Spatial-Semantic Collaborative Graph Network for Textbook Question Answering

Yaxian Wang, Bifan Wei, Jun Liu, Qika Lin, Lingling Zhang, and Yaqiang Wu

***Abstract*— Textbook Question Answering (TQA) task requires answering questions by reasoning based on both the given diagrams and text context. There are mainly two challenges for the task. First, the diagrams are different from the natural images. Similar shapes or color blocks may express different semantics and there is also a large intra-topic variation for diagrams. Hence, the characteristics of visual semantic ambiguity and variable visual appearance make the diagram understanding more challenging. Second, for the text, the specific education domain with terminologies exists a great gap with the general domain. Therefore, it is difficult to represent the text semantics effectively using a text encoder pretrained in the general domain. In this paper, we propose a Spatial-Semantic Collaborative Graph Network (SSCGN) for TQA task, which can help enhance the diagram and text understanding and facilitate multimodal reasoning. Specifically, the Spatial-guided Semantic Enhancing (SSE) module fully exploits the spatial and semantic relationships between visual objects and OCR tokens to collaboratively enhance the diagram semantic understanding. Moreover, based on the semantically enhanced region representations of the SSE module, the Fine-grained Spatial-Aware Graph Network (FSA-GN) can help obtain richer relation-aware region representations for joint reasoning by capturing more fine-grained spatial relationships. We further propose multiple self-supervised auxiliary tasks to enhance the initial diagram and text semantic representations by pretraining the diagram encoder and text encoder. Extensive experiments and ablation studies are conducted to validate the effectiveness of SSCGN.**

***Index Terms*— Textbook question answering, multi-modal machine comprehension, diagram understanding.**

Manuscript received 27 June 2022; revised 20 September 2022; accepted 17 December 2022. Date of publication 21 December 2022; date of current version 3 July 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2020AAA0108800; in part by the National Natural Science Foundation of China under Grant 62137002, Grant 61937001, Grant 62192781, Grant 62176209, Grant 62176207, Grant 62106190, and Grant 62250009; in part by the Innovative Research Group of the National Natural Science Foundation of China under Grant 61721002; in part by the Innovation Research Team of Ministry of Education under Grant IRT_17R86; in part by the Consulting Research Project of Chinese Academy of Engineering The Online and Offline Mixed Educational Service System for The Belt and Road Training in MOOC China; in part by the LENOVO-XJTU Intelligent Industry Joint Laboratory Project; in part by the CCF-Lenovo Blue Ocean Research Fund; in part by the Project of China Knowledge Centre for Engineering Science and Technology; in part by the Foundation of Key National Defense Science and Technology Laboratory under Grant 6142101210201; in part by the Natural Science Basic Research Program of Shaanxi under Grant 2023-JC-YB-293; and in part by the Youth Innovation Team of Shaanxi Universities. This article was recommended by Associate Editor B. Sheng. *(Corresponding author: Jun Liu.)*

Yaxian Wang, Bifan Wei, and Jun Liu are with the Shaanxi Provincial Key Laboratory of Big Data Knowledge Engineering and the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: wyx1566@stu.xjtu.edu.cn; weibifan@xjtu.edu.cn; liukeen@xjtu.edu.cn).

Qika Lin and Lingling Zhang are with the National Engineering Laboratory for Big Data Analytics and the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: qikalin@foxmail.com; zhanglling@xjtu.edu.cn).

Yaqiang Wu is with the Lenovo Research, Beijing 100094, China, and also with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, Shaanxi 710049, China (e-mail: wuyqe@lenovo.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3231463.

Digital Object Identifier 10.1109/TCSVT.2022.3231463

I. Introduction

MULTI-MODAL Machine Reading Comprehension (M^3C) [1] is a new challenging and interesting task that has received growing attention from both the natural language processing (NLP) and the computer vision (CV) communities. Different from some mainstream multi-modal tasks, such as visual question answering [2], [3], image captioning [4], [5] and image-text matching [6], [7], M^3C aims to answer questions according to a multimodal context including the textual context and images. Textbook Question Answering (TQA) [8] is a concrete M^3C task in the field of education. The multiple modalities including text context and diagrams for TQA task are all from a specific education domain such as middle school science. Specifically, a diagram question in TQA task consists of a diagram, text context, a question about the diagram, and several candidate answers.

There are mainly two challenges for TQA task. First, the visual modality in textbooks is mostly in the form of diagrams, which demonstrate the complicated phenomena involving life sciences, earth sciences, physical sciences, etc. Existing natural image understanding methods [2], [3], [4], [5], [6], [7] cannot be directly applied to understand this type of diagram due to the following two reasons: **(1) Visual semantic ambiguity.** Diagrams are usually an abstract expression of knowledge. They are mostly composed of simple shapes and color blocks, which omit or suppress information such as background, textures and shadings [9]. Similar shapes or color blocks may express different semantics in different topics or even within the same topic. For example, as shown in Fig. 1, in different topics, a circle can represent multiple objects including the *earth*, *moon* or an *electron*, and the yellow blocks can represent objects such as the *sun* or Na^+ in Fig. 1(b) and (c). In the same *earth moon phases* topic, the same shape can represent the *earth* or the *first quarter phase*. **(2) Significant intra-topic variation.** Diagrams from the same topic may have large visual differences. In Fig. 1, we display three topics of diagrams including *earth moon*

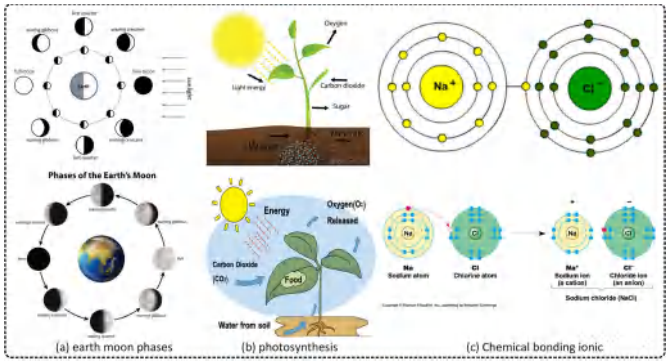

Fig. 1. Examples of diagrams from three different topics in the TQA task demonstrate that (1) similar shapes or color blocks may express different semantics, such as circles and yellow blocks; (2) diagrams in the same column have significant intra-topic variation.

phases, *photosynthesis* and *chemical bonding ionic*. It is obvious that the diagrams of the same topic (in the same column) have a significant intra-topic variation in the expression form. Specifically, the same objects in natural images such as *dogs* always have similar characteristics to each other, although they have different appearances. Thus, their visual representations can express explicit semantics. Conversely, the same objects in diagrams such as *earth* may have significantly different visual appearances in an abstract form, leading to a larger gap between their visual representations and semantics, which conveys more irregular and elusive knowledge than natural images. In summary, due to visual semantic ambiguity and variable characteristics of visual objects in diagrams, it is difficult to convey explicit semantics with only their visual appearance. However, most existing diagram understanding methods only consider the visual appearance information of objects, which may have certain limitations.

Second, the textual modality in the textbooks is about a specific education domain, which exists a great gap with the general domain. Therefore, it is insufficient to represent the knowledge using a text encoder pretrained by a large dataset from the general domain, such as English Wikipedia. For example, the terminologies *waxing crescent* and *waning gibbous* in Fig. 1(a) are rare and obscure in the general domain. The TQA task requires understanding the text semantics, including questions, answers, the text context and even the OCR tokens in the diagram. It is necessary but challenging to obtain effective text semantic representations.

To address the above challenges, we propose a new model, Spatial-Semantic Collaborative Graph Network (SSCGN) for the TQA task. Regarding the problem of diagram understanding, we find that visual objects are semantically related to spatially adjacent OCR tokens, which can transmit richer semantic information to humans when combined with visual appearance features. Inspired by that, we design a Spatial-guided Semantic Enhancing (SSE) module to enhance diagram semantic understanding, which exploits the spatial relationships between visual objects and OCR tokens to generate a spatial relation graph. The spatial relation graph acts as the surrogate ground truth to supervise the generation of the semantic relation graph. In the training process, we design a relation graph consistency loss to constrain the feature propagation between visual objects and OCR tokens, which enables fusion and exchange of multi-modal information. Moreover, we fully exploit spatial relationships to progressively enrich the region representations for joint reasoning. Specifically, based on the semantically enhanced region representations of the SSE module, the Fine-grained Spatial-Aware Graph Network (FSA-GN) can help obtain richer relation-aware region representations by capturing more fine-grained spatial relationships.

For the great gap between the education domain and the general domain, we also propose multiple self-supervised auxiliary tasks including two diagram auxiliary tasks to pretrain the diagram encoder and a text auxiliary task to pretrain the text encoder. Importantly, our proposed new self-supervised text auxiliary task, QA-Text Infomax maximizes the mutual information between the question-answer and summarized text context to force the model to incorporate the text context into the decision process, by which the pretrained language model helps bridge the semantic gap between education and general domains and obtain more robust text representations for the education domain.

Our main contributions can be summarized as follows:

- We propose a novel model for the TQA task, in which the spatial and semantic relationships between visual objects and OCR tokens are fully exploited to enhance the understanding of diagram semantics collaboratively. Moreover, the more fine-grained spatial relationships enable us to progressively enrich the visual representations for joint reasoning.
- We propose three specialized self-supervised auxiliary tasks, in which the first two for diagrams and the latter for text, which aim to enhance the initial semantic representations of diagrams and text effectively in the education domain, and lay an essential foundation for subsequent question answering.
- Extensive experiments show that SSCGN achieves state-of-the-art performance on the TQA task compared to baselines. Furthermore, ablation studies are conducted to verify the effectiveness of each module in SSCGN.

The rest of the paper is organized as follows. Section II introduces the related work about diagram understanding, self-supervised auxiliary task and textbook question answering. We present the details of our model in Section III. In Section IV, we analyze the results for experiments and ablation studies on TQA task. Finally, we conclude our work and make a plan for future work in Section V.

II. Related Work

Our work is closely related to diagram understanding, self-supervised auxiliary task and textbook question answering.

A. Diagram Understanding

Diagram understanding [10], [11] has received great interest in recent years. The work can be divided into the following three categories according to the types of objects in diagrams: scientific-style objects based, geometric objects based and natural-style objects based.

Scientific-style objects based work is about line plots, dot-line plots, vertical and horizontal bar graphs, or pie charts. Kafle et al. [12] presented a DVQA dataset to solve the problem of bar chart understanding. The work proposes a multi-output model for DVQA to generate generic answers using classification sub-network and chart-specific answers using OCR sub-network, and proposes a stacked attention network with a dynamic encoding model to handle chart-specific words. Kahou et al. [13] presented FigureQA, which focuses on scientific-style figures from five classes including line plots, dot-line plots, vertical and horizontal bar graphs, and pie charts. The work [13] takes all pairs of object presentations concatenated with the question as input to perform relation reasoning with a relation network [14].

Geometric objects based work is about visual elements such as points, lines, circles. Seo et al. [11] presented a dataset of high school plane geometry questions where every question has a textual description in English accompanied by a diagram. The work proposes G-ALIGNER for diagram understanding by discovering visual elements such as lines, circles, polygons and aligning them with their corresponding textual mentions. However, it does not identify geometric relations and does not solve geometry problems. Therefore, Seo et al. [15] presented GeoS, the first automated system to solve unaltered SAT-level geometry questions, which parses the question text and the diagram into a logical representation. However, it is only examined in a small dataset with 185 problems. Chen et al. [16] proposed a geometric question answering dataset GeoQA with 5,010 geometric problems and a neural geometric solver to address geometric problems by parsing multimodal information and generating interpretable programs.

Natural-style objects based work involves the education domain such as geography, biology, and physics. Kembhavi et al. [9] presented AI2 Diagrams (AI2D) of over 5,000 grade school science diagrams and proposed Diagram Parse Graphs (DPGs) to model the structure of diagrams, in which nodes correspond to the constituents including blobs, text boxes, arrows and arrowheads, and edges denote the relationships between the constituents. Kim et al. [17] proposed a unified diagram parsing network (UDPnet) to solve object detection with the SSD model [18] and relation matching with an RNN-based dynamic graph generation network (DGGN). Kembhavi et al. [8] translated DPGs into factual sentences through several translation rules to describe the graph.

In general, the diagram style in the TQA dataset is closest to that of AI2D but more complex and rich, hence the existing methods for AI2D are not sufficient to solve the TQA task. In addition, the TQA dataset contains extra text context, which requires joint reasoning over diagram and text context.

B. Self-Supervised Auxiliary Tasks

Self-supervised pretraining [19], [20], [21] has attracted great attention, which is widely used to avoid the cost of annotating large-scale datasets and improve model performance by designing auxiliary tasks. Auxiliary tasks act as an important strategy to learn representations of the data using pseudo labels [22]. The work can be divided into the following two categories according to the modalities: visual self-supervised auxiliary tasks and text self-supervised auxiliary tasks.

1) Visual Self-Supervised Auxiliary Tasks: There has been some work to enhance visual features including patch-based spatial composition prediction tasks [23], [24], [25], [26] and rotation prediction tasks [27]. The jigsaw puzzle [23], [24] was designed to recover the positions of shuffled segments of an image. Relative position prediction [25] was proposed to predict the relative position of two patches randomly sampled from a single image. Selfie [26] is a method used to determine the correct patch to be filled in the masked location. Rotation [27] was designed to infer the rotation angle's degree of an image. All these typical methods are for spatial relation contrast.

2) Text Self-Supervised Auxiliary Tasks: For the self-supervised methods in NLP, the core idea is to improve the text representation by pretraining the text encoder. The transformer is pretrained on a large-scale text corpus via various self-supervised tasks, such as predicting the random masked tokens in Bert [28], predicting whether two augmented sentences originate from the same sentence in CERT [29] and reconstructing the corrupted text in BART [30]. Wang et al. [31] proposed three self-supervised pretraining methods, including mask, replace and switch to capture the document-level context and improve the extractive summarization task. Vu et al. [32] formulated a probe task to measure the extent to which the sentence content property is captured in a paragraph embedding, which can help increase the downstream classification accuracy.

Inspired by these works, we design three self-supervised auxiliary tasks for diagram and text context.

C. Textbook Question Answering

In recent years, there have been some works on TQA task, which can be roughly divided into the following three categories: graph-based [8], [33], [34], [35], [36], [37], pretraining-based [38], [39], [40], and interpretability-based [41].

1) Graph-Based Methods: This type of method has been proven to be powerful for describing and modeling complex information, which has been widely used in various tasks [42], [43], [44]. Many existing methods also utilize graph-based methods to solve textbook question answering task. MemN+DPG [8] adopts memory networks to encode text and exploits the diagram parse graph based on DSDP-NET [9] to represent the diagram. EAMB [33] proposes an essay-anchor attentive multi-modal bilinear pooling method to address the long essay and multi-modal issues of the TQA task, which first uses GraphSAGE [45] to encode the essay information by learning the essay-anchor embeddings and then uses the bilinear pooling method MFB for multi-modal fusion. MoQA [34] proposes selecting the set of top-k supporting sentences/graph nodes of text and diagram for a certain question and comparing the performance of different knowledge representations. IGMN [35] utilizes semantic rules and spatial analysis rules to comprehend large contexts and diagrams respectively via a contradiction entity-relationship

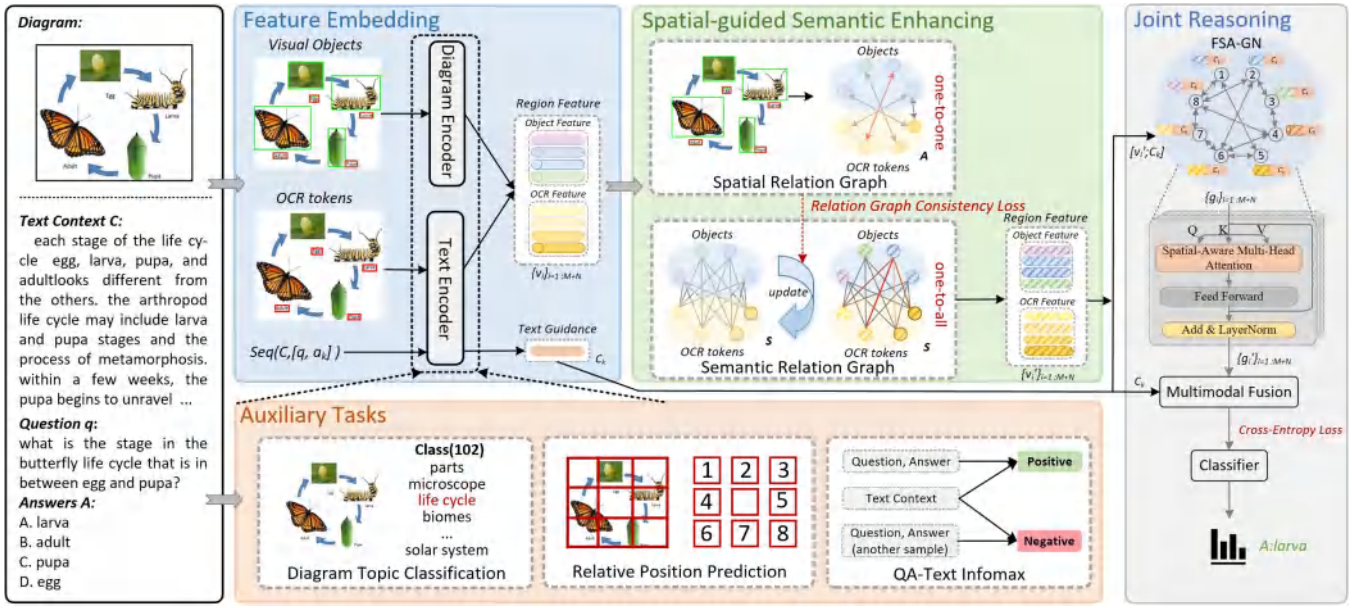

Fig. 2. Overall framework of Spatial-Semantic Collaborative Graph Network (SSCGN) which consists of four modules: a feature embedding module, a spatial-guided semantic enhancing module, a joint reasoning module and an auxiliary tasks module.

graph. F-GCN1+SSOC [36] suggests a fusion GCN [46] to extract knowledge about textual context and diagrams from the multi-modal context graph as background knowledge related to the question. In addition, the work introduces a novel self-supervised learning process into TQA training to tackle out-of-domain issues. RAFR [37] proposes a relation extraction method to construct a relation graph based on semantic dependency relations and relative positions between nodes to understand diagrams, which only considers the text information and ignores the visual information in the diagram.

2) Pretraining-Based Methods: Gomez-Perez and Ortega [38] proposed pretraining the models in a multistage process, finetuning them on TQA dataset and finally performing ensemble learning. The work presents an intelligent system for automatically answering textbook questions (ISAAQ) to overcome critical challenges such as the complexity and relatively small size of the TQA dataset or the scarcity of large diagram datasets. MHTQA [39] proposes a multi-head TQA architecture to address all types of questions and a contextualized iterative dual fusion network to learn a contextualized diagram representation. WSTQ [40] applies text matching and relation detection to deepen the text understanding and learn the diagram semantics respectively.

3) Interpretability-Based Methods: XTQA [41] generates span-level evidence based on a coarse-to-fine algorithm while answering the question, which can provide textual explanations for question answering to achieve interpretability to some extent.

The existing TQA models do not take full advantage of the OCR tokens in the diagram and the interaction of the visual objects and OCR tokens for diagram understanding. We explore and solve TQA task from this perspective.

III. Proposed Method

In this section, we start with the problem definition of TQA task. Given a diagram d, text context c and a question q about the diagram d, our task aims to select an answer $\hat{a}$ amongst the candidate answer set A, which can be defined as follows:

$$\hat{a} = \arg\max_{a \in A} p(a|d, c, q; \theta_d), \tag{1}$$

where θ_d is the trainable parameter, A is the candidate answer set and $\hat{a} \in A$ best matches the ground-truth answer. This type of questions about diagrams is called diagram multiple choice (DMC). In addition, there is also a subset of text questions without the above-mentioned diagram d, which includes text true/false (T/F) and text multiple choice (TMC) questions.

A. Overview

In this work, we propose a novel Spatial-Semantic Collaborative Graph Network (SSCGN) as shown in Fig. 2, which is mainly composed of the following four modules:

1) A feature embedding module extracts a set of visual regions including visual objects and OCR tokens, and then obtains their representations.

2) A spatial-guided semantic enhancing (SSE) module exploits the spatial relationships between visual objects and OCR tokens to enhance the region semantic representations and diagram understanding.

3) A joint reasoning module exploits twelve fine-grained spatial relationships to model the interaction between the regions under the guidance of questions and the text context, and generates a joint embedding to predict the final answer.

4) An auxiliary tasks (Aux) module helps to pretrain the diagram and text encoders to provide more robust features for the downstream TQA task.

For convenience, the explanations for some important notations are summarized in Table I.

B. Feature Embedding

Since both the visual objects and OCR tokens are significant constituents in the diagram, we first obtain a set of visual

TABLE I
IMPORTANT NOTATIONS AND EXPLANATIONS

Notations	Explanations
v_m^{obj}, v_n^{ocr}	the feature of $m-th$ visual object, $n-th$ OCR token.
$\{v_i\}_{i=1:M+N}$	the region feature set for M visual objects and N OCR tokens.
v_i	the initial feature of $i-th$ region, which can be object feature v_m^{obj} or OCR feature v_n^{ocr}.
v_i'	the enhanced feature of $i-th$ region, obtained by SSE module.
A	the spatial relation matrix.
S	the semantic relation matrix.
C_k	text guidance, encoded by RoBERTa with text context C, question q and one of the answers a_k as input.
g_i	the $i-th$ initial node feature for FSA-GN, obtained by concatenating v_i' and C_k.
g_i'	the relation-aware feature of $i-th$ region, obtained by FSA-GN.

regions including visual objects and OCR tokens by employing an object detector YOLOv3 [47] and an external Optical Character Recognition[1] (OCR) detector.

1) Embedding for Visual Objects: First, we adopt a pre-trained ResNet-101 [48] backbone to embed the visual objects detected by YOLOv3 to obtain M appearance features $\{v_m^a\}_{m=1:M}$. Each visual object is represented by the combination of appearance feature v_m^a and the positional feature (bounding box feature) v_m^p. We first project the above two types of features into a d_h-dimensional space. The final visual object feature is computed as follows:

$$v_m^{obj} = f(LN(W^a v_m^a), LN(W^p v_m^p)), \tag{2}$$

where $v_m^p = [x_m^{min}, y_m^{min}, x_m^{max}, y_m^{max}]$ is the top-left and bottom-right coordinates of the $m-th$ bounding box, $W^a \in \mathsf{R}^{d_h \times d_a}$ and $W^p \in \mathsf{R}^{d_h \times d_p}$ are learnable linear projection matrices, and $f(\cdot)$ is a feature fusion scheme, which can be an equal weighted sum, adaptive weighted sum or concat+MLP (the details can be found in Section IV-D). A layer normalization (LN) operation is applied to ensure that the feature representations are the same scale.

2) Embedding for OCR Tokens: By adopting an OCR detector, we can obtain the OCR tokens and their bounding boxes in the diagram. Since OCR tokens contain not only visual appearance information but also textual semantic information, we adopt the same embedding method as visual objects to obtain N appearance features for the OCR tokens $\{v_n^a\}_{n=M+1:M+N}$. For the text feature, we use a pretrained language model RoBERTa [49] to embed the text. Then we combine the appearance feature v_n^a, positional feature v_n^p and text feature v_n^t, which is computed as follows:

$$v_n^{ocr} = f(LN(W^a v_n^a), LN(W^p v_n^p), LN(W^t v_n^t)), \tag{3}$$

where W^a and W^p are the same as the parameter matrices in Eq. 2, $W^t \in \mathsf{R}^{d_h \times d_t}$ is also a linear projection matrix, $v_n^{ocr} \in \mathsf{R}^{d_h}$ and $f(\cdot)$ is the same as the feature fusion scheme for visual objects. Given the representations of visual objects and OCR tokens, we can denote the region set as $\{v_i\}_{i=1:M+N}$, where v_i represents the feature embedding of a visual object v_m^{obj} or an OCR token v_n^{ocr}.

[1] https://ocr.space/

C. Spatial-Guided Semantic Enhancing

The visual objects are semantically related to the spatially-adjacent OCR tokens, which can transmit richer semantic information to humans combined with the visual appearance feature. Inspired by that, we consider enhancing the representation of the visual objects with the help of surrounding OCR tokens. There are mainly two steps. The first step is to generate a spatial relation graph G^{sp} for the visual objects and OCR tokens. We can regard the generated spatial relation graph as a type of prior knowledge to act as a surrogate ground truth. The second step is to generate a semantic relation graph G^{se} under the supervision of the surrogate ground truth to constrain the consistency between G^{se} and G^{sp} by a relation graph consistency loss. We yield supervision for relation graph generation process. The additional supervision helps constrain the feature propagation between visual objects and OCR tokens, enabling fusion and exchange of multi-modal information to optimize the construction of the semantic relation graph and further enhance the region semantic representations.

1) Generation of Surrogate Spatial Relation Graph for Diagram: We first construct a spatial relation graph $G^{sp} = \{N, E\}$ according to the spatial relationships between the visual objects and OCR tokens, where N and E denote the node and edge set. The node set N includes two types of nodes, the visual objects $N_v = \{v_i^{obj}\}_{i=1:M}$ and the OCR tokens $N_t = \{v_i^{ocr}\}_{i=1:N}$. The edge set E includes the edges between the visual object and the OCR token $E_{obj-ocr}$. In other words, we only reserve the relationships between the heterogeneous nodes. In addition, the presence or absence of edges depends on the distance between the nodes. Specifically, the distance between two nodes is computed by:

$$\begin{aligned} d_I &= \sqrt{W^2 + H^2}, \\ d_{bb} &= \sqrt{(x_i^c - x_j^c)^2 + (y_i^c - y_j^c)^2}/d_I, \end{aligned} \tag{4}$$

where W and H are the width and height of a diagram, (x_i^c, y_i^c) and (x_j^c, y_j^c) are center point coordinates for the $i-th$ and

$j-th$ bounding boxes. The distance d_{bb} is normalized to the diagonal length of the diagram d_I.

The normalized distance is used to generate the spatial relation graph based on a threshold T as follows:

$$A_{ij} = \begin{cases} 1, & d_{bb} \leq T \\ 0, & d_{bb} > T. \end{cases} \tag{5}$$

When $A_{ij} = 1$, there is an edge between the node i and node j. The distances between all pairs of nodes are computed. In this way, we can obtain a sparse spatial relation matrix $A = \{A_{ij}\}_{i=1:M}^{j=1:N}$ as the surrogate ground-truth. The spatial relation using a hard mask is one-to-one, in which each visual object has an edge only with the spatially nearest OCR token.

2) Generation of Semantic Relation Graph for Diagram: The semantic graph is formalized as $G^{se} = \{N, E_s\}$. Edges in G^{se} represent the similarity between visual objects and OCR-tokens. Instead of the hard binary masks [0, 1] in the spatial relation graph G^{sp}, we adopt soft masks to build the semantic graph. Hence, we first project the embeddings of each visual object v_i^{obj} and OCR token v_j^{ocr} into the same embedding space. Then we compute the similarity between the node representations as follows:

$$S_{ij} = sigmoid((W^{obj} v_i^{obj})^\top \cdot (W^{ocr} v_j^{ocr})), \tag{6}$$

where $W^{obj}, W^{ocr} \in \mathsf{R}^{d_o \times d_h}$ are transformation matrices. After operating on all heterogeneous graph nodes, we can obtain a relation matrix $S = \{S_{ij}\}_{i=1:M}^{j=1:N}$ as the semantic relation graph. The semantic relation using a soft mask is one-to-all. The higher the similarity score is, the more related the visual object and OCR token become.

3) Relation Graph Consistency Loss: We hope to encourage the semantic relation graph to be consistent with the spatial relation graph, so that the model achieves feature propagation between the visual objects and OCR tokens in close proximity and obtains the enhanced region representations $\{v_i'\}_{i=1:M+N}$. The relation graph consistency loss is designed as follows:

$$L_{sup} = -\sum_{i=1}^{M}\sum_{j=1}^{N}(A_{ij} \cdot logS_{ij} + (1 - A_{ij}) \cdot log(1 - S_{ij})). \tag{7}$$

D. Joint Reasoning

In this module, first based on the semantically enhanced region representations $\{v_i'\}_{i=1:M+N}$ of the SSE module, the Fine-grained Spatial-Aware Graph Network (FSA-GN) is designed to help obtain richer relation-aware region representations and then the multimodal fusion is performed to predict a final answer.

1) FSA-GN: Unlike the only considered spatial relative distance in the SSE module, inspired by the work [50], [51], we consider more aspects, such as the relative distance, relative angle and Intersection over Union (IoU) to further exploit fine-grained spatial relationships between the regions. Specifically, we exploit twelve relatively complete spatial relationships (C1: inside, C2: covering, C3: overlap with IoU above 0.5, C4-C11: relative angle relation and an overlap with IoU below 0.5, C12: self) to establish the connection between the regions. Relationships are middle-level representations that connect local regions and the global understanding of images. To take full advantage of the fine-grained spatial relationships, we proposed a text-guided spatial-aware multimodal graph network to learn a more holistic view of the visual scene in the diagram, which mines the interactions between the regions in the diagram under the guidance of the question and text context. In the graph network, we encode the different types of spatial information into the edges by modifying the attention mechanism, named Spatial-Aware Multi-Head Attention (SA-MHA).

First, we use RoBERTa [49], a transformer language model to encode the input $s_k = seq(C, [q, a_k])$, where C is the text context, q is the question and a_k is one of its possible answers. In this way, we obtain a representation C_k of s_k as our text guidance. Then by concatenating the text guidance C_k with each region representation v_i', we introduce the text guidance into the relation learning process:

$$g_i = [v_i'; C_k], \tag{8}$$

where [;] denotes the concatenation operation.

By computing the twelve relative spatial relationships, we build a fine-grained spatial relation graph G^f with the visual objects and OCR tokens as nodes, in which the edges are aware of the direction and spatial relation types. The visual objects and the OCR tokens are initialized as $\{g_i\}_{i=1:M+N}$. Then we adopt a graph attention network with SA-MHA to obtain relation-aware node representations by capturing the spatial relationships. The update method is given by:

$$\begin{aligned} g_i' &= \sigma(\sum_{j \in N(i)} \alpha_{ij}(W_{dir(ij)}^V g_j' + b_{type(ij)})), \\ \alpha_{ij} &= \underset{j \in N(i)}{\text{softmax}}((W^Q g_i')^\top \cdot (W_{dir(ij)}^K g_j') + b'_{type(ij)}), \end{aligned} \tag{9}$$

where W^Q, W^K, and W^V denote the transformation matrices, and the subscript $dir(ij)$ indicates the direction of edge ij, which is used to select the transformation matrix related to the edge direction. Specifically, there are three edge directions including "from i to j","from j to i" and "from i to i". The subscript $type(ij)$ represents the label of edge ij, which is is used to select the bias terms $b_{type(ij)}$ and $b'_{type(ij)}$ related to twelve edge types, corresponding to twelve fine-grained spatial relationships. And $N(i)$ represents the node set in the neighborhood of nodes i. Additionally, α_{ij} acts as a gate to measure the relation between the nodes and control the contribution of node j to the feature representation of node i. Specifically, the stronger the relationship, the more information of node j is transmitted to the receiver node i. In this way, we can obtain the relation-aware region representation g_i', which implies spatial and semantic relations that are useful for question answering.

2) Answer Prediction: We first perform multimodal fusion following [3]. Specifically, for each choice a_k, we concatenate the region representation g_i' with the text representation C_k. The attention weight β_{ki} on each region and the whole

representation of the diagram $\hat{g}_k$ are calculated by:

$$\beta_{ki} = \text{softmax}(w_a \cdot tanh([g'_i; C_k])),$$
$$\hat{g}_k = \sum_{i=1}^{M+N} \beta_{ki} g'_i, \tag{10}$$

where w_a is a learned vector.

Finally, we compute the final probability of the candidate answer a_k by jointly embedding the text representation and diagram representation:

$$\hat{y}_k = \text{softmax}(w_c(C_k \cdot \hat{g}_k)), \tag{11}$$

where w_c is a learnable parameter vector.

Algorithm 1 Spatial-Semantic Collaborative Graph Network (SSCGN)

Input: Embedding for regions $\{v_i\}_{i=1:M+N}$ including embedding for visual objects $\{v_m^a\}_{m=1:M}$ and embedding for OCR tokens $\{v_n^a\}_{n=1:N}$.
Output: SSCGN model E_θ.

for number of training iterations **do**
  for each batch in training set **do**
    $A \leftarrow$ generate a spatial relation graph by Eq.(4)(5).
    $S \leftarrow$ generate a semantic relation graph by Eq.(6).
    $\{v'_i\}_{i=1:M+N} \leftarrow$ update S with the supervision of A via supervision loss L_{sup} by Eq.(7).
    $\{g'_i\}_{i=1:M+N} \leftarrow$ learn more robust region representations by Eq.(8)(9).
    $\hat{g}_k \leftarrow$ generate attended region representations by Eq.(10).
    Calculate question answering loss L_{qa} by Eq.(12).
    $L \leftarrow$ a weighted sum of L_{qa} and L_{sup} by Eq.(13).
    Update model parameters θ through back propagation: $\theta \leftarrow \theta - \eta \frac{\partial L}{\partial \theta}$.
  end for
end for
return E_θ.

We regard the TQA task as a multi-class classification problem. Therefore, SSCGN is optimized by minimizing the standard cross-entropy criterion L_{qa}, which can be denoted as follows:

$$L_{qa} = -\sum_{k=1}^{K} a_k \cdot \log \hat{y}_k, \tag{12}$$

where $a_k = [0, 1](k \in [1, \ldots, K])$ and the candidate answer set includes a correct answer.

The loss for the SSE module between the surrogate spatial relation graph and the semantic relation graph predicted by SSCGN and the loss for question answering are balanced by a multiplicative factor γ as follows:

$$L = L_{qa} + \gamma * L_{sup}. \tag{13}$$

The learning process of the proposed Spatial-Semantic Collaborative Graph Network (SSCGN) is summarized in Algorithm 1.

In fact, the modules corresponding to the two parts of loss are complementary to each other: the spatial-guided semantic enhancing module learns robust representations good for question answering, while the question answering task is used for testing the effect of the region representations and reasoning ability of the whole model.

E. Auxiliary Tasks

It is well known that self-supervised pretraining is an effective way to deal with data scarcity and enhance the model representation ability. Here, we design three self-supervised auxiliary tasks including two self-supervised diagram auxiliary tasks and a self-supervised text auxiliary task to pretrain our diagram encoder and text encoder offline. Note that the module is independent of the other three modules, which is a pretraining process.

1) Self-Supervised Diagram Auxiliary Task: Although the spatial-guided semantic enhancing module and fine-grained spatial-aware graph network can contribute a lot, a powerful diagram encoder is still necessary to improve the diagram understanding and question answering accuracy. To obtain diagram features of higher quality, we design two auxiliary tasks to pretrain the diagram encoder, including Diagram Topic Classification and Relative Position Prediction.

a) Diagram topic classification (DTC): First, we regard the prefix name of each diagram as a topic, and each topic contains multiple diagrams. After de-duplication, we obtain the diagram topics, including 102 classes. Then we further collect diagrams for each topic from the Google website. Finally, approximately 20,000 diagrams of higher quality were obtained by manual screening and filtering. Compared with the 3,455 diagrams in TQA dataset, we greatly expand the dataset with only diagrams, which do not have any other annotations. We regard it as a multi-class classification task using the image topic automatically generated by data as the label and use cross-entropy loss as the loss function to train the diagram encoder.

b) Relative position prediction (RPP): In this task, we first crop a diagram to 9 blocks and select the center block as the target. Then we train the diagram encoder with a random block and the target block as input to predict their relative position using a cross-entropy loss. Additionally, there are only eight possible relative positions for a pair of samples, meaning the output is a simple eight-way softmax classification.

2) Self-Supervised Text Auxiliary Task: We hope to bridge the gap between the education domain and general domain and learn text representations that are more suitable for the education domain, so we propose a new self-supervised text auxiliary task, **QA-Text Infomax (QTI)** to incorporate as much text context information as possible into the question and answers. The model focuses on the relation between the question-answer and text context and forces the question-answer representation to carry information shared with the text context. The maximization of mutual information (MI) [52] needs to extract positive samples and negative samples. Specifically, a positive sample is obtained by pairing the question and answer with the corresponding text context, which is represented as (f_{qa}, f_t). A negative sample is

TABLE II
DATASET OVERVIEW

Question Type	Candidate	Total Number	Dataset Split		
			Train	Val	Test
T/F	2	5,400	3,490	998	912
TMC	4-7	8,293	5,163	1,530	1,600
DMC	4	12,567	6,501	2,781	3,285
All	-	26,260	15,154	5,309	5,797

obtained by pairing the question and answer with the random sampling text context in a batch represented as $(f_{qa}, \overline{f_t})$ or pairing the text context with random sampling question and answer in a batch represented as $(\overline{f_{qa}}, f_t)$. Here, we maximize the mutual information between the question-answer f_{qa} and summarized text context f_t to force the model to incorporate the text context into the decision process, by which the pretrained language model can help obtain more robust text representations. The objective for the f_{qa}, f_t, $\overline{f_{qa}}$ and $\overline{f_t}$ is given by:

$$L_{aux} = -\frac{1}{N}\sum_{i=1}^{N}(log(g(f_{qa}, f_t)) + log(1 - g(f_{qa}, \overline{f_t})) + log(1 - g(\overline{f_{qa}}, f_t))), \quad (14)$$

where N is the batch size and $g(\cdot)$ denotes the bilinear operation function to score the question-answer and text context.

IV. EXPERIMENTS

In this section, we first briefly introduce the TQA dataset in Section IV-A. Then, we demonstrate the implementation details and experimental setup in Section IV-B. Subsequently, we analyze the experimental results for SSCGN in Section IV-C, evaluate the effectiveness of SSCGN based on several ablation studies and conduct a simple hyperparameter analysis in Section IV-D. Finally, in Section IV-E we make some case studies to demonstrate the ability of SSCGN.

A. TQA Dataset

We evaluate our proposed SSCGN on the TQA dataset [8], which consists of 1,076 lessons containing 78,338 sentences and 3,455 diagrams from life science, earth science and physical science textbooks of middle school science curricula. Table II demonstrates the question types, the corresponding number of candidate answers and the dataset split.

B. Implementation Details

In this work, our proposed SSCGN is based on the Pytorch platform. The diagram encoder is implemented by the pretrained ResNet-101 backbone. The object detector YOLOv3 was pretrained on the AI2D datasets [9]. The text encoder is implemented by the RoBERTa large [49], which is pretrained on a collection of scientific MC question datasets following [38]. The long text context provided by TQA contains only limited content relevant to the question. Therefore, we can adopt three methods, namely Information Retrieval (IR), Next Sentence Prediction (NSP) and Nearest Neighbors (NN), as the text background retrieval methods to retrieve the text that is most relevant to questions and answers, and an ensemble method for them following the work [38]. For a fair comparison with baselines, we adopt IR as the final text background retriever in most experiments. We set the maximum length of the input text sequence to 180 and the OCR token sequence to 30. All the inputs are truncated or padded to the same length. In addition, we set the maximum number of visual regions as 32 and the maximum number of OCR tokens to 20. The feature fusion scheme $f(\cdot)$ adopts an equal weighted sum (weight=1). The threshold T for spatial relationship generation is set to 0.3. For the joint reasoning module, we empirically set the head number for attention to 16 and the number of layers to 2. The multiplicative factor γ for the loss function is set to 0.3. The initial learning rate is set to $1e^{-5}$ for T/F, $2e^{-6}$ for TMC and $3e^{-6}$ for DMC. For optimization, Adam [53] with a linearly decayed learning rate and warm-up is used. To avoid overfitting, we choose a dropout value of 0.1.

C. Experimental Results and Analysis

We compare our proposed SSCGN with several existing methods including six graph-based methods, three pretraining-based methods and one interpretability-based method, which were introduced in Section II-C. Note that most of the methods only report the results on the validation set. Therefore, we compare SSCGN with ten of them on the validation set and four of them on the test set. For a fair comparison, we also adopt IR as the text background retriever to extract text that is most relevant to questions and answers from the long text context, the same as ISAAQ_{IR} [38] and WSTQ_{IR} [40]. Table III and Table IV show the experimental results on validation set and test set of TQA dataset respectively. From the two tables, we make the following three observations:

- For all questions (All), SSCGN achieves the best performance on both the validation and test set. Specifically, SSCGN attains 0.90% higher accuracy than the SOTA baseline MHTQA on the validation set and 2.65% higher accuracy than ISAAQ_{IR} on the test set. It indicates that SSCGN is effective in solving all types of questions.
- For all text questions (Text All), SSCGN obtains 1.49% and 2.27% accuracy improvement over MHTQA and ISAAQ_{IR} on the validation and test set respectively. Among them, for T/F questions, SSCGN has a lower performance than the previous best model MHTQA on the validation set, but is able to outperform it on all other question types. We analyze that the multi-type question learning strategy in MHTQA may be more suitable for T/F questions, however our method takes the performance of both types of text questions into account and achieves better performance on all text questions. In addition, for T/F questions, SSCGN outperforms ISAAQ_{IR} by 1.42% on the test set. For TMC questions, SSCGN improves the best baseline MHTQA by 3.36% on the validation set and outperforms ISAAQ_{IR} by 2.74% on the test

TABLE III

EXPERIMENTAL RESULTS (% ACCURACY) OF DIFFERENT TYPES OF QUESTIONS ON VALIDATION SET FOR TQA TASK. WE PRESENT THE ACCURACY OF TEXT TRUE/FALSE QUESTIONS (T/F), TEXT MULTIPLE CHOICE QUESTIONS (TMC), ALL TEXT QUESTIONS (TEXT ALL = T/F ∪ TMC), AND DIAGRAM MULTIPLE CHOICE QUESTIONS (DMC) AND ALL QUESTIONS (ALL = TEXT ALL ∪ DMC)

Model	T/F	TMC	Text All	DMC	All
MemN+DPG (2017) [8]	50.50	30.98	38.69	32.83	35.62
EAMB (2018) [33]	55.31	33.27	41.97	35.10	38.37
MoQA (2018) [34]	61.90	36.20	46.40	33.40	39.56
IGMN (2018) [35]	57.41	40.00	46.88	36.35	41.36
f-GCN1+SSOC (2019) [36]	62.73	49.54	54.75	37.61	45.77
XTQA (2020) [41]	58.24	30.33	41.32	32.05	36.46
ISAAQ_{IR} (2020) [38]	78.26	67.52	71.76	53.83	62.37
RAFR (2021) [37]	53.63	36.67	43.35	32.85	37.85
MHTQA (2021) [39]	**82.87**	69.22	74.61	54.87	64.27
WSTQ_{IR} (2022) [40]	77.15	56.80	64.87	50.92	57.17
SSCGN	81.50	**72.58**	**76.10**	**55.24**	**65.17**

TABLE IV

EXPERIMENTAL RESULTS (% ACCURACY) OF DIFFERENT TYPES OF QUESTIONS ON TEST SET FOR TQA TASK

Model	T/F	TMC	Text All	DMC	All
XTQA (2020) [41]	56.22	33.40	41.67	33.34	36.95
ISAAQ_{IR} (2020) [38]	77.74	68.94	72.13	50.50	59.88
RAFR (2021) [37]	52.75	34.38	41.03	30.47	35.04
WSTQ_{IR} (2022) [40]	77.28	58.27	65.15	44.28	53.24
SSCGN	**79.16**	**71.68**	**74.40**	**53.46**	**62.53**

set. This indicates that SSCGN has great potential for robust text understanding to comprehensively improve the performance for all types of text questions.

- For DMC questions, SSCGN achieves the best performance whether on the validation or test set, significantly improving the accuracy on the test set by 2.96%, which indicates that SSCGN can effectively alleviate visual semantic ambiguity and enhance the diagram understanding in TQA task.

To sum up, we believe that our proposed SSCGN can work well on TQA task.

To further fully compare the performance of ISAAQ and SSCGN, we have conducted multiple experiments on the setting of w/o pretraining and w/ pretraining. The setting w/ pretraining refers to using the multi-stage pretrained RoBERTa on a collection of scientific MC question datasets. The setting w/ pretraining refers to using the multi-stage pretrained RoBERTa on a collection of scientific MC question datasets. The setting w/o pretraining refers to using RoBERTa without the above pretraining. Note that besides the above pretraining on a collection of scientific MC questions, ISAAQ [38] uses additional AI2D [9] pretraining for the whole model and SSCGN uses additional auxiliary tasks. Therefore, for equality, we eliminate all pretrainings in SSCGN and ISAAQ, and then compare them under our settings. Due to the different settings, we rerun the ISAAQ to obtain results for comparison.

Table V shows the individual results of each text background retrieval method including IR, NN and NSP and their ensemble results. Experimental results on DMC demonstrate that our proposed SSCGN significantly outperforms ISAAQ on the same setting whether on the validation set or test set. In general, we improve the overall DMC performance of SOTA by 3.96% and 3.85% on the validation set and test set respectively on the setting of w/o pretraining. We also obtain 1.20% and 1.11% accuracy improvement on the validation set and test set respectively on the setting of w/ pretraining. We find that SSCGN has a more significant effect on the setting of w/o pretraining. It is speculated that ISAAQ may rely more on the representation ability of the language model, and our SSCGN has a stronger diagram understanding ability.

D. Ablation Studies

To analyze the contribution of each designed module, we conduct ablation studies on the validation set for the TQA task, as shown in Table VI and Table VII. Here, we verify the effectiveness of each module by adding each module or module combination to the Baseline. The Baseline uses a pretrained Resnet-101 backbone to encode the regions, a multi-stage pretrained RoBERTa on a collection of scientific MC question datasets to encode the question and text context and then performs mutimodal fusion [3] for answer prediction.

1) The Effectiveness of the Main Modules: We report the ablation results for main modules, as shown in Table VI. The specific analyses are as follows:

- **w/ SSE.** This method aims to verify the effectiveness of introducing OCR tokens to enhance the region representations for diagram understanding. We see that Baseline w/ SSE increases by 1.69 % in row 1 of Table VI, which proves that OCR tokens are good complements to the semantic features of visual objects. Compared with only the visual appearance or only text semantics, the integration of the semantically high-level OCR token features and region features can help alleviate visual semantic ambiguity in diagrams to some extent and obtain richer and more comprehensive visual representations.
- **w/ SA-MHA.** This method aims to verify the effectiveness of introducing spatial location into the atten-

TABLE V

EXPERIMENTAL RESULTS (% ACCURACY) OF DIAGRAM QUESTIONS ON VALIDATION SET AND TEST SET FOR TQA TASK, WHICH ARE MAINLY COMPARISONS OF ISAAQ AND OUR MODEL SSCGN ON DIFFERENT SETTINGS. IR, NSP AND NN ARE THREE TEXT BACKGROUND RETRIEVERS, AND ALL IS THE ASSEMBLING OF THEM

Setting	Model	Dataset	DMC			
			IR	NSP	NN	All
w/o pre-training	ISAAQ	val	44.34	47.79	44.08	47.82
		test	38.93	41.67	41.28	43.20
	SSCGN	val	52.24	51.92	50.37	51.78
		test	45.72	44.98	43.32	47.05
w/ pre-training	ISAAQ	val	52.64	51.35	50.31	53.07
		test	51.02	49.86	48.46	51.72
	SSCGN	val	54.12	53.86	51.93	54.27
		test	52.18	51.32	50.68	52.83

TABLE VI

MODULES ABLATION RESULTS (% ACCURACY) OF DIAGRAM QUESTIONS ON VALIDATION SET FOR TQA TASK

#	SSE	SA-MHA	FSA-GN	Aux	DMC	Δ
Baseline					52.37	-
1	✓				54.06	+1.69
2		✓			53.13	+0.76
3			✓		53.38	+1.01
4				✓	53.72	+1.35
5		✓	✓	✓	53.89	+1.52
6	✓		✓	✓	54.28	+1.91
7	✓	✓		✓	54.36	+1.99
8	✓	✓	✓		54.42	+2.05
SSCGN	✓	✓	✓	✓	**55.24**	**+2.87**

TABLE VII

AUXILIARY TASKS ABLATION RESULTS (% ACCURACY) OF DIAGRAM QUESTIONS ON VALIDATION SET FOR TQA TASK. ROW 7 OF THE GRAY COLOR HAS THE SAME SETTING AS ROW 4 IN TABLE VI

#	DTC	RPP	QTI	DMC	Δ
Baseline				52.37	-
1	✓			53.12	+0.75
2		✓		53.05	+0.68
3			✓	53.38	+1.01
4	✓	✓		53.49	+1.12
5	✓		✓	53.63	+1.26
6		✓	✓	53.54	+1.17
7	✓	✓	✓	53.72	+1.35

tion mechanism for diagram understanding and joint reasoning. In row 2, Baseline w/ SA-MHA obtains a 0.76% performance improvement, which proves that introducing spatial relationships into multi-head attention mechanism can help enrich the region representation and enhance their spatial expressive ability to help answer the spatial type questions involving reasoning about relative positions.

- **w/ FSA-GN.** This method aims to verify the effectiveness of our relation-aware region representation method for diagram questions. In row 3, Baseline w/ FSA-GN increases by 1.01%, which indicates that message passing in the form of the spatial-aware graph network can help enrich visual representations and is effective for joint reasoning to obtain better performance.
- **w/ Aux.** This method aims to demonstrate their effectiveness in improving the representation ability. In row 4, Baseline w/ Aux increases by 1.35%, which demonstrates that our diagram and text based pretraining methods can help obtain more robust visual and text representations for the education domain.

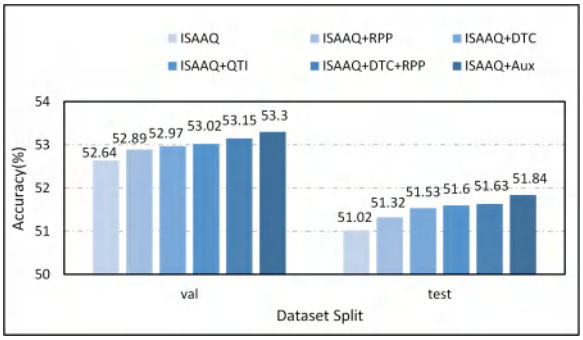

Fig. 3. $ISAAQ_{IR}$ Evolution vesions for DMC which add different auxiliary task or their combinations.

In rows 5-8, any three modules are combined based on Baseline, which is equivalent to removing one module from the full model SSCGN. We can observe that compared with Baseline, the module combinations can have higher performance improvement than the individual module, which indicates that different modules can complement each other to boost model performance. The combination of all modules forms SSCGN in the last row, which obtains the best performance. Consistently, compared with SSCGN, the performance of these ablation models in rows 5-8 all degrades. In summary, each module has a positive effect on model performance.

2) The Effectiveness of Auxiliary Tasks: To evaluate the contribution of each auxiliary task, we conduct some ablation studies by adding each of the auxiliary tasks or their combination to Baseline as shown in Table VII.

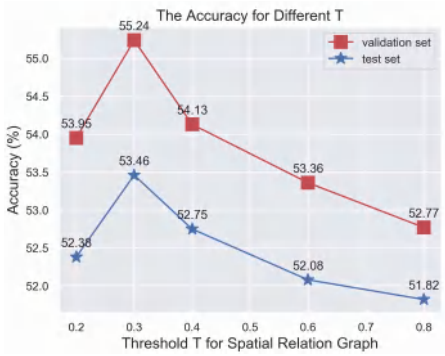

Fig. 4. Accuracy of DMC using spatial relation graphs constructed by using different thresholds T.

- **w/ DTC.** This method aims to verify the effectiveness of DTC for the visual representation. In row 1 of Table VII, Baseline w/ DTC obtains a 0.75% increase, which proves that the DTC task can enhance visual representation ability by improving the diagram classification accuracy. Intuitively, the higher the classification accuracy is, the better the diagram representation ability will be. In this case, the pretrained diagram encoder can contribute to the diagram representation, which is important to the question answering task.
- **w/ RPP.** This method aims to verify the effectiveness of RPP for the visual representation. In row 2, Baseline w/ DTC increases by 0.68%, which indicates that the feature representation learned through the within-diagram context can capture visual characteristics. The method can help enhance the visual representation ability by exploring to use spatial context as a source of the free and plentiful supervisory signal and lay a good foundation for our downstream task.
- **w/ QTI.** This method aims to verify the effectiveness of QTI for obtaining useful text features to help text understanding. In row 3, Baseline w/ QTI increases by 1.01%, which proves that the pretrained language model on our task can help obtain more robust text representations. QTI can help bridge the gap between the education domain and general domain and learn text representations that are more suitable for the education domain.

In rows 4-7, different combinations of the three auxiliary tasks on Baseline can achieve a performance improvement and outperform the performance when the auxiliary tasks are adopted alone. All three auxiliary tasks are combined together, which achieves the best gain.

3) The Scalability of Auxiliary Tasks: The auxiliary task module is scalable, which can also be used in ISAAQ. To evaluate the scalability of each auxiliary task, we conduct some studies by adding different auxiliary tasks or their combination on ISAAQ_{IR} with IR as the text background retriever. The results of ISAAQ_{IR} gradually evolution versions for diagram multiple choice questions are shown in Fig. 3. For ISAAQ_{IR}, the accuracy gains of DTC, RPP, QTI, DTC+RPP and the combination of all three tasks are 0.33%, 0.25%, 0.38%, 0.51%, and 0.66% respectively on the validation set, and 0.51%, 0.30%, 0.58%, 0.61%, and 0.82% respectively on the test set.

TABLE VIII
DMC PERFORMANCE OF DIFFERENT FEATURE FUSION SCHEMES FOR INITIAL REGION FEATURES

Method	DMC
Equal Weighted Sum (weight=1)	**55.24**
Adaptive Weighted Sum	54.88
Concat+MLP	54.63

4) Different Thresholds for Spatial Relation Graph Construction: We conduct a set of experiments using the IR as the text background retriever to analyze the effect of threshold T for the construction of the spatial relation graph and question answering. Specifically, we keep all other settings the same and only modify the threshold T within $\{0.2, 0.3, 0.4, 0.6, 0.8\}$. As shown in Fig. 4, it is obvious that $T = 0.3$ is the best choice for SSCGN, which can bring the best performance and best generalization.

5) Comparison of Different Feature Fusion Schemes: We choose different fusion schemes to control the fusion of multiple features, forming initial region features including visual object features and OCR token features. Here, we compare three different fusion schemes as follows:

- Equal Weighted Sum: This method considers that each type of feature contributes equally to the initial region feature, then carries a weighted sum on them to form the initial region feature.
- Adaptive Weighted Sum: This method learns different weights for the different types of features and carries a weighted sum on them to form the initial region feature.
- Concat+MLP: This method combines the multiple features by a concatenation operation, and then uses an MLP to encode the concatenated features.

We evaluate their performances on the validation set of the TQA dataset, and the performances are shown in Table VIII. We can observe that the equal weighted sum method achieves the best accuracy with a minor increase. Surprisingly, although using the adaptive weighted sum method to combine multiple features performs better than the concat+MLP method, it fails to achieve better performance compared to the equal weighted sum method. We speculated that due to the small amount of data, the adaptive weighted sum method is difficult to optimize for better results.

6) Comparison of Different Maximum Numbers for Visual Objects and OCR Tokens: We set the maximum number based on the percentage of diagrams that contain visual objects or OCR tokens less than the set threshold (the percentage = the number of diagrams containing visual objects or OCR tokens less than the set threshold / the number of all diagrams). Specifically, according to different percentages, we set different maximum numbers for visual objects and OCR tokens to analyze their impact on our model. As shown in Table IX, we can make the following three observations as follows:

- When the setting for the maximum number of visual objects and OCR tokens is small (e.g. row 1 and row 5),

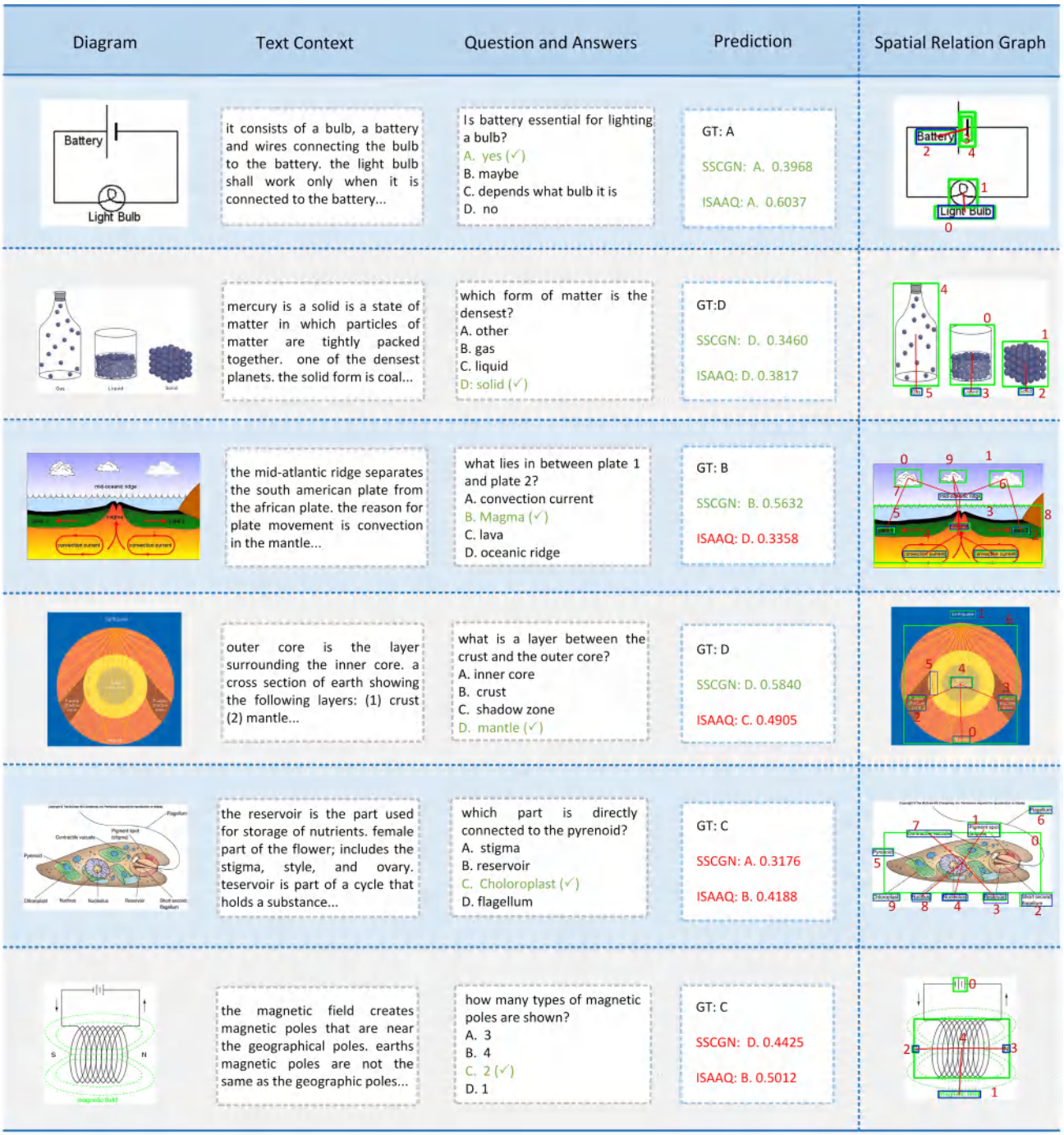

Fig. 5. Examples from the validation set for case studies. The text context is most relevant to the question and answers, which is obtained with IR as the text background retriever. The spatial relation graphs are constructed by SSCGN. The success and failure cases of our proposed SSCGN, the baseline ISAAQ and the ground truth (GT) are compared.

many diagrams need to truncate the input visual objects and OCR tokens, leading to necessary information loss. Thus, the model cannot learn effective diagram representation, which further results in degraded performance of question answering.

- When the setting for a maximum number of visual objects and OCR tokens is large enough as shown in row 4 and row 8, although all visual objects and OCR tokens for any diagram are taken into account, the model performance drops slightly. We analyze that for these diagrams, too many visual objects or OCR tokens are detected, some of which may have low confidence, leading to introducing noisy information in the question answering process.
- When $M = 32$ and $N = 20$ in row 3 and row 7, this setting can achieve the best performance.

To sum up, the OCR tokens and visual objects are the important elements, it is necessary to set an appropriate maximum number for them.

7) Comparison of Different Detectors for Visual Objects and OCR Tokens: For object detection model of visual objects, besides YOLOv3 [47], we select the classic Faster R-CNN [54] and recent mainstream SAM-DETR [55] as

TABLE IX
DMC PERFORMANCE WITH DIFFERENT MAXIMUM NUMBERS *M* FOR VISUAL OBJECTS OR DIFFERENT MAXIMUM NUMBERS *N* FOR OCR TOKENS

#		Percentage	M	DMC
1	N=20	60%	12	54.01
2		80%	16	54.53
3		98%	32	**55.24**
4		100%	123	55.02
#		Percentage	N	DMC
5	M=32	60%	8	53.48
6		80%	12	54.13
7		98%	20	**55.24**
8		100%	61	54.89

TABLE X
DMC PERFORMANCE WITH DIFFERENT DETECTORS FOR VISUAL OBJECTS AND OCR TOKENS

#	Object Detectors	OCR Detectors	DMC
1	YOLOv3	EasyOCR	54.67
2		PaddleOCR	55.06
3		OCRSpace	**55.24**
#	OCR Detectors	Object Detectors	DMC
4	OCRSpace	Faster RCNN	54.74
5		SAM-DETR	54.93
6		YOLOv3	**55.24**

two variants. For OCR detectors of OCR tokens, besides OCRSpace[1], we select two popular detectors PaddleOCR[2] and EasyOCR[3] as two variants. As shown in Table X, we observe that these results just fluctuate slightly. We analyze that current object and OCR token detection tasks have been widely explored and achieved satisfactory performance, which can provide good foundations for our downstream TQA task by detecting relatively accurate visual objects and OCR tokens. Thus, the task performance is still dominated by the TQA model design. Moreover, compared with existing models, the better performance demonstrates better representation and reasoning ability of our model.

E. Case Studies

To better analyze our model, we select four success examples and two failure examples as shown in Fig. 5. We present the spatial relation graph generated by SSCGN besides the text context, diagrams, questions and answers. For the spatial relation graphs, the green boxes indicate the visual objects detected by YOLOv3, the blue boxes indicate the OCR tokens detected by the OCR detector, and the red lines represent the relation generated by our spatial relation algorithm. In the top four rows are success examples, we can conclude that SSCGN can mostly generate an ideal spatial relation graph as the surrogate ground truth to supervise the generation of the semantic relation graph. For example, the OCR tokens “Gas”, “Liquid” and “Solid” can connect to the most related visual objects, and further enhance their representations. In addition, we find that the extraction of accurate relevant text context is also important for question answering. For the third and fourth rows, the questions are related to the relative position such as “between”, our fine-grained spatial relation graph can help with joint reasoning. For the first wrong example, failure to match the meaning “is directly connected to” with the right spatial relationship may be the cause of this incorrect answer. The reason is that SSCGN cannot understand this implicit positional relationship. In addition, the relevant textual context is also not extracted accurately. For the second wrong example, failure to count is the main cause of the incorrect answer. In fact, numerical reasoning and counting are key points we need to improve further. Compared with ISAAQ, it is obvious that SSCGN can perform better on the questions related to spatial relationships and questions requiring a robust diagram understanding.

[2]https://github.com/PaddlePaddle/PaddleOCR
[3]https://github.com/JaidedAI/EasyOCR

V. CONCLUSION

In this paper, we propose a novel model Spatial-Semantic Collaborative Graph Network (SSCGN), which explores the spatial and semantic relationships between visual objects and OCR tokens in a diagram for textbook question answering. To enhance the representation of the visual objects via surrounding OCR tokens, we devise a spatial-guided semantic enhancing module to establish a semantic relation graph under the supervision of the generated surrogate spatial relation graph. We also design a fine-grained spatial-aware graph network to learn richer relation-aware region representations. Furthermore, we design three self-supervised auxiliary tasks to enhance the diagram and text semantic representations. Experimental results on TQA certify the effectiveness of each proposed module, and SSCGN achieves the SOTA performance compared with all other baselines.

In the future, we will explore the following directions: (1) We are going to realize the interpretability of textbook question answering tasks. (2) We are going to further study the self-supervised diagram understanding methods.

REFERENCES

[1] C. Zeng, S. Li, Q. Li, J. Hu, and J. Hu, “A survey on machine reading comprehension—Tasks, evaluation metrics and benchmark datasets,” *Appl. Sci.*, vol. 10, no. 21, p. 7640, Oct. 2020.
[2] S. Antol et al., “VQA: Visual question answering,” in *Proc. IEEE ICCV*, Jun. 2015, pp. 2425–2433.
[3] P. Anderson et al., “Bottom-up and top-down attention for image captioning and visual question answering,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6077–6086.
[4] Y. Cui, G. Yang, A. Veit, X. Huang, and S. Belongie, “Learning to evaluate image captioning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5804–5812.
[5] J. Yu, J. Li, Z. Yu, and Q. Huang, “Multimodal transformer with multi-view visual representation for image captioning,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4467–4480, Dec. 2020.
[6] J. Wu, C. Wu, J. Lu, L. Wang, and X. Cui, “Region reinforcement network with topic constraint for image-text matching,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 388–397, Jan. 2022.

[7] K. Wen, X. Gu, and Q. Cheng, "Learning dual semantic relations with graph attention for image-text matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2866–2879, Jul. 2021.
[8] A. Kembhavi, M. Seo, D. Schwenk, J. Choi, A. Farhadi, and H. Hajishirzi, "Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension," in *Proc. CVPR*, Jul. 2017, pp. 4999–5007.
[9] A. Kembhavi, M. Salvato, E. Kolve, M. Seo, H. Hajishirzi, and A. Farhadi, "A diagram is worth a dozen images," in *Proc. Eur. Conf. Comput. Vis.* Amsterdam, The Netherlands: Springer, Oct. 2016, pp. 235–251.
[10] X. Lin, S. Shimotsuji, M. Minoh, and T. Sakai, "Efficient diagram understanding with characteristic pattern detection," *Comput. Vis., Graph., Image Process.*, vol. 30, no. 1, pp. 84–106, Apr. 1985.
[11] M. J. Seo, H. Hajishirzi, A. Farhadi, and O. Etzioni, "Diagram understanding in geometry questions," in *Proc. 28th AAAI Conf. Artif. Intell. (AAAI)*, Quebec City, QC, Canada, Jul. 2014, pp. 2831–2838.
[12] K. Kafle, B. Price, S. Cohen, and C. Kanan, "DVQA: Understanding data visualizations via question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 5648–5656.
[13] S. E. Kahou, V. Michalski, A. Atkinson, A. Kádár, A. Trischler, and Y. Bengio, "FigureQA: An annotated figure dataset for visual reasoning," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr. 2018.
[14] A. Santoro et al., "A simple neural network module for relational reasoning," in *Proc. Adv. Neural Inf. Process. Syst. Annu. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, Dec. 2017, pp. 4967–4976.
[15] M. Seo, H. Hajishirzi, A. Farhadi, O. Etzioni, and C. Malcolm, "Solving geometry problems: Combining text and diagram interpretation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1466–1476.
[16] J. Chen et al., "GeoQA: A geometric question answering benchmark towards multimodal numerical reasoning," in *Proc. Findings Assoc. Comput. Linguistics (ACL/IJCNLP)*, Aug. 2021, pp. 513–523.
[17] D. Kim, Y. Yoo, J. Kim, S. Lee, and N. Kwak, "Dynamic graph generation network: Generating relational knowledge from diagrams," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4167–4175.
[18] W. Liu et al., "SSD: Single shot multibox detector," in *Proc. Eur. Conf. Comput. Vis.*, Aug. 2016, pp. 21–37.
[19] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang, "Adversarial robustness: From self-supervised pre-training to fine-tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 699–708.
[20] L. Tao, X. Wang, and T. Yamasaki, "An improved inter-intra contrastive learning framework on self-supervised video representation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5266–5280, Aug. 2022.
[21] J. Huang, Y. Huang, Q. Wang, W. Yang, and H. Meng, "Self-supervised representation learning for videos by segmenting via sampling rate order prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3475–3489, Jun. 2022.
[22] A. Jaiswal, A. R. Babu, M. Z. Zadeh, D. Banerjee, and F. Makedon, "A survey on contrastive self-supervised learning," *Technologies*, vol. 9, no. 1, p. 2, Dec. 2020.
[23] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *Proc. Eur. Conf. Comput. Vis.*, Oct. 2016, pp. 69–84.
[24] F. M. Carlucci, A. D'Innocente, S. Bucci, B. Caputo, and T. Tommasi, "Domain generalization by solving jigsaw puzzles," in *Proc. CVPR*, Jun. 2019, pp. 2229–2238.
[25] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1422–1430.
[26] T. H. Trinh, M.-T. Luong, and Q. V. Le, "Selfie: Self-supervised pretraining for image embedding," 2019, *arXiv:1906.02940*.
[27] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *Proc. 6th Int. Conf. Learn. Represent. (ICLR)*, Vancouver, BC, Canada, Apr. 2018.
[28] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. ACL Conf. NAACL HLT*, Minneapolis, MN, USA, Jun. 2019, pp. 4171–4186.
[29] H. Fang, S. Wang, M. Zhou, J. Ding, and P. Xie, "CERT: Contrastive self-supervised learning for language understanding," 2020, *arXiv:2005.12766*.
[30] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics, (ACL)*, D. Jurafsky, J. Chai, N. Schluter, and J. R. Tetreault, Eds., Jul. 2020, pp. 7871–7880.
[31] H. Wang et al., "Self-supervised learning for contextualized extractive summarization," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 2221–2227.
[32] T. Vu and M. Iyyer, "Encouraging paragraph embeddings to remember sentence identity improves classification," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 6331–6338.
[33] J. Li, H. Su, J. Zhu, and B. Zhang, "Essay-anchor attentive multi-modal bilinear pooling for textbook question answering," in *Proc. IEEE Int. Conf. Multimedia Expo. (ICME)*, Jul. 2018, pp. 1–6.
[34] M. Haurilet, Z. Al-Halah, and R. Stiefelhagen, "MoQA—A multi-modal question answering architecture," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Munich, Germany, Sep. 2018, pp. 1–8.
[35] J. Li, H. Su, J. Zhu, S. Wang, and B. Zhang, "Textbook question answering under instructor guidance with memory networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3655–3663.
[36] D. Kim, S. Kim, and N. Kwak, "Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3568–3584.
[37] J. Ma, J. Liu, Y. Wang, J. Li, and T. Liu, "Relation-aware fine-grained reasoning network for textbook question answering," *IEEE Trans. Neural Netw. Learn. Syst.*, early access, Jun. 28, 2021, doi: 10.1109/TNNLS.2021.3089140.
[38] J. M. Gomez-Perez and R. Ortega, "ISAAQ—Mastering textbook questions with pre-trained transformers and bottom-up and top-down attention," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, Nov. 2020, pp. 5469–5479.
[39] J. He, X. Fu, Z. Long, S. Wang, C. Liang, and H. Lin, "Textbook question answering with multi-type question learning and contextualized diagram representation," in *Proc. Int. Conf. Artif. Neural Netw.*, Bratislava, Slovakia, Sep. 2021, pp. 86–98.
[40] J. Ma, Q. Chai, J. Huang, J. Liu, Y. You, and Q. Zheng, "Weakly supervised learning for textbook question answering," *IEEE Trans. Image Process.*, vol. 31, pp. 7378–7388, 2022.
[41] J. Ma et al., "XTQA: Span-level explanations of the textbook question answering," 2020, *arXiv:2011.12662*.
[42] Y. Luo, R. Ji, T. Guan, J. Yu, P. Liu, and Y. Yang, "Every node counts: Self-ensembling graph convolutional networks for semi-supervised learning," *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107451.
[43] Y. Luo, P. Liu, L. Zheng, T. Guan, J. Yu, and Y. Yang, "Category-level adversarial adaptation for semantic segmentation using purified features," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 8, pp. 3940–3956, Aug. 2022.
[44] F. Shao et al., "Active learning for point cloud semantic segmentation via spatial-structural diversity reasoning," 2022, *arXiv:2202.12588*.
[45] W. Hamilton, Z. Ying, and J. Leskovec, "Inductive representation learning on large graphs," in *Proc. Adv. Neural. Inf. Process. Syst.*, vol. 30. Red Hook, NY, USA: Curran Associates, Dec. 2017, pp. 1024–1034.
[46] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*, Toulon, France, Apr. 2017.
[47] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," 2018, *arXiv:1804.02767*.
[48] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 770–778.
[49] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.
[50] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 684–699.
[51] J. Lee, H. Lee, H. Lee, and K. Jung, "Learning to select question-relevant relations for visual question answering," in *Proc. 3rd Workshop Multimodal Artif. Intell.*, 2021, pp. 87–96.
[52] I. Belghazi, S. Rajeswar, A. Baratin, R. D. Hjelm, and A. C. Courville, "MINE: Mutual information neural estimation," 2018, *arXiv:1801.04062*.
[53] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent. (ICLR)*, Diego, CA, USA, May 2015.

[54] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real time object detection with region proposal networks," in *Proc. Adv. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2015, pp. 91–99.
[55] G. Zhang, Z. Luo, Y. Yu, K. Cui, and S. Lu, "Accelerating DETR convergence via semantic-aligned matching," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 949–958.

Yaxian Wang received the B.E. degree from Hunan Normal University, Changsha, China, in 2019. She is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, China. Her research interests include visual question answering, cross-modal learning, and interpretable reasoning.

Bifan Wei received the Ph.D. degree in computer science and technology from Xi'an Jiaotong University in 2014. He is currently a Research Fellow in computer science and technology with Xi'an Jiaotong University. His research interests include web data mining, taxonomy learning, and distance education.

Jun Liu received the B.S. and Ph.D. degrees in computer science and technology from Xi'an Jiaotong University in 1995 and 2004, respectively. He is currently a Professor with the School of Computer Science and Technology, Xi'an Jiaotong University, in China. He has authored more than 70 research papers in various journals and conference proceedings. His current research focuses on data mining and text mining. He served as a Guest Editor for many technical journals, such as *Information Fusion*, IEEE SYSTEMS JOURNAL, and *Future Generation Computer Systems*. He also acted as the conference/workshop/track chair at numerous conferences.

Qika Lin received the B.E. and M.E. degrees from the Beijing Institute of Technology, Beijing, China, in 2016 and 2019, respectively. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Xi'an Jiaotong University, Shaanxi, China. His research interests include knowledge graph, representation learning, and logic reasoning.

Lingling Zhang received the Ph.D. degree in computing science from Xi'an Jiaotong University in 2020. She was a Visiting Student at the School of Computer Science, Carnegie Mellon University, working with Prof. A. Hauptmann. She is currently an Assistant Professor in computer science with Xi'an Jiaotong University. Her research interests include cross-media information mining, computer vision, zeroshot learning, and few-shot learning.

Yaqiang Wu is currently a Research Director with Lenovo and oversee multiple research areas including cloud computing, soft engineering, and AI empowered applications. His own research has focused on computer vision, multimedia, and machine learning. He and his team won several competitions in text detection and recognition at ICDAR and ICPR in recent years.