



## MuRelSGG: Multimodal relationship prediction for neurosymbolic scene graph generation

Title	MuRelSGG: Multimodal relationship prediction for neurosymbolic scene graph generation
Author(s)	Khan, Muhammad Junaid;Siddiqui, Adil Masood;Khan, Hamid Saeed;Akram, Faisal;Khan, Muhammad Jaleed
Publication Date	2025-03-14
Publisher	Institute of Electrical and Electronics Engineers
Repository DOI	<a href="https://doi.org/10.1109/ACCESS.2025.3551267">https://doi.org/10.1109/ACCESS.2025.3551267</a>

## RESEARCH ARTICLE

# MuRelSGG: Multimodal Relationship Prediction for Neurosymbolic Scene Graph Generation

MUHAMMAD JUNAID KHAN<sup>1</sup>, ADIL MASOOD SIDDIQUI<sup>1</sup>, HAMID SAEED KHAN<sup>2</sup>,  
FAISAL AKRAM<sup>1</sup>, AND M. JALEED KHAN<sup>3</sup>

<sup>1</sup>Department of Electrical Engineering, Military College of Signals, National University of Sciences and Technology, Islamabad 46000, Pakistan

<sup>2</sup>siParadigm Diagnostic Informatics, Montville, NJ 07058, USA

<sup>3</sup>Data Science Institute, University of Galway, Galway, H91 TK33 Ireland

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

**ABSTRACT** Neurosymbolic Scene Graph Generation (SGG) is a promising approach that jointly leverages the perception capabilities of deep neural networks and the reasoning capabilities of symbolic techniques for scene understanding and visual reasoning. SGG systematically captures semantic components, including objects and their relationships, in images, enabling structured representations of visual data. However, existing SGG methods exhibit constrained accuracy and limited expressiveness, particularly in long-tail relationship prediction. To address these limitations, we present MuRelSGG, a novel neurosymbolic SGG framework that integrates a Transformer-based multimodal relationship prediction pipeline with common sense knowledge enrichment. This synergistic combination encapsulates global context, long-range dependencies, and complex object interactions to enhance relationship prediction in SGG. The proposed neurosymbolic architecture begins with object detection via Faster R-CNN, followed by a cascade of Multi-Head Attention Transformers (M-HAT) and Vision Transformers (ViT) for relationship prediction. Subsequently, CSKG enrichment refines and augments visual relationships, improving both accuracy and expressiveness. We conduct extensive evaluations on both the Visual Genome (VG) and GQA datasets to assess performance and generalizability. MuRelSGG achieves substantial gains in recall rates (VG:  $R@100 = 43.2$ ,  $mR@100 = 14.9$ ; GQA:  $R@100 = 42.1$ ), outperforming state-of-the-art SGG techniques. Ablation studies confirm the critical contributions of M-HAT, ViT, linguistic features, CSKG enrichment and embedding similarity thresholds, demonstrating the effectiveness of structured knowledge integration for long-tail relationship prediction. These findings underscore the potential of combining deep learning architectures with structured knowledge bases to advance visual scene representation and reasoning.

**INDEX TERMS** Common sense knowledge, neurosymbolic integration, scene graph, scene understanding, visual reasoning.

## I. INTRODUCTION

Over the last decade, visual intelligence has advanced significantly with deep learning, improving image classification, object recognition, and segmentation. These advancements now enable fine-grained scene understanding, such as identifying specific objects in crowded scenes. However, true scene comprehension requires more than just object detection—it necessitates the ability to reason about complex interactions,

relationships, and contextual dependencies. For example, recognizing a family picnic involves not only detecting objects (e.g., people, food, blankets) but also understanding their relationships (e.g., “mother serving child,” “dog eyeing sandwich”).

There is an increasing focus on neurosymbolic and knowledge-driven approaches for visual understanding and reasoning to address semantic and relational gaps. Neurosymbolic integration endeavours to amalgamate the profound learning capabilities and generalizability inherent in neural methodologies with the reasoning proficiency and

The associate editor coordinating the review of this manuscript and approving it for publication was Bhaskar P. Rimal.

interpretability characteristic of symbolic methodologies [1]. Such hybrid paradigms serve to alleviate the particular limitations associated with each methodology while simultaneously leveraging the unique advantages of both to enhance their scope and utility. For instance, structured knowledge bases and symbolic reasoning are instrumental in augmenting the operational efficacy of black-box neural networks [2]. Conversely, the facilitation of extensive symbolic reasoning and the completion of knowledge bases are rendered feasible through the application of neural networks and machine learning techniques [3]. The efficacy of tasks related to visual comprehension and reasoning is contingent upon the quality of image representation. As a result, numerous initiatives have been undertaken to systematically capture visual attributes and the interrelationships among objects. The scene graph, which systematically organizes objects along with their interrelations in a coherent, semantic format, has emerged as a prominent symbolic representation for visual data [4].

imperative for a multitude of subsequent visual reasoning tasks that necessitate an elevated comprehension and interpretation of the contents and context of images. SGG's downstream tasks include image captioning [6], visual question answering (VQA) [7], image retrieval [8], robot task planning [9], human action recognition [10], context-aware augmented reality [11], and multimedia event processing (MEP) [12].

Scene graphs are centred on visual relationships, and the ability to accurately forecast these relationships is essential to both visual reasoning and scene understanding. Large-scale labelled datasets are used by the majority of data-driven SGG approaches to train models for relationship prediction and object detection. Although object detection accuracy is high, visual relationship prediction is still complicated because there are so many potential associations, and they might appear and be interpreted differently depending on the environment and context. This is manifested as a long-tailed distribution of relationships and a higher frequency of generic relationships relative to significant ones in datasets such as Visual Genome [13], [14]. Numerous relationship predicates, especially those that are relevant (such as “part of,” “between,” “made of,” and “looking at,” as seen in Fig 3), have few appearances in datasets, which makes it more difficult for existing SGG algorithms to learn their feature representations. Furthermore, visual characteristics of relationships can differ significantly between settings and scenes (e.g. As shown in Fig 2 “Riding Bicycle,” “Riding Elephant,” and “Riding Surfboard” all have distinct visual characteristics despite having the same relationship predicate).

It is almost impossible to get enough training examples for every potential combination of visual relationships. While data-centric SGG approaches perform well in object detection and well-represented relationship prediction, they are less effective in rare or under-represented relationship prediction [15]. Another major worry is the robustness of SGG, which is its ability to function consistently in both familiar and new settings and independent of the frequency of visual correlations in datasets. To get beyond these challenges, a lot of work has been done. New aspects of visual associations in images, such as heterophily [16] and saliency [17], have been explored. Advanced methods like knowledge transfer [18], language supervision [19], and zero-shot learning [20] have also been used.

In particular, the infusion of common sense knowledge (CSK) has emerged as a promising approach to address these issues. In SGG, language and statistical priors have been widely utilised as sources of CSK; however, the performance of language priors is impacted by the limitations of semantic word embeddings, especially when dealing with rare or unknown relationships, and the statistical priors' heuristics do not generalise well [22]. Based on their breadth and focus, different KGs have been used in SGG to capture different kinds of common sense information. For example, ConceptNet [23], ATOMIC [24], Visual Genome [13], and

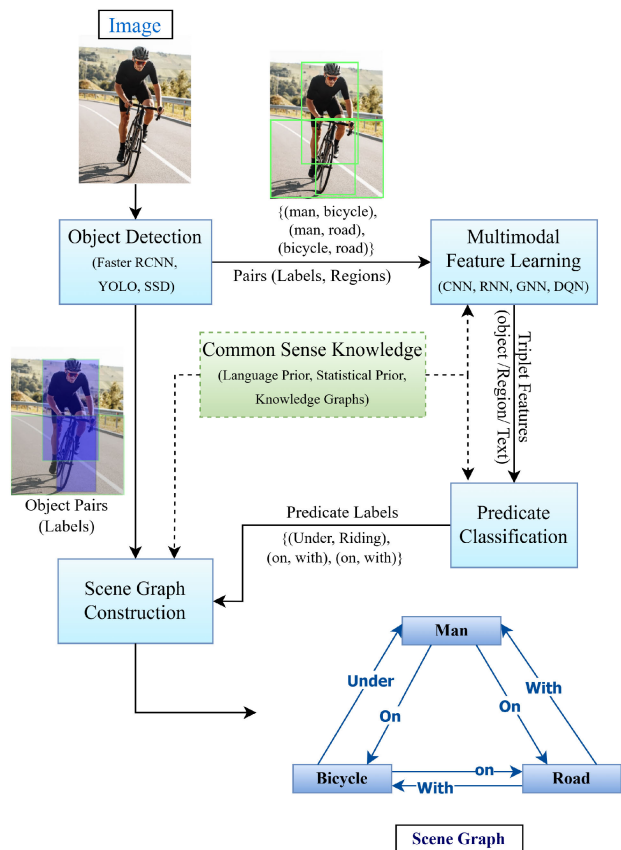
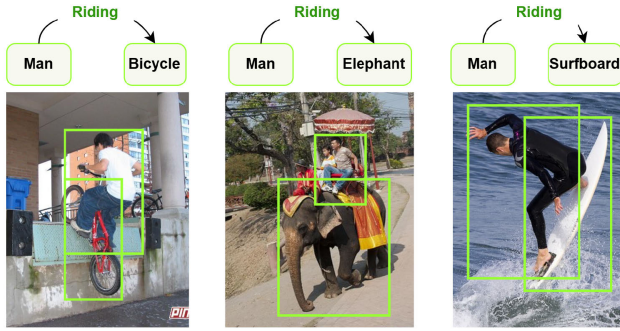


FIGURE 1. An illustration of the general procedure for creating a scene graph, starting at the top left and ending at the bottom right.

Scene Graph Generation (SGG), as illustrated in Fig 1, encompasses the identification and localization of entities within an image through deep learning-based detection and classification methodologies, succeeded by the prediction of visual relationships utilizing visual-linguistic multimodal strategies [5]. The semantic representation of images is



**FIGURE 2.** Same relationship predicate used in different appearances in VG dataset.

WordNet [25] provide general, procedural, visual, and lexical knowledge, respectively. A heterogeneous knowledge graph (KG), like the Common Sense Knowledge Graph (CSKG) [26], unifies and richly represents CSK by combining the distinct knowledge from several KGs. In semi-supervised expressive SGG [27] and generalised common sense question answering [28], CSKG has demonstrated efficacy. Nevertheless, they are still underappreciated in terms of improving scene representations for visual thinking.

The proposed methodology blends multimodal vision-language approaches with neurosymbolic techniques for SGG. First, the technique uses Faster RCNN for detecting objects, with ResNeXt-101-FPN serving as its foundation. Following object detection, contextual features are extracted from the image and object interactions are anticipated by utilising a cascade of Multi-Head Attention Transformers (M-HAT) and Vision Transformers (ViT). To improve the expressiveness of the resulting scene graph and address any potential errors, the approach incorporates background knowledge from the CSKG in the form of triples. The scene graph is filtered and enhanced during the refinement phase by adding pertinent triplets from CSKG, which increases the final representation's accuracy and comprehensiveness. By capturing the scene's intricate object interactions, long-range dependencies, and global context, this method produces an SGG framework that is more resilient and adaptable. Using the benchmark Visual Genome (VG) dataset, we assessed the proposed neurosymbolic and vision-language multimodal method for SGG and found substantial enhancements in relationship prediction performance. Ablation studies demonstrated the critical contributions of M-HAT and ViT, with their combined usage leading to the highest performance ( $R@100 = 43.2$ ), whereas removing either component resulted in a performance drop. In addition, the ablation studies highlighted the significance of linguistic features (leading to an improvement of 2.8 in  $R@100$ ), and the importance of carefully chosen embedding similarity thresholds. The positive experimental results highlight the effectiveness of combining common sense knowledge with advanced deep learning approaches for visual perception and reasoning. Moreover, the integration of multimodal

large language models (LLMs) with this framework offers promising potential for downstream reasoning tasks, such as generating not only fluent but also semantically rich captions for images or providing insightful answers to complex, open-ended questions about visual content. By leveraging enriched scene graphs, multimodal LLMs can achieve a deeper comprehension and more nuanced understanding of visual scenes, further enhancing their ability to articulate intricate relationships and contextual subtleties within the data. Major contributions of this paper include:

- 1) We present a novel neurosymbolic Transformer-based SGG methodology (illustrated in Figure 4 and Algorithm 1) that integrates multimodal relationship prediction using transformers with knowledge-based refinement and enrichment via CSKG [26]. This combination leverages neural perception for scene understanding while incorporating structured common sense knowledge for enhanced relational reasoning, leading to more expressive and context-aware scene graphs. This methodology demonstrates the potential of synergizing vision transformers with knowledge graphs to improve visual reasoning capabilities beyond conventional SGG models.
- 2) We provide extensive empirical validation of the proposed approach through rigorous comparative analysis on multiple datasets, including the Visual Genome (VG) dataset [13] and Generalized Question Answering (GQA) dataset [29]. Our method consistently outperforms data-centric baselines, as demonstrated by higher recall rates ( $R@100 = 43.2$  on VG, 44.5 on GQA) (Table 1 and Table 2). In addition, the improved mean recall ( $mR@100 = 14.1$ ) highlights the ability of our approach to mitigate data bias and perform consistently across frequent and infrequent relationships in the dataset. All comparative experiments were conducted under standard conditions, ensuring fair evaluation.
- 3) We perform comprehensive ablation studies to analyze the impact of key components in our framework. The results confirm that the joint utilization of M-HAT and ViT leads to superior scene graph representations (Table 3). The visual-linguistic multimodal feature representation significantly outperformed visual features only (Table 6). Our findings also demonstrate that CSKG-based knowledge enrichment substantially improves the prediction of long-tail relationships, outperforming common alternatives like statistical priors and ConceptNet (Table 5). These insights establish a novel contribution in structured knowledge-enhanced relationship prediction, which has not been comprehensively explored in prior SGG research.

## II. LITERATURE REVIEW

### A. SCENE GRAPH GENERATION

The scene graph constitutes a systematic representation of an image, incorporating comprehensive semantic data

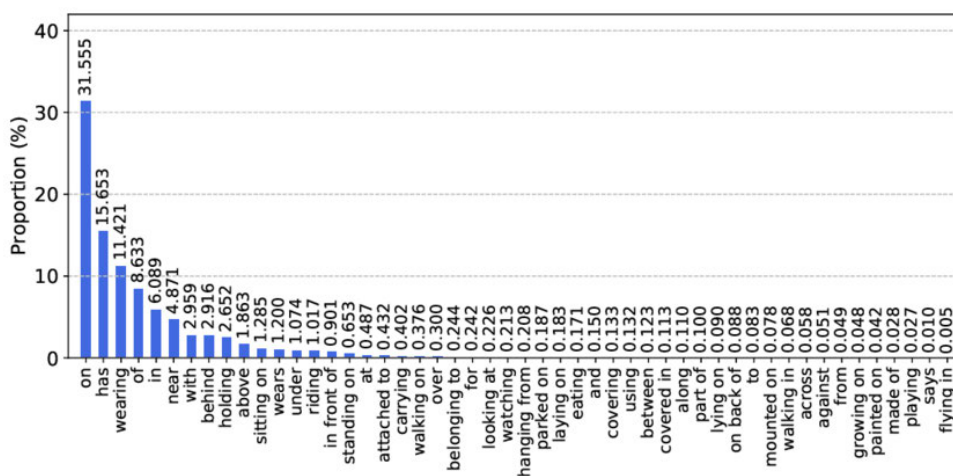


FIGURE 3. Long tailed distribution of relationship predicates in visual genome [21].

regarding a visual scenario, which encompasses objects, their attributes, and the visual interrelationships among them. The methodologies employed in SGG typically adhere to a methodical framework, as illustrated in Figure 1, wherein entities present within an image are recognized through the application of Deep Neural Network (DNN)-driven object detection methodologies, and the pairwise relationships between these entities are deduced via DNN-based vision-language hybrid features. The categorization of object pairs, along with their associated relational predicates, is fundamental to the construction of the symbolic scene graph relevant to the visual representation.

A particularly formidable challenge in SGG pertains to the forecasting of pairwise visual relationships among objects, primarily attributable to the pronounced imbalance within training datasets concerning relationship predicates (refer to Figure 3), the significantly diverse visual feature representation of these relationships across varying scenes (as depicted in Figure 2), and the inadequacy of training samples concerning the extensive array of potential triplet combinations, all of which severely constrains the robustness, expressiveness and accuracy of SGG methodologies. The prevailing shortcomings of SGG, in conjunction with its potential applications across various visual reasoning tasks, have fostered substantial academic interest within the field of visual intelligence research [5]. The two evident approaches for SGG include Data Centric methodologies, which utilizes the model's ability to infer and represent the relationships between objects directly from the data and the Neuro-symbolic integration that aims to enhance the neural capability of DNNs with logical reasoning of symbolic approaches through the use of external CSK [30], [31].

### B. DATA-CENTRIC SGG METHODS

Enhancing scene knowledge through the quality, diversity, and richness of the data used to train models is the main goal of data-centric SGG approaches. By using these techniques,

the model tries to deduce and depict the relationships between items from the data itself. These methods focus on increasing the accuracy and robustness of SGG by utilising the underlying structure and patterns within the data. Li et al. presented MSDN [32], which aligns object, phrase, and caption areas with a dynamic graph based on spatial and semantic relationships, therefore simultaneously solving object identification, SGG, and region captioning tasks in an end-to-end way. Direct SGG without explicit object detection has been proposed in SS-RCNN [33]. The model makes use of a structured triplet detector together with learnable triplet queries. To enhance training difficulty, it makes use of a relaxed and improved training technique based on knowledge distillation from a Siamese Sparse R-CNN.

Many Data-Centric methods have focused on refining message passing and relationship filtering using varying techniques. For example, Yang et al.'s [34] model prunes unlikely relationships between objects and captures contextual information within the scene graph using the Relation Proposal Network (RePN) and attentional Graph Convolutional Network (aGCN), respectively. A subgraph-based representation was used by FactorizableNet [35] to minimise the number of intermediate representations needed for inference. Their approach includes a Spatial-weighted Message-passing structure and a Spatial-sensitive Relation Inference module to use the spatial structure of the feature maps. Additionally, it makes use of 2-D feature maps to maintain spatial data inside the subgraph area. Zhang et al. [17] proposed visual Saliency-guided Message Passing to enhance scene graphs' capacity for relationship reasoning and generalisability by utilising ordinal regression to isolate the most relevant visual relationships. To improve message passing and refine relationship representation in a Graph Neural Network (GNN), Lin et al. [16] utilized the principle of heterophily within visual relationships, in combination with an adaptive reweighting transformer module, to facilitate the integration of information across various layers. Xu et al.

[36] introduced a recursive message-passing methodology that leverages conventional recurrent neural networks (RNN) in conjunction with contextual indicators for a unified inference mechanism for visual relationship forecasting in SGG.

Techniques focusing on countering the bias in data include NICEST [37], which addresses noisy label concerns through the generation of high-quality samples and by incorporating NICE and NIST elements. NICE encompasses the identification of noisy samples and the reassignment of soft predicate labels. NIST employs a multi-teacher knowledge distillation approach for unbiased knowledge fusion. Transformers have proven to be effective in SGG as they evaluate the relationships between all elements within an image and provide a more global representation of the scene than CNNs. Additionally, transformer models better capture contextual information and long-range dependencies. EGTR [38] utilized DCNN and transformer in a one-stage SGG model to generate SG from multi-head self-attention by-products derived from the object detector. With ResNet-50 and Swin Transformer, DSGG [39] approaches scene graph detection as a direct graph prediction problem transformer network that employs graph-aware queries to get a compact representation of the node and its relationships in the graph. In RepSGG [40], novel representations of entities and relationships for SGG and performance-guided logit adjustment strategy for longtailed learning are proposed. The majority of previous research has concentrated on the linguistic and visual patterns found in images, largely ignoring the contextual details and associated facts regarding the concepts found in images as well as the structural patterns of scene graph elements in commonsense knowledge graphs, which hold great promise for aiding in the comprehension and interpretation of visual concepts.

### C. COMMON SENSE KNOWLEDGE (CSK) BASED SGG METHODS

SGG is a challenging endeavour because of the large semantic space of interactions. This semantic space encompasses all potential relationships among objects in a scene. It includes both simple and complex relationships, demonstrating significant variability. Capturing all these relationships in a finite dataset proves challenging. The integration of CSK, reflecting human understanding, is therefore essential. CSK consists of foundational information and reasoning employed by humans [41]. In SGG, this involves recognizing typical associations, such as birds in trees rather than fish. Incorporating CSK into SGG enhances the representation of relationships in scenes. This integration facilitates a more accurate depiction despite limitations in training data. CSK sources for SGG can be categorized into statistical and language priors, and KG [30], [31].

The kind of CSK that depends on the structural regularities and statistical correlations seen in visual scenarios is

Statistical Priors. MotifNet [42] uses statistical priors to capture interdependence between objects and relationships in visual scenes. The SGG model categorises graph probability into three elements: bounding boxes followed by objects and then relations without making any independent assumptions. The DRM [43] framework utilizes statistical priors through the calibration of distributions associated with predicate and triplet features, which is informed by the classifications of head predicates, thus augmenting the representational patterns of tail predicates. The intrinsic and extrinsic CFA [44] modules augment the diversity of features by drawing upon similarities in patterns and contexts, thereby leveraging the statistical characteristics inherent in the data. Hierarchical clustering methodologies are employed to ascertain appropriate substitutions informed by statistical patterns.

Language priors leverage the semantic info inherent in lexical items to augment the prediction of relational dynamics. They facilitate the recognition of visual relationships by analyzing objects that exhibit semantic correlation. The VRD model [13] discerns visual relations in an image by capitalizing on both visual characteristics and linguistic components. The linguistic module utilizes pre-trained word vectors (word2vec) to map relationships onto an embedding space in which semantically analogous relationships are positioned in close proximity to one another. This capability enables the model to deduce less prevalent relationships based on more frequent, analogous ones, thereby effectively harnessing language priors as a repository of commonsense knowledge. According to DeepVRL [45], visual relationship and attribute recognition is a step-by-step method that uses linguistic priors to gradually reveal the object attributes and relationships in an image. It builds a directed semantic action graph that encapsulates the semantic links across object classes, attributes, and predicates using linguistic priors as the cornerstone.

KGs are formal databases that represent real-world entities and their relationships. KGs are crucial for providing structured information. In SGG, KGs provide CSK for improved accuracy. For instance, a KG may state relationships like “birds are often in trees,” allowing SGG systems to deduce these relationships in images even if such instances are not found in the training dataset. Incorporating KGs into SGG systems enhances contextual understanding and performance in scene interpretation. NeuSyRE [15] employs a DNN-based approach for the purposes of object detection and the establishment of pairwise associations. Subsequently, commonsense knowledge is integrated into the scene graph through the utilization of the CSKG [26] to incorporate relevant facts and contextual information represented as triplets. The regional characteristics and visual contextual attributes of objects are encoded through the application of Bi-LSTM networks and are subsequently amalgamated into a cohesive set of pairwise object features. A deep sparse graph attention network (DSGAN) for SGG was developed by Zhou et al. [46]. It learns object and predicate characteristics

and generates a sparse KG representation using statistical co-occurrence data.

The IRT-MSK [47] methodology employed GCNs and integrated a diverse array of structured knowledge repositories, with a particular emphasis on relational and commonsense knowledge, to encode intuitive understanding and to elucidate the interconnections between various entities. In pursuit of modelling the neighborhoods and pathways associated with items within ConceptNet and integrating them into the SGG architecture, COACHER [48] employed graph mining methodologies. COACHER capitalizes on ConceptNet to furnish embeddings of common-sense knowledge, which are subsequently harnessed to enhance zero-shot relation prediction. KnowZRel [49] leveraged a common sense knowledge-based approach for object refinement and zero-shot relationship retrieval for generalised scene graph generation.

The most recent and comprehensive consolidated dataset that combines commonsense info from seven different sources is the CSKG [26]. These sources include ConceptNet [23], Wikidata [50], ATOMIC [24], VG [13], Wordnet [25], Roget [51], and FrameNet [52]. CSKG possesses the capacity to significantly improve visual comprehension and reasoning activities. Nevertheless, it has yet to attract the requisite scholarly attention.

### III. PROPOSED METHOD

The proposed SGG methodology utilises a neurosymbolic framework alongside a vision-language multimodal strategy, integrating Faster R-CNN for the purpose of object detection, a transformer-based deep learning cascade for the prediction of pairwise visual relationships, and knowledge-driven refinement and enrichment techniques to construct a scene graph derived from an image. The neural and symbolic elements of the proposed framework are characterized by loose coupling, which facilitates the independent functioning and modification of each component without imposing alterations on the others, thereby promoting adaptability. Algorithm 1 presents the entire process of our method.

A detailed overview of the proposed methodology is provided in Figure 4. Given an input image, the first step involves Object detection & feature map extraction. Faster RCNN framework [53] was employed for the task of object detection. The foundational feature extractor for the Faster RCNN is the ResNeXt-101-FPN architecture [54]. For every recognised item in an input image  $I$ , the Faster RCNN produces object bounding boxes  $b$  and matching object class labels  $l$ . Furthermore, feature maps  $F$  are taken out of the CNN used as a foundation for the Faster RCNN and used for further processing to capture features relevant to a particular location.

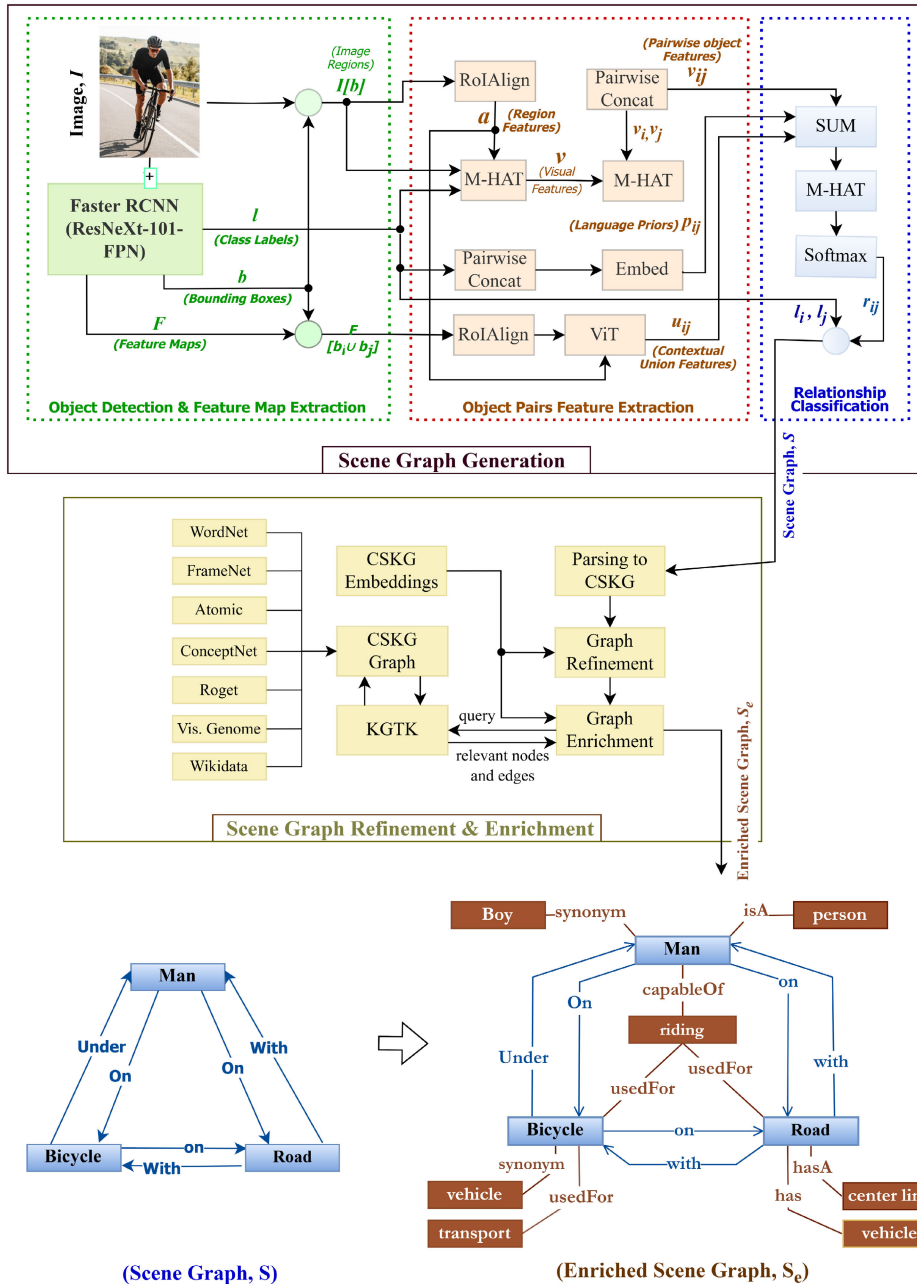
The second step includes relationship prediction between object pairs. Using RoIAlign [55], the region features  $a$  for every detected object are calculated and then applied to the image regions  $I[b]$  that have been cropped utilising object bounding boxes. These region features were processed by

a Vision Transformer (ViT) block [56] to extract contextual union features  $(u_{ij}|i \neq j; i, j = 1, \dots, n)$ . We leveraged the ViT's capability to capture global context by treating image patches as tokens and applying self-attention across the entire image, which results in a richer, more holistic representation of the regions. Following this, the visual context features  $v$  for each detected object are encoded with Multi-Head Attention Transformer (M-HAT) blocks [57]. This allows the model to better capture complex dependencies and interactions between objects and encode contextual information through parallel processing and self-attention mechanisms. For predicting relationships between object pairs, the combined pairwise object features  $v_{ij}|i \neq j; i, j = 1, \dots, n$  are obtained through encoding individual visual context features  $(v_i, v_j)$  using another Multi-Head Attention Transformer (M-HAT) block [57]. Further to compute the language prior  $p_{ij}$ , the pairwise object labels  $(l_i, l_j)$  are encoded through an embedding layer.

In the next stage, the three categories of extracted features of the object pairs, i.e.,  $v_{ij}$ ,  $p_{ij}$ , and  $u_{ij}$ , are fused using a summation function [58]. To further refine the fused features before relationship classification, we introduce an additional Multi-Head Attention Transformer block after the summation. This Transformer block re-attends to the fused features, enabling the model to dynamically focus on the most relevant aspects of the combined information, thereby improving the accuracy of the relationship prediction. Afterwards, a softmax function is employed to predict the relationship class labels  $r_{ij}$  and probabilities  $c_{ij}$ . Ultimately, the first scene graph  $S$  is created by connecting the pairwise objects and relationship predicates into a structured representation. The proposed method effectively leverages attention-based transformers at various stages of the SGG pipeline to capture global context, long-range dependencies, and complex interactions between objects.

Upon the creation of the preliminary scene graph  $S$  utilizing Transformer-based modules for the prediction of relationships, the scene graph representation encapsulates the entities present in an image along with their visual interrelations. Nonetheless, this representation can contain less relevant or out-of-context information, and it can be deficient in the requisite expressiveness necessary for thorough visual comprehension, as it might not adequately reflect the meanings and connections among the objects. We employ refinement and enrichment strategies [27] to the probabilistic representations generated in the relationship prediction process to mitigate these shortcomings. The predicted triplets (subject, predicate, object) are assigned confidence scores, and only those exceeding a predefined threshold are retained. The refined scene graphs are then augmented by incorporating common sense knowledge, thereby enhancing their expressiveness and rectifying any inaccuracy in predicted visual relationships.

To achieve this objective, we employ CSKG [26], which supplies contextual knowledge and pertinent information in the format of triplets. We parse the scene graph to a



**FIGURE 4.** Proposed framework - By Integrating neurosymbolic reasoning with multimodal transformers, our proposed method enhances scene graph generation by combining object detection, relationship prediction, and knowledge-driven enrichment for comprehensive visual understanding.

format that is in line with the CSKG data model in order to incorporate pertinent triplets that reflect background information and related facts from the CSKG. For this purpose, we utilize graph embeddings to assess the similarity among nodes during the processes of refinement and enrichment, as entities that are alike frequently exhibit comparable vector representations within the embedding space. To mitigate prediction inaccuracies and resolve any overlapping objects and conflicting relationships, a two-step filtering approach is applied: (1) bounding box filtering using an Intersection over

Union (IoU) threshold to discard redundant object detections, and (2) graph refinement based on embedding similarity to merge semantically similar entities. Any two bounding boxes with IoU exceeding the threshold are considered redundant, and the less confident detection is removed. If two objects have high embedding similarity, they are merged into a single entity. This prevents duplication in the scene graph and ensures more compact and accurate representations.

Subsequent to the refinement phase, the scene graph undergoes enrichment through queries (using Knowledge

**Algorithm 1** Proposed Method**Require:** Image  $I$ **Ensure:** Enriched Scene Graph  $S_e$ 

```

1:  $\{b, l, F\} = \text{FasterRCNN}(I)$ 
2: for each bounding box  $b_i$  do
3:    $a_i = \text{RoIAlign}(I[b_i])$ 
4: end for
5: for each pair of objects  $(i, j)$  do
6:    $u_{ij} = \text{ViT}(\text{RoIAlign}(F[b_i \cup b_j]))$ 
7: end for
8: for each object  $i$  do
9:    $v_i = \text{Trans}(a_i, I[b_i], l_i)$ 
10: end for
11: for each pair of objects  $(i, j)$  do
12:    $v_{ij} = \text{Trans}(\text{concat}(v_i, v_j))$ 
13:    $p_{ij} = \text{embed}(\text{concat}(l_i, l_j))$ 
14: end for
15:  $f = \text{SUM}(v_{ij}, p_{ij}, u_{ij})$ 
16:  $f' = \text{Trans}(f)$ 
17:  $\{r_{ij}, c_{ij}\} = \text{softmax}(f')$ 
18:  $S = \{l_i, r_{ij}, l_j\}$ 
19:  $S_r = \{\}$ 
20: for each triplet  $t \in S$  do
21:    $e_1 = \text{csvg\_emb}(t[\text{nodeI}])$ 
22:    $e_2 = \text{csvg\_emb}(t[\text{nodeII}])$ 
23:    $b_1 = b[t[\text{nodeI}]]$ 
24:    $b_2 = b[t[\text{nodeII}]]$ 
25:    $\text{metric}_{\text{sim}} = \text{cosine\_sim}(e_1, e_2)$ 
26:    $\text{metric}_{\text{IoU}} = \text{compute\_IoU}(b_1, b_2)$ 
27:   if  $\text{metric}_{\text{sim}} \leq \tau_{\text{sim}} \wedge \text{metric}_{\text{IoU}} \leq \tau_{\text{iou}}$  then
28:      $S_r.\text{append}(t)$ 
29:   end if
30: end for
31:  $S_e = S_r$ 
32: for each node  $e \in S_r$  do
33:    $e_1 = \text{csvg\_emb}(e)$ 
34:    $t_{\text{csvg}} = \text{query}(G_{\text{csvg}}, e)$ 
35:   for each triplet  $t_{\text{csvg}}$  do
36:     if  $e == t[\text{nodeI}]$  then
37:        $e_2 = \text{csvg\_emb}(t[\text{nodeII}])$ 
38:     else
39:        $e_2 = \text{csvg\_emb}(t[\text{nodeI}])$ 
40:     end if
41:      $s = \text{cosine\_sim}(e_1, e_2)$ 
42:     if  $s > \tau \wedge t \notin S_e$  then
43:        $S_e.\text{append}(t)$ 
44:     end if
45:   end for
46: end for

```

Graph Toolkit (KGTK) [59] to CSKG to retrieve additional triplets that encompass either a subject or an object node present in the predicted scene graph. Duplicate triplets and those with identical nodes (e.g. (boy, synonym, boy) and (pizza, similarTo, pizza)) are eliminated in the preprocessing

step as they lack informative value, while extracted nodes exhibiting significant structural similarity to corresponding object nodes are interconnected through edges in the scene graph based on embedding similarity. If a node extracted is already in the scene graph, the new edge connects to it; if not, a new node is created and integrated. This enrichment step adds relevant background knowledge and relationships to the scene graph, making the representation more comprehensive and expressive.

To address polysemous concepts, our model leverages contextual information through graph embeddings and cosine similarity computations. When a node representing a concept (e.g., “bank”) is detected in a scene graph, our method retrieves the top candidate matches from CSKG based on similarity scores. This ensures that the correct meaning (e.g., “river bank” vs. “financial bank”) is selected based on contextual relevance. Additionally, linguistic features from object labels and relationship predicates play a key role in resolving ambiguity. Multi-head attention transformers process these linguistic cues alongside visual embeddings, helping to align polysemous terms with their appropriate meanings in the given scene context.

As illustrated in Figure 4, the initial Scene Graph  $S$  (depicted in blue) conveys details regarding the objects and their interrelationships within the scene. The pertinent nodes and edges derived from the CSKG (represented in brown) augment and enhance the scene graph by supplying essential information concerning the potential spatial relationships between objects and any conceivable interactions among them, such as (man, on, bicycle) and (man, on, road). More critically, it also includes contextual information and associated facts, for instance, (man, capableOf, riding), (bicycle, usedFor, riding) and (road, usedFor, riding), which facilitates advanced reasoning to infer that “the man is riding a bicycle on the road.” In the postprocessing, the enriched scene graph’s ( $S_e$ ) format is modified to align with the original representation for performance assessment or downstream reasoning applications.

#### IV. EXPERIMENTS AND RESULTS

The experiments were performed using an AWS EC2 instance equipped with a Tesla T4 GPU, 16 vCPUs, and 64GB RAM. The Visual Genome (VG) dataset [13] was used for a thorough evaluation and comparison of the proposed method. The VG dataset contains 108K labeled images with annotations for objects and their visual relationships, making it a comprehensive benchmark for SGG. The GQA dataset [29] is included to assess the generalizability of the proposed method beyond VG. To ensure a fair comparison, we followed the same experimental settings, dataset splits, and evaluation protocols as used in prior work, including the baseline methods. For training and evaluation, the standard subset [36] of VG was used, which consists of the 150 object classes and the 50 most common predicate classes. A training split of 70% was employed, with 5000 samples reserved for validation and the remaining 30% for testing. Addition-

**TABLE 1. Thorough comparison of the proposed approach with recent data-centric SGG approaches utilising standard split of the VG dataset and common evaluation metrics (R@K and mR@K).**

SGG Technique	Approach	SGG Results	
		R@20/50/100(%)	mR@20/50/100(%)
<b>Proposed Method</b>	Transformer-based Multimodal SGG	<b>27.9/35.2/41.7</b>	<b>8.3/11.5/14.1</b>
EGTR [38]	Transformer based one-stage SGG model	23.5/30.2/34.3	5.5/7.9/10.1
RepSGG [40]	performance-guided logit adjustment	22.5/29.6/34.8	6.7/9.3/11.4
HL-Net [16]	Heterophily Learning based on Transformer & MP	26.0/33.7/38.1	-/-/9.2
TDE [63]	Causal Inference and Total Direct Effect	25.8/33.3/37.8	6.9/9.3/11.1
SS-RCNN [33]	Structured Sparse R-CNN	25.8/32.7/36.9	6.1/8.4/10.0
SMP [17]	Vision-Based Relation Saliency-Informed Message Passing	-/32.6/36.9	-/-/-
NeuSyRE (w/o enrichment) [15]	Fusion of Visual-Textual Features underpins SGG	26.1/32.7/36.5	7.9/10.0/11.7
NICEST [37]	Noisy label correction & training for Robust SGG	-/29.0/32.7	-/10.4/12.4
VCTree [61]	Dynamic tree structures and Bi-dir TreeLSTM	22.0/27.9/31.3	5.2/6.9/8.0
IMP+ [36]	Object & relationship feature refinement via MP	14.6/20.7/24.5	-/3.8/4.8
FactorizableNet [35]	Clustering-based graph factorization	-/13.1/16.5	-/-/-
MSDN [32]	Description of scene at the object, phrase, and caption levels	-/10.7/14.2	-/-/-
Graph RCNN [34]	RPN followed by Attention GCN	-/11.4/13.7	-/-/-

ally, pre-trained CSKG embeddings [26] were utilized for computing node similarity during the graph refinement & enrichment.

The Scene Graph Detection (SGDet) setting of SGG was adopted, where models must detect objects and predict relationships without access to ground-truth object annotations. This ensures a consistent evaluation of all methods under identical constraints. We report mean Average Precision (mAP) for object detection and Recall@K (R@K) and mean Recall@K (mR@K) for relationship prediction, aligning with standard evaluation protocol in SGG research. R@K measures the proportion of correctly predicted relationships within the top K predictions, considering both the label correctness and the associated confidence scores [60]. This metric is widely used in SGG, as it evaluates both accuracy and confidence in relationship prediction. mR@K is the average of R@K scores across all relationship categories, mitigating the bias toward frequent relationships [61], [62]. By incorporating mR@K, we address concerns related to long-tailed distributions and imbalanced data, ensuring a more balanced evaluation across relationship categories. All baselines reported results under these same experimental conditions, ensuring direct comparability. In order to ensure fairness in comparison, the established evaluation protocol for SGG was adhered to, preserving the integrity of the ground truth scene graphs within the testing dataset, while only the predicted scene graphs were enriched in the proposed methodology.

### A. SGG TRAINING AND EVALUATION

The two primary SGG pipeline components—the Transformer-based deep learning cascade for relationship prediction and the Faster RCNN for object identification—were separately trained and assessed before the evaluation process began. With a batch size equal to 2 and an initial learning rate of 0.002, the Faster RCNN model was trained using stochastic gradient descent (SGD). At the 60,000 and 80,000 iteration marks, the learning rate was reduced

by a factor of 10. Using a 0.5 Intersection over Union (IoU) criteria, the model obtained a mean Average Precision (mAP) of 30.07 after training on the test dataset. On the test dataset, the Transformer-based SGG pipeline achieved an R@100 score of 41.7 after being trained with a batch size of 4 and an initial learning rate of 0.04. The initial evaluation using the Transformer-based SGG model without knowledge enrichment yielded R@20 of 27.9, R@50 of 35.2 and R@100 of 41.7 on VG dataset. On the GQA dataset, the same model achieved R@20 of 26.4, R@50 of 33.8, and R@100 of 42.1, validating the generalizability of the model. Table 1 provides a detailed comparison of the proposed method with existing data-centric SGG techniques, showing that the proposed method, without knowledge enrichment, outperforms the existing methods. The inclusion of attention-based Transformers allows for better capture of global context and long-range dependencies between objects, leading to superior relationship predictions.

### B. SCENE GRAPH ENRICHMENT WITH CSKG

After applying the scene graph enrichment with CSKG, the performance improved significantly, with R@K scores increasing to 31.6, 37.3 and 43.2 on VG and 29.4, 35.7, and 44.5 on GQA. This improvement highlights the contribution of common sense knowledge in providing additional context and correcting missed or incorrect predictions. The enrichment process effectively enhances the recall rates across almost all relationship predicates, including those underrepresented in the dataset, thus demonstrating its potential in mitigating the bias toward more frequent relationships. Table 2 compares the proposed method with the existing knowledge-based SGG approaches. The proposed method, leveraging both Transformers and a comprehensive KG like CSKG, outperforms state-of-the-art methods, with recall scores significantly higher across the board. This clearly underscores the effectiveness of combining advanced deep learning models with enriched knowledge sources in improving the quality of SGG. Beyond achieving higher

**TABLE 2. Thorough comparison of the proposed approach with state of the art common sense knowledge (CSK)-based techniques utilising standard split of the VG dataset and common assessment metrics (R@K and mR@K).**

SGG Technique	Methodology	Knowledge Source	SGG Results	
			R@20/50/100(%)	mR@20/50/100(%)
Proposed Method	Transformer & KG-based SGG	CSKG [26]	<b>31.6/37.3/43.2</b>	<b>9.6/12.4/14.9</b>
DRM [43]	Predicate learning & transfer knowledge	Statistical Prior	-/34.0/38.9	-/9.0/11.2
NeuSyRE [15]	DNN-based SGG and enrichment using CSKG	CSKG [26]	29.9/35.5/39.1	9.1/10.9/12.6
CFA [44]	Intrinsic & extrinsic feature augmentation	Statistical Prior	-/27.7/32.1	-/12.3/14.6
DSGAN [46]	Deep sparse graph attention network	Sparse KG & Statistical Prior	23.2/28.8/32.9	7.8/8.9/11.8
IRT-MSK [47]	Instance Relation Transformer with Multi Structure Knowledge	CN [23], VG [13]	21.9/27.8/31.0	-/-
MOTIFS [42]	Stacked Motif Networks basing on RNN-LSTM	Statistical Prior	21.4/27.2/30.3	4.2/5.7/6.6
GB-Net [64]	Message passing between common sense graphs & scene graph	CN [23], WN [25], VG [13]	-/26.4/30.0	-/6.1/7.3
KERN [62]	Routing Network integrated with knowledge	Statistical Prior	22.3/27.1/29.8	-/6.4/7.3
COACHER [48]	Predicting zero-shot relationships through common sense infusion	ConceptNet [23]	13.4/19.3/22.2	-/-
KB-GAN [65]	Refinement of objects and phrases basing on common sense, reconst.	ConceptNet [23]	-/13.6/17.6	-/-
DeepVRL [45]	Deep Q-network for variation/ structured reinforcement learning	Language Prior	-/13.3/12.6	-/-
VRD [60]	Semantic word embeddings for relationship prediction	Language Prior	-/0.3/0.5	-/-

R@K, our method also exhibits improved mR@K, which is the average recall across all relationship categories. Unlike R@K, which is biased toward frequent relationships, mR@K provides a more balanced evaluation by ensuring that rare relationships are accounted for. Our results show that CSKG enrichment leads to a noticeable improvement in mR@100 from 8.3, 11.5, and 14.1 to 9.6, 12.4, and 14.9, demonstrating that our method mitigates the bias arising from the long-tailed distribution problem. This improvement highlights the model's ability to generalise to less common relationships rather than overfitting to dominant classes.

### C. CONTRIBUTION OF M-HAT AND ViT

To evaluate the individual contributions of M-HAT and ViT in our framework, we performed systematic ablation studies by selectively disabling one component at a time. The goal is to quantify the impact of these components on relationship prediction and global context modelling in SGG. In the first experiment, we retained only M-HAT by removing the ViT block responsible for contextual feature extraction. In the second, we retained ViT while removing M-HAT, which is responsible for capturing inter-object relationships. Additionally, a baseline experiment was conducted where both M-HAT and ViT were removed to assess the combined effect. The results, shown in Table 3, indicate that the absence of either component causes a significant drop in both R@100 and mR@100, demonstrating that M-HAT and ViT synergistically capture global context and long-range dependencies effectively. M-HAT excels in modelling inter-object dependencies through self-attention mechanisms, while ViT enhances the contextual awareness of object relationships via global feature aggregation. Together, they enable the model to jointly learn both visual semantics and object interactions, leading to more accurate and expressive scene graphs.

**TABLE 3. Impact of M-HAT and ViT on relationship prediction performance.**

Component	R@100	mR@100
M-HAT + ViT	<b>43.2</b>	<b>14.9</b>
M-HAT Only	40.1	13.6
ViT Only	39.7	13.3

**TABLE 4. Computational overhead comparison with and without CSKG.**

Method	Average Inference Time (ms)
Baseline SGG (without CSKG)	521.3
SGG with CSKG Enrichment	537.8
Overhead Introduced	16.5

### D. COMPUTATIONAL ANALYSIS

To assess the computational overhead of incorporating knowledge enrichment, we measured the inference times of the proposed SGG method with and without CSKG enrichment. As shown in Table 4, the CSKG enrichment step adds an average of 16.5ms per image, which accounts for approximately 3.2% additional computational cost. The enrichment process is optimised through precomputed knowledge graph embeddings and batch query execution, ensuring minimal impact on inference speed. Given the significant boost in relationship prediction accuracy, this minor computational cost is an acceptable trade-off.

### E. COMPARISON OF ENRICHMENT METHODS

We compared the proposed CSKG-based enrichment with common alternative enrichment strategies including statistical priors and ConceptNet [23] knowledge graph. Additionally, we performed an ablation study where no knowledge enrichment was applied, serving as a baseline to evaluate the effectiveness of external common sense knowledge integration. For statistical priors, we utilised frequency-based co-occurrence information from the training

**TABLE 5.** Performance comparison of enrichment methods on SGG performance.

Knowledge Source	R@100	mR@100
CSKG	<b>43.2</b>	<b>14.9</b>
Statistical Prior	42.1	14.4
ConceptNet	41.9	14.3
None	41.7	14.1

**TABLE 6.** Impact of linguistic features on SGG performance.

Configuration	R@100	mR@100
With Linguistic Features	<b>43.2</b>	<b>14.9</b>
Without Linguistic Features	30.4	10.3

set. For ConceptNet, we used its pre-trained embeddings and queried relevant triples based on object labels in the scene graph. The results, shown in Table 5, indicate that CSKG-based enrichment provides the most comprehensive contextual knowledge, yielding higher recall and mean recall scores compared to alternatives. While statistical priors and ConceptNet provide useful information, they fall short in terms of generalizability and capturing complex visual relationships (e.g., spatial or functional interactions). CSKG, with its heterogeneous knowledge base, is better suited for enriching scene graphs in complex visual reasoning tasks.

#### F. IMPACT OF LINGUISTIC FEATURES

To assess the importance of incorporating linguistic features in multimodal relationship prediction, we conducted a targeted ablation study by removing object labels and textual features from the model inputs. In this setting, the model was forced to rely solely on visual features for relationship prediction. The results in Table 6 show that removing the linguistic features caused a significant drop in performance, confirming the critical role of multimodal visual-linguistic features in capturing semantic relationships. Linguistic features provide semantic cues that complement visual information, particularly for context-sensitive relationships that require word-level disambiguation.

#### G. EFFECT OF EMBEDDING SIMILARITY AND IOU THRESHOLDS

To assess the impact of cosine similarity ( $\tau_{sim}$ ) and IoU thresholds ( $\tau_{IoU}$ ) during graph refinement, we tested various combinations of these thresholds. We experimented with threshold values of  $\tau_{sim} = \{0.25, 0.5, 0.75\}$  and  $\tau_{IoU} = \{0.25, 0.5, 0.75\}$ . The results, summarized in Table 7, indicate that the default thresholds ( $\tau_{sim} = 0.5$ ,  $\tau_{IoU} = 0.5$ ) provided the best trade-off between precision and recall. Higher similarity thresholds (e.g.,  $\tau_{sim} = 0.75$ ) lead to stricter entity merging, ensuring that only highly similar objects are considered the same entity, but this may exclude relevant relationships, reducing recall. Conversely, lower similarity thresholds (for example,  $\tau_{sim} = 0.25$ ) lead to more aggressive merging, which introduces spurious relationships and increases false positives. A similar effect

**TABLE 7.** Effect of varying similarity thresholds on SGG performance.

Thresholds	$\tau_{sim}$	$\tau_{IoU}$	R@100	mR@100
Default	0.5	0.5	<b>43.2</b>	<b>14.9</b>
High Cosine Similarity Threshold	0.75	0.5	42.0	14.2
Low Cosine Similarity Threshold	0.25	0.5	40.9	13.7
High IoU Threshold	0.5	0.75	42.3	14.5
Low IoU Threshold	0.5	0.25	41.2	14.0

is observed for IoU thresholds: higher thresholds discard too many overlapping detections, removing potentially valid object instances, while lower thresholds retain excessive redundancy, causing duplicate object instances in the scene graph. The incorporation of IoU-based filtering and cosine similarity-based merging improves scene graph representation by resolving overlapping object conflicts. Performance improvement in R@100 and mR@100 confirms that handling redundant detections enhances the expressiveness of the scene graph.

#### V. CONCLUSION

In this paper, we introduced a novel neurosymbolic framework, MuRelSGG, that integrates transformer-based multimodal relationship prediction with common sense knowledge enrichment to enhance scene graph generation. This approach combines deep learning perception with structured knowledge reasoning, leading to more accurate and expressive scene representations. Our methodological contributions include multi-head attention transformers (M-HAT), vision transformers (ViT), and structured knowledge enrichment, which together significantly improves relationship prediction accuracy and long-tail relationship representation. Extensive experiments on the Visual Genome (VG) and GQA datasets confirm the generalizability and robustness of our framework. Our model outperformed the state-of-the-art methods in terms of recall rates and robustness to data bias (R@100 = 43.2, mR@100 = 14.9) on the VG benchmark and showed consistent performance on GQA dataset. Ablation studies revealed the critical roles of M-HAT, ViT, linguistic features, CSKG enrichment, and optimized similarity thresholds in enhancing SGG performance, validating the effectiveness of our proposed approach. Future work will explore zero-shot and few-shot SGG capabilities, applications in downstream tasks like image captioning and visual question answering, and integration of multimodal Large Language Models (LLMs) in the proposed method. Our work establishes a strong foundation for advancing neurosymbolic SGG, demonstrating that the synergistic combination of deep learning and structured knowledge can lead to more semantically rich and context-aware scene representation.

#### REFERENCES

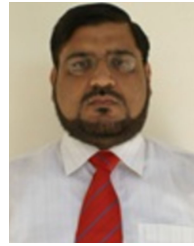
- [1] P. Hitzler, F. Bianchi, M. Ebrahimi, and M. K. Sarker, "Neural-symbolic integration and the semantic web," *Semantic Web*, vol. 11, no. 1, pp. 3–11, Jan. 2020.

- [2] A. Bennetot, J.-L. Laurent, R. Chatila, and N. Díaz-Rodríguez, “Towards explainable neural-symbolic visual reasoning,” 2019, *arXiv:1909.09065*.
- [3] W. W. Cohen, H. Sun, R. Alex Hofer, and M. Siegler, “Scalable neural methods for reasoning with a symbolic knowledge base,” 2020, *arXiv:2002.06115*.
- [4] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, “A comprehensive survey of scene graphs: Generation and application,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 1–26, Jan. 2023.
- [5] X. Chang, P. Ren, P. Xu, Z. Li, X. Chen, and A. Hauptmann, “A comprehensive survey of scene graphs: Generation and application,” 2021, *arXiv:2104.01111*.
- [6] Y. Zhong, L. Wang, J. Chen, D. Yu, and Y. Li, “Comprehensive image captioning via scene graph decomposition,” in *Proc. Eur. Conf. Comput. Vis. Cham, Switzerland: Springer*, Jan. 2020, pp. 211–229.
- [7] R. Koner, H. Li, M. Hildebrandt, D. Das, V. Tresp, and S. Günemann, “Graphhopper: Multi-hop scene graph reasoning for visual question answering,” in *Proc. Int. Semantic Web Conf. Cham, Switzerland: Springer*, Jan. 2021, pp. 111–127.
- [8] B. Schroeder and S. Tripathi, “Structured query-based image retrieval using scene graphs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2020, pp. 178–179.
- [9] C. Agia, K. M. Jatavallabhula, M. Khodeir, O. Mikšik, V. Vineet, M. Mukadam, L. Paull, and F. Shkurti, “Taskography: Evaluating robot task planning over large 3D scene graphs,” in *Proc. Conf. Robot Learn.*, Jun. 2021, pp. 46–58.
- [10] J. Ji, R. Krishna, L. Fei-Fei, and J. C. Niebles, “Action genome: Actions as compositions of spatio-temporal scene graphs,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10233–10244.
- [11] T. Tahara, T. Seno, G. Narita, and T. Ishikawa, “Retargetable AR: Context-aware augmented reality in indoor scenes based on 3D scene graph,” in *Proc. IEEE Int. Symp. Mixed Augmented Reality Adjunct (ISMAR-Adjunct)*, Nov. 2020, pp. 249–255.
- [12] E. Curry, D. Salwala, P. Dhingra, F. A. Pontes, and P. Yadav, “Multimodal event processing: A neural-symbolic paradigm for the Internet of Multimedia Things,” *IEEE Internet Things J.*, vol. 9, no. 15, pp. 13705–13724, Aug. 2022.
- [13] 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.
- [14] M. J. Khan, J. G. Breslin, and E. Curry, “Towards fairness in multimodal scene graph generation: Mitigating biases in datasets, knowledge sources and models,” in *Proc. 32nd ACM Int. Conf. Inf. Knowl. Manage. (CIKM) Workshops*, 2023.
- [15] M. J. Khan, J. G. Breslin, and E. Curry, “NeuSyRE: Neuro-symbolic visual understanding and reasoning framework based on scene graph enrichment,” *Semantic Web*, vol. 15, no. 4, pp. 1389–1413, Oct. 2024.
- [16] 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.
- [17] Y. Zhang, Y. Pan, T. Yao, R. Huang, T. Mei, and C. Chen, “Boosting scene graph generation with visual relation saliency,” in *Proc. ACM Trans. Multimedia Comput., Commun., Appl. (TOMM)*, vol. 19, Mar. 2022, pp. 1–17.
- [18] T. He, L. Gao, J. Song, J. Cai, and Y.-F. Li, “Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation,” 2020, *arXiv:2006.07585*.
- [19] K. Ye and A. Kovashka, “Linguistic structures as weak supervision for visual scene graph generation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8285–8295.
- [20] C.-W. Lee, W. Fang, C.-K. Yeh, and Y. 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.
- [21] V. Kumar, A. Mundu, and S. K. Singh, “Scene graph generation with geometric context,” in *Proc. Int. Conf. Comput. Vis. Image Process. Cham, Switzerland: Springer*, Jan. 2021, pp. 340–350.
- [22] W. Wang, Y. Yang, and F. Wu, “Towards data-and knowledge-driven artificial intelligence: A survey on neuro-symbolic computing,” 2022, *arXiv:2210.15889*.
- [23] R. Speer, J. Chin, and C. Havasi, “Conceptnet 5.5: An open multilingual graph of general knowledge,” in *Proc. 31st AAAI Conf. Artif. Intell.*, 2017, pp. 4444–4451.
- [24] M. Sap, R. L. Bras, E. Allaway, C. Bhagavatula, N. Lourie, H. Rashkin, B. Roof, N. A. Smith, and Y. Choi, “ATOMIC: An atlas of machine commonsense for if-then reasoning,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 33, Jul. 2019, pp. 3027–3035.
- [25] G. A. Miller, “WordNet: A lexical database for English,” *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.
- [26] F. Ilievski, P. Szekely, and B. Zhang, “Cskg: The commonsense knowledge graph,” in *Proc. Eur. Semantic Web Conf. Cham, Switzerland: Springer*, 2021, pp. 680–696.
- [27] 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*, Jan. 2022, pp. 93–112.
- [28] K. Ma, F. Ilievski, J. Francis, Y. Bisk, E. Nyberg, and A. Oltramari, “Knowledge-driven data construction for zero-shot evaluation in commonsense question answering,” in *Proc. 35th AAAI Conf. Artif. Intell.*, vol. 35, May 2021, pp. 13507–13515.
- [29] 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.
- [30] 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.
- [31] 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.*, pp. 1–24, May 2024.
- [32] 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. 1270–1279.
- [33] 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.
- [34] 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)*, Jan. 2018, pp. 690–706.
- [35] 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)*, Jan. 2018, pp. 335–351.
- [36] 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.
- [37] 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*.
- [38] J. Im, J. Nam, N. Park, H. Lee, and S. Park, “EGTR: Extracting graph from transformer for scene graph generation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 24229–24238.
- [39] Z. Hayder and X. He, “DSGG: Dense relation transformer for an end-to-end scene graph generation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 28317–28326.
- [40] H. Liu and B. Bhanu, “RepSGG: Novel representations of entities and relationships for scene graph generation,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 8018–8035, Dec. 2024.
- [41] F. Ilievski, A. Oltramari, K. Ma, B. Zhang, D. L. McGuinness, and P. Szekely, “Dimensions of commonsense knowledge,” 2021, *arXiv:2101.04640*.
- [42] 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.
- [43] J. Li, Y. Wang, X. Guo, R. Yang, and W. Li, “Leveraging predicate and triplet learning for scene graph generation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 28369–28379.
- [44] L. Li, G. Chen, J. Xiao, Y. Yang, C. Wang, and L. Chen, “Compositional feature augmentation for unbiased scene graph generation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21685–21695.
- [45] 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.
- [46] 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.

- [47] 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.
- [48] 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*. Cham, Switzerland: Springer, Jan. 2021, pp. 466–482.
- [49] M. J. Khan, J. G. Breslin, and E. Curry, "KnowZRel: Common sense knowledge-based zero-shot relationship retrieval for generalised scene graph generation," *IEEE Trans. Artif. Intell.*, early access, Feb. 21, 2025, doi: [10.1109/TAI.2025.3544177](https://doi.org/10.1109/TAI.2025.3544177).
- [50] D. Vrandečić and M. Krötzsch, "Wikidata: A free collaborative knowledgebase," *Commun. ACM*, vol. 57, no. 10, pp. 78–85, Sep. 2014.
- [51] B. Kipper, *Rogee's 21st Century Thesaurus in Dictionary Form*, 3rd ed., New York, NY, USA: The philip lief group, 2005.
- [52] H. Wan, J. Ou, B. Wang, J. Du, J. Z. Pan, and J. Zeng, "Iterative visual relationship detection via commonsense knowledge graph," in *Proc. Joint Int. Semantic Technol. Conf.* Cham, Switzerland: Springer, 2019, pp. 210–225.
- [53] 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.
- [54] 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. 2117–2125.
- [55] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2961–2969.
- [56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.
- [57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.
- [58] 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.
- [59] 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.* Cham, Switzerland: Springer, Jan. 2020, pp. 278–293.
- [60] C. Lu, R. Krishna, M. S. Bernstein, and L. Fei-Fei, "Visual relationship detection with language priors," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Jan. 2016, pp. 852–869.
- [61] 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. 6612–6621.
- [62] 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.
- [63] 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. 3713–3722.
- [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, Jan. 2020, pp. 606–623.
- [65] 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.



**MUHAMMAD JUNAID KHAN** received the master's degree in electrical engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2020. He is currently a Ph.D. Scholar with NUST. His research is focused on Improved Visual Reasoning: A Neurosymbolic Approach With Scene Graph Enrichment. He was a Faculty Member with the Military College of Signals, NUST. His research interests include computer vision, image processing, deep learning, and embedded systems



**ADIL MASOOD SIDDIQUI** received the Ph.D. degree in electrical engineering from the University of Engineering and Technology, Taxila, Pakistan, in 2009. He is currently an Associate Professor and the Dean of the Military College of Signals, National University of Sciences and Technology (NUST), Pakistan. He has contributed more than 40 research publications which include 25 high-impact factor journals. His research interests include computer vision, image registration, image de-noising, image enhancement, and defogging.



**HAMID SAEED KHAN** received the B.S. degree in software engineering from the University of Engineering and Technology, and the M.S. degree in engineering project management from the National University of Sciences and Technology. He is currently a Software Engineering Manager with siParadigm Diagnostic Informatics, USA. His research interests include artificial intelligence, software engineering, engineering management, and digital pathology.



**FAISAL AKRAM** received the bachelor's degree in telecommunication engineering from the National University of Sciences and Technology (NUST), Pakistan, in 2005, the master's degree in communication technology from the University of Ulm, Germany, in 2013, and the Ph.D. degree from NUST. He is currently an Associate Professor with the Electrical Engineering Department, Military College of Signals, NUST. His research interests include compressed sensing, wireless communications, mm-wave hybrid MIMO systems, machine learning, deep reinforcement learning, and channel coding.



**M. JALEED KHAN** received the Ph.D. degree in artificial intelligence from the University of Galway, Ireland. He is currently a Senior Researcher in AI and data analytics with Fujitsu Research and an Honorary Research Fellow with the University of Oxford, U.K. His research interests include neurosymbolic AI, machine learning, and computer vision has resulted in more than 40 highly-cited publications and several book chapters and open-source projects. He is actively involved in the AI research community as a PC member of top-tier conferences (ECAI and ECML), a Reviewer for journals (TPAMI, IJCV, and TNNLS), a Professional Member of ACM and IAPR, and a Grant Panelist for funding bodies (FFG Austria and NCN Poland).

...