Applied Intelligence (2023) 53:16706–16723
https://doi.org/10.1007/s10489-022-04355-w

# Local self-attention in transformer for visual question answering

Xiang Shen[1] · Dezhi Han[1] · Zihan Guo[1] · Chongqing Chen[1] · Jie Hua[2] · Gaofeng Luo[3]

Accepted: 19 November 2022 / Published online: 15 December 2022

## Abstract

Visual Question Answering (VQA) is a multimodal task that requires models to understand both textual and visual information. Various VQA models have applied the Transformer structure due to its excellent ability to model self-attention global dependencies. However, balancing global and local dependency modeling in traditional Transformer structures is an ongoing issue. A Transformer-based VQA model that only models global dependencies cannot effectively capture image context information. Thus, this paper proposes a novel Local Self-Attention in Transformer (LSAT) for a visual question answering model to address these issues. The LSAT model simultaneously models intra-window and inter-window attention by setting local windows for visual features. Therefore, the LSAT model can effectively avoid redundant information in global self-attention while capturing rich contextual information. This paper uses grid visual features to conduct extensive experiments and ablation studies on the VQA benchmark datasets VQA 2.0 and CLEVR. The experimental results show that the LSAT model outperforms the benchmark model in all indicators when the appropriate local window size is selected. Specifically, the best test results of LSAT using grid visual features on the VQA 2.0 and CLEVR datasets were **71.94**% and **98.72**%, respectively. Experimental results and ablation studies demonstrate that the proposed method has good performance. Source code is available at https://github.com/shenxiang-vqa/LSAT.

**Keywords** Transformer · Local self-attention · Grid/regional visual features · Visual question answering

## 1 Introduction

Transformer [1] has achieved state-of-the-art results on a wide range of natural language processing (NLP) tasks. Many researchers have also successfully applied it to vision and language (V&L) tasks (*e.g.*, visual question answering [2, 3], cross-modality information retrieval [4, 5]), etc. Many researchers have proposed various multi-modal networks based on Transformer, achieving state-of-the-art performance on various benchmark datasets. Much of the success is attributed to the global dependency modeling capability of the self-attention component, which enables the network to capture contextual information from the entire sequence of inputs while to a certain extent facilitating modal alignment in language and vision.

Inspired by the excellent performance of Transformer in natural language processing tasks, Yu et al. [2] used the pure Transformer architecture for VQA tasks, and the experimental results showed that the Transformer architecture also has good performance in visual tasks. In multimodal learning tasks such as visual reasoning (*e.g.*, semantic sementation [6, 7], image understanding [8]) models need to process visual information from different receptive fields. In the VQA task, to correctly answer the questions input by the user, the model should pay attention to the global visual information when understanding image features. At the same time, it should pay attention to the local visual information of the image according to the semantic features of the question and use the local information interaction to capture richer contextual image feature information.

Some studies [3, 9, 32, 34] have demonstrated that it is difficult to achieve satisfactory performance using traditional Transformer models that only model global self-attention in VQA tasks. In addition, the relatively expensive

✉ Dezhi Han
dzhan@shmtu.edu.cn

Xiang Shen
202240310002@stu.shmtu.edu

[1] College of Information Engineering, Shanghai Maritime University, 1550 Haigang Avenue, Shanghai, 201306, China

[2] TD School, University of Technology, Sydney, Ultimo, NSW, 2007, Australia

[3] College of Information Engineering, Shaoyang University, Shaoyang, 422099, China

computational cost of modeling global attention makes it challenging when being flexibly applied to various VQA tasks, especially in high-resolution image scenarios. The self-attention (SA) unit is one of the critical components of Transformer, which is used to compute self-attention scores for the input sequence. These attention scores are collected in a global attention map containing the correlation of each input element in the entire sequence and are normalized. Therefore, these attention maps describe strong global dependencies across the sequence. The problem where global self-attention can't adequately model image context information is more prominent in end-to-end visual, multimodal reasoning models. Related research [10, 11] shows that well-pretraining grid visual features have better expression ability. However, grid visual features contain more fragmented semantic information than the widely-used regional visual features [12]. As shown in Fig. 1, the global attention of the traditional SA unit while modeling the input image is more likely to introduce noise information. The model pays attention to the image regions that are not relevant while attempting to correctly answer the question and this ultimately negatively affects the model's performance. However, utilizing the local self-attention mechanism improves attention to critical areas of the image according to question features. It can better utilize the visual information around important image areas to oblige the model's reasoning and prediction. Advanced VQA models are gradually replacing global self-attention with local self-attention. A key and challenging problem is enhancing the modeling ability of local self-attention-based models. For example, Swin [13] and Shuffle Transformer [14] proposed shift windows and shuffle windows, respectively, and alternately used two different window partitions in consecutive blocks (*e.g.*, rule windows and proposal windows) to build cross-window connections. The MSG Transformer [15] manipulates messenger tokens to exchange information between windows. Wang et al. [17] proposed axial self-attention to treat local attention regions as a single row or as a column of feature maps. CSWin [18] proposed cross-shaped window self-attention, which can be considered a multi-row and multi-column expansion of axial self-attention. While these methods perform well and even outperform their CNN counterparts, the underlying self-attention and dependencies of Transformer are not rich enough to capture rich contextual information and pass it to the top layers for effective modeling.

We propose a novel local self-attention scheme to address the above-mentioned problems. The scheme, Local Self-Attention in Transformer (LSAT), models local self-attention at the bottom layer while still modeling global attention at the upper layer to integrate the findings of the previous layers. LSAT provides a solid baseline for image self-attention modeling by setting local windows for visual features to simultaneously model intra-window attention and inter-window attention to capture rich contextual visual information features. In order to verify the effectiveness and practicability of the LSAT model, this paper utilizes regional and grid visual features extracted from region proposals to conduct experiments on the VQA benchmark datasets. The experimental results show that the LSAT model has good performance. The main contributions of this paper are as follows:

(1) Proposal of a novel local self-attention mechanism, which can more effectively capture rich image context-dependent features. The LSAT model can also represent the interaction between windows and can model image self-attention learning by establishing an adjustable window size. Moreover, the LSAT model overcomes the shortcomings of global self-attention (such as high computational overhead and lack of interaction between regions), which is critical for designing end-to-end vision tasks based on Transformer.

(2) This paper also employs regional and grid visual features to conduct experiments to verify the effectiveness of the local self-attention mechanism. We find that using a local attention mechanism on-grid visual

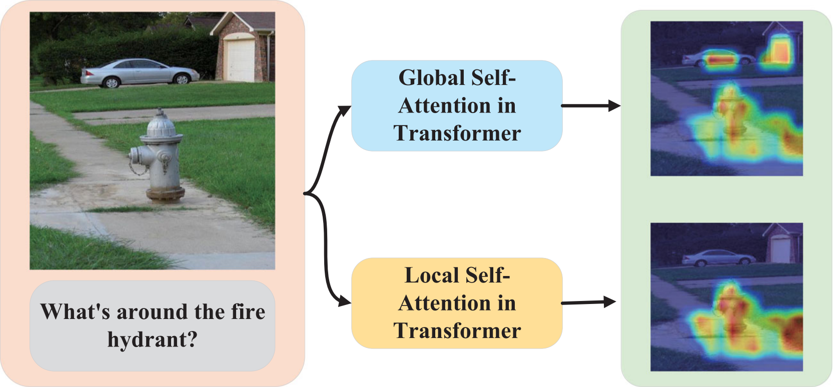

What's around the fire hydrant?
Global Self-Attention in Transformer
Local Self-Attention in Transformer

**Fig. 1** Transformer attention results on the same image using global self-attention and local self-attention mechanism respectively

features can more effectively reduce the introduction of irrelevant information and effectively capture contextual information compared with regional visual features.

(3) Experimental results on benchmark VQA 2.0 and CLEVR datasets show that the LSTA model achieves better results than current state-of-the-art models. We analyzed the effect of parameter settings on the LSAT model through ablation experiments and further reveal the interpretability of the model employing visual examples.

The rest of the paper is organized as follows: Section 2 introduces the visual question answering model and local self-attention mechanism related to the research in this paper, and Section 3 details the overall structure and specific design details of the LSAT model. Section 4 introduces the experimental results and ablation studies using regional and grid visual features, respectively, based on the VQA 2.0 and CLEVR benchmark datasets under different parameter settings and visualizes the model through attention. Section 5 summarizes our work and points out future research directions.

## 2 Related work

### 2.1 Visual question answering

The Visual Question Answering (VQA) task has been gained increased attention in the past few years. VQA takes images and natural language questions about images as input and generates natural language answers as output is usually considered a classification task with fixed categories [19]. Traditional VQA models represent input images and questions as global features. However, global features introduce considerable noise that affects the model's ability to predict the correct answer. Therefore, most advanced VQA models utilize an attention mechanism to reduce irrelevant information. For example, many studies [20, 22, 25] employ CNN-based network structures to model various visual attention mechanisms to localize image regions relevant to the input question. Other related work [22, 26] propose methods that combine both question-guided visual attention and image-guided question attention. This is also vital in the multimodal fusion method in the VQA model. The traditional multimodal fusion method maps the image and question features into a common space and uses concatenation, addition, and other methods for simple fusion. Recent works [27–29] has explored a more complex and advanced (effective) multimodal fusion methods.

With the development of the Transformer network [1] and ViT [30, 31] network structure, Yu et al. applied the Transformer structure to the VQA task for the first time and proposed the deep modular co-attention network (MCAN) [2]. Later, researchers proposed various variant structures based on the MCAN network [3, 32–34]. Recent advanced VQA methods concentrate on stacked multi-layer attention and improvements in fusion mechanisms. Yang et al. [35] proposed a stacking attention mechanism (SAN) that employs the semantic representation of the question as a query to search for regions in the image relevant to the question's answer. Lu et al. [21] proposed a hierarchical question-image co-attention mechanism (HieCoAtt), which alternately learns visual and question attention. Okatani et al. [36] proposed a dense symmetric co-attention (DSCA) model, which exploits the co-attention mechanism of densely stacked layers. Kim et al. [28] proposed a bilinear attention network (BAN) to improve model performance by establishing associations between regions and images. Gao et al. [37] proposed dynamic fusion with an intra- and inter-modality attention flow (DFAF) model that explores information interactions within and between modalities. Although these methods to a certain extent improve the performance of VQA models, they are not effective in exploring the intrinsic dependencies between questions and images, resulting in suboptimal model performance. With the success of unsupervised pre-training in NLP, the development of Transformer and its variants in VQA tasks has been further promoted, making pre-training a new trend [38, 39]. Researchers realize that effectively utilizing question and image features in VQA tasks is crucial to predicting the correct answer. Hence, the latest VQA methods concentrate on improving attention mechanisms and fusion methods. For example, Zhao et al. [10] proposed the TRAnsformer Routing (TRAR) model, which uses grid image features that contain more visual semantic information. The fragmentation of grid visual features will lead to more noise in the visual information modeling. Avoiding this fragmented information in the model during the modeling process is very important for the reasoning and prediction of the model. Employing the local window mechanism in the LSAT model can effectively avoid introduction of fragmented information in the image modeling process. Using the local information interaction between windows in the end-to-end training process can improve the prediction probability of visual reasoning tasks.

### 2.2 Local self-attention mechanism

In recent years, local self-attention mechanisms have been widely used in computer vision. Unlike CNN, traditional Transformers do not involve inductive biases on local connections, which may lead to insufficient extraction of local features, such as connections of lines, edges, and colors of images. The self-attention mechanism was initially proposed in machine translation methods for NLP tasks. It belongs to a branch of attention mechanism, which reduces the dependence on external information and better captures the internal correlation of data or features. Therefore, the self-attention mechanism plays an increasingly significant role in computer vision and is widely used in many visual tasks, including object detection [40, 41] and image classification [42, 43]. The earliest approach was to replace the single-scale structure of ViT with a hierarchical architecture to obtain multi-scale features [44]. For example, Parmar et al. [45] applied a local attention mechanism to the visual part of the Transformer to focus on the domain information of the local window. Watzel et al. [46] proposed a dynamic fusion method between global and local attention scores based on Gaussian masks. The small networks for learning the fusion process and the Gaussian masks require only a few additional parameters and are simple to add to current transformer architectures. Swin [13] and Shuffle Transformer [14] propose shifted and shuffled windows, respectively, and alternately employ two different window partitions (the regular window and the regional window) to build cross-window connections in consecutive blocks. MSG Transformer [15] can flexibly exchange visual information across regions and reduce computational complexity by manipulating messenger tokens to exchange information intra-windows. Wu et al. [47] proposed Pale-Shaped Attention (PS-Attention) to parallelize row attention and column attention of attention map, which captures richer information while keeping the computational complexity similar to previous local attention mechanisms utilizing context information. Zhou et al. [16] proposed a multi-scale deep contextual convolutional network, which can fully utilize the local and global contextual information of the entire scene to each pixel, thereby enriching the semantic information of the image. Wang et al. [17] proposed axial self-attention to treat local regions as a single row or column of feature maps. CSWin Transformer [18] proposed cross-window self-attention, considered a multi-row and multi-column expansion of axial self-attention. Guo et al. [3] proposed multi-modal explicit sparse attention networks (MESAN) to efficiently filter features on feature maps using a ranking and selection method. Although these methods perform well, the dependencies of the modeled self-attention layers and the information interaction between windows are still insufficient.The LSAT model utilizes a novel image local modeling approach, significantly different from previously studied local self-attention mechanisms. We design local windows to model image self-attention to help locate critical regions of the image while designing local windows to interactively capture rich image feature information (relational properties of different locations and objects).

## 3 Method

The LSAT model first introduces the extraction and encoding of image and question features, secondly introduces the mechanism of local self-attention in the decoder in detail, and finally describes modality fusion and answer prediction. The overall architecture of LSAT is shown in Fig. 2.

### 3.1 Feature extraction of question and image

**Grid/Region image feature** We use grid and regional visual features in image feature extraction. In the VQA task, most existing methods use regional visual features. However, compared with regional visual features, grid visual features contain more image information, so this paper uses grid or regional image features to verify the effectiveness of

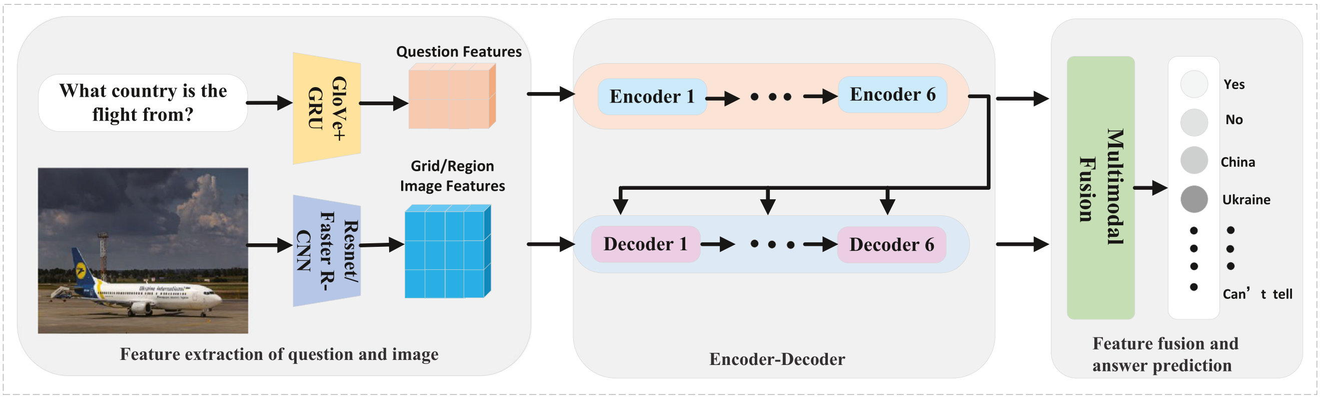


Fig. 2 Overall flowchart of the Local Self-Attention in Transformer (LSAT). Image feature extraction uses regional image features or grid image features

the LSAT model. The grid visual features are extracted employing the visual backbone network ResNext152 [11] pre-trained on the Visual Genome dataset [48]. When extracting grid features, the input image is first padded to 16×16. Then the input image features are pooled using a method with a convolution kernel size of 2×2 and a stride of 2, and the final resolution of grid visual features is 8×8. Extraction of visual region features is a bottom-up approach [12]. We use Faster R-CNN (with ResNet-101 as its backbone) pre-training from the Visual Genome dataset to extract regional image features. A confidence threshold is set for the anchor box of the detected object selection probability, and a dynamic number of visual objects $m_R \in [10, 100]$ can be obtained. The $i$-th regional visual feature of the input image is $x_i^r \in \mathbb{R}^{2048}$, and the available feature matrix after convolution pooling all regional visual features is $X^R \in \mathbb{R}^{m \times 2048}$.

**Question feature** For each input question, we first tokenize and crop each question to a maximum of 14 words and pad with 0 if there are less than 14 words when the model calculates. Then using the 300-D GloVe word embedding algorithm pre-trained on a large-scale corpus, each word in the question is further converted into a vector to obtain a word sequence of size $n \times 300$, where $n \times 300$ is the number of words in the question. Finally, the word vector is input into the GRU network to obtain the final question feature matrix $Y \in \mathbb{R}^{n \times 512}$.

## 3.2 Enocder and decoder

The LSAT model uses the encoder and decoder structure to model the question and image features. This section mainly introduces the decoder's principle of using the local self-attention mechanism. As shown in Fig. 3, considering the small number of input question words, it is the same as related research [2, 10]. We model the question with a global self-attention mechanism in the encoder and model the image with a local self-attention mechanism in the decoder to capture richer visual information.

### 3.2.1 Encoder

As shown on the left of Fig. 3, LSAT uses proportional dot-product attention in the encoder for self-attention learning to learn fine-grained question features. The encoder consists of $N$ stacked self-attention (SA) units, each with two sub-layers. The first sublayer is a multi-head attention layer, and the second sublayer is a fully connected feedforward layer. The first SA unit takes the question feature $Y = \{y_1, y_2, \cdots y_n\} \in \mathbb{R}^{n \times d_v}$ as input, and its multi-head attention layer learns the correlation between each word pair $< y_i, y_j >$. The feedforward layer of the SA unit further transforms the output of the previous sub-layer through two fully connected layers ReLU and Dropout. The output of the previous SA unit is used as the input of the next SA unit, and the output of the encoder is concerned with the question feature is represented by $F_1$, and the specific equation is as follows:

$$Q_E = YW_Q, K_E = YW_K, V_E = YW_V, \tag{1}$$

$$\begin{aligned} F_1 &= MHA\left(Q_E, K_E, V_E\right) \\ &= concat\left(head_1, head_2, \cdots, head_h\right) W^O, \end{aligned} \tag{2}$$

$$head_i = Att\left(Q_E W_i^{Q_e}, K_E W_i^{K_e}, V_E W_i^{V_e}\right), \tag{3}$$

$$S_i = soft\max\left(\frac{\left(Q_E W_i^{Q_e}\right) \cdot \left(K_E W_i^{K_e}\right)^T}{\sqrt{d_h}}\right), \tag{4}$$

$$head_i = S_i\left(V_E W_i^{V_e}\right), \tag{5}$$

where $W_i^{Q_e}, W_i^{K_e}$ and $W_i^{V_e}$ are learnable parameter matrices, and $concat\left(\cdot\right)$ represents connecting all heads. In order to facilitate the calculation, $d_h = d_v/h$ is usually set, and the *soft*max function is used for normalization. The question features obtained by the absolute encoder can be defined as:

$$FFN\left(F_1\right) = \max\left(0, F_1 W_1 + b_1\right) W_2 + b_2, \tag{6}$$

where $W_i$ and $b_i$ represent weight coefficients and biased variable respectively.

### 3.2.2 Decoder

This section details the principle of the local self-attention mechanism in the decoder. The decoder in Fig. 3 is composed of LSA and LSGA units, respectively. The LSA unit is the critical component of the LSAT model. The LSA unit first employs a local window to realize effective modeling of image self-attention learning for the input image. The LSA unit can obtain the image feature information with rich semantics. The LSGA unit achieves proper attention to key regions of the image according to the semantic features of the question input by the encoder. Figure 4 compares the global self-attention and local self-attention mechanisms, respectively. The standard decoder uses the global attention mechanism to model the input image features. The decoder mainly models image features rich in input semantic information. The disadvantage of global attention is that it is challenging to realize the attention of the local domain information of the image, and the local domain information is often the key to answering the input question correctly. Furthermore, modeling image features with global self-attention introduces much-fragmented information.

**Fig. 3** Schematic diagram of implementation of encoder and decoder architecture. $X^G$, $X^R$ and $Y$ represent visual features and question features, respectively. SA represents a Self-Attention unit in the encoder. LSA and LSGA represent Local Self-Attention and Local Self-Attention Guided unit respectively in the decoder. Add & Norm represent addition and layer normalization, respectively, and $FFN$ is a feedforward network

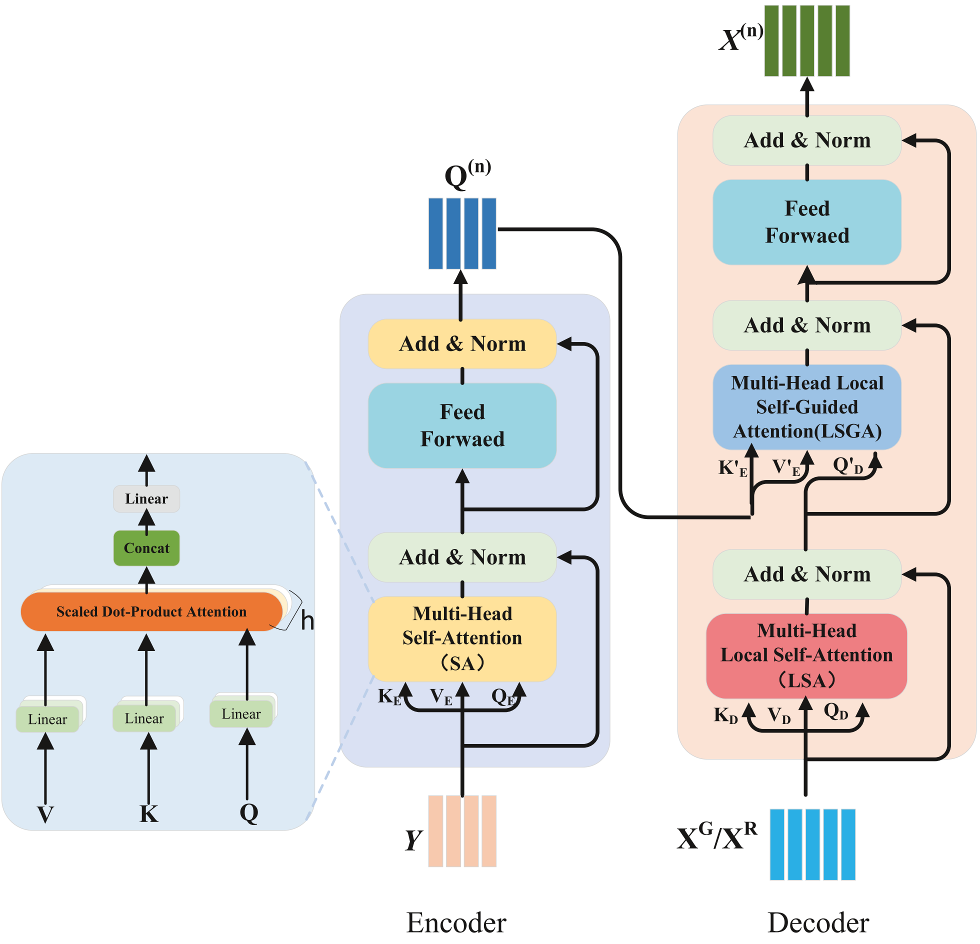
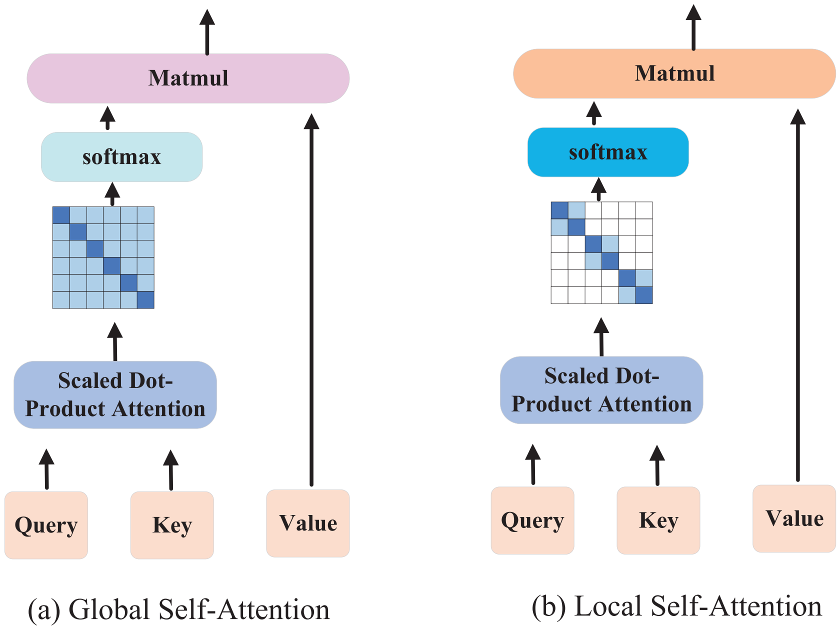


**Fig. 4** (a) Global self-attention mechanism in decoder. (b) Local self-attention mechanism in decoder


Matmul
softmax
Scaled Dot-Product Attention
Query
Key
Value
(a) Global Self-Attention
Matmul
softmax
Scaled Dot-Product Attention
Query
Key
Value
(b) Local Self-Attention

Therefore, the decoder in the LSAT model utilizes local self-attention to achieve interactive modeling learning within and between windows. Specifically, the local self-attention mechanism divides a global window of image feature size $t$ into $m$ local windows, where each image feature block contains $t/m$ local image features. We propose two forms to model local self-attention mechanisms. As shown in Fig. 5 (b), the image is modeled employing local pixel feature blocks instead of global image attention features. The local self-attention mechanism computes the interactions between all image features within each feature block when modeling visual attention within each local image feature. The local self-attention mechanism calculates the interaction between each feature block and all image features within adjacent feature blocks when modeling visual attention between feature blocks. The information of local pixels can be used to focus on the domain information, thereby capturing richer visual context information.

The decoder part of the Transformer model is shown on the right side of Fig. 3. The decoder consists of $N$ stacked identical LSGA (local self-guided attention) units. The decoder consists of $N$ stacked identical Local Self-Attention (LSA) and Local Self & Guided-Attention (LSGA) units. The input region or grid visual features are denoted as $X^R = \left\{x_1^r, x_2^r, \cdots, x_{m_R}^r\right\} \in \mathbb{R}^{m_R \times d_v}$ or $X^G = \left\{x_1^g, x_2^g, \cdots, x_{m_g}^r\right\} \in \mathbb{R}^{m_g \times d_k}$, respectively. When modelling image features, self-attention to local image features is achieved using a proportional dot-product attention unit. We take the grid visual feature as an example, which is expressed explicitly as:

$$
\begin{aligned}
Q_D &= X^G W_Q, K_D = X^G W_K, V_D = X^G W_V, && (7)\\
A_i &= \left( soft\max \left( \frac{Q_D K_D^T}{\sqrt{d_h}} \right) V_D \right), && (8)\\
H_i &= LA\,(A, w, f, b)\,, && (9)
\end{aligned}
$$

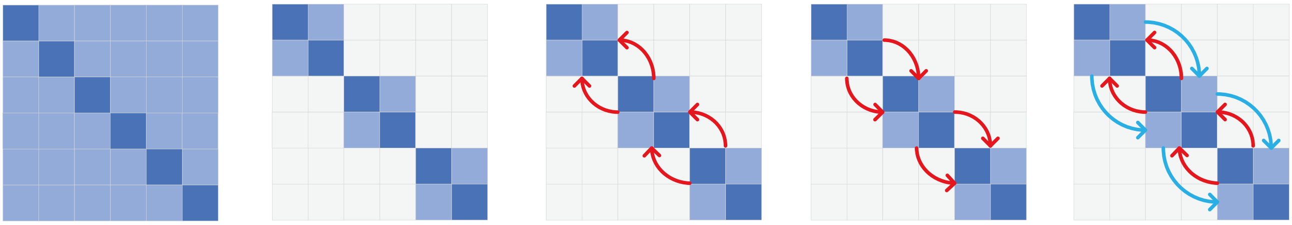
where $w$=$t/m$ represents the local window size, $f$ represents backward attention, and $b$ represents forward attention. For example, when $f$=2 is set, it means to pay attention to the number of front/back adjacent windows. Considering the computational cost, we set $f$ and $b$ to be 0 or 1, respectively, in our experiments. If $f$=1 is set, the partial window focuses on the previous window and interacts with it. In the same way, if $b$=1 is set, it means that the local window pays attention to a window behind and interacts with it. If $f$=$b$=1 is set at the same time, it means that the windows can follow each other and learn interactively.

As shown in Fig. 5, global self-attention and local window attention under different settings are compared, respectively. Among them, Fig. 5(a) represents the global self-attention mechanism of traditional methods in modelling visual feature self-attention, calculating the interaction between all image features. Figure 5(b) shows that the local self-attention only pays attention to the local window and does not interact with other windows for modelling learning. Figure 5(c) shows that interactions with subsequent adjacent windows are also calculated in addition to the current window's internal self-attention. Similarly, Fig. 5(d) shows that in addition to the self-attention of the current window, the interaction with the previous adjacent window is also calculated. Figure 5(e) shows that in addition to the current window self-attention, the interaction modelling of the current window's adjacent front and back windows is also calculated.

The LSGA unit obtains the question features through the encoder to guide the decoder modeling to focus on key image features. In order to improve and improve the expressiveness of image features, the model can obtain feature representations in different locations and subspaces. In the decoder, $h$ parallel attention heads are employed, and each head scales the attention individually using the dot-product attention (as in (4)). Finally, all the heads are connected to get the image feature representation:

$$
\begin{aligned}
F_2 &= concat\,(H_1, H_2, \cdots, H_h)\, W^O, && (10)\\
FFN\,(F_2) &= \max\,(0, F_2 W_3 + b_3) + b_4, && (11)
\end{aligned}
$$

where $W_i$ and $b_i$ denote the learnable parameter matrix and bias weights, $H_i$ denotes the $i$-$th$ local self-attention image feature, and $concat(.)$ denotes concatenating all attention



(a) Global self-attention (b) Local self-attention (0-0) (c) Local self-attention (0-1) (d) Local self-attention (1-0) (e) Local self-attention (1-1)

**Fig. 5** Comparing the global attention pattern and the configuration of local self-attention patterns in our LSAT

heads. The image features obtained by the multi-head self-attention mechanism and the original input grid image features are added to obtain $Q'_D$:

$$Q'_D = LN\left(X^G + FFN\left(F_2\right)\right), \tag{12}$$

$$\begin{aligned} F_3 &= MHAtt\left(Q'_D, K'_E, V'_E\right) \\ &= concat\left(head_1, head_2, \cdots, head_h\right) W^O, \end{aligned} \tag{13}$$

$$head_i = Att\left(Q'_D W_i^{Q_d}, K'_E W_i^{K_e}, V'_E W_i^{V_e}\right), \tag{14}$$

where $W_i^{Q_d}$, $W_i^{K_e}$ and $W_i^{V_e}$ represent learnable parameter matrices, and $LN(\cdot)$ represents layer normalization. The definition of the $Att(\cdot)$ function is the same as that of (4).

### 3.3 Feature fusion and answer prediction

The image and question features output by the encoder and decoder are fused to predict the correct answer. Following the method in [2], we use a two-layer $MLP$ (*e.g.*, $FC(512)$-$ReLU$-$Dropout(0.1)$-$FC(1)$). The network structure compresses two different features to the same dimension and fuses question $Y$ with regional visual feature $X^R$ or grid visual feature $X^G$. In order to more succinctly present this process, a visual grid feature is provided as an example. Specifically, input question feature $Y$ is input into the $MLP$ network layer to align the language and visual modalities. Then the softmax function calculates the attention weight value, and finally, multiplies and adds each corresponding attention weight to get the final question feature. Similarly, the fusion of question features and regional visual features is similar. This is shown in (15) and (16):

$$\lambda = soft\max\left(MLP\left(Y^{(l)}\right)\right), \tag{15}$$

$$\bar{Y} = \sum_{j=1}^{n} \lambda_i y_j^{(l)}, \tag{16}$$

Where $\lambda = [\lambda_1, \lambda_2, \cdots, \lambda_n] \in R^n$ is the question of learning weight, the visual grid and area features can be obtained as $\bar{X}^G$ and $\bar{X}^R$, respectively. Finally, the fusion feature can be obtained by the linear fusion method, and the specific expression is shown in (17):

$$F_G = layerNorm\left(W_{xg}^T \bar{X}^G + W_y^T \bar{Y}\right), \tag{17}$$

$F_G$ denotes the fusion feature of visual grid features, question features, $W_{xg}^T$ and $W_y^T$ the learning parameter matrix. Finally, $F_G$ is classified and predicted by the nonlinear function ReLU and sigmoid function, and binary cross-entropy (BCE) [25] is used as a loss function during training.

## 4 Experiments and discussion

### 4.1 Datasets

**VQA 2.0** The LSAT model is trained, validated, and tested on the VQA 2.0 [56] dataset, which is based on Microsoft COCO image data and is currently the most commonly utilized large-scale dataset for evaluating the performance of VQA models. The model- tries to minimize the effectiveness of the model learning dataset bias by balancing the answers to each question. The VQA 2.0 dataset contains 1.1M questions posed by humans. It consists of three parts: training set, validation set, and testing set. Each valid piece of data is represented by a three-element question and answer group composed of the dataset (image, question, answer). The training set contains 82,783 images and 443,757 question and answer groups corresponding to the images. The verification set includes 40,504 images and a corresponding 214,354 question and answer groups, and the testing set contains 81,434 images and 447,793 questions and answers groups. According to the categories of answers, questions can be divided into three types: yes/no (Yes/No), count (Number), and Other. We show the results on test-dev and test-standard on the VQA evaluation server. To be consistent with 'human accuracy', the accuracy matric is $\min\left(\frac{\#humans\ that\ provided\ that\ answer}{3}, 1\right)$, showing that an answer is regarded as 100% accurate if at least three annotations exactly match the predicted answer.

**CLEVR** CLEVR [24] tests the reasoning ability of visual diagnostic datasets, including counting, comparing, logical reasoning, and storing information in memory. It consists of 100,000 3D images of random shapes, sizes, materials, colors, and in rendered images in the dataset (70,000 for training, 15,000 for validation, and 15,000 for testing). The dataset contains nearly one million natural language questions, and 853,554 unique questions. Questions can be grouped into five general types: Exist, Count, Compare Numbers, Query Attributes, and Compare Attributes. There are 699,989 training questions, 149,991 validation, and 149,988 testing questions. The size of the vocabulary questions and answers is 82 and 28, respectively.

### 4.2 Experimental setup

All experiments in this paper are based on Linux Ubuntu system, the GPU is NVIDIA TITAN V 12GB, the deep learning framework is Pytorch, and the CUDA version is 10.0. For a fair comparison, LSAT follows most of the parameter settings in MCAN and TRAR. The question encoder adopts GRU, the dimension is 512, and the input question word is initialized by GLOVE embedding, which size is set to 300. The encoder and decoder layers in Transformer are both $N = 6$, used for the question and visual modeling, respectively, and the hidden dimension size is 1024. To effectively illustrate our method's effectiveness, we conducted experiments with regional visual features and grid visual features, respectively. The question words and image features are insufficient to fill 14 and 100 with 0 padding. The number of regional visual features $m_R$ is set to 100, and the number of divided windows is $w_R \in [10, 20, 25, 50]$. The number of grid visual features $m_G$ is set to 64, and the corresponding number of division windows is $w_G \in [4, 8, 16, 32]$. The CLEVR dataset uses more minor image grid features, the number of grid features $m_C$ is set to 224, and the corresponding local window is divided into $w_C \in [28, 56, 112]$. The model training parameters are optimized using the Adam [23] optimizer, where $\beta_1 = 0.9$ and $\beta_2 = 0.999$ are set. We adopted a learning rate warm-up strategy, where the learning rate is set to $\min\left(2.5te^{-5}, 1e^{-4}\right)$, where $t$ is the current epoch number starting from 1. After 10 epochs, the learning rate is decayed by 1/5 every 2 epochs. The number of training epochs on the VQA 2.0 and CLEVR datasets is 13 and 16, respectively. The batch size is set to 64 and uses the binary cross-entropy loss (BCE) function. A gradient clipping strategy with a threshold of 0.25 is used to prevent exploding gradients during training. We use weight normalization and dropout for each linear map to stabilize training and prevent overfitting.

## 4.3 Ablation analysis

In this section, we mainly discuss the test results of the LSAT model on the VQA 2.0 and CLEVR benchmark datasets. The region and grid visual features are employed for training and testing on the VQA 2.0 dataset, respectively. Different sizes of windows and interaction methods are set for specific image feature sizes to test the impact on the model performance. The *train+val+vg* method was used to train on the VQA 2.0 dataset and tested on the test dataset, and get the results. Visual Genome (VG) is an ongoing effort to connect structured image concepts and language datasets. We utilize the *train+val* method to train and test On the CLEVR dataset. We explore the effect of the local self-attention mechanism on grid visual features in Section 4.3.1 and the impact on regional visual features in Section 4.3.2. Section 4.3.3 studies the effect of the local self-attention mechanism on the CLEVR dataset.

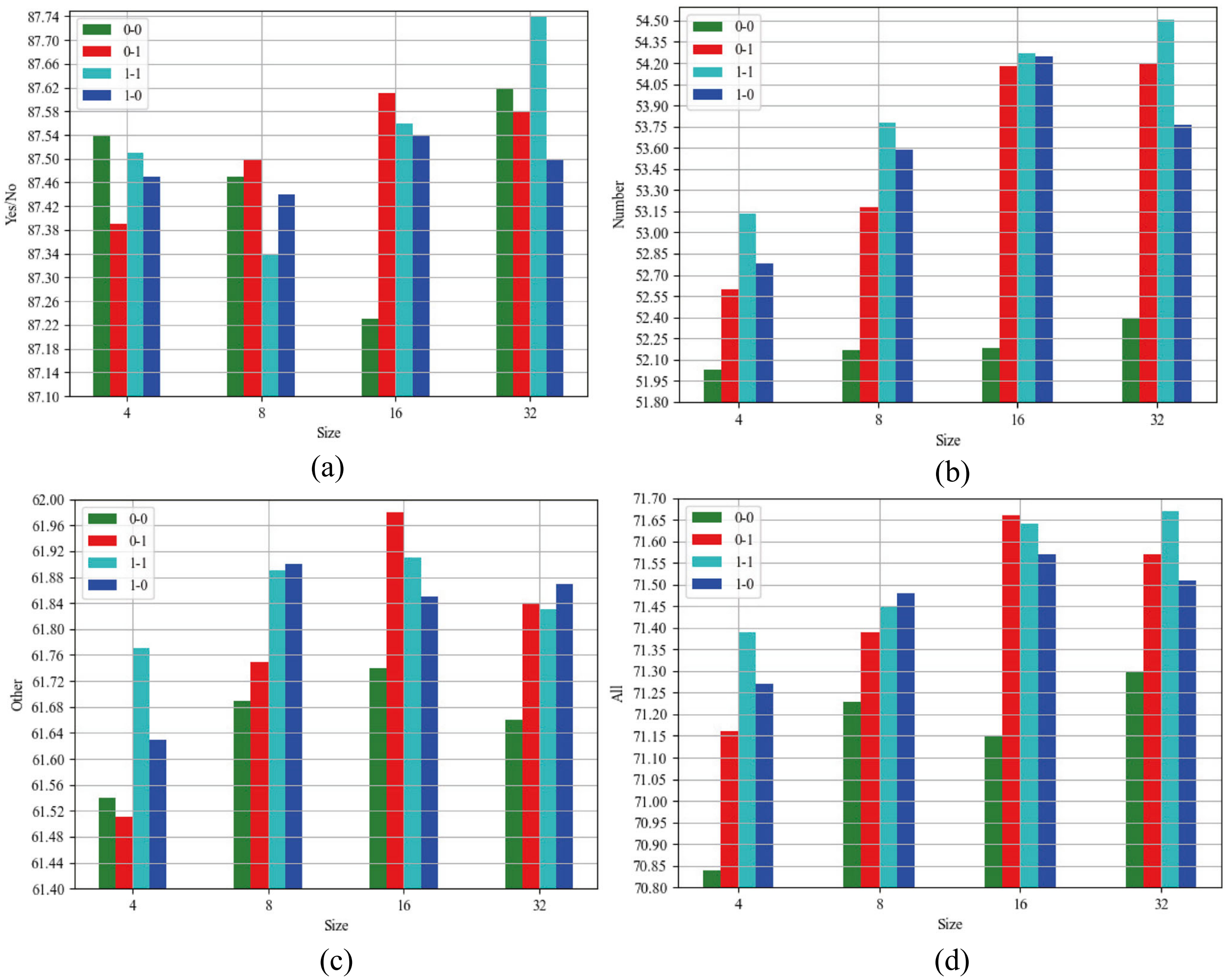
### 4.3.1 Grid image features with local self-attention

As shown in Fig. 6, we report our experimental data results with different interaction modes and different window sizes. Where '0-0' denotes $f$=0, $b$=0, each local window only realizes the self-attention learning modeling within the window. It can be seen from the experimental results that the local window is used but the interaction between neighboring windows is not modeled. The experimental results are not very good. '0-1' represents $f$=0, $b$=1, which is to model the interactive attention learning inside each local window and between its preceding neighboring local windows. '1-0' means $f$=1, $b$=0, which is to model the interactive attention learning inside each local window and adjacent local windows. '1-1' means $f$=$b$=1, (*i.e.*, modeling interactive attention learning inside each local window and the neighboring local windows in front and behind it). The experimental results show that better results can be achieved in modeling the interaction between the local window and the adjacent windows before and after. For example, when the window size is set at $w$=32, the highest accuracy is 71.67% when answering the question type 'All'.

Table 1 lists the best experimental results with different interaction methods and window sizes and the experimental results for comparison of the advanced benchmark TRAR_Base model. TRAR_Base [10] represents a baseline model based on the traditional Transformer structure global attention mechanism and uses grid visual features as input. In order to facilitate writing, use LSAT$_G$-[($f$, $b$),$w$] to represent the experimental method using grid image features in different windows and different interaction methods. Where $f$=$b$=1 indicates backward or forward attention, $w$ denotes the window size. In particular, the best experimental results can be obtained when both the front and rear of the local window are concerned with each other, and the test result of answering the question type 'All' on test-std is 71.94%. The above experiments demonstrate the effectiveness of the LSAT model on grid visual features.

### 4.3.2 Region image features with local self-attention

This section discusses the experimental effect of using the local self-attention mechanism on regional visual features. We use different window sizes and interaction methods to explore the effectiveness of local self-attention. Since the number of regional features is $m_R \in [10, 100]$ , the window size is set to $w_R \in [10, 20, 25, 50]$ .The properties of the experimental parameters $f$ and $b$ discussed in this section are defined similarly as in Section 4.3.1. As shown in Fig. 7, we model the interaction between windows and lean forward or backward attention. Although the accuracy of the LSAT model in answering the question types 'Other' and 'All' is significantly improved, the experimental effect of regional visual features is not as good as that of grid visual features. We think this is because the number of regional visual features is not fixed. The grid feature is to grid all the image features. The number of features is fixed, and the number of features containing visual information



(a)

(b)

(c)

(d)

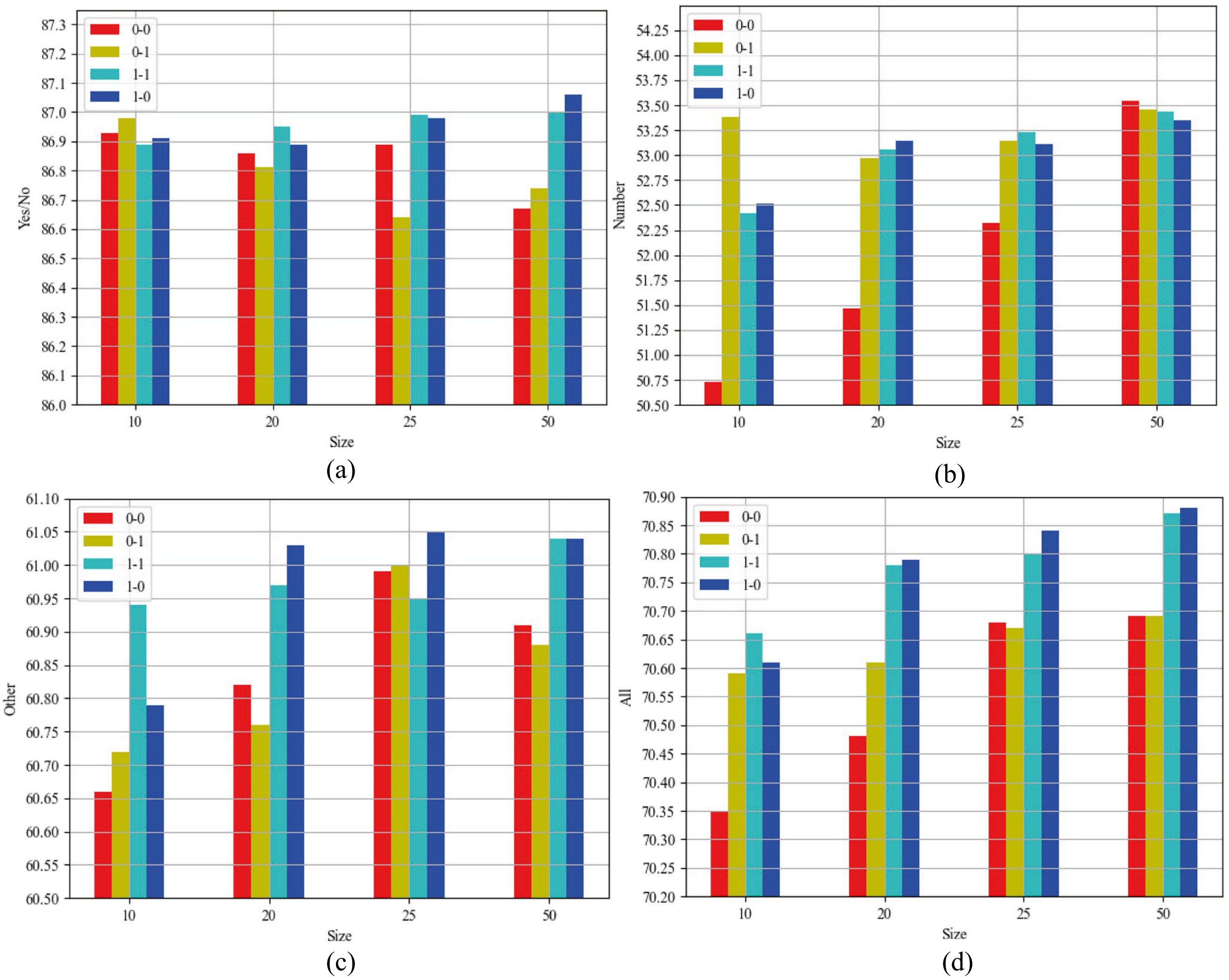
**Fig. 6** The experimental results of the LSAT model are based on grid visual features on the VQA 2.0 dataset, and the windows sizes are set to $w_G \in [4, 8, 16, 32]$: (a) represents the accuracy of "Yes/No" using different interaction methods, (b) represents the accuracy of "Number" using different interaction methods, (c) represents the accuracy of "Other" using different interaction methods, and (d) represents the accuracy of "All" using different interaction methods

related to correctly answering the question is relatively stable. Therefore, the local self-attention mechanism is more suitable for a model using grid features. This also proves the necessity of modelling the window's interior and interaction with other windows.

Table 2 lists the experimental results of using a the window size $w$ of 50 and under different interaction modes to compare with MCAN [2]. MCAN is a model that uses the global self-attention mechanism in traditional transformers under regional visual features. For the convenience of

**Table 1** Represents the best results using local attention with different interaction styles and window sizes under visual grid features

| | Test-dev | | | | Test-std |
|---|---|---|---|---|---|
| Model | Yes/No | Number | Other | All | All |
| TRAR_Base [10] | 87.43 | 53.80 | 61.81 | 71.45 | – |
| LSAT$_G$-[(0,0),32] | 87.62 | 52.39 | 61.66 | 71.30 | – |
| LSAT$_G$-[(0,1),16] | 87.61 | 54.18 | 61.98 | 71.66 | – |
| LSAT$_G$-[(1,1),32] | 87.74 | 54.51 | 61.83 | 71.67 | 71.94 |
| LSAT$_G$-[(1,0),16] | 87.54 | 54.25 | 61.85 | 71.57 | – |



(a)

(b)

(c)

(d)

**Fig. 7** The size of the window is set to $w_R \in [10, 20, 25, 50]$: (a) represents the accuracy of 'Yes/No' using different interaction methods, (b) represents the accuracy of 'Number' using different interaction methods, (c) represents the accuracy of 'Other' using different interaction methods, and (d) represents the accuracy of 'All' using different interaction methods

writing, LSAT$_R$-[$f$-$b$,$w_R$] represents the use of different local windows and interaction methods under the regional visual feature. When $f$ or $b$ is set to 1, there is an interaction between windows. Otherwise, there is no interaction. The experimental results in Table 2 show that when only the a local window is used without interaction, the results are better than in the MCAN model, where the LSAT$_R$-[0-0,50] of test result on the *test-std* dataset is 71.14% ( 0.24% higher than MCAN). We believe that the critical target features detected using regional visual features are concentrated in a specific image region. The local self-attention mechanism can more effectively locate the target region. Therefore,

**Table 2** Represents the best results using local attention with different interaction styles and window sizes under regional image features

| Model | Test-dev | | | | Test-std |
|---|---|---|---|---|---|
| | Yes/No | Number | Other | All | All |
| MCAN [2] | 86.82 | 53.26 | 60.72 | 70.63 | 70.90 |
| LSAT$_R$-[(0,0),50] | 86.87 | 53.54 | 60.91 | 70.69 | 71.14 |
| LSAT$_R$-[(0,1),50] | 86.74 | 53.46 | 60.88 | 70.69 | 70.93 |
| LSAT$_R$-[(1,1),50] | 87.00 | 53.43 | 61.04 | 70.87 | 70.99 |
| LSAT$_R$-[(1,0),50] | 87.06 | 53.32 | 61.04 | 70.88 | 71.13 |

only modelling the interior of the local window is better than that of the MCAN model, and the experimental results in the case of interaction also exceed the MCAN baseline model, which fully demonstrates the effectiveness and feasibility of the LSTA model.

#### 4.3.3 Local self-attention on CLEVR

This section mainly discusses the experimental results using the local attention mechanism on the CLEVR dataset. The images in the CLEVR [24] dataset have 224×224 grid image features, and we use different windows and interaction methods to test the model effects on the CLEVR dataset. As shown in Table 3, the window size was set to 112, 56, and 28, respectively, to compare with the TRAR_Base model in different interaction modes. LSAT$_C$-[$f$-$b$, $w$] is defined as the size of the global image feature window in the CLEVR dataset, and is 224, and the local window size can be divided under different interaction methods as $w$. $f$-$b$ represents the interaction mode before the window, which is the same as the definition used in the above experiment. Table 3 shows that when the window size is set reasonably, the effect of using the local self-attention mechanism is significantly better than the global self-attention model TRAR_Base [10] model even without interaction(0-0). Since the image feature information in the CLEVR dataset is more fragmented, it can perform well when modeled with local windows. It also illustrates the critical role of the local self-attention mechanism in the end-to-end visual reasoning task.

### 4.4 Comparison with state-of-the-art

#### 4.4.1 Comparison on VQA 2.0 dataset

The experimental data are based on the VQA 2.0 dataset test results. We compare the effects of the LSAT with the current advanced VQA methods. As shown in Table 4, the LSAT model surpasses the performance of these traditional Transformer-based VQA methods and achieves new state-of-the-art performance on this benchmark. All results were obtained by a single model. Table 4 is divided into three blocks in the row. The first part summarizes several models that do not use Faster-RCNN to extract features, and the second part uses pre-trained Faster-RCNN to detect salient objects while using Glove for word vector encoding. Our results use the same pre-trained Faster-RCNN and Glove models to experiment with the last block's grid and region visual features. The experimental results demonstrate the effectiveness and applicability of our method.

Among these state-of-the-art VQA models, Bottom-up [25] and Bottom-up+MFH [25] are models that combine regional visual features with question-guided visual attention, which takes into account the natural underpinnings of attention. BAN [28] is a bilinear attention network that considers bilinear interactions between input multimodalities to fully exploit question and image feature information. BAN-Counter [28] combines BAN with Counter [28], a neural network component that allows robust counting between visual object proposals, further improving the model's accuracy on counting metrics.

**Table 3** Comparison of experimental results using different window sizes and interaction methods on the CLEVR dataset

| Model | f-b | All | Count | Exist | CN | QA | CA |
|---|---|---|---|---|---|---|---|
| TRAR_Base [10] | – | 98.54 | 96.34 | 99.24 | 98.60 | 99.43 | 98.93 |
| LSAT$_C$-[$f$-$b$,28] | 0-0 | 98.55 | 96.30 | 99.20 | 98.58 | 99.46 | 99.13 |
| | 0-1 | 98.59 | 96.45 | 99.26 | 98.75 | 99.52 | 98.96 |
| | 1-0 | 98.30 | 95.79 | 99.13 | 97.86 | 99.44 | 98.81 |
| | 1-1 | 98.27 | 95.40 | 99.18 | 98.51 | 99.50 | 98.80 |
| LSAT$_C$-[$f$-$b$,56] | 0-0 | 98.72 | 96.81 | 99.31 | 98.77 | 99.52 | 99.12 |
| | 0-1 | 98.63 | 96.53 | 99.21 | 98.76 | 99.48 | 99.14 |
| | 1-0 | 98.52 | 96.04 | 99.30 | 98.83 | 99.48 | 99.11 |
| | 1-1 | 98.63 | 96.53 | 99.25 | 98.83 | 99.50 | 99.02 |
| LSAT$_C$-[$f$-$b$,112] | 0-0 | 98.50 | 96.22 | 99.15 | 98.69 | 99.50 | 99.13 |
| | 0-1 | 98.54 | 96.49 | 99.26 | 97.88 | 99.50 | 98.96 |
| | 1-0 | 98.55 | 96.10 | 99.22 | 98.68 | 99.46 | 99.06 |
| | 1-1 | 98.41 | 95.61 | 99.22 | 98.78 | 99.52 | 99.02 |

CI, QA, CA stand for Count Integer, Query Attribute and Compare and Compare Attribute, respectively

**Table 4** Performance comparison results on VQA 2.0 with SOTAs

| | Test-dev | | | | Test-std |
|---|---|---|---|---|---|
| Model | Yes/No | Number | Other | All | All |
| Language only [19] | – | – | – | – | 44.26 |
| LSTM+CNN [19] | – | – | – | – | 54.22 |
| MCB reported in [19] | – | – | – | – | 62.27 |
| DCN [36] | 83.50 | 46.60 | 56.72 | 66.60 | 67.00 |
| Bottom-up [25] | 81.82 | 44.21 | 56.05 | 65.32 | 65.67 |
| Bottom-up+MFH [25] | 84.27 | 49.56 | 59.89 | 68.76 | – |
| MFH [29] | 85.31 | 49.56 | 59.89 | 68.76 | – |
| BAN [28] | 85.42 | 50.93 | 60.26 | 69.52 | – |
| BAN-Counter [28] | 85.42 | 54.04 | 60.52 | 70.04 | 70.35 |
| VRR [34] | 83.31 | 45.51 | 58.41 | 67.20 | 67.34 |
| DFAF [37] | 86.09 | 53.32 | 60.49 | 70.22 | 70.34 |
| MuRel [50] | 84.77 | 49.84 | 57.85 | 68.03 | 68.41 |
| ReGAT [49] | 86.08 | 54.42 | 60.33 | 70.27 | 70.58 |
| MCAN [2] | 86.82 | 53.26 | 60.72 | 70.63 | 70.90 |
| TRAR_Base [10] | 87.43 | 53.80 | 61.81 | 71.45 | – |
| LSAT-R(ours) | 87.06 | 53.32 | 61.04 | 70.88 | 71.13 |
| LSAT-G(ours) | **87.74** | **54.51** | **61.83** | **71.67** | **71.94** |

The blackened part indicates that the accuracy is the highest among all the comparison data, emphasizing the meaning of higher accuracy than other data

DCN [36], DFAF [37] and MCAN [2] have similar network architectures, as these models all build a deep co-attention network to mine dense interactions among multimodalities so that the models achieve the best performance. MuRel [50] and ReGAT [49] utilize graph neural networks to build deep inference networks, which build graph inference networks based on relationships between objects and achieve impressive experimental results, especially on the type for "Number" metrics. TRAR_Base [10] is a model that employs the global self-attention mechanism in a traditional transformer under grid visual features, which can further refine image features by employing grid visual features. We utilized the local self-attention mechanism to capture richer visual context information based on regional and grid visual features. The performance of our proposed LSAT model compared with current state-of-the-art visual question answering models on the VQA 2.0 dataset is shown in Table 4. The MCAN model is the champion model of the 2019 VQA Challenge. Compared with the baseline model MCAN, the LSAT-R model has higher accuracy. The accuracy of answering the question type "All" is improved by 0.25% and 0.20% on test-dev and test-std, respectively. Compared with the benchmark TRAR_Base model, the LSAT-G model is also significantly improved, and the accuracy of answering the question type "All" is improved by 0.22% in test-dev. It is worth noting that the accuracy of the LSAT model in the "Number" type has been significantly improved, and it surpassed the existing models BAN-Counter [28] and ReGAT [49], which are good at answering the count type.

### 4.4.2 Comparison on CLEVR dataset

To further evaluate the generalization ability of LSAT, we also validated LSAT on the another widely used benchmark, CLEVR. CLEVR is primarily a visual reasoning dataset, and the questions involve complex reasoning. As shown in Table 5, the LSAT model outperforms existing state-of-the-art models on all question attribute accuracy metrics tested on the CLVER dataset. FILM [51] is very effective for visual reasoning models, demonstrating that answering image-related questions requires multi-step reasoning. TBD [52] proposes a complex visual question answering model for visual primitive reasoning. The recently proposed SNAMT [53] model is based on an encoder and decoder structure, which can adaptively adjust the question feature encoding and layout, and decoding by considering intermediate question results. v-VRANet [54] proposes a novel encoder-decoder visual relational reasoning module that can reason about object-relational visual and textual information guided by textual information. RWSAN [55] proposes a residual weight-sharing attention network based on the encoder and decoder structure, which achieves competitive performance with lower cost and fewer parameters. The

**Table 5** Comparison of experimental results using different window sizes and interaction methods on the CLEVR dataset

| Methods | All | Count | Exist | CN | QA | CA |
|---|---|---|---|---|---|---|
| Human [24] | 92.60 | 86.70 | 96.60 | 86.40 | 95.00 | 96.00 |
| FILM [51] | 97.60 | 94.50 | 99.20 | 93.80 | 99.20 | 99.00 |
| TBD [52] | 98.70 | 96.80 | 98.90 | 99.10 | 99.40 | 99.20 |
| SNAMT [53] | 96.60 | 91.50 | 97.90 | 98.70 | 98.60 | 98.00 |
| v-VRANet [54] | 96.10 | 93.40 | 97.10 | 95.10 | 97.40 | 96.80 |
| RWSAN [55] | 98.42 | 96.34 | 99.27 | 97.66 | 99.45 | 98.71 |
| TRAR_Base [10] | 98.54 | 96.34 | 99.24 | 98.60 | 99.43 | 98.93 |
| LSAT-C (ours) | **98.72** | **96.81** | **99.31** | **98.77** | **99.52** | **99.22** |

The blackened part indicates that the accuracy is the highest among all the comparison data, emphasizing the meaning of higher accuracy than other data

TRAT-Base [10] model is a global self-attention visual reasoning model based on an encoder and decoder. The experimental results in Table 5 show that the LSAT model employing the local self-attention mechanism for visual feature inference outperforms existing state-of-the-art models, proving that the LSAT model has good performance.

## 4.5 Qualitative analysis

Figure 8 compares global/local self-attention for regional and global/local self-attention for grid visual features by visualization. As shown in Fig. 8, when global self-attention is used for modelling, although the image has a wide range of attention, it also contains considerable noise and irrelevant information. Especially when utilizing grid visual features, the image is divided into different feature blocks. Although the image features are fine-grained, much irrelevant information and noise are introduced in the self-attention learning, which affects the model performance. As shown in Fig. 8(a), when using global self-attention, the regional visual features focus on the person's action "eat" and the integrity of the "pizza" part, resulting in the wrong final answer. The main focus of the global mesh feature is the integrity of the entire pizza. When using a local self-attention mechanism, interactive learning can be modelled both within a window and across windows. The model can pay attention to the action of 'eat' and the change of 'pizza' simultaneously to accurately understand the image feature information. Although the final answer

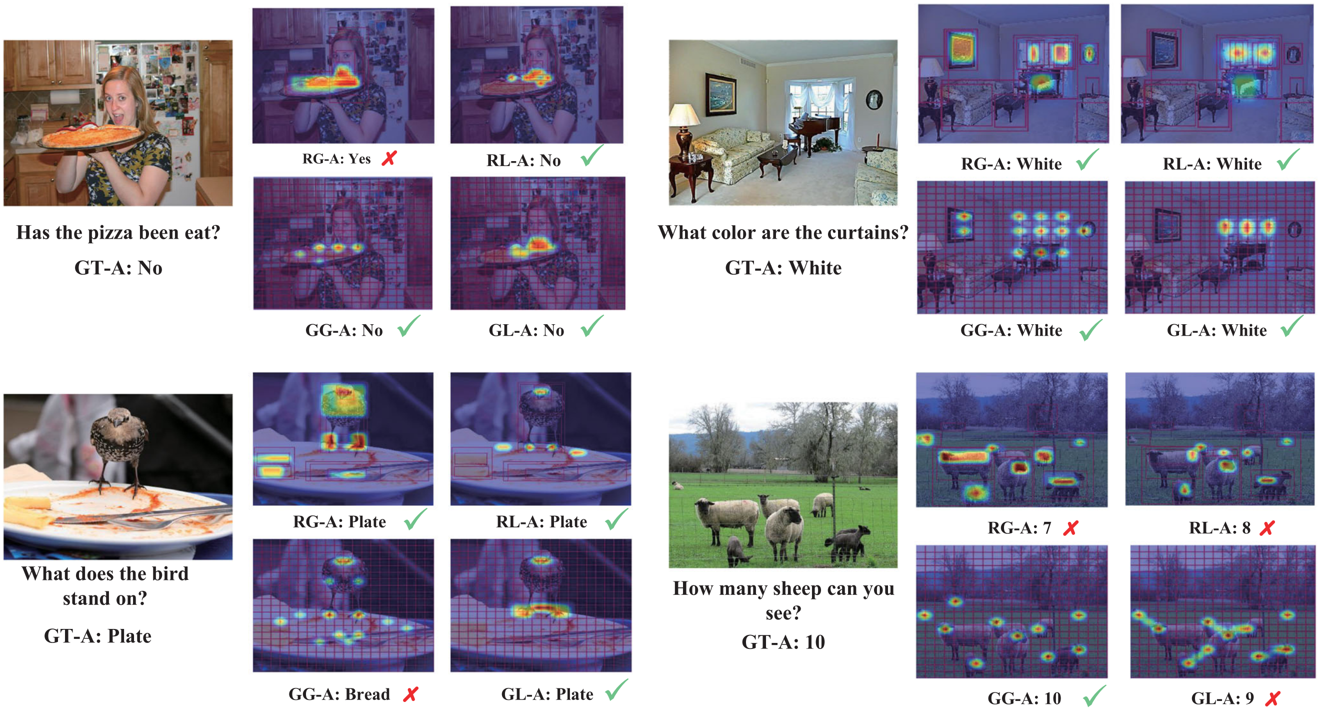


**Fig. 8** Model visualization results, source images taken from VQA 2.0 dataset [56]. We show ground-truth answers (GT-A), and region global answers (RG-A) region local answers (RL-A). At same time, we also show grid global answers (GG-A) and grid local answers (GL-A)

of the model in Fig. 8(b) is correct, we can see from the figure that the object regions that the model pays attention to are different. Many visual objects in the region are concerned with the global self-attention redundant, which will increase the computational complexity of the model and reduce performance of the model. When using the global grid visual feature in Fig. 8(c), the image features are more fine-grained because the grid visual feature is used. The model mainly focuses on learning the target objects in the 'plate' so that the model understands that there is a deviation in locating the target object. Figure 8(d) examines the counting ability of the model. The figure shows that although the global self-attention using grid visual features gives the correct answer type, it is not the correct answer. It is also wrong in other examples. For the counting problem, the model must not only focus precisely on the target object in the image, but also deal with some overlapping, tiny target objects. As shown in Fig. 8, we employ the local attention mechanism in the grid visual features. The model can precisely locate the counting target and capture rich contextual information through the mutual interaction between windows, effectively handling some small, overlapping target objects, thereby improving the model's counting ability. Although our model does not answer correctly, it can effectively approximate the correct answer. The following visualizations demonstrate the effectiveness and interpretability of the LSAT model.

## 5 Conclusion

Using global self-attention to model images in traditional Transformers cannot capture rich contextual information and also increases the model's computational overhead. This traditional approach also introduces considerable irrelevant information and noise. Accordingly, this paper proposes an LSAT model based on a local self-attention mechanism. LSAT reduces the introduction of irrelevant information and noise by using adjustable local windows and reduces the computational complexity of the model's global attention. In addition, LSAT can model the interaction between windows to better capture visual context information with local information. To verify the effectiveness of our method, we conducted validation tests on two benchmark datasets, VQA 2.0 and CLEVR, utilizing regional and grid visual features, proving the vital role of the local self-attention mechanism in end-to-end visual task reasoning. Finally, the superiority of the LSAT model is demonstrated through ablation experiments and attention visualization.

We hope that the local self-attention mechanism can be widely used in future work developing the Transformer vision field. The local attention window we designed is a manually set local window with a fixed size. In future research, we will explore a self-adjustable size of for the window. Adapting sliding windows can help models to better achieve self-attention learning.

**Acknowledgements** This research is supported by the National Natural Science Foundation of China under Grant 61873160 and Grant No.61672338, and the Natural Science Foundation of Shanghai under Grant 21ZR1426500. This research is also supported by Scientific Research Fund of Hunan Provincial Education Department (Grant No.21A0470). We thank all the reviewers for their constructive comments and helpful suggestions.

## References

1. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. Adv Neural Inf Process Syst (NIPS) 30:5998–6008
2. Yu Z, Yu J, Cui Y, Tao D, Tian Q (2019) Deep modular co-attention networks for visual question answering. In: IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6281–6290
3. Guo Z, Han D (2022) Sparse co-attention visual question answering networks based on thresholds. Appl Intell :1–15
4. Chen H, Ding G, Lin Z, Zhao S, Han J (2019) Cross-modal image-text retrieval with semantic consistency. In: Proceedings of the 27th ACM international conference on multimedia, pp 1749–1757
5. Zhang Z, Lin Z, Zhao Z, Xiao Z (2019) Cross-modal interaction networks for query-based moment retrieval in videos. In: Proceedings of the 42nd international ACM SIGIR conference on research and development in information retrieval, pp 655–664
6. Zhou Q, Qiang Y, Mo Y, Wu X, Latecki LJ (2022) Banet: Boundary-assistant encoder-decoder network for semantic segmentation. IEEE Transactions on Intelligent Transportation Systems
7. Zhou Q, Wu X, Zhang S, Kang B, Ge Z, Latecki LJ (2022) Contextual ensemble network for semantic segmentation. Pattern Recogn 122:108290
8. Al-Malla MA, Jafar A, Ghneim N (2022) Image captioning model using attention and object features to mimic human image understanding. J Big Data 9(1):1–16
9. Mei Y, Fan Y, Zhou Y (2021) Image super-resolution with non-local sparse attention. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 3517–3526
10. Zhou Y, Ren T, Zhu C, Sun X, Liu J, Ding X, Xu M, Ji R (2021) Trar: Routing the attention spans in transformer for visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 2074–2084
11. Jiang H, Misra I, Rohrbach M, Learned-Miller E, Chen X (2020) In defense of grid features for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 10267–10276
12. Anderson P, He X, Buehler C, Teney D, Johnson M, Gould S, Zhang L (2018) Bottom-up and top-down attention for image captioning and visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 6077–6086
13. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Y, Lin S, Guo B (2021) Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 10012–10022

14. Huang Z, Ben Y, Luo G, Cheng P, Yu G, Fu B (2021) Shuffle transformer: Rethinking spatial shuffle for vision transformer. arXiv:2106.03650
15. Fang J, Xie L, Wang X, Zhang X, Liu W, Tian Q (2022) Msg-transformer: Exchanging local spatial information by manipulating messenger tokens. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12063–12072
16. Zhou Q, Yang W, Gao G, Ou W, Lu H, Chen J, Latecki JL (2019) Multi-scale deep context convolutional neural networks for semantic segmentation. World Wide Web 22(2):555–570
17. Wang H, Zhu Y, Green B, Adam H, Yuille A, Chen L-C (2022) Axial-deeplab: Stand-alone axial-attention for panoptic segmentation. In: European conference on computer vision (ECCV). Springer, pp 108–126
18. Dong X, Bao J, Chen D, Zhang W, Yu N, Yuan L, Chen D, Guo B (2022) Cswin transformer: A general vision transformer backbone with cross-shaped windows. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12124–12134
19. Goyal Y, Khot T, Summers-Stay D, Batra D, Parikh D (2017) Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 6904–6913
20. Lee K-H, Chen X, Hua G, Hu H, He X (2018) Stacked cross attention for image-text matching. In: Proceedings of the European conference on computer vision (ECCV), pp 201–216
21. Lu J, Yang J, Batra D, Parikh D (2016) Hierarchical question-image co-attention for visual question answering. Advances in neural information processing systems, 29
22. Nam H, Ha J-W, Kim J (2017) Dual attention networks for multimodal reasoning and matching. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 299–307
23. Kingma DP, Ba J (2014) Adam: A method for stochastic optimization. arXiv:1412.6980
24. Johnson J, Hariharan B, Van Der Maaten L, Fei-Fei L, Lawrence Zitnick C, Girshick R (2017) Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 2901–2910
25. Teney D, Anderson P, He X, Van Den Hengel A (2018) Tips and tricks for visual question answering: Learnings from the 2017 challenge. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4223–4232
26. Fan H, Zhou J (2018) Stacked latent attention for multimodal reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 1072–1080
27. Yu Z, Yu J, Fan J, Tao D (2017) Multi-modal factorized bilinear pooling with co-attention learning for visual question answering. In: Proceedings of the IEEE international conference on computer vision (ICCV), pp 1821–1830
28. Kim J-H, Jun J, Zhang B-T (2018) Bilinear attention networks. Advances in neural information processing systems, 31
29. Yu Z, Yu J, Xiang C, Fan J, Tao D (2018) Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. IEEE Trans Neural Netw Learn Syst 29(12):5947–5959
30. Guo J, Han K, Wu H, Tang Y, Chen X, Wang Y, Xu C (2022) CMT: Convolutional neural networks meet vision transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 12175–12185
31. Han K, Xiao A, Wu E, Guo J, Xu C, Wang Y (2021) Transformer in transformer. Adv Neural Inf Process Syst (NIPS) 34:15908–15919
32. Chen C, Han D, Chang C-C (2022) CAAN: Context-aware attention network for visual question answering. Pattern Recogn 132:108980
33. Liu Y, Zhang X, Zhang Q, Li C, Huang F, Tang X, Li Z (2021) Dual self-attention with co-attention networks for visual question answering. Pattern Recogn 117:107956
34. Shen X, Han D, Chang C-C, Zong L (2022) Dual self-guided attention with sparse question networks for visual question answering. IEICE Trans Inf Syst 105(4):785–796
35. Yang Z, He X, Gao J, Deng L, Smola A (2016) Stacked attention networks for image question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 21–29
36. Nguyen D-K, Okatani T (2018) Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 6087–6096
37. Gao P, Jiang Z, You H, Lu PC, Hoi S, Wang X, Li H (2019) Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 6639–6648
38. Lu J, Batra D, Parikh D, Lee S (2019) Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32
39. Zhou L, Palangi H, Zhang L, Hu H, Corso J, Gao J (2020) Unified vision-language pre-training for image captioning and VQA. In: Proceedings of the AAAI conference on artificial intelligence, vol 34(7), pp 13041–13049
40. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: European conference on computer vision (ECCV). Springer, pp 213–229
41. Zhu X, Su W, Lu L, Li B, Wang X, Dai JFDD (2021) Deformable transformers for end-to-end object detection. In: Proceedings of the 9th international conference on learning representations virtual event, Austria: OpenReview. net
42. Touvron H, Cord M, Douze M, Massa F, Sablayrolles A, Jégou H (2021) Training data-efficient image transformers & distillation through attention. In: International conference on machine learning, pp 10347–10357
43. Yuan L, Chen Y, Wang T, Yu W, Shi Y, Jiang Z-H, Tay EF, Feng J, Yan S (2021) Tokens-to-token vit: Training vision transformers from scratch on imagenet. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp 558–567
44. Wang W, Xie E, Li X, Fan D-P, Song K, Liang D, Lu T, Luo P, Shao L (2021) Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 568–578
45. Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A, Tran D (2018) Image transformer. In: International conference on machine learning (PMLR), pp 4055–4064
46. Watzel T, Kürzinger L, Li L, Rigoll G (2021) Induced local attention for transformer models in speech recognition. In: International conference on speech and computer. Springer, pp 795–806

47. Wu S, Wu T, Tan H, Guo G (2021) Pale transformer: A general vision transformer backbone with pale-shaped attention. arXiv:2112.14000
48. Krishna R, Zhu Y, Groth O, Johnson J, Hata K, Kravitz J, Chen S, Kalantidis Y, Li L-J, Shamma AD et al (2017) Visual genome: Connecting language and vision using crowdsourced dense image annotations. Int J Comput Vis (IJCV) 123(1):32–73
49. Li L, Gan Z, Cheng Y, Liu J (2019) Relation-aware graph attention network for visual question answering. In: Proceedings of the IEEE/CVF international conference on computer vision (ICCV), pp 10313–10322
50. Cadene R, Ben-Younes H, Cord M, Thome N (2019) Murel: Multimodal relational reasoning for visual question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 1989–1998
51. Gulrajani I, Ahmed F, Arjovsky M (2017) Vincent, dumoulin, and aaron c courville. Improved training of, wasserstein gans. In: NeurIPS, p 3
52. Mascharka D, Tran P, Soklaski R, Majumdar A (2018) Transparency by design: Closing the gap between performance and interpretability in visual reasoning. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), pp 4942–4950
53. Zhong H, Chen J, Shen C, Zhang H, Huang J, Hua X-S (2020) Self-adaptive neural module transformer for visual question answering. IEEE Trans Multimed 23:1264–1273
54. Yu J, Zhang W, Lu Y, Qin Z, Hu Y, Tan J, Wu Q (2022) Reasoning on the relation: Enhancing visual representation for visual question answering and cross-modal retrieval. IEEE Trans Multimed 22(12):3196–3209
55. Qin B, Hu H, Zhuang Y (2022) Deep residual weight-sharing attention network with low-rank attention for visual question answering. IEEE Transactions on Multimedia
56. Antol S, Agrawal A, Lu J, Mitchell M, Batra D, Zitnick CL, Parikh D (2015) VQA: Visual question answering. In: Proceedings of the IEEE international conference on computer vision (CVPR), pp 2425–2433

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Xiang Shen** is currently pursuing the Ph.D. degree with Shanghai Maritime Unversity. His research interests include visual question answering and maching learning.

**Dezhi Han** received the B.S. degree in applied physics from the Hefei University of Technology, Hefei, China, in 1990, and the M.S. and Ph.D. degrees in computing science from the Huazhong University of Science and Technology, Wuhan, China, in 2001 and 2005, respectively.

He is currently a Professor with the Department of Computer, Shanghai Maritime University, Pudong, China, in 2006. His current research interests include cloud and outsourcing security, wireless communication security, network and information securityCVisual Question Answering. He is currently a Member of the IEEE.

**Zihan Guo** is currently pursuing the Ph.D. degree in Shanghai Maritime University. His research interests include computer vision and natural language processing methods related to visual question answering.

**Chongqing Chen** received his B.E. degree from Guangdong University of Technology, China, in 2016, and now is currently pursuing the Ph.D. degree at the School of Information Engineering, Shanghai Maritime University, China. His research interests include visual question answering, visual grounding.

**Jie Hua** received his B.S. degree in software engineering from Wuhan University, Wuhan, China, in 1999 and a PhD degree in software engineering from University of Technology Sydney (UTS), Sydney, Australia, in 2014. He is currently working as Professor at Shaoyang University and a researcher at UTS. His research interests include graph drawing, visualisation, big data, and deep learning.

**Gaofeng Luo** is currently a Professor with the College of Information Engineering, Shaoyang University, China. He received an M.S. degree in information and communication engineering from Central South University, China, in 2010; and a Ph.D. degree from the College of Information Engineering of Shanghai Maritime University, China, in 2020.

His main research interests include Quantum Machine Learning and Quantum Information Processing. Prof. Luo is also a member of the China Computer Federation (CCF).