

Decoupled Global-Local Collaborative Network for Visual Question Answering

Ganlong Zhou¹, Dezhi Han², Xiang Shen^{1,3}, and Xiaoqi Xin¹

¹ School of Information Engineering, Shanghai Maritime University
Shanghai 201306, China

ganlongzhou@163.com (corresponding author)

² School of Information Engineering, Shanghai Maritime University
Shanghai 201306, China

dzhan@shmtu.edu.cn (corresponding author)

³ School of Computer Science, The University of Sydney
NSW 2006, Sydney, Australia
xiang.shen@sydney.edu.au

Abstract. Visual Question Answering (VQA) aims to achieve cross-modal semantic understanding through joint modeling of visual content and natural language. Although existing attention-based approaches effectively align features, they struggle to simultaneously accommodate global semantic modeling and local fine-grained perception within single-path architectures, leading to semantic drift and underutilization of information when addressing multi-perspective reasoning tasks. To mitigate this limitation, we propose a Decoupled Global-Local Collaborative Network (DGL-Net) that structurally partitions the cross-modal reasoning process into two complementary branches: global semantic understanding and local fine-grained modeling. Specifically, the global branch employs hierarchical attention fusion to model long-range dependencies and holistic scene semantics, providing contextual constraints for reasoning. The local branch introduces semantic relation enhancement and relevant region screening to explicitly model intra-modal structural relationships, strengthening fine-grained region-word alignment by injecting structural information. During training and inference, a branch-decoupling and collaborative fusion strategy enables independent yet complementary feature learning. Extensive experimental results demonstrate accuracies of 72.15% and 57.87% on VQA-v2 and GQA benchmarks, respectively. Notably, the proposed method exhibits enhanced robustness and generalization in complex semantic and compositional reasoning tasks, validating the efficacy of the decoupled collaborative strategy.

Keywords: Visual Question Answering, Transformer, Decoupled strategy, Cross-Modal Alignment.

1. Introduction

In recent years, multimodal learning has advanced significantly, driven by large-scale pretraining and generative models [30][36][9]. However, the semantic gap between vision and language remains a fundamental challenge constraining the advancement of

cross-modal understanding. Among various cross-modal tasks, Visual Question Answering (VQA) stands out as a particularly demanding benchmark: it requires models to simultaneously perform holistic scene comprehension (e.g., inferring overall scene atmosphere) and fine-grained entity recognition (e.g., localizing subtle visual attributes) within a unified framework, performing semantic reasoning based on their interaction [34][42]. Nevertheless, the crux of this task lies not merely in cross-modal feature alignment, but more fundamentally in the collaborative modeling of information at heterogeneous semantic scales—specifically, how to simultaneously accommodate global semantic understanding and local fine-grained perception within a single reasoning pathway. This inherent tension between global and local demands renders VQA an ideal lens through which to examine structural conflicts in modeling paradigms.

Most existing mainstream methods construct single-path cross-modal interactions based on attention mechanisms, implementing fusion of visual and linguistic information through unified attention allocation strategies. Although such methods have achieved remarkable progress in feature alignment and representation learning, they implicitly assume that information at different semantic scales can be uniformly modeled through the same reasoning pathway[40]. However, this assumption often fails in practice. Specifically, global semantic modeling relies on holistic perception of scene structure and long-range dependencies, whereas local reasoning requires high-resolution discriminative modeling of key regions; these two objectives exhibit essential differences in feature distribution, attention scope, and information granularity. In a single attention pathway, the model struggles to optimize both objectives simultaneously, leading to structural conflicts in attention allocation: when biased toward the global context, fine-grained information becomes diluted; when biased toward local details, overall semantic constraints become insufficient, resulting in semantic drift and unstable reasoning. This global-local conflict arising from the modeling paradigm has become a critical bottleneck limiting performance improvements in VQA models.

Previous studies have attempted to mitigate this issue by enhancing attention mechanisms, such as introducing spatial [28] or channel attention [25] to strengthen local and global information modeling, respectively. However, these methods essentially perform feature enhancement within a single representation space, failing to decouple modeling objectives for different semantic scales at the structural level, and thus cannot fundamentally eliminate conflicts during optimization. Further analysis reveals that different types of questions often correspond to distinct reasoning modes: fine-grained attribute recognition relies on local discriminative features, whereas complex semantic judgment depends on global contextual information. Single-path modeling cannot adaptively match this distinction, resulting in clear capability boundaries for models in complex scenarios. Specifically, as illustrated in Fig 1, for different questions regarding the same image, visual reasoning tasks exhibit significant modal differences: Q1 requires precise identification of fine-grained attributes of specific objects (e.g., “sunglasses” in the question), representing local detail reasoning; Q2 relies on comprehensive judgment of the overall scene atmosphere (e.g., “dangerous appearance” in the question), representing global semantic reasoning. Taking Q1 as an example, when employing methods biased toward global understanding, although the model captures the overall scene layout, it lacks precise anchoring capability for key semantics, causing attention to disperse to irrelevant objects mentioned in the question that appear in multiple regions, making it difficult to focus on fine-grained

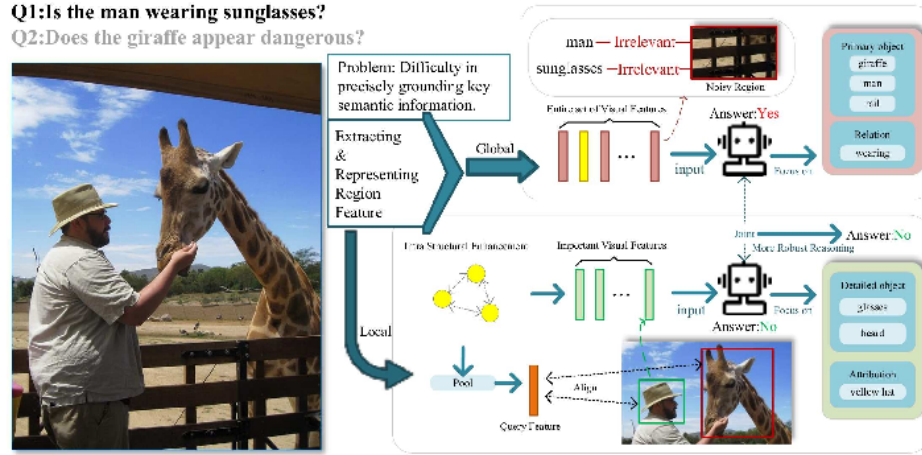


Fig. 1. Motivation illustration of the decoupled collaborative paradigm. The upper and lower panels depict the inference mechanisms of the global and local understanding branches for Q1 (fine-grained attribute recognition) and Q2 (holistic scene reasoning), respectively.

visual information such as texture attributes of small-region objects, ultimately leading to incorrect answers. Conversely, if employing methods biased toward local understanding, the model optimizes the representation capability of regions and words for high-level semantics through contextual semantic modeling, accurately localizes key objects based on cross-modal information interaction, concentrates attention within a streamlined feature subspace to execute fine-grained reasoning, and outputs correct answers. Notably, the correctness of the dual-branch joint inference further validates the efficacy of the decoupled collaborative paradigm: the late fusion mechanism leverages high-confidence predictions from either branch to compensate for the other's low-confidence errors, achieving more robust reasoning than single-path models. However, while this reasoning approach excels at detail capture, it may neglect the grasp of overall scene semantics, making it difficult to handle Q2-type global reasoning tasks requiring comprehensive judgment.

From an optimization perspective, the inherent global-local conflict in single-path architectures can be formalized as a multi-objective optimization dilemma within shared parameter spaces, analogous to the gradient interference phenomenon identified in multi-task learning[46]. Let Θ denote the unified parameter space, and let $\mathcal{L}_g(\theta)$ and $\mathcal{L}_l(\theta)$ denote the global semantic understanding objective and the local fine-grained alignment objective, respectively, where $\theta \in \Theta$. The standard joint-training paradigm optimizes the combined loss $\mathcal{L}(\theta) = \mathcal{L}_g(\theta) + \mathcal{L}_l(\theta)$, whose gradient decomposes as $\nabla_{\theta}\mathcal{L} = \nabla_{\theta}\mathcal{L}_g + \nabla_{\theta}\mathcal{L}_l$. When the two gradient components are negatively correlated, i.e., $\cos(\nabla_{\theta}\mathcal{L}_g, \nabla_{\theta}\mathcal{L}_l) < 0$, any parameter update that reduces $\mathcal{L}_g$ inevitably increases $\mathcal{L}_l$, and vice versa. This gradient-direction conflict constitutes the mathematical foundation of the global-local dilemma: the model cannot simultaneously accommodate holistic

scene perception and fine-grained region discrimination within the same parameter subspace without suffering from mutual interference during gradient descent.

Based on these observations, this paper proposes a Decoupled Global-Local Collaborative Network (DGL-Net) from a modeling paradigm perspective, explicitly partitioning the cross-modal reasoning process into two complementary subspaces: global semantic modeling and local fine-grained alignment. Specifically, the global branch considers visual features of the entire image, modeling long-range semantic dependencies and holistic scene structure through hierarchical attention mechanisms to provide stable contextual constraints for reasoning. The local branch models spatial positional information of image regions and semantic dependency structures of question text, excavates intra-modal semantic associations to extract structural information for structural enhancement of features, improves discriminability of key semantic regions and words, and filters visual features highly relevant to the question based on cross-modal alignment mechanisms, effectively suppressing interference from irrelevant regions, thereby supporting refined semantic reasoning. Furthermore, this paper adopts branch decoupling and collaborative reasoning strategies, enabling representations at different semantic scales to be independently optimized while achieving effective fusion, thereby alleviating multi-objective optimization conflicts and enhancing overall reasoning capability.

The main contributions of this paper are summarized as follows:

- From a modeling paradigm perspective, we reveal the inherent conflict between global and local semantic modeling in VQA and identify the structural limitations of single-path attention mechanisms.
- We propose a decoupled global-local collaborative modeling framework that characterizes information at different semantic scales through dual-branch structures, achieving structured modeling of the cross-modal reasoning process.
- Experimental results on VQA-v2 and GQA benchmark datasets demonstrate that the proposed method outperforms existing methods in overall performance and complex reasoning scenarios, validating the effectiveness and generalization capability of this modeling paradigm.

The remainder of this paper is organized as follows. Section 2 reviews related work pertinent to this study, surveying advances in VQA concerning the balance between global and local perception. Section 3 elaborates on the overall architecture and key module designs of the proposed Decoupled Global-Local Collaborative Network. Section 4 delineates the experimental setup and evaluation metrics, followed by extensive comparative experiments and analyses on the VQA-v2 and GQA datasets. Section 5 concludes this work and discusses future research directions.

2. Related Work

2.1. Advances in Visual Question Answering

Visual Question Answering as a pivotal task connecting computer vision and natural language processing, has evolved from simple multimodal fusion to deep collaborative reasoning since the proposal of standard datasets by Antol et al. [2]. Early VQA methods

primarily relied on basic multimodal fusion strategies. Kim et al. [19] proposed a multimodal residual learning framework that achieved efficient fusion of visual-linguistic features through bilinear pooling, pioneering end-to-end VQA training. Yu et al. [48] further proposed the Multimodal Factorized High-order Pooling method, reducing computational complexity of bilinear fusion through low-rank constraints. These early works laid the foundation for multimodal representation learning in VQA but lacked explicit cross-modal alignment mechanisms.

The introduction of attention mechanisms marked the first major leap in VQA technology. Yang et al. [45] proposed the Stacked Attention Networks, which progressively refined semantic localization of image regions through iterative visual focusing. Anderson et al. [1] pioneered the combination of bottom-up object detection and top-down attention mechanisms, achieving object-level precise attention allocation. Rahman et al. [31] further improved reasoning quality through Attention on Attention with contextual gating mechanisms. However, these unidirectional attention methods struggle to model bidirectional interaction between questions and images. To address these limitations, co-attention mechanisms emerged. Yu et al. [47] designed the Deep Modular Co-Attention Network (MCAN), exploring deep interaction of intra-modal and inter-modal relationships through compositions of self-attention and guided-attention units. Lu et al. [24] first proposed hierarchical co-attention frameworks, achieving bidirectional guidance through alternating optimization of visual and textual attention. Peng et al. [28] proposed the Multimodal Relation Attention Network, further enhancing cross-modal relationship modeling capabilities. With the rise of Transformer architectures, large-scale pretrained VQA methods have achieved significant progress. Tan and Bansal [39] proposed LXMERT, which achieved visual-linguistic pretraining through cross-modal Transformer encoders, substantially improving downstream VQA performance. Li et al. [21] developed Visual-BERT, utilizing self-attention mechanisms to mine deep associations between image regions and text. Recently, data augmentation strategies by Chen et al. [8] and dual-attention mechanisms by Mishra et al. [25] have continued to advance attention mechanisms in VQA. Eslami et al. [10] validated the effectiveness of CLIP pretrained models in domain-specific VQA. However, existing Transformer-based methods focus predominantly on global semantic alignment, remaining limited in handling local reasoning problems requiring fine-grained region perception.

2.2. Advances in Global and Local Perception

Existing VQA methods have long faced the inherent conflict between global semantic coverage and local detail focusing in attention allocation strategies [11]. Early research predominantly optimized along single dimensions: spatial attention mechanisms [20,43] strengthen local perception by localizing regions of interest, yet excessive focusing leads to loss of global context; channel attention methods [26] enhance global semantic representation but lack spatial constraints on fine-grained regions.

In recent years, researchers have increasingly recognized the necessity of balancing global and local perception. However, existing methods predominantly adopt shared-parameter or joint training strategies to simultaneously optimize global and local objectives [17,23,27,12], leading to mutual interference during gradient updates. Although Kil et al. [17] mined implicit knowledge for comprehensive representations, Liu et al. [23] employed prompt tuning to guide multi-scale attention, Peng et al. [27] revealed the

empirical importance of global-local fusion, Guo et al. [12] optimized feature balance through loss rescaling, and Niu et al. [26] separated confounding factors via causal intervention, none fundamentally decoupled perception requirements at different scales. The common limitation pervading these approaches is the implicit assumption that a single interaction pathway can uniformly accommodate multi-scale semantics, thereby forcing heterogeneous objectives to compete for the same parameter budget—a constraint equally present in co-attention frameworks [12,44] and recent compact Transformers [34,47,15,37]. More recently, the research community has shifted toward explicit multi-scale modeling and architectural decoupling. Guo et al. [13] and Liu et al. [22] validated the efficacy of decoupling visual interpretation from linguistic reasoning in mathematical problem solving and rationale generation, respectively, yet their scopes do not extend to fine-grained visual question answering. Shen et al. [32,34] further advanced this direction by demonstrating that hybrid or multi-branch architectures can effectively mitigate the inherent trade-off between global stability and local adaptability in cross-modal alignment.

Distinct from these works, the proposed DGL-Net physically separates global semantic reasoning and local fine-grained localization into independently optimized dual-branch architectures. During training, both branches optimize independently, fundamentally eliminating gradient interference inherent in shared-parameter models. During inference, global semantic constraints and local detail localization are organically unified through late fusion mechanisms. This “independent training–collaborative inference” paradigm breaks through the limitations of single pathways, providing a principled solution to the structural conflict between global and local perception during optimization and achieving more reliable multi-scale semantic modeling.

3. Method

This paper proposes the Decoupled Global-Local Collaborative Network, which achieves structural decoupling of cross-modal reasoning via explicitly separated global and local branches, with the overall architecture illustrated in Fig. 2. The framework comprises three core modules: Image and Question Representation (Section 3.1), Global-Local Perception Network (Section 3.2), and Multimodal Fusion and Reasoning (Section 3.3).

3.1. Image and Question Representation

Input Processing

Input images are represented as a series of regional visual features $V \in \mathbb{R}^{m \times d}$ using bottom-up attention mechanisms, where m is the number of detected objects and d is the feature dimension. Specifically, these features are extracted from a pretrained Faster R-CNN detection model with ResNet-101 as the backbone network, pretrained on the Visual Genome dataset[20]. By setting a threshold on detection confidence scores, a dynamic number of salient regions can be obtained, with quantities ranging from 36 to 100. For the i -th detected object, its feature vector $v_i \in \mathbb{R}^d$ is obtained through mean pooling of the convolutional feature map corresponding to that region. This object-detection-based visual feature representation method focuses on semantically salient regions in images, providing structured visual input for subsequent multimodal reasoning tasks.

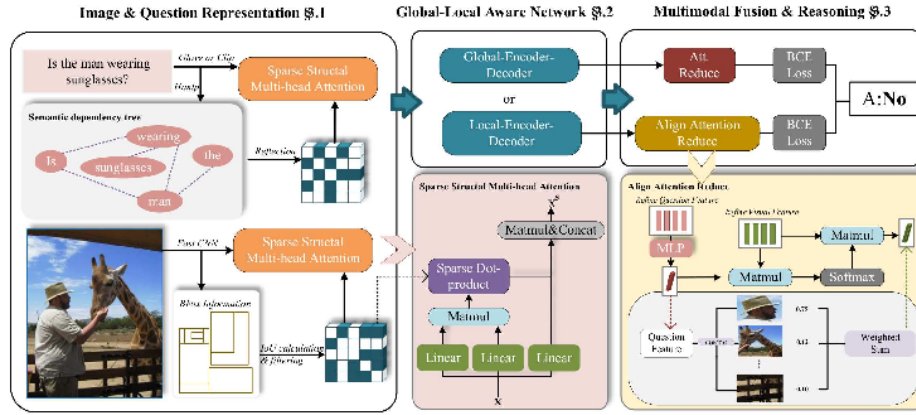


Fig. 2. Overall pipeline of the proposed DGL-Net. Structural information is employed for intra-modal feature enhancement. Two distinct branch backbones are devised for global and local understanding, respectively. Decoupling of global-local reasoning is achieved through the strategy of independent branch training and joint inference.

Question text features are obtained using the CLIP text encoder $T \in \mathbb{R}^{n \times d}$, where n is the sequence length and d is the feature dimension. Specifically, the CLIP employed in experiments utilizes a deep architecture of multi-layer Transformer blocks pretrained on large-scale corpora. Input text undergoes preprocessing and tokenization, converting to discrete token sequences, and is mapped to continuous vectors through a token embedding layer[29]. To preserve positional information between words, the model introduces learnable position encodings added to word embeddings, obtaining initial feature representations $T^{(0)} \in \mathbb{R}^{n \times d_q}$, where d_q is the initial embedding feature dimension. During encoding, a multi-level feature aggregation strategy is adopted, fusing output features from the last layer to obtain final text representations through linear layer dimension alignment.

Intra-Modal Structured Information Representation

In fine-grained local understanding, precise cross-modal alignment depends on the expressive capability of region and word features for complex semantics. However, treating regions and words as isolated feature vectors ignores their intrinsic topological structural relationships. This absence of structural information leads to subsequent self-attention and guided-attention interactions lacking explicit constraints, with attention weights easily dispersing to semantically irrelevant elements, making it difficult to form discriminative local representations. Visually, adjacent or overlapping regions typically exhibit semantic associations; in natural language, words exhibit long-distance dependencies, with semantically related words often adjacent in syntactic structures. Therefore, a mechanism is urgently needed to explicitly model sparse structural associations within modalities, providing prior topological guidance for attention interaction.

Visual Sparse Relation Construction: Given a set of image region features $\mathcal{V} = \{v_1, v_2, \dots, v_n\}$, where $v_i \in \mathbb{R}^d$ is the feature vector of the i -th region and $b_i = (x_1, y_1, x_2, y_2)$ represents its bounding box coordinates. The affinity matrix $A^{\text{vis}} \in \mathbb{R}^{n \times n}$

is calculated between regions, where each entry measures the spatial overlap between bounding boxes using the Intersection-over-Union (IoU) metric: $\bar{A}_{ij}^{\text{vis}} = \text{IoU}(b_i, b_j)$. Subsequently, threshold τ_v is applied for binarization, constructing the visual sparse adjacency matrix:

$$S_{ij}^v = \begin{cases} 1, & \text{if } A_{ij} > \tau_v \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

This matrix characterizes spatial co-occurrence relationships between regions: when two regions significantly overlap in the image, their semantic association probability is high (e.g., “person” and “hat”); conversely, they are considered weakly associated or irrelevant (e.g., “giraffe” and “background fence”).

Textual Sparse Relation Construction: For question text, Semantic Dependency Parsing (SDP) technology is employed to generate syntactic dependency graphs $G = (\mathcal{W}, \mathcal{E})$, where $\mathcal{W} = \{w_1, w_2, \dots, w_m\}$ is the word set and $\mathcal{E}$ is the word set and $S^t \in \mathbb{R}^{m \times m}$:

$$S_{ij}^t = \begin{cases} 1, & \text{if } (w_i, w_j) \text{ or } (w_j, w_i) \in \mathcal{E} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

This matrix captures syntactic-semantic associations between words (e.g., the verb-object relationship between “sunglasses” and “wearing”), filtering noise interference from distant vocabulary.

Structure-Aware Attention Calculation: Visual sparse matrix S^v and textual sparse matrix S^t are injected into corresponding modal multi-head attention mechanisms, respectively. Taking the visual branch as an example, for the h -th attention head, the query-key similarity matrix $V_h V_h^\top \in \mathbb{R}^{n \times n}$ is calculated, followed by application of the sparse mask:

$$M_h^v = \text{Softmax} \left(\frac{X_h X_h^\top \odot S^v}{\sqrt{d_k}} \right) \quad (3)$$

This operation forces attention weights to propagate only within predefined sparse structures, suppressing responses from irrelevant regions. The final output is:

$$V^{\text{str}} = \text{Concat} [M_1^v V_1, \dots, M_h^v V_h] \cdot W \quad (4)$$

Here, W is a learnable parameter matrix. The text side follows the same paradigm, generating T^{str} . Through this process, the SSMA module explicitly encodes geometric space constraints and syntactic structure constraints into feature representations, providing a structured discriminative foundation for subsequent attention interactions.

3.2. Global-Local Aware Network

Based on the image, question, and corresponding structured features generated by the above feature representation modules, this paper further constructs global understanding and local understanding branches, responsible for holistic semantic reasoning and fine-grained region localization tasks, respectively. Both branches adopt encoder-decoder architectures, implementing hierarchical feature interaction through stacking three types of attention units. The following first introduces the design of the three attention units, then details the network structures of both branches.

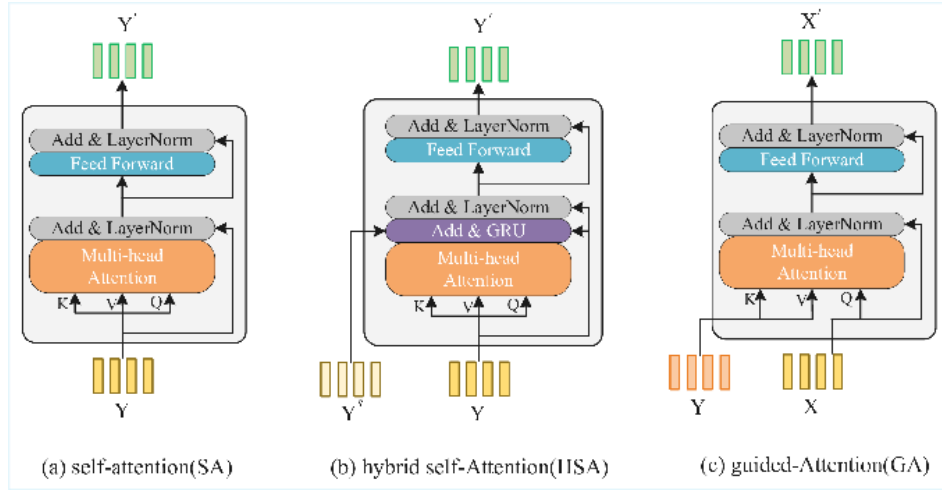


Fig. 3. Three fundamental attention units employed in the Global-Local Perception Network—Self-Attention (SA), Hybrid Self-Attention (HSA), and Guided-Attention (GA)—facilitating intra-modal feature refinement, structural information injection, and cross-modal semantic fusion, respectively.

Attention Modules

Self-Attention Unit (SA): Following the classic Transformer design[41], it utilizes multi-head attention with h parallel heads to compute intra-sequence self-dependencies, thereby refining features, as illustrated in Fig 3(a). Given input features $Y \in \mathbb{R}^{n \times d}$, for the h -th attention head, SA first generates query, key, and value matrices $Q, K, V \in \mathbb{R}^{n \times d}$ through linear projections, then executes scaled dot-product attention:

$$head_h = \text{Softmax} \left(\frac{QK^\top}{\sqrt{d}} \right) V_h \quad (5)$$

Concatenating outputs from all heads yields the complete result:

$$f = \text{Concat} [head_1 V_1, \dots, head_h V_h] \cdot W \quad (6)$$

where W is a learnable weight matrix. Followed by fully connected feed-forward networks and layer normalization to obtain the final output:

$$Y' = \text{LayerNorm} (\text{FFN} (f)) \quad (7)$$

where:

$$\text{FFN} (f) = \max (0, f \cdot W_1) W_2 \quad (8)$$

Here, W_1 and W_2 are learnable weight matrices. SA addresses the limitation of traditional LSTMs in capturing long-distance dependencies between sequence elements, effectively refining intra-modal contextual semantic representations.

Hybrid Self-Attention Unit (HSA): Addressing the structural sensitivity requirements of local understanding tasks, we propose the Hybrid Self-Attention (HSA) unit, as illustrated in Fig 3(b), which incorporates structural features $Y^{str} \in \mathbb{R}^{n \times d}$ on the basis of standard self-attention, introducing GRU to control the injection ratio of structural information into attention outputs:

$$f = \text{GRU}(\text{SA}(Y) + Y^{str}) \quad (9)$$

Modeling its sequence dependencies and potential multi-semantic expressive capabilities through multi-layer recurrence enhances the model's perception capability of key topological relationships. After further processing through fully connected feed-forward networks and layer normalization, more discriminative local representations are obtained.

Guided-Attention Unit (GA): Following the Bottom-Up Top-Down Attention cross-modal interaction paradigm[1], GA implements bidirectional semantic guidance between vision and language modalities, as illustrated in Fig 3(c). Taking visual-guided text as an example, for the h -th attention head, given query question features $T \in \mathbb{R}^{m \times d}$ and guiding visual features $V \in \mathbb{R}^{n \times d}$, GA computes text-to-visual guided attention:

$$\text{head}_h = \text{Softmax}\left(\frac{TW_Q(VW_K)^\top}{\sqrt{d}}\right)V \quad (10)$$

Concatenating results from all heads followed by fully connected feed-forward networks and layer normalization yields the final output. GA achieves efficient alignment and fusion of cross-modal information by actively retrieving visual content with question features as queries, or constraining text reasoning with visual features as queries.

Stacked Encoders and Decoders

To fully excavate hierarchical semantic associations of multimodal features, this paper adopts competitive encoder-decoder architectures to construct global and local understanding branches. This architecture gradually refines intra-modal high-order features and achieves deep cross-modal interaction through hierarchical stacking of attention units.

Global Understanding Encoder-Decoder: As illustrated in Fig 4(a), the global branch adopts a stacked structure combining pure self-attention and guided attention. The encoder consists of L layers of SA units, processing visual features V and textual features T layer by layer:

$$X^{(l)} = \text{SA}\left(X^{(l-1)}\right) \quad (11)$$

The encoder captures potential semantic dependencies through multi-layer self-interaction, gradually establishing global semantic representations. The decoder similarly contains L layers, each composed of SA units cascaded with GA units, implementing cross-modal bidirectional guidance:

$$X^{(l)} = \text{GA}\left(\text{SA}\left(X^{(l-1)}\right), Y^{(L)}\right) \quad (12)$$

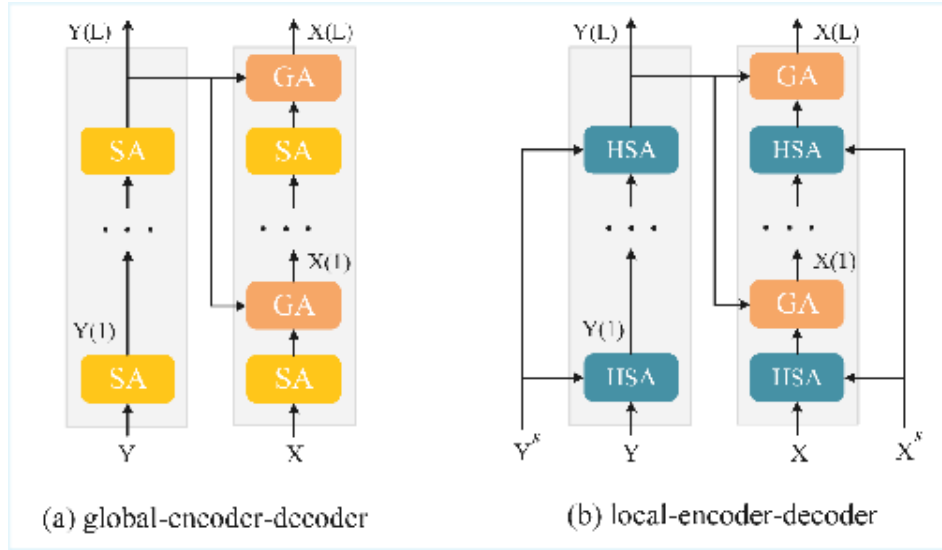


Fig. 4. Illustration of the encoder-decoder architectures in the Global-Local Perception Network.

where $X^{(L)}$ is the top-level feature output by the encoder. Through cross-guidance of GA units, the decoder maintains intra-modal consistency while achieving efficient alignment and fusion of visual-linguistic semantics.

Local Understanding Encoder-Decoder: As illustrated in Fig 4(b), the local branch replaces SA units with HSA units to continuously inject sparse structure priors. Each encoder layer receives parallel inputs of dynamic features and static structure features:

$$X^{(l)} = \text{HSA} \left(X^{(l-1)}, X^{str} \right) \quad (13)$$

where X^{str} represents sparse structural features generated by SSMA, encoding spatial positional relationships between visual regions and semantic dependency relationships between text words, respectively. The decoder adopts an alternating stacked structure of HSA and GA, maintaining structural constraints during cross-modal interaction:

$$X^{(l)} = \text{GA} \left(\text{HSA} \left(X^{(l-1)}, X^{str} \right), Y^{(L)} \right) \quad (14)$$

Notably, throughout the forward propagation process, structural features X^{str} remain constant in the network and are not associated with layer numbers. This design is based on the fact that structural information originates from inherent topological attributes of the original semantic space: visual end from physical spatial relationships determined by object bounding box geometric positions, textual end from grammatical associations defined by dependency syntax graphs—these relationships are statically determined by data prior structures and do not change with high-level feature abstract transformations. Therefore, structural features serve as cross-layer shared constant prior information, providing stable structural constraints for attention calculations at each layer, avoiding topological

drift during dynamic feature learning, and ensuring that the local understanding branch always follows original semantic associations for discriminative feature refinement.

3.3. Modal Fusion and Reasoning

After deep co-attention learning, the two branches adopt different feature compression strategies for answer reasoning: the global branch captures overall semantic profiles through adaptive attention pooling, while the local branch focuses on key local regions through explicit similarity calculation and low-temperature sharpening mechanisms.

Global Understanding Branch: This branch aims to encode the overall semantic layout of scenes. For output visual features $V^{(L)} \in \mathbb{R}^{m \times d}$ and textual features $T^{(L)} \in \mathbb{R}^{n \times d}$, MLP-based attention compression models are employed to aggregate them into global vectors. Taking the visual end as an example, importance weights for each region are first calculated: $a = \text{Softmax}(\text{MLP}(X^{(L)}))$, $a \in \mathbb{R}^m$, where the MLP structure is $\text{FC}(d) \rightarrow \text{ReLU} \rightarrow \text{Dropout}(0.1) \rightarrow \text{FC}(1)$. Subsequently, global visual representations are generated through soft weighted summation:

$$\bar{x}_g = \sum_{i=1}^m \alpha_i x_i^{(L)} \quad (15)$$

This mechanism adaptively aggregates multi-region information, suitable for comprehensive reasoning dependent on holistic scene cognition. Similarly, global textual representations $\bar{y}_g$ are obtained. Finally, answer distributions are generated through linear fusion and classification heads:

$$Z_g = \text{LayerNorm}(W_x^\top \bar{x}_g + W_y^\top \bar{y}_g), \quad \text{score}_g = \sigma(W_p^\top Z_g) \quad (16)$$

Local Understanding Branch: Unlike the distributed aggregation of the global branch, the local branch aims to precisely anchor a few key regions highly relevant to the question through explicit semantic similarity screening, suppressing background noise. First, textual features are compressed into holistic representations $\bar{y}_l$ through MLP attention models. Subsequently, cosine similarity with each visual region feature $v_i^{(L)}$ is calculated to measure semantic relevance:

$$s_i = \frac{\bar{y}_l^\top v_i^{(L)}}{|\bar{y}_l| |v_i^{(L)}|}, \quad i = 1, 2, \dots, m \quad (17)$$

To achieve sharpened selection of a few highly relevant regions, a temperature coefficient τ is introduced for low-temperature Softmax scaling :

$$\alpha_i = \frac{\exp(s_i/\tau)}{\sum_{j=1}^m \exp(s_j/\tau)} \quad (18)$$

When $\tau \rightarrow 0$, this distribution approaches one-hot form, retaining only regions most relevant to the question, achieving a soft approximation of hard selection. Final aggregated local visual representation: We set $\tau = 0.10$ as the default temperature, which achieves a favorable balance between selective precision and semantic completeness.

$$\bar{x}_l = \sum_{i=1}^m \alpha_i x_i^{(L)} \quad (19)$$

Finally, answer distributions are generated through linear fusion and classification heads:

$$Z_l = \text{LayerNorm} (W_x^\top \bar{x}_l + W_y^\top \bar{y}_l), \quad \text{score}_l = \sigma (W_p^\top Z_l) \quad (20)$$

We employ binary cross-entropy loss functions to optimize the distribution of predicted answer scores:

$$\mathcal{L} = - \sum_{j=1}^N [t_j \log (\text{score}_j) + (1 - t_j) \log (1 - \text{score}_j)] \quad (21)$$

Our strategy involves independently training classifiers on two understanding branches from different perspectives. Through arithmetic averaging operations, the model achieves adaptive integration of dual-branch predictions without introducing additional fusion parameters or complex gating networks, reducing overfitting risks while improving inference robustness. The answer corresponding to maximum confidence in $\text{score}_{joint} = \text{score}_g + \text{score}_l$ is selected as the system prediction output.

4. Experiments

4.1. Datasets

Unlike research paradigms relying on large-scale pretrained models to extract generic representations, this paper adopts end-to-end training strategies, directly optimizing visual-linguistic feature alignment on VQA-specific data distributions. This strategy offers three advantages: First, feature extractors jointly optimized for the VQA task can more precisely capture question-oriented visual semantic associations, avoiding domain gaps between features learned by pretrained models on generic corpora and downstream task requirements. Second, end-to-end training enables the model to adaptively learn dataset-specific statistical characteristics, reducing dependence on language priors that may exist in pretraining corpora. Finally, this paradigm enables lightweight architecture designs, achieving competitive performance without loading massive pretrained weights. To ensure fair comparison and validate model effectiveness, this paper selects two mainstream benchmarks: VQA 2.0 and GQA.

VQA 2.0 [11] and GQA [16] are employed as the primary evaluation benchmarks. Derived from MS COCO, VQA 2.0 comprises 1.1 million natural language questions with standard training, validation, and test splits. Its complementary image-pair design mitigates language priors by enforcing image-grounded reasoning, and questions are categorized as “Yes/No”, “Number”, or “Other”. Accuracy is reported on the official test-dev and test-standard sets.

Built upon Visual Genome scene graphs, GQA encompasses 113K images and 22 million structured questions targeting compositional reasoning and relational understanding. The dataset emphasizes multi-step inference and alleviates language bias through

Table 1. Hyperparameter settings for DGL-Net training and inference.

Hyperparameter	Value	Description
Hidden dimension d	512	Model embedding size
Attention heads h	8	Parallel self-attention heads
Head dimension	64	d/h
Encoder/Decoder layers L	6	Stacked layer depth
FFN structure	FC($4d$) $\rightarrow$ ReLU Dropout(0.1) $\rightarrow$ FC(d)	$\rightarrow$ Feed-forward network
MLP compression	FC(d) $\rightarrow$ ReLU Dropout(0.1) $\rightarrow$ FC(1)	$\rightarrow$ Attention reduction
Visual IoU threshold τ_v	0.4	Sparse adjacency construction
Visual regions N	36–100	Dynamic, confidence-thresholded
Text parser	HanLP (Biaffine)	Semantic dependency parsing
Optimizer	AdamW	$\beta_1 = 0.9, \beta_2 = 0.999$
Weight decay	10^{-4}	L2 regularization
Batch size	64	Training samples per step
Loss function	Binary cross-entropy	Multi-label answer classification
Warmup epochs	4	Linear LR warmup
Initial learning rate	2.5×10^{-5}	Warmup starting value
Peak learning rate	1×10^{-4}	Post-warmup maximum
Early stopping	3 epochs	Patience on validation accuracy
Gradient clipping	0.25	Max gradient norm threshold
Dropout rate	0.1	Applied to all linear projections

balanced answer distributions. Beyond standard accuracy, GQA evaluates Consistency, Plausibility, Validity, and Distribution to comprehensively measure reasoning robustness. We utilize the balanced training and validation sets for model development, with all metrics reported on the test set.

4.2. Experimental Setup

All experiments are conducted on NVIDIA L40 24GB GPUs. The complete hyperparameter configurations are summarized in Table 1, ensuring full reproducibility of our results. The semantic dependency trees are parsed using the HanLP toolkit pretrained on SemEval-2015 Task 18 English corpora. Visual region quantities are dynamically adjusted based on Faster R-CNN detection confidence scores.

To ensure reproducibility and fair comparison, the two branches follow a strictly decoupled training protocol. Both branches share the same data loader and batch order to guarantee identical training distributions, yet maintain completely separate optimizer states and parameter gradients. Specifically, each branch is assigned an independent AdamW optimizer with identical initial hyperparameters including peak learning rate, warmup schedule, and weight decay, but their gradient updates are computed and applied in isolation. Backpropagation from the global branch loss never reaches local branch parameters, and vice versa. The two branches are trained in parallel on separate GPUs for

the same number of epochs, with early stopping determined independently based on each branch’s validation accuracy. After both branches converge, their parameters are frozen and only the arithmetic averaging operation is applied during joint inference. This protocol ensures that each branch optimizes its specific objective without gradient interference, while maintaining strictly comparable training conditions with single branch baselines.

4.3. Model Performance Comparison

To comprehensively validate the effectiveness of the proposed method, we systematically compare DGL-Net and its single-branch variants DGL-Global and DGL-Local with representative methods in the visual question answering field in recent years, with experimental results shown in Table 2. All comparison experiments are conducted on the official test sets of VQA 2.0 dataset (test-dev and test-standard) to ensure evaluation fairness and reproducibility.

Table 2. Performance comparison on the test-dev and test-std splits.

Method	Image Feature	Test-dev (%)				Test-std
		Yes/No	Number	Other	All	All
MCAN(’19)[47]	Region	87.08	53.16	60.66	70.69	-
CAAN(’22)[3]	Region	87.09	53.37	61.13	70.94	71.31
CLVIN(’23)[6]	Region	87.09	53.65	60.89	70.86	71.16
RRAF(’24)[35]	Region	87.03	55.39	60.95	71.06	71.33
VMAN(’24)[37]	Region	86.95	53.28	61.16	70.89	71.34
MPCCT(’24)[4]	Region	86.96	55.57	61.27	71.20	-
$RMCB_G$ (’25)[5]	Region	87.26	55.21	61.34	71.32	-
VWAN(’24)[37]	Grid	87.63	53.85	61.89	71.58	-
LCRN(’24)[14]	Grid	87.74	53.99	62.29	71.83	72.12
GFSNet(’25)[33]	Grid	87.78	54.91	61.94	71.79	71.99
DGL-Global	Region	87.46	53.95	60.96	71.08	-
DGL-Local	Region	87.60	53.49	60.80	71.01	-
DGL-Net	Region	87.99	54.42	61.75	71.73	72.15

Comparison with single-branch baselines shows that DGL-Net significantly outperforms single global or local branches through independent training and joint inference strategies. Specifically, compared with DGL-Global and DGL-Local, DGL-Net brings absolute performance improvements of 0.65% and 0.72% on test-dev, respectively, validating the effectiveness of the dual-branch decoupled collaborative mechanism. This improvement primarily stems from the complementary fusion of scene contextual constraints provided by the global branch and fine-grained region localization capabilities contributed by the local branch. Performance comparison with recent methods shows that DGL-Net achieves 72.15% accuracy on test-standard, outperforming compared contemporary methods including LCRN and GFSNet based on Grid features. Notably, both

LCRN and GFSNet rely on grid-based visual features, while DGL-Net achieves comparable generalization performance using only regional features, indicating that the proposed decoupled collaborative paradigm has advantages in feature utilization efficiency. On test-dev, DGL-Net achieves 71.73% accuracy, demonstrating strong competitiveness among region-based methods. Specifically, compared with recently published region-based methods RMCBG and VMAN, DGL-Net improves by 0.41% and 0.84%, respectively. Particularly, on “Other” categories representing open-ended questions requiring fine-grained reasoning, DGL-Net achieves 61.75%, surpassing RMCBG, indicating that the local understanding branch’s modeling capability for fine-grained semantic associations significantly enhances the model’s open-domain reasoning performance. Comprehensive comparison shows that as an end-to-end trained region-based method, DGL-Net’s performance approaches current state-of-the-art grid-based methods. Considering the advantages of region features in computational efficiency and interpretability, this method demonstrates strong competitiveness in the trade-off between accuracy and practicality. Furthermore, DGL-Net achieves competitive scores on both “Yes/No” and “Number” categories, indicating that the decoupled architecture can effectively handle heterogeneous question types, validating the model’s robustness under different reasoning modes.

To further validate the model’s compositional reasoning capability and generalization, supplementary experiments are conducted on the GQA benchmark. As shown in Table 3, DGL-Net demonstrates advantages on multiple evaluation metrics, particularly in open-ended questions requiring complex reasoning and logical consistency.

Table 3. Performance comparison on multiple evaluation metrics.

Method	Binary	Open	Consistency	Validity	Plausibility	Distribution	Accuracy
Human[16]	87.12	90.40	99.40	99.20	97.60	-	89.30
BAN[18]	73.31	41.13	84.64	96.77	85.58	-	56.19
DMFNet[49]	73.98	41.86	86.98	97.62	84.87	-	57.05
LXMERT[39]	75.33	41.89	91.00	95.36	83.5	6.42	60.83
MAGM[43]	70.33	40.65	88.05	96.17	84.22	6.43	57.57
SGCN[44]	68.20	32.80	83.49	95.90	84.10	8.00	54.56
SelfGraphVQA[38]	68.20	32.80	64.30	95.00	90.00	-	54.5
LION[7]	-	-	-	-	-	-	51.56
DGL-Global	76.23	40.84	87.25	96.49	84.94	1.65	57.08
DGL-Local	76.01	40.90	87.16	96.37	84.88	1.7	57.01
DGL-Net	76.56	42.01	87.55	97.12	85.72	1.82	57.87

DGL-Net achieves 42.01% on the Open metric. It is worth noting that LXMERT[33] and DMFNet [46] employ large-scale vision-language pretraining followed by task-specific fine-tuning, whereas DGL-Net is trained end-to-end from scratch on VQA-specific data. The comparison herein serves as a reference rather than direct competition: it demonstrates that the proposed decoupled architecture, despite using significantly fewer pre-training resources, achieves competitive compositional reasoning performance through

Table 4. Comparison of parameter count (Params), computational complexity (FLOPs), and inference speed of DGL-Net and its single-branch variants on the VQA 2.0 test-dev set.

Models	Params (M)	FLOPs (G)	Time (ms/img)	All (%)
baseline	51.61	3.67	2894	70.75
DGL-Global	52.58	3.81	2931	71.08
DGL-Local	75.70	7.09	3803	71.01
DGL-Net	-	-	-	71.73

explicit global-local collaborative mechanisms. This highlights the efficacy of structural decoupling as a lightweight alternative to massive pretraining for fine-grained reasoning tasks. Meanwhile, the model achieves leading performance on Validity and Plausibility, with the global branch’s overall scene constraints and local branch’s precise semantic localization synergistically reducing logical conflicts during reasoning, validating the advantages of the dual-branch architecture in generating logically reasonable and semantically credible answers. DGL-Net achieves 57.87% accuracy through end-to-end training strategies, demonstrating excellent performance among non-pretrained methods. Compared with single-branch baselines, DGL-Net improves by 0.79% and 0.86% over DGL-Global and DGL-Local, respectively, while the Consistency metric increases from approximately 87.2% to 87.55%, further corroborating the effectiveness of the decoupled collaborative training strategy in improving model reasoning consistency. These results indicate that even without dependence on large-scale pretraining, the proposed decoupled architecture can still achieve competitive compositional reasoning performance through explicit global-local collaborative mechanisms.

Table 4 compares parameter counts, computational complexity, and inference speed of the single-branch baselines. DGL-Net is not independently listed because its inference mechanism introduces no additional fusion parameters or gating networks; the final output derives from simple arithmetic averaging of the two independently trained branch predictions. Thus, the overall complexity is fully determined by the two decoupled branches without extra overhead from joint training or fusion modules. The global branch incurs modest computational increase relative to the baseline, while the local branch demands higher resources owing to its structural sparse attention and hybrid self-attention components. Despite the dual-branch architecture, the collaborative inference achieves absolute improvements of 0.65% and 0.72% over the global and local branches respectively, representing a favorable accuracy-efficiency trade-off. Relative to large-scale pretrained methods, DGL-Net maintains a significantly lighter footprint, underscoring the practical viability of structural decoupling for resource-constrained deployment.

4.4. Ablation Studies

To deeply analyze the contribution of each module to model performance, systematic ablation experiments are conducted on VQA 2.0 test-dev, with results shown in Table 5.

Module Ablation Experiments

Table 5. Ablation study results on the test-dev split.

Methods	Test-dev			
	Yes/No	Number	Other	All
DGL-Global (GED+AR)	87.46	53.95	60.96	71.08
DGL-Global w/ AAR	87.30	53.14	60.79	70.85
DGL-Local (LED+AAR)	87.60	53.49	60.80	71.01
DGL-Coupled (Multi-task)	87.54	53.77	60.67	70.95
DGL-Net	87.99	54.42	61.75	71.73

First, experimental settings are explained: DGL-Global (GED+AR) represents the global understanding branch using only global encoder-decoder and MLP attention compression; DGL-Global w/AAR represents the variant where alignment attention compression from the local understanding branch is incorrectly applied to the global architecture; DGL-Local (LED+AAR) represents the local understanding branch using only local encoder-decoder and alignment attention compression; DGL-Coupled represents the coupled paradigm where dual branches adopt multi-task joint training; DGL-Net represents the complete decoupled collaborative network with independent training and joint inference proposed in this paper.

Using only the global branch or local branch alone achieves competitive baseline performance, but both achieve only 60.96% and 60.80% on “Other” open-ended questions, respectively, with overall accuracy failing to break through 71.1%. This indicates inherent bottlenecks in single-dimensional reasoning capabilities: while the global branch excels at capturing holistic scene semantics, it lacks precise localization of fine-grained regions; while the local branch can screen key regions, it may be constrained by neglecting overall context.

To validate the compatibility between attention compression mechanisms and network architectures, we apply the local branch’s alignment attention compression to the global architecture. Results show that this mismatched scheme performance drops to 70.85%, a decrease of 0.23% compared with DGL-Global, with particularly significant losses in the “Number” category, dropping from 53.95% to 53.14%. This indicates that cosine similarity-based sharpening selection mechanisms are more suitable for local reasoning requiring precise region screening, while MLP-based adaptive pooling better satisfies global scene summarization needs, validating the importance of component adaptation in divide-and-conquer strategies.

Comparison of dual-branch training paradigms shows that DGL-Coupled is significantly lower than DGL-Net, with a performance gap of 0.78%. This result indicates that simple parameter sharing or multi-task optimization leads to mutual interference between the two branches during gradient updates, making it difficult to balance heterogeneous optimization objectives for global and local features. In contrast, DGL-Net’s independent

training strategy enables each branch to focus on capability learning in specific dimensions during optimization, achieving complementary advantages through late fusion and effectively avoiding optimization conflicts.

Comprehensive results show that DGL-Net improves by 0.79% over the best single branch in the “Other” category and 0.47% in the “Number” category, validating the gains of the dual-branch collaborative mechanism for complex reasoning tasks. Particularly in open-ended questions requiring fine-grained attribute recognition, the complementary fusion of global and local representations significantly enhances the model’s discriminative capability. These ablation results fully demonstrate the rationality and effectiveness of each component design in the proposed decoupled global-local collaborative architecture.

Table 6. Threshold ablation.

Parameter	Test-dev			
	Yes/No	Number	Other	All
$\tau_v=0.2$	87.48	53.29	60.92	71.00
$\tau_v=0.4$	87.60	53.49	60.80	71.01
$\tau_v=0.6$	87.47	53.65	60.61	70.89
$\tau_v=0.8$	86.77	53.99	60.85	70.75

Threshold Selection Ablation

To validate the rationality of the region association discrimination threshold in visual sparse adjacency matrix construction, ablation experiments are conducted on VQA 2.0 test-dev, with results shown in Table 6.

Experimental results show that the model achieves optimal overall performance when $\tau_v = 0.4$. When $\tau_v = 0.2$, the threshold is too low, causing the sparse matrix to contain too many weakly associated regions, introducing structural noise and degrading overall performance. Conversely, when $\tau_v = 0.6$ or $\tau_v = 0.8$, the threshold is too high, causing excessive sparsity that destroys local structure integrity, resulting in decreased overall accuracy. Notably, the counting category shows a monotonically increasing trend with threshold elevation. This indicates that for counting problems requiring precise enumeration of specific objects, high thresholds can effectively filter out core regions highly relevant to query targets, suppressing duplicate counting risks brought by overlapping regions or background interference. However, this high sparsity comes at the cost of overall structure perception: when $\tau_v = 0.8$, “Yes/No” category performance significantly decreases, and the “Other” category shows obvious decline at the 0.6 threshold, indicating that extreme sparsity damages the model’s reasoning capability for complex semantic relationships.

Comprehensive analysis shows that $\tau_v = 0.4$ maintains appropriate structural sparsity while preserving necessary inter-region semantic associations, achieving optimal balance across “Yes/No”, “Number”, and “Other” question types. Therefore, this paper fixes

Table 7. Sensitivity analysis of the temperature coefficient on VQA 2.0 test-dev.

τ	Yes/No	Number	Other	All
0.01	87.41	53.66	60.61	70.95
0.10	87.60	53.49	60.80	71.01
0.50	87.36	53.37	60.68	70.88
1.00	86.87	53.14	60.72	70.79

$\tau_v = 0.4$ as the default threshold for visual sparse structure construction in subsequent experiments.

Sensitivity Analysis of Temperature Coefficient τ

The temperature coefficient τ modulates the sharpness of the local Softmax distribution in the region screening module. Its empirical impact is systematically assessed on the VQA 2.0 test-dev split by varying τ over $\{0.01, 0.1, 0.5, 1.0\}$ while all other hyperparameters remain at their defaults. The results are reported in Table 7.

The results indicate a non-monotonic response of the local branch to τ , with $\tau = 0.10$ yielding the optimal overall accuracy of 71.01%. At $\tau = 0.01$, aggressive sharpening suppresses duplicate background counts, marginally improving the “Number” metric to 53.66% while degrading “Yes/No” to 87.41% and “Other” to 60.61% due to excessive contextual truncation. As τ increases beyond 0.10, the Softmax distribution progressively flattens: “Yes/No” drops monotonically to 86.87% at $\tau = 1.00$, while “Other” converges toward the single-branch baseline, indicating loss of discriminative screening. Accordingly, we adopt $\tau = 0.10$ as the default, achieving the best compromise between selective precision and context preservation.

4.5. Visualization Analysis

To intuitively demonstrate the reasoning mechanism differences between the two branches, this paper conducts visualization analysis on cross-modal attention weights generated by the final decoder layer of both branches. Fig 5 shows attention heatmap comparisons between Global Awareness (global understanding branch) and Local Awareness (local understanding branch) in typical samples, revealing significant division of labor between the two in semantic attention levels.

For the first image with question “Is the man wearing sunglasses?”, the global understanding branch’s attention mainly concentrates on the salient subject “man”, with heatmaps showing high response values uniformly distributed across region rows and word columns representing this target, reflecting its macroscopic grasp of scene dominant objects. However, this holistic perception mode cannot precisely anchor the fine-grained attribute “sunglasses”, causing key details to be submerged in strong responses from the subject object. In contrast, the local understanding branch focuses high weights on specific regions corresponding to “sunglasses” and the semantic association of “wearing” through sparse structure constraints and alignment attention mechanisms, showing selective enhancement of local details, validating its localization precision in fine-grained

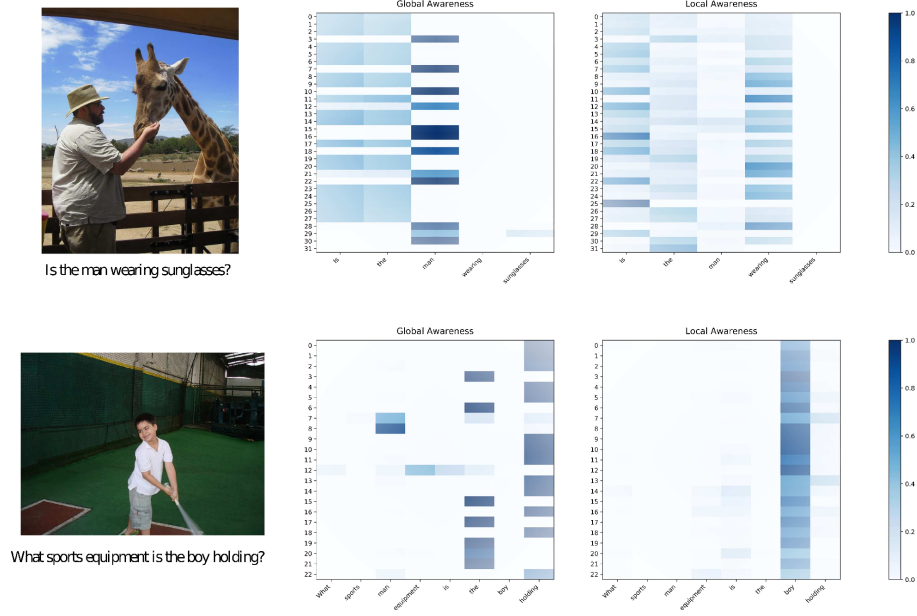


Fig. 5. Visualization of attention heatmaps.

attribute recognition. For the second image with question “What sports equipment is the boy holding?”, both branches show more significant semantic hierarchy differences. The global understanding branch shows strongest response to the verb “holding”, focusing on overall scene semantics constituted by action relationships; while the local understanding branch simultaneously focuses on two noun entities “boy” and “equipment”, supporting answer inference through recognition of specific objects and their attributes. This difference reflects the complementarity between the global branch’s summarization capability for high-order semantic relationships such as actions and scene atmospheres, and the local branch’s precise localization capability for basic visual concepts such as object categories and spatial attributes.

Attention visualization results indicate that global and local understanding branches indeed learn heterogeneous attention allocation strategies: the former tends to capture dominant semantics and relationship structures in overall scenes, while the latter focuses on fine-grained regions and entity concepts directly related to queries. Collaborative reasoning achieved through late fusion effectively compensates for limitations of single perception modes, providing more robust decision foundations for complex visual question answering tasks.

Furthermore, to validate the collaborative mechanism, qualitative analysis is conducted on diverse question sets for the same image. As shown in Fig 6, joint reasoning achieves averaging fusion of dual-branch predictions: when one branch has high confidence in the correct answer, its advantage can effectively mask the near-random low-confidence distribution produced by the other branch due to understanding failure,

thereby synthesizing dual-branch understanding capabilities. Specifically, for fine-grained questions such as “Is the man wearing sunglasses?”, the local branch has explicit high-confidence predictions for correct answers, while although the global branch produces incorrect judgments, its error distribution has relatively low confidence; after averaging, correct answers still dominate. Conversely, for global reasoning questions such as “Does the giraffe appear dangerous?”, the global branch’s high confidence in “No” successfully suppresses the local branch’s erroneous low-confidence responses caused by excessive attention to local textures. This “high-confidence dominance, low-confidence dilution” fusion mechanism enables joint reasoning to stably inherit advantages from single branches across different question types.

	Question	Joint Awareness	Global Awareness	Local Awareness
	1) Is the man wearing sunglasses?	No	Yes	No
	2) What animal is this?	Giraffe	Giraffe	Giraffe
	3) What color are the spots on the animal?	Brown	Brown	Brown
	4) Does the giraffe appear dangerous?	No	No	Yes
	5) What color is the man's shirt?	White	White	White
	Question	Joint Awareness	Global Awareness	Local Awareness
	1) Is this picture in color?	Yes	Yes	Yes
	2) What sports equipment is the boy holding?	Bat	Bat	Tennis racket
	3) What do you call this sports location?	Tennis court	Tennis court	Tennis court
	4) What is the name of this sport?	Baseball	Baseball	Tennis
	5) What is the boy wearing?	Shorts	Shorts	Shorts

Fig. 6. Qualitative comparison of the joint model and its decoupled branches.

However, when both branches produce errors or when original visual features have inherent ambiguities, joint reasoning cannot overcome cognitive limitations. For example, for “What do you call this sports location?”, all three branches incorrectly judge it as “Tennis court”, indicating that underlying feature extraction fails to distinguish visual similarities between baseball fields and tennis courts. At this point, regardless of fusion strategy, correct semantics cannot be recovered from ambiguous original representations. Failure cases reveal the current method’s dependence on visual feature quality. In the future, more powerful visual backbone networks can be introduced or visual feature adaptive enhancement mechanisms can be explored, dynamically adjusting feature extraction granularity at the encoder front-end according to question types rather than relying on fixed region proposals, to further enhance model reasoning robustness.

5. Conclusion

This paper proposes a Decoupled Global-Local Collaborative Network for Visual Question Answering. Addressing the structural conflict between global semantics and local

fine-grained perception in VQA, it achieves decoupled modeling of cross-modal reasoning through explicit dual-branch architectures. This method breaks through the limitations of traditional single-path attention mechanisms that trade off global and local objectives within single parameter spaces, constructing independently optimized global understanding and local understanding branches: the former captures holistic scene semantics through hierarchical attention fusion, while the latter achieves precise localization of key regions based on sparse structural multi-head attention and alignment compression mechanisms. The two branches train independently to avoid gradient interference, with late fusion achieving complementary advantages. Distinguished from existing methods that indirectly balance global and local information through multi-task learning within shared parameter spaces, DGL-Net fundamentally eliminates optimization conflicts through physical separation of parameter spaces. Experimental results demonstrate that this paradigm achieves competitive performance on VQA 2.0 and GQA benchmarks, with particular advantages in fine-grained reasoning tasks.

However, the proposed method still exhibits two intrinsic limitations. First, the static arithmetic averaging fusion ignores the dynamic correlation between branch predictive uncertainties. When one branch produces high-entropy distributions due to ambiguous inputs, the fixed fusion weight cannot adaptively suppress noisy contributions, rendering the model prone to erroneous consensus. Second, the reliance on pre-extracted visual features limits the model's generalization capability. When both branches simultaneously err owing to inherently ambiguous visual features, the late-fusion mechanism cannot recover correct semantics from corrupted representations. In light of these limitations, future research will focus on two core directions. First, addressing the static fusion bottleneck, we will explore uncertainty-driven adaptive fusion paradigms by establishing dynamic evaluation mechanisms for branch confidence, enabling fusion weights to be adaptively regulated according to input reliability, thereby transcending the robustness ceiling of fixed fusion strategies. Second, addressing the fixed backbone dependency, we will advance toward end-to-end multi-scale visual representation learning, investigating mechanisms that dynamically adjust feature extraction granularity and receptive field scope in accordance with reasoning task characteristics, so as to reduce reliance on pretrained detectors while enhancing the generalization and discriminability of visual inputs.

References

1. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., Zhang, L.: Bottom-up and top-down attention for image captioning and visual question answering. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 6077–6086. Computer Vision Foundation / IEEE Computer Society (2018), http://openaccess.thecvf.com/content_cvpr_2018/html/Anderson_Bottom-Up_and_Top-Down_CVPR_2018_paper.html
2. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C.L., Parikh, D.: VQA: visual question answering. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015. pp. 2425–2433. IEEE Computer Society (2015), <https://doi.org/10.1109/ICCV.2015.279>
3. Chen, C., Han, D., Chang, C.: CAAN: context-aware attention network for visual question answering. *Pattern Recognit.* 132, 108980 (2022), <https://doi.org/10.1016/j.patcog.2022.108980>

4. Chen, C., Han, D., Chang, C.: MPCCT: multimodal vision-language learning paradigm with context-based compact transformer. *Pattern Recognit.* 147, 110084 (2024), <https://doi.org/10.1016/j.patcog.2023.110084>
5. Chen, C., Han, D., Guo, Z., Chang, C.: Towards bias-aware visual question answering: Rectifying and mitigating comprehension biases. *Expert Syst. Appl.* 264, 125817 (2025), <https://doi.org/10.1016/j.eswa.2024.125817>
6. Chen, C., Han, D., Shen, X.: CLVIN: complete language-vision interaction network for visual question answering. *Knowl. Based Syst.* 275, 110706 (2023), <https://doi.org/10.1016/j.knosys.2023.110706>
7. Chen, G., Shen, L., Shao, R., Deng, X., Nie, L.: LION: Empowering multimodal large language model with dual-level visual knowledge. In: *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. pp. 26530–26540. IEEE (2024), <https://doi.org/10.1109/CVPR52733.2024.02506>
8. Chen, L., Zheng, Y., Xiao, J.: Rethinking data augmentation for robust visual question answering. In: Avidan, S., Brostow, G.J., Cissé, M., Farinella, G.M., Hassner, T. (eds.) *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*. pp. 95–112. *Lecture Notes in Computer Science*, Springer (2022), https://doi.org/10.1007/978-3-031-20059-5_6
9. Chowdhury, S., Soni, B.: Handling language prior and compositional reasoning issues in visual question answering system. *Neurocomputing* 635, 129906 (2025), <https://doi.org/10.1016/j.neucom.2025.129906>
10. Eslami, S., Meinel, C., de Melo, G.: Pubmedclip: How much does CLIP benefit visual question answering in the medical domain? In: Vlachos, A., Augenstein, I. (eds.) *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*. pp. 1151–1163. *Findings of ACL, Association for Computational Linguistics* (2023), <https://doi.org/10.18653/v1/2023.findings-eacl.88>
11. Goyal, Y., Khot, T., Agrawal, A., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA matter: Elevating the role of image understanding in visual question answering. *Int. J. Comput. Vis.* 127(4), 398–414 (2019), <https://doi.org/10.1007/s11263-018-1116-0>
12. Guo, Y., Nie, L., Cheng, Z., Tian, Q., Zhang, M.: Loss re-scaling VQA: revisiting the language prior problem from a class-imbalance view. *IEEE Trans. Image Process.* 31, 227–238 (2022), <https://doi.org/10.1109/TIP.2021.3128322>
13. Guo, Z., Liu, M., Ji, Z., Bai, J., Zhang, L., Zuo, W.: Decoupled visual interpretation and linguistic reasoning for math problem solving. *CoRR abs/2505.17609* (2025), <https://doi.org/10.48550/arXiv.2505.17609>
14. Han, D., Shi, J., Zhao, J., Wu, H., Zhou, Y., Li, L., Khan, M.K., Li, K.: LRCN: layer-residual co-attention networks for visual question answering. *Expert Syst. Appl.* 263, 125658 (2025), <https://doi.org/10.1016/j.eswa.2024.125658>
15. He, H., Choi, J.D.: The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. pp. 5555–5577. *Association for Computational Linguistics* (2021), <https://doi.org/10.18653/v1/2021.emnlp-main.451>
16. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. pp. 6700–6709. *Computer Vision Foundation / IEEE* (2019), http://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_paper.html

17. Kil, J., Zhang, C., Xuan, D., Chao, W.: Discovering the unknown knowns: Turning implicit knowledge in the dataset into explicit training examples for visual question answering. In: Moens, M., Huang, X., Specia, L., Yih, S.W. (eds.) *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*. pp. 6346–6361. Association for Computational Linguistics (2021), <https://doi.org/10.18653/v1/2021.emnlp-main.512>
18. Kim, J., Jun, J., Zhang, B.: Bilinear attention networks. In: Bengio, S., Wallach, H.M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*. pp. 1571–1581 (2018), <https://proceedings.neurips.cc/paper/2018/hash/96ea64f3a1aa2fd00c72faacf0cb8ac9-Abstract.html>
19. Kim, J., Lee, S., Kwak, D., Heo, M., Kim, J., Ha, J., Zhang, B.: Multimodal residual learning for visual QA. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. pp. 361–369 (2016), <https://proceedings.neurips.cc/paper/2016/hash/9b04d152845ec0a378394003c96da594-Abstract.html>
20. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L., Shamma, D.A., Bernstein, M.S., Fei-Fei, L.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* 123(1), 32–73 (2017)
21. Li, L.H., Yatskar, M., Yin, D., Hsieh, C., Chang, K.: Visualbert: A simple and performant baseline for vision and language. *CoRR abs/1908.03557* (2019), <http://arxiv.org/abs/1908.03557>
22. Liu, W., Wang, J., Wang, H., Li, R., Qiu, Y., Zhang, Y., Han, J., Zou, Y.: Decoupled rationalization with asymmetric learning rates: A flexible lipschitz restraint. In: Singh, A.K., Sun, Y., Akoglu, L., Gunopulos, D., Yan, X., Kumar, R., Ozcan, F., Ye, J. (eds.) *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, KDD 2023, Long Beach, CA, USA, August 6-10, 2023*. pp. 1535–1547. ACM (2023), <https://doi.org/10.1145/3580305.3599299>
23. Liu, Y., Wei, W., Peng, D., Zhu, F.: Declaration-based prompt tuning for visual question answering. In: Raedt, L.D. (ed.) *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*. pp. 3264–3270. *ijcai.org* (2022), <https://doi.org/10.24963/ijcai.2022/453>
24. Lu, J., Yang, J., Batra, D., Parikh, D.: Hierarchical question-image co-attention for visual question answering. In: Lee, D.D., Sugiyama, M., von Luxburg, U., Guyon, I., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*. pp. 289–297 (2016), <https://proceedings.neurips.cc/paper/2016/hash/9dcb88e0137649590b755372b040afad-Abstract.html>
25. Mishra, A., Anand, A., Guha, P.: Dual attention and question categorization-based visual question answering. *IEEE Trans. Artif. Intell.* 4(1), 81–91 (2023), <https://doi.org/10.1109/TAI.2022.3160418>
26. Niu, Y., Tang, K., Zhang, H., Lu, Z., Hua, X., Wen, J.: Counterfactual VQA: A cause-effect look at language bias. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. pp. 12700–12710. Computer Vision Foundation / IEEE (2021), https://openaccess.thecvf.com/content/CVPR2021/html/Niu_Counterfactual_VQA_A_Cause-Effect_Look_at_Language_Bias_CVPR_2021_paper.html
27. Peng, D., Wei, W., Mao, X., Fu, Y., Chen, D.: An empirical study on the language modal in visual question answering. In: *Proceedings of the Thirty-Second International Joint Conference*

- on Artificial Intelligence, IJCAI 2023, 19th-25th August 2023, Macao, SAR, China. pp. 4109–4117. *ijcai.org* (2023), <https://doi.org/10.24963/ijcai.2023/457>
28. Peng, L., Yang, Y., Wang, Z., Huang, Z., Shen, H.T.: Mra-net: Improving VQA via multi-modal relation attention network. *IEEE Trans. Pattern Anal. Mach. Intell.* 44(1), 318–329 (2022), <https://doi.org/10.1109/TPAMI.2020.3004830>
 29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*. pp. 8748–8763. *Proceedings of Machine Learning Research, PMLR* (2021), <http://proceedings.mlr.press/v139/radford21a.html>
 30. Ren, L., Liu, Y., Ouyang, C., Yu, Y., Zhou, S., He, Y., Wan, Y.: Dylas: A dynamic label alignment strategy for large-scale multi-label text classification. *Inf. Fusion* 120, 103081 (2025), <https://doi.org/10.1016/j.inffus.2025.103081>
 31. Sharma, H., Jalal, A.S.: An improved attention and hybrid optimization technique for visual question answering. *Neural Process. Lett.* 54(1), 709–730 (2022), <https://doi.org/10.1007/s11063-021-10655-y>
 32. Shen, X., Chen, C., Han, D., Xu, Y., Wang, X., Zhou, H.: A triple-branch hybrid dynamic-static alignment strategy for vision-language tasks. *Neural Networks* 191, 107871 (2025), <https://doi.org/10.1016/j.neunet.2025.107871>
 33. Shen, X., Han, D., Chang, C., Oad, A., Wu, H.: Gfsnet: Gaussian fourier with sparse attention network for visual question answering. *Artif. Intell. Rev.* 58(6), 159 (2025), <https://doi.org/10.1007/s10462-025-11163-4>
 34. Shen, X., Han, D., Chang, C., Xu, Y., Chen, C.: Multimodal context-aware consistency alignment for vision-language tasks. *Expert Syst. Appl.* 295, 128857 (2026), <https://doi.org/10.1016/j.eswa.2025.128857>
 35. Shen, X., Han, D., Zong, L., Guo, Z., Hua, J.: Relational reasoning and adaptive fusion for visual question answering. *Appl. Intell.* 54(6), 5062–5080 (2024), <https://doi.org/10.1007/s10489-024-05437-7>
 36. Sobo, A., Mubarak, A., Baimagambetov, A., Polatidis, N.: Evaluating llms for code generation in HRI: A comparative study of chatgpt, gemini, and claude. *Appl. Artif. Intell.* 39(1) (2025), <https://doi.org/10.1080/08839514.2024.2439610>
 37. Song, X., Han, D., Chen, C., Shen, X., Wu, H.: Vman: visual-modified attention network for multimodal paradigms. *Vis. Comput.* 41(4), 2737–2754 (2025), <https://doi.org/10.1007/s00371-024-03563-4>
 38. Souza, B., Aasan, M., Pedrini, H., Rivera, A.R.: Selfgraphvqa: A self-supervised graph neural network for scene-based question answering. In: *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*. pp. 4642–4647. *IEEE* (2023), <https://doi.org/10.1109/ICCVW60793.2023.00499>
 39. Tan, H., Bansal, M.: LXMERT: learning cross-modality encoder representations from transformers. In: Inui, K., Jiang, J., Ng, V., Wan, X. (eds.) *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*. pp. 5099–5110. *Association for Computational Linguistics* (2019), <https://doi.org/10.18653/v1/D19-1514>
 40. Tan, S., Zhang, Y., Li, Y., Shin, B.: Alignnet: spatiotemporal alignment and multi-scale feature fusion for enhanced lidar semantic segmentation. *Appl. Intell.* 56(4), 119 (2026), <https://doi.org/10.1007/s10489-025-07021-z>
 41. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Guyon, I., von Luxburg, U., Bengio, S., Wallach, H.M., Fergus, R., Vishwanathan, S.V.N., Garnett, R. (eds.) *Advances in Neural Information Processing Systems 30: Annual Conference on Neural*

- Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA. pp. 5998–6008 (2017), <https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html>
42. Xu, Y., Liu, G., Shen, X., Wang, X., Zhou, H.: Multi-granularity modal interaction and fusion framework for vision-language tasks. *Engineering Applications of Artificial Intelligence* 178, 114849 (2026), <https://www.sciencedirect.com/science/article/pii/S0952197626011310>
 43. Xu, Y., Zhang, L., Shen, X.: Multi-modal adaptive gated mechanism for visual question answering. *PLOS ONE* 18(6), 1–24 (06 2023), <https://doi.org/10.1371/journal.pone.0287557>
 44. Yang, Z., Qin, Z., Yu, J., Wan, T.: Prior visual relationship reasoning for visual question answering. In: *IEEE International Conference on Image Processing, ICIP 2020, Abu Dhabi, United Arab Emirates, October 25-28, 2020*. pp. 1411–1415. IEEE (2020), <https://doi.org/10.1109/ICIP40778.2020.9190771>
 45. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.J.: Stacked attention networks for image question answering. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*. pp. 21–29. IEEE Computer Society (2016), <https://doi.org/10.1109/CVPR.2016.10>
 46. Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K., Finn, C.: Gradient surgery for multi-task learning. In: *Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020)*, <https://proceedings.neurips.cc/paper/2020/hash/3fe78a8acf5fda99de95303940a2420c-Abstract.html>
 47. Yu, Z., Yu, J., Cui, Y., Tao, D., Tian, Q.: Deep modular co-attention networks for visual question answering. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. pp. 6281–6290. Computer Vision Foundation / IEEE (2019), http://openaccess.thecvf.com/content_CVPR_2019/html/Yu_Deep_Modular_Co-Attention_Networks_for_Visual_Question_Answering_CVPR_2019_paper.html
 48. Yu, Z., Yu, J., Xiang, C., Fan, J., Tao, D.: Beyond bilinear: Generalized multimodal factorized high-order pooling for visual question answering. *IEEE Trans. Neural Networks Learn. Syst.* 29(12), 5947–5959 (2018), <https://doi.org/10.1109/TNNLS.2018.2817340>
 49. Zhang, W., Yu, J., Zhao, W., Ran, C.: Dmrnet: Deep multimodal reasoning and fusion for visual question answering and explanation generation. *Inf. Fusion* 72, 70–79 (2021), <https://doi.org/10.1016/j.inffus.2021.02.006>

Ganlong Zhou is currently pursuing the M.S. degree with the School of Computer Science, Shanghai Maritime University, Shanghai, China. His research focuses on multi-modal cognitive intelligence, with particular emphasis on image-text matching, visual question answering, and cross-modal alignment mechanisms.

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, from 2006. His current research interests include cloud and outsourcing security, wireless communication security, network and information security, visual question answering. He is currently a Member of the IEEE.

Xiang Shen is currently pursuing his Ph.D. at Shanghai Maritime University and is also a joint Ph.D. student at The University of Sydney, supported by the China Scholarship Council (CSC). His research focuses on multimodal learning, with particular emphasis on visual question answering, multimodal fusion strategies, and the development of large-scale multimodal models. His current work also explores the integration of AI agents for adaptive perception and reasoning in complex environments, aiming to bridge the gap between visual understanding and natural language processing.

Xiaoqi Xin is a Ph.D. student at Shanghai Maritime University researching multimodal emotion recognition. Her research combines affective computing and multi-source signal fusion to improve emotion perception under realistic conditions.

Received: April 3, 2026; Accepted: June 14, 2026.