relationship recall rates (Table 3).
3) We compared the performance of SGG using different KGs and different KG embedding models and observed that ComplEx [34] embeddings of CSKG [30] yielded the highest performance (Table 1 and Table 2).
4) In our VQA evaluation, the proposed approach achieved 29% higher accuracy on open-ended questions, 16% higher accuracy on binary questions and over 10% higher consistency, validity and plausibility scores, compared to the traditional approach data-centric approach without knowledge enrichment (Figure 4). The proposed approach also surpassed the existing state-of-the-art method with a 4% higher accuracy on open-ended questions as well as improved accuracy on binary questions, validity, consistency and plausibility scores (Table 4).

The rest of the paper is organised as follows: Section II presents the related work on this topic, the proposed framework is explained in Section III, and the comprehensive experimental analysis and discussion are presented in Section IV, which is followed by the conclusion in Section V.

II. RELATED WORK

In this section, we review the related works on scene graph generation, visual question answering and knowledge enrichment.

A. SCENE GRAPH GENERATION

Most SGG methods merge visual and semantic representations within deep neural networks to predict visual relationship triples on a large scale. Zhang et al. [35] proposed a tripartite approach for capturing visual representations, blending streams dedicated to subjects and objects with a stream for predicates to enhance the interactions between subjects and objects. During the learning process, text-based features were fused as visual feature labels. In a similar way, Peyre et al. [36] utilized a space for visual phrase embeddings throughout the learning stage, which aids in the prediction of previously unseen relationships and mitigates the impact of changes in appearance. For visual relationship prediction, Zellers et al. [19] and Chen et al. [18] utilized pre-calculated frequency priors to infuse common sense knowledge derived from dataset statistics. Xu et al. [37] introduced an iterative message passing (IMP) technique for refining object and relationship features in SGG. VCTree [38] harnessed dynamic tree structures and Bi-directional TreeLSTM for effective SGG. Tang et al. [39] adopted causal inference for predicting relationship triples and proposed the Total Direct Effect (TDE) approach to mitigate bias caused by imbalanced datasets. EBM [40] is an energy-based learning framework for scene graph generation, integrating scene graph structures into the output space and enabling effective learning from a limited label set due to its inductive bias. SVRP [41] allows for the inference of relations for unseen object classes, employing a two-step method that involves pre-training on coarse-grained region-caption data, followed by fine-tuning using prompt-based techniques without updating the model parameters. FGPL-A [42] leverages an adaptive predicate lattice and entity discriminating loss functions to dynamically identify and refine hard-to-distinguish predicates in SGG

models. Zhang et al. [43] developed the Saliency-guided Message Passing (SMP) technique to improve relationship reasoning and the adaptability of scene graphs by prioritizing the most significant visual relationship triples. In another work, Lin et al. [44] used the concept of heterophily within visual relationship triples to enhance the representation of relationships and enhance message passing within a Graph Neural Network (GNN) through an adaptive re-weighting transformer module aimed at facilitating the fusion of information. Zhou et al. introduced the Debiased Scene Graph Generation (DSDI) framework [45], which applies dual imbalance learning to balance head–tail relations. Subsequently, the Causal Feature Enhancement Network (CFEN) [46] refined relational representations through causal feature disentanglement. These causal or debiasing strategies contribute to relational completeness in SGG. With their focus on visual and language features only, existing methods neglect the impact of common sense knowledge and the structural features of visual concepts present in heterogeneous KGs.

B. VISUAL QUESTION ANSWERING

Anderson et al. [47] used Faster R-CNN to propose image regions and integrated bottom-up and top-down attention mechanisms to enhance the interpretability of attention weights and unified visual-linguistic understanding for VQA (UpDown). Tan and Bansal [48] proposed the LXMERT framework employing a large-scale Transformer model with three encoders for scene graph-based VQA based on the understanding of visual concepts and language semantics, as well as, intra- and cross-modal relationships. Meta Module Network (MNM) [49] addresses the scalability and generalizability in VQA using a metamorphic meta module, which dynamically morphs into diverse instance modules, offers flexibility and allows for complex visual reasoning, while preserving the same model complexity as the function set expands. MDETR [50] is an end-to-end modulated detector that leverages a transformer-based architecture to fuse image and text modalities at an early stage for efficient extraction of visual concepts from the free-form text in multi-modal reasoning systems including VQA. Zhang et al. [51] performed visual reasoning for VQA based on their object detection model designed for visual-language tasks with richer visual representations of objects and concepts. Among the scene graph-based VQA methods, Hudson and Manning [52] presented a visual reasoning approach based on Neural State Machines (NSM) integrating visual and linguistic inputs into semantic concepts via a probabilistic scene graph for sequential reasoning and inference. Zhang et al. [53] embedded the structural features of scene graphs into a Graph Neural Network (GNN) for downstream VQA. Yang et al. [54] proposed Scene Graph Convolutional Network (SceneGCN) that incorporates object properties and semantic relationships into a structured scene representation for enhanced VQA via visual context and language priors. Graphhopper (GH) [55] addresses the challenge of performing multi-hop knowledge graph reasoning over complex visual scenes to predict reasoning paths that lead to the answer in VQA. Dual Message-passing enhanced Graph Neural Network (DM-GNN) [56] encodes multi-scale scene graph information into two diversified graphs focused on objects and relations, and uses a dual structure to encode them to achieve a balanced representation of object, relation, and attribute features in VQA. The Scene Graph Refinement network (SGR) [57] propose a transformer-based refinement network to enhance object and relation feature learning in VQA, utilizing question semantics to jointly learn multimodal representations and select the most relevant relations for question answering. Eiter et al. [58] introduced a neurosymbolic VQA pipeline, integrating neural network predictions with logic programming to handle imperfect data and compute answers. Yi et al. [59] combined deep learning for visual and language processing with symbolic reasoning, executing derived programs on structural scene representations to answer questions. Dahlgren and Dan [60] investigated the compositional generalization in multimodal mathematical reasoning, revealing limitations in current VQA models, and suggested knowledge-based curriculum learning for enhanced reasoning. The Question-aware Dynamic Scene Graph (QDSG) method introduces a dynamic scene graph refinement mechanism that leverages question-specific word-level co-attention to adaptively refine both node and edge features for improved scene understanding and iterative reasoning in VQA [61]. Recent works such as Core-to-Global Reasoning (CTGR) [62] attempt to unify local object interactions and global context reasoning through a hierarchical neural module. While CTGR achieves strong compositional reasoning, it remains limited to data-driven inference within the visual domain. The lack of explicit integration of external common-sense knowledge into the scene representation and reasoning for improved downstream reasoning and interpretability remains a gap.

C. KNOWLEDGE ENRICHMENT

Early knowledge-based methods leaned on statistical priors [17], [18], [19], [20] and language priors [21], [22] for common sense knowledge infusion. While priors helped slightly advance the prediction performance of relationship triples in SGG, they possess significant limitations that restrict their expressivity and applicability in downstream visual reasoning, particularly in VQA. Statistical priors are often based on heuristic methods, limiting their generalizability, while language priors are susceptible to the constraints of word embeddings, especially when applied to underrepresented objects in benchmark datasets. These shortcomings are exploited by downstream VQA models, which tend to rely heavily on such statistical biases and trends within the answer distribution, thereby significantly limiting their visual reasoning capabilities. Recognizing the necessity of understanding visual scenes beyond mere image-question

pairs for comprehensive visual reasoning, few strategies have been employed. Some approaches aim to retrieve basic factual information to answer questions [23], [24], while others focus on actively acquiring additional information and predicting the answer [63]. Fact-based VQA (FVQA) [23] attempts to address this by learning query mappings and retrieving information from a knowledge base with limited factual information. Similarly, Narasimhan and Schwing [24] answer questions by predicting the key triple associated with the question. These methods rely on ground truth triples for supervision and are limited to one-hop reasoning only. Their effectiveness is limited due to the lack of extensive coverage and diversity of general common sense knowledge about visual scenes, leading to difficulties in answering questions that necessitate implicit and background knowledge about the visual concepts present in the scenes. KGs have become a promising source of common sense knowledge within knowledge-based scene representation and visual reasoning techniques. Certain SGG methods [19], [24], [25], [26] incorporate related knowledge extracted from KGs at various stages of the SGG process. Other techniques utilize message passing approaches [18], [27], [28], [29] to encode KG structural features into the model. Integrating knowledge from multiple KGs into a heterogeneous knowledge source with enhanced the coverage and diversity of common sense knowledge. Zareian et al. [64] proposed the Graph Bridging Network (GB-Net), a framework designed to construct a scene graph and associate its nodes and edges with associated entities in a common-sense graph sourced from VG, WordNet, and ConceptNet. It utilizes message passing in GNNs for the iterative enhancement of the relationships within the scene graph. Similarly, Guo et al. [65] harnessed relational and common-sense knowledge extracted from VG and ConceptNet, integrating this information into an Instance Relation Transformer (IRT) for predicting relationship triples in SGG. The Gaussian Distribution-Aware SGG (GDA-SGG) method models commonsense knowledge using Gaussian distributions, providing a probabilistic space for uncertainty in visual context and commonsense, coupled with multi-expert classifiers for denoising, enabling more accurate scene graph generation [66]. The Bipartite Graph Neural Network (BGNN) leverages confidence-aware adaptive message propagation and bi-level data resampling to mitigate the challenges of long-tailed class distributions and intra-class variations for unbiased scene graph generation [67]. The Confidence-Aware Commonsense Integration SGG (CA-SGG) method incorporates a hybrid-attention module to minimize uncertainty in representation learning, paired with a confidence estimation branch to dynamically adjust the need for commonsense knowledge in relation recognition tasks [68]. CSKG [30], a systematically consolidated common sense knowledge source, was used for knowledge infusion in SGG [31], and the resulting scene graphs were used for downstream image synthesis. While these SGG methodologies leverage multiple knowledge graphs, their application in visual reasoning techniques, particularly VQA, remains unexplored. Evaluating the impact of integrating common sense knowledge from various KGs is essential for advancing visual reasoning capabilities. Certain VQA methodologies [28], [32] have employed a restricted selection [23] from KGs like DBpedia, ConceptNet, and WebChild. However, these approaches did not leverage scene graphs, missing out on the valuable structural features related to visual concepts. To address this gap, Ziaeefard and Lécué [69] introduced a VQA technique based on Graph Attention Networks, which incorporates both scene graphs and contextual knowledge from ConceptNet, aiming to enrich the understanding and processing of visual information. It is crucial to assess the efficacy of leveraging associated background and factual knowledge from multiple KGs in visual reasoning. Some VQA methods [28], [32] have used a limited subset [23] of ConceptNet, WebChild and DBpedia, but these methods did not utilize scene graphs, thereby overlooking the structural and relational information about image content. Ziaeefard and Lécué [69] proposed a Graph Attention Networks-based VQA technique that encodes scene graphs along with external knowledge extracted from ConceptNet. The VQA-GNN model proposed by Wang et al. [70] introduced a bidirectional fusion technique to merge structured knowledge from ConceptNet and scene graphs and leveraged GNNs to perform inter-modal message passing for question answering. There is a significant need to explore the use of rich common sense knowledge to enrich scene graph-based VQA methods to alleviate the existing challenges. This will enable the VQA methods to jointly leverage the complementary structural features of scene graphs and rich common sense knowledge in heterogeneous KGs for visual reasoning.

III. NEUROSYMBOLIC VISUAL QUESTION ANSWERING FRAMEWORK

The proposed Neurosymbolic Visual Question Answering (NeSyVQA) framework employs CNN- and LSTM-based object detection and multimodal predicate classification, and heterogeneous KG-based enrichment of scene graphs, which is followed by attention-based scene graph reasoning for VQA. The NeSyVQA framework is presented in Figure 2, and each module is presented in detail in the following subsections.

A. SCENE GRAPH GENERATION

Algorithm 1 presents the scene graph generation process. Faster RCNN [71], with ResNeXt-101-FPN [72] backend, is employed to detect objects. It provides a label l and a bounding box b for each object within an image I, as well as the feature maps F generated by the backend network. The detected bounding boxes b and labels l from Faster RCNN serve as the initial object nodes in the scene graph, while the feature maps F provide the context for generating relationships between these objects. Potential redundancy in object prediction, indicated by significant bounding box overlap, similar labels, or identical structural patterns within

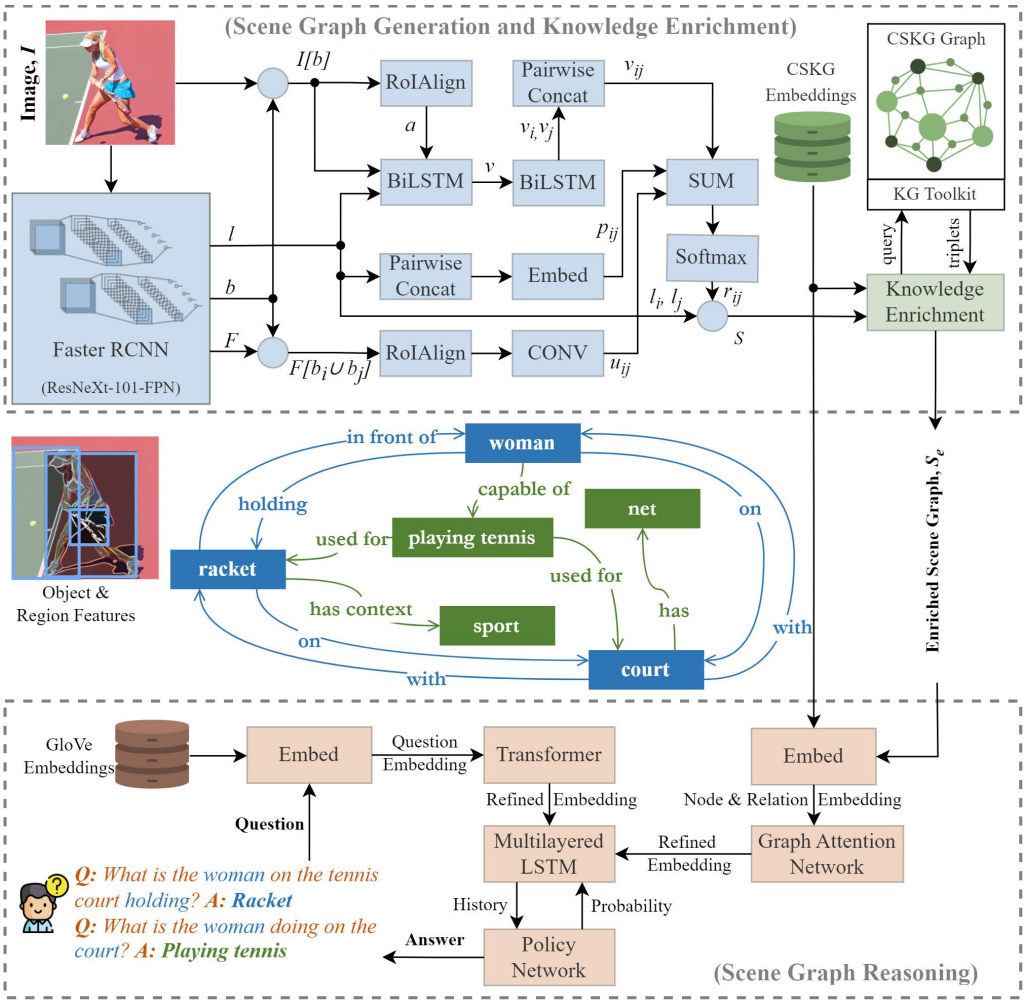


FIGURE 2. Proposed NeSyVQA framework comprising CNN- and LSTM-based multimodal scene graph generation, scene graph enrichment using a heterogeneous knowledge source, and an attention-based scene graph reasoning network for VQA.

Algorithm 1 Scene Graph Generation

Input: Image I
Output: Scene Graph S
$\{b, l, F\} \leftarrow \text{FasterRCNN}(I)$
$a \leftarrow \text{RoIAlign}(I[b])$
$v \leftarrow \text{BiLSTM}(a, I[b], l)$
for each pair of objects i, j **do**
$v_{ij} \leftarrow \text{concat}(\text{BiLSTM}(v_i), \text{BiLSTM}(v_j))$
$p_{ij} \leftarrow \text{embed}(\text{concat}(l_i, l_j))$
$u_{ij} \leftarrow \text{conv}(\text{RoIAlign}(F[b_i \cup b_j]))$
$\{r_{ij}, c_{ij}\} \leftarrow \text{softmax}(\text{SUM}(v_{ij}, p_{ij}, u_{ij}))$
end for
$S \leftarrow \{l_i, r_{ij}, l_j\}$

the CSKG, is mitigated at this stage. To minimize prediction errors, object nodes with a high Intersection over Union (IoU) of bounding boxes or substantial cosine similarity in CSKG embeddings compared to another object node are systematically discarded.

To extract region features a from specific image areas $I[b]$ defined by object bounding boxes, we apply the RoIAlign method [73]. These features a serve as a basis for computing the visual context features v for each object, which are then encoded using the combination of a, the cropped image regions $I[b]$, and labels l through Bi-LSTM layers [19]. The Bi-LSTM architecture, noted for its ability to process sequences with variable lengths and manage long-term dependencies due to its bidirectional nature, considering both previous and upcoming contexts of objects, is adept at predicting the pairwise visual relationships.

To derive the combined pairwise object features v_{ij} for distinct object pairs ($i \neq j; i, j = 1, \ldots, n$), where n is the number of detected objects, the process involves encoding the individual visual context features (v_i, v_j) using Bi-LSTM and subsequent concatenation. The language prior p_{ij} is determined through the embedding of pairwise object labels (l_i, l_j). The contextual union features u_{ij} are extracted by applying RoIAlign to the combined regions of object pairs within the feature maps F. These three distinct feature sets, v_{ij}, p_{ij} and u_{ij}, are integrated via a summation operation [74],

and are employed in softmax classification to predict the relationship predicates r_{ij} alongside their confidence values c_{ij}. By combining visual features (v_{ij}), language priors (p_{ij}), and union features (u_{ij}), the model captures both the visual and semantic context, improving its ability to predict accurate relationships between object pairs. This process concludes by connecting the pairwise objects and their associated relationship predicates into a structured representation to create the scene graph S.

B. KNOWLEDGE ENRICHMENT

The Knowledge Graph Toolkit (KGTK) [75] is used to extract new relationship triples from CSKG, specifically targeting those that share a subject or object node with the existing elements in the scene graph. Each scene-graph node v and candidate CSKG node u are embedded as $e_{\text{sg}}(v)$ and $e_{\text{kg}}(u)$, respectively, and compared via cosine similarity. Each scene-graph node v is represented as $e_{\text{sg}}(v) = \alpha\,\phi_{\text{text}}(\text{name}(v)) + (1-\alpha)\,\phi_{\text{vis}}(\text{roi}(v))$, where ϕ_{text} and ϕ_{vis} denote the textual and visual encoders respectively. Each CSKG node u is represented as $e_{\text{kg}}(u) = \psi_{\text{text}}(\text{name}(u))$. The similarity score is $s = \cos\big(e_{\text{sg}}(v), e_{\text{kg}}(u)\big)$; triples are retained if $s \geq \tau$. Cosine similarity is used because it is scale-invariant and effectively measures angular proximity between heterogeneous embedding spaces. Unless stated otherwise, $\alpha = 0.5$ and $\tau = 0.8$, which provided the best trade-off between textual–visual alignment and enrichment coverage. (For brevity, Algorithm 2 uses $e_1 = e_{\text{sg}}(v)$ and $e_2 = e_{\text{kg}}(u)$.)

The enrichment process proceeds through five main stages: (1) Candidate mining — retrieve CSKG nodes lexically or semantically related to scene-graph nodes. (2) Embedding scoring — compute $s = \cos(e_{\text{sg}}, e_{\text{kg}})$. (3) Filtering — retain top-k candidates with $s \geq \tau$. (4) Alignment — map CSKG predicates to scene-graph relations. (5) Post-processing — remove duplicates and self-loops.

The next step involves establishing links between the nodes in the newly extracted triples and the associated object nodes within the scene graph, focusing on those pairs that demonstrate a significant degree of structural similarity. In cases where a node from the newly identified triples already exists within the scene graph, the strategy is to directly connect the edge of the triple to this pre-existing node, thereby reducing redundancy. A cosine similarity score is computed between node embeddings e_1 and e_2 to determine whether to introduce a new triple from the CSKG into the enriched scene graph, and the triple is added only if $s \geq \tau$. After the process of enriching scene graphs, they are aligned with the representation model, facilitating their straightforward incorporation into the scene graph-based VQA model. To align newly introduced triples with the scene-graph relation vocabulary, each predicate is mapped using a data-driven probability model:

$$r^* = \arg\max_{r \in R_{\text{SG}}} \frac{\text{count}_{VG}(h, r, t) + \lambda}{\sum_{r' \in R_{\text{SG}}} \text{count}_{VG}(h, r', t) + \lambda\,|R_{\text{SG}}|},$$

Algorithm 2 Scene Graph Enrichment

Input: Scene Graph S, Common Sense Knowledge Graph G_{CSKG}
Output: Enriched Scene Graph S_e
$S_e \leftarrow S$
for each node in S **do**
$e_1 \leftarrow$ sg_emb(*node*)
$triples_{CSKG} \leftarrow$ query(G_{CSKG}, *node*)
$triples_{CSKG} \leftarrow$ preprocess($triples_{CSKG}$)
for each triple in $triples_{CSKG}$ **do**
if *node* == *triple*[node1] **then**
$e_2 \leftarrow$ cskg_emb(*triple*[node2])
else
$e_2 \leftarrow$ cskg_emb(*triple*[node1])
end if
$s \leftarrow$ cosine_sim(e_1, e_2)
if $s \geq \tau \wedge triple \notin S_e$ **then**
$S_e.append(triple)$
end if
end for
end for
$S_e \leftarrow$ postprocess(S_e)

Algorithm 3 Scene Graph Reasoning

Input: Enriched Scene Graph S_e, Question Q
Output: Answer *Ans*
Initialize node embeddings in S_e using label embeddings
$S'_e \leftarrow$ GAT(S_e)
Initialize words in Q using GloVe embeddings
$Q' \leftarrow$ transformer(Q)
Initialize agent state at hub node with Q'
$n_{current} \leftarrow$ get_current_node(*state*)
while $n_{current} \neq$ pred_answer(*state*)**and***Ans* $\neq \emptyset$ **do**
history $\leftarrow$ LSTM(*state*)
prob $\leftarrow$ policy_network(*history*)
action $\leftarrow$ sample(*prob*)
state $\leftarrow$ update(*state*, *action*)
$n_{current} \leftarrow$ get_current_node(*state*)
end while
Ans $\leftarrow$ pred_answer(*state*)

where λ is a smoothing constant. If no reliable mapping is found, a generic relation such as *LocatedNear* is assigned to maintain graph connectivity. New entities from CSKG are aligned to detector labels through string matching and synonym expansion using WordNet; if no direct match exists, the entity is linked to its nearest semantic neighbor in the embedding space.

C. DOWNSTREAM REASONING

The downstream scene-graph reasoning module for VQA, presented in Algorithm 3, uses an agent-based mechanism that begins at a central node connected to most nodes in the enriched scene graph and traverses to adjacent nodes

until it reaches the node representing the final answer. Each entity and relation in the enriched scene graph is initialized with the embeddings of their textual labels, providing the initial semantic representation of the scene. A Graph Attention Network (GAT) [76] then refines these node embeddings by aggregating information from neighboring nodes to incorporate relational context. GAT enables the model to focus on the most relevant relations for answering the question by weighting neighboring nodes according to their importance. Relations and inverse relations allow bidirectional context flow, yielding a more comprehensive understanding of the scene.

The question Q is encoded by first initializing its words using GloVe embeddings [77] to capture lexical semantics, followed by a transformer [78] that produces a context-aware question representation Q'. The reasoning process maintains a query or state vector q_t that evolves as the agent moves through the graph. At each reasoning step t, attention weights are computed as

$$\alpha_{t,j} = \text{softmax}_j\left(q_t^\top W h_j\right),$$

where h_j is the embedding of a neighboring node j and W is a learnable attention weight matrix. The agent then updates its internal state according to

$$q_{t+1} = f(q_t, h_{i^*}), \quad i^* = \arg\max_j \alpha_{t,j},$$

where the most relevant neighbor i^* is selected based on the highest attention weight. After T reasoning steps, the final answer logits are computed as $y = W_o q_T$, where W_o is an output projection matrix, and the network is trained using cross-entropy loss with ground-truth answers.

As LSTMs are suitable for modeling temporal dependencies, a multilayer LSTM [79] encodes the traversal history of the agent on the enriched scene graph. The LSTM processes the embeddings of previously taken actions to form a history representation, which is fed into a policy network that outputs a probability distribution over the next possible actions. The policy network guides the agent through the graph by selecting the most promising neighbor at each step. The agent continues to traverse until it reaches a terminal node predicted to represent the answer. The state of the agent—initialized with the question representation Q'—thus evolves through iterative attention, state updates, and history encoding, integrating both the enriched scene structure and linguistic context to infer the final answer.

The neural (SGG and VQA pipelines) and symbolic (structured representation and knowledge enrichment) components in the NeSyVQA framework are loosely coupled as per the taxonomy of neurosymbolic approaches in [13] and [14]. These components operate in tandem to enhance collective performance, i.e. the accuracy of the initial scene graph plays a crucial role in effective knowledge enrichment, which ultimately impacts the performance of downstream reasoning for VQA.

IV. EXPERIMENTS AND RESULTS

A. EXPERIMENTAL SETUP

1) PLATFORM AND TOOLS

We used PyTorch[1] and KGTK[2] for implementation of the proposed framework and conducted experiments on a machine with AMD Ryzen 7 1700 Eight-Core Processor, 16 GB RAM, NVIDIA TITAN Xp GPU (with 12 GB memory) and Ubuntu 18.04.

2) DATASETS

The General Question Answering (GQA) dataset [3] is the standard dataset for scene graph-based VQA. It contains 113,018 images, 22 million questions, 1702 object classes and 310 relationship types, with an 80-10-10 split for training, validation and testing. Given the long-tailed distribution of objects and relationships in the dataset that impacts the performance of SGG, we used the common subset of the dataset with the most frequent 800 object classes and 170 relationship classes that account for more than 95% of their instances in the dataset. We conducted additional experiments on the Visual Genome [33] dataset to validate the results further. It contains 108K labelled images, 1.7 million open-ended question and answer pairs, and annotations for objects and visual relationships, with the most frequent 150 object classes and 50 relationship classes included in the standard split [37] we used.

3) EVALUATION METRICS

We used the following standard metrics to evaluate the performance of SGG:

1) ''Recall@K ($R@K$)'' [21] measures the fraction of times the correct relationship is among the top K confident relationship predictions, considering not just the correctness of predicted relationship labels, but also their confidence scores.
2) ''mean Recall@K ($mR@K$)'' [18], [38] computes the mean of R@K values computed separately for each relationship category, with the aim to mitigate bias towards dominant relationships during the evaluation.

Unless stated otherwise, we evaluate SGG performance under the Scene Graph Detection (SGDet) setting, in which the model detects objects and predicts pairwise relationships among them.

In VQA, the ''binary'' type questions are designed to have a 'yes' or 'no' answer, for example, questions that involve checking the presence, absence, or relationship between objects in the image. On the other hand, the ''open'' type questions require a more elaborate answer that needs deeper reasoning about the semantics of the visual content, usually involving identifying, describing, or explaining objects and relationships in the image. Apart from the standard accuracy metric, the performance metrics in GQA [3] are more robust to informed guesses as they need a deeper semantic

[1] https://pytorch.org/
[2] https://kgtk.readthedocs.io/

understanding of questions and visual content. The following performance metrics are used to quantify the reasoning capabilities of the VQA methods:

1) "Accuracy" (Top-1) is the fraction of times the predicted answer with the highest probability matches the groundtruth, separately calculated for binary and open questions.
2) "Consistency" measures the ability to answer multiple related questions consistently, indicating the level of understanding of the semantics of a question within the scene.
3) "Validity" evaluates whether an answer aligns with the scope of the question, reflecting the ability to comprehend the question.
4) "Plausibility" measures if an answer is reasonable within the context of the question and in line with real-world knowledge.
5) "Distribution" (lower is better) checks the match between the distributions of predicted answers and groundtruth, showing the ability to predict the less frequent answers in addition to the common ones.

Unless otherwise noted, all results on GQA are computed on the official test-dev split using the public GQA evaluation server and official metric implementation.[3]

Consider an image showing a picnic scene with a red apple on a blue blanket. Responding to the question "What colour is the apple on the picnic blanket?" with an answer "red" demonstrates accuracy. Maintaining the same answer across related questions, such as "Is there a red object on the blue blanket?" and "Is the fruit on the blanket red?" indicates "consistency". The answer "green" is inaccurate in this scene, yet it is "valid" and "plausible". Conversely, the answer "blue" is neither valid nor plausible as apples are not naturally blue. The diversity, compositionality, broader semantic space and real-world content of GQA, coupled with its comprehensive suite of performance metrics, make it well-suited for the benchmark evaluation of enriched scene graph-based VQA. Other datasets either include synthetic images [80], lack semantic representation [81] or confine knowledge enrichment to the limited background knowledge embedded within the datasets [23], [82], [83].

4) KG EMBEDDING MODELS

We compared the performance of the following four KG embedding models in scene graph enrichment:

1) TransE [84] is a KG embedding model that uses a translation operation to model relationships between entities.
2) RESCAL [85] represents entities as vectors and relationships as matrices, allowing it to capture higher-order relationships.
3) DistMult [86] uses a distance-based multiplication operation to model relationships.
4) ComplEx [34] extends DistMult by using complex-valued vectors.

The specific approach used by each model differs; however, they all aim to map entities and relationships in the KG to points in a high-dimensional vector space, such that the geometric relationships between the points reflect the relationships between the entities and relationships in the KG.

5) BASELINES

We compared the SGG performance of our method with

1) existing common sense knowledge-based SGG methods, including DSGAN [17], IRT-MSK [65], MOTIFS [19], GB-Net [64], KERN [18], COACHER [87], KB-GAN [25], DeepVRL [22] and VRD [21], and
2) conventional data-centric SGG methods, including HL-Net [44], TDE [39], SS-RCNN [88], SMP [43], NICEST [89], VCTree [38], IMP+ [37], FactorizableNet [90], MSDN [91], Graph RCNN [92], FGPL-A [42], EBM [40], SVRP [41], and DSDI [45].

We compared the VQA performance of our method with

1) existing scene graph-based VQA methods, including Graphhopper [55], DM-GNN [56], NSM [52], SGR [57], CTGR [62] and SceneGCN [54], and
2) conventional multimodal VQA methods including VinVL [51], UpDown [47], MDETR [50], MMN [49], and LXMERT [48].

The performance of these methods is reported in the same setting, using the standard dataset split.

B. RESULTS AND DISCUSSION

1) SGG EVALUATION

The Faster RCNN was trained using image data and object annotations using a learning rate of 0.003 (decreased by a factor of ten after 70k and 100k iterations), SGD optimizer and a batch size of 2. Post-training, the Faster RCNN was fixed, and the whole SGG model was trained using the image data and annotations of relationship triples using a learning rate of 0.04 (reduced by a factor of ten twice when validation performance plateaued), SGD optimizer and batch size of 4. We observed $R@100$ of 32.7 and $mR@100$ of 12.1 on the GQA test set. After enrichment of scene graphs, we noted a substantial improvement in recall rates, i.e. $R@100$ and $mR@100$ increased to 41.7 and 15.1 respectively on the GQA dataset, as shown in Figure 3. Similar results were noted for the VG dataset.

This advancement can be attributed to the additional visual cues injected by CSKG, particularly regarding the relative positions of objects and their potential interactions, which aid in minimizing errors and omissions during scene graph construction. Figure 5 demonstrates how the recall rate, particularly $R@100$, varies with the cosine similarity threshold, τ, used in Algorithm 2. As the threshold increases initially, recall rises significantly, capturing more relationships between the detected objects. Between the threshold values of 0.5 to 0.7, recall growth diminishes, and plateaus

[3]https://cs.stanford.edu/people/dorarad/gqa/evaluate.html

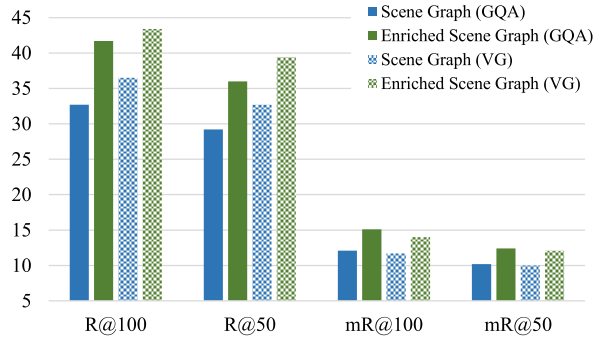


FIGURE 3. SGG performance measures observed before (blue) and after (green) semantic enrichment of scene graphs.

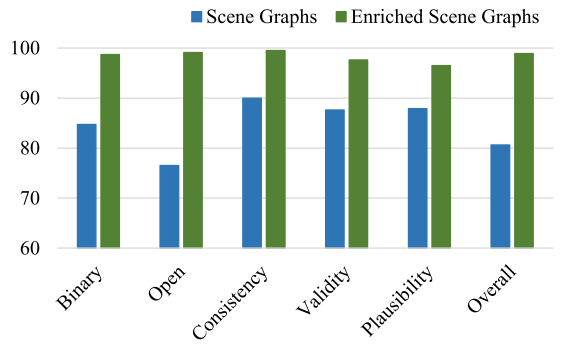


FIGURE 4. VQA performance measures observed before (blue) and after (green) semantic enrichment of scene graphs.

TABLE 1. Comparison between different KGs using ComplEx embeddings.

KG	*R*@100 (GQA)	*R*@100 (VG)
CSKG [30]	41.7	43.4
ConceptNet [93]	33.2	37.6
WordNet [94]	32.9	36.6

TABLE 2. Comparison between different KG embeddings of CSKG.

KG Embedding Model	*R*@100 (GQA)	*R*@100 (VG)
ComplEx [34]	41.7	43.4
DistMult [86]	40.8	42.1
TransE [84]	37.9	39
RESCAL [85]	34.5	37

after 0.8, indicating the addition of more irrelevant rather than meaningful relationships. This suggests that fine-tuning the similarity threshold is essential for minimizing noise while maximizing the inclusion of meaningful relationships. A threshold value of 0.8 was adhered to throughout the experimentation phase due to the highest recall rate achieved at this threshold.

2) COMPARISON OF KGs AND EMBEDDING MODELS

The recall rates obtained by the proposed framework with ComplEx embeddings of different KGs are shown in Table 1. Due to its heterogeneous nature and broader coverage of common sense knowledge, CSKG achieved a significantly higher recall rate compared to ConceptNet and WordNet. The recall rates obtained by the proposed framework with different embedding models for CSKG are shown in Table 2. Due to their capability to represent complex-valued and multi-dimensional relationships between entities in CSKG, ComplEx and DistMult achieved higher recall rates than TransE and RESCAL in the same setting, meaning that ComplEx and DistMult are more expressive and better suited for visual relationship prediction. ComplEx achieved the highest performance compared to the rest of the embedding methods.

3) SGG BENCHMARK COMPARISON

The performance of our method is compared with the baselines in Table 3, which shows that the proposed framework achieved a notably higher recall rate than all the comparative methods in terms of all metrics on both datasets. This significant advancement depicts that leveraging rich, diverse common sense knowledge in heterogeneous KGs enables the proposed framework to complement scene graphs with high-level semantics of visual concepts for more accurate and expressive scene representation.

4) VQA EVALUATION

The scene graph-based VQA model was trained and evaluated, both prior to and following scene graph enrichment. The performance metrics, as illustrated in Figure 4, demonstrate a significant improvement in accuracy for both binary and open-ended questions after scene graph enrichment. Binary question accuracy increased by 16%, while the accuracy for open-ended questions showed an even more pronounced gain of 29%. This greater improvement in open-ended questions can be attributed to the enhanced ability of enriched scene graphs to handle the complexity and diversity of open-ended questions, which require a deeper semantic understanding of the scene and the relationships between its components. The enriched scene graphs also demonstrated superior consistency, validity, and plausibility scores, further validating a more profound semantic understanding, better comprehension of the questions, and improved contextual reasoning abilities.

The largest gains appear in accuracy on open-ended questions and consistency, indicating that enrichment most benefits multi-step and compositional reasoning that relies on contextual knowledge beyond visual features alone. Improvements in validity and plausibility are smaller but consistent, suggesting better question grounding and more realistic answer distributions. Notably, these gains correlate with the observed increase in SGG recall after enrichment (Fig. 3), supporting the view that richer, more complete relational structure translates into stronger downstream reasoning. Overall, the results affirm that knowledge-enriched scene graphs improve both answer accuracy and the stability of reasoning across related questions.

The enrichment process provides a richer and more comprehensive representation of the scene, allowing the

TABLE 3. Benchmark comparison with existing SGG methods.

Dataset	Method	Knowledge Source	R@100	R@50	mR@100	mR@50
GQA [3]	**NeSyVQA (w/ enrichment)**	CSKG [30]	**41.7**	**36.0**	**15.1**	**12.4**
	SVRP [41]	-	35.8	31.8	12.8	10.5
	NeSyVQA (w/o enrichment)	-	32.7	29.2	12.1	10.2
	VCTree [38]	-	30.0	25.9	10.1	8.2
	TDE [39]	-	28.3	26.2	10.5	8.6
	MOTIF [19]	Statistical Prior	25.9	22.6	6.3	5.2
	IMP+ [37]	-	24.6	19.7	7.6	1.3
	FGPL-A [42]	-	23.8	-	3.2	-
VG [33]	**NeSyVQA (w/ enrichment)**	CSKG [30]	**43.4**	**39.4**	**14.0**	**12.1**
	HL-Net [44]	-	38.1	33.7	9.2	-
	TDE [39]	-	37.8	33.3	11.1	9.3
	CA-SGG [68]	ConceptNet [93]	37.3	32.5	7.3	6.3
	SS-RCNN [88]	-	36.9	32.7	10.0	8.4
	SMP [43]	-	36.9	32.6	-	-
	NeSyVQA (w/o enrichment)	-	36.5	32.7	11.7	10
	BGNN [67]	-	35.8	31	12.6	10.7
	EBM [40]	-	33.7	26.8	11.6	9.7
	DSGAN [17]	Statistical Prior	32.9	28.8	11.8	8.9
	NICEST [89]	-	32.7	29.0	12.4	10.4
	DSDI [45]	-	32.1	27.7	12.1	10.3
	GDA-SGG [66]	ConceptNet [93]	31.8	28.6	10.1	9.4
	VCTree [38]	-	31.3	27.9	8.0	6.9
	IRT-MSK [65]	CN [93] and VG [33]	31.0	27.8	-	-
	MOTIF [19]	Statistical Prior	30.3	27.2	6.6	5.7
	GB-Net [64]	CN [93], WN [94] and VG [33]	30.0	26.4	7.3	6.1
	KERN [18]	Statistical Prior	29.8	27.1	7.3	6.4
	IMP+ [37]	-	24.5	20.7	4.8	3.8
	COACHER [87]	ConceptNet [93]	22.2	19.3	-	-
	KB-GAN [25]	ConceptNet [93]	17.6	13.6	-	-
	FactorizableNet [90]	-	16.5	13.1	-	-
	MSDN [91]	-	14.2	10.7	-	-
	Graph RCNN [92]	-	13.7	11.4	-	-
	DeepVRL [22]	Language Prior	12.6	13.3	-	-
	VRD [21]	Language Prior	0.5	0.3	-	-

TABLE 4. Benchmark comparison with existing VQA methods on GQA dataset [3].

Method	Scene Graph	Knowledge Enrichment	Binary	Open	Consistency	Validity	Plausibility	Distribution	Overall
Humans [3]	-	-	91.2	87.4	98.4	**98.9**	**97.2**	-	89.3
NeSyVQA (w/ enrichment)	✓	✓	**98.7**	**99.1**	**99.5**	97.6	96.5	0.07	**98.9**
QDSG [61]	-	-	94.5	96.7	**99.5**	95.3	95.3	**0.05**	96.4
VQA-GNN [70]	✓	✗	–	–	–	–	–	–	90.3
Graphhopper [55]	✓	✗	85.8	77.3	92.9	92.3	89.5	–	81.4
NeSyVQA (w/o enrichment)	✓	✗	84.7	76.5	89.9	87.6	87.9	3.1	80.6
CTGR [62]	✓	✗	–	–	–	–	–	–	73.3
DMGNN [56]	✓	✗	69.8	72.2	-	93.8	–	3.8	71.2
VinVL [51]	✗	✗	82.6	48.7	94.4	96.6	84.9	4.7	64.7
UpDown [47]	✗	✗	66.6	34.8	78.7	96.2	84.6	5.9	64.7
NSM [52]	✓	✗	78.9	49.3	93.3	96.4	84.3	3.7	63.2
MDETR [50]	✗	✗	80.9	46.2	93.9	96.3	84.2	5.4	62.5
SGR [57]	✓	✗	78.9	46.4	92.7	96.8	86.4	**1.6**	61.7
MMN [49]	✗	✗	78.9	44.9	92.5	96.2	84.5	5.5	60.8
LXMERT [48]	✗	✗	77.8	45	93.1	96.4	85.2	6.4	60.3
SGCN [54]	✓	✗	70.3	40.6	83.5	95.9	84.2	6.4	54.6

VQA model to better understand the semantics and context of the scene and the question, leading to more valid and plausible answers and consistent responses to related questions. For instance, the increase in consistency from 89.9% to 99.5% suggests that enriched scene graphs not only improve the direct prediction of relationships but also allow the system to reason more coherently across multiple, related questions.

Figure 5 demonstrates how the overall accuracy varies with the cosine similarity threshold, τ, used in Algorithm 2. As the threshold increases initially, the accuracy of VQA increases significantly with the increase in the recall rate and expressivity of SGG. Between the threshold values of 0.5 to 0.7, this increase diminishes, and plateaus after 0.8 due to the addition of more irrelevant rather than meaningful relationships in the scene graphs. A threshold value of 0.8 was

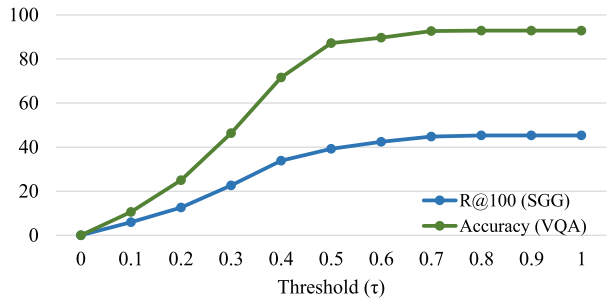


FIGURE 5. The effect of varying cosine similarity threshold on the recall rate in SGG and the overall accuracy of VQA.

adhered to throughout the experimentation phase due to highest accuracy achieved at this threshold.

5) VQA BENCHMARK COMPARISON

The performance of our enriched scene graph-based VQA method is compared to existing baselines, as presented in Table 4. Our framework surpasses the current methods in terms of accuracy for open-ended questions, consistency, and plausibility while maintaining comparable performance in accuracy for binary questions, validity, and distribution. This improvement underscores the benefits of incorporating rich and diverse common sense knowledge into scene graphs, enabling the VQA model to comprehend the high-level semantics of visual concepts in images and questions and provided more accurate and expressive responses to posed questions.

C. QUALITATIVE ANALYSIS

The qualitative results of our framework on three representative images from the GQA dataset [3] are shown in Figure 6. These examples were intentionally selected to include complex, multi-object scenes with overlapping entities, compositional or multi-hop questions, and occlusion or contextual ambiguity, to illustrate how knowledge enrichment enhances the interpretability and reasoning capability of the framework.

The proposed framework effectively enriches the scene graphs, enabling deeper reasoning by providing commonsense and functional background knowledge about the depicted objects beyond their directly observable appearance. This enriched information allows the model to infer implicit relationships that are essential for answering reasoning-intensive questions.

In the first example, multiple interacting objects (man, bag, frisbee) form a scene with high relational density. The enrichment process adds relations such as *used for recreation* and *capable of carrying something*, allowing the system to correctly infer that the bag can carry the frisbee and that the frisbee is used for recreation. This demonstrates the model's ability to reason over functional roles rather than relying purely on visual co-occurrence.

In the second example, the scene contains partial occlusion and multiple correlated wearable items. The enriched triples (e.g., *gloves are part of skiing gear*; *goggles used for protecting eyes*) enable the system to integrate distributed cues—ski, gloves, goggles—and infer a higher-level concept: that the woman is wearing skiing gear and engaging in skiing. This illustrates the framework's capacity for multi-hop compositional reasoning across semantically linked
objects.

In the third example, fine-grained attribute reasoning is required to distinguish between digital and non-digital objects in an indoor workspace. The enriched graph injects background knowledge that monitors and tablets are digital devices, while notebooks are used for writing, not for computation. This context allows the model to answer accurately that the notebook is not digital and confirms the indoor scene type.

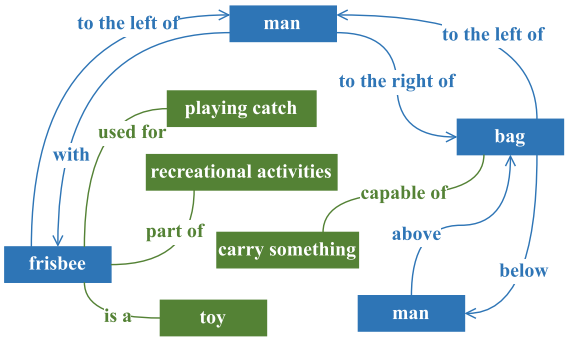
Across all three cases, the enriched scene graphs exhibit denser, semantically coherent relational structures, yielding more accurate and contextually consistent answers. These examples collectively demonstrate that knowledge enrichment not only improves factual correctness but also supports reasoning about object functionality, compositional context, and implicit scene semantics, which are essential for robust visual question answering.

D. LIMITATIONS AND FUTURE DIRECTIONS

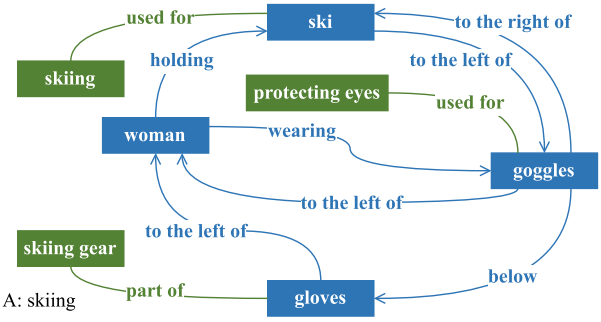
The proposed framework surpasses the VQA baselines and narrows the gap with human-level reasoning performance. However, it still falls short of human performance in terms of validity and plausibility by a small margin, as indicated in Table 4, suggesting potential areas for improvement and highlighting the importance of incorporating even richer knowledge representations and reasoning strategies. Heterogeneous KGs, though extensive sources of common sense knowledge that have improved the overall performance of SGG and VQA, are somewhat constrained by their limited contextual understanding [95]. This becomes a drawback when the KGs lack contextually valid information about a specific scene [15].

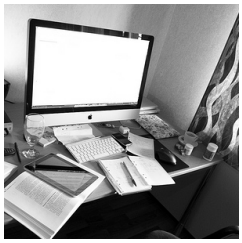
Despite our method outperforming existing SGG methods, this limitation leads to difficulties with fine-grained predicate disambiguation in complex scenes (for example, distinguishing between 'throwing' and 'passing'), indicating the need for more sophisticated relationship reasoning approaches. These limitations in SGG, particularly in visual relationship prediction, directly impact downstream VQA. The inherent complexity and the limited recall rate of SGG ultimately lead to suboptimal performance in VQA, as the ability to accurately interpret and reason about the visual scene is critical for VQA. Therefore, developing more robust and accurate knowledge-based SGG and VQA methods remains an open research problem. Addressing this could significantly enhance visual reasoning capabilities. Moreover, the current

Q1: Do all these people have the same gender? A: yes
Q2: Which object can carry another object? A: bag
Q2: Is there an object that can be used for recreation? A: yes, the frisbee



Q1: What is the woman wearing? A: goggles
Q2: Is the woman wearing a skiing gear? A: yes
Q3: What activity is the woman likely engaging in? A: skiing





Q1: Is the tablet to the right of a notebook? A: no
Q2: Is it indoors? A: yes
Q3: Which of the items on the table are not digital? A: notebook

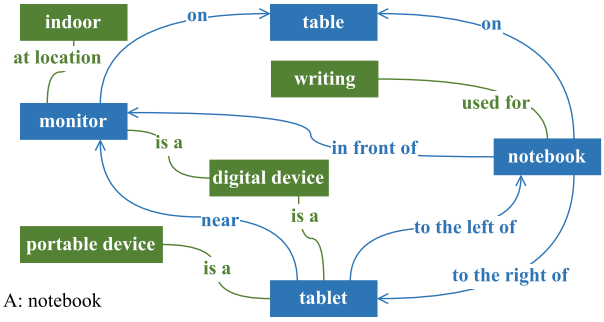


FIGURE 6. Qualitative results of scene graph enrichment and VQA on example images from the GQA dataset [3].

evaluation methods for SGG and VQA do not directly assess the accuracy and relevance of external knowledge. This highlights the need for new evaluation methods that quantify the quality of knowledge infused from external sources.

A stronger neurosymbolic integration [96] between data-centric and knowledge-based modules of the framework could potentially complement the individual strengths of these modules and enable complex visual reasoning capabilities. Incorporating external knowledge into the training regime of deep neural networks has shown promise in enhancing the understanding of visual relationships [97], [98]. While some initial studies have made progress in the context of scene graph generation [25], [64], the potential of incorporating heterogeneous common sense knowledge needs to be explored. Leveraging heterogeneous KGs for rule extraction and integration into neural networks [99], [100] could significantly advance scene understanding and visual reasoning.

Knowledge transfer and distillation techniques [36], [101] present a promising direction in which the knowledge of previously seen visual relationships and questions can be used to guide unseen relationship prediction and answer prediction, enhancing the performance and practicality of SGG and VQA in real-world scenarios. Foundation models [102], with their vast pre-trained knowledge bases and advanced reasoning capabilities, offer extensive world knowledge that can be leveraged to infer missing information and generate hypotheses [103] about unseen or ambiguous parts of a scene, thereby further enriching the scene graph with inferred knowledge that goes beyond the explicit content of KGs. With their structured and semantically rich representation, enriched scene graphs can reciprocally guide foundation models to focus on specific objects and their spatial and relational dynamics within images for improved visual comprehension in generating precise, context-aware responses.

V. CONCLUSION

We proposed the NeSyVQA framework, which generates scene graphs using a multimodal deep learning cascade, enriches the scene graphs with rich common sense knowledge extracted from a heterogeneous KG and employs the enriched scene graphs in an attention-based reasoning network for VQA. NeSyVQA demonstrated significant improvement over the traditional approach lacking heterogenous knowledge enrichment, achieving over 19% higher recall rates in SGG and a 29% increase in accuracy for open-ended questions in VQA. Furthermore, NeSyVQA outperformed the existing state-of-the-art SGG and VQA methods with over 13% higher relationship recall rates in SGG and a 4% higher accuracy on open-ended questions in VQA. The promising results demonstrate the effectiveness of leveraging heterogeneous KGs for complex visual reasoning, paving the way for future research towards more accurate and intuitive VQA systems.

ACKNOWLEDGMENT

This publication has emanated from research conducted with the financial support of Research Ireland under Grant number 18/CRT/6223 and 12/RC/2289_P2. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

REFERENCES

[1] S. Manmadhan and B. C. Kovoor, "Visual question answering: A state-of-the-art review," *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5705–5745, Dec. 2020.

[2] K. Kafle and C. Kanan, "Visual question answering: Datasets, algorithms, and future challenges," *Comput. Vis. Image Understand.*, vol. 163, pp. 3–20, Oct. 2017.

[3] D. A. Hudson and C. D. Manning, "GQA: A new dataset for real-world visual reasoning and compositional question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6693–6702.

[4] D. Gurari, Q. Li, A. J. Stangl, A. Guo, C. Lin, K. Grauman, J. Luo, and J. P. Bigham, "VizWiz grand challenge: Answering visual questions from blind people," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3608–3617.

[5] A. B. Abacha, S. A. Hasan, V. V. Datla, J. Liu, D. Demner-Fushman, and H. Müller, "VQA-med: Overview of the medical visual question answering task at ImageCLEF 2019," in *Proc. CLEF (Work. Notes)*, 2024, pp. 1–11.

[6] D. Li, Z. Zhang, K. Yu, K. Huang, and T. Tan, "ISEE: An intelligent scene exploration and evaluation platform for large-scale visual surveillance," *IEEE Trans. Parallel Distrib. Syst.*, vol. 30, no. 12, pp. 2743–2758, Dec. 2019.

[7] B. He, M. Xia, X. Yu, P. Jian, H. Meng, and Z. Chen, "An educational robot system of visual question answering for preschoolers," in *Proc. 2nd Int. Conf. Robot. Autom. Eng. (ICRAE)*, Dec. 2017, pp. 441–445.

[8] M. Stefanini, M. Cornia, L. Baraldi, M. Corsini, and R. Cucchiara, "Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain," in *Proc. 20th Int. Conf. Image Anal. Process. (ICIAP)*, Trento, Italy. Cham, Switzerland: Springer, 2019, pp. 729–740.

[9] Y. Zhou, S. Mishra, M. Verma, N. Bhamidipati, and W. Wang, "Recommending themes for ad creative design via visual-linguistic representations," in *Proc. Web Conf.*, Apr. 2020, pp. 2521–2527.

[10] D. Teney and A. van den Hengel, "Actively seeking and learning from live data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1940–1949.

[11] 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," 2021, *arXiv:2101.05479*.

[12] P. Hitzler, F. Bianchi, M. Ebrahimi, and M. K. Sarker, "Neural-symbolic integration and the semantic web," *Semantic Web*, vol. 11, no. 1, pp. 3–11, Jan. 2020.

[13] W. Wang, Y. Yang, and F. Wu, "Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing," 2022, *arXiv:2210.15889*.

[14] A. D. Garcez and L. C. Lamb, "Neurosymbolic AI: The 3rd wave," *Artif. Intell. Rev.*, vol. 56, no. 11, pp. 1–20, Nov. 2023.

[15] M. J. Khan, J. G. Breslin, and E. Curry, "Common sense knowledge infusion for visual understanding and reasoning: Approaches, challenges, and applications," *IEEE Internet Comput.*, vol. 26, no. 4, pp. 21–27, Jul. 2022.

[16] M. J. Khan, F. Ilievski, J. G. Breslin, and E. Curry, "A survey of neurosymbolic visual reasoning with scene graphs and common sense knowledge," *Neurosymbolic Artif. Intell.*, vol. 1, pp. 1–24, Mar. 2025.

[17] H. Zhou, Y. Yang, T. Luo, J. Zhang, and S. Li, "A unified deep sparse graph attention network for scene graph generation," *Pattern Recognit.*, vol. 123, Mar. 2022, Art. no. 108367.

[18] T. Chen, W. Yu, R. Chen, and L. Lin, "Knowledge-embedded routing network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6156–6164.

[19] R. Zellers, M. Yatskar, S. Thomson, and Y. Choi, "Neural motifs: Scene graph parsing with global context," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5831–5840.

[20] B. Dai, Y. Zhang, and D. Lin, "Detecting visual relationships with deep relational networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3298–3308.

[21] C. Lu, R. Krishna, M. S. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 852–869.

[22] X. Liang, L. Lee, and E. P. Xing, "Deep variation-structured reinforcement learning for visual relationship and attribute detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4408–4417.

[23] P. Wang, Q. Wu, C. Shen, A. Dick, and A. van den Hengel, "FVQA: Fact-based visual question answering," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 10, pp. 2413–2427, Oct. 2018.

[24] M. Narasimhan and A. G. Schwing, "Straight to the facts: Learning knowledge base retrieval for factual visual question answering," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 460–477.

[25] J. Gu, H. Zhao, Z. Lin, S. Li, J. Cai, and M. Ling, "Scene graph generation with external knowledge and image reconstruction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1969–1978.

[26] Z. Su, C. Zhu, Y. Dong, D. Cai, Y. Chen, and J. Li, "Learning visual knowledge memory networks for visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7736–7745.

[27] A. Zareian, S. Karaman, and S.-F. Chang, "Weakly supervised visual semantic parsing," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3733–3742.

[28] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6857–6866.

[29] C.-W. Lee, W. Fang, C.-K. Yeh, and Y.-C.-F. Wang, "Multi-label zero-shot learning with structured knowledge graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1576–1585.

[30] F. Ilievski, "CSKG: The commonsense knowledge graph," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2022, pp. 680–696.

[31] M. J. Khan, J. G. Breslin, and E. Curry, "Expressive scene graph generation using commonsense knowledge infusion for visual understanding and reasoning," in *Proc. Eur. Semantic Web Conf.* Cham, Switzerland: Springer, 2022, pp. 93–112.

[32] L. Liu, M. Wang, X. He, L. Qing, and H. Chen, "Fact-based visual question answering via dual-process system," *Knowl.-Based Syst.*, vol. 237, Feb. 2022, Art. no. 107650.

[33] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, no. 1, pp. 32–73, May 2017.

[34] T. Trouillon, J. Welbl, S. Riedel, É. Gaussier, and G. Bouchard, "Complex embeddings for simple link prediction," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2016, pp. 2071–2080.

[35] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny, "Large-scale visual relationship understanding," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 9185–9194.

[36] J. Peyre, J. Sivic, I. Laptev, and C. Schmid, "Detecting unseen visual relations using analogies," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1981–1990.

[37] D. Xu, Y. Zhu, C. B. Choy, and L. Fei-Fei, "Scene graph generation by iterative message passing," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3097–3106.

[38] K. Tang, H. Zhang, B. Wu, W. Luo, and W. Liu, "Learning to compose dynamic tree structures for visual contexts," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 6619–6628.

[39] K. Tang, Y. Niu, J. Huang, J. Shi, and H. Zhang, "Unbiased scene graph generation from biased training," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3716–3725.

[40] M. Suhail, A. Mittal, B. Siddiquie, C. Broaddus, J. Eledath, G. Medioni, and L. Sigal, "Energy-based learning for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 13931–13940.

[41] T. He, L. Gao, J. Song, and Y.-F. Li, "Towards open-vocabulary scene graph generation with prompt-based finetuning," in *Proc. 17th Eur. Conf. Comput. Vis. (ECCV)*. Cham, Switzerland: Springer, Oct. 2022, pp. 56–73.

[42] X. Lyu, L. Gao, P. Zeng, H. T. Shen, and J. Song, "Adaptive fine-grained predicates learning for scene graph generation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 11, pp. 13921–13940, Nov. 2023.

[43] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C.-W. Chen, "Boosting scene graph generation with visual relation saliency," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–17, Jan. 2023.

[44] X. Lin, C. Ding, Y. Zhan, Z. Li, and D. Tao, "HL-Net: Heterophily learning network for scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19454–19463.

[45] H. Zhou, J. Zhang, T. Luo, Y. Yang, and J. Lei, "Debiased scene graph generation for dual imbalance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 4, pp. 4274–4288, Apr. 2023.

[46] H. Zhou, T. Luo, J. Zhang, and L. Liu, "Exploring the essence of relationships for scene graph generation via causal features enhancement network," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 8, pp. 6616–6630, Aug. 2025.

[47] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang, "Bottom-up and top-down attention for image captioning and visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.

[48] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," 2019, *arXiv:1908.07490*.

[49] W. Chen, Z. Gan, L. Li, Y. Cheng, W. Wang, and J. Liu, "Meta module network for compositional visual reasoning," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 655–664.

[50] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "MDETR—Modulated detection for end-to-end multi-modal understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 1760–1770.

[51] P. Zhang, X. Li, X. Hu, J. Yang, L. Zhang, L. Wang, Y. Choi, and J. Gao, "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 5575–5584.

[52] D. A. Hudson and C. D. Manning, "Learning by abstraction: The neural state machine," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 5901–5914.

[53] C. Zhang, W.-L. Chao, and D. Xuan, "An empirical study on leveraging scene graphs for visual question answering," 2019, *arXiv:1907.12133*.

[54] Z. Yang, Z. Qin, J. Yu, and T. Wan, "Prior visual relationship reasoning for visual question answering," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2020, pp. 1411–1415.

[55] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp, and S. Günnemann, "Graphhopper: Multi-hop scene graph reasoning for visual question answering," in *Proc. Int. Semantic Web Conf.*, 2021, pp. 111–127.

[56] H. Li, X. Li, B. Karimi, J. Chen, and M. Sun, "Joint learning of object graph and relation graph for visual question answering," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2022, pp. 1–6.

[57] T. Qian, J. Chen, S. Chen, B. Wu, and Y.-G. Jiang, "Scene graph refinement network for visual question answering," *IEEE Trans. Multimedia*, vol. 25, pp. 3950–3961, 2022.

[58] T. Eiter, N. Higuera, J. Oetsch, and M. Pritz, "A neuro-symbolic ASP pipeline for visual question answering," *Theory Pract. Log. Program.*, vol. 22, no. 5, pp. 739–754, Sep. 2022.

[59] K. Yi, J.-J. Wu, C. Gan, A. Torralba, P. Kohli, and J. B. Tenenbaum, "Neural-symbolic VQA: Disentangling reasoning from vision and language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 1–12.

[60] A. Dahlgren and S. Dan, "Compositional generalization in neuro-symbolic visual question answering," in *Proc. Int. Joint Conf. Artif. Intell. Workshop Knowl.-Based Compositional Generalization*, 2023, pp. 1–7.

[61] J. Wu, F. Ge, H. Hong, Y. Shi, Y. Hao, and L. Ma, "Question-aware dynamic scene graph of local semantic representation learning for visual question answering," *Pattern Recognit. Lett.*, vol. 170, pp. 93–99, Jun. 2023.

[62] H. Zhou, T. Luo, and Z. Jiang, "Core-to-global reasoning for compositional visual question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 39, 2025, pp. 10770–10778.

[63] Y. Zhu, J. J. Lim, and L. Fei-Fei, "Knowledge acquisition for visual question answering via iterative querying," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1154–1163.

[64] A. Zareian, S. Karaman, and S.-F. Chang, "Bridging knowledge graphs to generate scene graphs," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2020, pp. 606–623.

[65] Y. Guo, J. Song, L. Gao, and H. T. Shen, "One-shot scene graph generation," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 3090–3098.

[66] H. Tian, N. Xu, M. Kankanhalli, and A.-A. Liu, "Gaussian distribution-aware commonsense knowledge learning for scene graph generation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 12, pp. 13044–13057, Dec. 2024.

[67] R. Li, S. Zhang, B. Wan, and X. He, "Bipartite graph network with adaptive message passing for unbiased scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11104–11114.

[68] H. Tian, N. Xu, Y. Wang, C. Yan, B. Zheng, X. Li, and A.-A. Liu, "Towards confidence-aware commonsense knowledge integration for scene graph generation," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2023, pp. 2255–2260.

[69] M. Ziaeefard and F. Lécué, "Towards knowledge-augmented visual question answering," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 1863–1873.

[70] Y. Wang, M. Yasunaga, H. Ren, S. Wada, and J. Leskovec, "VQA-GNN: Reasoning with multimodal knowledge via graph neural networks for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21525–21535.

[71] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[72] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 936–944.

[73] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.

[74] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 1933–1941.

[75] F. Ilievski, D. Garijo, H. Chalupsky, N. T. Divvala, Y. Yao, C. M. Rogers, R. Li, J. Liu, A. Singh, D. Schwabe, and P. Szekely, "KGTK: A toolkit for large knowledge graph manipulation and analysis," in *Proc. Int. Semantic Web Conf.*, 2020, pp. 278–293.

[76] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," 2017, *arXiv:1710.10903*.

[77] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[78] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2025, pp. 5998–6008.

[79] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997.

[80] J. Johnson, B. Hariharan, L. van der Maaten, L. Fei-Fei, C. L. Zitnick, and R. Girshick, "CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1988–1997.

[81] Y. Goyal, T. Khot, D. Summers-Stay, D. Batra, and D. Parikh, "Making the V in VQA matter: Elevating the role of image understanding in visual question answering," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6325–6334.

[82] K. Marino, M. Rastegari, A. Farhadi, and R. Mottaghi, "OK-VQA: A visual question answering benchmark requiring external knowledge," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3190–3199.

[83] S. Shah, A. Mishra, N. Yadati, and P. Talukdar, "KVQA: Knowledge-aware visual question answering," in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, 2019, pp. 8876–8884.

[84] A. Bordes, N. Usunier, A. García-Durán, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi-relational data," in *Proc. Adv. Neural Inf. Process. Syst.*, 2015, pp. 1–9.

[85] M. Nickel, V. Tresp, and H. Kriegel, "A three-way model for collective learning on multi-relational data," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 809–816.

[86] B. Yang, W.-T. Yih, X. He, J. Gao, and L. Deng, "Embedding entities and relations for learning and inference in knowledge bases," 2014, *arXiv:1412.6575*.

[87] X. Kan, H. Cui, and C. Yang, "Zero-shot scene graph relation prediction through commonsense knowledge integration," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2021, pp. 466–482.

[88] Y. Teng and L. Wang, "Structured sparse R-CNN for direct scene graph generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 19415–19424.

[89] L. Li, J. Xiao, H. Shi, H. Zhang, Y. Yang, W. Liu, and L. Chen, "NICEST: Noisy label correction and training for robust scene graph generation," 2022, *arXiv:2207.13316*.

[90] Y. Li, W. Ouyang, B. Zhou, J. Shi, C. Zhang, and X. Wang, "Factorizable Net: An efficient subgraph-based framework for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 335–351.

[91] Y. Li, W. Ouyang, B. Zhou, K. Wang, and X. Wang, "Scene graph generation from objects, phrases and region captions," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1261–1270.

[92] J. Yang, J. Lu, S. Lee, D. Batra, and D. Parikh, "Graph R-CNN for scene graph generation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 670–685.

[93] R. E. Speer, J. Chin, and C. Havasi, "ConceptNet 5.5: An open multilingual graph of general knowledge," in *Proc. 31st AAAI Conf. Artif. Intell.*, vol. 31, 2017, pp. 4444–4451.

[94] G. A. Miller, "Wordnet: A lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[95] A. Ettorre, A. Bobasheva, C. Faron, and F. Michel, "A systematic approach to identify the information captured by knowledge graph embeddings," in *Proc. IEEE/WIC/ACM Int. Conf. Web Intell.*, Dec. 2021, pp. 617–622.

[96] U. Kursuncu, M. Gaur, and A. Sheth, "Knowledge infused learning (K-IL): Towards deep incorporation of knowledge in deep learning," 2019, *arXiv:1912.00512*.

[97] H. Dai, Y. Tian, B. Dai, S. Skiena, and L. Song, "Syntax-directed variational autoencoder for structured data," 2018, *arXiv:1802.08786*.

[98] M. Allamanis, P. Chanthirasegaran, P. Kohli, and C. Sutton, "Learning continuous semantic representations of symbolic expressions," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 80–88.

[99] M. Nayyeri, C. Xu, M. M. Alam, J. Lehmann, and H. S. Yazdi, "LogicENN: A neural based knowledge graphs embedding model with logical rules," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 7050–7062, Jun. 2023.

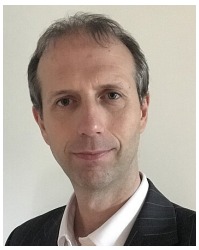
[100] N. Hoernle, R. M. Karampatsis, V. Belle, and K. Gal, "MultiplexNet: Towards fully satisfied logical constraints in neural networks," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, 2022, pp. 5700–5709.

[101] X. Yang, H. Zhang, and J. Cai, "Auto-encoding and distilling scene graphs for image captioning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2313–2327, May 2022.

[102] C. Zhou, Q. Li, C. Li, J. Yu, Y. Liu, G. Wang, K. Zhang, C. Ji, Q. Yan, L. He, H. Peng, J. Li, J. Wu, Z. Liu, P. Xie, C. Xiong, J. Pei, P. S. Yu, and L. Sun, "A comprehensive survey on pretrained foundation models: A history from BERT to ChatGPT," 2023, *arXiv:2302.09419*.

[103] C. Raffel, N. Shazeer, A. P. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2019.

M. JALEED KHAN received the Ph.D. degree in artificial intelligence from the University of Galway, Ireland. He is currently a Lead AI/Data Scientist with Elsewhen, U.K., and an Honorary Research Fellow with the University of Oxford, U.K. With an H-index of 20, his research interests include neurosymbolic and multimodal AI, has resulted in around 50 peer-reviewed publications and several book chapters and open source projects. He is actively involved in the AI research community, as a PC member of major international conferences (ECAI and ECML), a Reviewer of top-tier journals (IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE and IJCV), a Professional Member of ACM and IAPR, and a Grant Panelist for funding bodies (FFG Austria and NCN Poland).



JOHN G. BRESLIN (Senior Member, IEEE) received the bachelor's and Ph.D. degrees in electronic engineering from the University of Galway, in 1994 and 2002, respectively. He is currently an Established Professor in electronic engineering with the University of Galway, where he is also the Director of the TechInnovate and AgInnovate Entrepreneurship Programs. He is a Principal Investigator with the Research Ireland Insight Centre for Data Analytics and the Data2Sustain European Digital Innovation Hub. He has co-authored around 350 publications, including the books *The Social Semantic Web*, *Social Semantic Web Mining*, and *Old Ireland in Colour Trilogy*. He co-created the SIOC framework, implemented in hundreds of applications (by Yahoo, Boeing, and Vodafone) on at least 65 000 websites with 35 million data instances. He is also the Co-Founder of PorterShed, boards.ie and adverts.ie.

EDWARD CURRY is currently an Established Professor of Data Science and the Director of the Data Science Institute and the Insight Research Ireland Centre for Data Analytics, University of Galway. With around 350 publications, he has made substantial contributions to semantic technologies, incremental data management, event processing middleware, software engineering, and distributed systems and information systems. He combines strong theoretical results with high-impact practical applications. The excellence and impact of his research have been acknowledged by numerous awards, including best paper awards and the University of Galway President's Award for Societal Impact, in 2017. His team's technology enables intelligent systems for smart environments in collaboration with several industrial partners. He is an Organizer and the Program Co-Chair of major international conferences, including ESWC 2025, CIKM 2020, ECML 2018, the IEEE Big Data Congress, and European Big Data Value Forum. He is the Co-Founder and the elected Vice President of the Big Data Value Association, an industry-led European big data community, which has built consensus on a joint European big data research and innovation agenda, and influenced European data innovation policy to deliver on the agenda.