Reasoning Visual Dialog with Sparse Graph Learning and Knowledge Transfer





Gi-Cheon Kang Junseok Park Hwaran Lee Byoung-Tak Zhang† Jin-Hwa Kim†

EMNLP 2021 Findings

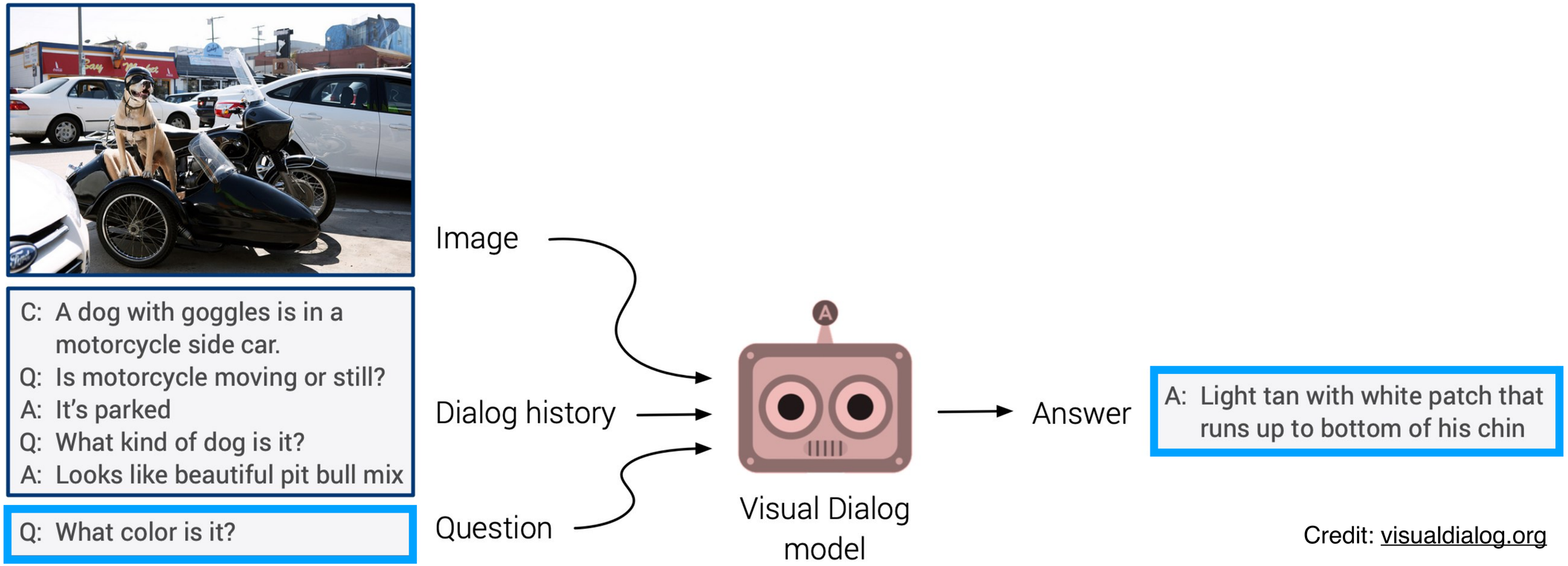
(† corresponding authors)

SEOUL NATIONAL UNIVERSITY

NAVER AI LAB

What is Visual Dialog?

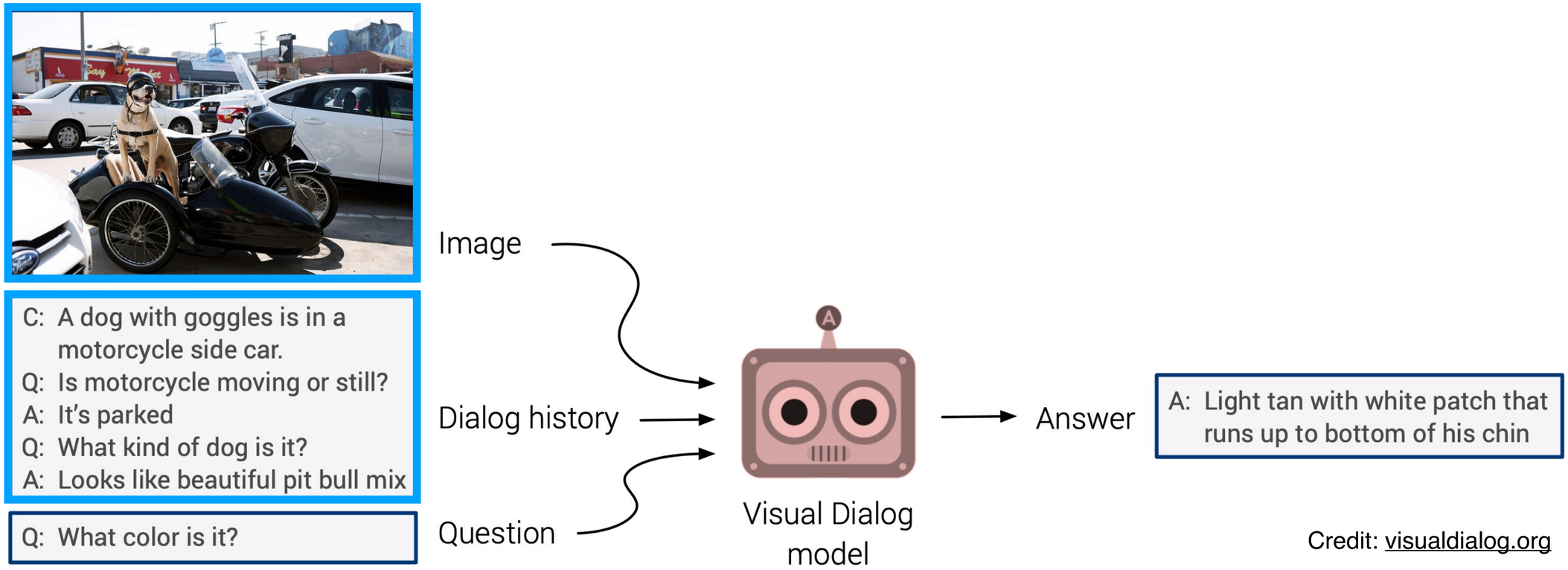
- **Answer a sequence of questions grounded in an image**
- Image and dialog history as a context



Credit: visualdialog.org

What is Visual Dialog?

- Answer a sequence of questions grounded in an image
- **Image and dialog history as context**



Credit: visualdialog.org

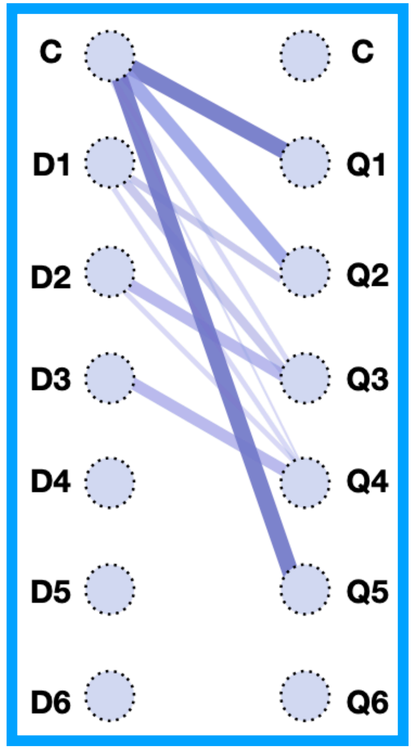
Motivation

- Challenge 1: reasoning over underlying **semantic structures** among a series of utterances

Image

C	The little girl with the pink jacket leads the tan colored horse
D1	How old does the girl look? ↳ Three
D2	Is she alone? ↳ No
D3	How many other people are there? ↳ Two other adults
D4	Do you think they are her parents? ↳ No
D5	What color is the horse? ↳ Light brown
D6	Do you see a fence anywhere? ↳ Yes

Dialog



Sparse structure
(Ours)

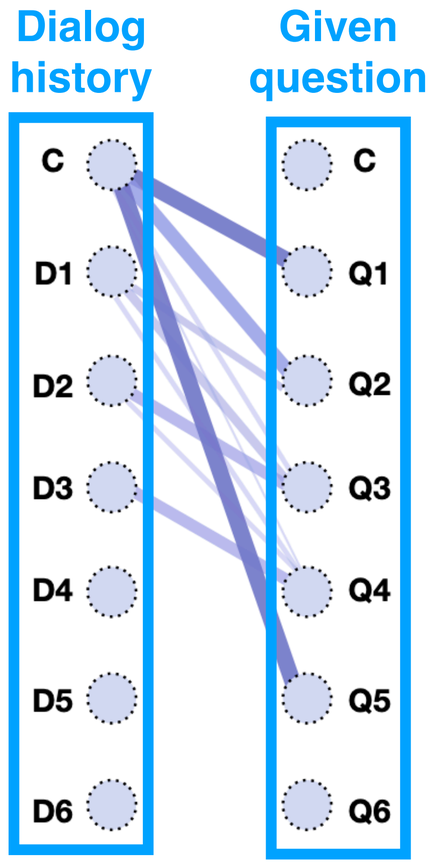
Motivation

- Challenge 1: reasoning over underlying **semantic structures** among a series of utterances

Image

C	The little girl with the pink jacket leads the tan colored horse
D1	How old does the girl look? ↳ Three
D2	Is she alone? ↳ No
D3	How many other people are there? ↳ Two other adults
D4	Do you think they are her parents? ↳ No
D5	What color is the horse? ↳ Light brown
D6	Do you see a fence anywhere? ↳ Yes

Dialog



Sparse structure
(Ours)

Motivation

- Challenge 1: reasoning over underlying semantic structures among a series of utterances
- Previous methods typically inferred the **dense structures** with **soft-attention**
  (fully-connected)

C	The little girl with the pink jacket leads the tan colored horse
D1	How old does the girl look? ↳ Three
D2	Is she alone? ↳ No
D3	How many other people are there? ↳ Two other adults
D4	Do you think they are her parents? ↳ No
D5	What color is the horse? ↳ Light brown
D6	Do you see a fence anywhere? ↳ Yes

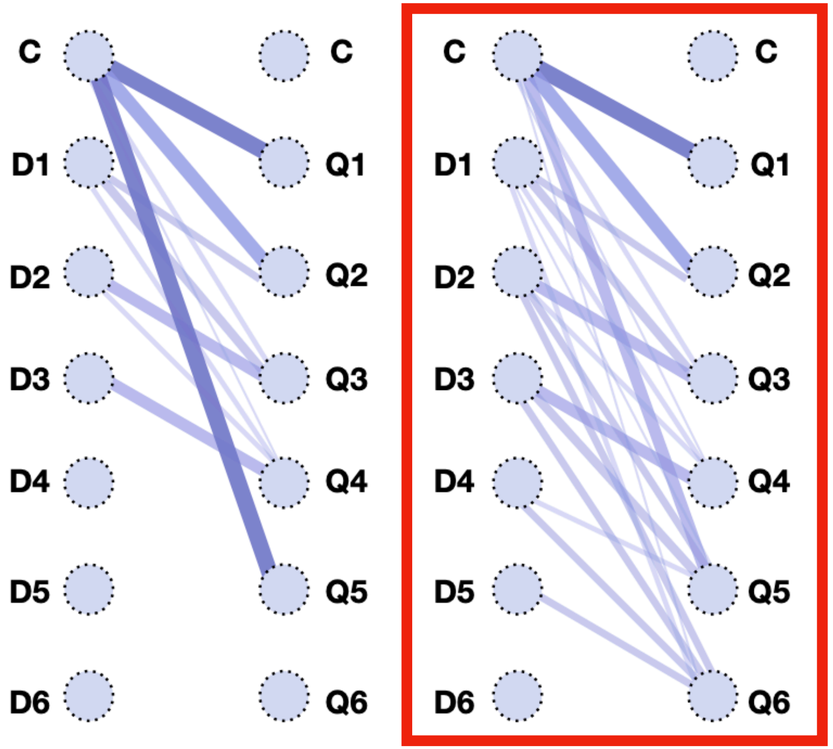


Image **Dialog** **Sparse structure** **(Ours)** **Dense structure**

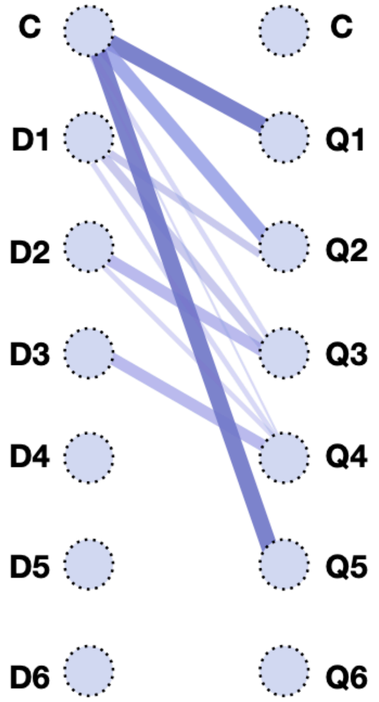
Motivation

- Challenge 1: reasoning over underlying semantic structures among a series of utterances
- Previous methods typically inferred the dense structures with soft-attention
- **Should we use the context, even when the context-free question is given?**

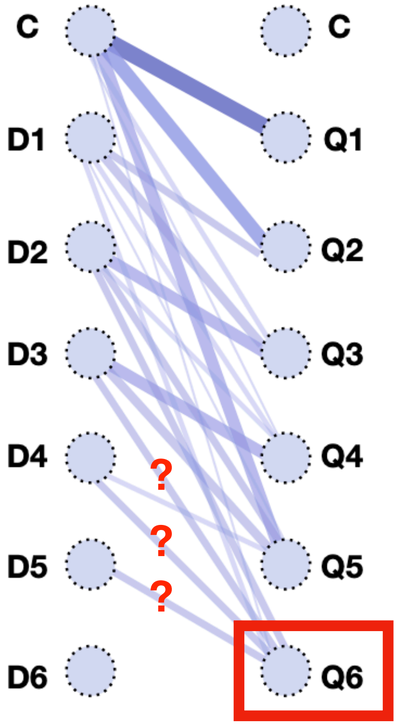
Image

C	The little girl with the pink jacket leads the tan colored horse
D1	How old does the girl look? ↳ Three
D2	Is she alone? ↳ No
D3	How many other people are there? ↳ Two other adults
D4	Do you think they are her parents? ↳ No
D5	What color is the horse? ↳ Light brown
D6	Do you see a fence anywhere? ↳ Yes

Dialog



Sparse structure
(Ours)



Dense structure

Motivation

- Challenge 2: Identifying **several appropriate answers** to the given question

Very tall-looking girl posing with a skateboard propped in front of her

D1 : What color is the girl's hair? Dark blonde
D2 : What color is the skateboard? Green

...

D4 : Is she wearing pants? Yes, jeans
Q5 : Are the jeans blue?

Top-k Predicted Answers

- Yes they are (Ground-truth answer)
- Yes
- Blue
- Yes I think so
- …

Motivation

- Challenge 2: Identifying several appropriate answers to the given question
- Existing studies focus on finding the single ground-truth answer with one-hot labels
- **The one-hot encoded labels could suppress several plausible answers**

Very tall-looking girl posing with a skateboard propped in front of her

D1 : What color is the girl's hair? Dark blonde
D2 : What color is the skateboard? Green

...

D4 : Is she wearing pants? Yes, jeans
Q5 : Are the jeans blue?

Top-k Predicted Answers

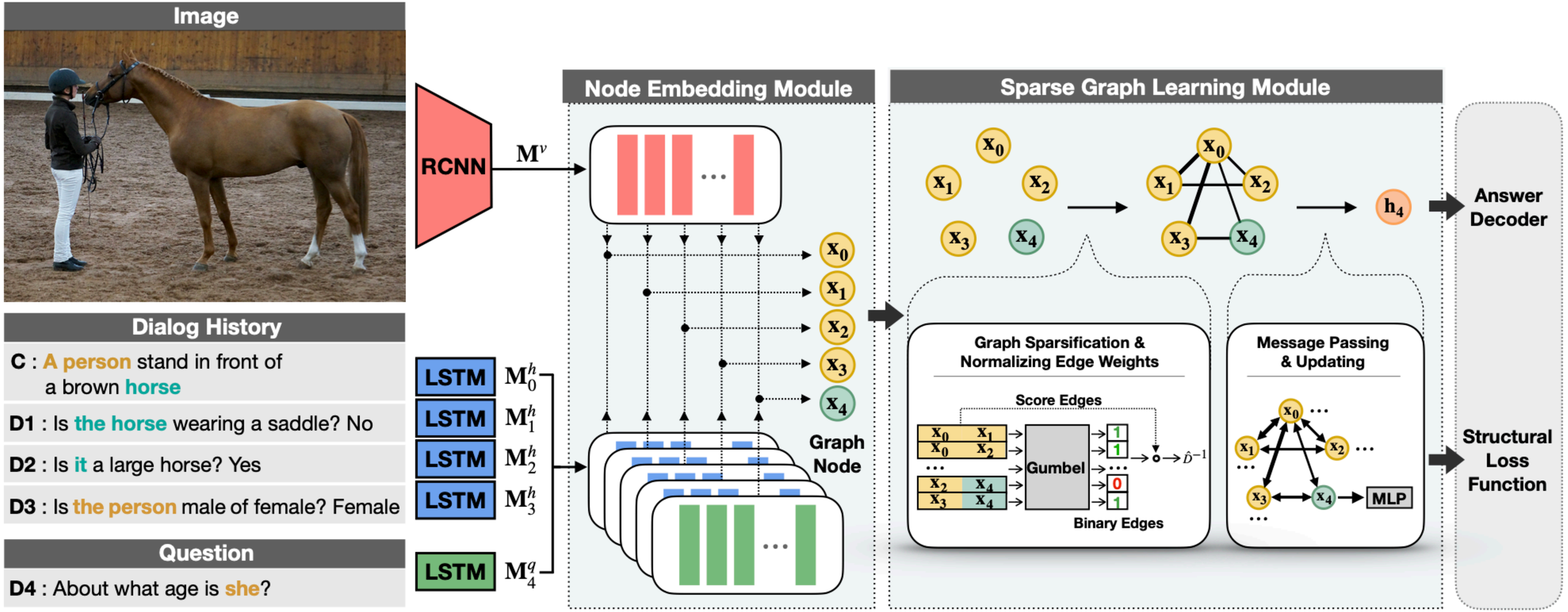
- Yes they are (Ground-truth answer)
- No
- No it's gray
- One is
- …
- Yes
- Blue
- Yes I think so

Proposed Methods

- Challenge 1: reasoning over underlying semantic structures among a series of utterances
  ↳ **Sparse Graph Learning**

- Challenge 2: Identifying several appropriate answers to the given question
  ↳ **Knowledge Transfer**

Sparse Graph Learning



Sparse Graph Learning

Input Features

Visual features: $\mathbf{M}^v \in \mathbb{R}^{K \times d_h}$

Language features (dialog history): $\{\mathbf{M}_i^h\}_{i=0}^{t-1} \in \mathbb{R}^{t \times L \times d_h}$

Language features (question): $\mathbf{M}_t^q \in \mathbb{R}^{L \times d_h}$

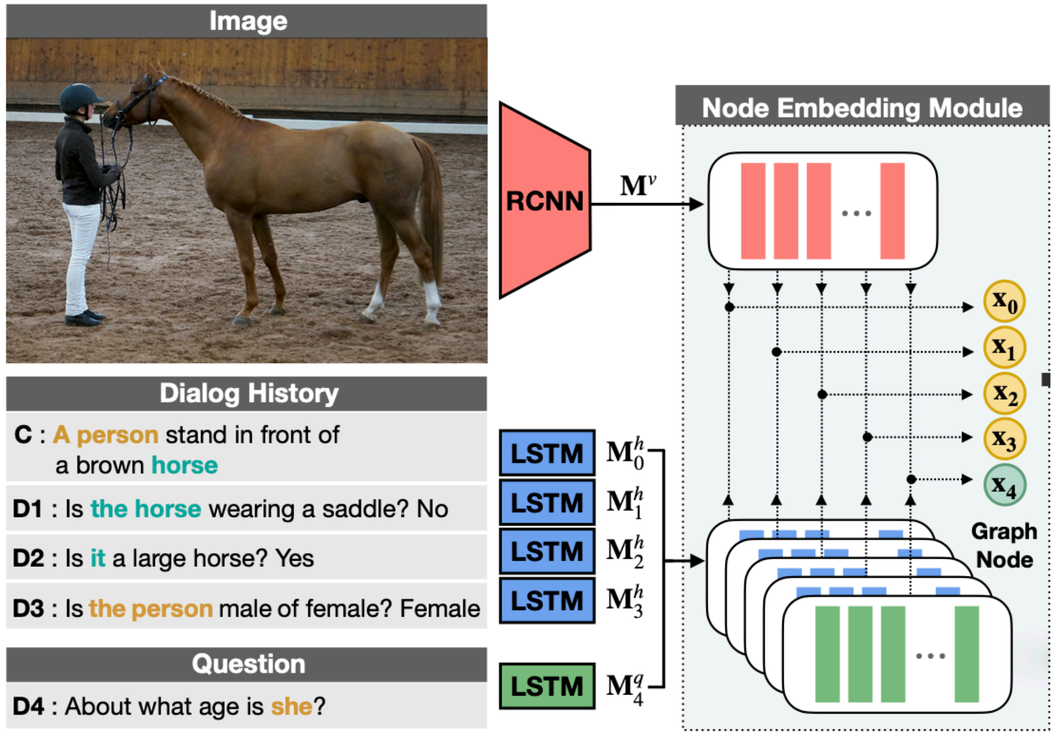
Node Embedding Module

Joint embedding of visual and linguistic representations

$$\mathbf{x}_t = f_{ne}(\mathbf{M}^v, \mathbf{M}_t^q)$$

Each round of the dialog history is also embedded. Finally we get t+1 nodes of the graph represented as a set of vectors

$$\mathbf{X} \in \mathbb{R}^{(t+1) \times d_h}$$



Sparse Graph Learning

Sparse Graph Learning Module

Input: $\mathbf{X} \in \mathbb{R}^{(t+1)\times d_h}$

Binary edges: $\mathbf{A}^b_{ij} = z_{ij} \sim Categorical(\mathbf{p}_{ij})$

$$\mathbf{p}_{ij} = \text{softmax}\Big(\mathbf{W}_c(\mathbf{x}_i \circ \mathbf{x}_j)^\top / \tau\Big)$$

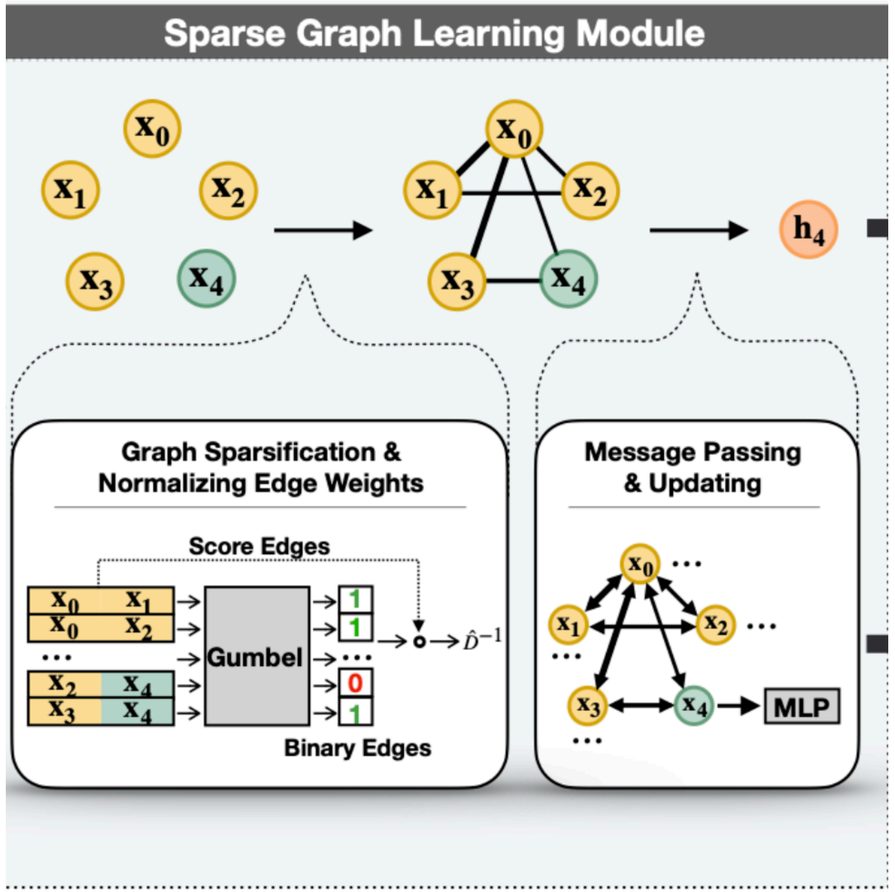
$$z_{ij} = \begin{cases} 1, & \text{if } \underset{k\in\{0,1\}}{\text{argmax}}\big(\log(p_k) + g_k\big) = 1 \\ 0, & \text{otherwise} \end{cases}$$

Score edges: $\mathbf{A}^s_{ij} = (\mathbf{x}_i\mathbf{x}_j^\top)^2$

Sparse weighted edges: $\hat{\mathbf{A}}_{ij} = \mathbf{A}^b_{ij}\mathbf{A}^s_{ij} = z_{ij}(\mathbf{x}_i\mathbf{x}_j^\top)^2$

Message passing & update: $\mathbf{M} = F_M(\mathbf{X}, \hat{\mathbf{A}}) = \hat{\mathbf{D}}^{-1}\hat{\mathbf{A}}\mathbf{X}\mathbf{W}_m$

$$\mathbf{H} = F_U(\mathbf{X}, \mathbf{M}) = f_u(\mathbf{X} + \mathbf{M})$$

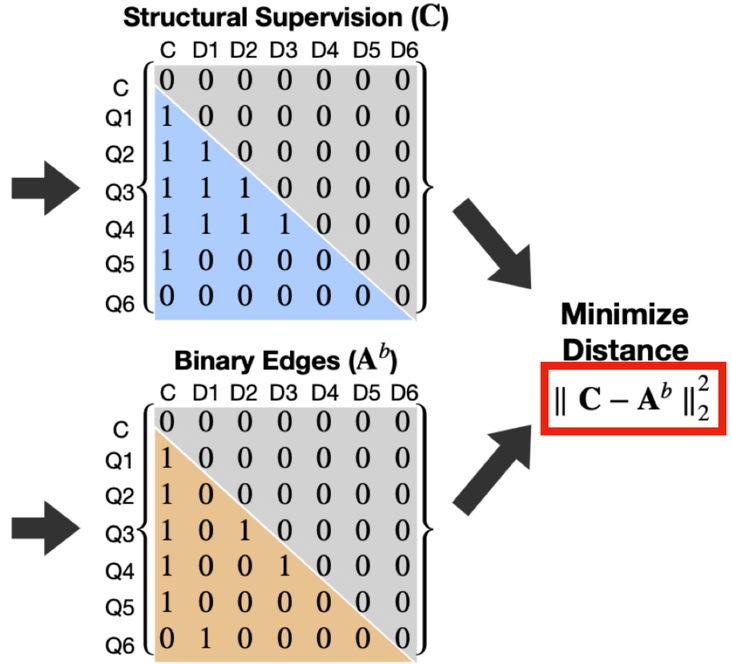


Sparse Graph Learning

Structural Learning

Object function for reasoning improved binary edges

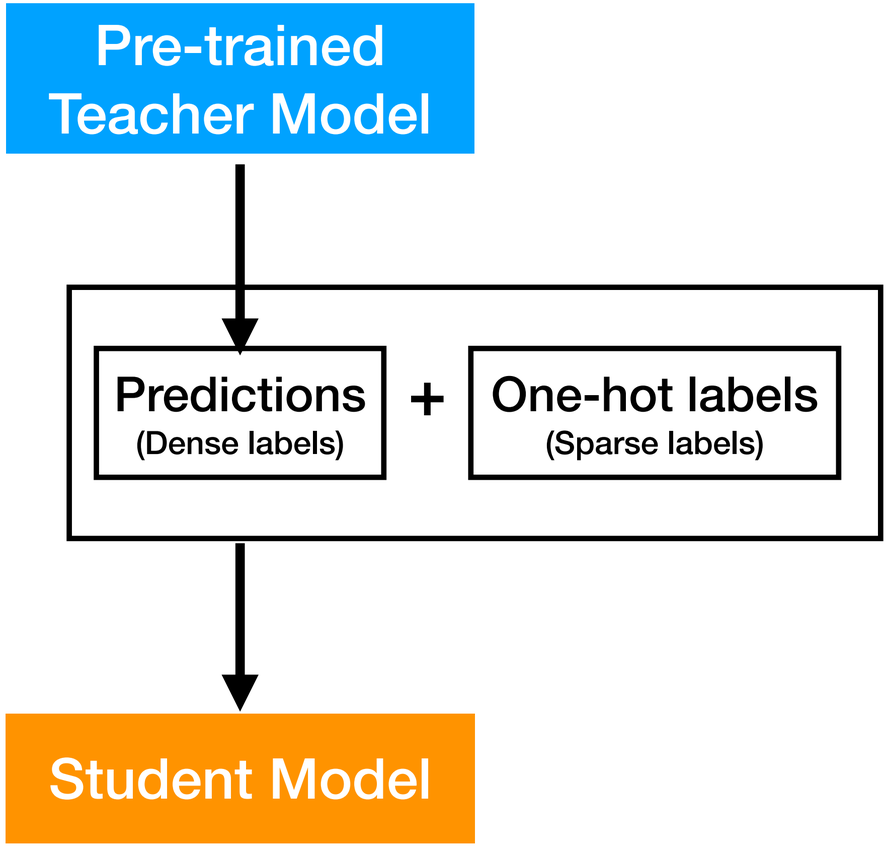
C	The little girl with the pink jacket leads the tan colored horse
D1	How old does the girl look? ↳ Three
D2	Is she alone? ↳ No
D3	How many other people are there? ↳ Two other adults
D4	Do you think they are her parents? ↳ No
D5	What color is the horse? ↳ Light brown
D6	Do you see a fence anywhere? ↳ Yes



(a) Image (b) Dialogs (c) Structural Loss Function

Knowledge Transfer

- Knowledge Transfer technique extracts the soft scores of each candidate answer from the pre-trained teacher model (Qi et al., 2020), and uses these scores as **pseudo labels**
- Combining predictions from the teacher model and the one-hot labels
- We cast visual dialog as a **regression task** that predicts the correctness of each candidate answers individually



Evaluation Metrics

★ **Mean Reciprocal Rank (MRR)** - $\mathrm{MRR} = \frac{1}{\mathrm{Q}} \sum_{i=1}^{\mathrm{Q}} \frac{1}{rank_i^{gt}}$

Recall@k, k ∈ {1, 5, 10} - existence of ground truth answer in top-k ranked list

Mean Rank (Mean) - mean rank of the ground truth answer

★ **Normalized Discounted Cumulative Gain (NDCG)** - answer *relevance*

Answer options : [“two”, “yes”, “probably”, “no”, “yes it is”]

Ground-truth relevances : [0, 1.0, 0.5, 0, 1.0] (collecting dense annotations)

Ideal ranking of answer options : [“yes”, “yes it is”, “probably”, “two”, “no”]

Submitted ranking of answer options : [“yes”, “yes it is”, “two”, “probably”, “no”]

$$\mathrm{NDCG} = \frac{DCG_{submitted}}{DCG_{ideal}} \approx \frac{1.63}{1.88} \approx 0.87 \qquad \mathrm{DCG} = \sum_{j=1} \frac{relevance_j}{log_2(j+1)}$$

NDCG penalizes the lower rank of candidates with high relevance scores !

★ **Overall** - mean value of MRR and NDCG

Quantitative Results

Model	Overall↑	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
GNN	57.10	52.82	61.37	47.33	77.98	87.83	4.57
CorefNMN	58.10	54.70	61.50	47.55	78.10	88.80	4.40
RvA	59.31	55.59	63.03	49.03	80.40	89.83	4.18
Synergistic	59.76	57.32	62.20	47.90	80.43	89.95	4.17
Synergistic‡	60.65	57.88	63.42	49.30	80.77	90.68	3.97
ReDAN	57.50	61.86	53.13	41.38	66.07	74.50	8.91
ReDAN+‡	59.10	64.47	53.73	42.45	64.68	75.68	6.63
DAN	60.40	57.59	63.20	49.63	79.75	89.35	4.30
DAN‡	62.14	59.36	64.92	51.28	81.60	90.88	3.92
HACAN	60.70	57.17	64.22	50.88	80.63	89.45	4.20
FGA	57.90	52.10	63.70	49.58	80.97	88.55	4.51
FGA‡	60.90	54.50	**67.30**	**53.40**	**85.28**	**92.70**	**3.54**
MCA†	55.08	72.47	37.68	20.67	56.67	72.12	8.89
P1+P2†	60.09	71.60	48.58	35.98	62.08	77.23	7.48
P1+P2†‡	63.32	74.02	52.62	40.03	68.85	79.15	6.76
VisDial-BERT†	62.60	74.47	50.74	37.95	64.13	80.00	6.28
VD-BERT	62.70	59.96	65.44	51.63	82.23	90.68	3.90
VD-BERT†	60.63	74.54	46.72	33.15	61.58	77.15	7.18
VD-BERT†‡	63.26	**75.35**	51.17	38.90	62.82	77.98	6.69
SGL	62.13	61.97	62.28	48.15	79.65	89.10	4.34
SGL+KT†	65.31	72.60	58.01	46.20	71.01	83.20	5.85
SGL+KT†‡	**66.03**	73.70	58.36	46.63	71.28	84.15	5.57

Table 1: Test-std performance of the discriminative model on the VisDial v1.0 dataset. ↑ indicates higher is better. ↓ indicates lower is better. † denotes the use of dense labels. ‡ denotes ensemble model.

Quantitative Results

Model	Overall	NDCG	MRR
Edgeless	60.75	61.96	59.54
Dense	61.05	58.85	63.25
Sparse-hard	61.44	59.71	63.16
P1+P2† (teacher model)	61.65	73.42	49.88
SGL w/o RPN	61.56	61.25	61.86
SGL w/o SS	61.66	62.46	60.85
SGL w/o MR	62.11	62.42	61.79
SGL	63.38	63.41	**63.34**
SGL+KT†	**66.82**	**74.54**	59.10

Table 2: Comparison with the baseline models on the VisDial v1.0 validation split. MR, SS, and RPN denote the use of multi-step reasoning, structural supervision, and region proposal network, respectively. † denotes the use of dense labels.

Baseline Models

Edgeless: a model that does not use the dialog history

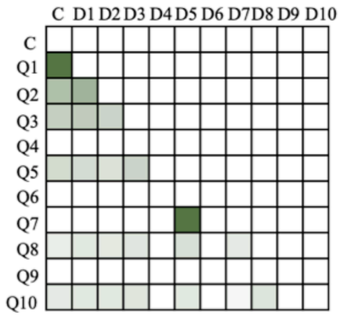
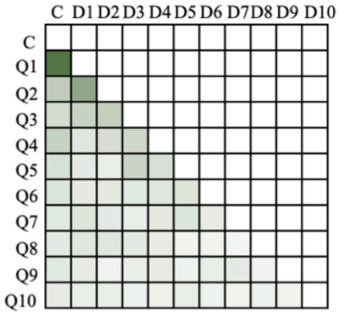
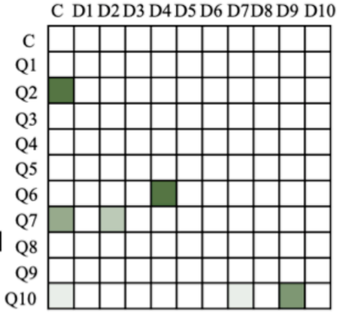
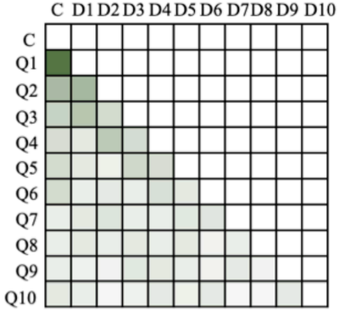
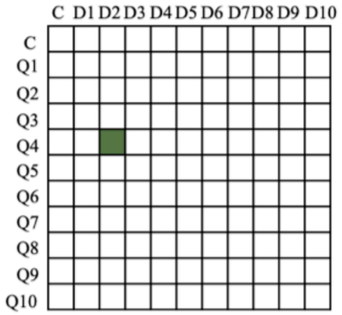
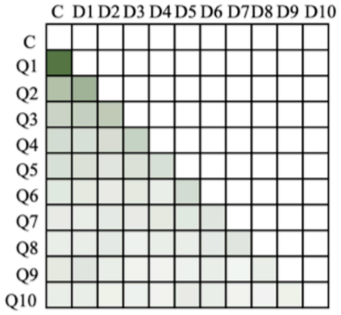
Dense: a model that infers dense structures

Sparse-hard: a model that picks exactly one edge weights

Model	F1-Score
Edgeless	0.0
Dense	0.246
Sparse-hard	0.279
SGL	0.714
SGL+KT	**0.748**

Table 3: Graph inference on VisDial v1.0 val split.

Qualitative Results

Image & Caption	Dialog	Sparse Structure (Ours)	Dense Structure
 The dog is pacing along the very grassy area	D1 : Is the dog large? No, very small D2 : What color is the dog? Tan and gray D3 : Is the dog a puppy? Nope D4 : Are there any people? I can't see any, no D5 : Does the dog have on a collar? Yes, or a bandana D6 : Are there any trees? Not that I can see D7 : What color is the collar? Red D8 : Is the dog fenced in? I can't tell, but I don't think so D9 : Can you see a house? No D10 : Is the dog standing? Yes		
 A bunk bed is in an old, wooden cabin with no sheets	D1 : Are there any people? No D2 : What color is the bunk bed? Brown D3 : What is the floor made of? Carpet D4 : Is there a window? Yes D5 : Is it night time? No D6 : What do you see through the window? A tree D7 : Is there a pillow on the bed? No D8 : Do you see other furniture? A table and another bed D9 : Is there a chair? Yes D10 : What's the chair made of? Wood		
 The red and yellow buses are on the sidewalk by a busy street	D1 : Can you see any people? Lots of people D2 : Can you see the sky? Yes a little D3 : Is there any logos or writing? Some logo on the car D4 : What color is the sky? Light blue D5 : Do you see any animals? No it is in the big city D6 : Do you see any trees? Lots of trees behind the bus D7 : Anything made out of wood? No D8 : Anything made out of glass? Cars, bus windows D9 : Anything made out of metal? Cars, bus, motorcycle D10 : Anything made out of plastic? Helmet, signs		

Qualitative Results

Image & Caption	Dialog History & Current Question	Predicted Answers	Predicted Answers

A brown bear sits on a big rock

D1 : Is this a real **bear**, rather than a toy? …
D2 : Is **it** very wooly? Very!
D3 : What season does it seem to be? …
…
D6 : Are any **other animals** visible? …
Q7 : Is there any trees or other greenery?

Top-5 Answers for Q7 (SGL)		Top-5 Answers for Q7 (Dense)	
Barely see some grass (Ground-truth)	0.62	It's definitely a real **bear**	0.46
No	0.21	Tons of underbrush	0.40
Tons of underbrush	0.05	In the background	0.07
No trees	0.03	Some trees in the distance	0.01
Some leaves at the top	0.02	No trees	0.01

Very tall-looking girl posing with a skateboard propped in front of her

D1 : What color is the girl's hair? Dark blonde
D2 : What color is the skateboard? Green
…
D4 : Is she wearing pants? Yes, jeans
Q5 : Are the jeans blue?

Top-5 Answers for Q5 (SGL)		Top-5 Answers for Q5 (SGL+KT)	
No	0.37	Yes	0.71
Yes	0.35	Yes they are (Ground-truth)	0.31
No it's grey	0.07	Yep	0.28
One is	0.03	Blue	0.21
Yes they are (Ground-truth)	0.02	Yes I think so	0.19

Figure 5: A visualization of the top five predicted answers from SGL+KT, SGL, and Dense. Note that SGL+KT utilizes the sigmoid activation function to compute the answer scores while the others use the softmax function.

Conclusions

- Address the Visual Dialog as a structure inference problem
- Propose Sparse Graph Learning to infer the sparse structures of the dialog
- **The Softmax-based dense structures distracts the AI agent from answering the correct answers**

- Introduce Knowledge Transfer technique to answer several plausible answers
- Knowledge Transfer retrieves consistent answers and improves the performance of visual dialog dramatically

Thank You !

Code: https://github.com/gicheonkang/sglkt-visdial
Paper: https://arxiv.org/pdf/2004.06698.pdf