



# DualVD: An Adaptive Dual Encoding Model for Deep Visual Understanding in Visual Dialogue

蒋萧泽

北京航空航天大学

自动化科学与电气工程学院 智能计算与机器学习实验室

xzjiang@buaa.edu.cn



蒋萧泽

北京航空航天大学

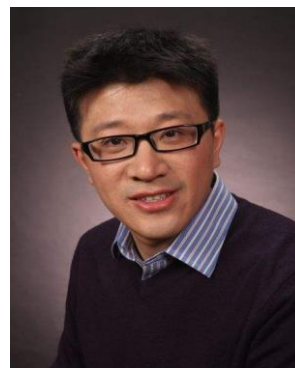
xzjiang@buaa.edu.cn



于静

中科院信工所

yujing02@iie.ac.cn



秦曾昌

北京航空航天大学

zcqin@buaa.edu.cn



张星星

微软亚洲研究院

xizhang@microsoft.com



吴琦

阿德莱德大学

qi.wu01@adelaide.edu.au

论文链接: <https://arxiv.org/pdf/1911.07251.pdf>  
代码链接: <https://github.com/JXZe/DualVD>



# 目录

1. 视觉对话任务概述
2. 研究动机
3. 模型设计
4. 实验结果
5. 总结及展望

# 视觉对话任务概述——研究背景与意义



北京航空航天大学  
BEIHANG UNIVERSITY



Is there smoke in any room around you?



Yes, in one room

Go there and look for people



...

## ▲ 视觉导航



Peter just uploaded a picture from his vacation in Hawaii

Great! Is he at the beach?



No, on a mountain



...

## ▲ 视觉信息不对等的智能聊天（小爱同学，天猫精灵）



Did anyone enter the room last week?



Yes, 127 instances logged on camera

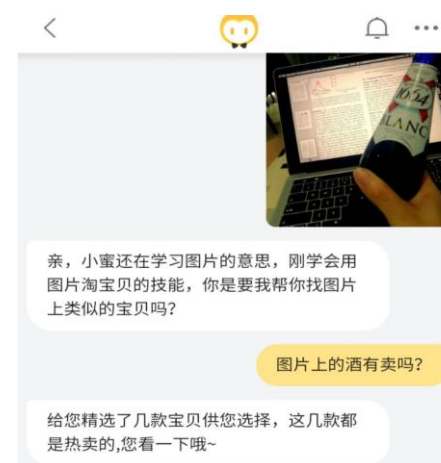


Were any of them carrying a black bag?



...

## ▲ 安防领域



亲，小蜜还在学习图片的意思，刚学会用图片淘宝贝的技能，你是要我帮你找图片上类似的宝贝吗？

图片上的酒有卖吗？

给您精选了几款宝贝供您选择，这几款都是热卖的，您看一下哦~



Is my pose right?

No, you should raise your arms higher.

## ▲ 基于多模态信息的智能客服 (阿里小蜜，度蜜，智能健身教练)

## ➤ 视觉对话任务

### 输入

- 图像  $I$
- 图像描述  $C$  和  $t-1$  轮的对话历史

$$H = \{C, (Q_1, A_1), \dots, (Q_{t-1}, A_{t-1})\}$$

- 当前轮问题  $Q_t$

### 输出

- 视觉对话任务要求从100个候选答案集合  $A_t = \{A_1, A_2, \dots, A_{100}\}$  中选择最佳答案（每一轮的问题均有自己的候选答案集合）



#### VQA

Q: How many people on wheelchairs ?

A: Two

Q: How many wheelchairs ?

A: One

#### Captioning

Two people are in a wheelchair and one is holding a racket.

#### Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman



#### Visual Dialog

Q: What is the gender of the one in the white shirt ?

A: She is a woman

Q: What is she doing ?

A: Playing a Wii game

Q: Is that a man to her right ?

A: No, it's a woman



## 挑战1：同一问题有多种正确答案



Q: Do you see any birds?

### valid answers

No No birds

I do not see any birds

Nope Not at all

No, Not that I can see

## 挑战2：问题中的指代理解



### Visual Dialog

Q: How many people are on wheelchairs ?

A: Two

Q: What are their genders ?

A: One male and one female

Q: Which one is holding a racket ?

A: The woman

## 挑战3：对话中涉及的图像内容不断变化



C: A man doing a grind on a skateboard.

Q1: Is the man on the skateboard?

A1: Yes, he is.

...

Q4: Is he younger or older?

A4: He is in the middle-aged.

Q5: Is there sky in the picture?

A5: Yes, the sky is deep blue.



Image

**C:** A man doing a grind on a skateboard.

**Q1:** Is the man on the skateboard?

**A1:** Yes, he is.

...

**Q4:** Is he younger or older?

**A4:** He is in the middle-aged.

**Q5:** Is there sky in the picture?

**A5:** Yes, the sky is deep blue with some clouds.

History

- 视觉对话需要模型根据对话的推进，不断调整视角，关注问题涉及的多样的视觉信息
- 如何在对话过程中自适应地捕获回答问题所需的视觉线索是视觉对话中的重要挑战之一



Image

C: A man doing a grind on a skateboard.

Q1: Is the man on the skateboard?

A1: Yes, he is.

...

Q4: Is he younger or older?

A4: He is in the middle-aged.

Q5: Is there sky in the picture?

A5: Yes, the sky is deep blue with some clouds.

History



the man



skateboard

前景信息

- 视觉对话需要模型根据对话的推进，不断调整视角，关注问题涉及的多样的视觉信息
- 如何在对话过程中自适应地捕获回答问题所需的视觉线索是视觉对话中的重要挑战之一





Image

C: A man doing a grind on a skateboard.  
Q1: Is the man on the skateboard?  
A1: Yes, he is.  
...  
Q4: Is he younger or older?  
A4: He is in the middle-aged.  
Q5: Is there sky in the picture?  
A5: Yes, the sky is deep blue with some clouds.

History



sky

背景信息

- 视觉对话需要模型根据对话的推进，不断调整视角，关注问题涉及的多样的视觉信息
- 如何在对话过程中自适应地捕获回答问题所需的视觉线索是视觉对话中的重要挑战之一



Image

C: A man doing a grind on a skateboard.

Q1: Is the man on the skateboard?

A1: Yes, he is.

...

Q4: Is he younger or older?

A4: He is in the middle-aged.

Q5: Is there sky in the picture?

A5: Yes, the sky is deep blue with some clouds.

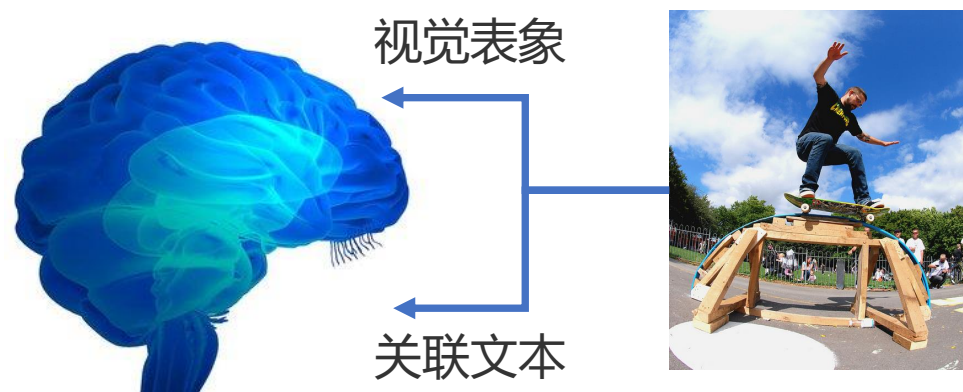
History



the middle-aged man

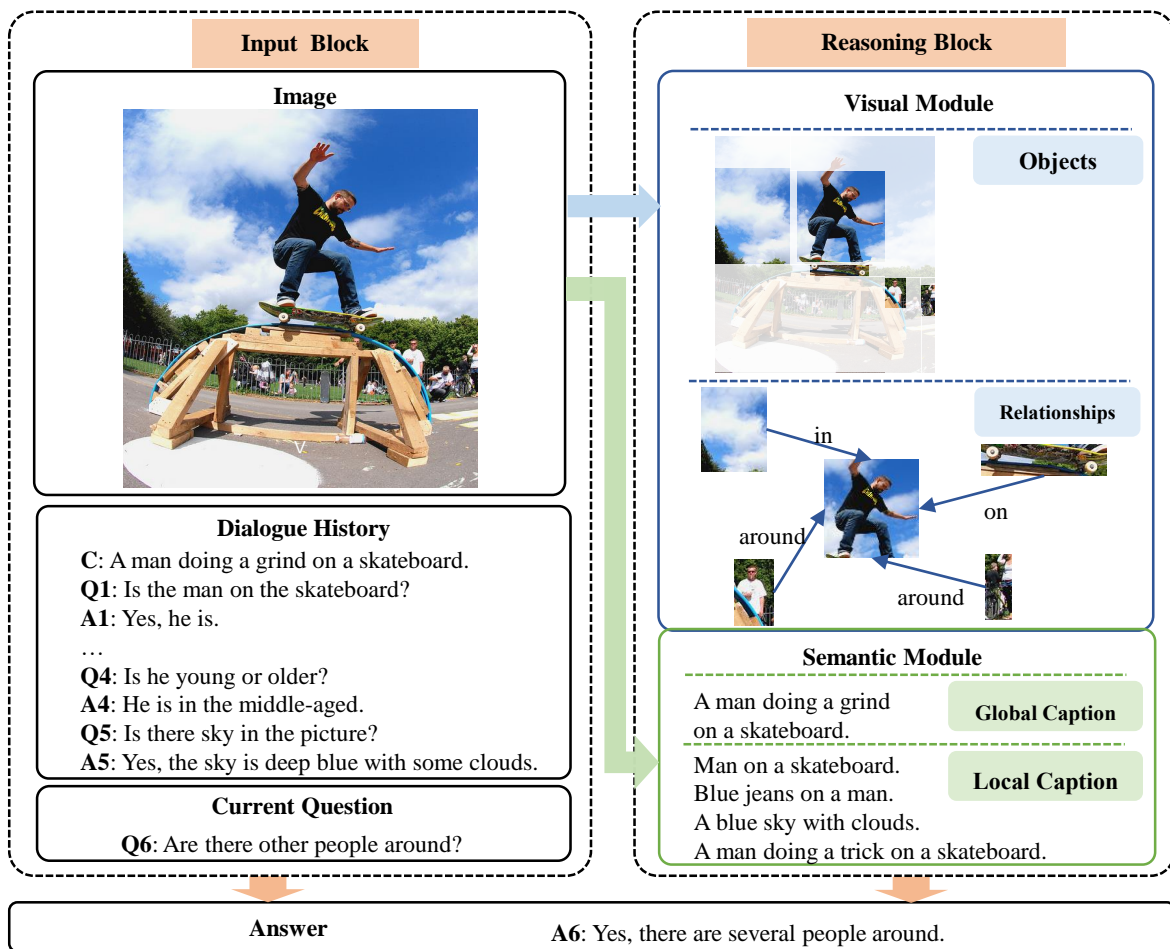
高层语义信息

- 视觉对话需要模型根据对话的推进，不断调整视角，关注问题涉及的多样的视觉信息
- 如何在对话过程中自适应地捕获回答问题所需的视觉线索是视觉对话中的重要挑战之一



- 认知学中**双向编码理论**<sup>[1]</sup>认为：  
人类大脑编码信息包括两种方式，即视觉表象和关联文本
- 当被问到某个概念时，大脑会检索相关的视觉信息、语言信息或综合考虑上述两种信息
- 这种双向编码方式能够增强大脑的记忆和理解能力

[1] A. Paivio, “*Imagery and Verbal Process*.” New York: Holt, Rinehart and Winston., 1971.

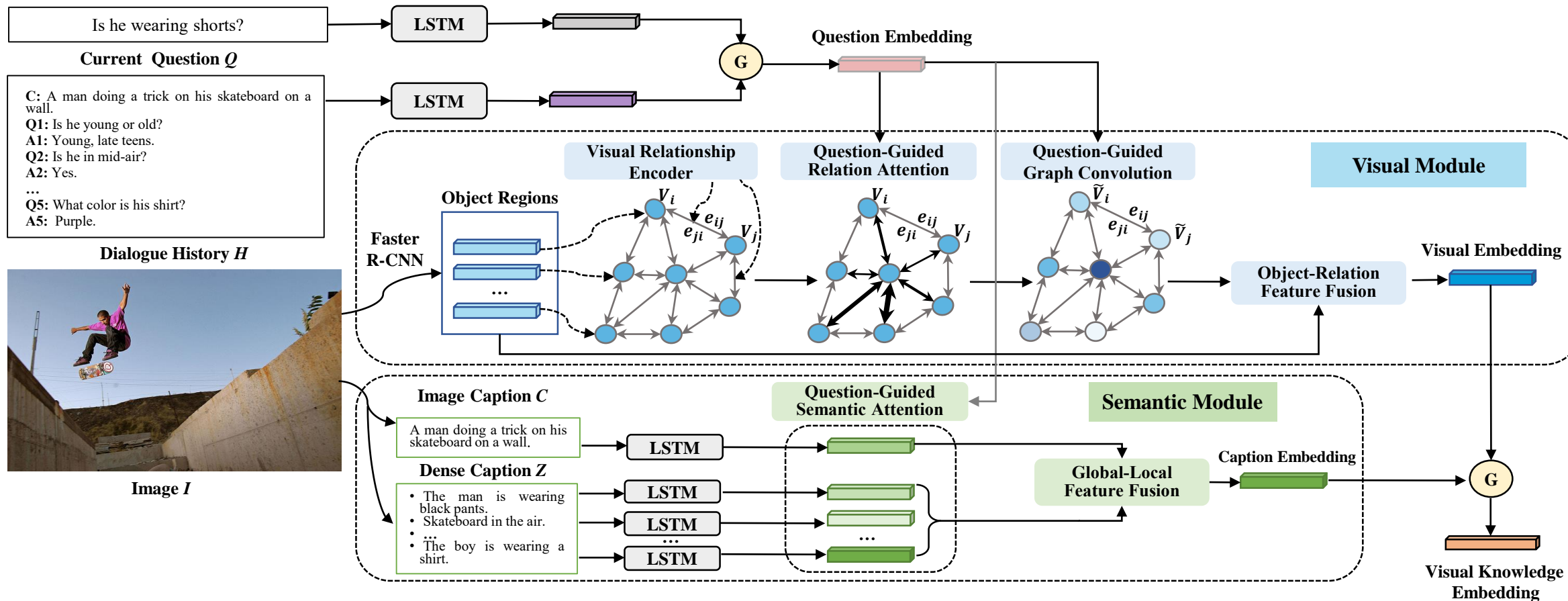


➤ 受双向编码理论启发，提出了一种从视觉和语义两方面刻画图像信息的新框架

➤ 基于双编码框架，提出了一种自适应视觉信息选择模型

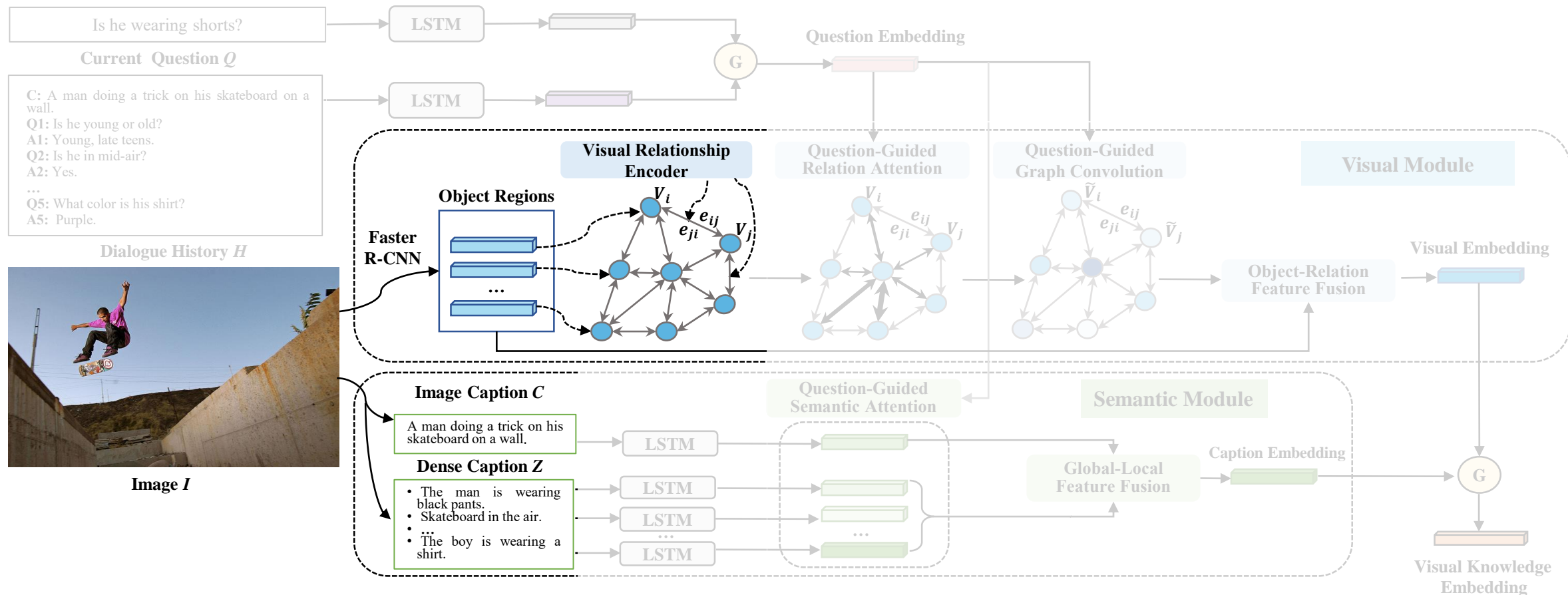
(1) 模态内信息选择：由问题驱动，分别在视觉模块和语义模块中获得独立线索

(2) 模态间信息选择：由问题驱动，获得视觉-语义的联合线索

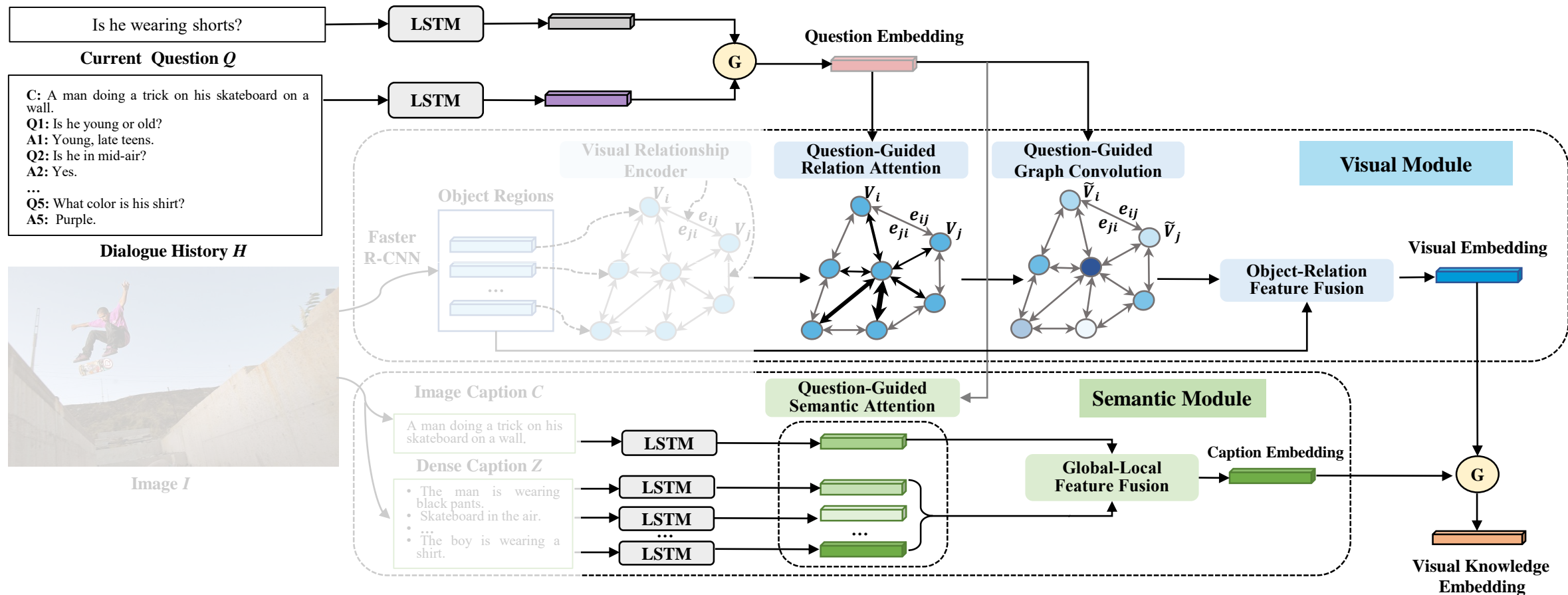


- 模型核心结构分为两部分：
- Visual-Semantic Dual Encoding
  - Adaptive Visual-Semantic Knowledge Selection





## ➤ Visual-Semantic Dual Encoding



## ➤ Adaptive Visual-Semantic Knowledge Selection

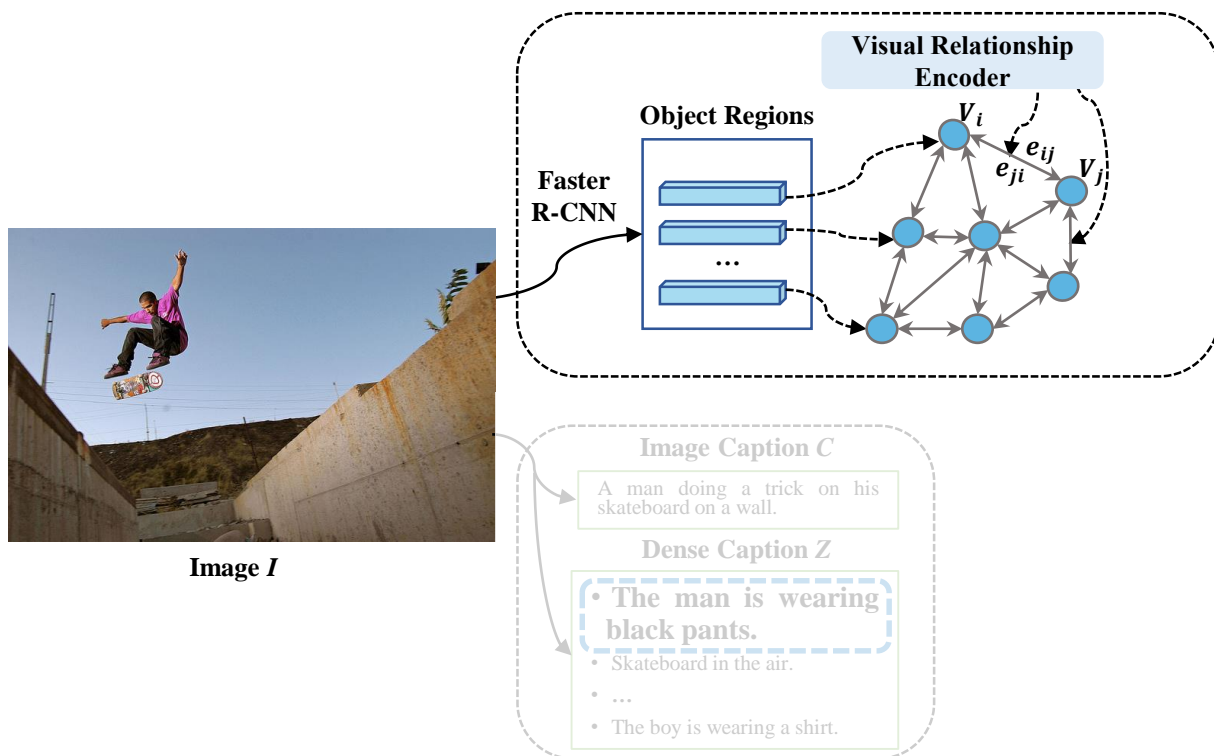
## ● Visual-Semantic Dual Encoding 视觉编码 和 语义编码

- 提出刻画图像的视觉信息和语义信息的新框架，其中视觉信息采用场景图表示，语义信息采用多层面语义描述表示

### • Scene Graph Construction

Step1 :采用Faster-RCNN提取图像中N个目标区域，构成场景图上的节点

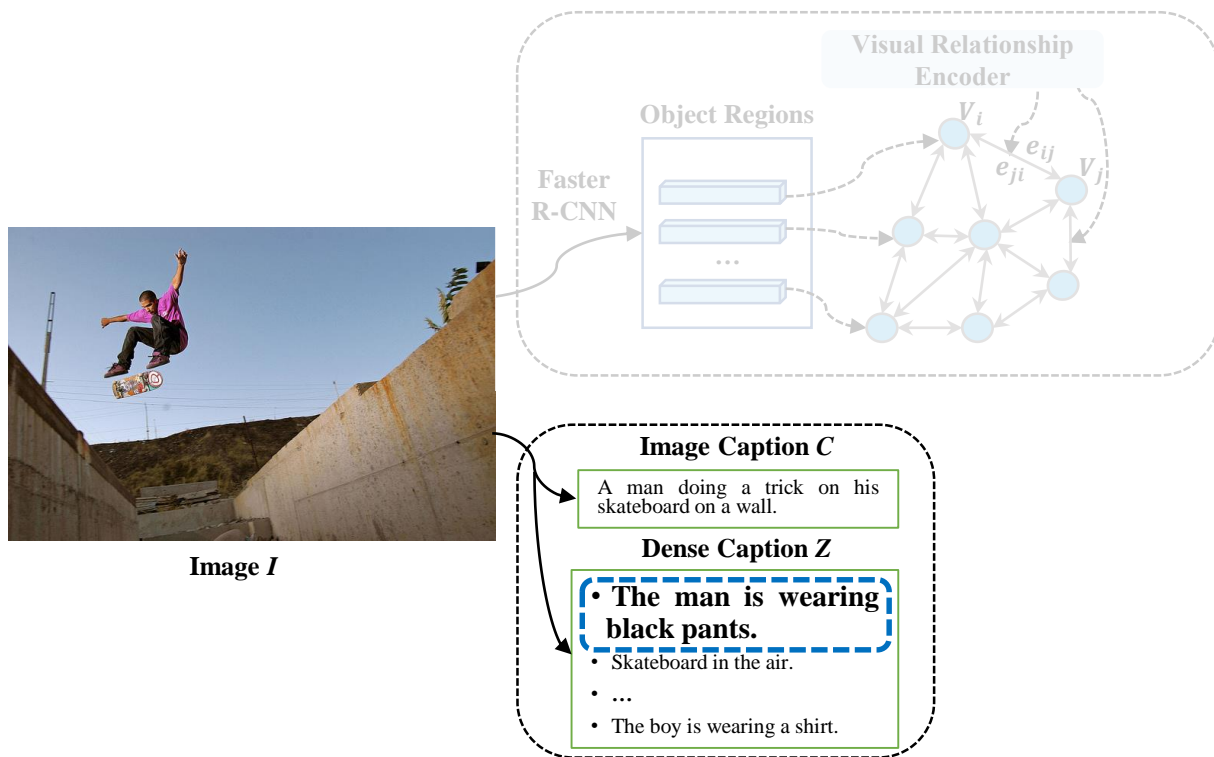
Step2: 采用Zhang等提出的视觉关系编码器<sup>[1]</sup>在GQA数据集上预训练，将给定图像中任何两个目标区域间的视觉关系编码为关系向量。



[1] J. Zhang, Y. Kalantidis, M. Rohrbach, M. Paluri, A. Elgammal, and M. Elhoseiny "Large-scale visual relationship understanding," in AAAI, 2019.

## ● Visual-Semantic Dual Encoding 视觉编码 和 语义编码

- 提出刻画图像的视觉信息和语义信息的新框架，其中视觉信息采用场景图表示，语义信息采用多层面语义描述表示



### • Multi-level Image Captions

将每幅图像表示为多层面的语义描述，同时刻画图像的局部和全局语义信息。

- ✓ 采用**数据集提供的图像描述**作为图像的全局语义信息；
- ✓ 采用Feifei Li等提出的 **DenseCap<sup>[1]</sup>** 提取描述细节的 *k* 条 dense captions 作为图像的局部语义信息。

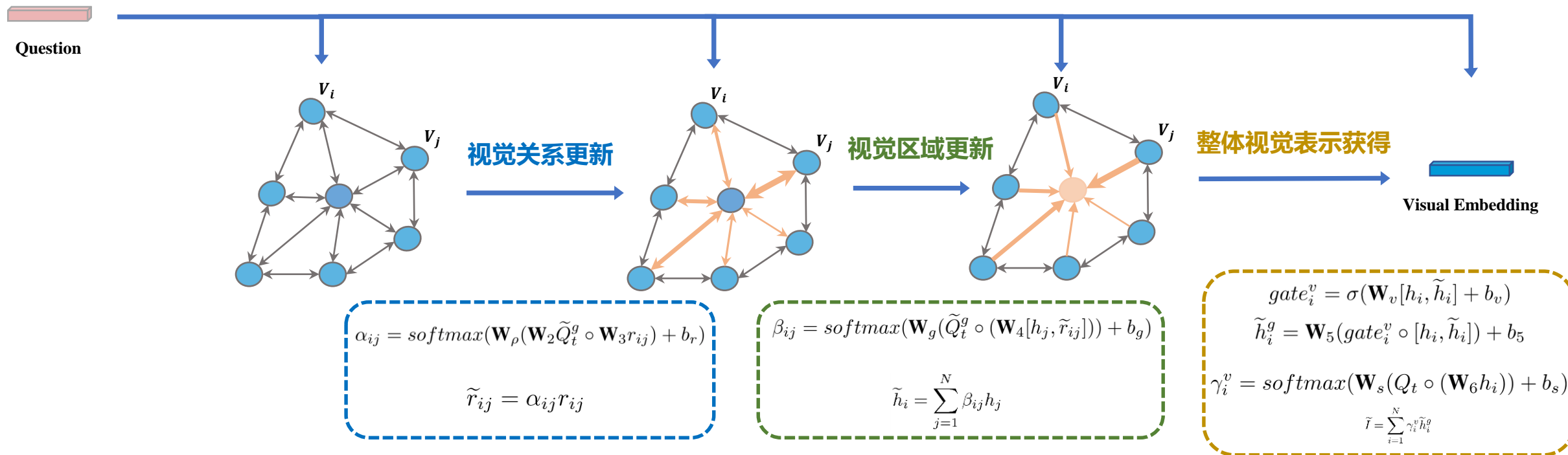
[1] Johnson, J.; Karpa- thy, A.; and Fei-Fei, L. 2016. Denscap: Fully convolu- tional localization networks for dense captioning. In *CVPR*, 4565–4574.

## ● Visual-Semantic Knowledge Selection

- 基于问题的引导，DualVD的信息选择过程分两步： **(1) 模态内信息选择** (2) 模态间信息选择

模态内信息选择分别通过**视觉模块 (Visual Module)** 和语义模块 (Semantic Module) 提取视觉和语义信息；

### • 视觉模块内部选择



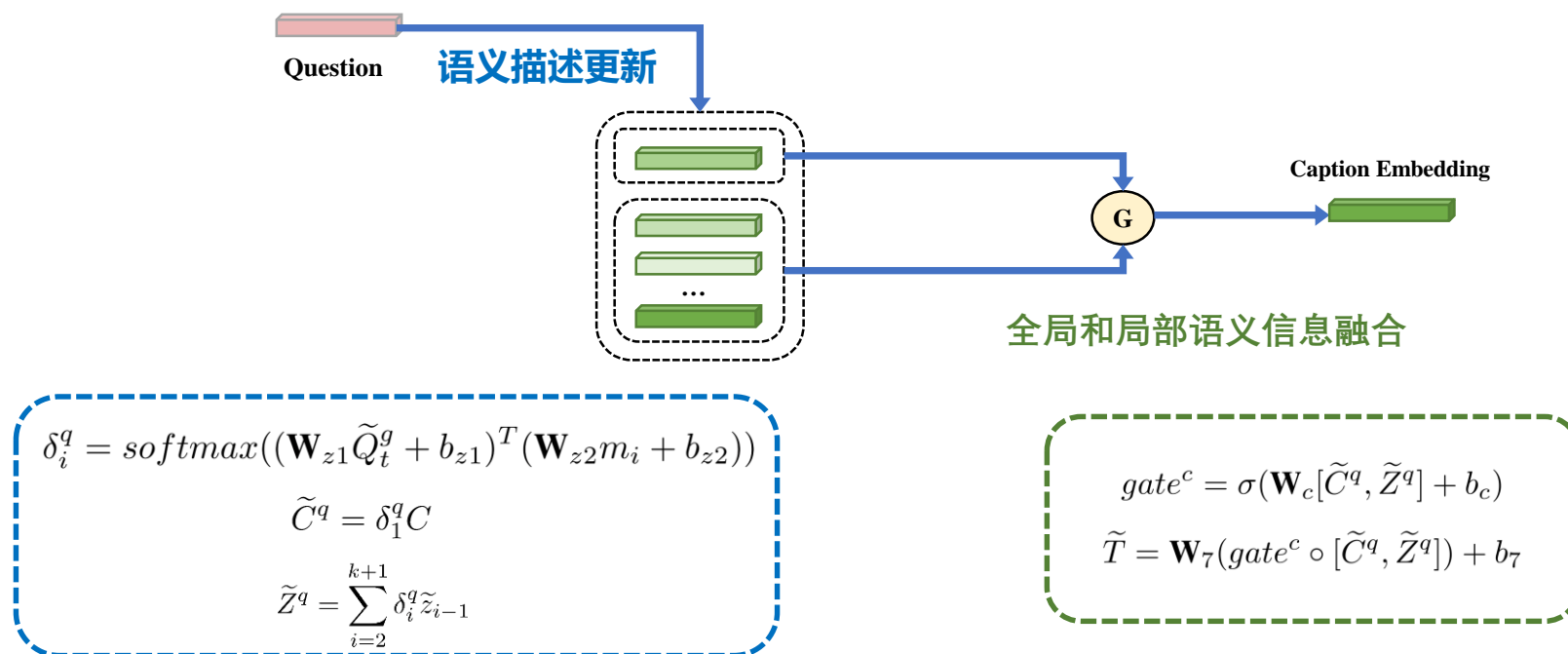


## ● Visual-Semantic Knowledge Selection

- 基于问题的引导，DualVD的信息选择过程分两步： **(1) 模态内信息选择** (2) 模态间信息选择

模态内信息选择分别通过视觉模块（Visual Module）和**语义模块（Semantic Module）**提取视觉和语义信息；

- **语义模块**内部选择

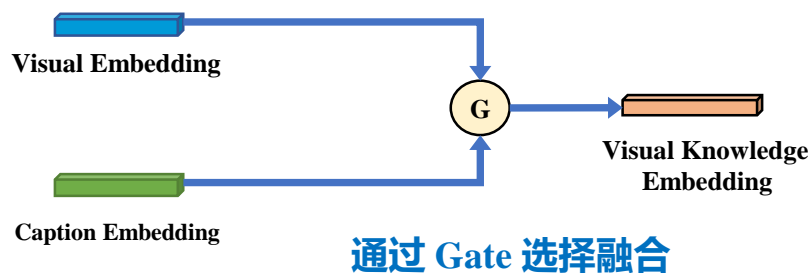


## ● Visual-Semantic Knowledge Selection

- 基于问题的引导，DualVD的信息选择过程分两步： （1）模态内信息选择 （2）模态间信息选择

模态间特征选择通过选择性视觉-语义融合（Selective visual-semantic fusion）汇聚视觉模块和语义模块中问题相关的线索

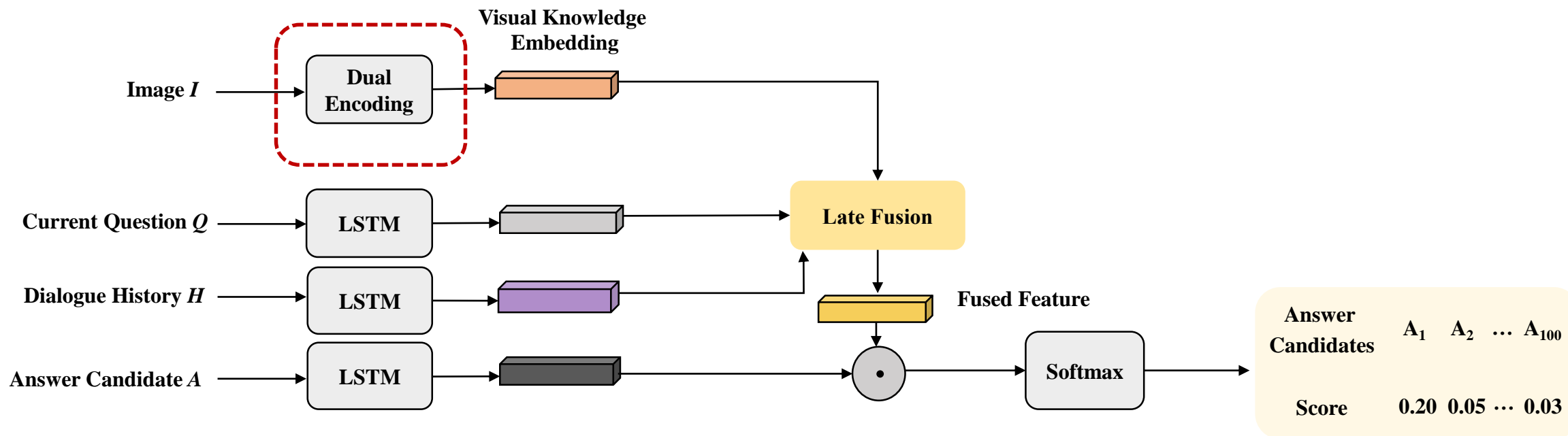
- 模态间选择



$$gate^s = \sigma(\mathbf{W}_s[\tilde{I}, \tilde{T}] + b_s)$$

$$S = gate^s \circ [\tilde{I}, \tilde{T}]$$

## ● Late Fusion Encoder



- 模型和现有针对对话历史的研究工作具有互补优势，可以应用于更加复杂的 encoder
- 本篇论文重点证明所提出的视觉建模方法的有效性，因此采用了简单的 Late Fusion encoder

## ➤ 数据集

VisDial v0.9: 图像数据全部来源于MSCOCO,  
分为训练集 (train)、测试集 (test) 以及验证集 (val)

VisDial v1.0: 将VisDial v1.0 的所有数据均作为其训练集 (train, 120k),  
测试集 (test, 8k) 和验证集 (val, 2k) 的图像数据 来自于 Flickr

## ➤ 评价指标

VisDial v0.9: 采用检索的评价指标 MRR,  $R@k$  ( $k = 1, 5, 10$ ), Mean

VisDial v1.0: 增加 NDCG 评价指标

只有 Mean 值越小越好, 其他的评价指标值越大越好

## ➤ 实验内容

- (1) 在 VisDial v0.9 和 VisDial v1.0 上比较state-of-the-art的结果
- (2) 对模型核心组成部分的消融实验
- (3) 对实验结果的可视化实验

## ● Compare with State-of-the-art

Table 1: Comparison on validation split of VisDial v0.9.

Model	MRR	R@1	R@5	R@10	Mean
LF	58.07	43.82	74.68	84.07	5.78
HRE	58.46	44.67	74.50	84.22	5.72
MN	59.65	45.55	76.22	85.37	5.46
SAN-QI	57.64	43.44	74.26	83.72	5.88
HieCoAtt-QI	57.88	43.51	74.49	83.96	5.84
AMEM	61.60	47.74	78.04	86.84	4.99
HCIAE	62.22	48.48	78.75	87.59	4.81
SF	62.42	48.55	78.96	87.75	4.70
CoAtt	63.98	50.29	80.71	88.81	4.47
CorefMN	<b>64.10</b>	<b>50.92</b>	80.18	88.81	4.45
VGNN	62.85	48.95	79.65	88.36	4.57
<b>DualVD</b>	62.94	48.64	<b>80.89</b>	<b>89.94</b>	<b>4.17</b>

Table 2: Comparison on test-standard split of VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
LF	55.42	40.95	72.45	82.83	5.95	45.31
HRE	54.16	39.93	70.47	81.50	6.41	45.46
MN	55.49	40.98	72.30	83.30	5.92	47.50
LF-Att	57.07	42.08	74.82	85.05	5.41	40.76
MN-Att	56.90	42.43	74.00	84.35	5.59	49.58
CorefMN	61.50	47.55	78.10	88.80	4.40	54.70
VGNN	61.37	47.33	77.98	87.83	4.57	52.82
RvA	63.03	49.03	80.40	89.83	4.18	55.59
DL-61	62.20	47.90	<b>80.43</b>	<b>89.95</b>	4.17	<b>57.32</b>
<b>DualVD</b>	<b>63.23</b>	<b>49.25</b>	80.23	89.70	<b>4.11</b>	56.32

- 与现有算法相比，DualVD的结果超过现有大多数模型，略低于采用了多步推理和复杂attention机制的模型



➤ Ablation Study

- ObjRep:

只用 Faster R-CNN 提取的图像特征作为图像 的表示
- RelRep:

在 ObjRep 的基础上增加 *question-guided relation attention* 和 *question-guided graph convolution* 操作
- VisNoRel:

图像的每个节点之间有边, 但是没有预训练的信息
- VisMod:

只用视觉模块编码图像的模型

Table 3: Ablation study of DualVD on VisDial v1.0.

Model	MRR	R@1	R@5	R@10	Mean	NDCG
ObjRep	63.84	49.83	81.27	90.29	4.07	55.48
RelRep	63.63	49.25	81.01	90.34	4.07	55.12
VisNoRel	63.97	49.87	81.74	90.60	4.00	56.73
VisMod	64.11	50.04	81.78	90.52	3.99	56.67
GlCap	60.02	45.34	77.66	87.27	4.78	50.04
LoCap	60.95	46.43	78.45	88.17	4.62	51.72
SemMod	61.07	46.69	78.56	88.09	4.59	51.10
<b>DualVD</b>	<b>64.64</b>	<b>50.74</b>	<b>82.10</b>	<b>91.00</b>	<b>3.91</b>	<b>57.30</b>

- GlCap:

只用全局的图像描述
- LoCap:

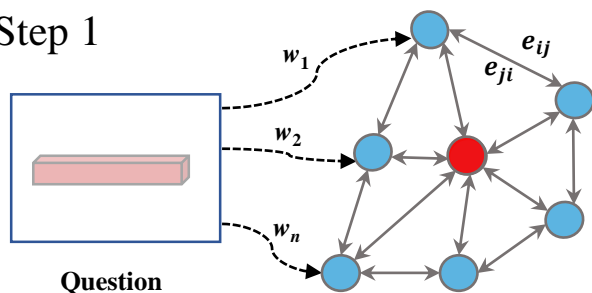
只用局部的图像描述
- SemMod:

只用语义模块编码图像的模型

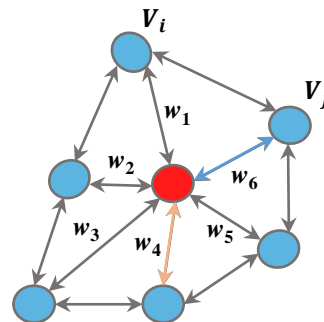
- 该模型的一个优势是具有较强的可解释性，通过对 Attention Weight、Gate Value 的可视化，能够显示分析模型特征选择的过程。

- 图像区域可视化

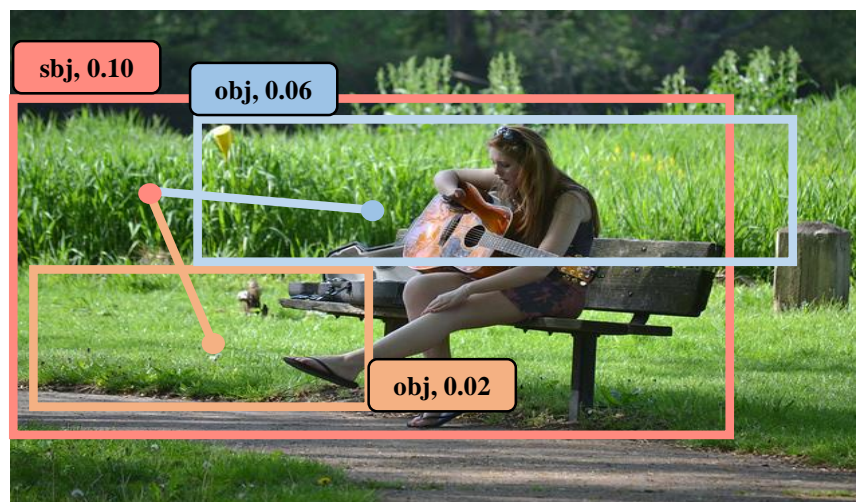
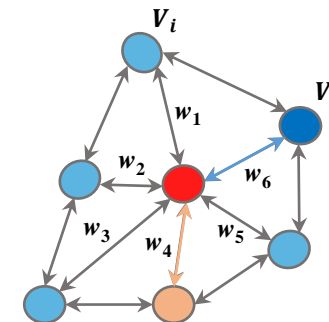
Step 1



Step 2



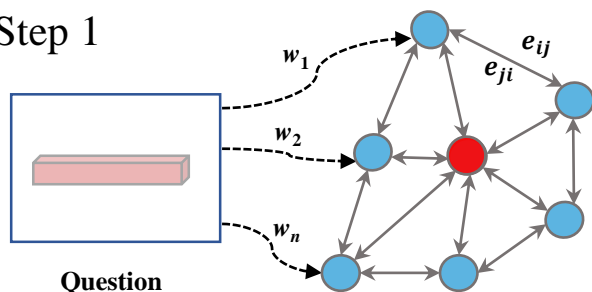
Step 3



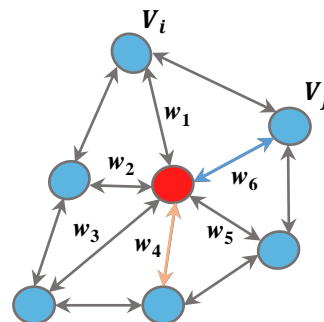
- 该模型的一个优势是具有较强的可解释性，通过对 Attention Weight、Gate Value 的可视化，能够显示分析模型特征选择的过程

- 图像区域可视化

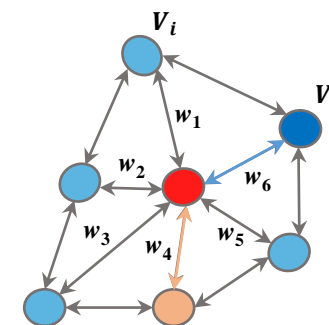
Step 1



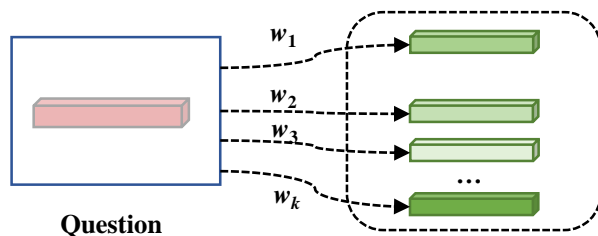
Step 2



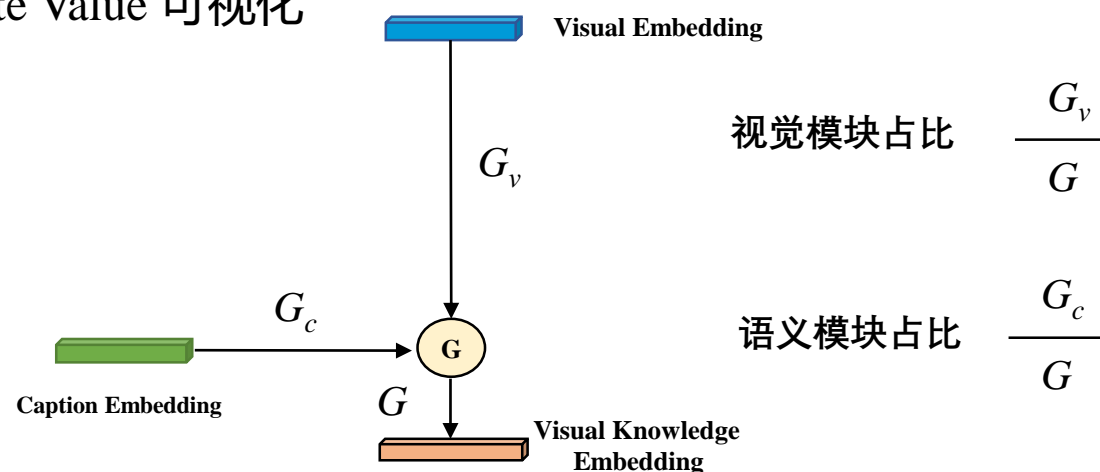
Step 3



- Caption 可视化



- Gate Value 可视化



- Case

Image



Dialogue History

C: 2 boys playing disc golf in a forest.

Question1	Are the boys teenagers?
Answer1	They are young boys.
Question2	Do you see a lot of trees?
Answer2	Yes, a ton of trees.
Question3	Dose 1 of the boys holding the disc?
Answer3	They are both holding discs.




• Case

Question1	Are the boys teenagers?
Answer1	They are young boys.

Visual Module

Semantic Module

Ratio of total gate values: 55.96%



Ratio of total gate values: 44.04%

2 boys playing disc golf in a forest.

A man wearing blue shorts.

Boy holding blue frisbee.

Two people playing with a frisbee.

A blue shirt on a man.

Boy wearing blue shirt.

Blue shorts on the man.

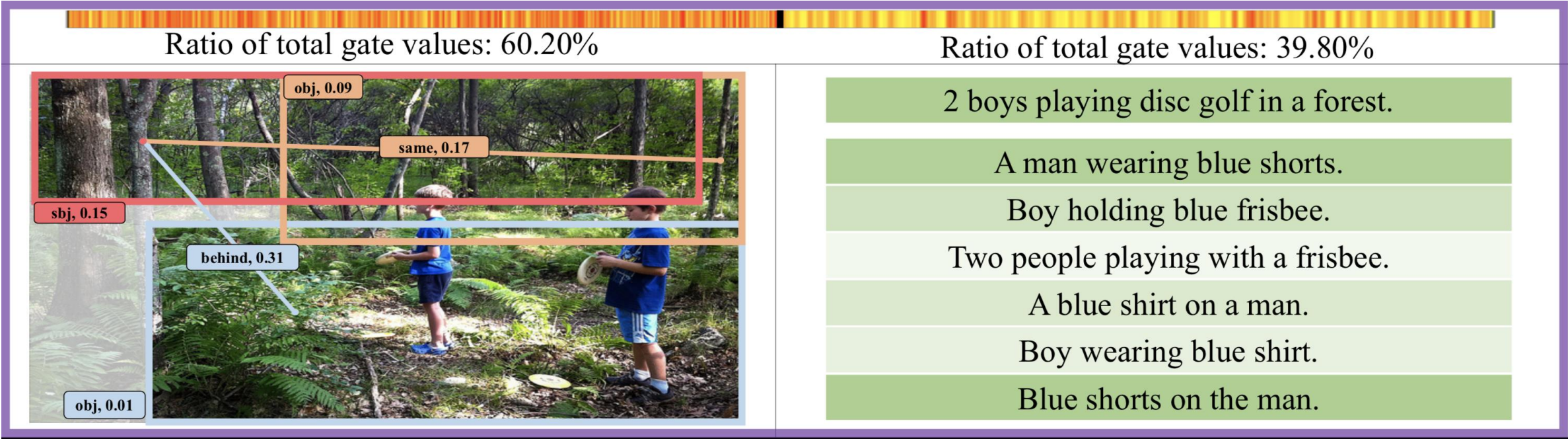


- Case

Question2	Do you see a lot of trees?
Answer2	Yes, a ton of trees.

Visual Module

Semantic Module



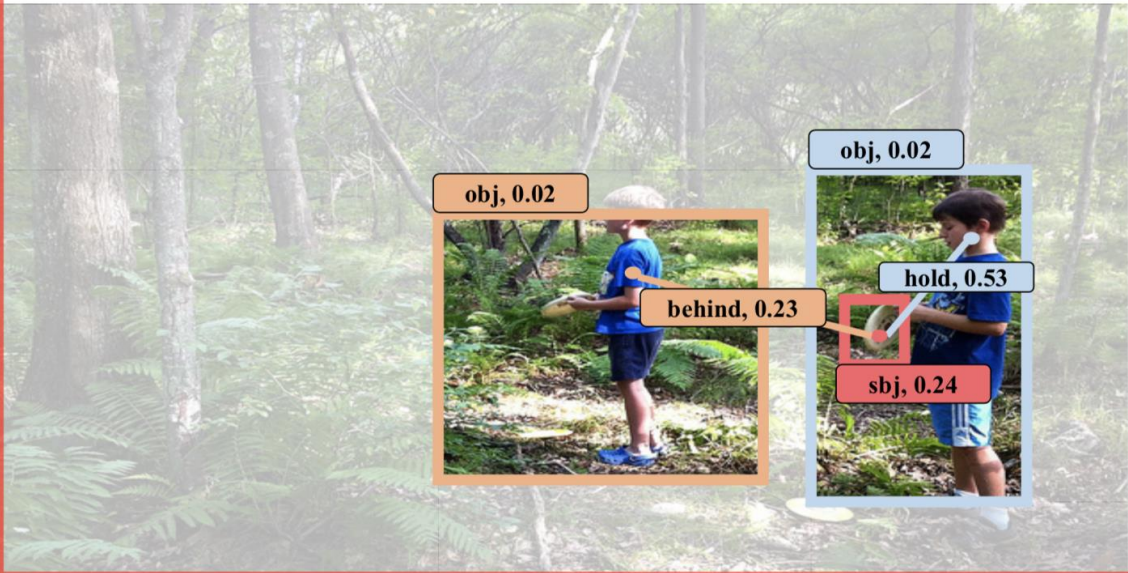
• Case

Question3	Dose 1 of the boys holding the disc?
Answer3	They are both holding discs.

Visual Module

Semantic Module

Ratio of total gate values: 54.90%



Ratio of total gate values: 45.10%

2 boys playing disc golf in a forest.

A man wearing blue shorts.

Boy holding blue frisbee.

Two people playing with a frisbee.

A blue shirt on a man.

Boy wearing blue shirt.

Blue shorts on the man.

- 一般性结论
  - 视觉信息和语义信息对于回答问题的贡献取决于问题的复杂性和信息源的相关性
  - 视觉信息将为回答问题提供更重要的依据
  - 模型能够根据问题的变化，自适应调整关注的信息

- 提出了一种应用于视觉对话领域对于图像信息进行双编码的 DualVD 模型
- 通过可视化, DualVD 具有良好的可解释性
- 本研究工作和视觉对话领域中的其他工作 (比如致力于对历史信息进行建模的工作) 具有互补性
- 将 DualVD 与现有的模型框架进行结合是我们将来的一项工作





# 感谢聆听！

---

蒋萧泽

北京航空航天大学

自动化科学与电气工程学院 智能计算与机器学习实验室

xzjiang@buaa.edu.cn





蒋萧泽

北京航空航天大学

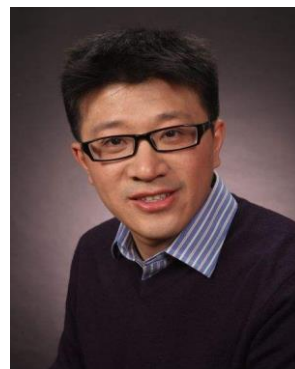
xzjiang@buaa.edu.cn



于静

中科院信工所

yujing02@iie.ac.cn



秦曾昌

北京航空航天大学

zcqin@buaa.edu.cn



张星星

微软亚洲研究院

xizhang@microsoft.com



吴琦

阿德莱德大学

qi.wu01@adelaide.edu.au

论文链接: <https://arxiv.org/pdf/1911.07251.pdf>  
代码链接: <https://github.com/JXZe/DualVD>