Heterogeneous Graph Learning for Visual Commonsense Reasoning

Weijiang Yu

Jingwen Zhou

Weihao Yu

Xiaodan Liang

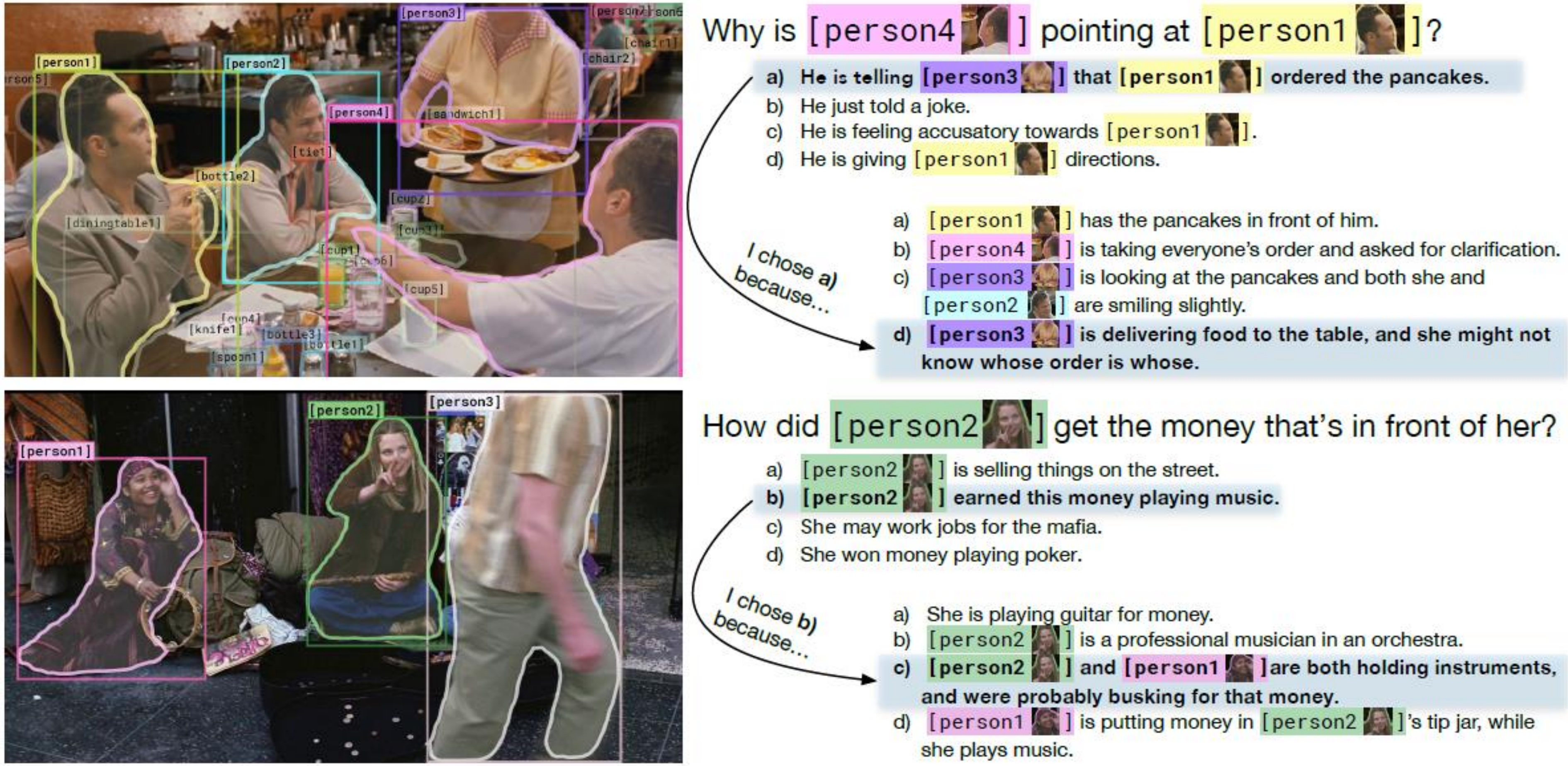
Nong Xiao

Sun Yat-sen University (SYSU), China

NeurIPS 2019 Spotlight Presentation
Vancouver, Canada

December 2019

Task Definition



Why is [person4] pointing at [person1]?

a) **He is telling [person3] that [person1] ordered the pancakes.**
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

I chose a) because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes and both she and [person2] are smiling slightly.
d) **[person3] is delivering food to the table, and she might not know whose order is whose.**

How did [person2] get the money that's in front of her?

a) [person2] is selling things on the street.
b) **[person2] earned this money playing music.**
c) She may work jobs for the mafia.
d) She won money playing poker.

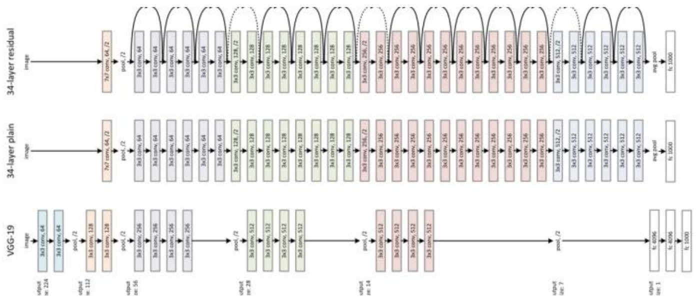
I chose b) because...

a) She is playing guitar for money.
b) [person2] is a professional musician in an orchestra.
c) **[person2] and [person1] are both holding instruments, and were probably busking for that money.**
d) [person1] is putting money in [person2]'s tip jar, while she plays music.

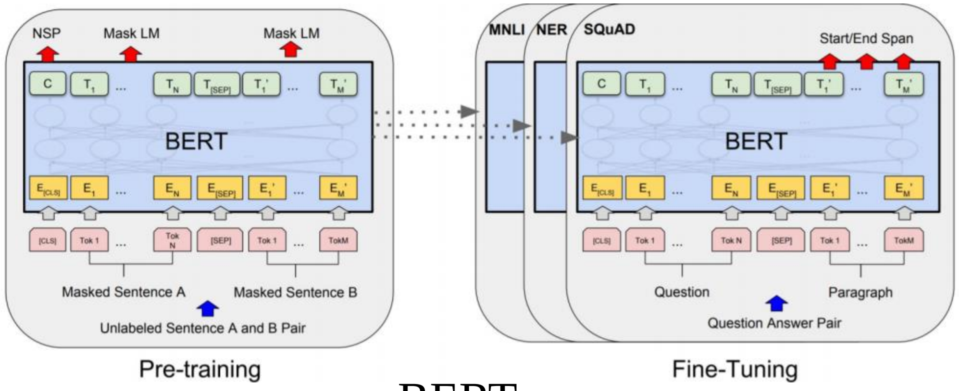
R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi. From recognition to cognition: Visual commonsense reasoning. In The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019.

Previous Works

- ➢ Powerful Backbone, such as resnet152, bert-large.

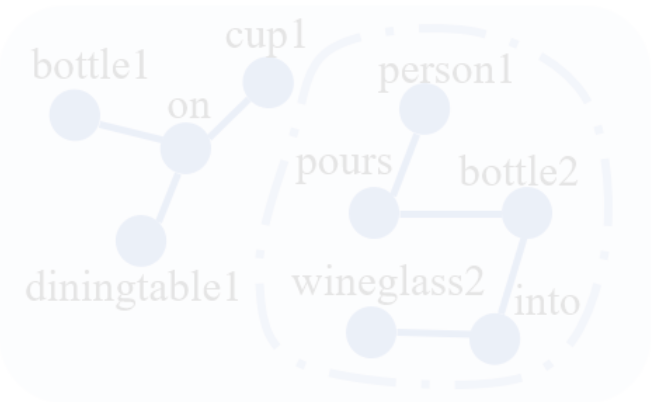


Resnet

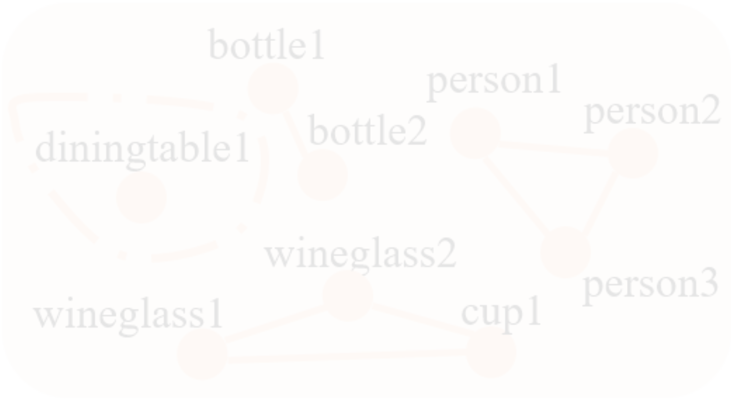


BERT

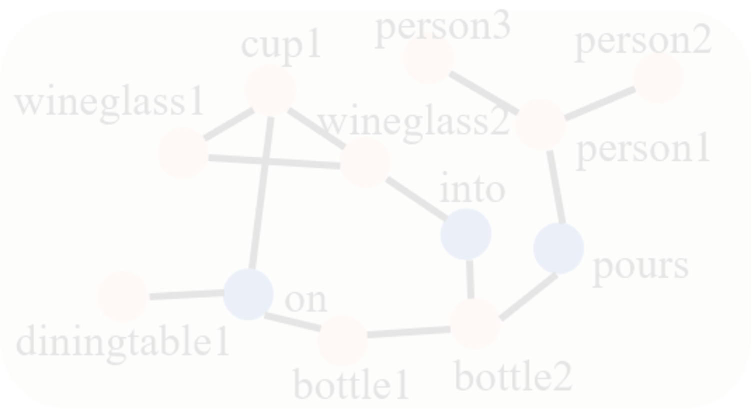
- ➢ Graph-based Methods



(a) Answer-to-Answer Homogeneous Graph



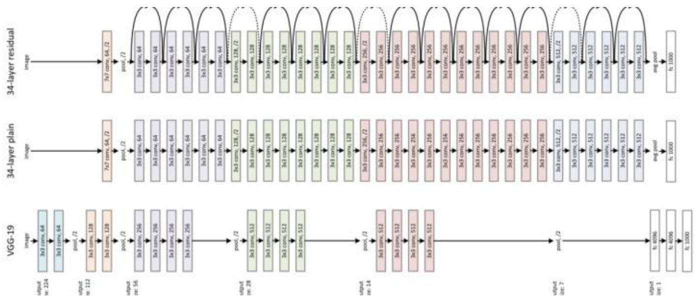
(b) Vision-to-Vision Homogeneous Graph



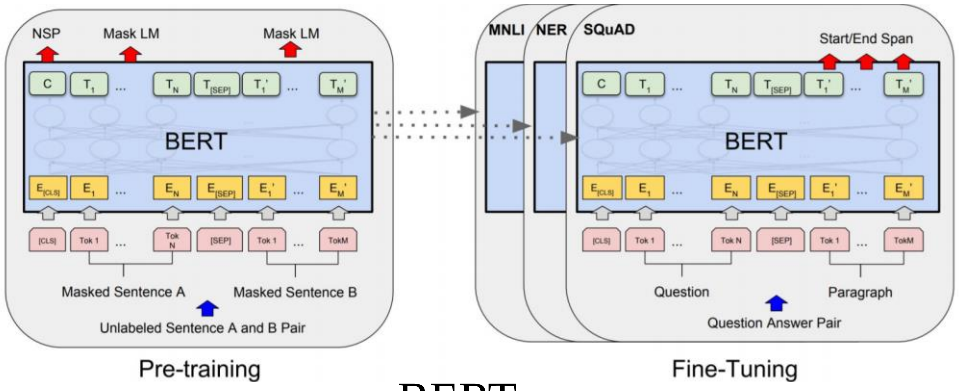
(c) Vision-to-Answer Heterogeneous Graph

Previous Works

- Powerful Backbone, such as resnet152, bert-large.

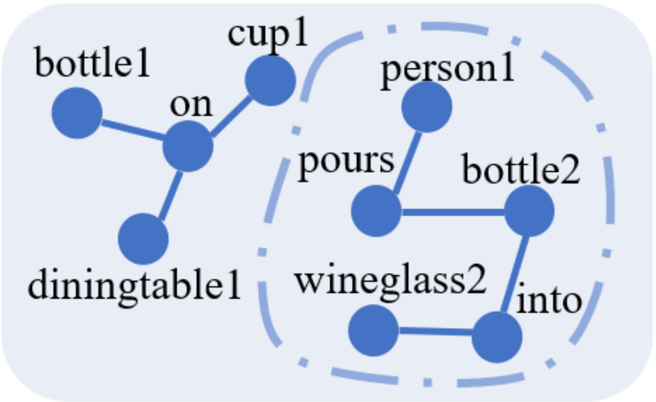


Resnet

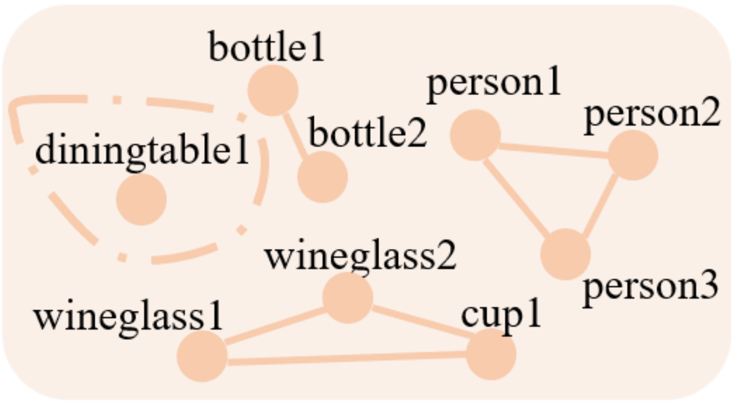


BERT

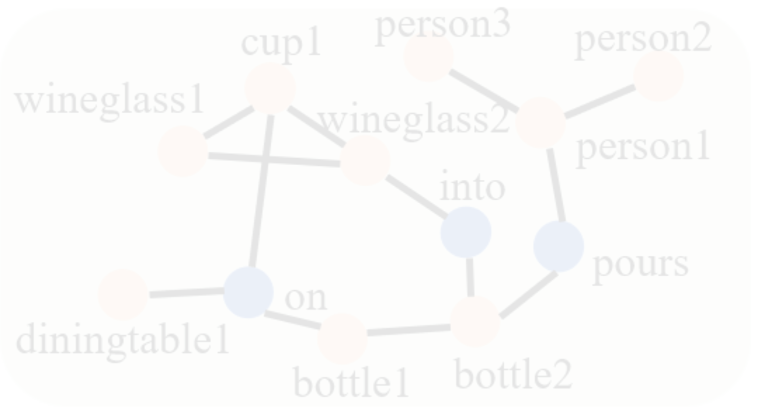
- Graph-based Methods



(a) Answer-to-Answer Homogeneous Graph



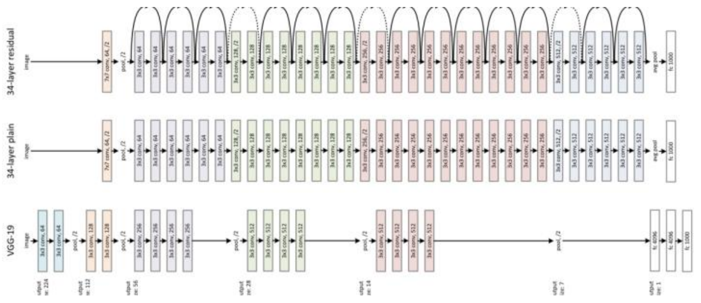
(b) Vision-to-Vision Homogeneous Graph



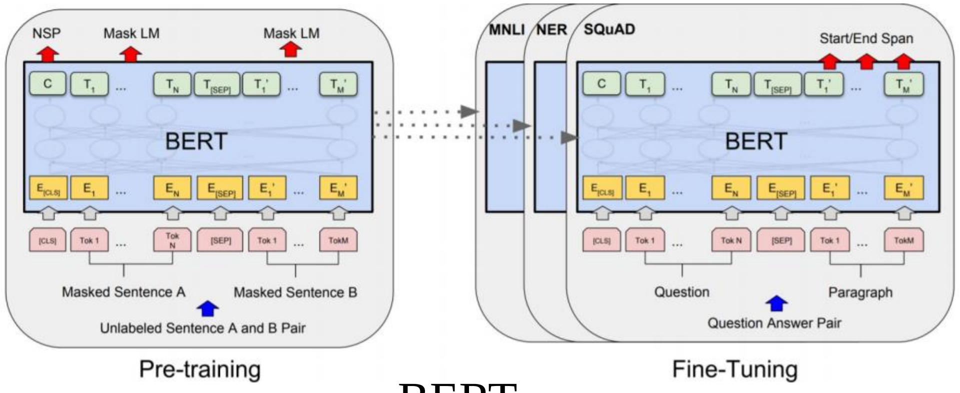
(c) Vision-to-Answer Heterogeneous Graph

Previous Works

- Powerful Backbone, such as resnet152, bert-large.

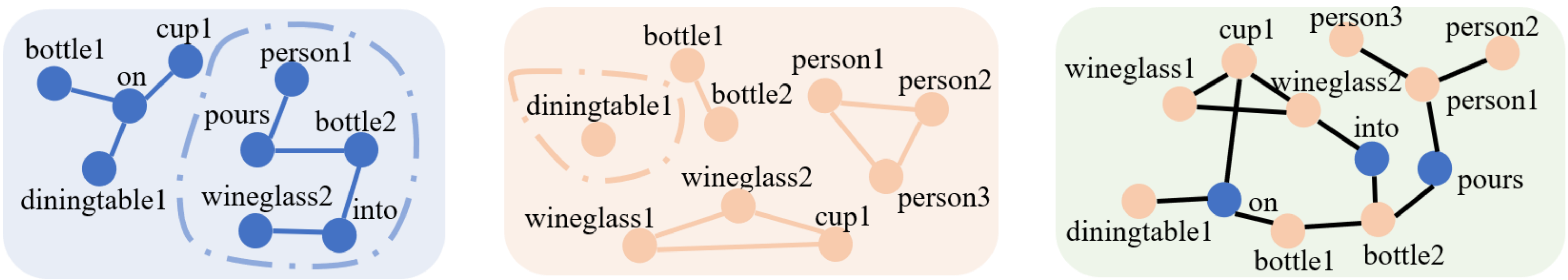


Resnet



BERT

- Graph-based Methods



(a) Answer-to-Answer Homogeneous Graph (b) Vision-to-Vision Homogeneous Graph (c) Vision-to-Answer Heterogeneous Graph

Our Method – HGL (Heterogeneous Graph Learning)

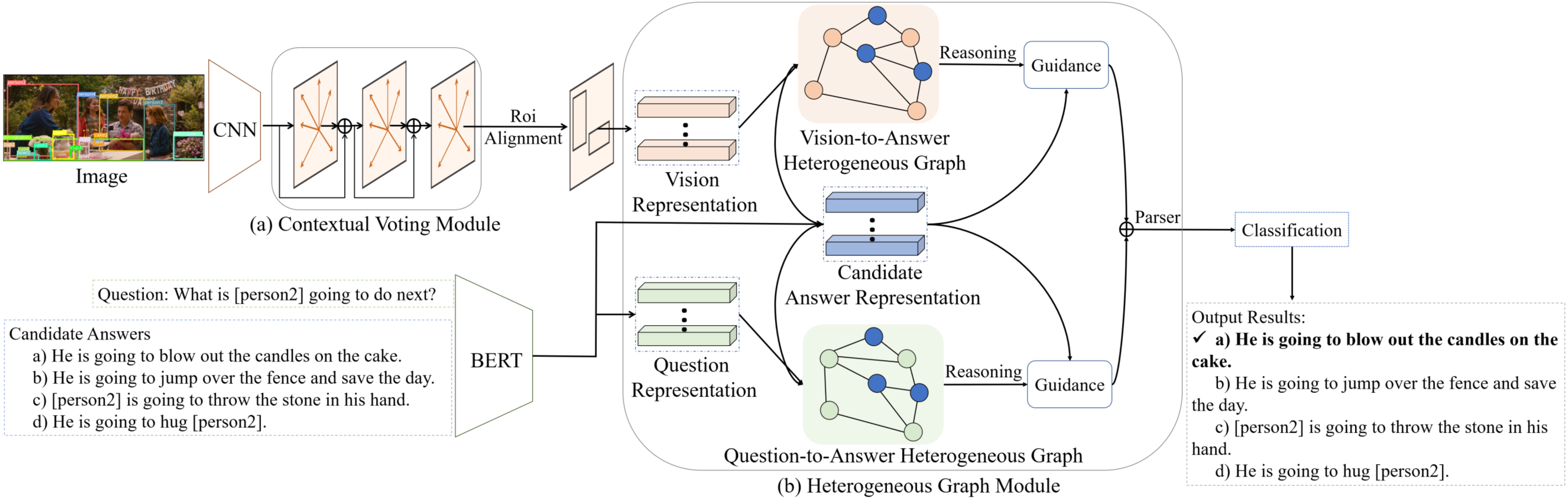


Image
CNN
(a) Contextual Voting Module
Roi Alignment
Vision Representation
Vision-to-Answer Heterogeneous Graph
Reasoning
Guidance
Candidate Answer Representation
Question Representation
Question-to-Answer Heterogeneous Graph
Reasoning
Guidance
Parser
Classification
(b) Heterogeneous Graph Module
BERT

Question: What is [person2] going to do next?

Candidate Answers
a) He is going to blow out the candles on the cake.
b) He is going to jump over the fence and save the day.
c) [person2] is going to throw the stone in his hand.
d) He is going to hug [person2].

Output Results:
✓ **a) He is going to blow out the candles on the cake.**
b) He is going to jump over the fence and save the day.
c) [person2] is going to throw the stone in his hand.
d) He is going to hug [person2].

- The goal of heterogeneous graph is to explore proper semantic alignment between and linguistic domains and knowledge reasoning to generate persuasive reasoning paths.
- The contextual voting module is for visual scene understanding with a global perspective at the low-level features. Some ambiguous semantics (rainy day) that lack of specific labels for detection and can not benefit from the labeled object bounding boxes and categories such as “person” and “dog” during training.

Our Method – HGL (Heterogeneous Graph Learning)

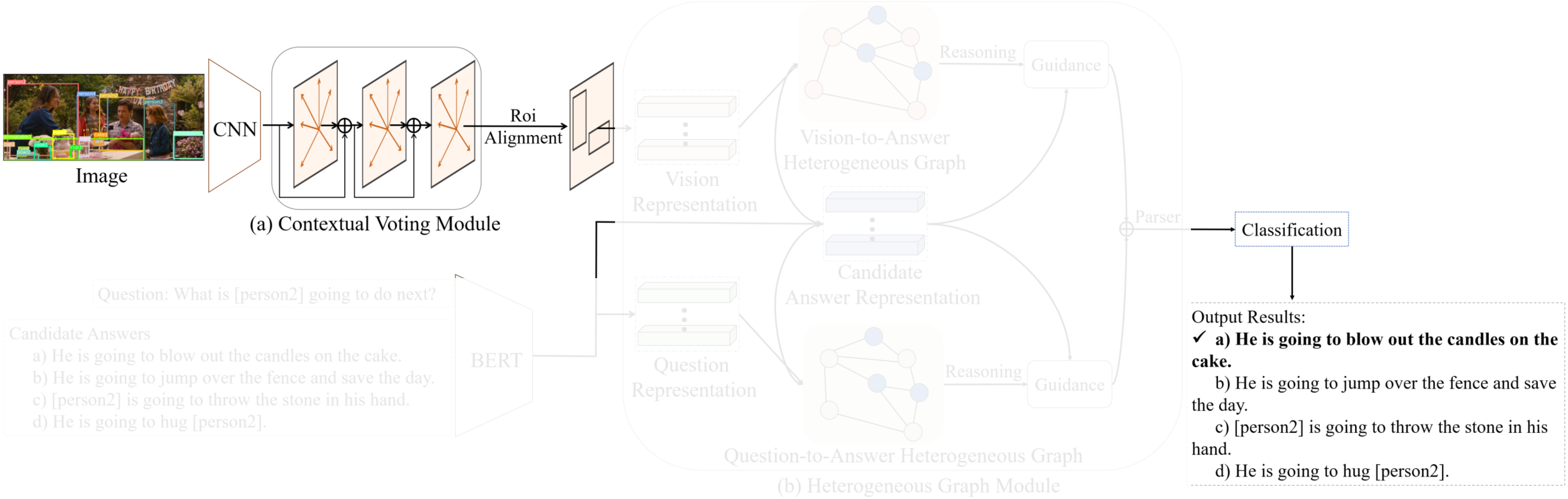


Image
CNN
Roi Alignment
(a) Contextual Voting Module
Question: What is [person2] going to do next?
Candidate Answers
a) He is going to blow out the candles on the cake.
b) He is going to jump over the fence and save the day.
c) [person2] is going to throw the stone in his hand.
d) He is going to hug [person2].
BERT
Vision Representation
Question Representation
Vision-to-Answer Heterogeneous Graph
Candidate Answer Representation
Question-to-Answer Heterogeneous Graph
Reasoning
Guidance
Reasoning
Guidance
Parser
(b) Heterogeneous Graph Module
Classification
Output Results:
✓ a) He is going to blow out the candles on the cake.
b) He is going to jump over the fence and save the day.
c) [person2] is going to throw the stone in his hand.
d) He is going to hug [person2].

- The goal of heterogeneous graph is to explore proper semantic alignment between and linguistic domains and knowledge reasoning to generate persuasive reasoning paths.
- The contextual voting module is for visual scene understanding with a global perspective at the low-level features. Some ambiguous semantics (rainy day) that lack of specific labels for detection and can not benefit from the labeled object bounding boxes and categories such as “person” and “dog” during training.

Our Method – HGL (Heterogeneous Graph Learning)

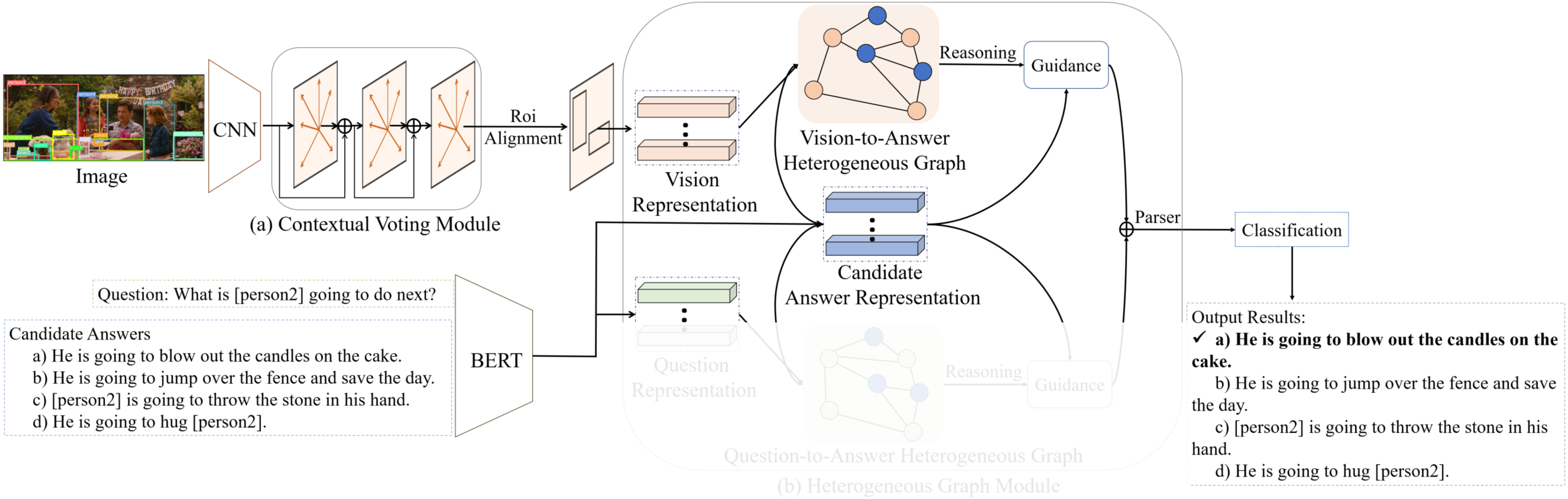


Image
CNN
(a) Contextual Voting Module
Roi Alignment
Vision Representation
Vision-to-Answer Heterogeneous Graph
Reasoning
Guidance
Candidate Answer Representation
BERT
Question Representation
Reasoning
Guidance
Question-to-Answer Heterogeneous Graph
(b) Heterogeneous Graph Module
Parser
Classification

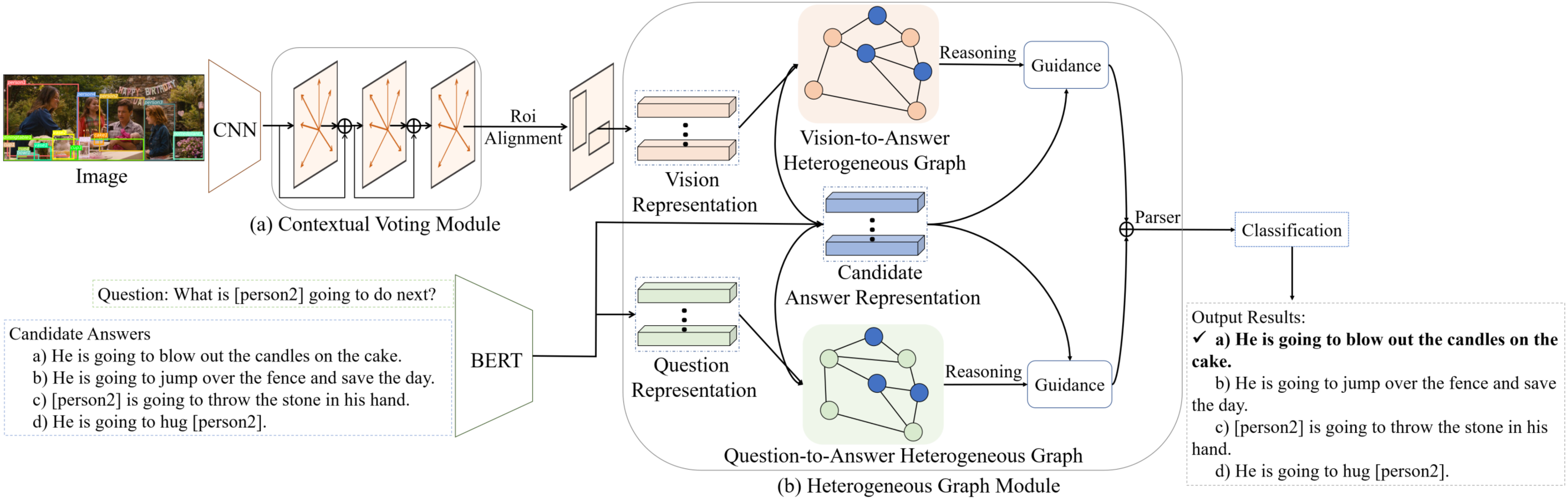
Question: What is [person2] going to do next?

Candidate Answers
a) He is going to blow out the candles on the cake.
b) He is going to jump over the fence and save the day.
c) [person2] is going to throw the stone in his hand.
d) He is going to hug [person2].

Output Results:
✓ **a) He is going to blow out the candles on the cake.**
b) He is going to jump over the fence and save the day.
c) [person2] is going to throw the stone in his hand.
d) He is going to hug [person2].

- The goal of heterogeneous graph is to explore proper semantic alignment between and linguistic domains and knowledge reasoning to generate persuasive reasoning paths.
- The contextual voting module is for visual scene understanding with a global perspective at the low-level features. Some ambiguous semantics (rainy day) that lack of specific labels for detection and can not benefit from the labeled object bounding boxes and categories such as “person” and “dog” during training.

Our Method – HGL (Heterogeneous Graph Learning)



(a) Contextual Voting Module

(b) Heterogeneous Graph Module

Question: What is [person2] going to do next?

Candidate Answers

a) He is going to blow out the candles on the cake.

b) He is going to jump over the fence and save the day.

c) [person2] is going to throw the stone in his hand.

d) He is going to hug [person2].

Output Results:

✓ **a) He is going to blow out the candles on the cake.**

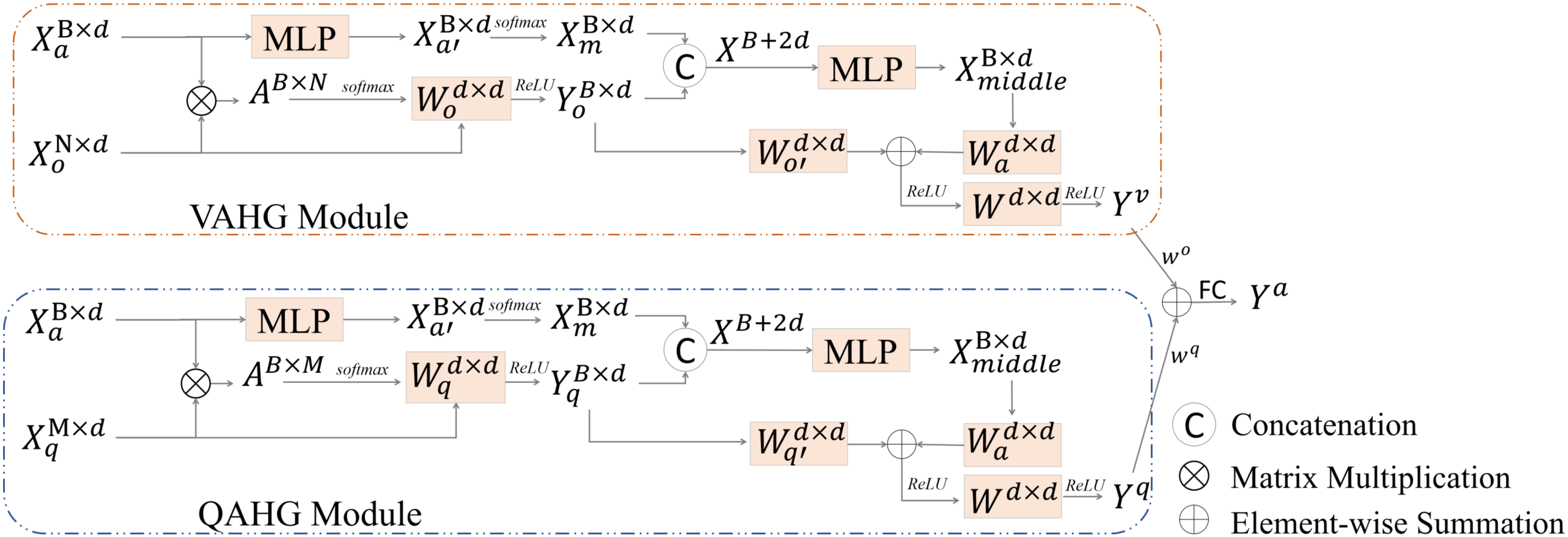
b) He is going to jump over the fence and save the day.

c) [person2] is going to throw the stone in his hand.

d) He is going to hug [person2].

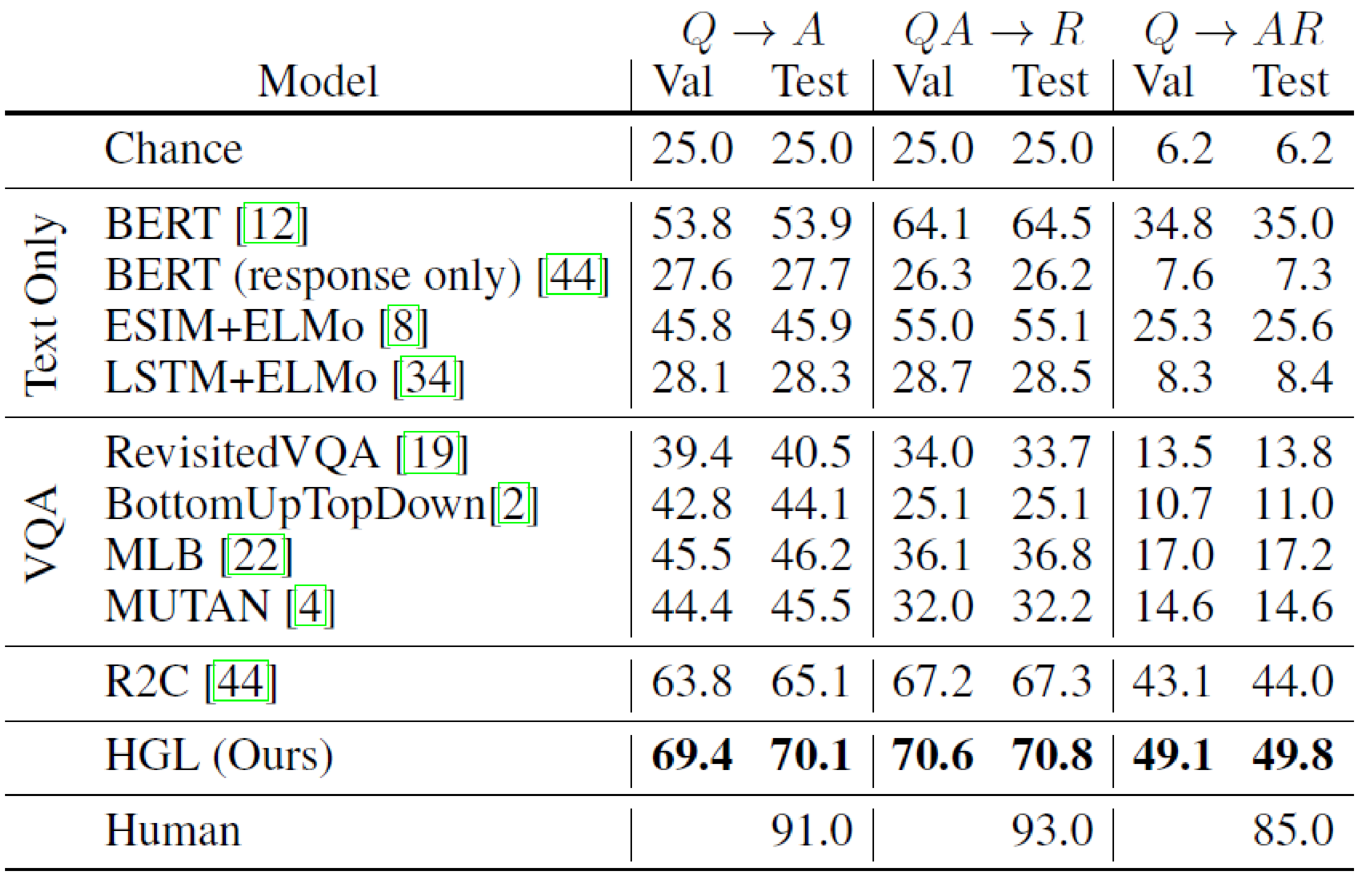
- The goal of heterogeneous graph is to explore **proper semantic alignment** between and linguistic domains and knowledge reasoning to generate persuasive reasoning paths.
- The contextual voting module is for visual scene understanding with a global perspective at the low-level features. Some **ambiguous semantics** (rainy day) that **lack of specific labels** for detection and can not benefit from the labeled object **bounding boxes and categories** such as "person" and "dog" during training.

Our Method – HGL (Heterogeneous Graph Learning)



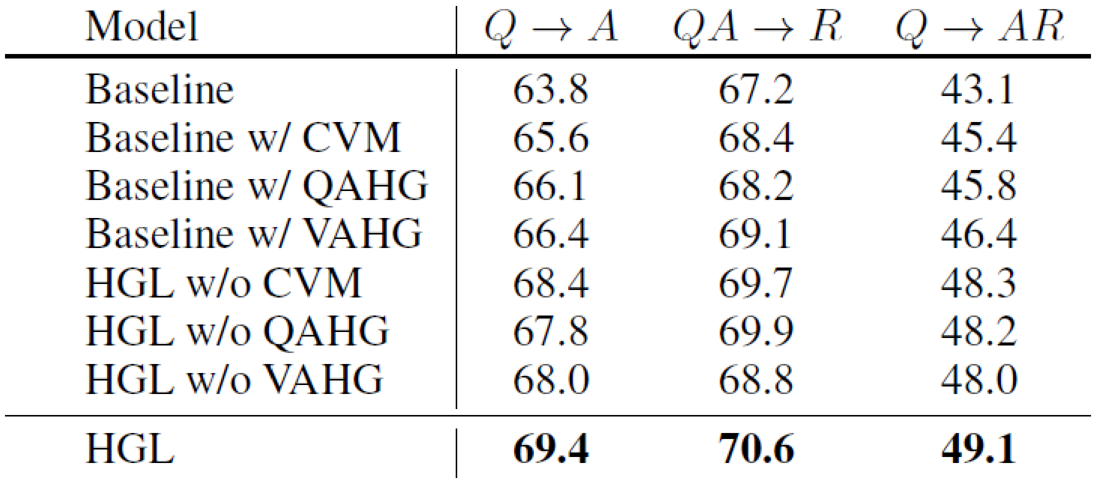
- The implementation details of our **heterogeneous graphs** by taking the representation of image, question and answer as inputs.

Experimental Results



	Model	$Q \rightarrow A$		$QA \rightarrow R$		$Q \rightarrow AR$	
		Val	Test	Val	Test	Val	Test
	Chance	25.0	25.0	25.0	25.0	6.2	6.2
Text Only	BERT [12]	53.8	53.9	64.1	64.5	34.8	35.0
	BERT (response only) [44]	27.6	27.7	26.3	26.2	7.6	7.3
	ESIM+ELMo [8]	45.8	45.9	55.0	55.1	25.3	25.6
	LSTM+ELMo [34]	28.1	28.3	28.7	28.5	8.3	8.4
VQA	RevisitedVQA [19]	39.4	40.5	34.0	33.7	13.5	13.8
	BottomUpTopDown[2]	42.8	44.1	25.1	25.1	10.7	11.0
	MLB [22]	45.5	46.2	36.1	36.8	17.0	17.2
	MUTAN [4]	44.4	45.5	32.0	32.2	14.6	14.6
	R2C [44]	63.8	65.1	67.2	67.3	43.1	44.0
	HGL (Ours)	**69.4**	**70.1**	**70.6**	**70.8**	**49.1**	**49.8**
	Human		91.0		93.0		85.0

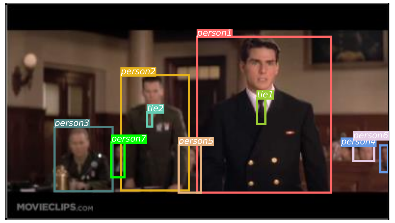
Table 1: Main results of validation and test dataset on VCR with respect to three tasks. Note that we do not need any extra information such as additional data or features.



Model	$Q \rightarrow A$	$QA \rightarrow R$	$Q \rightarrow AR$
Baseline	63.8	67.2	43.1
Baseline w/ CVM	65.6	68.4	45.4
Baseline w/ QAHG	66.1	68.2	45.8
Baseline w/ VAHG	66.4	69.1	46.4
HGL w/o CVM	68.4	69.7	48.3
HGL w/o QAHG	67.8	69.9	48.2
HGL w/o VAHG	68.0	68.8	48.0
HGL	**69.4**	**70.6**	**49.1**

Table 2: Ablation studies for our HGL on three tasks over the validation set.

Experimental Results

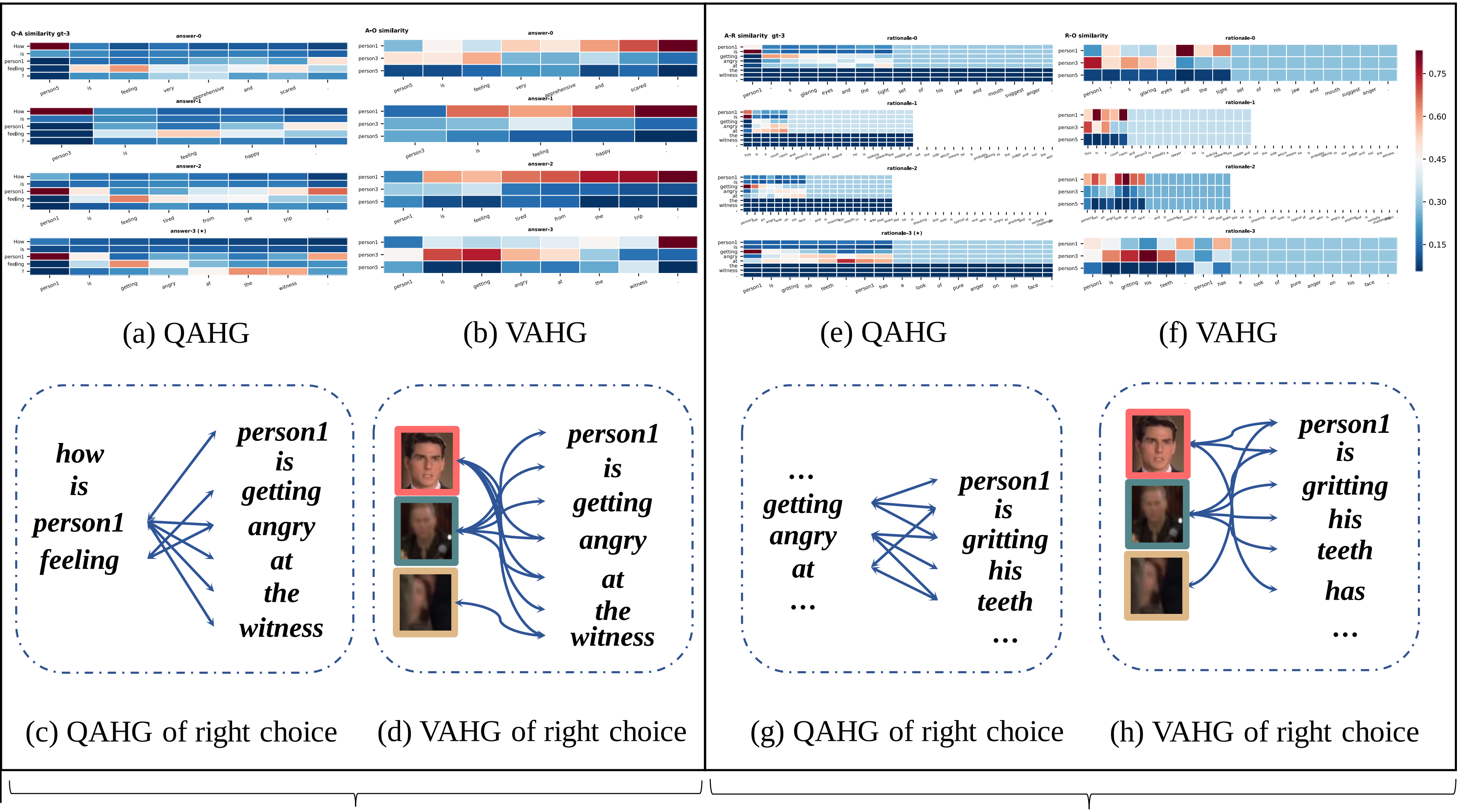


Q: How is [person1] feeling?

a) [person5] is feeling very apprehensive and scared.
b) [person3] is feeling happy.
c) [person1] is feeling tired from the trip.
d) [person1] is getting angry at the witness. ✓

R: d) is right because…

a) [person1]'s glaring eyes and the tight set of his jaw and mouth suggest anger.
b) This is a courtroom and [person3] is probably a lawyer. He is looking towards the middle and not the side which means he is probably talking to the judge and not the witness.
c) [person1] has an angry look on his face, and is moving his mouth in a way that looks like he is shouting, this look is typical of one who is angry at another and is verbally challenging them.
d) [person1] is gritting his teeth. [person1] has a look of pure anger on his face. ✓



(a) QAHG (b) VAHG (e) QAHG (f) VAHG

(c) QAHG of right choice (d) VAHG of right choice (g) QAHG of right choice (h) VAHG of right choice

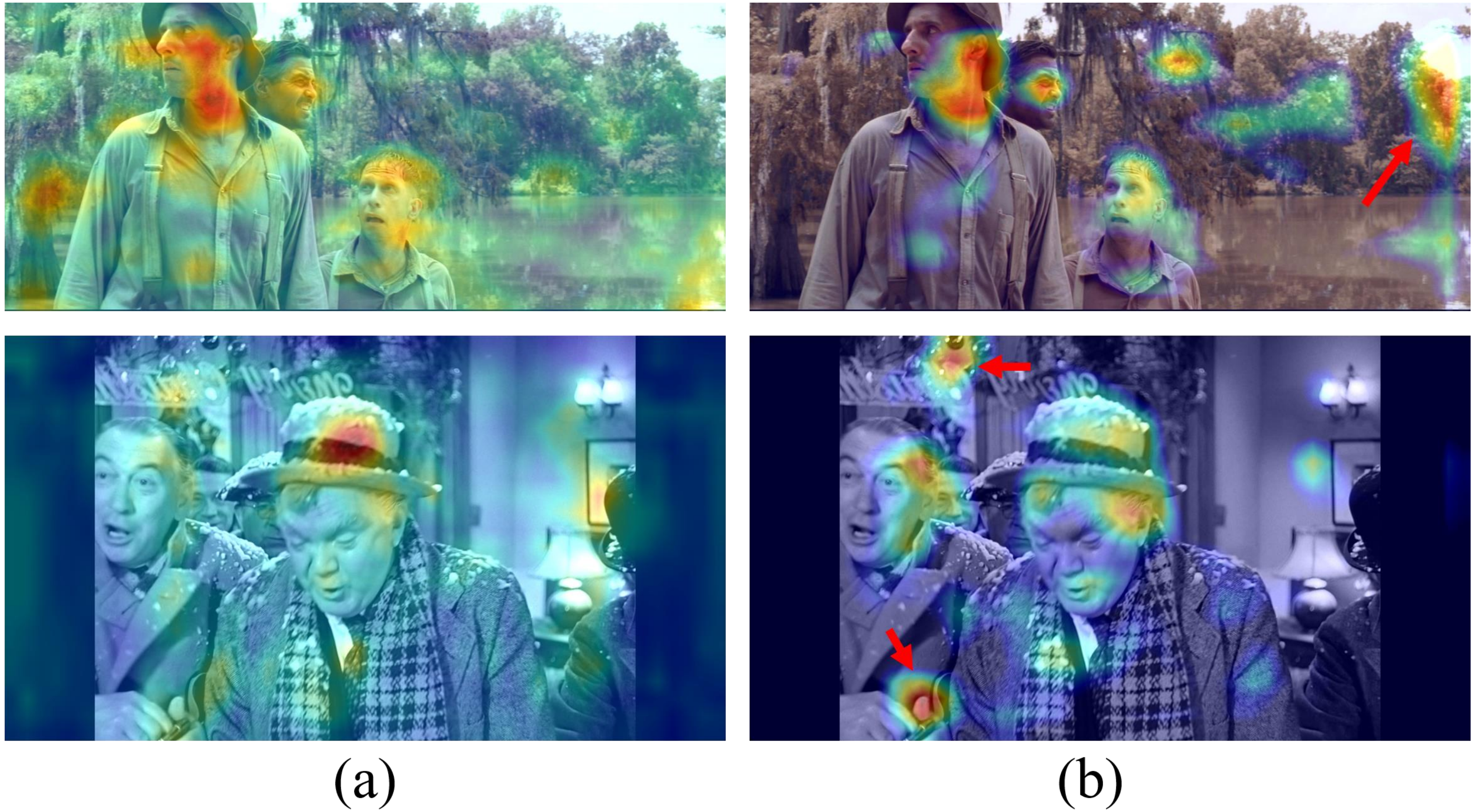
Answer Reason

- ➢ The predicted result is shown as **bold** font, and the ground truth (GT) is shown as ✓.

Experimental Results

Q: What if [person2] fell?
A: Person2 would get wet.
R: Preson2 is surrounded by water.

Q: Is it snowing outside?
A: Yes, it is snowing.
R: [person4] is dressed in a hat, scarf and a big jacket, his hat and shoulders are covered in white snowflakes.



(a) Baseline (b) our HGL

Conclusion & Future Work

The key merits of our work lie in four aspects:

- a framework called HGL is introduced to seamlessly integrate the intra-graph and inter-graph in order to bridge vision and linguistic domain, which consists of a heterogeneous graph module and a CVM;
- a heterogeneous graph module is proposed including a primal VAHG and a dual QAHG to collaborate with each other via heterogeneous graph reasoning and guidance mechanism;
- a CVM is presented to provide a new perspective for global reasoning;
- extensive experiments have demonstrated the state-of-the-art performance of our proposed HGL on three cognition-level tasks.

Several thoughts:

- Characteristics of natural language, such causal relationship.
- The reasoning for the specific number, such as 2 > 1.
- The interaction between visual instance relationships and linguistic contextual semantics

Our code is available in https://github.com/yuweijiang/HGL-pytorch *thank you for your listening*