# Focal and Composed Vision-semantic Modeling for Visual Question Answering

Yudong Han[1], Yangyang Guo[1], Jianhua Yin[1]*, Meng Liu[2]*, Yupeng Hu[1], Liqiang Nie[1]
[1]Shandong University, Qingdao, China
[2]Shandong Jianzhu University, Jinan, China
hanyudong.sdu@gmail.com,guoyang.eric@gmail.com,jhyin@sdu.edu.cn
mengliu.sdu@gmail.com,huyupeng@sdu.edu.cn,nieliqiang@gmail.com

## ABSTRACT

Visual Question Answering (VQA) is a vital yet challenging task in the field of multimedia comprehension. In order to correctly answer questions about an image, a VQA model requires to sufficiently understand the visual scene, especially the vision-semantic reasonings between the two modalities. Traditional relation-based methods allow to encode the pairwise relations of objects to boost the VQA model performance. However, this simple strategy is deficient to exploit the abundant concepts expressed by the composition of diverse image objects, leading to sub-optimal performance. In this paper, we propose a focal and composed vision-semantic modeling method, which is a trainable end-to-end model, for better vision-semantic redundancy removal and compositionality modeling. Concretely, we first introduce the LENA cell, a plug-and-play reasoning module, which removes redundant semantic by a focal mechanism in the first step, followed by the vision-semantic compositionality modeling for better visual reasoning. We then incorporate the cell into a full LENA network, which progressively refines multimodal composed representations, and can be leveraged to infer the high-order vision-semantic in a multi-step learning way. Extensive experiments on two benchmark datasets, i.e., VQA v2 and VQA-CP v2, verify the superiority of our model as compared with several state-of-the-art baselines.

## CCS CONCEPTS

• **Computing methodologies** → **Computer vision**; **Scene understanding**.

## KEYWORDS

Visual Question Answering; Vision-semantic Redundancy; Vision-semantic Compositionality

**ACM Reference Format:**
Yudong Han, Yangyang Guo, Jianhua Yin, Meng Liu, Yupeng Hu, Liqiang Nie. 2021. Focal and Composed Vision-semantic Modeling for Visual Question Answering. In *Proceedings of the 29th ACM International Conference on Multimedia (MM '21), October 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3474085.3475609

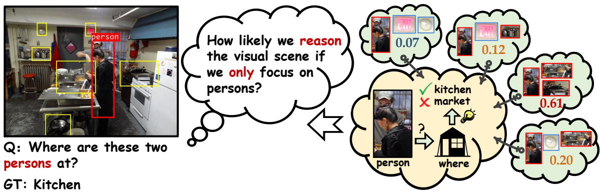
person
Q: Where are these two persons at?
GT: Kitchen
How likely we reason the visual scene if we only focus on persons?
0.07
0.12
0.61
0.20
✓ kitchen
✗ market
?
person
where

**Figure 1: Illustration of the vision-semantic compositionality. The key concept "person" can be combined with multiple objects based on the vision-semantic compositionality, which is beneficial to infer the correct answer - kitchen.**

## 1 INTRODUCTION

Visual Question Answering (VQA) has long been recognized as a fundamental task of multi-modal comprehension. It targets at correctly answering a natural language question about a given image, which has profound effect on various applications including visually impaired assistance and human-machine interaction. With the increasing prosperity of both computer vision and natural language processing communities, numerous well-designed multimodal methods [2, 3, 17, 18, 49] have been proposed to facilitate this task.

To obtain expressive image features, most existing models [2, 13, 27, 28, 53, 55] utilize the top-down soft attention mechanism, which assigns different attention weights to image regions according to their relevance with the given question. Nevertheless, these methods only focus on refining the superficial visual features, ignoring the sophisticated interactions between image objects. To tackle this, some relation-based models [7, 24, 31, 32, 47] have been presented recently, which leverage the intra-modality interactions to improve the question answering performance. Typically, they encode the object-pair relations by considering the semantic and/or spatial interactions between image objects, thereby enhance the visual feature learning.

Despite improved results on benchmark datasets have being observed, existing VQA methods show certain limitations in the following aspects: 1) **vision-semantic redundancy.** The soft attention mechanism is generally adopted to compute the inter-modality relevance between image objects and questions, where objects relevant to questions obtain more attention, otherwise less. However, some negligible objects, such as "clock" and "sign" in Figure 1, are actually inefficacious for question answering, would produce disturbance to some extent and lead to vision-semantic misunderstanding.

*Corresponding author.

*MM '21, October 20–24, 2021, Virtual Event, China*
© 2021 Association for Computing Machinery.
ACM ISBN 978-1-4503-8651-7/21/10...$15.00
https://doi.org/10.1145/3474085.3475609

2) **lack of vision-semantic compositionality.** Existing relation-based methods generally leverage object-pair relations to perform visual reasoning. However, simply encoding the pairwise relations of objects cannot sufficiently exploit the complex evidence composed by diverse objects. As the example illustrated in Figure 1, these models cannot accurately capture the vision-semantic compositionality expressed by "people", "cooker", and "hearth", which inevitably limits the model to efficaciously reason the correct answer "kitchen".

To settle the aforementioned issues, in this paper, we propose a novel foca**L** and compos**E**d visio**N**-sem**A**antic modeling network, dubbed as LENA, for better vision-semantic redundancy removal and compositionality modeling in VQA. Specifically, we first apply the bottom-up attention [2] to identify the salient objects, followed by our LENA cell equipped with two novel units. Concretely, a Focal Attention Unit (FAU) is leveraged to selectively remove redundant objects for subsequent object composing process, while a Composed Semantic-Aware Unit (CAU) is thereafter introduced to aggregate diverse objects into multiple composed embeddings, whereby we adaptively select the most relevant compound via the attention learning. Afterwards, we design an iterative reasoning process to progressively refine the high-order composed representations.

To verify the effectiveness of the proposed method, we conducted extensive experiments on two popular datasets, i.e., VQA v2 [12] and VQA-CP v2 [37]. Experimental results demonstrate that our proposed method can yield better performance as compared to a series of baseline models.

In summary, the contributions of this work are three-fold:

- We propose a focal and composed vision-semantic modeling network, i.e., LENA network, which iterates through the LENA cell to progressively refine high-order vision-semantic representations and reason over the visual scene with respect to the question.
- We introduce the LENA cell, where we elaborately devise two novel units. The former unit effectively removes the redundant semantic, followed by the latter one to exploit the high-order vision-semantic compositionality.
- We conducted extensive comparative experiments on two publicly available datasets to validate the effectiveness of vision-semantic redundancy removal and compositionality modeling. As a side contribution, we have released our code to benefit other researchers[1].

# 2 RELATED WORK

## 2.1 Visual Attention in VQA

The visual attention mechanism plays an indispensable part in current VQA methods, which is leveraged to learn the fine-grained visual features. Initially, the visual attention is introduced into VQA to assign distinctive weights to image regions according to the relevance with given questions [41]. To pinpoint the regions that are most indicative to the question, the multi-glimpse attention [49] is proposed to iteratively infer the answer by fusing the visual features with updated question features. Different from the above top-down visual attention methods, Anderson et al. [2] introduced a united bottom-up and top-down attention (UpDn) mechanism which firstly detects salient image objects and then leverages visual attention to attend on the crucial objects. This technique has been widely adopted and applied in recent studies [10, 34, 47], boosting the performance of a series of VQA models. For instance, BAN [20] adopts the bilinear attention between the detected image objects and question words to strengthen the fine-grained visual representations. Moreover, DFAF [10] introduces the self-attention mechanism into VQA, where the intra-modality interactions are effectively modeled.

## 2.2 Visual Relation Modeling in VQA

Recently, exploring the visual relations between objects has constituted a key research direction in VQA, and a series of well-devised relation-based models have been proposed [10, 24, 33, 34]. For instance, in [34], a graph learner is introduced to reason the relation representation through the pairwise attention modeling and spatial graph convolution. Analogously, DCN [33] and DFAF [10] leverage the self-attention mechanism to encode the pairwise relation, thereby enhance the visual feature learning. To further model multi-type inter-object interactions, ReGAT [24] considers both explicit and implicit relations to enrich visual representation. ODA was proposed in [39], which encodes the object-pair relations by implementing difference operator between regions. Moreover, MuRel [47] firstly merges spatial and semantic representations with the BLOCK fusion approach [6], and then employs the max operator for neighbours aggregation.

# 3 OUR PROPOSED METHOD

As shown in Figure 2, our proposed LENA model comprises of three components: Visual and Textual Encoder (Section 3.1), LENA Network (Section 3.2), and Answer Reasoning (Section 3.3). In the following, we elaborate these three components sequentially.

## 3.1 Visual and Textual Encoder

*3.1.1 Visual Encoder.* Following [7, 24], we extract image features using a pre-trained Faster-RCNN method [38]. Concretely, we first select $n$ bounding boxes with the highest class confidence scores for each image, where the obtained visual region feature and corresponding semantic indicator are respectively denoted as $\mathbf{f}_i \in \mathbb{R}^{d_v}$ and $\mathbf{h}_i \in \mathbb{R}^{d_c}$. Meanwhile, the position information of each image objects can be represented by $\mathbf{p}_i = (x_i, y_i, w_i, h_i)$, where $x_i$ and $y_i$ are top left coordinates, as well $w_i$ and $h_i$ denote width and height, respectively. To fully explore linguistic cues, we then introduce a semantic embedding matrix $\mathbf{W}_h \in \mathbb{R}^{d_e \times d_c}$, where each column represents the embedding of a semantic concept. Afterwards, based on the semantic indicator $\mathbf{h}_i$ and $\mathbf{W}_h$, we could derive the semantic embedding $\mathbf{e}_i$ for the $i$-th visual region as follows,

$$\mathbf{e}_i = \mathbf{W}_h \mathbf{h}_i. \tag{1}$$

Thereafter, we feed the concatenation of the semantic embedding $\mathbf{e}_i \in \mathbb{R}_e$ and the visual embedding $\mathbf{f}_i$ into a fully-connected layer, obtaining the final representation $\mathbf{v}_i \in \mathbb{R}^d$ of the $i$-th visual region,

$$\mathbf{v}_i = \mathbf{W}_v \ [\mathbf{f}_i; \mathbf{e}_i] + \mathbf{b}_v, \tag{2}$$

where $\mathbf{W}_v \in \mathbb{R}^{d \times (d_v + d_e)}$ and $\mathbf{b}_v \in \mathbb{R}^d$ are learnable parameters.

[1] https://github.com/Einstones/LeNa-Net

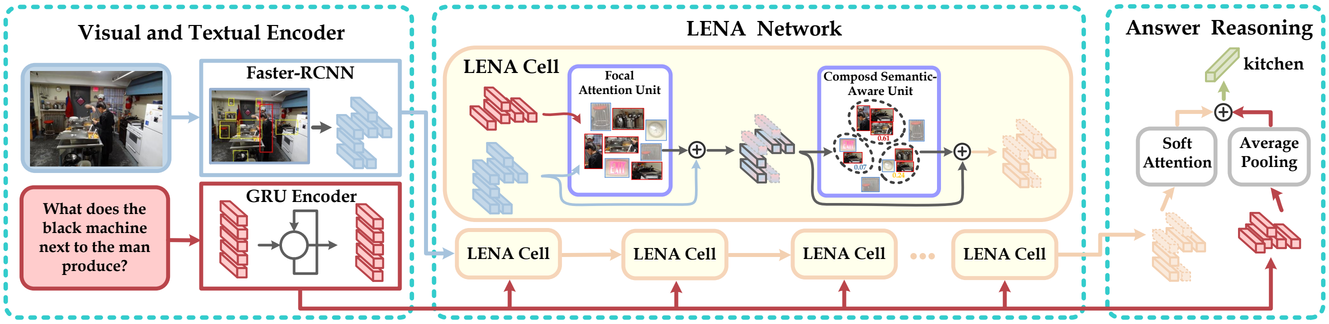

Figure 2: Schematic illustration of our proposed LENA model.

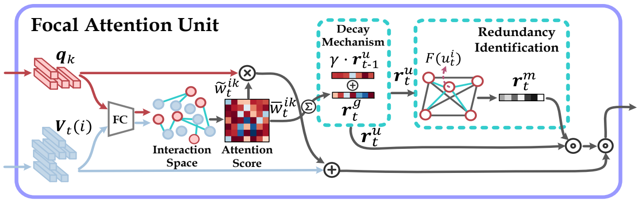

Figure 3: Focal Attention Unit.

*3.1.2 Texual Encoder.* For the question encoder, we adopt GRU [9] followed by the self-attention mechanism [45] to generate the context-aware embedding for each word, denoted as $\{\mathbf{q}_k\}_{k=1}^m$, where $m$ represents the length of the question and $\mathbf{q}_k \in \mathbb{R}^{d_q}$ $(d_q = d)$ is the representation of the $k$-th word.

## 3.2 LENA Network

We propose the LENA Network, a vision-semantic modeling architecture that iterates through a series of LENA cells. Each cell takes as input the semantic features $\{\mathbf{V}_t(i)\}_{i=1}^n$ and the word embeddings $\{\mathbf{q}_k\}_{k=1}^m$, where $\mathbf{V}_t(i) \in \mathbb{R}^d$ denotes initial representation of the $i$-th visual region in the $t$-th LENA cell, which is initialized as $\mathbf{v}_i$, i.e., $\mathbf{V}_0(i) = \mathbf{v}_i$. The output of the $t$-th cell is represented as $\{\mathbf{V}_t(i)''\}_{i=1}^n$.

*3.2.1 LENA Cell.* As shown in Figure 2, there are two key parts in the LENA cell: Focal Attention Unit and Composed Semantic-Aware Unit.

**Focal Attention Unit**. This unit aims to reduce the vision-semantic redundancy via the focal mechanism, including three steps: attention estimation, redundancy identification, and redundancy removal. The details are showed in Figure 3. Specifically, we first calculate the region-by-word attention scores,

$$w_t^{ik} = (\mathbf{W}_v \mathbf{V}_t(i))^\top (\mathbf{W}_q \mathbf{q}_k), \tag{3}$$

where $\mathbf{W}_q \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_v \in \mathbb{R}^{d\times d}$ are projection matrixes. Afterwards, we apply a softmax function over $w_t^{ik}$ by column and then sum the normalized scores $\bar{w}_t^{ik}$ by row, obtaining $\mathbf{r}_t^g = \left[g_t^1, ..., g_t^n\right] \in \mathbb{R}^{1\times n}$, where $g_t^i = \sum_{k=1}^m \bar{w}_t^{ik}$ represents the obtained attention of $\mathbf{V}_t(i)$ from the given question. To maintain a relatively long-term memory for attention estimation, we adopt a decay mechanism by considering its previous attention score $\mathbf{r}_{t-1}^u$ and the current one, i.e., $\mathbf{r}_t^u = \gamma \cdot \mathbf{r}_{t-1}^u + \mathbf{r}_t^g$, where $\gamma$ is a decay parameter and $\mathbf{r}_t^u = \left[u_t^1, ..., u_t^n\right] \in \mathbb{R}^{1\times n}$, which is initialized as $\mathbf{r}_0^u = \mathbf{0}$.

In the second step, to better identify the redundant semantic, we score each region based on its allocated attention from question and its relative attention to other regions,

$$F(u_t^i) = \frac{1}{n} \sum_{l=1, l\neq i}^{n} (u_t^i - u_t^l) \cdot \sqrt{u_t^l}, \tag{4}$$

where $\sqrt{u_t^i}$ represents the normalized attention and $(u_t^i - u_t^l)$ represents the mathematical difference, which determines the relative attention of the $l$-th region to the $i$-th region. The regions with scores smaller than zero are deemed as redundant regions, otherwise important, and a truncation function is adopted as $\mathbf{r}_t^m = \mathbb{I}(F(\mathbf{r}_t^u) > 0)$ to distinguish them, where $F(\mathbf{r}_t^u) = \left[F(u_t^1), \ldots, F(u_t^n)\right] \in \mathbb{R}^{1\times n}$, $\mathbf{r}_t^m \in \mathbb{R}^{1\times n}$, and $\mathbb{I}(x)$ returns 1 if $x$ is true, otherwise 0.

Then we regularize the attention by generated importance score

$$\tilde{\mathbf{r}}_t^u = \mathbf{r}_t^u \odot \mathbf{r}_t^m, \tag{5}$$

where $\odot$ denotes the element-wise multiplication, and $\mathbf{1} \in \mathbb{R}^n$ denotes the vector with all elements set to 1.

Finally, we remove the vision-semantic redundancy via $\tilde{\mathbf{r}}_t^u \in \mathbb{R}^{1\times n}$, which can be formulated as follows,

$$\tilde{\mathbf{V}}_t(i) = \tilde{u}_t^i \cdot (\sum_{k=1}^m \tilde{w}_t^{ik} \mathbf{q}_k + \mathbf{V}_t(i)), \tag{6}$$

where $\tilde{u}_t^i$ represents the $i$-th element of $\tilde{\mathbf{r}}_t^u$, and $\sum_{k=1}^m \tilde{w}_t^{ik} \mathbf{q}_k + \mathbf{V}_t(i)$ denotes the vision-semantic representation of the $i$-th region, as well $\tilde{w}_t^{ik}$ is obtained by performing softmax function over $w_t^{ik}$ by row. By this means, the FAU provides a distilled vision-semantic context for subsequent composing process.

**Composed Semantic-Aware Unit (CAU)**. To exploit the high-order vision-semantic compositionality, we forgo the traditional pairwise relation-based methods [10, 34, 47]. In this part, we introduce a composed semantic-aware unit, as shown in Figure 4. We first produce different compounds expressed by diverse image objects, where each compound is adaptively assigned a distinctive score by the attention learning. Then, we perform compounds aggregation to further refine the high-order composed semantic.

As the dilated convolution [51] can aggregate multi-scale contextual semantic by using different kernels, we design a dilated convolution layer with $K$ parallel kernels, as reported in Table 1.

**Table 1: Dilated convolution configurations, where $z^k$, $d^k$, and $p^k$ denote kernel size, dilation rate, and padding size, respectively.**

| kernel | $k_1$ | $k_2$ | $k_3$ | $k_4$ | $k_5$ | $k_6$ | $k_7$ |
|---|---|---|---|---|---|---|---|
| $z^k$ | 1 | 3 | 3 | 3 | 5 | 5 | 5 |
| $d^k$ | 1 | 1 | 2 | 3 | 1 | 2 | 3 |
| $p^k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 |

Specifically, for each vision-semantic representation $\tilde{\mathbf{V}}_t(i)$ derived from Eq. (6), we use $K$ kernels to produce $K$ compounds. We denote the $k$-th kernel as $\phi_k(\cdot)$, the output of $k$-th kernel is described as,

$$\tilde{\mathbf{V}}_t^k(i) = \phi_k(\tilde{\mathbf{V}}_t(i)). \tag{7}$$

where $\tilde{\mathbf{V}}_t^k(i)$ is uniquely determined by the $k$-th kernel and the $i$-th region, and we define $\vec{\mathbf{V}}_t(i) = \left[\tilde{\mathbf{V}}_t^1(i), ..., \tilde{\mathbf{V}}_t^K(i)\right] \in \mathbb{R}^{d\times K}$. Afterwards, a fully-connected layer followed by a softmax function is adopted to compute the importance score for each compound. In formula, we summarize this process as follows,

$$\begin{cases} s_t^k(i) = \mathbf{W}_s\, \tilde{\mathbf{V}}_t^k(i) + b_s, \\ \tilde{s}_t^k(i) = \frac{\exp(\beta \cdot s_t^k(i))}{\sum_{i=1}^K \exp(\beta \cdot s_t^k(i))}, \end{cases} \tag{8}$$

where $\mathbf{W}_s \in \mathbb{R}^{d\times 1}$ and $b_s \in \mathbb{R}$ are learnable weights and bias, respectively. We define $\tilde{\mathbf{S}}_t(i) = \left[\tilde{s}_t^1(i), ..., \tilde{s}_t^K(i)\right] \in \mathbb{R}^{1\times K}$, which is denoted as the corresponding scores vector. $\beta$ is the temperature parameter to increase the gap between the optimal compounds and the suboptimal ones.

Thereafter, a linear combination of $K$ compounds is computed by their corresponding scores as follows,

$$\mathbf{V}_t^{'}(i) = \vec{\mathbf{V}}_t(i)\, \tilde{\mathbf{S}}_t(i)^\top. \tag{9}$$

With Eq. (9), we could capture the most salient composed information related to the $i$-th region, which is summarized by $\mathbf{V}_t^{'}(i) \in \mathbb{R}^d$.

To further refine the composed vision-semantic representations, we conduct high-order aggregation to link up the similar compounds. Specifically, we merge spatial and semantic representations to build compounds interactions,

$$a_{ik,t} = \frac{a_{ik,t}^p \exp(a_{ik,t}^s)}{\sum_{k=1}^n a_{ik,t}^p \exp(a_{ik,t}^s)}, \tag{10}$$

where $a_{ik,t}^s \in \mathbb{R}$ and $a_{ik,t}^p \in \mathbb{R}$ measures the vision-semantic correlation and relative geometric correlation between the $i$-th and the $k$-th compound, respectively, which are computed as,

$$\begin{cases} a_{ik,t}^s = (\mathbf{W}_{s1}\mathbf{V}_t^{'}(i))^\top\, (\mathbf{W}_{s2}\mathbf{V}_t^{'}(k)), \\ a_{ik,t}^p = \max(0, \mathbf{W}_p\, f(\mathbf{p}_i, \mathbf{p}_k)), \end{cases} \tag{11}$$

where $\mathbf{W}_{s1} \in \mathbb{R}^{d\times d}$ and $\mathbf{W}_{s2} \in \mathbb{R}^{d\times d}$ are projection matrixes, $f(\cdot,\cdot)$ first computes a 4-dimensional relative geometry feature $(\log(\frac{|x_i - x_k|}{w_i}), \log(\frac{|y_i - y_k|}{h_i}), \log(\frac{w_k}{w_i}), \log(\frac{h_k}{h_i}))$ based on their position information, i.e., $\mathbf{p}_i$ and $\mathbf{p}_k$, and embeds it into a $d_p$-dimensionality feature by computing cosine and sine functions of different wavelengths [45], and then $\mathbf{W}_p$ transforms the feature $f(\mathbf{p}_i, \mathbf{p}_k)$ into a geometry attention map, which is further trimmed at 0. Finally, similar compounds are aggregated and mutually strengthened by $a_{ik,t}$,

$$\mathbf{V}_t^{''}(i) = \sum_{k=1}^n a_{ik,t}\, \mathbf{W}_{s3}\mathbf{V}_t^{'}(k) \tag{12}$$

where $\mathbf{W}_{s3} \in \mathbb{R}^{d\times d}$ is a projection matrix, similar to $\mathbf{W}_{s1}$ and $\mathbf{W}_{s2}$, $\mathbf{V}_t^{''}(i) \in \mathbb{R}^d$ represents the refined composed representation related to $i$-th region.

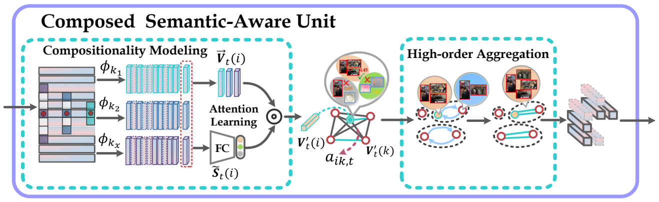

**Figure 4: Composed Semantic-Aware Unit.**

### 3.2.2 *LENA Network.*

The LENA network mimics a simple form of iterative reasoning by leveraging the power of the LENA cell to iteratively exploit the high-order semantic of visual scenes. As we can see from Figure 2, the vision-semantic representation is refined by the LENA cell through multiple steps. More specifically, for each step $t \in [1, T]$, where $T$ is the total number of steps, a LENA cell processes and updates the representation as follows,

$$\mathbf{V}_t(i) = \text{LENACell}(\mathbf{q}_k, \mathbf{V}_{t-1}(i)). \tag{13}$$

At each step $t$ ($t > 1$), the LENA cell reduces the redundant regions (objects or compounds) generated by last cell, and further exploits the high-order vision-semantic representations.

The residual nature of this module makes it possible to align multiple cells without being affected by gradient vanishing. Moreover, each cell has independent parameters due to that the cell in different depth tends to capture different levels of information, which enables to reveal deeper vision-semantic representations.

## 3.3 Answer Reasoning

To generate the holistic representation for the whole visual scene, we adopt the soft attention mechanism to refine $\mathbf{V}_T(i)$. For the question, we use average pooling to aggregate the word features into a sentence feature. We then adopt a simple additive fusion to obtain the final vision-semantic representation. We summarize the above process as follows,

$$\mathbf{y} = \delta(\mathbf{W}_y(\sum_{i=1}^n a_i \cdot \mathbf{V}_T(i) + \frac{1}{m}\sum_{k=1}^m \mathbf{q}_k)), \tag{14}$$

where $\delta(\cdot)$ is a sigmoid activation function, $a_i = softmax(\mathbf{W}_a \mathbf{V}_T(i))$, as well $\mathbf{W}_y$ and $\mathbf{W}_a$ are learnable parameters. Following [2], we use the binary cross-entropy loss function to train an answer classifier.

# 4 EXPERIMENTS

In this section, we conducted experiments to evaluate the performance of our proposed LENA model on two VQA benchmark datasets, i.e., VQA v2 [12] and VQA-CP v2 [37].

**Table 2: Performance comparison on the *val*, *test-dev*, and *test-std* splits of the VQA v2 dataset. For *val* results (the left column), all models are trained on the *train* split. For *test* results (the middle and the right column), all model are trained on the *train* + *val* splits. We directly quoted the results of these baselines from their original papers except the ones marked by “*”, which are obtained by running their released code. The best performance is highlighted in bold.**

| Methods | VQA v2 *val* (Acc. %) | | | | VQA v2 *test-dev* (Acc. %) | | | | VQA v2 *test-std* (Acc. %) | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | All | Y/N | Num. | Other | All | Y/N | Num. | Other | All | Y/N | Num. | Other |
| MLB [21] | 62.91 | - | - | - | 66.27 | 83.58 | 44.92 | 56.34 | 66.62 | - | - | - |
| MUTAN [5] | 63.61 | - | - | - | 66.01 | 82.88 | 44.54 | 56.50 | 66.38 | - | - | - |
| DCN [33] | 63.83 | - | - | - | 66.83 | 84.48 | 41.66 | 57.44 | 66.66 | 84.61 | 41.27 | 56.83 |
| DA-NTN [4] | - | - | - | - | 67.56 | 84.29 | 47.14 | 57.92 | 67.94 | - | - | - |
| BLOCK [6] | - | - | - | - | 67.58 | 83.60 | 47.33 | 58.51 | 67.92 | 83.98 | 46.77 | 58.79 |
| UpDn [2] | 63.15 | 80.07 | 42.87 | 55.81 | 65.32 | 81.82 | 44.21 | 56.05 | 65.67 | 82.20 | 43.90 | 56.26 |
| BAN [20]* | 65.18 | - | - | - | 68.16 | - | - | - | 68.38 | - | - | - |
| CGN [34] | - | - | - | - | 66.45 | - | - | - | 66.18 | 82.91 | 47.13 | 56.22 |
| CoR-3 [47]* | 64.92 | - | - | - | 68.19 | 84.96 | 47.11 | 58.64 | 68.59 | 85.16 | 47.19 | 59.07 |
| ODA [48]* | - | - | - | - | 68.17 | 84.66 | 48.04 | 58.68 | - | - | - | - |
| ReGAT [24] | 65.30 | - | - | - | - | - | - | - | - | - | - | - |
| CRA-Net [36] | 65.84 | 83.63 | 46.85 | 57.26 | 68.61 | 84.87 | 49.46 | 59.08 | 68.92 | 85.21 | 48.43 | 59.42 |
| MRA-Net [35] | 66.08 | - | - | - | 69.02 | 85.58 | 48.92 | 59.46 | 69.46 | 85.83 | 49.22 | 59.86 |
| SceneGCN [50] | - | - | - | - | 66.81 | 82.72 | 46.85 | 57.77 | 67.14 | 83.16 | 46.61 | 57.89 |
| TRN+UpDn [16] | - | - | - | - | 67.00 | 83.83 | 45.61 | 57.44 | 67.21 | - | - | - |
| VCTREE+HL [44] | 65.10 | 82.61 | 45.10 | 57.10 | 68.19 | 84.28 | 47.78 | 59.11 | 68.49 | 84.55 | 47.36 | 59.34 |
| **Ours (LENA)** | **66.59** | **84.35** | **48.71** | **57.79** | **69.39** | **85.87** | **49.97** | **59.52** | **69.70** | **85.95** | **49.90** | **59.87** |

**Table 3: Performance comparison on the *test-dev* and *test-std* splits of the VQA v2 dataset. All models are trained on the *train* + *val* splits with extra Visual Genome augmentation [23]. The symbol “*” refers to models that are pre-trained on the large corpus *Conceptual Captions* [40].**

| Methods | *test-dev* (Acc. %) | | | | *test-std* (Acc. %) |
|---|---|---|---|---|---|
| | All | Y/N | Num. | Other | All |
| MFH [54] | 68.76 | 84.27 | 49.56 | 59.89 | - |
| BAN [20] | 69.66 | 85.46 | 50.66 | 60.66 | - |
| Counter [55] | 68.09 | 83.14 | 51.62 | 58.97 | 68.41 |
| BAN+Counter [20] | 70.04 | 85.42 | 54.04 | 60.52 | 70.35 |
| MuRel [47] | 68.03 | 84.77 | 49.84 | 57.85 | 68.41 |
| Erase-Att [29] | 70.07 | 85.87 | 50.28 | 61.10 | 70.36 |
| MLIN [11] | 70.18 | 85.96 | 52.93 | 60.40 | 70.28 |
| DFAF [10] | 70.22 | 86.09 | 53.32 | 60.49 | 70.34 |
| MCAN [52] | **70.63** | **86.82** | 53.26 | **60.72** | **70.63** |
| **Ours (LENA)** | 70.31 | 86.63 | **54.26** | 60.22 | 70.48 |
| VL-BERT [43]* | 70.50 | - | - | - | 70.83 |
| ViLBERT [30]* | 70.55 | - | - | - | 70.92 |
| VisualBERT [25]* | **70.80** | - | - | - | **71.00** |
| **Ours (LENA)** | 70.31 | 86.63 | 54.26 | 60.22 | 70.48 |

## 4.1 Datasets and Evaluation Metric

*4.1.1 Datasets.* **VQA v2**: It is the most commonly used VQA benchmark dataset. The images are from the Microsoft COCO dataset [26], and each question has answers from ten different annotators. The answers are divided into three categories: *Yes/No*, *Number*, and *Other*. Besides, the dataset is split into *train*, *val*, and *test* (or *test-std*) splits, and 25% of the *test-std* set is reserved as the *test-dev* set. The ground truth answers are only available for the first two splits.

**VQA-CP v2**: This dataset is curated from the VQA v2 dataset, which is introduced to evaluate the question-oriented bias reduction capability in VQA models. Due to the significant difference of distribution between the *train* set and the *test* set, the VQA-CP v2 dataset is much more challenging than the VQA v2.

*4.1.2 Evaluation Metric.* We adopted the standard VQA accuracy metric for evaluation [3]. Given an image and a corresponding question, for a predicted answer $a$, the accuracy is computed as,

$$Acc_a = \min(1, \frac{\#\text{humans that provide } a}{3}). \tag{15}$$

Note that each question is answered by ten participants, and this metric takes the disagreement in human individuals into consideration.

## 4.2 Implementation Details

The hyper-parameters of our model used in the experiments are summarized as follows. For each image, we set a threshold for Faster RCNN to obtain $n = 36$ object features. All questions are padded or truncated to the same length $m = 14$. The size of the answer vocabulary is set to M=3,129, as used in [2]. The dimensionality of $d_v$, $d_q$, $d_e$, $d_c$, $d_p$, and $d$ are set to 2,048, 512, 300, 1600, 64, and 512, respectively. The number of the LENA cell is $T \in \{1, 2, 4, 6, 8, 10\}$ and we experimentally set $T$ to 8 in our best report. To reasonably control the redundancy removal, the decaying parameter $\gamma$ is set to 0.80 in our best record.

To train the LENA network, we use the Adam optimizer [22] with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The base learning rate is set to $\min(2.5te^{-5}, 1e^{-4})$, where $t$ is the current epoch number starting from 1. After 10 epochs, the learning rate is decayed by 1/4 every two epochs to $2.5te^{-5}$. The batch size is set to 64.

## 4.3 Performance Comparison

In this section, we compare the performance of the proposed method against state-of-the-art methods on two popular benchmark datasets, VQA v2 and VQA-CP v2. Table 2, Table 3, and Table 4 summarize the accuracy comparison results, where *Y/N*, *Num.*, *Other*, and *All* represent the accuracy of the yes/no questions, the counting questions, other kinds of questions, and all the questions, respectively.

Table 2 shows the results trained on the *train* (the left column) and *train+val* (the middle and right column) splits of the VQA v2 dataset. We can see that our LENA model achieves the best performance, substantially surpassing all the baselines. Specifically, UpDn [2] adopts the soft attention mechanism to leverage the salient image regions for question answering, and BAN [20] employs the bilinear attention to further enhance the fine-grained visual representations, both of which ignore the vision-semantic redundancy problem. By effectively performing vision-semantic redundancy removal, our LENA model outperforms UpDn and BAN by 4.03% and 1.32% on the *test-std* set, respectively. In addition, CGN [34] and ODA [48] explore object-pair relations to enhance the feature learning, which cannot sufficiently exploit the complex semantic composed by diverse objects. Nevertheless, our LENA model outperforms CGN and ODA by 2.94% and 1.22% on the *test-dev* set respectively, which demonstrates the superiority of our model in vision-semantic compositionality modeling.

Table 3 shows the results trained on the *train+val* splits of the VQA v2 dataset with extra Visual Genome augmentation [23]. We can observe that our LENA model performs better than the module-based model like MuRel [47], which iterates through multiple steps of a reasoning cell. This adequately shows the power of our LENA cell in high-order semantic modeling. Inspiringly, we are comparable to several formidable pre-trained models in the VQA task, which are first trained on the large corpus [40] and then fine-tuned on the VQA v2 dataset.

To demonstrate the generalizability of our LENA model, we also conducted experiments on the VQA-CP v2 dataset [14, 15]. The corresponding results are reported in Table 4. Agrawal *et al.* [1] curated the GVQA model to reduce the influence of language bias, although our proposed method is not specialized to tackle this problem, it still has 1.09% advantage over the GVQA model. In addition, DC-GCN [19] considers fine-grained pairwise interactions, where their dependency parsing mechanism can effectively reduce question-based overfitting. We observe that our LENA model provides a substantial gain over this strong baseline by 0.79%. In summary, the results on the VQA-CP v2 dataset show that our well-devised LENA cell enables our model to effectively perform redundancy removal and enhance the vision-semantic composed representation, whereby reduces the language bias to some extent.

**Table 4: Performance evaluated on the *test* split of the VQA-CP v2 dataset. Models with "*" are designed for solving the language bias problem. The best results are in bold.**

| Method | VQA-CP v2 *test* (Acc. %) | | | |
|---|---|---|---|---|
| | All | Y/N | Num. | Other |
| RAMAN [42] | 39.21 | - | - | - |
| BAN [20] | 39.31 | - | - | - |
| MuRel [47] | 39.54 | 42.85 | 13.17 | 45.05 |
| ReGAT-Sem [24] | 39.54 | - | - | - |
| ReGAT-Imp [24] | 39.58 | - | - | - |
| ReGAT-Spa [24] | 40.30 | - | - | - |
| ReGAT [24] | 40.42 | - | - | - |
| MRA-Net [35] | 40.45 | 44.53 | 13.05 | 45.83 |
| DC-GCN [19] | 41.47 | - | - | - |
| UpDn [2] | 39.74 | 42.70 | 11.93 | 46.05 |
| GVQA [1]* | 31.30 | 57.99 | 13.68 | 22.14 |
| UpDn+AdvReg. [37]* | 41.17 | 65.49 | **15.48** | 35.48 |
| **Ours (LENA, T=4)** | 41.51 | 65.93 | 12.31 | 46.59 |
| **Ours (LENA, T=6)** | **42.26** | **67.74** | 14.71 | **46.78** |
| **Ours (LENA, T=8)** | 41.62 | 66.12 | 12.64 | 46.62 |

**Table 5: Ablation studies of our proposed model on the *val* split of the VQA v2 dataset. The results of the full model are highlighted in bold.**

| Method | VQA v2 *val* (Acc. %) | | | |
|---|---|---|---|---|
| | All | Y/N | Num. | Other |
| **Ours (LENA)** | **66.59** | **84.35** | **48.71** | **57.79** |
| | (Focal Attention Unit) | | | |
| $w/o$ FAU | $65.07^{\downarrow 1.52}$ | 82.57 | 45.01 | 56.97 |
| $r.w.$ Soft Att. | $66.01^{\downarrow 0.58}$ | 83.62 | 47.21 | 57.38 |
| $r.w.$ Gating Att. | $66.18^{\downarrow 0.41}$ | 83.92 | 47.83 | 57.43 |
| $r.w.$ M-LSTM | $65.68^{\downarrow 0.91}$ | 83.09 | 46.62 | 57.28 |
| | (Composed Semantic-Aware Unit) | | | |
| $w/o$ CAU | $65.88^{\downarrow 0.71}$ | 83.81 | 46.81 | 57.35 |
| $r.w.$ GCN | $65.45^{\downarrow 1.14}$ | 82.98 | 45.76 | 56.21 |

## 4.4 Ablation Study

In this section, we design some ablation experiments to verify the effectiveness of FAU and CAU in our LENA model. For fair comparison, all the evaluated models have the same experimental settings.

**Focal Attention Unit.** To justify the effectiveness of FAU, we experimented with three variants: 1) ***w/o* FAU** refers to the LENA model without FAU; 2) ***r.w.* Gating Att.** replaces FAU with a Gating Attention Unit [8], which can be described as the following form: $\beta \cdot \mathbf{V}_t(i) + (1-\beta) \cdot \mathbf{q}_k$, where $\beta$ is a real-valued number parameterized by $\mathbf{V}_t(i)$ and $\mathbf{q}_k$; 3) ***r.w.* M-LSTM** replaces FAU with M-LSTM [8], a multimodal gating fusion function that is similar to LSTM, which is defined as $\mathbf{g} \odot \mathbf{V}_t(i) + (\mathbf{1} - \mathbf{g}) \odot (\mathbf{W}\left[\mathbf{V}_t(i); \mathbf{q}_k\right])$, where $\mathbf{g}$ is the gating vector parameterized by $\mathbf{V}_t(i)$ and $\mathbf{q}_k$, and $\mathbf{W}$ is the learnable parameter. The comparison results are summarized in Table 5. We could see that the performance of ***w/o* FAU** drops 1.42% compared with that of our full model. Meanwhile, our full model with FAU achieves substantially better performance than other two basic variants. These experiments verified the effectiveness

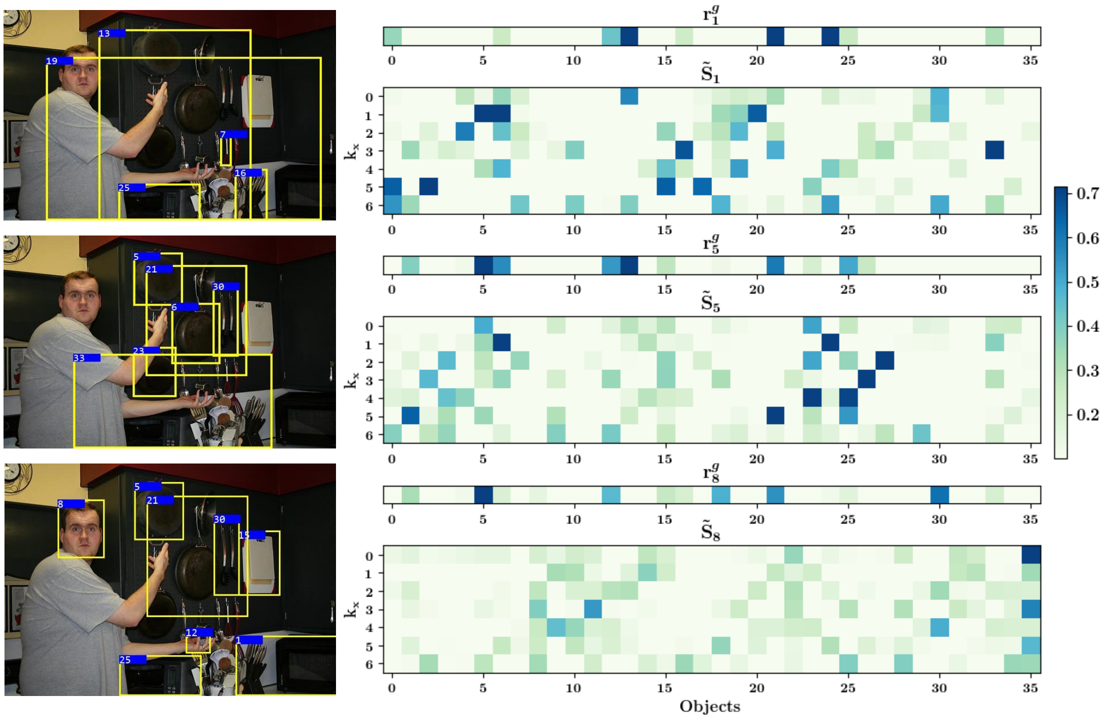

**Figure 5: Visualization of LENA-1, LENA-5, and LENA-8. The index within [0-35] shown on the horizontal axes of the attention maps corresponds to each object in the image, and the index within [0-6] represents different composed mode determined by corresponding convolutional kernels.**

of FAU, and these superior performance can be attributed to the effective pruning of the semantic redundancy achieved by the focal mechanism.

**Composed Semantic-Aware Unit.** To explore how CAU affects the results, we experimented with two variants: 1) ***w/o* CAU** refers to the LENA model without CAU; 2) ***r.w* GCN** replaces CAU with a Graph Convolutional Network [46] for pairwise relation modeling. By jointly analyzing the results in Table 5, we can observe that our LENA model displays consistent improvements over these two variants, which manifests that exploiting diverse composed semantic can provide more sophisticated evidence for question answering.

## 4.5 Visualization

To intuitively assess which regions (objects or compounds) are more faithful to model decisions, we utilized the visualization strategy similar to [20]. As shown in the rightmost of Figure 5, we visualized the learned attentions $\mathbf{r}_t^g \in \mathbb{R}^n$ and $\tilde{\mathbf{S}}_t \in \mathbb{R}^{K \times n}$ from three LENA cells, i.e., LENA-1, LENA-5, and LENA-8. Here, LENA-$t$ is the LENA cell at time step $t$, $\mathbf{r}_t^g$ represents the importance of each region in LENA-$t$, and $\tilde{\mathbf{S}}_t$ represents the different composed modes associated with regions. Particularly, in $\tilde{\mathbf{S}}_t$, each row denotes one composed mode determined by the corresponding convolution kernel, and each column corresponds to each region in the image. Given the form of $(x, y)$, $x$ represents the $x$-th composed mode, i.e., convolution kernel $k_x$, as shown in Table 1, and $y$ represents the $y$-th region.

Taking LENA-8 as an example, the large values in $\tilde{\mathbf{S}}_8$ occur on (3, 11), (4, 9), (4, 30), (6, 25), (3, 35), and (5, 35), which indicates that the model tends to concern these composed results. Particularly, different convolution kernels tend to capture different region sets. For example, (3, 11) tends to capture the region sets consisting of $\{8, 11, 14\}$, and (5, 25) tends to capture $\{21, 23, 25, 27, 29\}$. Thereafter, we merge all the region sets into one set, from which we further remove the redundant regions based on $\mathbf{r}_8^g$. Finally, we find out the important regions and mark them with yellow boxes, as shown in the leftmost of Figure 5.

## 4.6 Qualitative Results

We show the behavior of the focal attention unit and the composed semantic-aware unit in Figure 6. In the first sample, our focal attention unit first removes the redundant regions, such as "clock" in the top left corner of the image. The boxes with yellow color in the second image refer to the reserved regions. Following that, our model leverages the composed semantic-aware unit to exploit the abundant composed concepts. Here, we list three compounds with different confidence scores generated in LENA-8. Taking the compound with score 0.44 as an example, we observe that the localized regions are "cook", "people", "pat", and "hand", their composed representation provides a more comprehensive information for

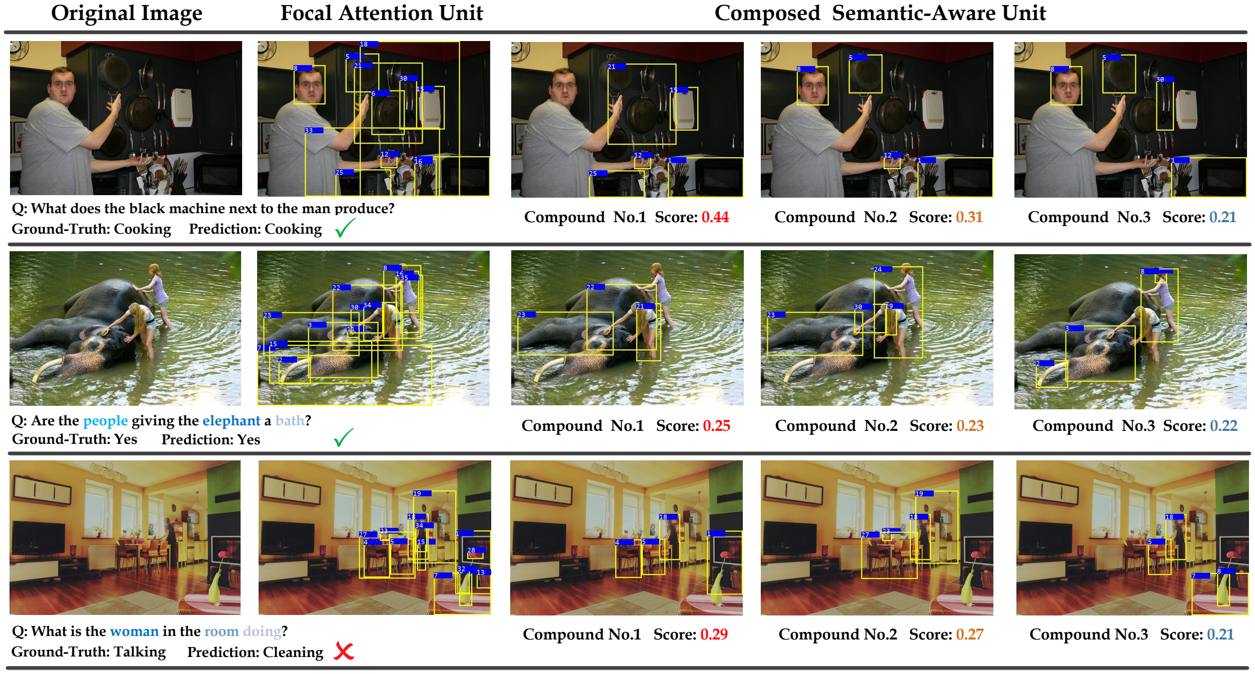
Original Image
Focal Attention Unit
Composed Semantic-Aware Unit
Q: What does the black machine next to the man produce?
Ground-Truth: Cooking Prediction: Cooking ✓
Compound No.1 Score: 0.44
Compound No.2 Score: 0.31
Compound No.3 Score: 0.21
Q: Are the people giving the elephant a bath?
Ground-Truth: Yes Prediction: Yes ✓
Compound No.1 Score: 0.25
Compound No.2 Score: 0.23
Compound No.3 Score: 0.22
Q: What is the woman in the room doing?
Ground-Truth: Talking Prediction: Cleaning ✗
Compound No.1 Score: 0.29
Compound No.2 Score: 0.27
Compound No.3 Score: 0.21

**Figure 6: Qualitative evaluation of our LENA model.**

answering “cooking”. Similarly, in the second sample, our LENA model first removes some irrelevant background regions. Then, the model captures the composed semantic, i.e., “people” + “ivory and water” + “hands and arms”, which plays a vital role for understanding the visual scene about bath. In the last sample, our LENA model predicts the wrong answer due to lacking of accurate detection on “woman” outside the window, which impedes the composing process between two women.

### 4.7 Parameter Sensitivity

We carried out experiments to explore how the number of LENA cells $T$ affects the model performance. Performance is reported on the *val* split of the VQA v2 dataset. In Figure 7, we train the six different LENA networks on the VQA v2 *train* split, each with a different number of LENA cells. Networks with six and eight LENA cells provide a gain of 1.07% and 1.37% in the overall accuracy over the network with one LENA cell, respectively. It indicates that understanding complex visual scene requires more incisive composed vision-semantic modeling. Moreover, with increasing $T$, the performance of our LENA network steadily improves and finally saturates at $T = 8$. The saturation can be explained by the unstable gradients during training when $T > 8$, which makes the optimization difficult.

## 5 CONCLUSION

In this paper, we present a focal and composed vision-semantic modeling network for VQA. It leverages a series of LENA cells to progressively perform vision-semantic redundancy removal and compositionality modeling for better visual reasoning. We exhibited various ablation studies, clearly demonstrating the gain of our focal mechanism to remove vision-semantic redundancy, the effectiveness of compositionality modeling, and the multi-step iterations in the whole process. Our final LENA network achieve very competitive performance with state-of-the-art methods on two popular used datasets.

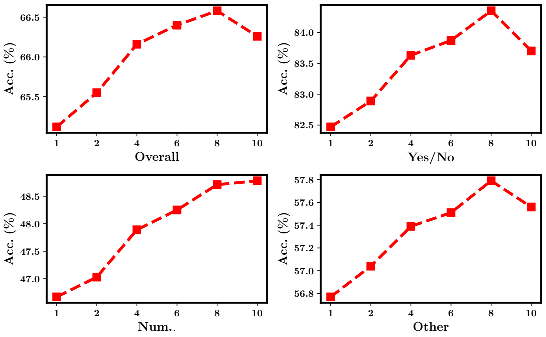

**Figure 7: Performance of our LENA network with different number of LENA cells $T$ on four types of questions.**

## 6 ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China, No.: 61802231; the Shandong Provincial Natural Science Foundation, No.: ZR2019QF001; the Shandong Provincial Natural Science Foundation, No.:ZR2019JQ23; the National Natural Science Foundation of China, No.: 62006142; Project of Qingdao Postdoctoral Applied Research.

# REFERENCES

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering. In *CVPR*. IEEE, 4971–4980.

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *CVPR*. IEEE, 6077–6086.

[3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*. IEEE, 2425–2433.

[4] Yalong Bai, Jianlong Fu, Tiejun Zhao, and Tao Mei. 2018. Deep Attention Neural Tensor Network for Visual Question Answering. In *ECCV*. Springer, 21–37.

[5] Hedi Ben-younes, Rémi Cadène, Matthieu Cord, and Nicolas Thome. 2017. MUTAN: Multimodal Tucker Fusion for Visual Question Answering. In *ICCV*. IEEE, 2631–2639.

[6] Hedi Ben-younes, Rémi Cadène, Nicolas Thome, and Matthieu Cord. 2019. BLOCK: Bilinear Superdiagonal Fusion for Visual Question Answering and Visual Relationship Detection. In *AAAI*. AAAI Press, 8102–8109.

[7] Rémi Cadène, Hedi Ben-younes, Matthieu Cord, and Nicolas Thome. 2019. MUREL: Multimodal Relational Reasoning for Visual Question Answering. In *CVPR*. IEEE, 1989–1998.

[8] Hui Chen, Guiguang Ding, Xudong Liu, Zijia Lin, Ji Liu, and Jungong Han. 2020. IMRAM: Iterative Matching With Recurrent Attention Memory for Cross-Modal Image-Text Retrieval. In *CVPR*. IEEE, 12652–12660.

[9] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*. ACL, 1724–1734.

[10] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic Fusion With Intra- and Inter-Modality Attention Flow for Visual Question Answering. In *CVPR*. IEEE, 6639–6648.

[11] Peng Gao, Haoxuan You, Zhanpeng Zhang, Xiaogang Wang, and Hongsheng Li. 2019. Multi-Modality Latent Interaction Network for Visual Question Answering. In *ICCV*. IEEE, 5824–5834.

[12] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *CVPR*. IEEE, 6325–6334.

[13] Wenya Guo, Ying Zhang, Xiaoping Wu, Jufeng Yang, Xiangrui Cai, and Xiaojie Yuan. 2020. Re-Attention for Visual Question Answering. In *AAAI*. AAAI Press, 91–98.

[14] Yangyang Guo, Zhiyong Cheng, Liqiang Nie, Yibing Liu, Yinglong Wang, and Mohan Kankanhalli. 2019. Quantifying and alleviating the language prior problem in visual question answering. In *SIGIR*. ACM, 75–84.

[15] Yangyang Guo, Liqiang Nie, Zhiyong Cheng, Feng Ji, Ji Zhang, and Alberto Del Bimbo. 2021. AdaVQA: Overcoming language priors with adapted margin cosine loss. In *IJCAI*. ACM.

[16] Xinzhe Han, Shuhui Wang, Chi Su, Weigang Zhang, Qingming Huang, and Qi Tian. 2020. Interpretable Visual Reasoning via Probabilistic Formulation Under Natural Supervision. In *ECCV*. Springer, 553–570.

[17] Yupeng Hu, Meng Liu, Xiaobin Su, Zan Gao, and Liqiang Nie. 2021. Video Moment Localization via Deep Cross-modal Hashing. *IEEE Transactions on Image Processing* 30 (2021), 4667–4677.

[18] Yupeng Hu, Liqiang Nie, Meng Liu, Kun Wang, Yinglong Wanga, and Xiansheng Hua. 2021. Coarse-to-Fine Semantic Alignment for Cross-modal Moment Localization. *IEEE Transactions on Image Processing* 30 (2021), 5933–5943.

[19] Qingbao Huang, Jielong Wei, Yi Cai, Changmeng Zheng, Junying Chen, Ho-fung Leung, and Qing Li. 2020. Aligned Dual Channel Graph Convolutional Network for Visual Question Answering. In *ACL*. ACL, 7166–7176.

[20] Jin-Hwa Kim, Jaehyun Jun, and Byoung-Tak Zhang. 2018. Bilinear Attention Networks. In *NIPS*. MIT Press, 1571–1581.

[21] Jin-Hwa Kim, Kyoung Woon On, Woosang Lim, Jeonghee Kim, Jung-Woo Ha, and Byoung-Tak Zhang. 2017. Hadamard Product for Low-rank Bilinear Pooling. In *ICLR*.

[22] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*. 1–15.

[23] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *IJCV* 123 (2017), 32–73.

[24] Linjie Li, Zhe Gan, Yu Cheng, and Jingjing Liu. 2019. Relation-Aware Graph Attention Network for Visual Question Answering. In *ICCV*. IEEE, 10312–10321.

[25] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A Simple and Performant Baseline for Vision and Language. arXiv:1908.03557

[26] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*. Springer, 740–755.

[27] M. Liu, L. Nie, X. Wang, Q. Tian, and B. Chen. 2019. Online Data Organizer: Micro-Video Categorization by Structure-Guided Multimodal Dictionary Learning. *IEEE Transactions on Image Processing* 28, 3 (2019), 1235–1247.

[28] Meng Liu, Xiang Wang, Liqiang Nie, Qi Tian, Baoquan Chen, and Tat-Seng Chua. 2018. Cross-Modal Moment Localization in Videos. In *MM*. ACM, 843–851.

[29] Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving Referring Expression Grounding With Cross-Modal Attention-Guided Erasing. In *CVPR*. IEEE, 1950–1959.

[30] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. In *NIPS*. MIT Press, 13–23.

[31] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. 2016. Hierarchical Question-Image Co-Attention for Visual Question Answering. In *NIPS*. MIT Press, 289–297.

[32] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual Attention Networks for Multimodal Reasoning and Matching. In *CVPR*. IEEE, 2156–2164.

[33] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved Fusion of Visual and Language Representations by Dense Symmetric Co-Attention for Visual Question Answering. In *CVPR*. IEEE, 6087–6096.

[34] Will Norcliffe-Brown, Stathis Vafeias, and Sarah Parisot. 2018. Learning Conditioned Graph Structures for Interpretable Visual Question Answering. In *NIPS*. MIT Press, 8344–8353.

[35] Liang Peng, Yang Yang, Zheng Wang, Zi Huang, and Heng Tao Shen. 2020. MRA-Net: Improving VQA via Multi-modal Relation Attention Network. *TPAMI* (2020), 1–1.

[36] Liang Peng, Yang Yang, Zheng Wang, Xiao Wu, and Zi Huang. 2019. CRA-Net: Composed Relation Attention Network for Visual Question Answering. In *MM*. ACM, 1202–1210.

[37] Sainandan Ramakrishnan, Aishwarya Agrawal, and Stefan Lee. 2018. Overcoming Language Priors in Visual Question Answering with Adversarial Regularization. In *NIPS*. MIT Press, 1548–1558.

[38] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In *NIPS*. MIT Press, 91–99.

[39] Adam Santoro, David Raposo, David G. T. Barrett, Mateusz Malinowski, Razvan Pascanu, Peter W. Battaglia, and Tim Lillicrap. 2017. A simple neural network module for relational reasoning. In *NIPS*. MIT Press, 4967–4976.

[40] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *ACL*. ACL.

[41] Kevin J. Shih, Saurabh Singh, and Derek Hoiem. 2016. Where to Look: Focus Regions for Visual Question Answering. In *CVPR*. IEEE, 4613–4621.

[42] Robik Shrestha, Kushal Kafle, and Christopher Kanan. 2019. Answer Them All! Toward Universal Visual Question Answering Models. In *CVPR*. IEEE, 10472–10481.

[43] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of Generic Visual-Linguistic Representations. In *ICLR*.

[44] Kaihua Tang, Hanwang Zhang, Baoyuan Wu, Wenhan Luo, and Wei Liu. 2019. Learning to Compose Dynamic Tree Structures for Visual Contexts. In *CVPR*. IEEE, 6619–6628.

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NIPS*. MIT Press, 5998–6008.

[46] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. 2018. Graph Attention Networks. In *ICLR*.

[47] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. 2018. Chain of Reasoning for Visual Question Answering. In *NIPS*. MIT Press, 273–283.

[48] Chenfei Wu, Jinlai Liu, Xiaojie Wang, and Xuan Dong. 2018. Object-Difference Attention: A Simple Relational Attention for Visual Question Answering. In *MM*. ACM, 519–527.

[49] Zichao Yang, Xiaodong He, Jianfeng Gao, Li Deng, and Alexander J. Smola. 2016. Stacked Attention Networks for Image Question Answering. In *CVPR*. IEEE, 21–29.

[50] Zhuoqian Yang, Zengchang Qin, Jing Yu, and Yue Hu. 2019. Scene Graph Reasoning with Prior Visual Relationship for Visual Question Answering. arXiv:1812.09681

[51] Fisher Yu and Vladlen Koltun. 2016. Multi-Scale Context Aggregation by Dilated Convolutions. In *ICLR*.

[52] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. 2019. Deep Modular Co-Attention Networks for Visual Question Answering. In *CVPR*. IEEE, 6281–6290.

[53] Zhou Yu, Jun Yu, Jianping Fan, and Dacheng Tao. 2017. Multi-modal Factorized Bilinear Pooling with Co-attention Learning for Visual Question Answering. In *ICCV*. IEEE, 1839–1848.

[54] Zhou Yu, Jun Yu, Chenchao Xiang, Jianping Fan, and Dacheng Tao. 2018. Beyond Bilinear: Generalized Multimodal Factorized High-Order Pooling for Visual Question Answering. *TNNLS* 29 (2018), 5947–5959.

[55] Yan Zhang, Jonathon S. Hare, and Adam Prügel-Bennett. 2018. Learning to Count Objects in Natural Images for Visual Question Answering. In *ICLR*.