Exploring Sparse Spatial Relation in Graph Inference for Text-Based VQA

Sheng Zhou, Dan Guo, *Member, IEEE*, Jia Li, Xun Yang, and Meng Wang, *Fellow, IEEE*

***Abstract*— Text-based visual question answering (TextVQA) faces the significant challenge of avoiding redundant relational inference. To be specific, a large number of detected objects and optical character recognition (OCR) tokens result in rich visual relationships. Existing works take all visual relationships into account for answer prediction. However, there are three observations: (1) a single subject in the images can be easily detected as multiple objects with distinct bounding boxes (considered repetitive objects). The associations between these repetitive objects are superfluous for answer reasoning; (2) two spatially distant OCR tokens detected in the image frequently have weak semantic dependencies for answer reasoning; and (3) the co-existence of nearby objects and tokens may be indicative of important visual cues for predicting answers. Rather than utilizing all of them for answer prediction, we make an effort to identify the most important connections or eliminate redundant ones. We propose a sparse spatial graph network (SSGN) that introduces a spatially aware relation pruning technique to this task. As spatial factors for relation measurement, we employ spatial distance, geometric dimension, overlap area, and DIoU for spatially aware pruning. We consider three visual relationships for graph learning: object-object, OCR-OCR tokens, and object-OCR token relationships. SSGN is a progressive graph learning architecture that verifies the pivotal relations in the correlated object-token sparse graph, and then in the respective object-based sparse graph and token-based sparse graph. Experiment results on TextVQA and ST-VQA datasets demonstrate that SSGN achieves promising performances. And some visualization results further demonstrate the interpretability of our method.**

***Index Terms*— Visual question answering, text-based visual question answering, graph inference, spatial relation, relation learning.**

I. Introduction

SCENE text expresses rich information in human activities, such as numeric symbols [14], advertisement slogans [46], traffic signs [3], and price tags in shops [3]. Text-based visual question answering (TextVQA) becomes an emerging hot topic in the field of vision and language. The TextVQA models have a wide range of applications, such as visual impairment assistance, online education, online shopping [60], driving assistance [60], *etc*.

With the advancement of artificial intelligence technology, many multimodal analysis models have been rapidly developed with visual understanding capabilities [16], [17], [18], [53], such as the tasks of image captioning [26], visual grounding [38], and visual question answering (VQA) [19]. The most relevant research to our work is the VQA task. General VQA models [15] have powerful reasoning capabilities to answer object-based visual questions regarding people, scenes, motifs, and even plot comprehension. However, the TextVQA models are dedicated to discovering the scene texts and utilizing them to answer the text-based visual questions, where scene texts may involve small, fuzzy, and illegible text fonts. To make up this research gap, Singh et al. [46] have released a novel TextVQA task and a new TextVQA dataset referring to both object-related and text-related visual questions. Meanwhile, Bitten et al. [3] have released another new dataset, ST-VQA, which could only answer questions using the scene text in the image. By comparison, scene text is extremely critical in the TextVQA task, for example, all questions in the ST-VQA [3] dataset are related to scene text.

Many efforts have been made to solve this task. Just in time, optical character recognition (OCR) tasks [2], [12] have made significant progress in the field of computer vision. Under the research background, some researchers [3], [14], [22], [30], [39], [46] apply this technique to existing VQA models, enabling the models with the ability to read scene text accurately. For example, LoRRA [46] is the first backbone for TextVQA which extends the VQA model Pythia [27] with a new OCR attention branch; the model is enabled to select answer words from a predefined vocabulary set of objects and an online set of OCR tokens, where the vocabulary set and the OCR set are collected from the training set and each image itself respectively. Besides, based on the success of Transformer [47] and BERT [10], M4C [22] implements a multi-modal transformer with a multi-step response which serves as another backbone and is widely-used for existing methods. Based on above two backbones, some graph models have been introduced into TextVQA task because of its outstanding relation reasoning ability, *e.g*, MM-GNN [14], SA-M4C [30], and CRN [39]. For a better exploration

Manuscript received 18 July 2022; revised 20 May 2023; accepted 18 August 2023. Date of publication 5 September 2023; date of current version 11 September 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB4500600; in part by the National Natural Science Foundation of China under Grant 62020106007, Grant 62272144, Grant U20A20183, Grant 72188101, Grant 62272435, Grant U22A2094, and Grant 62202139; and in part by the Major Project of Anhui Province under Grant 202203a05020011. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tao Mei. *(Corresponding authors: Dan Guo; Xun Yang; Meng Wang.)*

Sheng Zhou, Dan Guo, Jia Li, and Meng Wang are with the School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230601, China (e-mail: hzgn97@gmail.com; guodan@hfut.edu.cn; jiali@hfut.edu.cn; eric.mengwang@gmail.com).

Xun Yang is with the School of Information Science and Technology, University of Science and Technology of China, Hefei 230026, China (e-mail: xyang21@ustc.edu.cn).

Digital Object Identifier 10.1109/TIP.2023.3310332

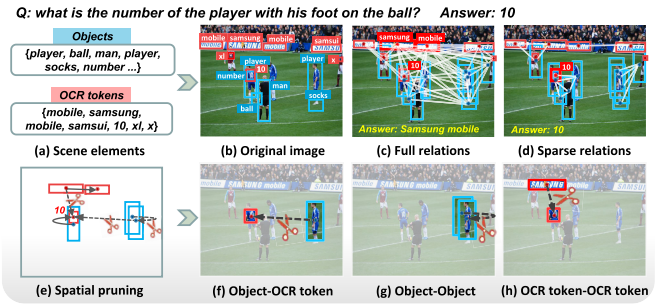

Fig. 1. Full (dense but redundant) relations vs. sparse relations in TextVQA task. The OCR system and the object detector have their own characteristics. The OCR system performs well at scene text recognition (*e.g*, numbers, symbols, characters), whereas the object detector excels at identifying visual subjects (*e.g*, people, animals, substances). We have to understand the scene texts or objects queried by the question both well. We take all the detected objects and OCR tokens in the image as visual entity nodes. Our purpose is to build an effective sparse spatial graph on the basis of spatial-aware relation pruning for answer prediction.

of visual relations, we adopt the graph structure in this work.

Almost the aforementioned works share a common feature in that the full relations among the visual entities including objects and OCR tokens are exploited to predict answers. Differently, we argue that excessive redundant relations exist among these visual entities, which may mislead the prediction of answers. As shown in Fig. 1 (c), driven by the question of "what is the number of the player with his foot on the ball?", under the full relations, "samsung mobile" is predicted as the answer due to the dense relations with the other visual entities, while the correct answer is "10".

This study is proposed to solve the relation redundancy problem for the TextVQA task. To be specific, we propose a sparse graph network inspired by three observations: (1) multiple bounding boxes covering a same visual entity are interpreted as different objects by the object detectors accompanied with redundant connections, *e.g*, the rightmost player is labeled with three bounding boxes in Fig. 1 (c); (2) distant OCR tokens often have weak or no semantic dependencies, *e.g*, the two OCR tokens "samsung" and "10" in Fig. 1 (c); (3) association between remote objects and tokens are useless, and the visual regions detected with both object and OCR token labels are informative for answer prediction, *e.g*, "10" is detected as "number" (object label) and number "10" (OCR token label) in Fig. 1 (c). In short, the idea originates from the native perspective of spatial relation. As shown in Fig. 1 (e), the spatial factors in our work refer to Distance-IoU [59], relative distance, geometric size, and overlap area. We work to use the spatial factors to eliminate the redundant relations between remote object and OCR token in Fig. 1 (f), much-overlapping objects in Fig. 1 (g), and distant OCR tokens in Fig. 1 (h).

To achieve effective relation reasoning, we propose a **S**parse **S**patial **G**raph **N**etwork **(SSGN)**, which cuts off useless or negative relations (edges in the graph) between numerous visual entities (object and OCR token nodes) to suppress the redundant message passing. As shown in Fig. 2, SSGN is a hierarchical graph learning architecture by excluding redundant relations (we deem these as noises) in three sparse sub-graphs. (1) **O**bject-**T**oken **S**parse **G**raph, **OTSG** first removes redundant relations to learn object-OCR token correlation verification, (2) **O**bject-based **S**parse **G**raph, **OSG**, and (3) **T**oken-based **S**parse **G**raph, **TSG** then verify the pivotal relations of **OTSG** in each visual entity space (object or OCR token). All the above three sub-graphs are implemented under the guidance of question. Based on this framework, SSGN progressively updates object and OCR token features by the graph learning of OTSG, OSG and TSG for answer prediction. Experiments conducted on TextVQA and ST-VQA datasets show that SSGN outperforms other comparative methods in the TextVQA task. Extensive ablation studies and visualization results demonstrate the validity, robustness, and interpretability of our method.

The main contributions are summarized as follows.

- *For task*, we perform a sparse spatial graph learning by introducing a spatial-aware relation pruning method into the TextVQA task, which shields redundant relations among numerous visual entities in complex scenes.
- *About sparsity*, we propose a graph preprocessing method which utilizes spatial coordinates of visual entities to build sparse relations (edges). In this work, we primarily emphasize the spatial-constrained relation pruning to trim the redundant relations. Besides, during graph learning, the pruned relations pass messages between visual entities under the guidance of question semantics.
- *About spatiality*, we consider Distance-IoU (DIoU), spatial distance, geometric size, and overlap area as the primary spatial factors in relation modeling. The motivation is that spatially distant visual entities in natural images often have less or even no relations, and geometrically similar boundary boxes with large overlaps may exist as superfluous visual entities.
- *About graph Inference*, we propose a hierarchical graph learning solution with partial edges to achieve node update. We first perform double-sided correlation verification (object-OCR token) in the OTSG, and then further verify the pivotal relations in respective OSG (object) and TSG (OCR token).

The rest of the paper is organized as follows. We provide an overview of related work in Sec. II and detail the proposed SSGN method in Sec. III. Extensive experiments including quantitative comparison with state-of-the-art methods, ablation studies, and visualization analysis are presented in Sec. IV, followed by a brief summary of this work in Sec. V.

II. Related Work

A. Text-Based Visual Question Answering

With the development of visual question answering tasks, some models [24], [29] have been proposed to improve the reasoning ability. Recently, TextVQA task [46] has become a new research focus aimed at answering questions about scene text comprehension. To promote this study, there are two benchmark datasets, TextVQA [46] and ST-VQA [3], and two backbones, LoRRA [46] and M4C [22]. LoRRA [46] uses an off-the-shelf OCR system [6] to detect multiple

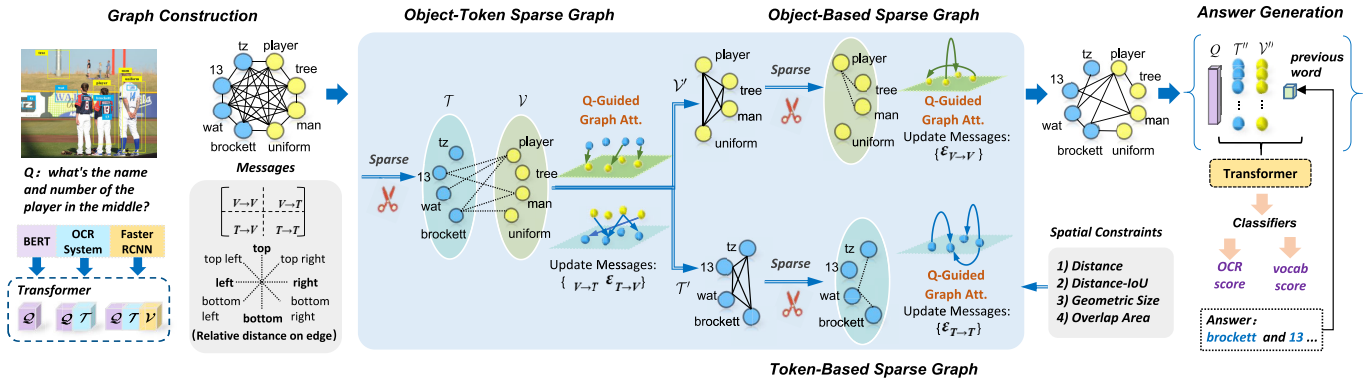

Fig. 2. The overall framework of Sparse Spatial Graph Network (SSGN). Given an image and a question, we prepare the features of the question ($\mathcal{Q}$), OCR tokens ($\mathcal{T}$), and objects ($\mathcal{V}$). After that, we construct graphs with $\mathcal{T}$ and $\mathcal{V}$ and build edges with the condition of spatial relation. The relations in the graph are defined as $\mathcal{E} = (\mathcal{E}_{\mathcal{V}\to\mathcal{V}}, \mathcal{E}_{\mathcal{T}\to\mathcal{T}}, \mathcal{E}_{\mathcal{V}\to\mathcal{T}}, \mathcal{E}_{\mathcal{T}\to\mathcal{V}})$. In this work, graph inference is used to update OCR token and object features under the guidance of the question $\mathcal{Q}$. We propose a novel progressive graph learning scheme with partial context update—the message ($\mathcal{E}_{\mathcal{V}\to\mathcal{T}}$, $\mathcal{E}_{\mathcal{T}\to\mathcal{V}}$) updates first and then $\mathcal{E}_{\mathcal{V}\to\mathcal{V}}$ and $\mathcal{E}_{\mathcal{T}\to\mathcal{T}}$ update in parallel. Finally, we use an answer generation module to iteratively predict answers.

OCR tokens in the image and extends a previous VQA model [27] to select a single OCR token as the answer. The methodological bottleneck of LoRRA is the inability to generate multiple words in an answer. To solve this problem, M4C [22] introduces a pointer-augmented multi-step decoder based on a transformer architecture to generate multiple words in the answer. Based on the above two backbones, existing methods could mainly be divided into four categories: (1) **Attention-based model.** SSBaseline [60] introduces an attention mechanism into M4C [22] to achieve feature aggregation, which greatly reduces the computation and obtains a good performance; (2) **Transformer-based model.** TAP [55] introduces large-scale data and pre-training techniques in a transformer architecture to improve performance. PAT [58] adopts a position-augmented transformer with entity-aligned mesh for comprehensively capturing position relations of visual entities; (3) **Graph-based model.** There are some graph networks focus on relation inference between objects and OCR tokens, including multi-modal graph (MM-GNN) [14], spatial-aware graph (SA-M4C) [30], role-aware graph (SMA) [13], and cascade reasoning network (CRN) [39]; (4) **Representation learning model.** BOV [57] proposes a visually enhanced scene text embedding and an object-oriented embedding to enhance the feature representation and improve the reasoning ability.

Previous works develop feature fusion [57], feature alignment [55], [58], feature attention [60], and relation learning [13], [14], [30], [39] to address the TextVQA task. Among them, graph model in TextVQA task shows outstanding advantage. Most graph models take full relations between or within objects and OCR tokens into account. Differently, we devote to removing redundant relations between object-object, OCR-OCR tokens, and object-OCR token to achieve effective answer reasoning.

B. Graph Inference Technique for Visual Reasoning

Plenty of works demonstrate that graph neural network (GNN) has a strong ability for relation reasoning [33], [37]. Both intra-modal and inter-modal graph models are applied for addressing the visual relation reasoning issue in various VQA and visual dialog tasks [7], [23], [61]. In general VQA task, Huang et al. [23] propose a novel dual-channel graph convolutional network (DC-GCN) to capture the visual relations between objects and the syntactic relations between question words separately. For Fact-based VQA [61], Zhu et al. first construct three intra-modal graphs to separately explore visual, semantic, and knowledge clues, and then aggregate them through cross-modal graph convolutions. As for visual dialog, Chen et al. design a graph-over-graph network named GoG [7] with three graphs for exploring co-reference relation among dialog history, dependency relation between question words, and spatial relation between visual objects. These relation-aware graphs are proposed to exploit consistent contexts from vision (image) and text (question) for visual reasoning. Similar to previous GNN works [7], [23], [61], we utilize the heterogeneous graph structures (intra- and inter-modality graphs) for relation inference. Besides, these previous GNN work [7], [23], [61] usually explores the fully-connected graph learning. The difference of this work is it extends a novel hierarchical graph learning scheme with sparse relations (edges) for achieving reasonable relation reasoning.

C. Exploitation of Spatial Relation in Vision

For visual reasoning, we insist that the exploitation of spatial relation can facilitate answer inference. We investigate the related work and find that the exploitation of visual-spatial relation is indeed beneficial for quite a few vision-related tasks [8], [25], [37], [50], [56]. For example, to fully understand the visual scene in the VQA task, ReGAT [37] utilizes a graph network to model the spatial relations between objects by measuring relative angles and overlapping areas of two objects in the image. For the TextCaps task [45], Wang et al. [50] strengthen the semantic correlation between two OCR tokens from both horizontal and vertical position dimensions to generate the captions, which is motivated by the spatial orientation relation of OCR token pairs. For object detection, Chen et al. [8] present a new object detector to locate the required partially-occluded object in the image,

which modifies the bounding box position by measuring the spatial distance of the occluded object and its neighbors. As for the visual relation detection task, Inayoshi et al. [25] propose a boundary-box channel-wise fusion method, which introduces object position and overlapping area into the image features for better identifying the relation between objects.

There is merely single object-object relation or OCR-OCR tokens relation discussed in the above tasks. We leverage the spatial relations between and within the two visual entities (*i.e*, object, and OCR token) for TextVQA. To explore the spatial relation for visual reasoning, we perform a multi-view spatial measurement by the spatial facts of DIoU, relative distance, geometric size, and overlapping area, which are well-designed for keeping the characteristic of the object and OCR token and refine the message passing between object-OCR token, object-object, and OCR-OCR tokens in the graph structure.

III. Method: Sparse Spatial Graph Network

In this work, we devote to addressing the relation reasoning between the detected objects and OCR tokens for the TextVQA task. We propose a Sparse Spatial Graph Network (named SSGN) for relation inference. As shown in Fig. 2, the question answering process involves three steps: (1) obtaining features of objects, OCR tokens, and the question, and then building a spatial-aware graph network with objects and OCR tokens (detailed in Sec. III-A), (2) performing spatial-aware relation pruning and implementing a hierarchical sparse spatial graph learning, where the graph involves the object-object, OCR-OCR tokens, and object-OCR token relations (the core part of our method introduced in Sec. III-B), and (3) updating the features of object and OCR token nodes in the graph and feeding them into an iterative answer decoder for answer prediction (detailed in Sec. III-C).

A. Preliminary

1) Feature Preparation: For an image I, we extract the initial object features by pre-trained Faster R-CNN [43]. The initial OCR token features are extracted by OCR systems, such as Rosetta-en [6], SBD-Trans [40], Google-OCR[1], and Microsoft-OCR[2]. Besides, we obtain a initial question feature by a fine-tuned three-layer BERT-BASE [10] where the number of hidden layer in BertEncoder is set to 3. Following the previous work [39], we input the initial features of the objects, OCR tokens, and the question into a transformer-based encoder architecture and update the features to $\mathcal{Q}=\{q_i\}_{i=1}^{K}$, $\mathcal{V}=\{v_i\}_{i=1}^{N}$, and $\mathcal{T}=\{t_i\}_{i=1}^{M}$, $v_i, t_i, q_i \in \mathbb{R}^d$, where N, M, and K is the number of objects, OCR tokens, question words, respectively.

2) Spatial-Aware Graph: To acquire the relations of object-object, OCR-OCR tokens, and object-OCR token, we construct a spatial-aware graph $\mathcal{G}=\{\mathcal{N}, \mathcal{E}\}$, where $\mathcal{N}$ is a node set that includes all the objects and OCR tokens, $\mathcal{N} = \mathcal{V} \cup \mathcal{T}$, and $\mathcal{E}$ is a directed edge set denoted as $\mathcal{E} = (\mathcal{E}_{\mathcal{V}\to\mathcal{V}}, \mathcal{E}_{\mathcal{T}\to\mathcal{T}}, \mathcal{E}_{\mathcal{V}\to\mathcal{T}}, \mathcal{E}_{\mathcal{T}\to\mathcal{V}})$. Among $\mathcal{E}$, $\mathcal{E}_{\mathcal{V}\to\mathcal{V}}$ and $\mathcal{E}_{\mathcal{T}\to\mathcal{T}}$ are two subsets that describe object-object and OCR-OCR tokens relations respectively; $\mathcal{E}_{\mathcal{V}\to\mathcal{T}}$ and $\mathcal{E}_{\mathcal{T}\to\mathcal{V}}$ are two edge subsets that contain the object-OCR edges with different message passing directions, *i.e*, from object to OCR token, and from OCR token to object. To make full use of the spatial information in the image, we use spatial coordinates of visual entities to model their explicit relations (edges in the graph). To be specific, we consider the relative distance and height-width ratio of the boundary boxes of two visual entities in the image. Taking $\mathcal{E}_{i\to j}$ as an example, we take node n_j as the reference node and measure the distance of node n_i to it. The edge feature vector $\mathcal{E}_{i\to j} \in \mathbb{R}^{d_e}$ is encoded below, where $d_e = 5$.

$$\mathcal{E}_{i\to j} = \left[\frac{x_i^{tl}-x_j^c}{w_j}, \frac{y_i^{tl}-y_j^c}{h_j}, \frac{x_i^{br}-x_j^c}{w_j}, \frac{y_i^{br}-y_j^c}{h_j}, \frac{w_i * h_i}{w_j * h_j}\right], \quad (1)$$

where $[x_i^{tl}, y_i^{tl}]$, $[x_i^{br}, y_i^{br}]$ denote the top-left and bottom-right coordinates of the bounding box of node n_i, $[x_i^c, y_i^c]$ and $[w_i, h_i]$ denote the center coordinate, width and height of bounding box of node n_i.

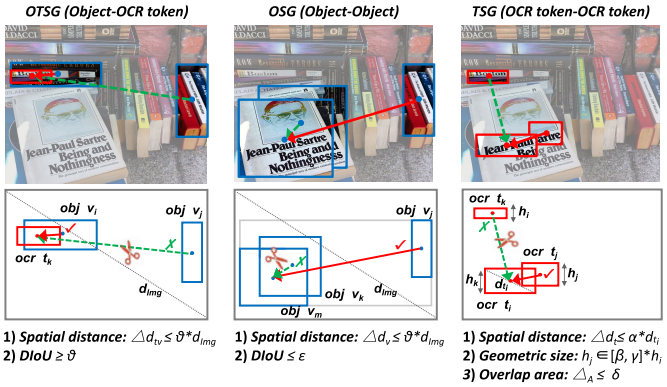

Fig. 3. Relation pruning based on spatial factors. We expect to establish the relations by co-occurring nearby object and token pair in OTSG and remove the redundancy of overlapping objects in OSG and distant tokens in TSG.

B. Sparse Spatial Graph

Each image contains a plethora of visual entities including objects and OCR tokens. The TextVQA task is characterized by reasoning out one or a few specified scene texts or objects in a local region of an image to answer the question. Driven by the question, there are many unnecessary visual relations in reasoning process that may cause interference. Motivated by the visual spatial relations in the natural image scene, as shown in Fig. 3, we seek to remove the redundant relations of much-overlapping objects and distant tokens, and enhance the relations by co-occurring nearby object and token pairs.

As shown in Fig. 2, we perform a hierarchical graph inference. The whole process is implemented under the question guidance. About the relation pruning, Object-Token Sparse Graph (Sec. III-B.1) first performs the correlation of nearby objects and OCR tokens to filter out useless object-OCR relations. Next, we conduct the Object-Based Sparse Graph (Sec. III-B.2) and Token-Based Sparse Graph (Sec. III-B.3) in each entity space (*i.e.*, object or OCR token) in parallel, which further refines the correlated relations for answer prediction.

[1]Google-OCR API: https://cloud.google.com/products/ai/https://cloud.google.com/products/ai/

[2]Microsoft-OCR API: https://azure.microsoft.com/en-us/services/cognitive-services/computer-vision/

1) Object-Token Sparse Graph (OTSG): We observe that an OCR token is informative for answer prediction if it is close to one or more objects, and vice versa for an object. This observation prompts us to judge the necessity of the relations between objects and tokens through spatial measurements. We try to cut off useless edges in the graph to facilitate graph relation reasoning, that is, each node in the graph only receives informative messages from its interactive neighbor nodes within a reasonable spatial scale. The implementation details are introduced as follows.

a) OTSG definition: There is a sub-graph $\mathcal{G}_{\mathcal{VT}} = \{\mathcal{N}, \mathcal{E}_{\mathcal{VT}}\} \subset \mathcal{G}$, where $\mathcal{N} = (\mathcal{V}, \mathcal{T})$, $\mathcal{E}_{\mathcal{VT}} = (\mathcal{E}_{\mathcal{V}\to\mathcal{T}}, \mathcal{E}_{\mathcal{T}\to\mathcal{V}})$, where $\mathcal{E}_{\mathcal{V}\to\mathcal{T}} \in \mathbb{R}^{N\times M\times d_e}$, $\mathcal{E}_{\mathcal{T}\to\mathcal{V}} \in \mathbb{R}^{M\times N\times d_e}$. Taking an edge $\mathcal{E}_{t\to v}$ from OCR token $\tilde{t}$ to object $\tilde{v}$ as an example, the edge feature $\mathcal{E}_{t\to v}$ is performed by Eq. 1, and the feature of reverse edge $\mathcal{E}_{v\to t}$ is calculated by Eq. 1 too.

b) Spatial relation pruning: In this case, we deem that the distant object and OCR token have less or even no correlated information. We design the sparsity function with two spatial pruning constraints: ① *Spatial distance*. It constrains the establishment of relations between objects and OCR tokens when the relative distance Δd_{tv} between token t_j and object v_i is greater than $\theta * d_{Img}$, where d_{Img} is the diagonal length of image and θ is a hyperparameter; ② *DIoU* (*Distance-IoU*) [59] is a good way to measure the spatial distance and geometric similarity of two boundary boxes. It is formulated in Eq. 2. Here, we use *DIoU* as a pruning criterion and its value is required to be greater than θ. In OTSG, the greater the overlap area of the object-OCR token pair is, the closer their relation is.

$$DIoU = IoU - \frac{\rho^2\left(n_i, n_j\right)}{c^2}, \tag{2}$$

where $\rho(\cdot)$ denotes the Euclidean distance and c is the smallest diagonal length of the bounding box covering both the two nodes n_i and n_j. IoU is the spatial criterion of *Intersection over Union* [43].

We use the *sparsity function* to update the edge set $\mathcal{E}_{\mathcal{T}\to\mathcal{V}}$ of $\mathcal{G}_{\mathcal{VT}}$ as follows.

$$\mathrm{E}_{t_j\to v_i} = \begin{cases} \mathcal{E}_{t_j\to v_i}, & if\ \Delta d_{tv} \leqslant \theta * d_{Img} \text{ or } DIoU \\ & \geqslant \theta; \\ [0,0,0,0,0], & else, \end{cases} \tag{3}$$

where $\Delta d_{tv} = \rho(t_j, v_i)$. Following the spatial setting in previous work [37], [54], [56], we keep the default setting of $\theta = 0.5$.

c) Graph inference: After relation pruning, we implement the graph learning to update all the object and token node features of sub-graph $\mathcal{G}_{\mathcal{VT}}$. Taking an object node in $\mathcal{V}$ as an example, we perform a question-guided correlation between object and OCR token. Specifically, we calculate the relation matrix $\mathcal{A}_{\mathcal{T}\to\mathcal{V}} \in \mathbb{R}^{N\times M}$ for message passing from OCR tokens to objects in $\mathcal{G}_{\mathcal{VT}}$. We encapsulate this process as a unified $MP(\cdot)$ function below.

$$\begin{aligned} \mathcal{A}_{\mathcal{T}\to\mathcal{V}} &= MP(\mathcal{G}_{\mathcal{VT}}, \mathcal{E}_{\mathcal{T}\to\mathcal{V}}, \mathcal{Q}) \\ &\Leftrightarrow \begin{cases} q = \sum_{i=1}^{K} softmax(\boldsymbol{W_{q_i}} q_i) \cdot q_i; \\ a = tanh\left(\boldsymbol{W_e}\mathcal{E}_{\mathcal{T}\to\mathcal{V}} + \boldsymbol{W_q} q\right); \\ \mathcal{A}_{\mathcal{T}\to\mathcal{V}} = softmax\left(\boldsymbol{W_a} a\right), \end{cases} \end{aligned} \tag{4}$$

where $\boldsymbol{W_{q_i}}$, $\boldsymbol{W_e}$, $\boldsymbol{W_q}$, $\boldsymbol{W_a}$ are learnable parameters. $\mathcal{A}_{\mathcal{T}\to\mathcal{V}}$ is a similarity matrix, where its element $\mathcal{A}_{\mathcal{T}\to\mathcal{V}ij}$ denotes the correlation weight of two nodes v_i and t_j, whose value is in the range of (0,1).

Then, we update the object node set $\mathcal{V}$ with the edge set $\mathcal{E}_{\mathcal{T}\to\mathcal{V}}$ and the relation matrix $\mathcal{A}_{\mathcal{T}\to\mathcal{V}}$. Concretely, the node v_i is updated by receiving the messages from its "connected" token neighbors in the sparse sub-graph. Please note that the token neighbors only exists in the edges $\mathcal{E}_{\mathcal{T}\to\mathcal{V}}$ updated by Eq. 3 rather than all the tokens. We encapsulate the graph inference process as a unified $GIN(\cdot)$ function below.

$$\begin{aligned} \mathcal{V}' &= GIN(\mathcal{G}_{\mathcal{VT}}, \mathcal{E}_{\mathcal{T}\to\mathcal{V}}, \mathcal{V}, \mathcal{T}, \mathcal{A}_{\mathcal{T}\to\mathcal{V}}) \\ &\Leftrightarrow \begin{cases} \mathcal{E}'_{\mathcal{T}\to\mathcal{V}} = \mathcal{A}_{\mathcal{T}\to\mathcal{V}} \cdot \boldsymbol{W_{\mathcal{E}}}\mathcal{E}_{\mathcal{T}\to\mathcal{V}}; \\ \mathcal{M}_{\mathcal{T}\to\mathcal{V}} = \boldsymbol{W_{\mathcal{T}}}\mathcal{A}_{\mathcal{T}\to\mathcal{V}}\mathcal{T}; \\ \mathcal{V}' = \boldsymbol{W_{\mathcal{V}}}\mathcal{V} + \boldsymbol{W_{\mathcal{E}'}}\mathcal{E}'_{\mathcal{T}\to\mathcal{V}} + \boldsymbol{W_{\mathcal{M}}}\mathcal{M}_{\mathcal{T}\to\mathcal{V}}, \end{cases} \end{aligned} \tag{5}$$

where $\boldsymbol{W_{\mathcal{E}}}$, $\boldsymbol{W_{\mathcal{T}}}$, $\boldsymbol{W_{\mathcal{V}}}$, $\boldsymbol{W_{\mathcal{E}'}}$, $\boldsymbol{W_{\mathcal{M}}}$ are learnable parameters. We obtain a new object representation $\mathcal{V}'=\{v'_i\}_{i=1}^{N} \in \mathbb{R}^{N\times d}$.

In the same way, we obtain the reverse relation $\mathcal{E}_{\mathcal{V}\to\mathcal{T}}$ and update any OCR token node in $\mathcal{T}$ by receiving messages from its connected object nodes. As a result, we obtain a new OCR token representation $\mathcal{T}'=\{t'_j\}_{j=1}^{M} \in \mathbb{R}^{M\times d}$.

2) Object-Based Sparse Graph (OSG): After the above double-sided verification of object and OCR token occurrences, here we narrow the range of relation learning in each visual entity space. In this part, we study the object-object relations in the graph. The objects are always densely detected by the pre-trained detection models (such as Faster RCNN [43] in this work). Superfluous bounding boxes with different sizes covering a same entity often appear, but they are taken as different objects. By observation, these close-by objects have similar geometric sizes and visual appearances, their high similarities result in the strongly intensified relations between them, which may cause imbalanced relations and negatively affect the answer reasoning. To achieve effective visual reasoning, we try to reduce this interference.

a) OSG definition: In this part, we focus on the object entities. Here is a sub-graph $\mathcal{G}_{\mathcal{V}} = \{\mathcal{N}_{\mathcal{V}}, \mathcal{E}_{\mathcal{V}}\}$, $\mathcal{N}_{\mathcal{V}} = \mathcal{V}'$, $\mathcal{E}_{\mathcal{V}} = \mathcal{E}_{\mathcal{V}\to\mathcal{V}} \in \mathbb{R}^{N\times N\times d_e}$, where the edges $\mathcal{E}_{v_i\to v_j}$ and $\mathcal{E}_{v_j\to v_i}$ are also calculated by Eq. 1.

b) Spatial relation pruning: To address the redundant relations of these similar and close-by objects, we employ *DIoU* [59] and *spatial distance* again as spatial pruning criteria. If DIoU is greater than ϵ, it means that two close objects exist and the connected edge between them has to be cut off. We use the *sparsity function* to update the edge set

$\mathcal{E}_{\mathcal{V}\to\mathcal{V}}$ of $\mathcal{G}_{\mathcal{V}}$ below.

$$\mathrm{E}_{v_j\to v_i} = \begin{cases} \mathcal{E}_{v_j\to v_i}, & if\ \Delta d_v \leqslant \theta * d_{Img}\ \text{and}\mathrm{DIoU} \leqslant \epsilon; \\ [0,0,0,0,0], & else, \end{cases} \tag{6}$$

where ϵ is a hyperparameter. We set $\Delta d_v = \rho(v_i', v_j')$ and θ =0.5 as the same in Eq. 3.

c) Graph inference: As the same to the above sub-graph $\mathcal{G}_{\mathcal{VT}}$, we perform the question-guided object graph learning on $\mathcal{G}_{\mathcal{V}}$. The relation matrix $\mathcal{A}_{\mathcal{V}} \in \mathbb{R}^{N\times N}$ of $\mathcal{G}_{\mathcal{V}}$ is calculated below.

$$\mathcal{A}_{\mathcal{V}} = MP(\mathcal{G}_{\mathcal{V}}, \mathcal{E}_{\mathcal{V}}, \mathcal{Q}) \tag{7}$$

For the object-object relations in $\mathcal{E}_{\mathcal{V}\to\mathcal{V}}$, we update object features from $\mathcal{V}'$ to $\mathcal{V}''=\{v_i''\}_{i=1}^{N} \in \mathbb{R}^{N\times d}$ by message passing from its "connected" neighbor objects in the sparse sub-graph OSG (please note that does not refer to all the objects in $\mathcal{V}'$) as follows.

$$\mathcal{V}'' = GIN(\mathcal{G}_{\mathcal{V}}, \mathcal{E}_{\mathcal{V}}, \mathcal{V}', \mathcal{V}', \mathcal{A}_{\mathcal{V}}) \tag{8}$$

3) Token-Based Sparse Graph (TSG): Here, we discuss the OCR-OCR tokens relations. In the previous work [21], [28], [50], [51], [52], scene text recognition has been proven to be significant for TextVQA. Unlike object detection, OCR tokens are detected in relatively small numbers and are independently dispersed, with few or no overlapping regions in the image. As shown in Fig. 3, a small overlap area exists between t_i "Jean-Paul" and t_j "Sartre". t_i and t_j are semantically relevant as they compose the name of a famous philosopher, where the distant t_k "Boston" is a city name. The close-by tokens are much more semantic-relevant than distant tokens. Under this consideration, we design a pruning rule for OCR tokens as below.

a) TSG definition: The OCR tokens sub-graph is denoted as $\mathcal{G}_{\mathcal{T}} = \{\mathcal{N}_{\mathcal{T}}, \mathcal{E}_{\mathcal{T}}\}$, $\mathcal{N}_{\mathcal{T}} = \mathcal{T}'$, $\mathcal{E}_{\mathcal{T}} = \mathcal{E}_{\mathcal{T}\to\mathcal{T}} \in \mathbb{R}^{M\times M\times d_e}$, where the edges $\mathcal{E}_{t_i\to t_j}$ and $\mathcal{E}_{t_j\to t_i}$ are also calculated by Eq. 1.

b) Spatial relation pruning: Taking $\mathcal{E}_{t_j\to t_i}$ as an example, we constrain it with the following factors: ① *Spatial distance*. It requires that $\Delta d_t \leqslant \alpha * d_{t_i}$, where Δd_t denotes the shortest bounding box distance between tokens t_i' and t_j', d_{t_i} represents the diagonal length of t_i''s bounding box, and α is a hyperparameter. Opposite to the redundancy of close-by objects, close-by tokens have a close relation. Thus, we cut off the $\mathcal{E}_{t_j\to t_i}$ if it does not meet this condition. ② *Geometric size*. We restrict the heights of close-by OCR tokens. By observation, the tokens gathering together to describe a phrase or sentence are often in similar font sizes. Another fact is that different font sizes reflect different semantic importance. We attempt to find out the scene texts at the same semantic level in the image. We set $\beta * h_i \leqslant h_j \leqslant \gamma * h_i$, where h_i and h_j denote the normalized height of t_i' and t_j' in the image, and β, γ are hyperparameters. ③ *Overlap area.* To make sure the readability of the OCR tokens, various OCR systems make efforts to output less or no overlap tokens. We follow this rule to calculate an overlap ratio $\Delta_A = max(\frac{A_{ij}}{A_i}, \frac{A_{ij}}{A_j})$ of tokens t_i' and t_j' and set it less than threshold δ, where A_i and A_j are bounding box areas of t_i' and t_j', respectively. A_{ij} is the intersection area between the bounding boxes of t_i' and t_j'. To summarize, we use the *sparsity function* to update the edge set $\mathcal{E}_{\mathcal{T}\to\mathcal{T}}$ of $\mathcal{G}_{\mathcal{T}}$ as follows.

$$\mathrm{E}_{t_j\to t_i} = \begin{cases} \mathcal{E}_{t_j\to t_i}, & if\ \Delta d_t \leqslant \alpha * d_{t_i},\ h_j \in [\beta,\gamma] * h_i, \\ & \quad \text{and}\ \Delta_A \leqslant \delta; \\ [0,0,0,0,0], & else, \end{cases} \tag{9}$$

where α, β and γ are hyperparameters. We set the threshold $\delta = 0.5$ following the OCR spatial setting [50].

c) Graph inference: Up to now, the sparse sub-graph $\mathcal{G}_{\mathcal{T}}$ referring to OCR token nodes is built. The token graph learning process of $\mathcal{G}_{\mathcal{T}}$ is performed the same as $\mathcal{G}_{\mathcal{T}}$. We first calculate the relation matrix $\mathcal{A}_{\mathcal{T}} = MP(\mathcal{G}_{\mathcal{T}}, \mathcal{E}_{\mathcal{T}}, \mathcal{Q}) \in \mathbb{R}^{M\times M}$ to perform the question-guided message passing among tokens. Any OCR token node t_j' is updated by its "connected" neighbor tokens. At last, we update the token node set by the unified graph inference function $\mathcal{T}'' = GIN(\mathcal{G}_{\mathcal{T}}, \mathcal{E}_{\mathcal{T}}, \mathcal{T}', \mathcal{T}', \mathcal{A}_{\mathcal{T}}) \in \mathbb{R}^{M\times d}$.

C. Answer Generation

1) Answer Prediction: Based on the final output node representations $\mathcal{V}''$ and $\mathcal{T}''$, we adopt an available text generator [22], [39] for answer prediction, which is made up of a transformer and two classifiers—the object classifier ψ_o and the OCR token classifier ψ_t. We concatenate $\mathcal{Q}$, $\mathcal{V}''$, $\mathcal{T}''$ and a hidden state $o \in \mathbb{R}^d$ and input them into a transformer module as follows.

$$[\check{\mathcal{Q}}, \check{\mathcal{V}}, \check{\mathcal{T}}, \check{o}] = \Psi([\boldsymbol{W_Q}\mathcal{Q}, \boldsymbol{W_{\mathcal{V}''}}\mathcal{V}'', \boldsymbol{W_{\mathcal{T}''}}\mathcal{T}'', \boldsymbol{W_o}o]), \tag{10}$$

where $\boldsymbol{W_Q}$, $\boldsymbol{W_{\mathcal{V}''}}$, $\boldsymbol{W_{\mathcal{T}''}}$, $\boldsymbol{W_o}$ are learnable parameters, $\Psi(\cdot)$ is a four-layer transformer, and o is initialized by positional embedding [22]. We perform Eq. 10 L times, thus we obtain a generated sequence $\check{\mathcal{O}} = [\check{o}_1, \cdots, \check{o}_L] \in \mathbb{R}^{d\times L}$.

This part can be regarded as a multi-label classification. At each l-th decoding time, the object classifier ψ_o is realized by a simple linear layer and predicts the probability score y_l^o over a pre-set object vocabulary. Another classifier, ψ_t, is proposed to compute the token score y_l^t by the dot product of the generated $\check{o}_l$ and OCR tokens $\check{\mathcal{T}}$, where $\check{\mathcal{T}}$ is a dynamic OCR token set detected in each image. Formally, the predicted scores y_l^o and y_l^t are calculated as below:

$$\begin{cases} y_l^o = \boldsymbol{W_l^o}\check{o}_l + \boldsymbol{b_l^o}; \\ y_l^t = (\boldsymbol{W_l^{t1}}\check{\mathcal{T}} + \boldsymbol{b_l^{t1}})^\top (\boldsymbol{W_l^{t2}}\check{o}_l + \boldsymbol{b_l^{t2}}), \end{cases} \tag{11}$$

where $\boldsymbol{W_l^o}$, $\boldsymbol{W_l^{t1}}$, $\boldsymbol{W_l^{t2}}$ are learnable parameters and $\boldsymbol{b_l^o}$, $\boldsymbol{b_l^{t1}}$, $\boldsymbol{b_l^{t2}}$ are scalar parameters at the l-th timestamp.

At last, we implement the $argmax$ function on y_l^o and y_l^t to predict the answer word y_l^{pred}. Thus, the answer sentence with length L, $y^{pred} = \{y_l^{pred}\}_{l=1}^{L}$ is represented as:

$$y_l^{pred} = argmax([y_l^o, y_l^t]), \tag{12}$$

2) Training Loss: Following the previous work [22], [39], binary cross-entropy loss $\mathcal{L}_{bce}$ is widely used for TextVQA. In real applications, the utterly correct answer sentence is expected but rarely occurs, whereas the answer with semantically similar words is acceptable. Following [39], a new auxiliary policy gradient loss $\mathcal{L}_{pg}$ based on ANLS (*Average Normalized Levenshtein Similarity* (stated in Eq. 15 [3]) is introduced into this task [46]. The ANLS measures the character-level composition similarity between the predicted and ground-truth answers as follows.

$$\begin{cases} \mathcal{L}_{bce} & = -y^{gt}\log(\sigma(y^{pred})) - (1-y^{gt})\log(1-\sigma(y^{pred})); \\ \mathcal{L}_{pg} & = -\log(\sigma(y^{pred}))\cdot \mathrm{ANLS}(y^{gt}, y^{pred}), \end{cases} \tag{13}$$

where $\sigma(\cdot)$ is sigmiod function, y^{gt} is the ground-truth.

By combing $\mathcal{L}_{bce}$ and $\mathcal{L}_{pg}$, the total loss is formulated as follows.

$$\mathcal{L} = \mathcal{L}_{bce} + \lambda\mathcal{L}_{pg}, \tag{14}$$

where λ is a trade-off hyperparameter.

IV. Experiment

A. Datasets

Experiments are conducted on two benchmark datasets of text-based visual question answering.

1) TextVQA [46]: In this dataset, images are collected from Open Images v3 [34]. It contains 28,408 images and 45,336 questions, which consists of a training set of 21,953 images and 34,602 question-answer (QA) pairs, a validation set of 3,166 images and 5,000 QA pairs, and a test set of 3,289 images and 5,734 QA pairs [22]. For each image, there are about one or two QA pairs. The average lengths of question and answer are 7.18 and 1.70 words, respectively. Questions in this dataset are interested in visual objects or scene texts in the images. Up to 39% (about 18K) of answers do not contain any OCR token.

2) ST-VQA [3]: The dataset comprises 23,038 images and 31,791 questions, which is collected from six different datasets of ICDAR 2013 [32], ICDAR 2015 [31], ImageNet [9], VizWiz [20], IIIT Scene Text Retrieval [42], Visual Genome [35], and COCO-Text [48]. Following the protocol [22], this dataset is divided into the train/val/test sets of 17,028/1,893/2,971 images and 23,446/2,628/4,070 QA pairs, respectively. Compared with the TextVQA dataset, ST-VQA more emphasizes scene texts as all the questions have to be answered with scene texts. Each image contains more than two scene texts, regardless of whether or not they can be detected by the OCR systems. ST-VQA introduces three novel tasks, namely the strongly contextualized task (Task 1), the weakly contextualized task (Task 2), and the open vocabulary task (Task 3). Specifically, Task 1 provides a dynamic candidate dictionary of 100 words defined for per image; Task 2 provides a fixed answer dictionary of 30, 000 words for the whole dataset; following [22], Task 3 provides a fixed answer dictionary of 5,000 words for the whole dataset.

TABLE I

THE STATISTICS OF OCR TOKEN OUTPUT BY DIFFERENT OCR SYSTEMS ON TEXTVQA AND ST-VQA DATASETS. *Total* REPRESENTS THE SUM NUMBER OF OCR TOKENS. *Mean*, *Min* AND *Max* REPRESENT THE AVERAGE, MINIMUM, AND MAXIMUM NUMBERS OF OCR TOKENS PER IMAGE. *Min* = 0 OCCURS IN THE CASES OF THE BLUR, PARTIALLY OCCLUDED SCENE TEXTS, OR SCENE TEXT IN ILLEGIBLE FANCY FONTS, *etc.* *Max* = 2694 OCCURS IN CASE OF SCENE TEXT FOR READING, SUCH AS BOOK PAGES. IN OUR EXPERIMENTS, WE CHOOSE M = 50 OCR TOKENS FOR EXPERIMENTS

Dataset	OCR System	Total	Mean	Min	Max
TextVQA	Rosetta-en	566,824	12.50	0	100
	SBD-Trans	914,521	20.17	0	247
	Google-OCR	1,304,155	28.77	0	2,657
	Microsoft-OCR	1,419,941	31.32	0	2,694
ST-VQA	Rosetta-en	226,643	7.52	0	99
	SBD-Trans	266,947	8.86	0	100
	Google-OCR	292,196	9.69	0	477
	Microsoft-OCR	359,137	11.91	0	358

B. Evaluation Metrics

We adopt accuracy (***Acc***) as a basic evaluation metric following [22], [30], [60]. Each question in the TextVQA dataset has ten human-annotated answers, and the final accuracy is the average score over these ten answers. As for ST-VQA [3], there is a new evaluation metric ***ANLS*** which measures the frequency of correct words in each generated answer as defined in Eq. 15. ***ANLS*** is calculated in term of the correct words, while ***Acc*** is calculated in term of the complete sentence.

$$ANLS(y^{pred}, y^{gt}) = 1 - \frac{NL(y^{pred}, y^{gt})}{max(|y^{pred}|, |y^{gt}|)}. \tag{15}$$

where *NL* denotes the normalized Levenshtein distance [36], y^{pred} and y^{gt} denote the predicted and ground-truth answers, respectively. As set in ST-VQA [3], the score *ANLS* is set to 0 if it is below 0.5.

C. Implementation Details

For feature extraction, we follow the experimental settings of [3] and [46]. We use the Faster R-CNN [43] pre-trained on Visual Genome dataset [35] to detect objects. Each object has a 2048-dim appearance feature and a 4-dim boundary box feature. They are encoded by a separate fully connected layer and then added into a 768-dim vector and used as an original object feature. As for OCR tokens, we conduct experiments on four OCR systems, *i.e*, Rosetta-en [6], SBD-Trans [40], Google-OCR [1], and Microsoft-OCR [2]. Each OCR token feature consists of four aspects, including 300-dim FastText feature [5], 604-dim PHOC (pyramidal histogram of characters) feature [1], 2048-dim appearance feature, and 4-dim bounding box feature. Following the previous work for TextVQA [22], [39], [58], all features are encoded by a separate fully connected layer and then added to obtain a 768-dim vector. Table I shows the statistics of OCR tokens detected on the TextVQA and ST-VQA datasets. In our experiments,

we choose $N = 100$ objects with the best probabilities and $M = 50$ OCR tokens following [22], [57], [58], [60].

About the other experiment setups, each question sentence is truncated with the length $K = 20$ and equipped with 768-dim word embedding. In this work, we perform a two-layer transformer with 12 heads for feature preparation in Sec. III-A.1 and a four-layer transformer with 12 heads for answer decoding in Sec. III-C. The maximum length of the output answer $L = 12$, and the trade-off parameter in the total loss objective is set to $\lambda = 1$. The threshold θ is set to 0.5 following the setting of spatial exploration [56]. We set empirical parameters with $\epsilon = 0.3$, $\alpha = 5$, $\beta = 0.3$ and $\gamma = 2.0$. We choose Adam as the optimizer, and the learning rate is set to 1e−4. During training, we multiply the learning rate by 0.1 at 10,000 and 21,000 iterations, respectively, for a total of 24,000 iterations.

D. Comparison With State-of-the-Art Results

In this subsection, we compare the proposed method with the state-of-the-art approaches—**Attention-based models** (LoRRA [46], SSBaseline [60]), **Transformer-based models** (M4C [22], LaAP-Net [21], PAT [58], TAP [55], LOGOS [41]), **Representation learning models** (BOV [57]), and **Graph-based models** (MM-GNN [14], CRN [39], SA-M4C [30], SMA [13]).

1) Results on TextVQA: As the experimental results shown in Table II, the proposed **SSGN** achieves a promising performance compared to state-of-the-art methods. Compared with **SSBaseline** [46] (an attention-based model), under the same setup of SBD-Trans features and ST-VQA [3] training data, **SSGN** improves 1.43% on the val set and 0.97% on the test set. Compared with **BOV** [57] (a representation learning model) based on SBD-Trans features, **SSGN (Ours)** improves 0.85% on the val set and 1% on the test set, and when taking ST-VQA as extra training data, **SSGN (Ours)** further improves 1.24% on the val set. Compared with **LaAP-Net** [21] (a transformer-based model without pre-trained techniques), **SSGN (Ours)** improves 1.32% on the val set and 1.06% on the test set. By introducing pre-trained techniques such as MLM (masked language modeling), ITM (image-text matching), RPP (relative position prediction) or large-scale extra data such as the Visual Genome dataset (which includes 108,000 images, 5.4 million descriptions, 1.7 million QA pairs and 2.3 million relation annotations) [35], **TAP** [55] and **LOGOS** [41] obtains higher performance than ours.

The fairest comparison exists among the graph-based models. Compared with **MM-GNN** [14] (a total fully-connected graph model), when using Rosetta-en features, our model achieves 10.56% and 10.50% improvements on the val and test sets respectively. **CRN** [39] constructs a fully-connected graph that merely explores the relations between objects and OCR tokens, namely ignoring object-object and token-token relations. Our model improves upon **CRN** by 1.61% on the val set and 0.64% on the test set. **SA-M4C** [30] introduces a spatial orientation factor into the graph modeling but does not consider the relation redundancy in the graph. Compared with **SA-M4C** [30], our model achieves 0.05% and 0.82% improvements on the val and test sets when using Google-OCR features and ST-VQA data for training. In **SMA** [13], each node adaptively selects the top-5 nearest neighbors for relation learning. When using the SBD-Trans features and extra training data, our model achieves 2.38% and 1.12% improvements on the val set and the test set than **SMA**.

TABLE II

MAIN COMPARISON ON TEXTVQA DATASET. THE GREY BLOCK MARKS THE PRE-TRAINING TECHNIQUES. * DENOTES THAT THE MODEL IS JOINTLY TRAINED WITH EXTRA PRE-TRAINING TASKS OF MLM, ITM, AND RPP. † DENOTES JOINTLY PRE-TRAINING WITH THE QUESTION-VISUAL GROUNDING TASK. VG DENOTES THE VISUAL GENOME DATASET [35]. "F" DENOTES FULL RELATIONS, AND "S" DENOTES SPARSE RELATIONS

Method	OCR System	Extra Data	Val Acc	Test Acc
Attention-based models				
LoRRA [46] (F)	Rosetta-ml	-	26.56	27.63
SSBaseline [60] (F)	Rosetta-en	-	40.38	40.92
SSBaseline [60] (F)	SBD-Trans	-	43.95	44.72
SSBaseline [60] (F)	SBD-Trans	ST-VQA	45.53	45.66
Transformer-based models				
M4C [22] (F)	Rosetta-en	-	39.40	39.01
M4C [22] (F)	Rosetta-en	ST-VQA	40.55	40.46
LaAP-Net [21] (F)	Rosetta-en	-	40.68	40.54
LaAP-Net [21] (F)	Rosetta-en	ST-VQA	41.02	40.54
PAT [58] (F)	Google-OCR	-	42.80	43.41
TAP* [55] (F)	Microsoft-OCR	-	49.91	49.71
TAP* [55] (F)	Microsoft-OCR	ST-VQA	50.57	50.71
LOGOS† [41] (F)	Microsoft-OCR	VG	50.79	50.65
LOGOS† [41] (F)	Microsoft-OCR	ST-VQA, VG	51.53	51.08
Representation learning models				
BOV [57] (F)	Rosetta-en	-	40.90	41.23
BOV [57] (F)	SBD-Trans	-	44.87	45.63
BOV [57] (F)	SBD-Trans	ST-VQA	46.24	46.96
Graph-based models				
MM-GNN [14] (F)	Rosetta-ml	-	31.44	31.10
CRN [39] (F)	Rosetta-en	-	40.39	40.96
SA-M4C [30] (F)	Google-OCR	-	43.90	-
SA-M4C [30] (F)	Google-OCR	ST-VQA	45.40	44.60
SMA [13] (S)	Rosetta-en	-	40.05	40.66
SMA [13] (S)	SBD-Trans	-	44.58	45.51
SMA [13] (S)	SBD-Trans	ST-VQA	44.58	45.51
SSGN (Ours) (S)	Rosetta-en	-	42.00	41.60
SSGN (Ours) (S)	SBD-Trans	-	45.72	46.63
SSGN (Ours) (S)	SBD-Trans	ST-VQA	**46.96**	46.63
SSGN (Ours) (S)	Google-OCR	ST-VQA	45.45	45.42
SSGN (Ours) (S)	**Microsoft-OCR**	**ST-VQA**	46.85	**47.16**

*2) **Results on ST-VQA**:* The questions in the ST-VQA dataset are answered more explicitly by utilizing the OCR tokens in the images than in the TextVQA dataset. From Table III, our method shows absolute superiority on the three tasks of ST-VQA dataset. In Task 1, **SSGN (Ours)** obtains 48.7% on Test ANLS and surpasses **MM-GNN** [14] by 28.7%. In Task 2, **SSGN (Ours)** reaches 49.5% on Test ANLS and outperforms **CRN** [39] by 1.3%. In Task 3, when using Microsoft-OCR, **SSGN (Ours)** achieves 58.9% on Val ANLS and 48.81% on Val Acc, surpassing all other methods on the val set. The results show that our model has competitive performance compared with other methods in all three tasks.

E. Role of Graph Module

In this subsection, we test the role of each graph module in our method. As shown in Table IV, **w/o graph** that removes

TABLE III

MAIN COMPARISON ON THREE TASKS OF ST-VQA DATASET. "F" DENOTES FULL RELATIONS, AND "S" DENOTES SPARSE RELATIONS

Method	OCR System	Extra Data	Task 1	Task 2	Task 3		
			Test ANLS	Test ANLS	Val Acc	Val ANLS	Test ANLS
Attention-based models							
SAN+STR [3] (F)	-	-	0.135	0.135	10.46	-	0.135
VTA [4] (F)	-	-	-	0.279	18.13	-	0.282
SSBaseline [60] (F)	SBD-Trans	-	0.506	0.505	-	-	0.509
Transformer-based models							
M4C [22] (F)	Rosetta-en	-	-	-	38.05	0.472	0.462
LaAP-Net [21] (F)	Rosetta-en	-	-	-	39.74	0.497	0.485
PAT [58] (F)	Google-OCR	-	-	-	41.10	-	0.508
TAP* [55] (F)	Microsoft-OCR	-	-	-	45.29	0.551	0.543
LOGOS† [41] (F)	Microsoft-OCR	VG	-	-	44.10	0.535	0.522
LOGOS† [41] (F)	Microsoft-OCR	TextVQA, VG	-	-	48.63	0.581	**0.579**
Representation learning models							
BOV [57] (F)	Rosetta-en	-			40.18	0.500	0.472
Graph-based models							
MM-GNN [14] (F)	Rosetta-ml	-	0.203	-	-	-	0.207
CRN [39] (F)	Rosetta-en	-	-	0.482	-	-	0.483
SA-M4C [30] (F)	Google-OCR	-	-	-	42.23	0.512	0.504
SMA [13] (S)	Rosetta-en	-	-	-	-	-	0.486
SSGN(Ours) (S)	Rosetta-en	-	0.487	0.495	40.12	0.493	0.490
SSGN (Ours) (S)	SBD-Trans	-	0.509	0.507	42.50	0.519	0.507
SSGN (Ours) (S)	SBD-Trans	TextVQA	0.547	0.550	44.72	0.548	0.535
SSGN (Ours) (S)	Google-OCR	TextVQA	0.520	0.526	44.69	0.537	0.523
SSGN (Ours) (S)	**Microsoft-OCR**	**TextVQA**	**0.570**	**0.573**	**48.81**	**0.589**	0.573

TABLE IV

THE PERFORMANCE OF GRAPH MODULES ON TEXTVQA DATASET

Method	OSG	TSG	OTSG	Val Acc	Test Acc
w/o graph	-	-	-	39.92	40.51
OSG	✓	-	-	40.00	41.32
TSG	-	✓	-	41.21	41.51
OTSG	-	-	✓	40.47	40.67
OSG&TSG	✓	✓	-	41.27	41.57
OTSG&OSG	✓	-	✓	40.49	41.18
OTSG&TSG	-	✓	✓	41.63	41.55
OSG&TSG&OTSG	✓	✓	✓	40.14	40.42
OSG&TSG→OTSG	✓	✓	✓	41.24	41.20
OTSG→OSG&TSG (Ours)	✓	✓	✓	**42.00**	**41.60**

TABLE V

ABLATION STUDIES OF GRAPH SPARSITY ON TEXTVQA DATASET

Method	OSG	TSG	OTSG	Val Acc	Test Acc
w/o sparsity	-	-	-	40.99	40.42
w/ OSG sparsity	✓	-	-	41.10	41.10
w/ TSG sparsity	-	✓	-	41.08	41.08
w/ OTSG sparsity	-	-	✓	41.32	41.57
w/o OTSG sparsity	✓	✓	-	41.43	41.32
w/o OSG sparsity	-	✓	✓	41.34	41.30
w/o TSG sparsity	✓	-	✓	41.49	41.51
SSGN (Ours)	✓	✓	✓	**42.00**	**41.60**

all the graph modules in our method shows the most severe performance degradation.

1) Single Sub-Graph: We test a single sub-graph and report the experimental results in Table IV. Single **OTSG** means that we merely save the OTSG graph module and remove the other graph modules in our method. Compared with the full model **SSGN**, **OTSG** decreases 1.53% on the val set and 0.93% on the test set obviously. This indicates that using object-OCR token interaction alone is not sufficient for answer prediction. Single **OSG** with object nodes achieves the worst performance, with a drop of 2% on the val set. The surprising case is the single **TSG** with OCR token nodes, which decreases slightly by 0.79% in the val set compared to the full model. This suggests that the OCR tokens indeed play an important role in this task.

2) Dual Sub-Graphs: We test various combinations of two sub-graphs. Among them, **OTSG&TSG** achieves the best performance, with a slight decrease of 0.37% on the val set compared with the full model **SSGN**. This result is reasonable because OCR tokens are important in both OTSG and TSG graphs. In contrast, **OTSG&OSG** has a substantial decrease of 1.51% on the val set, and even is much worse than the parallel learning of **OSG&TSG** with an accuracy reduction of 0.73%. We speculate that the object redundancy is much more serious than the OCR tokens. The redundant relations between objects may interfere with the reasoning process.

3) Hierarchical Graph Structure: Further, we test the hierarchical graph structure including the following three graph variants: ① A parallel learning of **OSG&TSG&OTSG**. In this case, we concatenate the object features output by OSG and OTSG as $\mathcal{V}''$ and the token features output by TSG and OTSG as $\mathcal{T}''$ for answer generation, ② **OTSG→OSG&TSG** (Ours), and ③ **OSG&TSG→OTSG**, a cascade graph learning approach in the opposite order of **OTSG→OSG&TSG**. Among these structures, **OSG&TSG&OTSG** has the worst performance. Compared to **OTSG→OSG&TSG** (Ours), it drops 1.86% on the val set and 1.18% on the test set. The **OSG&TSG&OTSG** performs even worse on any combination of dual sub-graphs. This may be because redundant relations are amplified in this three-sub-graph parallel learning. By comparing **OSG&TSG→OTSG** and **OTSG→OSG&TSG**, the latter is clearly more effective. We insist on the validity of **OTSG→OSG&TSG**, which first implements the correlation of object-OCR token and then examines the correlated clues in each visual space individually.

F. Ablation Studies

In this subsection, we conduct experiments with Rosetta-en OCR features on the TextVQA dataset to demonstrate the validity of our sparse method in Tables V ∼ Tables VII, and X, and to illustrate the role of heuristics in Tables VIII, IX, and Fig. 4.

1) Sparsity Test: In this part, we discuss the graph sparsity in Tables V and VI. There are different sparsity settings. "**w/ OTSG sparsity**" indicates that our approach just conducts the spatial sparsity of OTSG, while "**w/o OTSG sparsity**" indicates that we just cancel the sparsity operation of OTSG. And the definitions of "**w/ OSG sparsity**", "**w/ TSG sparsity**", "**w/o OSG sparsity**", and "**w/o TSG sparsity**" are similar. For "**w/o sparsity**", OTSG, OSG, and TSG are all performed on fully-connected graphs.

As shown in Table V, under only one sub-graph sparsity, the effect of **w/ OTSG sparsity** is most noticeable due to the importance of object-OCR token correlation. If a local region in the image is recognized by both objects and OCR tokens, it does deserve more attention. Building these relations

TABLE VI

STATISTICS OF SPARSITY RATIO (SR)[3] ON TEXTVQA AND ST-VQA TEST SETS UNDER DIFFERENT OCR SYSTEMS

Dataset	OCR System	OTSG SR (%)	OSG SR (%)	TSG SR (%)
TextVQA	Rosetta-en	10.32	14.66	54.13
	SBD-Trans	10.23	14.66	57.05
	Google-OCR	10.00	14.66	55.63
	Microsoft-OCR	9.85	14.66	53.68
ST-VQA	Rosetta-en	13.24	15.70	51.92
	SBD-Trans	32.36	15.70	55.81
	Google-OCR	13.26	15.70	55.51
	Microsoft-OCR	12.88	15.70	57.06

TABLE VII

STATISTICS OF SPARSITY RATIO (SR) OF TSG ON TEXTVQA AND ST-VQA TEST SETS UNDER DIFFERENT DATA DISTRIBUTIONS OF OCR TOKEN

Dataset	OCR System	TSG SR (%)	Data (%)	TSG SR (%)	Data (%)
		OCR $\leqslant$ 20		OCR $>$ 20	
TextVQA	Rosetta-en	56.29	86.57	40.91	13.43
	SBD-Trans	60.35	78.18	45.55	21.82
	Google-OCR	59.97	77.29	42.39	22.71
	Microsoft-OCR	57.79	73.18	42.98	26.82
		OCR $\leqslant$ 10		OCR $>$ 10	
ST-VQA	Rosetta-en	55.26	80.79	38.30	19.21
	SBD-Trans	58.13	76.24	48.51	23.76
	Google-OCR	58.00	78.38	47.26	21.62
	Microsoft-OCR	65.95	70.40	44.25	29.60

TABLE VIII

ABLATION STUDIES OF VARIOUS IOUS IN OSG ON TEXTVQA DATASET

Method	Threshold ϵ	SR (%)	Val Acc	Test Acc
DIoU	$\epsilon = 0.01$	24.59	41.15	40.93
	$\epsilon = 0.1$	19.71	41.51	41.02
	ϵ **= 0.3**	**14.66**	**42.00**	**41.60**
	$\epsilon = 0.8$	10.49	41.67	41.04
IoU	$\epsilon = 0.3$	15.33	41.47	40.66
GIoU	$\epsilon = 0.3$	14.35	41.03	40.72
CIoU	$\epsilon = 0.3$	14.61	41.66	41.57
DIoU (Ours)	ϵ **= 0.3**	**14.66**	**42.00**	**41.60**

TABLE IX

ABLATION STUDIES OF THRESHOLD θ IN OTSG AND OSG ON TEXTVQA DATASET

Method	Threshold θ	SR (%)	Val Acc	Test Acc
OTSG	$\theta = 0.3$	42.24	41.35	40.98
	$\theta = 0.4$	22.78	41.56	41.21
	θ **= 0.5**	**10.32**	**42.00**	**41.60**
	$\theta = 0.6$	00.04	41.11	40.69
OSG	$\theta = 0.3$	46.15	41.10	41.03
	$\theta = 0.4$	26.89	41.51	41.13
	θ **= 0.5**	**14.66**	**42.00**	**41.60**
	$\theta = 0.6$	00.08	41.33	40.97

with object-OCR co-occurrence are instructive for predicting answers. Besides, the performances are very close in the graph variants of two sub-graph sparsity, where **w/o TSG sparsity** performs marginally better than others. In fact, due to the characteristics of object detectors and OCR systems, dense object-object relations with large overlapping regions widely exist, while overlapping OCR tokens are much less frequent. Anyway, **SSGN** with the complete sparsity setting

[3]Sparsity Ratio (%) = $Avg(\frac{N_p}{N_I})$, where N_p and N_I denote the number of pruned edges and the total number of edges per image respectively.

TABLE X

ABLATION STUDIES OF SOFT SOLUTION BASED ON GAT [49] TECHNIQUE WITH ROSETTA-EN OCR FEATURES ON TEXTVQA DATASET

Method	Val Acc	Test Acc
SSGN-GAT	39.86	39.59
SSGN-GAT-Soft Sparse (Hyperparameter)	40.01	39.98
SSGN-GAT-Soft Sparse (Median)	39.99	39.87
SSGN-GAT-Soft Sparse (Mean)	39.94	39.73
SSGN-GAT-Spatial Sparse	40.70	40.35
SSGN (Ours)	**42.00**	**41.60**

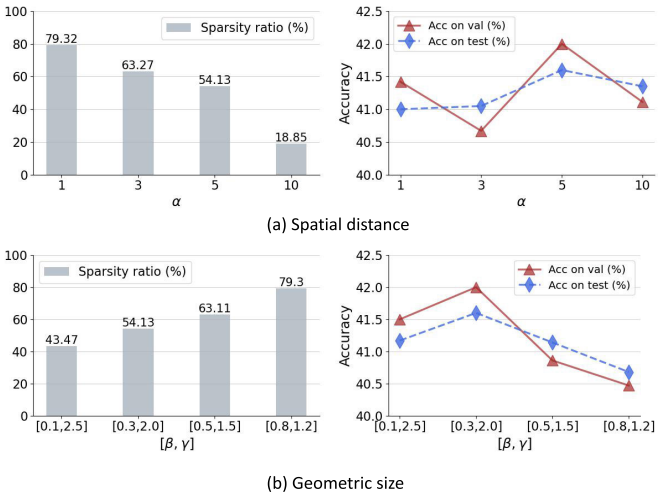

Fig. 4. Ablation studies of spatial factors in TSG.

achieves the best results. Furthermore, the sparsity ratios of OTSG, OSG, and TSG are reported in Table VI, around 10%, 15%, and 55%, respectively. These sparsity ratios are roughly stable, except for the case of 32.36% OTSG under SBD-Trans features on ST-VQA. We are surprised to find that this is actually caused by the SBD-Trans OCR system's mislabeling of object bounding boxes on the coco-text subset of the ST-VQA dataset. Among these, TSG has the greatest sparsity.

Back to the statistics of OCR tokens in Table I, various OCR systems output different but competitive token numbers with each other. Here, we discuss the effect of OCR systems on the sparsity rate in Table VII. We choose the *Mean* number of tokens per image as the cut-off point to observe the sparsity ratio of TSG, *i.e*, 20 for the TextVQA dataset and 10 for the ST-VQA dataset. As shown in Table VII, the TSG sparsity ranges from 55% to 66% for OCR$\leqslant$20/OCR$\leqslant$10 on the TextVQA / ST-VQA, while is about 45% for OCR$>$20/OCR$>$10, respectively. In conclusion, the sparsity patterns of the two datasets are similar and we consider the adaptive sparsity strategy in our approach to be stable and acceptable.

2) Impact of various IoUs: IoU is widely used for object detection. In this part, we discuss its effect in the OSG graph (Eq. 6). There are many variants of IoU, including IoU [11], GIoU [44], CIoU [59] and DIoU [59]. IoU [11] is a basic term that considers the overlap area of two object bounding boxes. With the basic IoU, GIoU (Generalized IoU) [44] considers

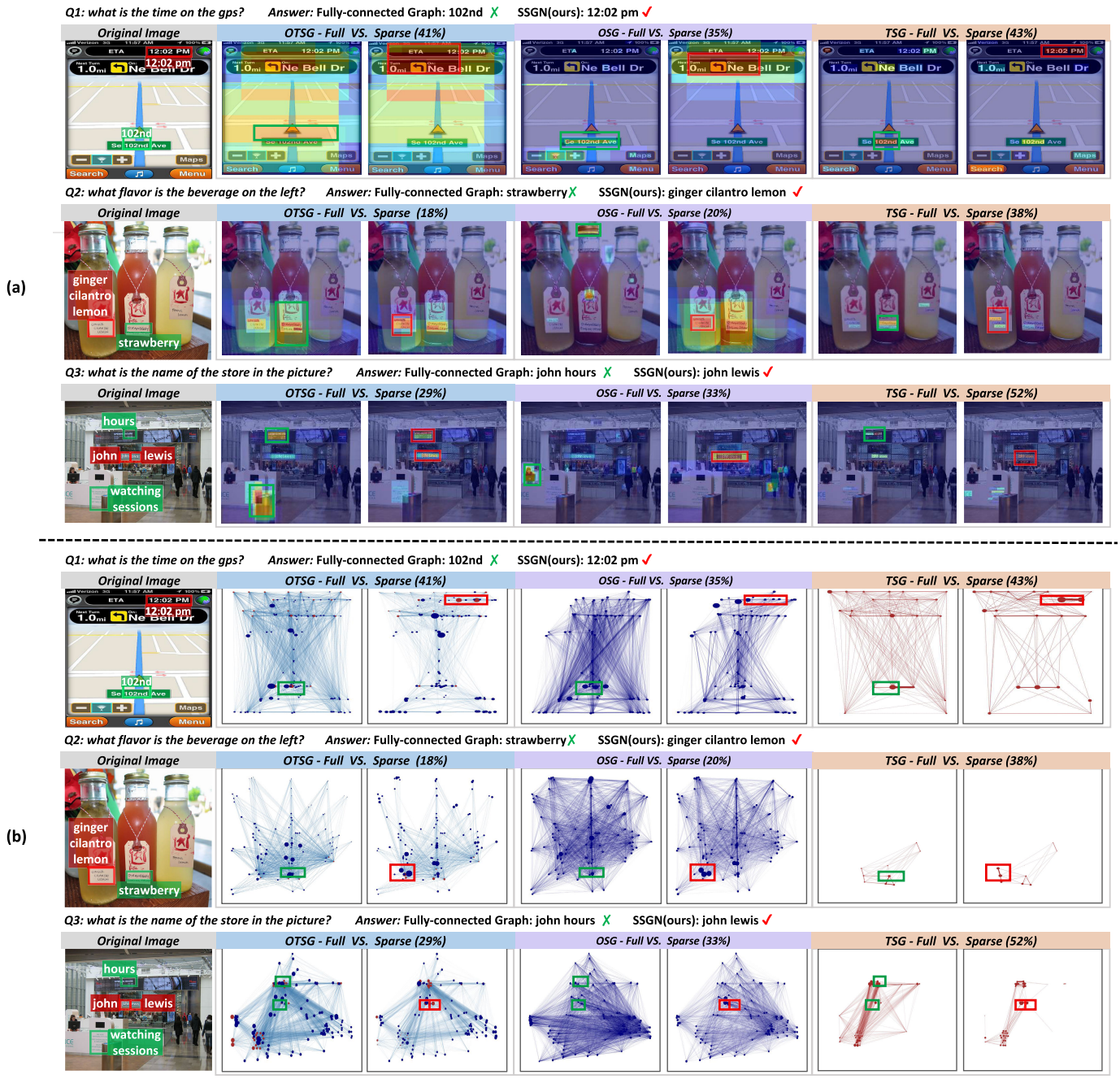

Fig. 5. Instantiation of fully connected and sparse graphs for OTSG, OSG, and TSG. The number in each parenthesis indicates the sparsity ratio. We show in (a) the visualization of nodes (objects and OCR tokens) overlapping in the image and in (b) the visualization of edges (relations). The node size reflects the sum of the connected edge weights and each edge weight is output by the message transition matrices $\mathcal{A}^{vt}$, $\mathcal{A}^{v}$, and $\mathcal{A}^{t}$ of each graph. By comparison, the sparse graph is more resistant to the interference of redundant relations and can generate accurate answers.

the relative direction, DIoU [59] (Distance-IoU) adds the center measurement, and CIoU (Complete-DIoU) [59] adds the length and width measurements of bounding box. As shown in Table VIII, **DIoU** considering overlap area and center distance performs the best at $\epsilon = 0.3$. Using only overlap area is not sufficient (**IoU** drops by 0.53%), and considering direction and angle is not appropriate for edge modeling in the TextVQA task (**GIoU** drops by 0.97 %). **CIoU** performs well (down 0.34%), but the length and width measurements are not effective to **DIoU** (considering center distance) in eliminating redundant spatial relations.

3) Impact of Spatial Factor: We analyze the role of spatial factors θ in graphs OTSG and OSG, and test $\{\alpha, \gamma\}$ in the graph TSG. In OTSG, θ is used to constrain the spatial distance or their overlap area of object-token pair, while θ is taken to constrain the spatial distance of object-object pair in OSG. As shown in Table IX, the greater the value of θ is, the fewer edges are pruned. The results show that the spatial distance is too close or too far to be suitable for relation inference. θ = 0.5 is the optimal setup for OTSG and OSG. We further test the spatial factors of distance and geometric size in the graph TSG. As shown in Fig. 4, the larger α is, the fewer

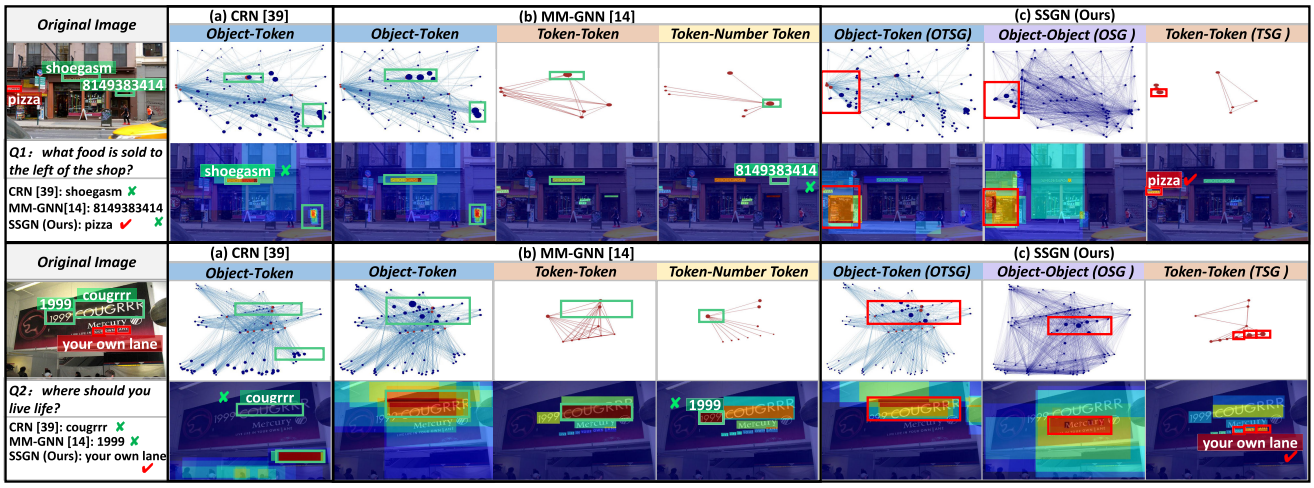

Fig. 6. Two visualization examples of our method compared with existing CRN [39] and MM-GNN [14]. The results show that through effective spatial pruning, our method SSGN performs accurate relation inference between object-token, object-object, and token-token for answer prediction.

edges are pruned. We set $\alpha = 5$ for the best performance. We consider that there is a balance between the sparsity ratio and distance. For the token's geometric size $[\beta, \gamma]$, the setting of $[0.3, 2.0]$ achieves the best performance. It seems that the broader ranges $[0.1, 2.5]$ and $[0.3, 2.0]$ filter out the relations useful for answer prediction better than $[0.5, 1.5]$ and $[0.8, 1.2]$.

4) Soft Relation Pruning: Here, we consider a flexible soft relational solution referring to GAT [49] technique. We replace the relational learning mode in the OTSG, OSG, and TSG graphs with soft weight learning in GAT [49]. As shown in Table X, the variants for ablation experiments are as follows: 1) **SSGN-GAT** is a fully-connected graph model that uses soft weights to combine full relations; 2) based on it, we test three soft pruning methods, *i.e*, **SSGN-GAT-Soft Sparse (Hyperparameter)**, **SSGN-GAT-Soft Sparse (Median)**, and **SSGN-GAT-Soft Sparse (Mean)**. The three sparse graph models respectively use an empirical *hyperparameter* (0.01), *median* and *mean* of edge weights to adaptively select sparse relations with different sparse criteria; 3) different from soft sparse solution, **SSGN-GAT-Spatial Sparse** is a sparse graph model that uses spatial conditions proposed in this work. Among these sparse graph variants, the baseline **SSGN-GAT** performs the worst. Compared with **SSGN-GAT**, the performances of all SSGN-GAT-Soft variants are slightly improved; for examples, **SSGN-GAT-Soft Sparse (Hyperparameter)** improves 0.15% and 0.39% on the val and test sets, respectively. Compared with **SSGN-GAT**, **SSGN-GAT-Spatial Sparse** with our spatial constraints increases 0.84% and 0.76% on the val and test sets, respectively. It proves that our spatial pruning method with sophisticated spatial conditions is effective and conducive to answer reasoning. Anyway, the proposed method **SSGN (Ours)** performs the best.

G. Visualization Analysis

To demonstrate the interpretability of our method, we visualize some examples below.

1) Graph Inference: As shown in Fig. 5 (a), we exhibit the graph learning process in the fully-connected graph and the sparse graph settings. The highly-responsive visual regions are quite different in the two graph settings. For the questions **Q1**∼**Q3**, in fully-connected graphs, the misleading answers are "102nd", "strawberry" and "john hour", while in sparse graphs the correct answers are "12:02 pm", "ginger cilantro lemon", and "john lewis". A remarkable observation is the existence of redundant relations in fully-connected graphs. For example in **Q1**, at the bottom of the image, the green road sign is recognized as a "sign" by the object detection model, which is covered by 11 different-sized bounding boxes that are considered as 11 objects. These close-by objects enhance the unnecessary visual relations between them. And the OCR token "Search" is far from the token "12:02 pm", while the relations between these two OCR tokens is no longer semantically needed. In the sparse graph, we cut off the redundant connections and make the inference of the answer more explicit. Fig. 5 gives an illustrative explanation of the effectiveness of sparse graph learning in this work.

2) Spatial Sparsity Analysis: As shown in Fig. 5 (b), in terms of the sparsity of OCR tokens, in **Q3**, the three tokens "john", "lewis" and "hours" are tightly connected in the fully-connected TSG graph. In fact, the two tokens "john" and "lewis" (the correct answer) are spatially close to each other but relatively distant from the token "hours". After relation pruning in the graph, the semantic difference between "hours" and "john lewis" is more explicit for the model. How about the other sparsity? Taking **Q1** "what is the time on the gaps" as an example, it is hard to distinguish between the numeric tokens "102nd" and "12:02 pm". In the fully-connected graph setting, the effect of "102nd" is enhanced by the redundant object-object and object-OCR token relations. After spatial pruning, "102nd" and "12:02 pm" can be evaluated fairly, and the correct answer "12:02 pm" is thus output.

3) Graph Model Comparison: Here, we display two examples to compare our model with two existing fully-connected graph methods **CRN** [39] and **MM-GNN** [14]. As shown in the Fig. 6, it can be found that both **CRN** [39] and

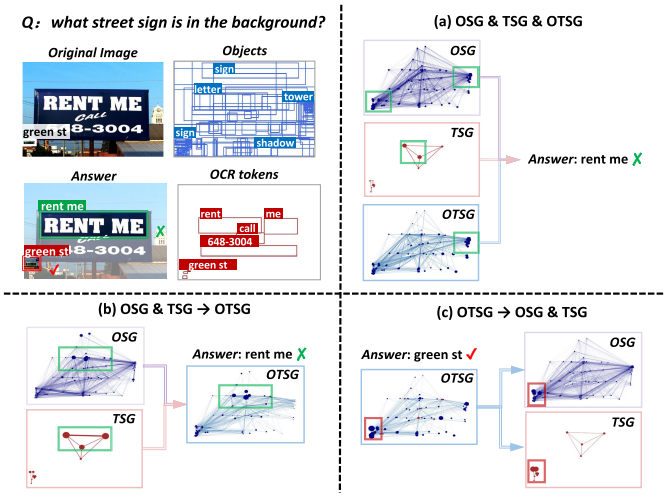

Fig. 7. Visualization of hierarchical graph structures for answer prediction. We discuss three solutions, namely OSG&TSG&OTSG, OSG&TSG→OTSG, and OTSG→OSG&TSG. By observing, in our method OTSG→OSG&TSG, the implementation of OTSG can first effectively discover the critical relations through the nearby object and OCR token co-occurrences, and then the parallel learning of OSG and TSG examines the critical relations under each space of objects and OCR tokens.

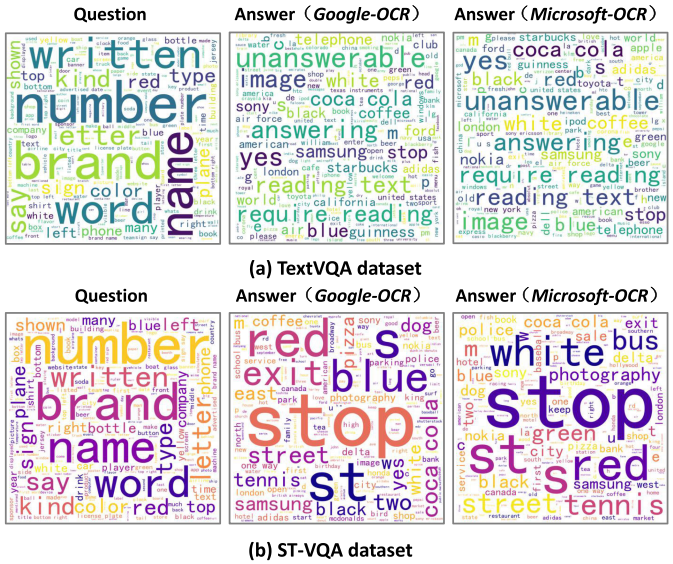

Fig. 8. Question and answer word clouds for TextVQA and ST-VQA datasets. We visualize the generated answers with Google-OCR and Microsoft-OCR as examples. For both questions and answers, stop words, *e.g*, "the", "is", "which", "what", "at", "on", *etc*, are removed from the statistics.

MM-GNN [14] are confused by the full relations and give wrong answers, while the proposed method performs well attributing to the sparse relation learning. As shown in **Q1 (a)** and **Q1 (b)**, the redundant relation between object "woman" and token "shoegasm" interferes with the answer reasoning of **CRN** [39] to the wrong answer "shoegasm", while the dense relations between token "shoegasm", digital token "8149383414" and the other tokens mislead the model **MM-GNN** [14], resulting the wrong answer "8149383414". In **Q1 (c)**, **SSGN (Our)** cuts off the relations between spatially distant object-token pairs and reduces the redundant associations between repeated objects and disconnected tokens by leveraging the customized spatial criteria. For example, in our TSG (token-token) graph, the token "pizza" has no connection to the tokens "shoegasm" and "8149383414". Finally, our model reasons out the correct answer "pizza". In addition, as shown in **Q2 (c)**, a similar conclusion can be drawn in **Q2 (c)**. The valuable and effective semantic associations are helpful for answer prediction rather than fully semantic associations established in the graphs.

4) Hierarchical Graph Structure: Here, we discuss the hierarchical graph structure for answer prediction. We visualize an example in Fig. 7. The proposed SSGN method is carried out with the parallel structure **OSG&TSG&OTSG**, as well as the cascading structure of the order **OSG&TSG→OTSG** and its reverse order **OTSG→OSG&TSG**, respectively. To answer the question "what street sign is in the background?" In **OSG&TSG&OTSG**, each sub-graph focuses on different entities in each visual space, such as object nodes "tower" and "sign" in OSG, OCR token nodes "rent me" and "call" in TSG and object node "tower" in OTSG. Unlike the discovery in **OSG&TSG&OTSG**, **OSG&TSG→OTSG** and **OTSG→OSG&TSG** consistently focus on "sign" in the image. But **OSG&TSG→OTSG** performs the graph learning in separate object and OCR token spaces in parallel first, resulting in a wrong focus on the foreground "sign". In contrast, our **OTSG→OSG&TSG** outputs the correct answer. By first implementing the object-OCR token correlation, our approach focuses directly on the background "sign".

5) Word Cloud Analysis: Here, we use the word clouds to visualize the high-frequency words in the questions and answers. As shown in Fig. 8, the questions in both TextVQA and ST-VQA datasets pay consistent attention to words "*name*", "*number*", "*word*", "*brand*", and "*written*". As for the answers, due to the difficulty of questions, "*unanswerable*" occurs with a remarkable frequency in TextVQA dataset. And the road sign word "*stop*" appears most frequently in the ST-VQA answers. It is also interesting to note that the ST-VQA prefers to ask questions about color accompanied with the answers "*blue*", "*red*", and "*white*".

V. Conclusion

In this paper, we propose a sparse spatial graph network (SSGN) for TextVQA, which focuses on edge pruning in graph learning. We investigate a depth study of graph sparsity from spatial factors, such as DIoU, distance, geometric size, and overlap area. We strive to prune redundant or useless relations. Extensive experiments are conducted on TextVQA and ST-VQA datasets under different OCR systems to validate the effectiveness of SSGN and to show interpretable visualization results.

References

[1] J. Almazán, A. Gordo, A. Fornés, and E. Valveny, "Word spotting and recognition with embedded attributes," *IEEE TPAMI*, vol. 36, no. 2, pp. 2552–2566, Dec. 2014.

[2] J. Baek et al., "What is wrong with scene text recognition model comparisons? Dataset and model analysis," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4714–4722.

[3] A. F. Biten et al., "Scene text visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4290–4300.

[4] A. F. Biten et al., "ICDAR 2019 competition on scene text visual question answering," in *Proc. Int. Conf. Document Anal. Recognit. (ICDAR)*, Sep. 2019, pp. 1563–1570.

[5] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *TACL*, vol. 5, pp. 135–146, Jun. 2017.

[6] F. Borisyuk, A. Gordo, and V. Sivakumar, "Rosetta: Large scale system for text detection and recognition in images," in *Proc. 24th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2018, pp. 71–79.

[7] F. Chen, X. Chen, F. Meng, P. Li, and J. Zhou, "GoG: Relation-aware graph-over-graph network for visual dialog," in *Proc. Findings Assoc. Comput. Linguistics ACL-IJCNLP*, 2021, pp. 230–243.

[8] Y. Chen, L. Tai, K. Sun, and M. Li, "MonoPair: Monocular 3D object detection using pairwise spatial relationships," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12090–12099.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[10] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL*, 2019, pp. 4171–4186.

[11] M. Everingham, L. V. Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL Visual Object Classes (VOC) challenge," *IJCV*, vol. 88, pp. 303–338, Sep. 2010.

[12] S. Fang, H. Xie, Y. Wang, Z. Mao, and Y. Zhang, "Read like humans: Autonomous, bidirectional and iterative language modeling for scene text recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 7094–7103.

[13] C. Gao et al., "Structured multimodal attentions for TextVQA," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 9603–9614, Dec. 2022.

[14] D. Gao, K. Li, R. Wang, S. Shan, and X. Chen, "Multi-modal graph neural network for joint reasoning on vision and scene text," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 12743–12753.

[15] M. Gu, Z. Zhao, W. Jin, R. Hong, and F. Wu, "Graph-based multi-interaction network for video question answering," *IEEE TIP*, vol. 30, pp. 2758–2770, 2021.

[16] D. Guo, K. Li, Z.-J. Zha, and M. Wang, "DADNet: Dilated-attention-deformable ConvNet for crowd counting," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1823–1832.

[17] D. Guo, S. Wang, Q. Tian, and M. Wang, "Dense temporal convolution network for sign language translation," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 744–750.

[18] D. Guo, W. Zhou, H. Li, and M. Wang, "Hierarchical LSTM for sign language translation," in *Proc. AAAI*, 2018, pp. 6845–6852.

[19] W. Guo, Y. Zhang, J. Yang, and X. Yuan, "Re-attention for visual question answering," *IEEE TIP*, vol. 30, pp. 6730–6743, 2021.

[20] D. Gurari et al., "VizWiz grand challenge: Answering visual questions from blind people," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3608–3617.

[21] W. Han, H. Huang, and T. Han, "Finding the evidence: Localization-aware answer prediction for text visual question answering," in *Proc. COLING*, 2020, pp. 3118–3131.

[22] R. Hu, A. Singh, T. Darrell, and M. Rohrbach, "Iterative answer prediction with pointer-augmented multimodal transformers for TextVQA," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9989–9999.

[23] Q. Huang et al., "Aligned dual channel graph convolutional network for visual question answering," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 7166–7176.

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

[25] S. Inayoshi, K. Otani, A. T.-D. Pablos, and T. Harada, "Bounding-box channels for visual relationship detection," in *Proc. ECCV*, 2020, pp. 682–697.

[26] J. Ji, C. Xu, X. Zhang, B. Wang, and X. Song, "Spatio-temporal memory attention for image captioning," *IEEE TIP*, vol. 29, pp. 7615–7628, 2020.

[27] Y. Jiang, V. Natarajan, X. Chen, M. Rohrbach, D. Batra, and D. Parikh, "Pythia v0.1: The winning entry to the VQA challenge 2018," 2018, *arXiv:1807.09956*.

[28] Z. Jin et al., "RUArt: A novel text-centered solution for text-based visual question answering," *IEEE TMM*, vol. 25, pp. 1–12, 2021.

[29] 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.

[30] Y. Kant et al., "Spatially aware multimodal transformers for TextVQA," in *Proc. ECCV*, 2020, pp. 715–732.

[31] D. Karatzas et al., "ICDAR 2015 competition on robust reading," in *Proc. 13th Int. Conf. Document Anal. Recognit. (ICDAR)*, Aug. 2015, pp. 1156–1160.

[32] D. Karatzas et al., "ICDAR 2013 robust reading competition," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 1484–1493.

[33] E.-S. Kim, W. Y. Kang, K.-W. On, Y.-J. Heo, and B.-T. Zhang, "Hypergraph attention networks for multimodal learning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 14569–14578.

[34] I. Krasin et al., (2017). *Openimages: A Public Dataset for Large-Scale Multi-Label and Multi-Class Image Classification*. [Online]. Available: https://github.com/openimages

[35] R. Krishna et al., "Visual Genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, vol. 123, pp. 32–73, Feb. 2017.

[36] V. I. Levenshtein et al., "Binary codes capable of correcting deletions, insertions, and reversals," *Sov. Phys. Doklady*, vol. 10, no. 8, pp. 707–710, 1966.

[37] L. Li, Z. Gan, Y. Cheng, and J. Liu, "Relation-aware graph attention network for visual question answering," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 10312–10321.

[38] Y. Liao, A. Zhang, Z. Chen, T. Hui, and S. Liu, "Progressive language-customized visual feature learning for one-stage visual grounding," *IEEE TIP*, vol. 31, pp. 4266–4277, 2022.

[39] F. Liu, G. Xu, Q. Wu, Q. Du, W. Jia, and M. Tan, "Cascade reasoning network for text-based visual question answering," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4060–4069.

[40] Y. Liu, S. Zhang, L. Jin, L. Xie, Y. Wu, and Z. Wang, "Omnidirectional scene text detection with sequential-free box discretization," in *Proc. 28th Int. Joint Conf. Artif. Intell.*, Aug. 2019, pp. 3052–3058.

[41] X. Lu, Z. Fan, Y. Wang, J. Oh, and C. P. Rosé, "Localize, group, and select: Boosting text-VQA by scene text modeling," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 2631–2639.

[42] A. Mishra, K. Alahari, and C. V. Jawahar, "Image retrieval using textual cues," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3040–3047.

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

[44] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: A metric and a loss for bounding box regression," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 658–666.

[45] O. Sidorov, R. Hu, M. Rohrbach, and A. Singh, "TextCaps: A dataset for image captioning with reading comprehension," in *Proc. ECCV*. Cham, Switzerland: Springer, 2020, pp. 742–758.

[46] A. Singh et al., "Towards VQA models that can read," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8309–8318.

[47] A. Vaswani et al., "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 5998–6008.

[48] A. Veit, T. Matera, L. Neumann, J. Matas, and S. Belongie, "COCO-Text: Dataset and benchmark for text detection and recognition in natural images," 2016, *arXiv:1601.07140*.

[49] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio, "Graph attention networks," in *Proc. ICLR*, 2018, pp. 1–12.

[50] J. Wang, J. Tang, and J. Luo, "Multimodal attention with image text spatial relationship for OCR-based image captioning," in *Proc. 28th ACM Int. Conf. Multimedia*, Oct. 2020, pp. 4337–4345.

[51] J. Wang, J. Tang, M. Yang, X. Bai, and J. Luo, "Improving OCR-based image captioning by incorporating geometrical relationship," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 1306–1315.

[52] X. Wang et al., "On the general value of evidence, and bilingual scene-text visual question answering," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 10123–10132.

[53] Y. Wang, J. Peng, H. Wang, and M. Wang, "Progressive learning with multi-scale attention network for cross-domain vehicle re-identification," *Sci. China Inf. Sci.*, vol. 65, Apr. 2022, Art. no. 160103.

[54] S. Yang, G. Li, and Y. Yu, "Relationship-embedded representation learning for grounding referring expressions," *IEEE TPAMI*, vol. 43, no. 8, pp. 2765–2779, Aug. 2021.
[55] Z. Yang et al., "TAP: Text-aware pre-training for text-VQA and text-caption," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8747–8757.
[56] T. Yao, Y. Pan, Y. Li, and T. Mei, "Exploring visual relationship for image captioning," in *Proc. ECCV*, 2018, pp. 684–699.
[57] G. Zeng, Y. Zhang, Y. Zhou, and X. Yang, "Beyond OCR + VQA: Involving OCR into the flow for robust and accurate TextVQA," in *Proc. 29th ACM Int. Conf. Multimedia*, Oct. 2021, pp. 376–385.
[58] X. Zhang and Q. Yang, "Position-augmented transformers with entity-aligned mesh for TextVQA," in *Proc. ACM MM*, 2021, pp. 2519–2528.
[59] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-IoU loss: Faster and better learning for bounding box regression," in *Proc. AAAI*, 2020, pp. 12993–13000.
[60] Q. Zhu, C. Gao, P. Wang, and Q. Wu, "Simple is not easy: A simple strong baseline for TextVQA and TextCaps," in *Proc. AAAI*, 2021, pp. 3608–3615.
[61] Z. Zhu, J. Yu, Y. Wang, Y. Sun, Y. Hu, and Q. Wu, "Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering," in *Proc. 29th Int. Joint Conf. Artif. Intell.*, Jul. 2020, pp. 1–9.

Sheng Zhou received the B.E. degree in the Internet of Things from Hengyang Normal University, China, in 2020. She is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Hefei University of Technology, China. Her research interests include computer vision, natural language processing, and multimodal machine learning.

Dan Guo (Member, IEEE) received the B.E. degree in computer science and technology from Yangtze University, China, in 2004, and the Ph.D. degree in system analysis and integration from the Huazhong University of Science and Technology, China, in 2010. She is currently an Associate Professor with the School of Computer Science and Information Engineering, Hefei University of Technology, China. Her research interests include computer vision, machine learning, and intelligent multimedia content analysis.

Jia Li received the B.E. degree in automation from the Hefei University of Technology (HFUT), Hefei, China, in 2016, and the Ph.D. degree from the School of Information Science and Technology, University of Science and Technology of China (USTC), Hefei, in 2021. He is currently a Lecturer with HFUT. His current research interests include computer vision and deep learning. He has published several papers in refereed journals and conferences, such as IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS, AAAI, and ACM MM.

Xun Yang received the Ph.D. degree from the Hefei University of Technology, Hefei, China, in 2017. He is currently a Professor with the Department of Electronic Engineering and Information Science, University of Science and Technology of China (USTC). From 2015 to 2017, he visited the University of Technology Sydney (UTS), Australia, as a joint Ph.D. student. He was a Research Fellow with the NExT++ Research Center, National University of Singapore (NUS), from 2018 to 2021. His current research interests include information retrieval, cross-media analysis and reasoning, and computer vision. He regularly serves as the PC member and an invited reviewer for top-tier conferences and prestigious journals in multimedia and artificial intelligence, such as the ACM Multimedia, IJCAI, AAAI, CVPR, and ICCV. He served as the Area Chair for the ACM Multimedia 2022. He also serves as the Associate Editor for the IEEE TRANSACTIONS ON BIG DATA journal.

Meng Wang (Fellow, IEEE) received the B.E. and Ph.D. degrees in the special class for the gifted young from the Department of Electronic Engineering and Information Science, University of Science and Technology of China, Hefei, China, in 2003 and 2008, respectively. He is currently a Professor with the Hefei University of Technology, China. He has authored over 200 book chapters, journals, and conference papers in his research areas. His current research interests include multimedia content analysis, computer vision, and pattern recognition. He was a recipient of the ACM SIGMM Rising Star Award 2014. He is an Associate Editor of the IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, the IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, and the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS.