Transformer-Based Relational Inference Network for Complex Visual Relational Reasoning

MINGKUI TAN and ZHIQUAN WEN, South China University of Technology, China
LEYUAN FANG, Hunan University, China
QI WU, The University of Adelaide, Australia

Visual Relational Reasoning is the basis of many vision-and-language based tasks (e.g., visual question answering and referring expression comprehension). In this article, we regard the complex referring expression comprehension (c-REF) task as the reasoning basis, in which c-REF seeks to localise a target object in an image guided by a complex query. Such queries often contain complex logic and thus impose two critical challenges for reasoning: (i) Comprehending the complex queries is difficult since these queries usually refer to multiple objects and their relationships; (ii) Reasoning among multiple objects guided by the queries and then localising the target correctly are non-trivial. To address the above challenges, we propose a Transformer-based Relational Inference Network (Trans-RINet). Specifically, to comprehend the queries, we mimic the language-comprehending mechanism of humans, and devise a language decomposition module to decompose the queries into four types, i.e., basic attributes, absolute location, visual relationship and relative location. We further devise four modules to address the corresponding information. In each module, we consider the intra-(i.e., between the objects) and inter-modality relationships(i.e., between the queries and objects) to improve the reasoning ability. Moreover, we construct a relational graph to represent the objects and their relationships, and devise a multi-step reasoning method to progressively understand the complex logic. Since each type of the queries is closely related, we let each module interact with each other before making a decision. Extensive experiments on the CLEVR-Ref+, Ref-Reasoning, and CLEVR-CoGenT datasets demonstrate the superior reasoning performance of our Trans-RINet.

CCS Concepts: • **Computing methodologies → Knowledge representation and reasoning;**

Additional Key Words and Phrases: Visual Relational Reasoning, complex referring expression comprehension, Gated Graph Neural Network

ACM Reference format:
Mingkui Tan, Zhiquan Wen, Leyuan Fang, and Qi Wu. 2023. Transformer-Based Relational Inference Network for Complex Visual Relational Reasoning. *ACM Trans. Multimedia Comput. Commun. Appl.* 20, 1, Article 10 (August 2023), 23 pages.
https://doi.org/10.1145/3605781

Mingkui Tan and Zhiquan Wen contributed equally to this research.

This work was partially supported by STI 2030–Major Projects (2022ZD0208900), National Natural Science Foundation of China (NSFC) 62072190, Key Realm R&D Program of Guangzhou 202007030007, Program for Guangdong Introducing Innovative and Entrepreneurial Teams 2017ZT0-7X183.

Authors' addresses: M. Tan and Z. Wen, South China University of Technology, Guangzhou, China; emails: mingkuitan@scut.edu.cn, sewenzhiquan@mail.scut.edu.cn; L. Fang, Hunan University, Changsha, China; email: fangleyuan@gmail.com; Q. Wu, The University of Adelaide, Adelaide, Australia; email: qi.wu01@adelaide.edu.au.

1551-6857/2023/08-ART10 $15.00
https://doi.org/10.1145/3605781

1 INTRODUCTION

Visual relational reasoning requires an agent to fully comprehend the textual information and then reason among multiple objects based on their relationships. This ability is vital for many vision-and-language based tasks, such as visual question answering (VQA) [22, 32, 35] and vision language navigation (VLN) [1, 34, 48]. However, reasoning can be very difficult because the visual and textual contents are often very complex. How to build a model to perform complex visual relational reasoning and how to validate the reasoning ability of such a model are still unclear.

Fortunately, the **complex** referring expression comprehension (c-REF) task [29, 56] is suitable as a test bed for visual relational reasoning methods. The reason is that c-REF requires an agent to perform reasoning among multiple objects and then localise a target object in an image, guided by a complex query (see Figure 1). More critically, the complex visual and textual contents in this task can be a simulation of real-world scenarios.

This task, however, is very challenging due to the following reasons: (a)The query typically contains multiple types of information, including the basic attributes, absolute location, visual relationship and relative location (refer to Figure 1). Comprehensively understanding the complex query is non-trivial. (b) Different from the general referring expression comprehension (g-REF) task (e.g., RefCOCO [25, 33]) that the queries usually describe objects without considering any relationships, the queries in c-REF often contain complex relationships among several objects (Refer to in Figure 2). Reasoning among multiple objects and then correctly localising the target object are very challenging.

Recently, Liu et al. [29] found some state-of-the-art g-REF methods [59] cannot achieve excellent reasoning performance on the c-REF task (e.g., a CLEVR-Ref+ dataset [29]), which may be attributed to the overlooking of the relationship among objects. Moreover, some methods [47, 58] seek to perform single-step reasoning only to model the relationships between the objects. However, as the real-world queries are usually complex and involve intricate relationships, adopting single-step reasoning to comprehend the queries is insufficient. As shown in Figure 1, reasoning is a multi-step process: **Step 1**, Find "*the first one of the yellow cylinders from the left*"; **Step 2**, Based on object selected in Step 1, find "*the thing that is on the left side of*" it; **Step 3**, Find "*the thing that is in front of*"the object selected in Step 2; **Step 4**, Find "*the sphere that has the same colour as*" the object selected in Step 3. Although some methods [3, 17] attempt to perform multi-step reasoning, these methods do not differentiate between the types of information in the queries. In this sense, they have insufficient query comprehension, resulting in capturing the complex logic insufficiently, and thus achieving unsatisfied performance.

To alleviate these issues, we propose a Transformer-based Relational Inference Network (Trans-RINet), which distinguishes different information in the query and models the relationships between the objects for multi-step reasoning. Specifically, to comprehend the sophisticated query, we devise a **language decomposition module** to decompose the query into four types, i.e., the basic attributes, absolute location, visual relationship and relative location. Then, we propose a **Transformer-based object attention module** to find the objects that are relevant to the information of the basic attributes and absolute location in the query. Moreover, to achieve multi-step reasoning, we propose a **Transformer-based relational inference module**. Specifically, we first transform an image into a relational graph, where the nodes correspond to the objects, and the edges denote their relationships. With the graph, we propose a multi-step reasoning method with Gated Graph Neural Networks (GGNNs) [28] to progressively understand the complex logic.

Note that in both the Transformer-based object attention module and relational inference module, we consider the intra- (i.e., between the objects) and inter-modality relationships (i.e.,

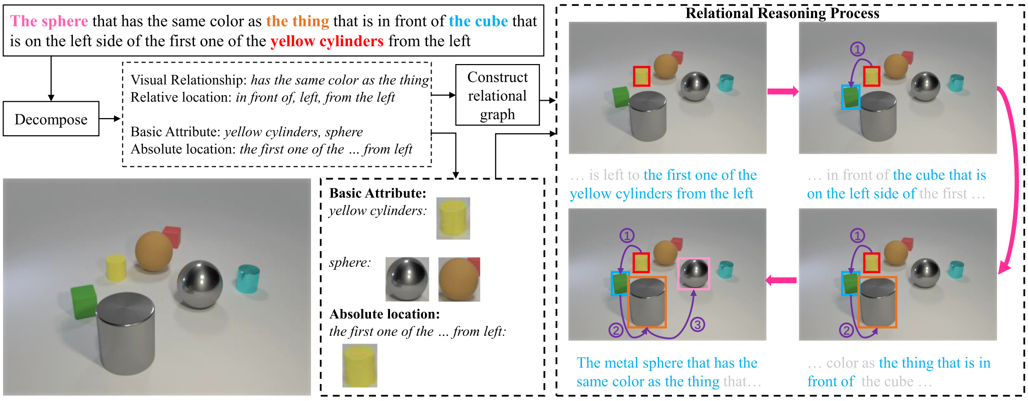
The sphere that has the same color as the thing that is in front of the cube that is on the left side of the first one of the yellow cylinders from the left
Decompose
Visual Relationship: has the same color as the thing
Relative location: in front of, left, from the left
Basic Attribute: yellow cylinders, sphere
Absolute location: the first one of the ... from left
Construct relational graph
Basic Attribute:
yellow cylinders:
sphere:
Absolute location:
the first one of the ... from left:
Relational Reasoning Process
... is left to the first one of the yellow cylinders from the left
... in front of the cube that is on the left side of the first ...
The metal sphere that has the same color as the thing that...
... color as the thing that is in front of the cube ...
①
②
③

Fig. 1. An example on the CLEVR-Ref+ dataset [29], where the query contains complex relationships, requiring multi-step reasoning to be solved. The aim of the example is to localise the target “metal sphere”based on the given complex query. To solve this, we propose to model the multi-step relationships via a relational graph step by step, and then reason to the target.

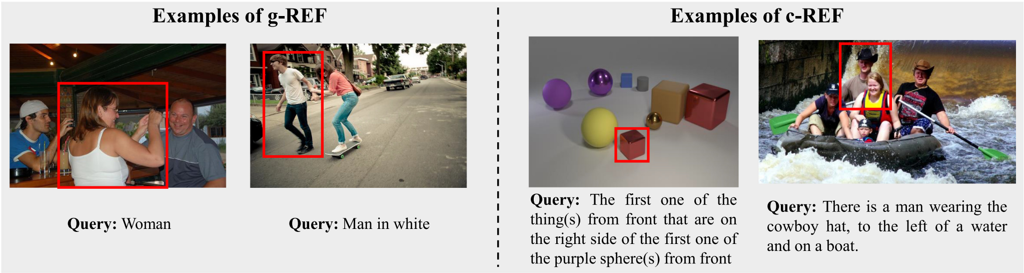
Examples of g-REF
Examples of c-REF
Query: Woman
Query: Man in white
Query: The first one of the thing(s) from front that are on the right side of the first one of the purple sphere(s) from front
Query: There is a man wearing the cowboy hat, to the left of a water and on a boat.

Fig. 2. Examples of the g-REF (e.g., RefCOCO dataset [25]) and c-REF tasks (e.g., CLEVR-Ref+ [29] and Ref-Reasoning [56] datasets). Note that the object referred to by the query is indicated by a red bold rectangle.

between the queries and objects). When localising one object in an image, only considering the inter-modality relationship is not enough. The intra-modality relationship is also important for localising the target. For example, if an image contains two similar objects, considering the inter-modality relationships only may result in localising the target insufficiently. But if further modelling the intra-modality relationships, the difference between these two objects is easy to be discovered.

We also propose a **multimodal fusion module** to let each module corresponding to each type of query interact with each other, since each type of query is closely related. Finally, we conduct experiments on some classical c-REF benchmarks, e.g., CLEVR-Ref+ [29], Ref-Reasoning [56], and CLEVR-CoGenT [29] datasets, where the queries in these datasets are complex, requiring powerful relational reasoning ability to localise the target of interest.

Summary of our key contributions:

- We propose a Transformer-based Relational Inference Network (Trans-RINet) to achieve long-chain reasoning, where Trans-RINet decomposes the query into four types to simplify the understanding of the query. Moreover, Trans-RINet performs multi-step reasoning with GGNNs to progressively comprehend the complex logic in the query.

- We propose to consider both the intra- and inter-modality relationships when localising the target. Moreover, we propose to promote the interaction between each module, since each type of query is closely related.
- Extensive experiments on the CLEVR-Ref+ [29], Ref-Reasoning [56], and CLEVR-CoGent [29] datasets demonstrate the superiority of our Trans-RINet.

This article extends our preliminary version MGA-Net [61] from several aspects. We propose an advanced version Trans-RINet by further: (1) considering the intra-modality relationships to improve the reasoning ability. Specifically, MGA-Net only considers inter-modality relationships when addressing the reasoning process. Instead, Trans-RINet considers intra- and inter-modality relationships when localising the target. In practice, due to the powerful ability to model relationships of Transformer, we substitute for the attention unit in object attention and relational inference modules of MGA-Net with the Transformer; (2) considering the interaction between each module to improve the reasoning ability. MGA-Net isolatedly deals with each module that corresponds to each type of query, ignoring that these modules are closely related. In Trans-RINet, we propose to promote the interaction between each module. Since the Transformer can make the entities connect and interact with each other freely, we adopt the Transformer as the interaction base; (3) providing more empirical results to show the effectiveness of our method. Specifically, Trans-RINet outperforms MGA-Net by more than 10% on the CLEVR-Ref+ and CLEVR-CoGent datasets.

2 RELATED WORK

2.1 Visual Relational Reasoning

Many vision-and-language tasks require visual relational reasoning to localise the referred object of the query, e.g., visual question answering (VQA) [30, 49–51, 57], visual language navigation (VLN) [21, 26, 38] and referring expression comprehension (REF) [39, 46, 47]. To address these high-level tasks, some recent works [18, 47, 58] tried to divide the query into components and performed the reasoning based on each component. Other works [6, 24, 43] exploited Neural Module Networks [2] to perform step-by-step reasoning by clearly dividing the queries into specific logic. Specifically, Chen et al.proposed a neural-symbolic framework called the Dynamic Concept Learner (DCL) [6], which achieves dynamic visual reasoning by recognising objects and events in videos and analysing their temporal and causal structures.

2.2 Referring Expression Comprehension

Referring Expression Comprehension (REF) task seeks to localise the referent (i.e., object) in a given image with a query. Many benchmark datasets were released for research (e.g., RefCOCO [25]). However, as discussed in [17], the queries in g-REF datasets did not require resolving relationships. Moreover, recent research [9] pointed out that the bias existed in the RefCOCO datasets. In this case, one may achieve high performance without any queries. To address this issue, Liu et al.constructed an approximately unbiased dataset called *CLEVR-Ref+* [29] to measure the reasoning ability of the models more accurately. Moreover, to evaluate the reasoning ability of the models across multiple similar images, Chen et al.introduced a new dataset called *Cops-Ref* [7] for the Compositional Referring Expression Comprehension task, which involves localising a region from a set of images. In this article, we evaluate our method on the complex referring expression comprehension (c-REF) task with CLEVR-Ref+ [29], Ref-Reasoning [56], and CLEVR-CoGenT [29] datasets. Note that these datasets are more challenging since they not only reduce the statistical biases but also require long-chain reasoning ability to comprehend complex queries.

2.3 Graph Neural Networks

Graph Neural Networks (GNNs) become one of the mainstream methods of reasoning gradually, and the variants (e.g., Graph Attention Network (GAT) [45] and Gated Graph Neural Networks (GGNNs) [28]) have been applied to several sophisticated tasks [17, 20, 60]. Recently, some works [23, 47, 54, 55] attempted to perform relational reasoning on the graphs. Specifically, LGRANs [47] adopted GAT to perform reasoning by aggregating the information for each graph node from its neighbourhoods. However, these methods performed reasoning in a one-step manner, while the complex queries require multi-step reasoning to be solved. DGA [55] and CM-RIN [54] considered the relationships among the objects in a neighbourhood distance, but they ignored the attribute relationships between the objects (e.g., two objects have the same colour). Moreover, LCGN [17] performed multi-step reasoning with a fully-connected graph. However, it did not distinguish different types of information in the query, leading to insufficient comprehension of the complex query. Jing et al. [23] modelled the relational reasoning process as reinforcement learning and conducted reasoning based on a graph. Unlike these methods, our Trans-RINet refines the relation representation with visual and location relations, which is helpful for complex relationship modelling. Moreover, we conduct multi-step reasoning with two relational graphs corresponding to visual and location relations, which is achieved by iteratively updating the graph representations.

2.4 Transformer

Transformer has first been proposed in [44] for addressing the neural machine translation task. Due to a self-attention mechanism in Transformer that can make the entities connect and interact with each other freely, Transformer has achieved excellent performance in language [10, 40], vision [4, 11], and vision&language [5, 19] fields.

2.4.1 Language Field.

Many transformer-based models [10, 40] have been proposed and achieved promising performance. Specifically, BERT [10] stacked the Transformer encoder, and achieved better performance due to the bidirectional pre-training for language representations. T5 [40] transformed many Natural Language Processing (NLP) tasks into generation tasks, which had the potential to address all of the NLP tasks in one model.

2.4.2 Vision Field.

Due to the success of Transformers in the language field, a series of transformers-based models [4, 11] have been applied to the vision field. Imitating the input of the Transformer in the language field, ViT [11] decomposed the image into some patches, and showed that a pure Transformer can achieve excellent performance on image classification. Moreover, DETR [4] formulated object detection as a set of prediction problems, and achieved better performance.

2.4.3 Vision & Language Field.

Inspired by BERT [10], some visual-linguistic pre-training (VLP) based methods [5, 15, 31] have been proposed to jointly learn the representation of images and texts. In general, these models contained several transformer encoder layers and took the object proposals and texts as inputs. These methods achieved good performance in many downstream tasks by fine-tuning. Other methods [19, 52] regarded Transformer as a feature updater, and performed the target task end-to-end, without pre-training. In Trans-RINet, we follow this paradigm and adopt Transformer to consider both the intra- and inter-modality relationships, and promote the interaction between the modules. Moreover, Guan et al. [12] adopted the same Transformer [44] to fuse the multiple embeddings of each user and item, respectively, where the fusion occurs intra-modality. Instead, in our Trans-RINet, we adopt a Transformer to fuse both objects and query information, which is cross-modality.

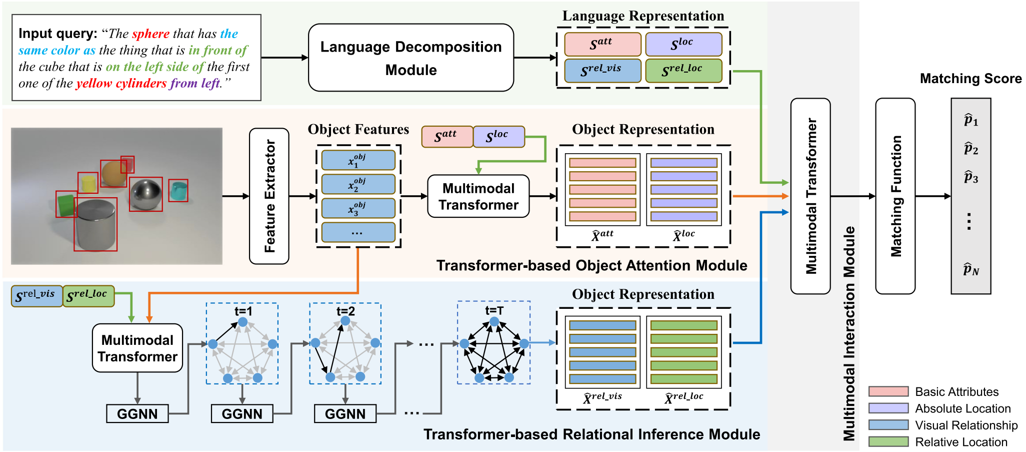

Fig. 3. Overview of Transformer-based Relational Inference Network (Trans-RINet) that contains four components. The language decomposition module decomposes the query into four types. The Transformer-based object attention module finds the relevant objects according to the information of the properties. With the relational graphs, the Transformer-based relational inference module captures the complex logic by a multi-step reasoning method. A multimodal interaction module lets each module interact with the other freely. The final score is obtained by matching the language and the corresponding object representations.

3 TRANSFORMER-BASED RELATIONAL INFERENCE NETWORK

In this article, we seek to construct a model to conduct visual relational reasoning guided by a complex query and then localise a target in an image. We find that the c-REF task requires a strong reasoning ability to localise a target, which is suitable as a test bed to evaluate our method. Formally, given a query r and its corresponding image I with N objects $\mathcal{O} = \{o_i\}_{i=1}^{N}$, the aim of c-REF is to localise the target object $o^{\star}$ by conducting reasoning among the objects $\mathcal{O}$, guided by r.

Since the queries usually contain rich information that refers to multiple objects and their relationships, how to fully comprehend them and how to perform reasoning over several objects guided by these queries are challenging. To address the above challenges, we first decompose the query into different types of information to reduce the difficulty of understanding. We further devise a functional module for each type of information. In each module, we consider both the intra- (i.e., between the objects) and inter-modality relationships (i.e., between the queries and the objects) to help localise the target. To comprehend the complex logic in a query, we first transform an image into a relational graph, where the nodes are the objects, and the edges are their relationships. Then, we propose a graph-based multi-step reasoning method to update the graph progressively, guided by the query. Since each type of query is closely related, we further consider letting each module interact with each other before making a decision.

As shown in Figure 3, our Trans-RINet contains four components. Specifically, the **language decomposition module** first decomposes the query r into four types, and obtains the corresponding language representations (i.e., $\mathbf{s}^{att}$, $\mathbf{s}^{loc}$, $\mathbf{s}^{rel\_vis}$, and $\mathbf{s}^{rel\_loc}$). Then, the **Transformer-based object attention module** adopts the visual features and location features to represent each object, and finds the candidate objects that are relevant to r guided by $\mathbf{s}^{att}$ and $\mathbf{s}^{loc}$. To perform multi-step reasoning, the **Transformer-based relational inference module** constructs two relational graphs among objects and their relationships, and then updates the corresponding node representations progressively with GGNNs under the guidance of $\mathbf{s}^{rel\_vis}$ and $\mathbf{s}^{rel\_loc}$, respectively.

Next, the **multimodal interaction module** promotes the communication between each module to achieve the aim of interaction. Last, we match the updated object representations with the corresponding language representations to obtain the final score.

3.1 Preliminary

3.1.1 Transformer. Due to the powerful relationship modelling ability of Transformer [44], it has achieved promising performance in vision&language [5, 19] field. We seek to take advantage of the strength of the Transformer to improve the reasoning ability. To detail our Trans-RINet, we first briefly review the main idea of the Transformer.

3.1.2 Self-Attention Mechanism. Different from the traditional attention mechanism [53], a self-attention mechanism is able to make the entities connect and interact with each other freely. Based on the multi-head self-attention mechanism, the Transformer enables to model both inter- and intra-modality relationships in a homogeneous way. For example, an object is allowed to attend to other objects and queries simultaneously. Specifically, given a feature vector $\mathbf{f} \in \mathbb{R}^{l \times d}$, the self-attention mechanism can be formulated as:

$$\tilde{\mathbf{f}} = \text{softmax}\left(\frac{(\mathbf{W}_Q\mathbf{f})(\mathbf{W}_K\mathbf{f})^\top}{\sqrt{d}}\right)(\mathbf{W}_V\mathbf{f}), \tag{1}$$

where $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_V$ are learnable parameters, $\frac{1}{\sqrt{d}}$ is the scaling factor, and $\tilde{\mathbf{f}} \in \mathbb{R}^{l \times d}$ is the updated feature. In practice, a standard Transformer usually contains multiple self-attention layers, feed-forward networks, and residual connections [14]. Formally, we use $\Phi(\mathbf{f}; \mathbf{W})$ to denote the Transformer layer, where $\mathbf{W}$ is the corresponding parameters.

3.2 Language Decomposition Module

As shown in Figure 1, queries in c-REF usually refer to multiple objects and their relationships, which can be decomposed into four types of information, i.e., (1) basic attributes: object category, size, colour and material; (2) absolute location: the absolute location of the objects in the image; (3) visual relationship: the relationships between objects (e.g., the objects have the same attribute or the interaction between the objects); and (4) relative location: the displacement between objects.

Due to the complexity of the queries, some methods [2] used an off-the-shelf language parser [62] to help comprehend the queries. However, as mentioned in [58], the off-the-shelf parser could raise irreparable parsing errors, which is harmful to the models. Thus, rather than relying on the off-the-shelf language parser, following [47] and [58], we adopt a self-attention mechanism to automatically parse the query.

Given a query with L words $r = \{w_l\}_{l=1}^L$, we first transform them into the word embeddings $\{\mathbf{e}_l\}_{l=1}^L$ by adopting a non-linear mapping function or pre-trained word embeddings (e.g., GloVe [37]). Then, we feed the word embeddings to a Bi-LSTM model [42] to obtain the forward and backward hidden state vectors, and the word representations $\mathbf{h} = \{\mathbf{h}_l\}_{l=1}^L$ are obtained by concatenating the forward and backward hidden vectors of the words. To obtain the language representations, we require calculating the attention score for each word in each module. For this purpose, we apply a fully connected layer to the representations, and then normalise the scores with a softmax function. In practice, we calculate the attention scores of basic attributes a_l^{att}, absolute location a_l^{loc}, visual relationship $a_l^{rel\_vis}$ and relative location $a_l^{rel\_loc}$ as follows:

$$a_l^{type} = \frac{\exp\left(\mathbf{w}_a^{type\top}\mathbf{h}_l\right)}{\sum_{k=1}^L \exp\left(\mathbf{w}_a^{type\top}\mathbf{h}_k\right)}, \tag{2}$$

where $type \in \{att, loc, rel\_vis, rel\_loc\}$, $\mathbf{w}_a^{type} \in \mathbb{R}^{d_w}$ is the parameters unique for each module, and d_w is the dimension of the word embeddings. Based on the attention scores for each module (i.e., $\mathbf{a}^{att}$, $\mathbf{a}^{loc}$, $\mathbf{a}^{rel\_vis}$, $\mathbf{a}^{rel\_loc} \in \mathbb{R}^L$), we obtain each type of the language representations by:

$$\mathbf{s}^{type} = \sum_{l=l}^{L} a_l^{type} \cdot \mathbf{e}_l. \tag{3}$$

3.3 Transformer-based Object Attention Module

In this module, we seek to find the relevant objects according to the information of the properties (i.e., the basic attributes and absolute location) in a query. To achieve this, we propose a Transformer-based object attention module. Specifically, we represent each object with its attribute and location features. Then, under the guidance of $\mathbf{s}^{att}$ and $\mathbf{s}^{loc}$, we update the object representations by considering the intra- and inter-modality relationships, which can be achieved in a Transformer.

3.3.1 Basic Attributes Representation. The basic attributes describe the object attributes (e.g., shape and material), which are included in the visual features. Thus, we adopt the visual features to represent the basic attributes. Specifically, the visual feature $\mathbf{u}_i$ of object i is obtained by adopting a pre-trained feature extractor (e.g., ResNet-101 [14]). Then, we obtain the basic attributes representation of object i by encoding the visual features by a multi-layer perception (MLP) f^u, i.e., $\mathbf{x}_i^{att} = f^u(\mathbf{u}_i)$.

3.3.2 Absolute Location Representation. The absolute location represents the absolute location of the objects in an image. Suppose the width and height of the image are represented as $[W, H]$, and the top-left corner coordinate, bottom-right corner coordinate, width and height of object i are represented as $[x_{tl_i}, y_{tl_i}, x_{br_i}, y_{br_i}, w_i, h_i]$, then the location feature of object i is represented as a 5-dimensional vector $\mathbf{l}_i = [\frac{x_{tl_i}}{W}, \frac{y_{tl_i}}{H}, \frac{x_{br_i}}{W}, \frac{y_{br_i}}{H}, \frac{w_i \cdot h_i}{W \cdot H}]$. It denotes the top-left and bottom-right corner coordinates of the object region (normalised between 0 and 1) and its relative area (i.e., the ratio of the bounding box area to the image area). Since the visual features may also contain the location information of the object from the background, we concatenate the visual feature with the location feature to obtain the location representation of object i as $\mathbf{x}_i^{loc} = [f^u(\mathbf{u}_i), f^l(\mathbf{l}_i)]$, where f^l is an MLP and $[\cdot, \cdot]$ denotes for concatenation.

3.3.3 Transformer-Based Object Attention Module. Considering the inter-modality relationships (i.e., between the queries and objects) only to localise the target may not be enough to achieve better performance, while introducing the intra-modality relationships (i.e., between the objects) would be helpful. Thus, it is necessary to consider both the inter- and intra-modality relationships when localising the target. Fortunately, the multimodal Transformer meet our need. Specifically, following M4C [19], we apply a stack of L Transformer layers over the list of the query features $\mathbf{s}^{obj}$ and object features $\mathbf{X}^{obj} = \{\mathbf{x}_i^{obj}\}_{i=1}^N$, where $obj \in \{att, loc\}$. In this way, the Transformer is able to take advantage of the self-attention mechanism to consider both the inter- and intra-modality relationships to update the features, which can be formulated as:

$$[\hat{\mathbf{s}}^{obj}, \hat{\mathbf{X}}^{obj}] = \Phi([\mathbf{s}^{obj}, \mathbf{X}^{obj}]; \mathbf{W}^{obj}), \tag{4}$$

where $[\cdot, \cdot]$ denotes for concatenation, $\mathbf{W}^{obj}$ is the learnable parameters of the Transformer. The overall process is shown in (a) of Figure 4.

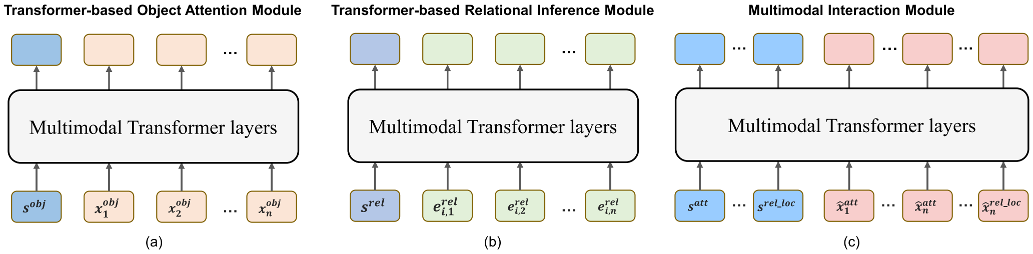
Transformer-based Object Attention Module
Transformer-based Relational Inference Module
Multimodal Interaction Module
Multimodal Transformer layers
Multimodal Transformer layers
Multimodal Transformer layers
s^{obj}
x_1^{obj}
x_2^{obj}
...
x_n^{obj}
s^{rel}
$e_{i,1}^{rel}$
$e_{i,2}^{rel}$
...
$e_{i,n}^{rel}$
s^{att}
...
s^{rel_loc}
$\hat{x}_1^{att}$
...
$\hat{x}_n^{att}$
...
$\hat{x}_n^{rel\_loc}$
(a)
(b)
(c)

Fig. 4. Overview of the Multimodal Transformer layers in Trans-RINet for visual relational reasoning.

3.4 Transformer-based Relational Inference Module

To fully comprehend the complex logic in the query, we consider constructing a relational graph as a reasoning basis to conduct reasoning. To this end, we first transform the image into a graph, where the nodes correspond to the objects and the edges denote their relationships. As referred to in Section 2.3, for each object, adopting one-step reasoning cannot guarantee to capture the multi-hop relationships between other objects, which makes it difficult to understand the complex logic. Thus, for each node, we devise a multi-step reasoning method that adopts GGNNs to progressively update the node representations by aggregating the information from the neighbourhoods, guided by $\mathbf{s}^{rel\_vis}$ and $\mathbf{s}^{rel\_loc}$, respectively. Before introducing the multi-step reasoning method, we first detail how to construct relational graphs.

3.4.1 Graph Construction. We construct a directed graph $\mathcal{G}=\{\mathcal{V},\mathcal{E}\}$ over the objects O, where $\mathcal{V}=\{v_i\}_{i=1}^{N}$ is the node set and $\mathcal{E}=\{e_{ij}\}_{i,j=1}^{N}$ is the edge set. Each node v_i corresponds to an object $o_i \in \{O\}$ and each edge e_{ij} denotes the relationships between the objects o_i and o_j.

3.4.2 Visual Relationship Representation. The visual relationship describes the relationships between objects, including but not limited to the attribute relationships between two objects (e.g., A and B are in the same colour) or the interaction between objects (e.g., A holds B). To this end, we concatenate these two object features to represent the visual relationship. As mentioned in Section 3.3, the feature of object i is obtained by concatenating its visual feature and location feature $\mathbf{x}_i^{vis}=[\mathbf{u}_i,\mathbf{l}_i]$. Then, the edge representation of the visual relationship between object i and object j is calculated by using an MLP f^{rel_vis} to encode these two object features:

$$\mathbf{e}_{ij}^{rel\_vis}=f^{rel\_vis}\left(\left[\mathbf{x}_i^{vis},\mathbf{x}_j^{vis}\right]\right). \tag{5}$$

3.4.3 Relative Location Representation. The relative location describes the displacement between two objects, implying the spatial correlation between the objects. Here, we represent the spatial relationship between two objects (e.g., v_i and v_j) as $\tilde{\boldsymbol{e}}_{ij}=[\frac{x_{tl_j}-x_{c_i}}{w_i},\frac{y_{tl_j}-y_{c_i}}{h_i},\frac{x_{br_j}-x_{c_i}}{w_i},\frac{y_{br_j}-y_{c_i}}{h_i},\frac{w_j\cdot h_j}{w_i\cdot h_i}]$, where $[x_{c_i},y_{c_i},w_i,h_i]$ is the centre coordinate, width and height of object i, and $[x_{tl_j},y_{tl_j},x_{br_j},y_{br_j},w_j,h_j]$ is the top-left corner coordinate, bottom-right corner coordinate, width and height of the j-th object, respectively. Considering the query like “A is at the right of B”, the object “B” plays an important role in the understanding of the location relationship. In this sense, we include the visual feature and location feature of the object j for the relative location representations of the object i. Thus, the edge representation of the relative location is calculated as follows:

$$\mathbf{e}_{ij}^{rel\_loc}=f^{rel\_loc}\left(\left[\tilde{\boldsymbol{e}}_{ij},\mathbf{x}_j^{vis}\right]\right), \tag{6}$$

where f^{rel_loc} is an MLP. After learning the edge representation, we would obtain two types of visual relational graphs.

3.4.4 Transformer-based Relational Inference Module. Similar to that in Section 3.3.3, we also introduce the multimodal Transformer to consider the inter- and intra-modality relationships. Specifically, we apply a stack of L Transformer layers over the list of the query features $\mathbf{s}^{rel}$ and edge features $\mathbf{e}^{rel}$, where $rel \in \{rel\_vis, rel\_loc\}$. For one sample, we suppose the inputs are $\mathbf{s}^{rel} \in \mathbb{R}^{d_e}$ and $\{\mathbf{e}_{ij}^{rel}\}_{i,j=1}^{N}$, where $\mathbf{e}_{ij}^{rel} \in \mathbb{R}^{d_e}$. We concatenate these two types of features and then send them into the multimodal Transformer to obtain the updated features, which can be formulated as:

$$\left[\hat{\mathbf{s}}^{rel}, \left\{\hat{\mathbf{e}}_{ij}^{rel}\right\}_{j=1}^{N}\right] = \Phi\left(\left[\mathbf{s}^{rel}, \left\{\mathbf{e}_{ij}^{rel}\right\}_{j=1}^{N}\right]; \mathbf{W}^{rel}\right), \tag{7}$$

where $[\cdot,\cdot]$ is a concatenation operation and $\mathbf{W}_{rel}$ is the learnable parameters of the Transformer. As shown in Equation (7), it updates the relational features for object i only. In this sense, we would repeat this operation for N times to obtain the updated relationships feature matrix $\{\hat{\mathbf{e}}_{ij}^{rel}\}_{i,j=1}^{N}$. The overall process is shown in (b) of Figure 4.

To achieve the multi-step reasoning, based on the relational graphs, we introduce GGNNs [28]to progressively update the node representations by aggregating the relational information from the neighbourhoods. Specifically, GGNNs are a propagation model, whose mechanism is similar to recurrent neural networks. Before performing the multi-step reasoning with GGNNs, we require a propagation matrix. As shown in Equation (7), the Transformer adopts the self-attention mechanism to calculate the similarity between each entity, which contains a wealth of guidance information. Thus, we seek to use the self-attention matrix from the Transformer as the propagation matrix. As shown in Equation (1), we would obtain the attention matrix $\mathbf{A} \in \mathbb{R}^{l\times l}$. Generally speaking, self-attention usually contains some heads H, aiming to focus on different directions. In this sense, the attention matrix can be reformulated as $\mathbf{A} \in \mathbb{R}^{H\times l\times l}$. Thus, the self-attention matrix in the Transformer-based relational inference module is $\mathbf{A} \in \mathbb{R}^{N\times H\times(N+1)\times(N+1)}$, where $(N+1)$ means the input contains N object relational features and 1 language feature. Note that we obtain the attention matrix at the last Transformer layer.

Obtain the Propagation Matrix. Based on the self-attention matrix $\mathbf{A}$ in the Transformer-based relational inference module, we provide two methods to obtain the propagation matrix. One intuitive way is to adopt mean operations on the dimension of the attention head as $\mathbf{A}_{mean} \in \mathbb{R}^{N\times 1\times(N+1)\times(N+1)}$. Moreover, according to the self-attention in Equation (1), the softmax operates on the features at the last dimension. To maintain the softmax operation at the last dimension and reduce the number of dimensions in the attention matrix, we would also adopt the mean operation at the penultimate dimension of the attention matrix. Finally, since the attention matrix contains the information about the language, we remove the dimension of the language feature in the attention matrix, and obtain the final propagation matrix $\mathbf{A}_{mean} \in \mathbb{R}^{N\times 1\times 1\times N}$.

Another method to obtain the propagation matrix is to imitate the multi-head self-attention fusion mechanism. Specifically, given the initial attention matrix $\mathbf{A} \in \mathbb{R}^{N\times H\times(N+1)\times(N+1)}$, to reduce the dimension, we first adopt the mean operation at the penultimate dimension of the attention matrix to retain the softmax operation at the last dimension (i.e., $\mathbf{A}_{head} \in \mathbb{R}^{N\times H\times 1\times(N+1)}$). Next, we remove the dimension of the language feature in the attention matrix, and obtain the final propagation matrix with attention head $\mathbf{A}_{head} \in \mathbb{R}^{N\times H\times 1\times N}$. We will conduct the ablation studies about these two propagation matrices in Section 4.6.

Multi-Step Reasoning. Based on the propagation matrix $\mathbf{A}_{mean}^{rel}$, for each object node in the graph, the propagation process of the tth step is defined as follows:

$$\begin{aligned}\mathbf{z}_i^{rel,(t)} &= \tanh\left({\mathbf{a}_{mean,i}^{rel}}^{\top}\left[\mathbf{h}_1^{rel,(t-1)}; \ldots; \mathbf{h}_N^{rel,(t-1)}\right]\right),\\ \mathbf{h}_i^{rel,(t)} &= \mathrm{GRUCell}\left(\mathbf{z}_i^{rel,(t)}, \mathbf{h}_i^{rel,(t-1)}\right),\end{aligned} \tag{8}$$

where $rel \in \{rel\_vis, rel\_loc\}$ and $\mathbf{a}^{rel}_{mean,i}$ is the ith row of the propagation matrix $\mathbf{A}^{rel}_{mean} \in \mathbb{R}^{N \times N}$. GRUCell is a one-step GRU [8] updating mechanism. $\mathbf{h}_i^{rel,(t)}$ is the hidden state of the ith object at step t.

Moreover, with the propagation matrix A^{rel}_{head} at hand, we first transpose it as $\mathbf{A}^{rel}_{head} \in \mathbb{R}^{H \times N \times N}$, and then split the last dimension of the hidden state $\mathbf{h}_i^{(rel),t}$ into $H \times \frac{d_e}{H}$. In this way, we obtain the hidden states $\mathbf{h}^{(rel),t} \in \mathbb{R}^{H \times N \times \frac{d_e}{H}}$. Similar to that using $\mathbf{A}^{rel}_{mean}$, for each node in the graph, the propagation process of the t-th step can be calculated as follows:

$$\begin{aligned} \mathbf{z}_i^{rel,(t)} &= \tanh\left(\mathbf{a}^{rel}_{head,i}\left[\mathbf{h}_1^{rel,(t-1)};\ldots;\mathbf{h}_N^{rel,(t-1)}\right]\right), \\ \mathbf{h}_i^{rel,(t)} &= \text{GRUCell}\left(\text{flatten}\left(\mathbf{z}_i^{rel,(t)}\right), \text{flatten}\left(\mathbf{h}_i^{rel,(t-1)}\right)\right), \end{aligned} \tag{9}$$

where $\mathbf{a}^{rel}_{head,i} \in \mathbb{R}^{H \times 1 \times N}$, $\mathbf{h}_i^{rel,(t-1)} \in \mathbb{R}^{H \times 1 \times \frac{d_e}{H}}$, flatten means that flattening the feature into one dimension.

Note that no matter which propagation matrix is used, we obtain the initial hidden state $\mathbf{h}_i^{rel,(0)}$ by adopting the mean operation on the penultimate dimension of the updated relationships feature matrix $\mathbf{h}_i^{rel,(0)} = \hat{\mathbf{e}}_i^{rel} = \text{mean}_j(\{\hat{\mathbf{e}}_{ij}^{rel}\}_{i,j=1}^N)$.

In this way, at each time step, we update the node representations by aggregating the information from the neighbourhoods guided by the propagation matrix. Thus, the multi-step updating method enables each object to aggregate the relational information from the neighbourhoods and progressively understand the complex logic in the queries. After T time steps for propagation, the final representation for the ith node can be obtained by:

$$\hat{\mathbf{x}}_i^{rel} = \mathbf{h}_i^{rel,(T)}. \tag{10}$$

3.5 Multimodal Interaction Module

Since we decompose the query into four types, each type of query is closely related. Thus, it is necessary to promote the interaction between each module corresponding to each type of information of the queries. Inspired by the self-attention mechanism of free interaction between each entity, we also adopt a multimodal Transformer to connect the outputs generated from each module. Specifically, with the four types of outputs at hand, namely, $\mathbf{s}^{type}$ and $\hat{\mathbf{x}}_i^{type}$, where $type \in \{att, loc, rel\_vis, rel\_loc\}$, we regard each feature as one unique entity, and then promote them to interact with each other in the stack of L Transformer layers, which can be formulated as:

$$\begin{aligned} \mathbf{F}_{input} &= \left[\mathbf{s}^{att}, \ldots, \mathbf{s}^{rel\_loc}, \hat{\mathbf{x}}_i^{att}, \ldots, \hat{\mathbf{x}}_i^{rel\_loc}\right], \\ \left[\tilde{\mathbf{s}}^{att}, \ldots, \tilde{\mathbf{s}}^{rel\_loc}, \tilde{\mathbf{x}}_i^{att}, \ldots, \tilde{\mathbf{x}}_i^{rel\_loc}\right] &= \Phi(\mathbf{F}_{input}; \mathbf{W}^{fuse}), \end{aligned} \tag{11}$$

where $[\cdot,\cdot]$ denotes for concatenation, and $\mathbf{W}^{fuse}$ is the learnable parameters of the Transformer. In this way, each module interacts with the other freely, improving information circulation, which helps make a prediction. The overall process is shown in (c) of Figure 4.

3.6 Matching Function and Loss Function

3.6.1 Matching Function. To localise the target object, we choose the object with the highest matching score. To this end, we require to obtain the matching scores for each object. Specifically, we devise a matching function to predict the final scores by matching the language representations and the corresponding object representations, where the object representations and the language representation are obtained by Equation (11). The matching score p_i^{type} between the language

ALGORITHM 1: Training details of our Trans-RINet.

```
Require: Training data $\{(I_k, r_k, \boldsymbol{y}_k)\}_{k=1}^{K}$, updating step $T$, training iterations $D$.
  repeat
    // Language Decomposition Module, type ∈ {att, loc, rel_vis, rel_loc}
    Calculate the word attention $\left\{a_l^{type}\right\}_{l=1}^{L}$ using Equation (2).
    Obtain the query representations $\boldsymbol{s}^{type}$ using Equation (3).
    // Transformer-based Object Attention Module, obj ∈ {att, loc}
    Obtain the object representation $\hat{\mathbf{x}}_i^{obj}$ using Equation (4).
    // Transformer-based Relational Inference Module, rel ∈ {rel_vis, rel_loc}
    Construct the relational graphs $\mathcal{G}$ via Equations (5) and (6).
    repeat
      Obtain the object representation $\hat{\mathbf{x}}_i^{rel}$ through GGNNs using Equation (8).
    until Reasoning steps reach T.
    Let each module interact with each other by Equation (11).
    Calculate the matching scores $p_i^{type}$ between $\tilde{\mathbf{s}}^{type}$ and $\tilde{\mathbf{x}}_i^{type}$ using Equation (12).
    Obtain the final scores $\boldsymbol{p}$ using Equation (14).
    Update the parameters of Trans-RINet by minimising the loss in Equation (15).
  until Iteration times reach D.
```

representation $\tilde{\mathbf{s}}^{type}$ and the ith object representation $\tilde{\mathbf{x}}_i^{type}$ can be calculated as follows:

$$p_i^{type} = (\tilde{\mathbf{s}}^{type})^\top \tilde{\mathbf{x}}_i^{type}, \tag{12}$$

where $type \in \{att, loc, rel\_vis, rel\_loc\}$. Following [47] and [58], we calculate four weights to represent the contributions of each module. To this end, we apply a fully connected layer to the mean embedding vector $\mathbf{e} = \text{mean}_l(\{\mathbf{e}_l\}_{l=1}^{L})$, where $\mathbf{e}_l$ is the word embedding of lth word in the query. The calculation of the weights is as follows:

$$[w^{att}, w^{loc}, w^{rel\_vis}, w^{rel\_loc}] = \text{softmax}\left(\mathbf{W}_s \mathbf{e}\right), \tag{13}$$

where $\mathbf{W}_s \in \mathbb{R}^{4 \times d_w}$ is the parameters of the fully connected layer, and d_w is the dimension of the word embedding. For object i, the final matching score p_i is calculated by weighted summing up of the p_i^{type} with four weights:

$$p_i = \sum_{type} w^{type} p_i^{type}. \tag{14}$$

3.6.2 Loss Function. Since we localise the referent from the candidate objects in the image, we regard this task as a multi-class classification task. The probability of object i being the referent is calculated as $\tilde{p}_i = \frac{\exp(p_i)}{\sum_{j=1}^{N} \exp(p_j)}$. Intuitively, we would use the cross-entropy loss as the loss function:

$$L = -\sum_{i=1}^{N} y_i \cdot \log\left(\tilde{p}_i\right), \tag{15}$$

where y_i is 1 when object i is the referent and 0 otherwise. The training details of our Trans-RINet can be found in Algorithm 1.

4 EXPERIMENTS

In this section, we evaluate our Trans-RINet on complex referring expression comprehension datasets, including the synthetic dataset (i.e., CLEVR-Ref+ [29]) and real-world dataset (i.e.,

Table 1. Details of CLEVR-Ref+ and CLEVR-CoGent Datasets

CLEVR-Ref+ dataset			CLEVR-CoGent dataset		
Split	#Expressions	#Images	Split	#Expressions	#Images
Train set	222569	63057	Train set	222474	62981
Val set	47731	13534	ValA set	47638	13518
			ValB set	47723	13489

Ref-Reasoning [56]). Moreover, we further conduct experiments on the CLEVR-CoGenT [29] dataset to demonstrate the effective generalisation ability of our Trans-RINet. Last, the ablation studies and visualisation analyses are conducted to show the contribution of each module in our method.

4.1 Datasets and Metrics

CLEVR-Ref+ [29] and **CLEVR-CoGenT** [29] are synthetic datasets whose images and queries are generated by the code automatically. These datasets are approximately unbiased by adopting a uniform sampling strategy. Moreover, the CLEVR-CoGent dataset has two different conditions (i.e., Condition A and Condition B) that contain different object attributes.

Ref-Reasoning [56] is a large-scale real-world dataset, which contains semantically rich queries that describe objects, attributes, and their direct or indirect relations. Specifically, it includes approximately 79k queries and 83k images.

Metrics. We adopt Top-1 Accuracy to evaluate the methods. Specifically, we conduct experiments on the referring expression comprehension task, which involves localising a single target object within an image. To evaluate the accuracy of our method on the CLEVR-Ref+ and CLEVR-CoGenT datasets, we follow the prior works [17] by collecting and using only those expressions that describe a single target object for both training and testing. The specifics of this data collection process can be found in Table 1.

4.2 Implementation Details

On the CLEVR-Ref+ and CLEVR-CoGenT dataset, we follow the settings in [17] and obtain the region feature of each object by using ResNet101 [14] pre-trained on ImageNet. We evaluate our method in two settings: (i) "det": bounding boxes detected by Mask-RCNN [13];[1] (ii)"gt": ground truth bounding boxes. The queries are encoded with a non-linear mapping function. On the Ref-Reasoning dataset, we adopt the 2048-dimensional visual features of objects provided by the dataset and encode the queries with GloVe word embedding [37]. During training, we use Adam [27] optimiser with the learning rate 1e-4. The updating step of GGNNs is set to 3. Following [47], we set the dimensions of the language and object representations to 512. We implement our method based on PyTorch [36].

On the CLEVR-Ref+ and CLEVR-CoGenT datasets, the batch size is set to 30 images, which means that we feed 30 images and all the queries associated with these images to the network for each training iteration. On the Ref-Reasoning dataset, for MGA-Net, due to the limitation of the GPU memory, we set the batch size of 32 queries, which means we feed the MGA-Net with 32 queries and the associated images as the input. Similarly, the batch size is set to 25 queries for Trans-RINet. For Trans-RINet, on the CLEVR-Ref+ and CLEVR-CoGent datasets, we set all the multimodal transformers with 2 layers and 4 self-attention heads. All the multimodal transformers are set to 3 layers and 4 self-attention heads on the Ref-Reasoning dataset. We conduct most

[1] We use the pre-trained Mask-RCNN from https://github.com/kexinyi/ns-vqa.

Table 2. Comparisons with State-of-the-Art Methods on the CLEVR-Ref+ Dataset in Terms of Accuracy (%)

Method	Accuracy (%)
Stack-NMN [16]	56.5
SLR [59]	57.7
MAttNet [58]	60.9
GroundeR [41]	61.7
LCGN [17]	74.8
LCGN [17] (gt)	76.0
RPR [23] (gt)	77.7
MGA-Net [61] (det)	80.1
MGA-Net [61] (gt)	80.8
Trans-RINet (det)	93.2
Trans-RINet (gt)	**94.1**

experiments on one TITAN Xp GPU, except that we use V100 GPU on the Ref-Reasoning dataset in Trans-RINet. The source code and the pre-trained models are available at Trans-RINet.

4.3 Evaluation on the CLEVR-Ref+ Dataset

We compare our Trans-RINet with several state-of-the-art methods, i.e., Stack-NMN [16], SLR [59], MAttNet [58], GroundeR [41], LCGN [17], RPR [23], and MGA-Net [61]. From Table 2, our method outperforms all compared methods. Specifically, MAttNet performs reasoning in a one-step manner, ignoring the issue of long-chain reasoning, and thus obtains the accuracy of 60.9%. Beneficial from the multi-step reasoning, LCGN and RPR surpass MAttNet by around 14%, and 17%, respectively. However, these methods encode the queries holistically, without distinguishing different types of information. Moreover, MGA-Net decomposes the query into four types and conducts multi-step reasoning on the relational graphs with GGNNs, and obtains an accuracy of 80.1% in the "det" setting. Moreover, MGA-Net would further improve the accuracy to 80.8% when in the "gt" setting.

However, MGA-Net still encounters two limitations: First, only considering the inter-modality relationships may not be enough to tackle the complex queries, while introducing the intra-modality relationships would improve the reasoning ability of the model. Second, MGA-Net tackles each module isolatedly, which however, may achieve unsatisfactory performance. To address the above issues, our Trans-RINet introduces the multimodal transformer that can promote the entities connecting and interacting with each other freely. Thus, as shown in Table 2, our Trans-RINet achieves an accuracy of 93.2%, which outperforms MGA-Net by approximately 13%. When using the ground truth bounding boxes, Trans-RINet further improves the performance to 94.1%. These results demonstrate the effectiveness and necessity to further consider the intra-modality relationships and the interaction between each module.

Moreover, we show the curve of training and testing Loss and Accuracy of Trans-RINet in Figure 5. From these results, we have the following observations: (1) As the training epoch increases, the training and testing loss converges gradually. (2) The training and testing accuracy are also converged with the training process going on. These results demonstrate our Trans-RINet has a stable training procedure.

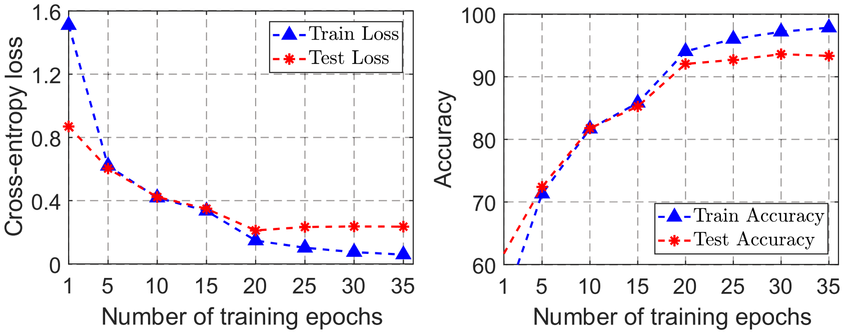

Fig. 5. The curve of training and testing Accuracy and Loss of Trans-RINet on the CLEVR-Ref+ dataset. Trans-RINet uses $\mathbf{A}_{head}$ as propagation matrix, and sets updating step T to 3.

Table 3. Comparisons of Different Methods on the Ref-Reasoning Dataset about Validation and Test Splits in Terms of Accuracy (%)

Method	val (%)	test (%)
CNN	12.36	12.15
CNN+LSTM	42.38	42.43
CMRIN [54]	47.40	47.69
DGA [55]	48.95	49.51
SGMN [56]	51.59	51.95
RPR [23]	52.22	53.02
MGA-Net [61]	56.37	56.67
Trans-RINet	**56.75**	**57.55**

4.4 Evaluation on the Ref-Reasoning Dataset

To further demonstrate the effectiveness of our methods, we compare some state-of-the-art methods (e.g., CMRIN [54], DGA [55], SGMN [56], and RPR [23]) on the large-scale real-world dataset (i.e., the Ref-Reasoning [56] dataset). We present the experimental results in Table 3. From these results, we have the following observations: (1) Our MGA-Net achieves the best performance compared with the state-of-the-art methods. Specifically, MGA-Net outperforms CMGIN, DGA, SGMN, and RPR by approximately 8%, 6%, 4%, and 3%, respectively, on both the validation and test split. (2) Moreover, our Trans-RINet further surpasses the MGA-Net by approximately 1% on the test split. These results demonstrate the effectiveness of our MGA-Net and Trans-RINet on the complex real-world dataset.

4.5 Evaluation on the CLEVR-CoGenT Dataset

To demonstrate the generalisation ability of our method, we further conduct experiments on the CLEVR-CoGent dataset. Specifically, the training set is in Condition A, while the validation set is in Conditions A and B. From the results in Table 4, our Trans-RINet outperforms SLR [59], MAttNet [58], and MGA-Net [61] by a large margin. Specifically, our Trans-RINet outperforms MGA-Net by approximately 10% on valA and valB, either ground truth or detected bounding box. These results demonstrate that further considering the intra-modality relationships and the interaction between each module would promote the generalisation ability of the models.

Table 4. Comparisons with Baselines on the CLEVR-CoGenT Dataset (valA & B) in Terms of Accuracy

Method	valA	valB
SLR [59]	0.63	0.59
MAttNet [58]	0.64	0.63
MGA-Net [61] (det)	0.82	0.76
MGA-Net [61] (gt)	0.83	0.78
Trans-RINet (det)	0.93	0.86
Trans-RINet (gt)	**0.94**	**0.87**

"det" means detected bounding boxes, while "gt" denotes ground-truth bounding boxes.

Table 5. Impact of the Four Modules Corresponding to Each Type of the Queries in Trans-RINet on the CLEVR-Ref+ Dataset in Terms of Accuracy (%)

Module	Accuracy (%)
att	73.52
att + loc	78.89
att + loc + rel_vis	80.95
att + loc + rel_loc	83.96
att + loc + rel_vis + rel_loc	**85.81**

Note that on the CLEVR-Ref+ and CLEVR-CoGenT datasets, our Trans-RINet achieves comparable performance between the "det" and "gt" settings. The main reason is that the scenes of the images are pure and simple, without introducing noise, and thus the detectors are able to detect the objects in the scenes accurately.

4.6 Ablation Study

4.6.1 Effect of Four Modules. Our Trans-RINet contains four modules corresponding to each type of the queries (i.e., the basic attributes (att), absolute location (loc), relative location (rel_loc), and visual relationship (rel_vis)). We conduct ablation studies on the CLEVR-Ref+ dataset to evaluate each module. We use the "gt" setting and set the updating step T to 3.

From the results in Table 5, directly adopting the basic attribute to localise the target obtains the accuracy of 73.52% (Row 1). When introducing the absolute location module, the accuracy would further improve (Row 2). Row 3 or Row 4 show the benefits brought by the relational reasoning modules, demonstrating the effectiveness of multi-step reasoning. Moreover, when combining four modules, our Trans-RINet obtains the best accuracy of 85.81%. The above results imply that distinguishing different types of information in the query is crucial for relational reasoning.

4.6.2 Effect of the Multimodal Interaction Module. To evaluate the multimodal interaction module mentioned in Section 3.5, we conduct ablation studies about with or without multimodal interaction module based on Trans-RINet that uses $\mathbf{A}_{head}$ as the propagation matrix. From the results in Table 6, we have the following observations: (1) our Trans-RINet outperforms that w/o interaction for any updating steps, which demonstrates the effectiveness and necessity of promoting the interaction between each module; (2) With the increase of the updating step, the performance gap

Table 6. Ablation Studies about the Multimodal Interaction Module on the CLEVR-Ref+ Dataset

Dataset	Setting	Trans-RINet w/o interaction	Trans-RINet
CLEVR-Ref+ gt bbox	$T = 0$	80.92	82.51
	$T = 1$	83.32	85.93
	$T = 3$	**85.81**	93.60
	$T = 5$	84.34	**94.02**

We use the Trans-RINet as the base model with A_{head} as the propagation matrix. Note that "w/o interaction" denotes without introducing the multimodal interaction module.

Table 7. Ablation Studies about Different Types of Propagation matrix (i.e., $\mathbf{A}_{mean}$ and $\mathbf{A}_{head}$) on the CLEVR-Ref+, CLEVR-CoGenT (valA & valB), and Ref-Reasoning Datasets

Method	Dataset	CLEVR-Ref+		CLEVR-CoGenT (valA)		CLEVR-CoGenT (valB)		Ref-Reasoning	
-	Setting	detected bbox	gt bbox	detected bbox	gt bbox	detected bbox	gt bbox	val	test
$\mathbf{A}_{mean}$	T=1	84.81	87.44	85.17	87.34	79.01	81.37	56.39	56.74
	T=3	91.50	**94.19**	**93.59**	94.65	**86.81**	87.55	56.60	57.21
	T=5	92.08	94.02	91.13	91.28	84.28	84.10	56.74	57.45
$\mathbf{A}_{head}$	T=1	85.73	85.93	85.43	88.71	78.74	80.95	56.59	56.78
	T=3	**93.21**	93.60	91.88	**94.82**	84.65	**87.90**	**56.75**	**57.55**
	T=5	91.19	94.02	91.58	92.11	84.54	84.87	56.65	57.40

between Trans-RINet and Trans-RINet w/o interaction gradually increases, which demonstrates the importance of promoting the interaction of each module in the reasoning process.

4.6.3 Effect of Different Propagation Matrices (i.e., $\mathbf{A}_{mean}$ and $\mathbf{A}_{head}$). As mentioned in Section 3.4.4, we provide two types of propagation matrices when performing multi-step reasoning. Thus, in this part, we provide the ablation study to evaluate different propagation matrices. We show the experimental results in Table 7. From these results, both $\mathbf{A}_{mean}$ and $\mathbf{A}_{head}$ achieve comparable performance. These results demonstrate that in the reasoning process, adopting the mean operation on the dimension of the self-attention head has a similar effect to that of retaining the self-attention head, to some extent.

4.7 Effectiveness of Multi-step Reasoning

4.7.1 Quantitative Results. We conduct ablation studies on our methods (i.e., MGA-Net and Trans-RINet) about the multi-step reasoning by setting different updating steps T in GGNNs. As shown in Table 8, our methods with GGNNs ($T > 1$) perform better than those without GGNNs ($T = 0$) significantly regardless of the datasets, demonstrating the necessity and superiority of relational reasoning. Moreover, by increasing the updating step (from $T = 1$ to $T = 3$), the performance would further improve, which demonstrates the effectiveness of multi-step reasoning. However, in most cases, our methods with updating step $T = 3$ perform better than that with updating step $T = 5$. The experiments demonstrate an appropriate updating step would obtain the best performance.

To further investigate the relationship between the number of reasoning steps (T) and the model's performance, we conducted additional experiments on the CLEVR-Ref+ dataset for $T > 5$ (e.g., $T = 6, 7, 8$) and presented the results in Table 9. From the results, we have the following observations: (1) As the number of reasoning steps increases (from $T = 1$ to $T = 5$), the model's performance improves, highlighting the effectiveness of multi-step reasoning. Furthermore, our Trans-RINet with $T = 3$ achieves a performance comparable to that of the model equipped with

Table 8. Impact of the Updating Step T

MGA-Net								
Dataset	CLEVR-Ref+		CLEVR-CoGenT (valA)		CLEVR-CoGenT (valB)		Ref-Reasoning	
Setting	detected bbox	gt bbox	detected bbox	gt bbox	detected bbox	gt bbox	val	test
T=0	75.81	76.51	76.00	76.24	71.37	72.32	55.05	55.38
T=1	79.52	80.25	79.01	79.15	74.26	74.53	56.08	56.12
T=3	**80.18**	**80.87**	**82.02**	**82.90**	**76.60**	**78.15**	**56.37**	56.55
T=5	79.05	79.65	79.95	80.36	74.69	76.00	56.19	**56.67**
Trans-RINet								
T=0	80.11	82.51	80.86	81.77	73.92	74.92	55.90	56.32
T=1	85.73	87.44	85.43	88.71	79.01	81.37	56.59	56.78
T=3	**93.21**	**94.19**	**93.59**	**94.82**	**86.81**	**87.90**	**56.75**	**57.55**
T=5	92.08	94.02	91.58	92.11	84.54	84.87	56.74	57.45

We report the accuracy (%) of our methods (i.e., MGA-Net and Trans-RINet) with different updating step T on the CLEVR-Ref+, CLEVR-CoGenT (valA & valB), and Ref-Reasoning datasets.

Table 9. Impact of the Updating Step T

Dataset	Setting	Trans-RINet
CLEVR-Ref+ gt bbox	$T = 3$	93.60
	$T = 5$	**94.02**
	$T = 6$	93.61
	$T = 7$	93.38
	$T = 8$	87.54

We report the accuracy (%) of our method on the CLEVR-Ref+ dataset. We use the Trans-RINet as the base model with A_{head} as the propagation matrix.

$T = 5$, while requiring fewer updates. Hence, selecting $T = 3$ may be a better choice; (2) When T is greater than 5, the performance gradually declines, which may be attributed to the introduction of noise by conducting more reasoning steps, since not all expressions require reasoning steps higher than 5.

Note that on the Ref-Reasoning dataset, the performance improvement of our Trans-RINet compared with MGA-Net is not as substantial as that on the CLEVR-Ref+ and CLEVR-CoGent datasets. This phenomenon may be caused by the following reason. On the CLEVR-Ref+ and CLEVR-CoGent datasets, the scene of the image is simple, and the maximum number of objects is 10 without noises. Thus, when introducing the Transformer, our Trans-RINet would accurately capture the relationships and achieve better performance. However, on the Ref-Reasoning dataset, the maximum number of objects is more than 100, and thus the noise objects (e.g., background) may exist. Since self-attention can achieve interaction between each entity freely, the noises may affect the model performance.

4.7.2 Visualisation. To evaluate the ability of long-chain reasoning of our Trans-RINet, we visualise the results at each reasoning step. Specifically, we train our Trans-RINet on the CLEVR-Ref+ dataset with the “gt” setting, and the updating step T in GGNNs is set to 3. Then, we obtain the score for each node by matching the node representations with the language representations at each updating step (i.e., $T = 0, 1, 2, 3$). Note that in the attention maps, the darker the colour, the higher the confidence. As shown in Figure 6, our Trans-RINet decomposes the queries reasonably, and achieves accurate localisation at each reasoning step, which demonstrates the superior reasoning ability of our Trans-RINet.

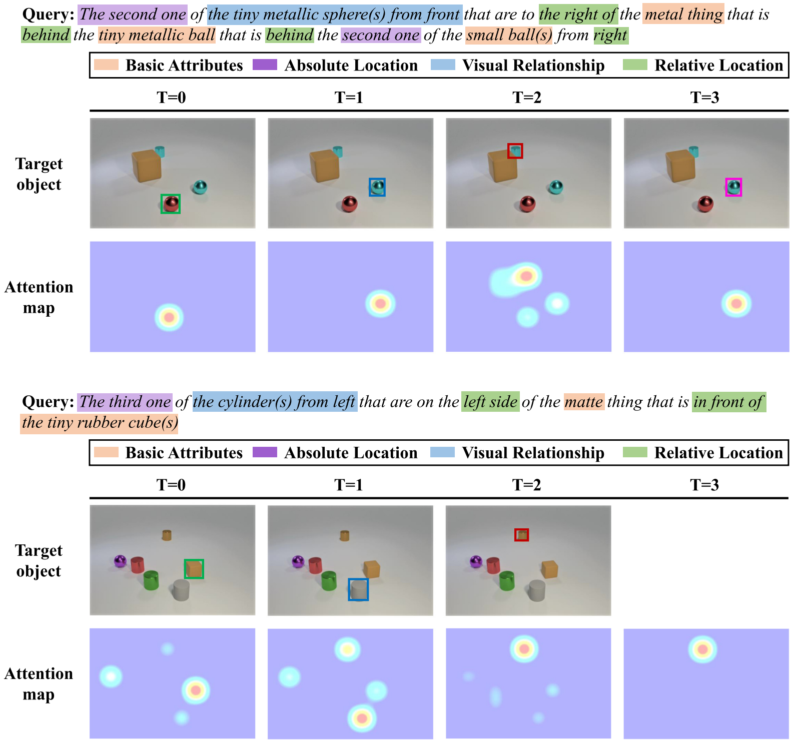
Query: The second one of the tiny metallic sphere(s) from front that are to the right of the metal thing that is behind the tiny metallic ball that is behind the second one of the small ball(s) from right
Basic Attributes
Absolute Location
Visual Relationship
Relative Location
T=0
T=1
T=2
T=3
Target object
Attention map
Query: The third one of the cylinder(s) from left that are on the left side of the matte thing that is in front of the tiny rubber cube(s)
Basic Attributes
Absolute Location
Visual Relationship
Relative Location
T=0
T=1
T=2
T=3
Target object
Attention map

Fig. 6. Examples of multi-step reasoning on complex queries with Trans-RINet. We visualise the attention maps and mark the target object with a bounding box for each step.

Moreover, to better comprehend our Trans-RINet, we show the failure cases of our Trans-RINet on the CLEVR-Ref+ dataset in Figure 7. From the results, we draw the following observations: (1) For simple samples that do not require too many reasoning steps, our Trans-RINet may make incorrect predictions even when the reasoning process is correct. For instance, as shown in Figures 7(a) and 7(b), the expressions require only one and two reasoning steps, respectively. Our Trans-RINet exhibits correct reasoning steps, but misses the correct answer due to excessive reasoning steps. These results indicate that conducting too many reasoning steps for simple samples could introduce noise, leading to incorrect predictions even with correct reasoning steps. In the future, we will extend our method to achieve the dynamic reasoning steps for different samples. (2) Although our Trans-RINet makes a correct final prediction, the intermediate reasoning process may be incorrect. For example, as shown in Figure 7(c), our Trans-RINet makes a wrong prediction at $T = 0$. These results demonstrate our Trans-RINet possesses the capability to correct errors to some extent.

5 CONCLUSION

In this article, we have proposed a Trans-RINet for complex visual relational reasoning. To fully comprehend the textual information, we decompose the complex query into four types and devise a module for each type. In each module, we consider both the intra- (i.e., between the objects) and inter-modality relationships (i.e., between the queries and the objects) to help improve the reasoning ability. Moreover, we construct relational graphs to represent the objects and their relationships. Based on the relational graphs, we devise a graph inference network that adopts

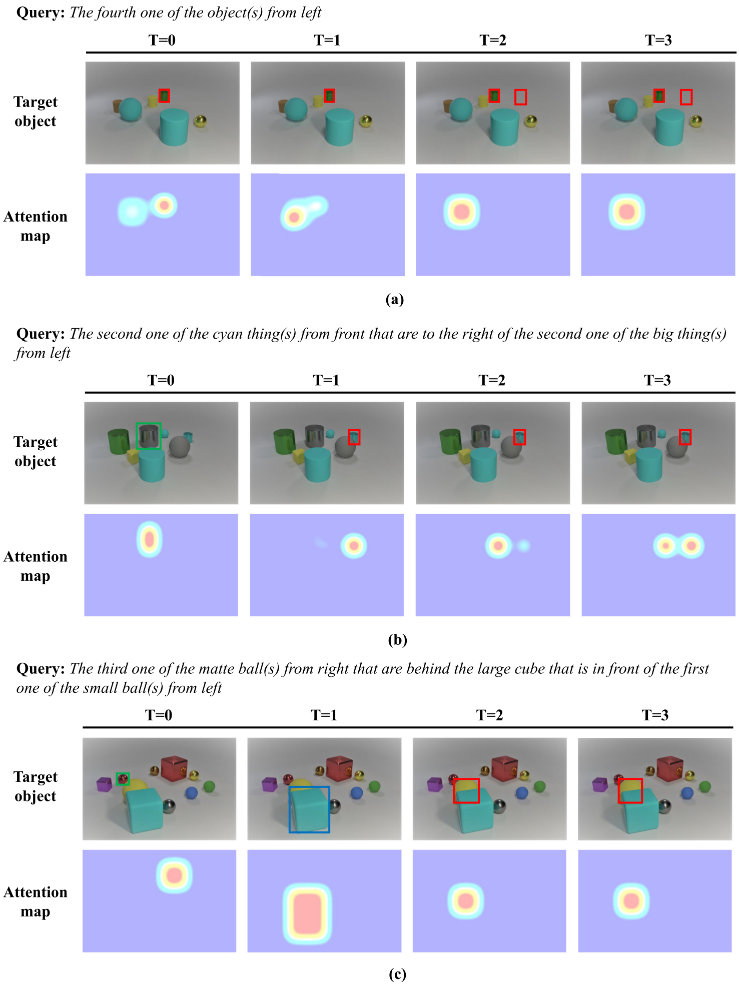
Query: The fourth one of the object(s) from left
T=0
T=1
T=2
T=3
Target object
Attention map
(a)
Query: The second one of the cyan thing(s) from front that are to the right of the second one of the big thing(s) from left
T=0
T=1
T=2
T=3
Target object
Attention map
(b)
Query: The third one of the matte ball(s) from right that are behind the large cube that is in front of the first one of the small ball(s) from left
T=0
T=1
T=2
T=3
Target object
Attention map
(c)

Fig. 7. Failure cases of our Trans-RINet on the CLEVR-Ref+ dataset [29]. We visualise the attention maps and mark the target object with a bounding box for each step.

GGNN to update the graph representations progressively. In this way, Trans-RINet encodes the multi-step relationships among objects. Before making the decision, we devise a multimodal fusion module to make each module communicate with each other, since each module is closely related. Extensive experiments show the effectiveness and the superior relational reasoning ability of our Trans-RINet.

REFERENCES

[1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. 2018. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3674–3683.

[2] Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 39–48.

[3] Mohit Bajaj, Lanjun Wang, and Leonid Sigal. 2019. G3raphGround: Graph-based language grounding. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 4281–4290.

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 213–229.

[5] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: UNiversal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 104–120.

[6] Zhenfang Chen, Jiayuan Mao, Jiajun Wu, Kwan-Yee K. Wong, Joshua B. Tenenbaum, and Chuang Gan. 2021. Grounding physical concepts of objects and events through dynamic visual reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[7] Zhenfang Chen, Peng Wang, Lin Ma, Kwan-Yee Kenneth Wong, and Qi Wu. 2020. Cops-Ref: A new dataset and task on compositional referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2020), 10083–10092.

[8] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1724–1734.

[9] Volkan Cirik, Louis-Philippe Morency, and Taylor Berg-Kirkpatrick. 2018. Visual referring expression recognition: What do systems actually learn?. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 781–787.

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 4171–4186.

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[12] Weili Guan, Fangkai Jiao, Xuemeng Song, Haokun Wen, Chung-Hsing Yeh, and Xiaojun Chang. 2022. Personalized fashion compatibility modeling via metapath-guided heterogeneous graph learning. In *Proceedings of the International Conference on Research on Development in Information Retrieval (ACM SIGIR)*. 482-âĂŞ491.

[13] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2980–2988.

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[15] Mengge He, Wenjing Du, Zhiquan Wen, Qing Du, Yutong Xie, and Qi Wu. 2023. Multi-granularity aggregation transformer for joint video-audio-text representation learning. *IEEE Transactions on Circuits and Systems for Video Technology (TCSVT)* 33, 6 (2023), 2990–3002.

[16] Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 55–71.

[17] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 10294–10303.

[18] Ronghang Hu, Marcus Rohrbach, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2017. Modeling relationships in referential expressions with compositional modular networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1115–1124.

[19] Ronghang Hu, Amanpreet Singh, Trevor Darrell, and Marcus Rohrbach. 2020. Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 9989–9999.

[20] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 11021–11028.

[21] Haoshuo Huang, Vihan Jain, Harsh Mehta, Alexander Ku, Gabriel Magalhaes, Jason Baldridge, and Eugene Ie. 2019. Transferable representation learning in vision-and-language navigation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 7404–7413.

[22] Drew A. Hudson and Christopher D. Manning. 2018. Compositional attention networks for machine reasoning. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[23] Chenchen Jing, Yunde Jia, Yuwei Wu, Chuanhao Li, and Qi Wu. 2022. Learning the dynamics of visual relational reasoning via reinforced path routing. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 2. 7.

[24] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Judy Hoffman, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. Inferring and executing programs for visual reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 2989–2998.

[25] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. 2014. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 787–798.

[26] Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha S. Srinivasa. 2019. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6741–6749.

[27] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[28] Yujia Li, Daniel Tarlow, Marc Brockschmidt, and Richard S. Zemel. 2016. Gated graph sequence neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[29] Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L. Yuille. 2019. CLEVR-Ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4185–4194.

[30] Yibing Liu, Yangyang Guo, Jianhua Yin, Xuemeng Song, Weifeng Liu, Liqiang Nie, and Min Zhang. 2022. Answer questions with right image regions: A visual attention regularization approach. *ACM Trans. Multimedia Comput. Commun. Appl. (ACM TOMMCCAP)* 18, 4 (2022).

[31] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 10434–10443.

[32] Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[33] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. 2016. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 11–20.

[34] Khanh Nguyen, Debadeepta Dey, Chris Brockett, and Bill Dolan. 2019. Vision-based navigation with language-based assistance via imitation learning with indirect intervention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 12527–12537.

[35] Yonghua Pan, Zechao Li, Liyan Zhang, and Jinhui Tang. 2022. Causal inference with knowledge distilling and curriculum learning for unbiased VQA. *ACM Trans. Multimedia Comput. Commun. Appl. (ACM TOMMCCAP)* 18, 3 (2022).

[36] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*. 8024–8035.

[37] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1532–1543.

[38] Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, and Qi Wu. 2021. The road to know-where: An object-and-room informed sequential BERT for indoor vision-language navigation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. 1655–1664.

[39] Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2020. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia (TMM)* (2020).

[40] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res. (JMLR)* 21 (2020), 140:1–140:67.

[41] Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 817–834.

[42] Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing (TSP)* 45, 11 (1997), 2673–2681.

[43] Jiaxin Shi, Hanwang Zhang, and Juanzi Li. 2019. Explainable and explicit visual reasoning over scene graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 8376–8384.

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*. 5998–6008.

[45] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph attention networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

[46] Jia Wang, Jingcheng Ke, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. 2022. Referring expression comprehension via enhanced cross-modal graph attention networks. *ACM Trans. Multimedia Comput. Commun. Appl. (ACM TOMMCCAP)* (2022).

[47] Peng Wang, Qi Wu, Jiewei Cao, Chunhua Shen, Lianli Gao, and Anton van den Hengel. 2019. Neighbourhood watch: Referring expression comprehension via language-guided graph attention networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 1960–1968.

[48] Xin Wang, Qiuyuan Huang, Asli Çelikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. 2019. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 6629–6638.

[49] Zhiquan Wen, Shuaicheng Niu, Ge Li, Qingyao Wu, Mingkui Tan, and Qi Wu. 2023. Test-time model adaptation for visual question answering with debiased self-supervisions. *IEEE Transactions on Multimedia (TMM)* (2023).

[50] Zhiquan Wen, Yaowei Wang, Mingkui Tan, Qingyao Wu, and Qi Wu. 2023. Digging out discrimination information from generated samples for robust visual question answering. In *Findings of the Association for Computational Linguistics: ACL.*

[51] Zhiquan Wen, Guanghui Xu, Mingkui Tan, Qingyao Wu, and Qi Wu. 2021. Debiased visual question answering from feature and sample perspectives. In *Advances in Neural Information Processing Systems (NeurIPS).*

[52] Guanghui Xu, Shuaicheng Niu, Mingkui Tan, Yucheng Luo, Qing Du, and Qi Wu. 2021. Towards accurate text-based image captioning with content diversity exploration. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 12637–12646.

[53] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the International Conference on Machine Learning (ICML).* 2048–2057.

[54] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Cross-modal relationship inference for grounding referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 4145–4154.

[55] Sibei Yang, Guanbin Li, and Yizhou Yu. 2019. Dynamic graph attention for referring expression comprehension. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 4644–4653.

[56] Sibei Yang, Guanbin Li, and Yizhou Yu. 2020. Graph-structured referring expression reasoning in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 9949–9958.

[57] Dongfei Yu, Jianlong Fu, Xinmei Tian, and Tao Mei. 2019. Multi-source multi-level attention networks for visual question answering. *ACM Trans. Multimedia Comput. Commun. Appl. (ACM TOMMCCAP)* 15, 2s (2019).

[58] Licheng Yu, Zhe Lin, Xiaohui Shen, Jimei Yang, Xin Lu, Mohit Bansal, and Tamara L. Berg. 2018. MAttNet: Modular attention network for referring expression comprehension. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 1307–1315.

[59] Licheng Yu, Hao Tan, Mohit Bansal, and Tamara L. Berg. 2017. A joint speaker-listener-reinforcer model for referring expressions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR).* 3521–3529.

[60] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph convolutional networks for temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV).* 7094–7103.

[61] Yihan Zheng, Zhiquan Wen, Mingkui Tan, Runhao Zeng, Qi Chen, Yaowei Wang, and Qi Wu. 2020. Modular graph attention network for complex visual relational reasoning. In *Proceedings of the Asian Conference on Computer Vision (ACCV).* 137–153.

[62] Muhua Zhu, Yue Zhang, Wenliang Chen, Min Zhang, and Jingbo Zhu. 2013. Fast and accurate shift-reduce constituent parsing. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL).* 434–443.

Received 20 January 2023; revised 16 June 2023; accepted 17 June 2023