VD-BERT: A Unified Vision and Dialog Transformer with BERT

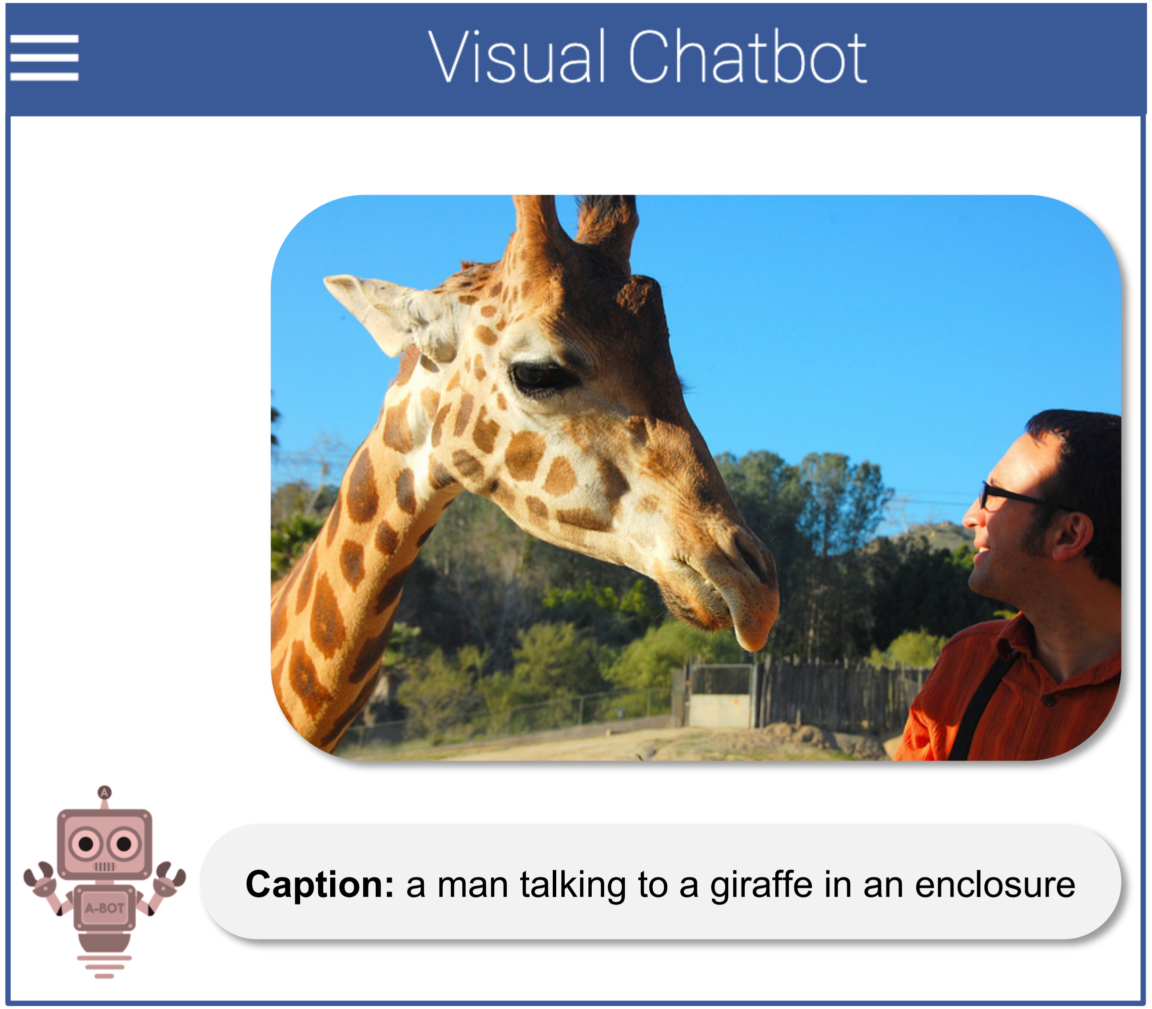
Yue Wang[1], Shafiq Joty[2], Michael R. Lyu[1], Irwin King[1], Caiming Xiong[2], Steven C.H. Hoi[2]

1. The Chinese University of Hong Kong   2. Salesforce Research

Code & Models: https://github.com/salesforce/VD-BERT

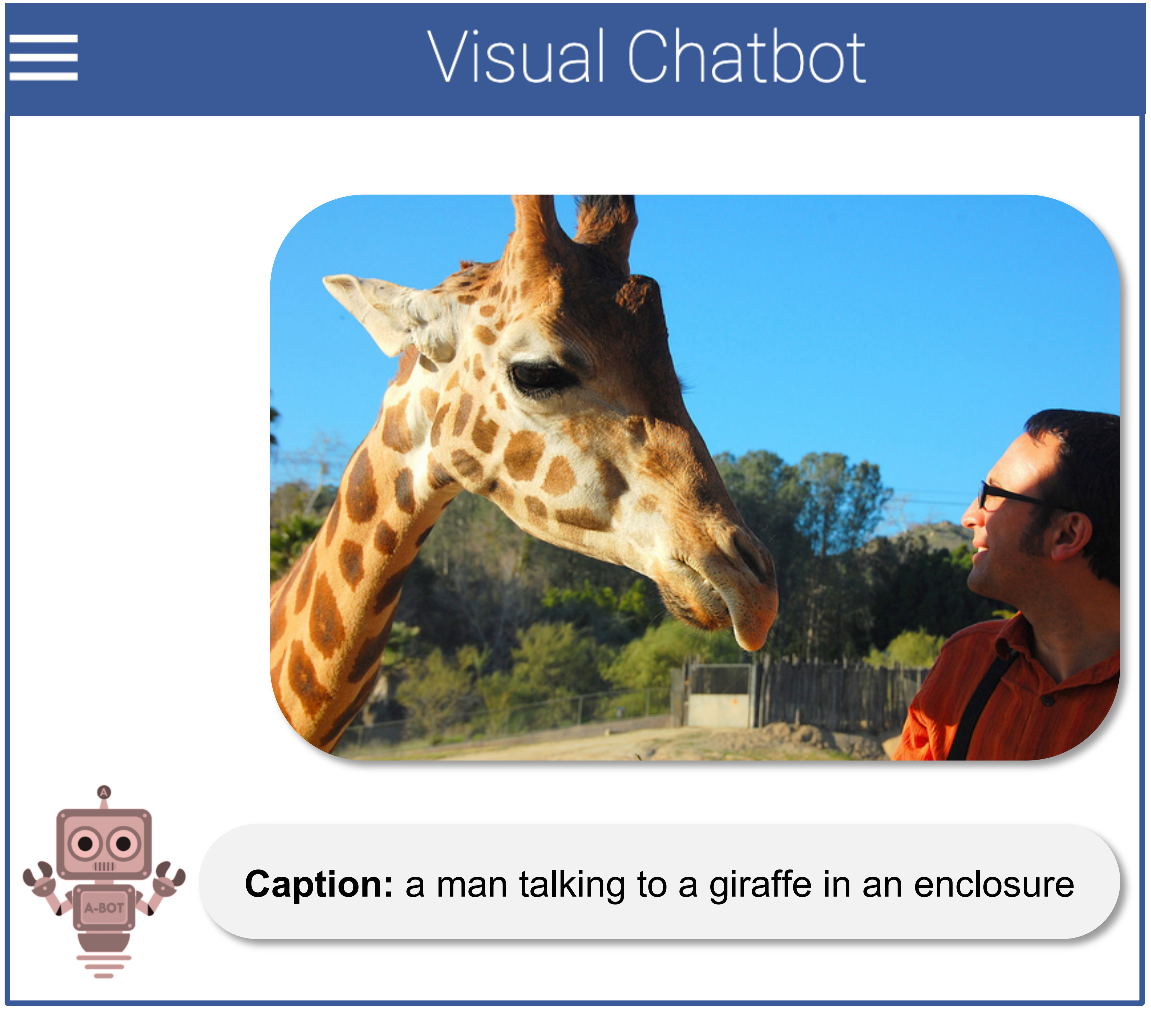
What is Visual Dialog?

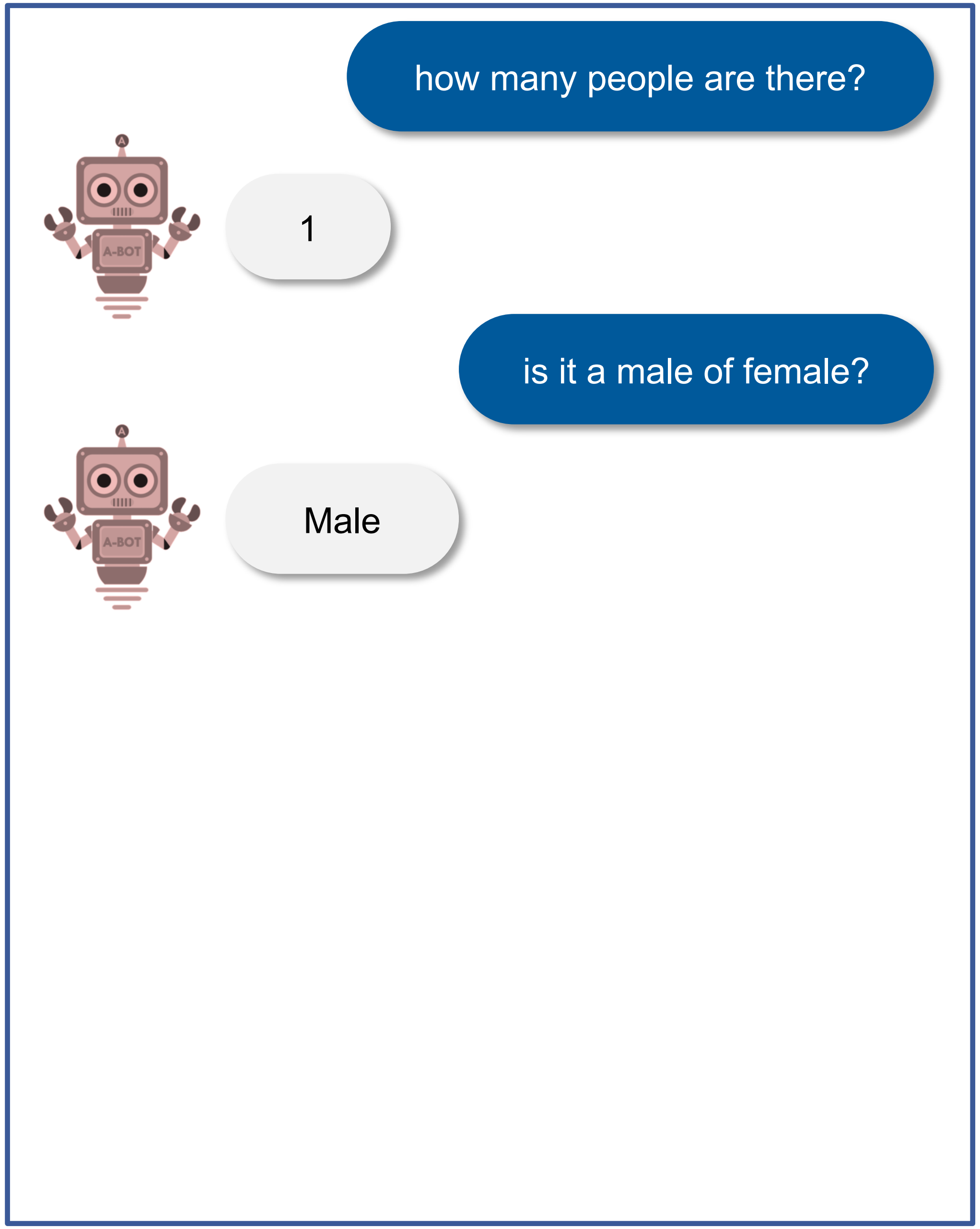
(Das et al., 2017)



What is Visual Dialog?

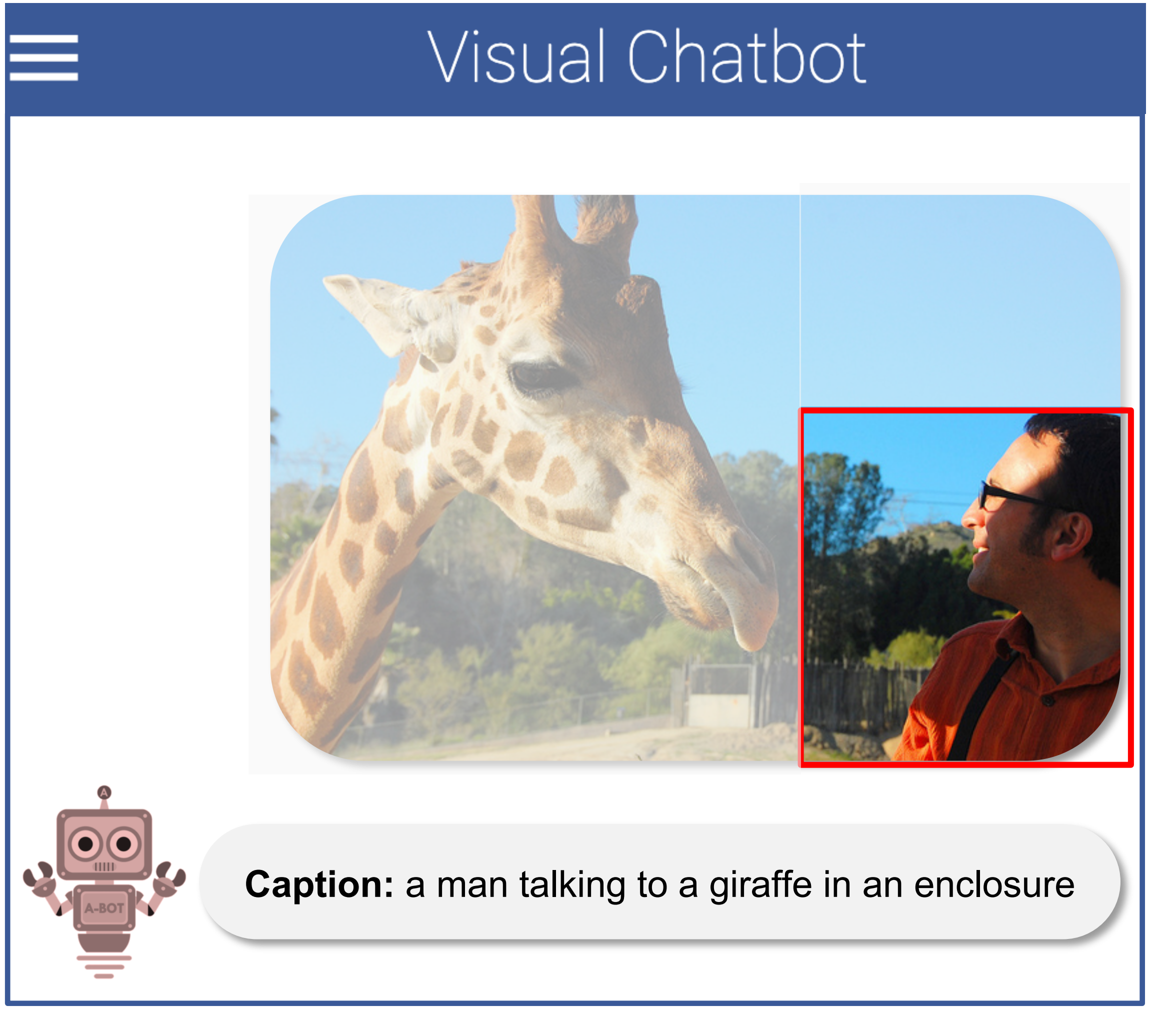
(Das et al., 2017)

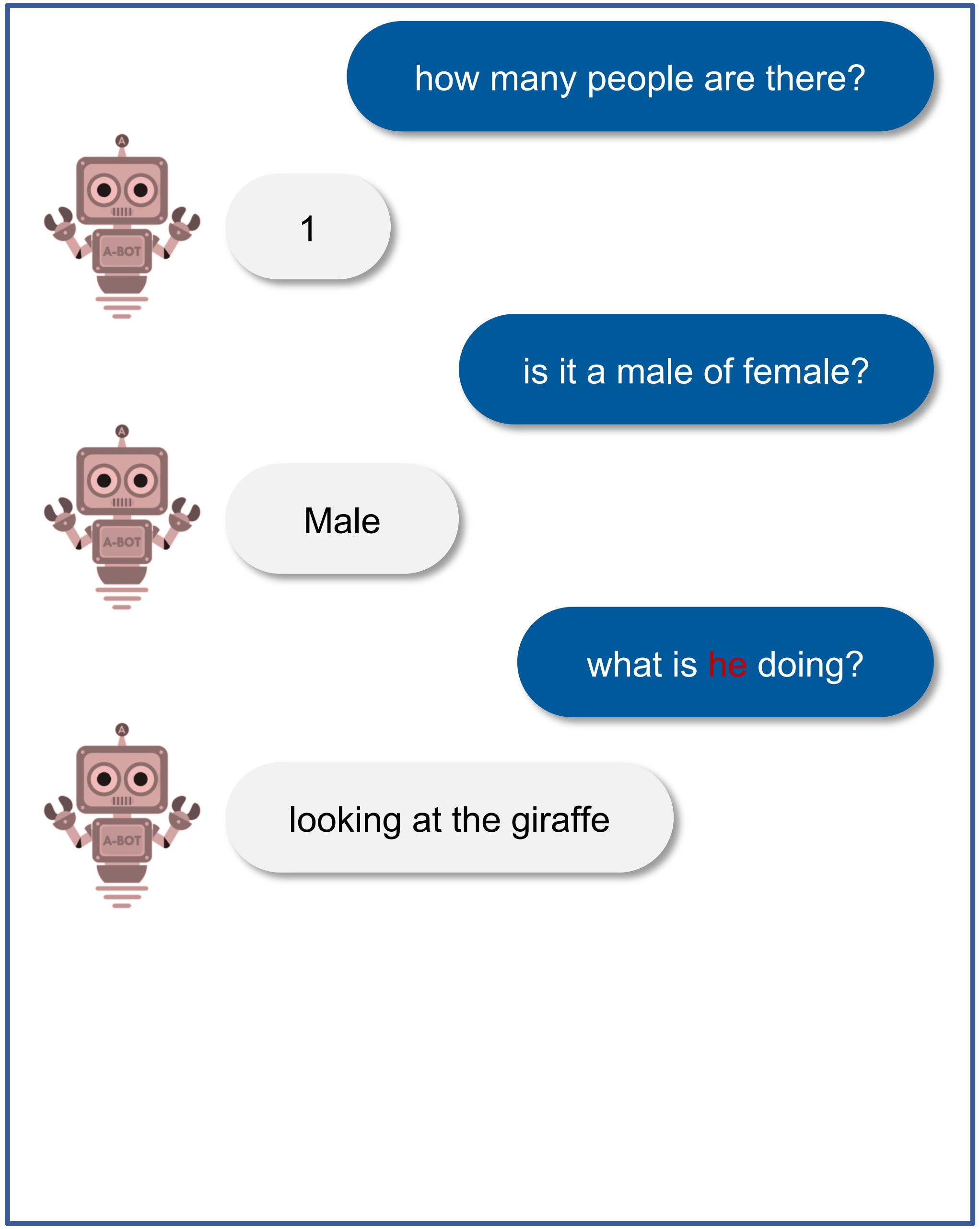




What is Visual Dialog?

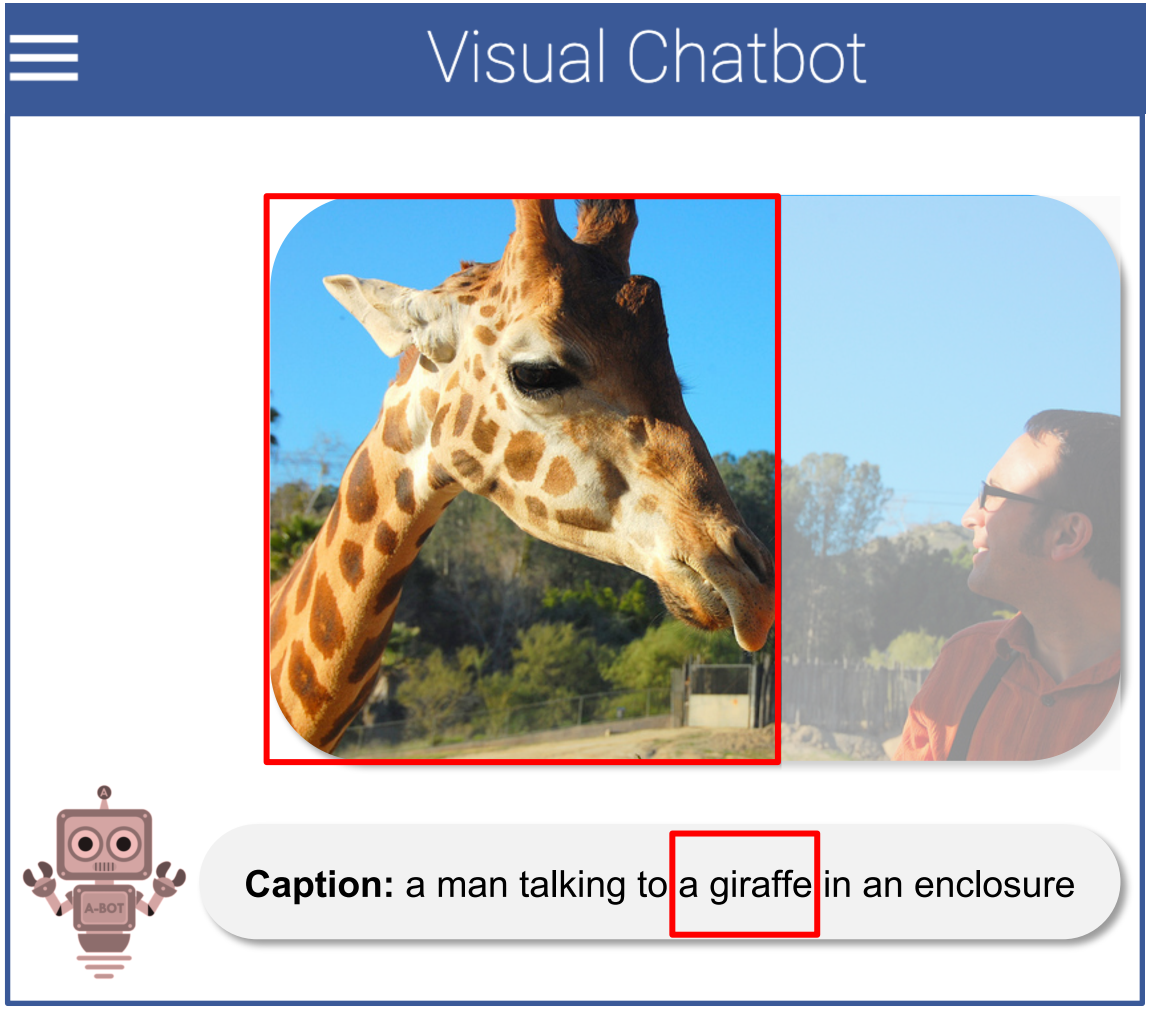
(Das et al., 2017)

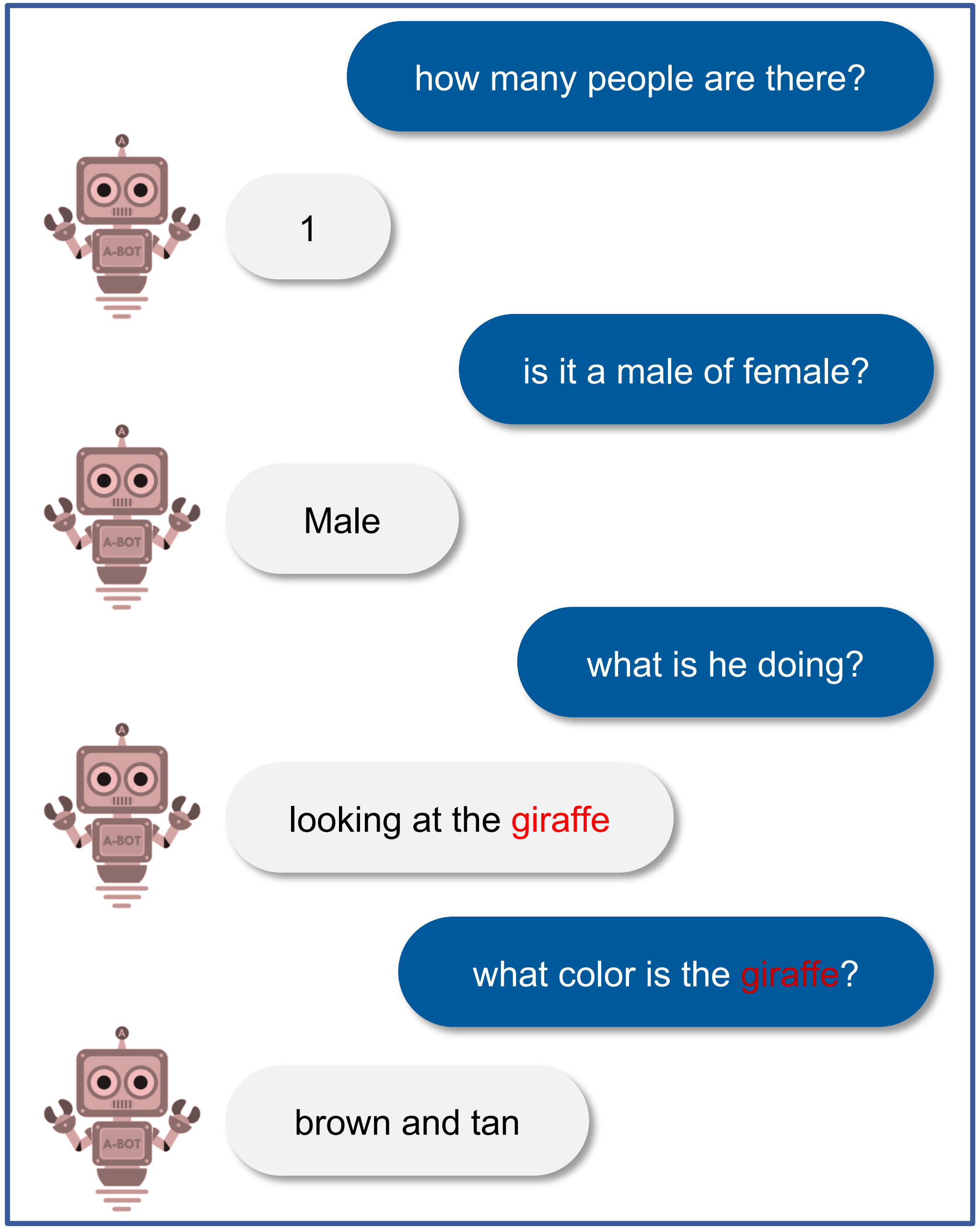




What is Visual Dialog?

(Das et al., 2017)





Visual Dialog (VisDial)

Task Definition

Input:

- An Image I
- Dialog history
  - $H_t = \{C, (Q_1, A_1), \dots, (Q_{t-1}, A_{t-1})\}$
- A follow-up question Q_t

Predict an answer $\hat{A}_\mathrm{t}$

- By ranking 100 candidates $\{\hat{A}_\mathrm{t}^1, \hat{A}_\mathrm{t}^2, \dots, \hat{A}_\mathrm{t}^{100}\}$

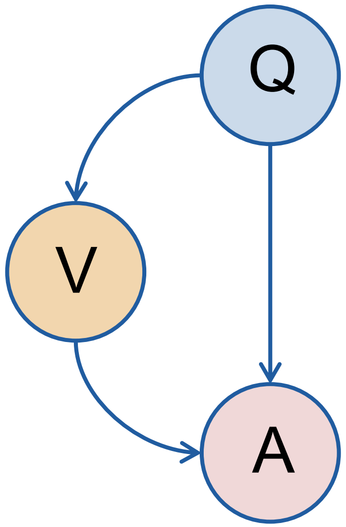
C : a man talking to a giraffe in an enclosure
Q_1 : how many people are there?
A_1 : 1
Q_2 : is it a male of female?
A_2 : Male
Q_3 : what is he doing?
A_3 : looking at the giraffe

Q_t : what color is the giraffe?

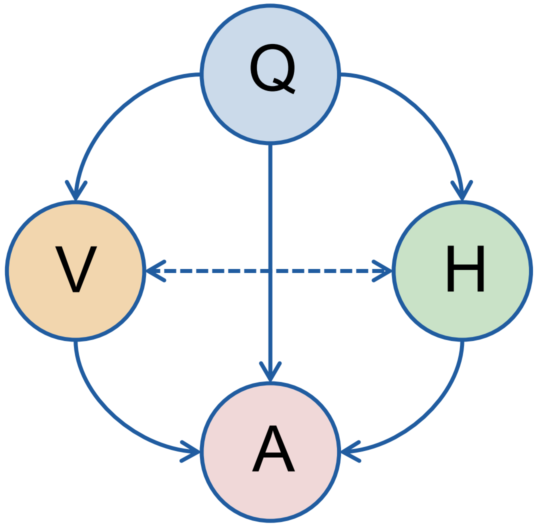
$\hat{A}_\mathrm{t}$: brown and tan

Visual Dialog is Challenging

- ❖ Reasoning not only on the image but also multi-rounds of dialog
- ❖ Primary method: attention mechanisms
  - V: vision, H: dialog history, Q: question, A: answer



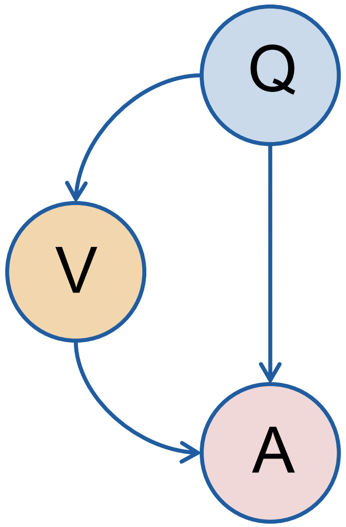
Visual Question Answering



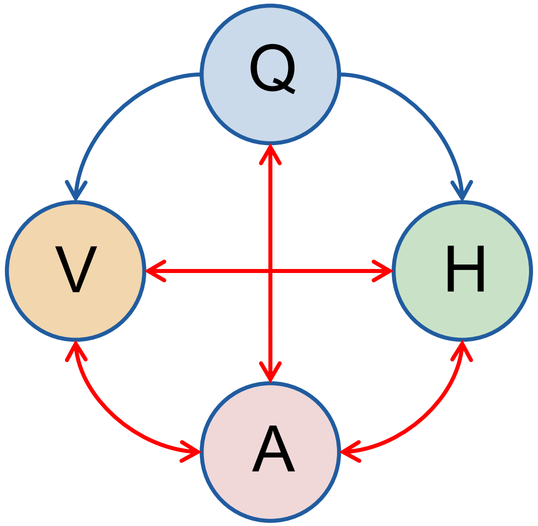
Prior Visual Dialog

Visual Dialog is Challenging

- ❖ Reasoning not only on the image but also multi-rounds of dialog
- ❖ Primary method: attention mechanisms
  - V: vision, H: dialog history, Q: question, A: answer

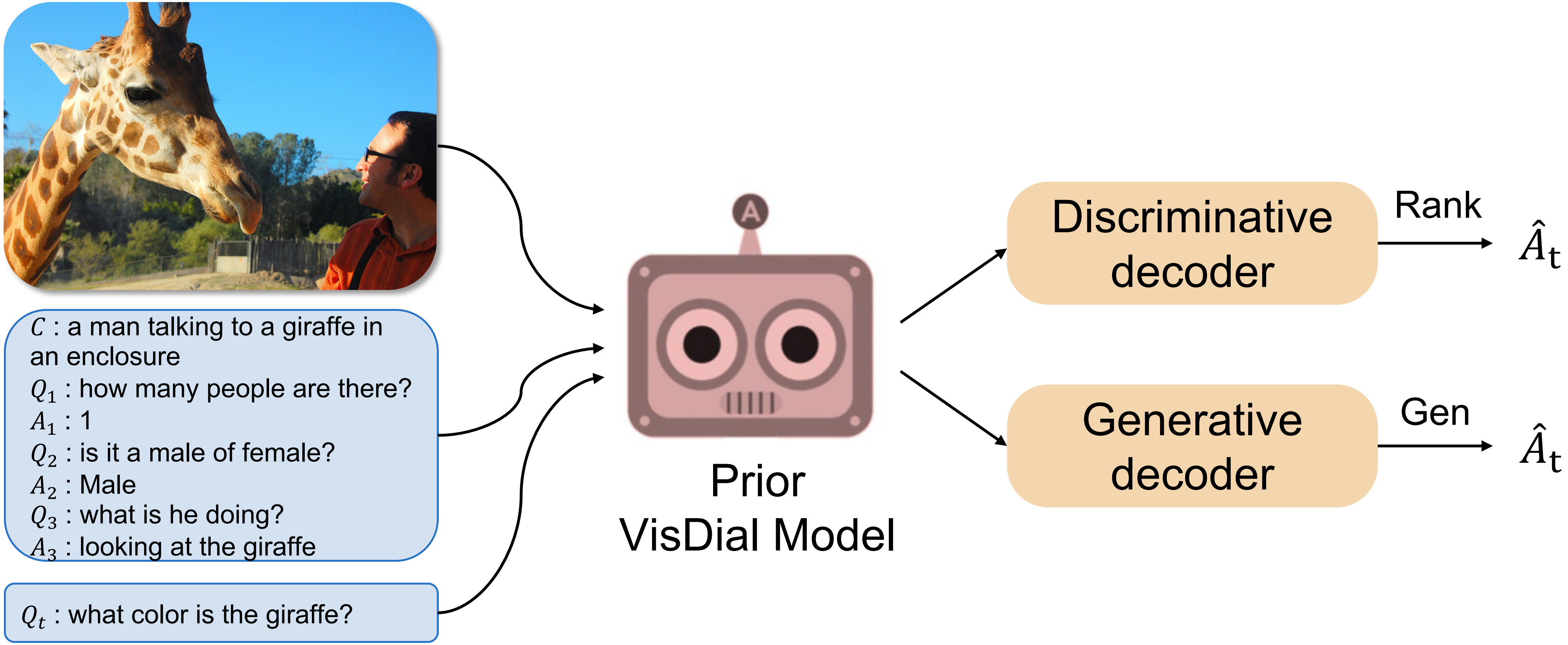


Visual Question Answering



Our Visual Dialog

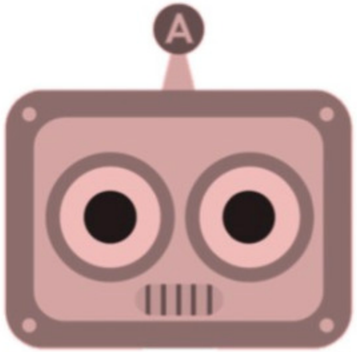
Decoding: Discriminative vs. Generative



Decoding: Discriminative vs. Generative

C : a man talking to a giraffe in an enclosure
Q_1 : how many people are there?
A_1 : 1
Q_2 : is it a male of female?
A_2 : Male
Q_3 : what is he doing?
A_3 : looking at the giraffe

Q_t : what color is the giraffe?



Our VisDial Model

Unified model

Rank $\hat{A}_t$

Gen $\hat{A}_t$

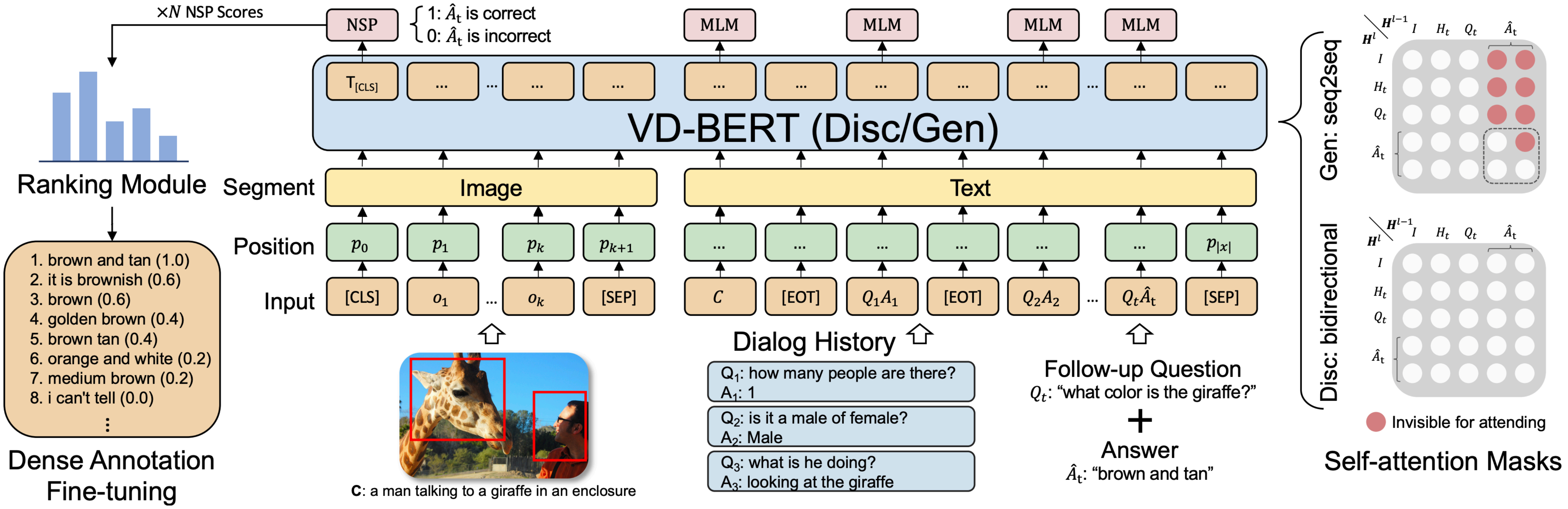
Proposed Solution

Contributions

- Unified Vision and Dialog Transformer with BERT (VD-BERT)
  - Employ self-attention to capture intricate vision-dialog interactions in a unified manner
  - Support both discriminative and generative settings seamlessly through a unified architecture
  - Extend BERT-like pretraining to achieve effective vision and dialog fusion

- Our proposed solution achieves new state-of-the-art results on the VisDial benchmark

Overview of VD-BERT



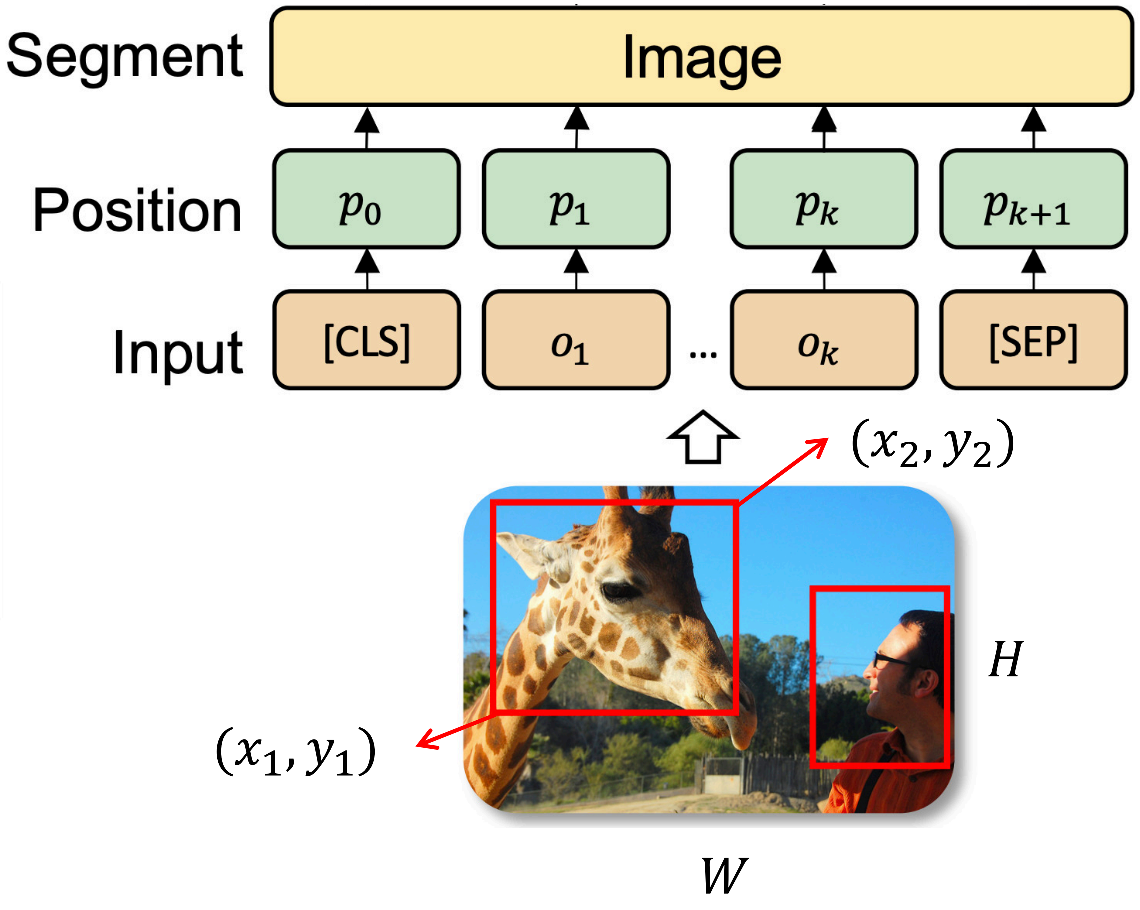
Proposed Solution

Encoding Image

- Visual feature
  - Use Faster R-CNN to detect k objects
    - $O_{\mathrm{I}} = \{o_1, \dots, o_k\}$
    - Each o_i is Region-of-Interest feature

- Position feature
  - Let (x_1, y_1) and (x_2, y_2) be the bottom-left and top-right corners of an object

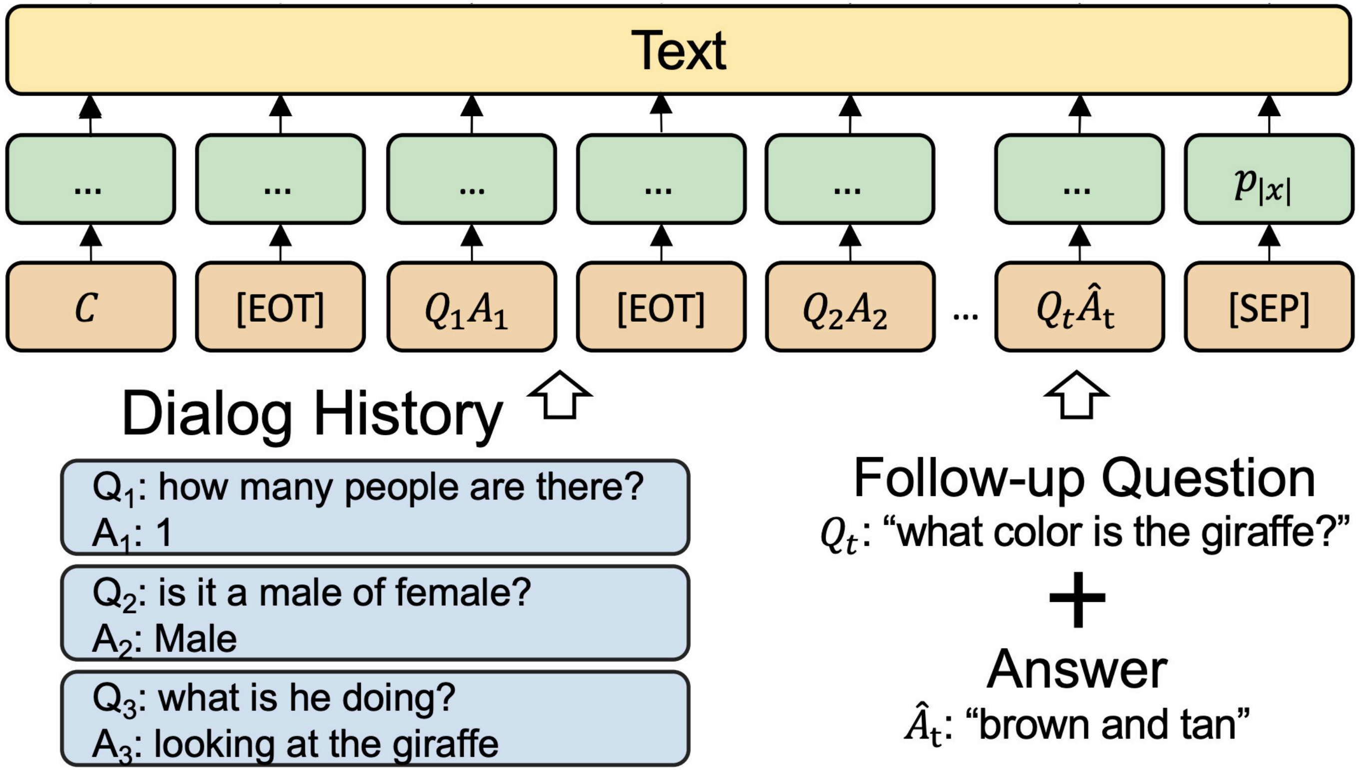
$$p_i = \left(\frac{x_1}{W}, \frac{y_1}{H}, \frac{x_2}{W}, \frac{y_2}{H}, \underbrace{\frac{(x_2-x_1)(y_2-y_1)}{WH}}_{\text{Relative area}}\right)$$



Proposed Solution

Encoding Language

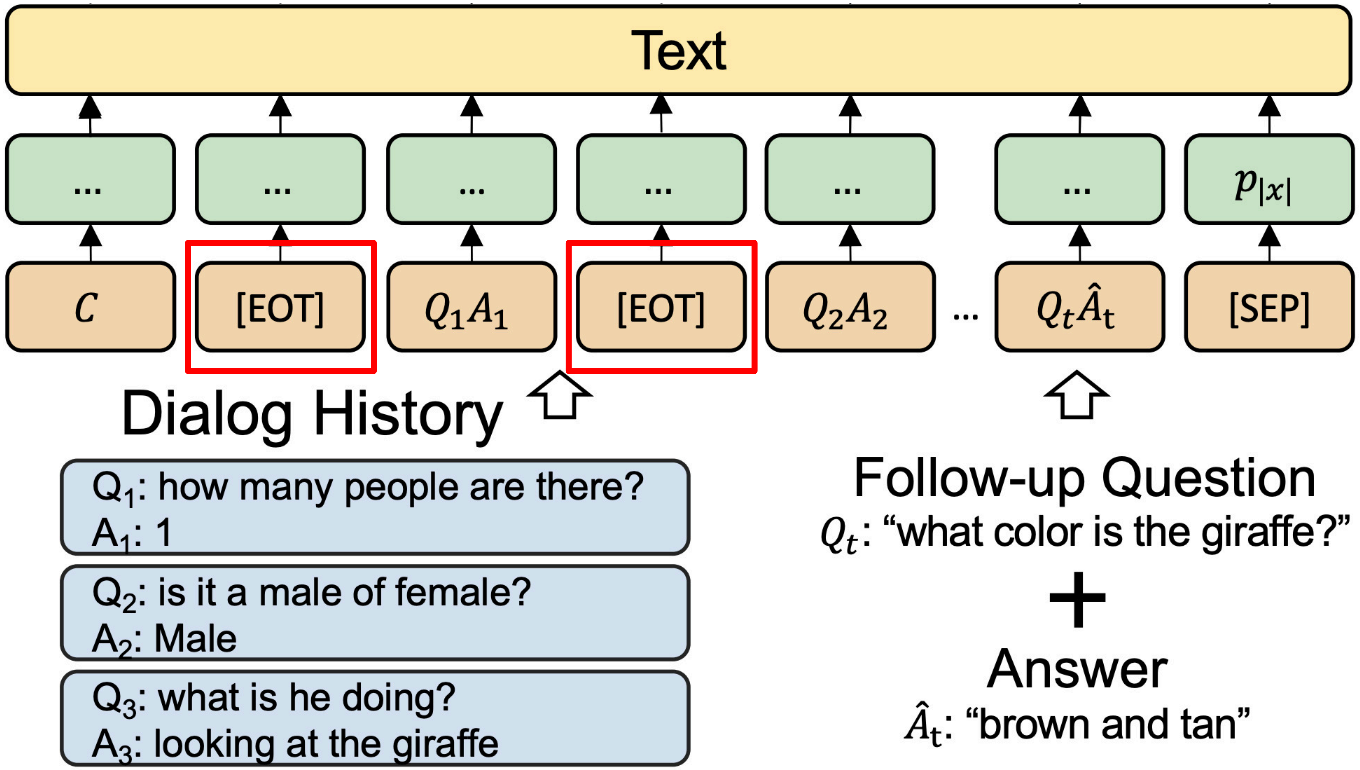
- Encode dialog structure
  - [EOT]: end of dialog turn
- Language feature (BERT)
  - WordPiece tokenization
  - Sinusoidal position embedding



Proposed Solution

Encoding Language

- Encode dialog structure
  - [EOT]: end of dialog turn
- Language feature (BERT)
  - WordPiece tokenization
  - Sinusoidal position embedding

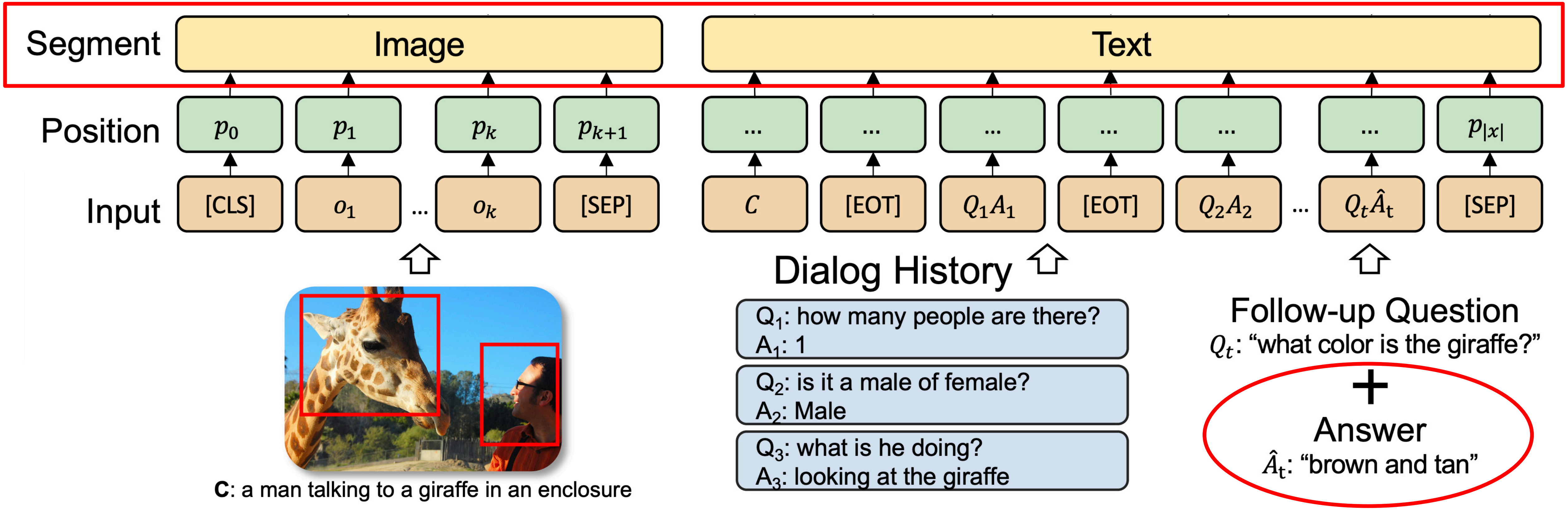


Proposed Solution

Combine Image and Text

Separate vision and language modalities

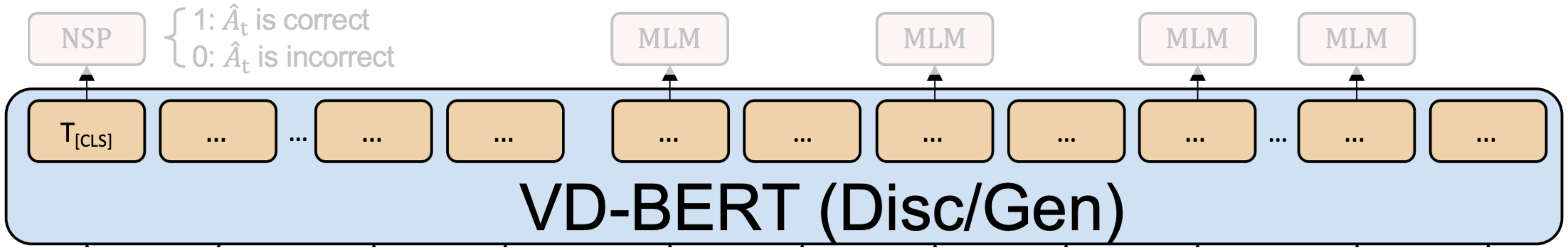
$$\mathbf{x} = (\texttt{[CLS]}, o_1, ..., o_k, \texttt{[SEP]}, C, \texttt{[EOT]}, Q_1A_1, \texttt{[EOT]}, ..., Q_t\hat{A}_t, \texttt{[SEP]})$$

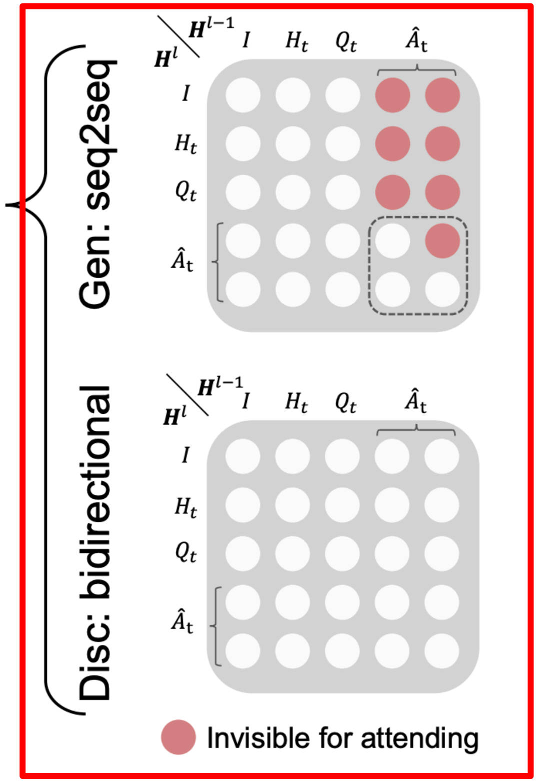


Early fusion of answer candidate

Proposed Solution

Single-stream Transformer Encoder





Self-attention Masks M

- Self-attention in Transformer

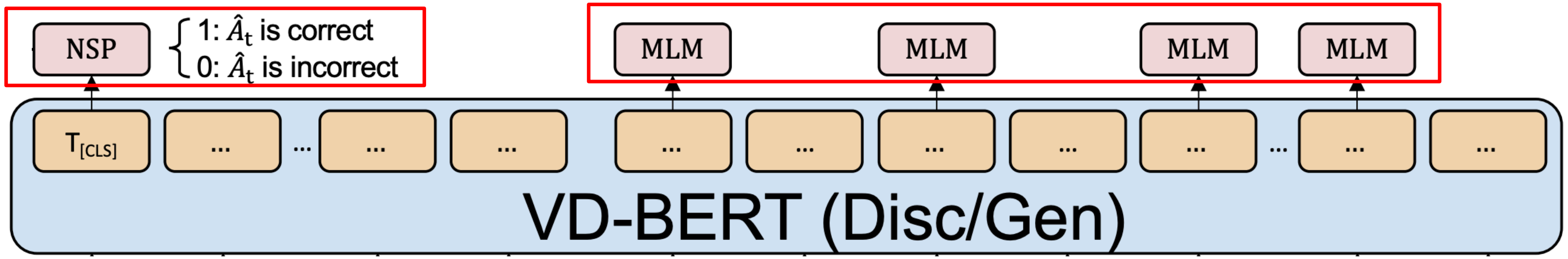
$$\mathbf{Q} = \mathbf{H}^{l-1}\mathbf{W}_l^Q, \mathbf{K} = \mathbf{H}^{l-1}\mathbf{W}_l^K, \mathbf{V} = \mathbf{H}^{l-1}\mathbf{W}_l^V, \quad (1)$$

$$\boxed{\mathbf{M}_{ij} = \begin{cases} 0, & \text{allow to attend,} \\ -\infty, & \text{prevent from attending,} \end{cases} \quad (2)}$$

$$\mathbf{A}_l = \text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}} + \mathbf{M})\mathbf{V}, \quad (3)$$

Proposed Solution

Visually Grounded Training Objectives



- Masked Language Modeling (MLM)
  - Predict masked tokens based on the image and other tokens

$$\mathcal{L}_{MLM} = -E_{(I,\mathbf{w})\sim D} \log P(w_m | \mathbf{w}_{\backslash m}, I)$$

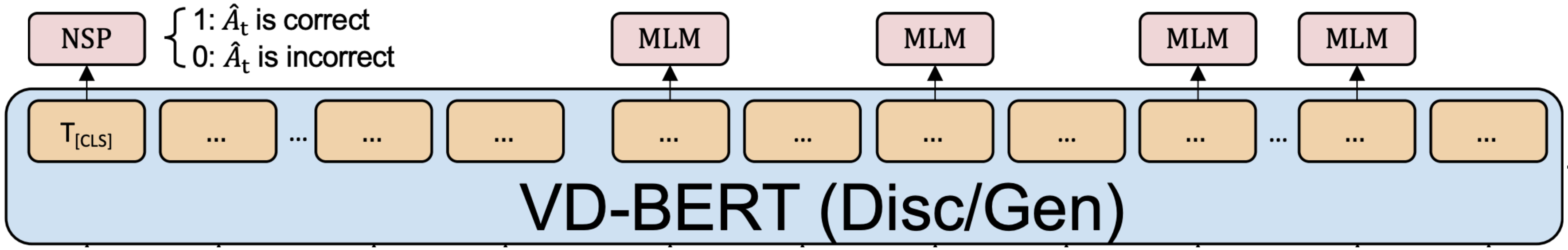
- Next Sentence Prediction (NSP)
  - Determine whether the appended $\hat{A}_t$ is correct or not

$$\mathcal{L}_{NSP} = -E_{(I,\mathbf{w})\sim D} \log P(y | S(I, \mathbf{w}))$$

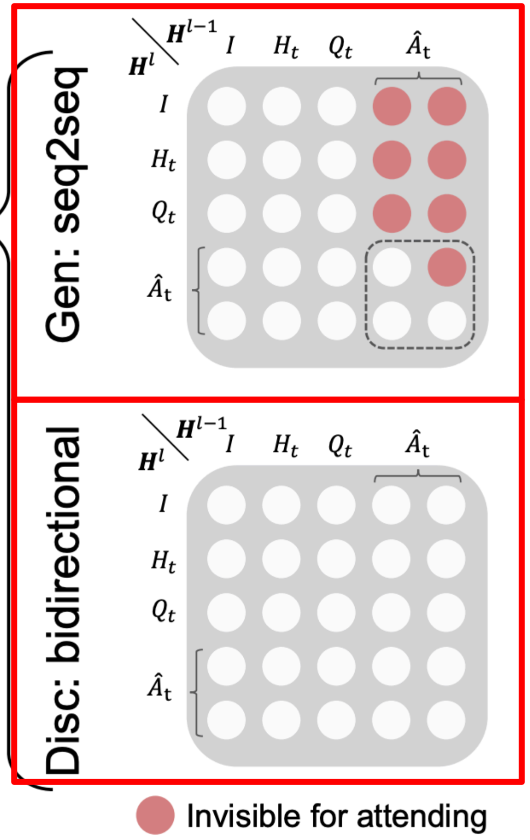
Vision and dialog fusion

Proposed Solution

Discriminative and Generative Settings



- ❖ Discriminative Setting
  - Bidirectional masks
  - Employ NSP head to predict scores for each $\hat{A}_t$
- ❖ Generative Setting
  - Seq2seq masks
  - Perform MLM recursively to generate $\hat{A}_t$



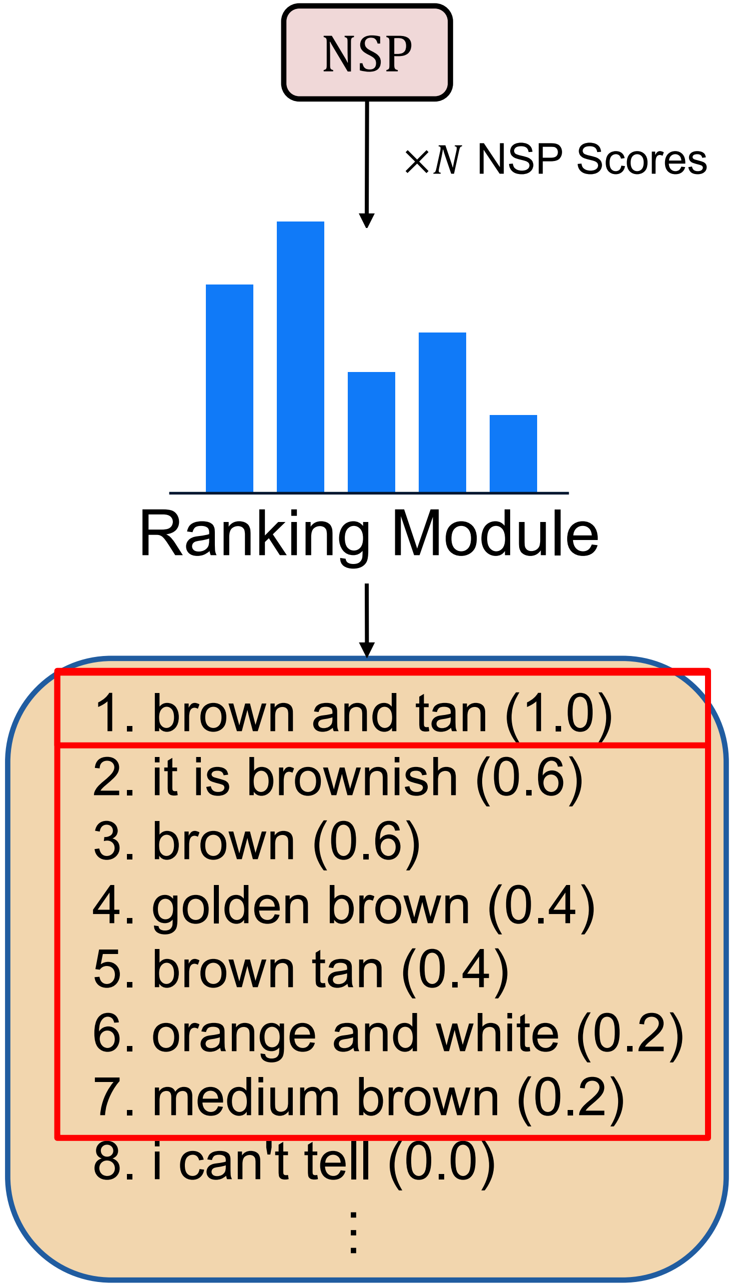
Self-attention Masks

Proposed Solution

Fine-tuning with Rank Optimization

- Dense annotations
  - Assign a continuous relevance score $s_i \in [0,1]$ to each $\hat{A}_t^i$

Q_t : what color is the giraffe?

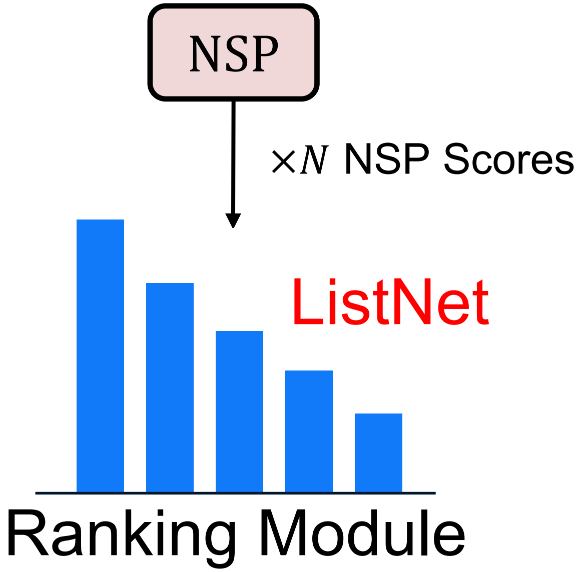


Proposed Solution

Fine-tuning with Rank Optimization

- Dense annotations
  - Assign a continuous relevance score $s_i \in [0,1]$ to each $\hat{A}_t^i$

Q_t : what color is the giraffe?



1. brown and tan (1.0)
2. it is brownish (0.6)
3. brown (0.6)
4. golden brown (0.4)
5. brown tan (0.4)
6. orange and white (0.2)
7. medium brown (0.2)
8. i can't tell (0.0)

⋮

Experiments

Experimental Setup

- VisDial Dataset
  - Image statistics of VisDial v0.9 and v1.0
  - Each image has 1 caption and 10 QA pairs

	Train	Val
v0.9	82,783	40,504

	Train	Val	Test
v1.0	123,287	2,064	8,000

The ground-truth answers are not public

- Metric
  - Sparse evaluation (only one correct)
    - Mean Reciprocal Rank (MRR)
    - Recall@K ($K \in \{1, 5, 10\}$)
    - Mean Rank
  - Dense evaluation (relevance score)
    - NDCG

Experiments

Experimental Setup

- VisDial Dataset
  - Image statistics of VisDial v0.9 and v1.0
  - Each image has 1 caption and 10 QA pairs

	Train	Val
v0.9	82,783	40,504

	Train	Val	Test
v1.0	123,287	2,064	8,000

- Metric
  - Sparse evaluation (only one correct)
    - Mean Reciprocal Rank (MRR)
    - Recall@K ($K \in \{1, 5, 10\}$)
    - Mean Rank
  - Dense evaluation (relevance score)
    - NDCG

Main focus!

Experiments

Full Comparison on VisDial v1.0

- Observations
  - New state of the art for both single-model and ensemble settings

Leaderboard:https://evalai.cloudcv.org/web/challenges/challenge-page/161/leaderboard/483

	Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
Published Results	NMN	58.10	58.80	44.15	76.88	86.88	4.81
	CorefNMN	54.70	61.50	47.55	78.10	88.80	4.40
	GNN	52.82	61.37	47.33	77.98	87.83	4.57
	FGA	52.10	63.70	49.58	80.97	88.55	4.51
	DVAN	54.70	62.58	48.90	79.35	89.03	4.36
	RvA	55.59	63.03	49.03	80.40	89.83	4.18
	DualVD	56.32	63.23	49.25	80.23	89.70	4.11
	HACAN	57.17	64.22	50.88	80.63	89.45	4.20
	Synergistic	57.32	62.20	47.90	80.43	89.95	4.17
	Synergistic†	57.88	63.42	49.30	80.77	90.68	3.97
	DAN	57.59	63.20	49.63	79.75	89.35	4.30
	DAN†	59.36	64.92	51.28	81.60	**90.88**	3.92
	ReDAN†	64.47	53.73	42.45	64.68	75.68	6.64
	CAG	56.64	63.49	49.85	80.63	90.15	4.11
	Square†	60.16	61.26	47.15	78.73	88.48	4.46
	MCA*	72.47	37.68	20.67	56.67	72.12	8.89
	MReal-BDAI†*	74.02	52.62	40.03	68.85	79.15	6.76
	P1_P2†*	74.91	49.13	36.68	62.98	78.55	7.03
Leaderboard Results	LF	45.31	55.42	40.95	72.45	82.83	5.95
	HRE	45.46	54.16	39.93	70.45	81.50	6.41
	MN	47.50	55.49	40.98	72.30	83.30	5.92
	MN-Att	49.58	56.90	42.42	74.00	84.35	5.59
	LF-Att	49.76	57.07	42.08	74.82	85.05	5.41
	MS ConvAI	55.35	63.27	49.53	80.40	89.60	4.15
	UET-VNU†	57.40	59.50	45.50	76.33	85.82	5.34
	MVAN	59.37	64.84	51.45	81.12	90.65	3.97
	SGLNs†	61.27	59.97	45.68	77.12	87.10	4.85
	VisDial-BERT*	74.47	50.74	37.95	64.13	80.00	6.28
	Tohoku-CV†*	74.88	52.14	38.93	66.60	80.65	6.53
Ours	VD-BERT	59.96	**65.44**	**51.63**	**82.23**	90.68	**3.90**
	VD-BERT*	74.54	46.72	33.15	61.58	77.15	7.18
	VD-BERT†*	**75.35**	51.17	38.90	62.82	77.98	6.69

"†" denotes ensemble model
"∗" denotes dense annotation fine-tuning

Experiments

Full Comparison on VisDial v1.0

- Observations
  - New state of the art for both single-model and ensemble settings
  - Inconsistency between NDCG and other metrics

Leaderboard:https://evalai.cloudcv.org/web/challenges/challenge-page/161/leaderboard/483

	Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
Published Results	NMN	58.10	58.80	44.15	76.88	86.88	4.81
	CorefNMN	54.70	61.50	47.55	78.10	88.80	4.40
	GNN	52.82	61.37	47.33	77.98	87.83	4.57
	FGA	52.10	63.70	49.58	80.97	88.55	4.51
	DVAN	54.70	62.58	48.90	79.35	89.03	4.36
	RvA	55.59	63.03	49.03	80.40	89.83	4.18
	DualVD	56.32	63.23	49.25	80.23	89.70	4.11
	HACAN	57.17	64.22	50.88	80.63	89.45	4.20
	Synergistic	57.32	62.20	47.90	80.43	89.95	4.17
	Synergistic†	57.88	63.42	49.30	80.77	90.68	3.97
	DAN	57.59	63.20	49.63	79.75	89.35	4.30
	DAN†	59.36	64.92	51.28	81.60	**90.88**	3.92
	ReDAN†	64.47	53.73	42.45	64.68	75.68	6.64
	CAG	56.64	63.49	49.85	80.63	90.15	4.11
	Square†	60.16	61.26	47.15	78.73	88.48	4.46
	MCA*	72.47	37.68	20.67	56.67	72.12	8.89
	MReal-BDAI†*	74.02	52.62	40.03	68.85	79.15	6.76
	P1_P2†*	74.91	49.13	36.68	62.98	78.55	7.03
Leaderboard Results	LF	45.31	55.42	40.95	72.45	82.83	5.95
	HRE	45.46	54.16	39.93	70.45	81.50	6.41
	MN	47.50	55.49	40.98	72.30	83.30	5.92
	MN-Att	49.58	56.90	42.42	74.00	84.35	5.59
	LF-Att	49.76	57.07	42.08	74.82	85.05	5.41
	MS ConvAI	55.35	63.27	49.53	80.40	89.60	4.15
	UET-VNU†	57.40	59.50	45.50	76.33	85.82	5.34
	MVAN	59.37	64.84	51.45	81.12	90.65	3.97
	SGLNs†	61.27	59.97	45.68	77.12	87.10	4.85
	VisDial-BERT*	74.47	50.74	37.95	64.13	80.00	6.28
	Tohoku-CV†*	74.88	52.14	38.93	66.60	80.65	6.53
Ours	VD-BERT	59.96	**65.44**	**51.63**	**82.23**	90.68	**3.90**
	VD-BERT*	74.54	46.72	33.15	61.58	77.15	7.18
	VD-BERT†*	**75.35**	51.17	38.90	62.82	77.98	6.69

"†" denotes ensemble model
"*" denotes dense annotation fine-tuning

Experiments

Discriminative and Generative Results on VisDial v0.9

Model	MRR↑	R@1↑	R@5↑	R@10↑	Mean ↓
	Discriminative/Generative				
LF	58.07/51.99	43.82/41.83	74.68/61.78	84.07/67.59	5.78/17.07
HRE	58.46/52.37	44.67/42.29	74.50/62.18	84.22/67.92	5.72/17.07
HREA	58.68/52.42	44.82/42.28	74.81/62.33	84.36/68.17	5.66/16.79
MN	59.65/52.59	45.55/42.29	76.22/62.85	85.37/68.88	5.46/17.06
HCIAE	62.22/54.67	48.48/44.35	78.75/65.28	87.59/71.55	4.81/14.23
CoAtt	63.98/55.78	50.29/46.10	80.71/**65.69**	88.81/71.74	4.47/14.43
RvA	66.34/55.43	52.71/45.37	82.97/65.27	90.73/**72.97**	**3.93/10.71**
DVAN	66.67/55.94	53.62/46.58	82.85/65.50	90.72/71.25	**3.93**/14.79
VD-BERT	**70.04/55.95**	**57.79/46.83**	**85.34**/65.43	**92.68**/72.05	4.04/13.18

Experiments

Ablation Study

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean ↓
No history	**64.70**	62.93	48.70	80.42	89.73	4.30
One previous turn	63.47	65.30	51.66	82.30	90.97	3.86
Full history	63.22	**67.44**	**54.02**	**83.96**	**92.33**	**3.53**
↪ only text	54.32	62.79	48.48	80.12	89.33	4.27

Training with various contexts

❖ Longer dialog history benefits most of metrics except NDCG

Experiments

Ablation Study

Model	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean ↓
No history	**64.70**	62.93	48.70	80.42	89.73	4.30
One previous turn	63.47	65.30	51.66	82.30	90.97	3.86
Full history	63.22	**67.44**	**54.02**	**83.96**	**92.33**	**3.53**
↪ only text	54.32	62.79	48.48	80.12	89.33	4.27

Training with various contexts

- Longer dialog history benefits most of metrics except NDCG
- Textual information dominates the VisDial task

Experiments

Case Study

A double decker bus sits empty at the station

Q1: are there any people?
A1: yes

Q2: are they on the bus?
A2: no, the bus is empty

Q3: are there any other buses?
A3: 1 other bus

Q4: are there people on bus?
A4: no it's empty (GT)

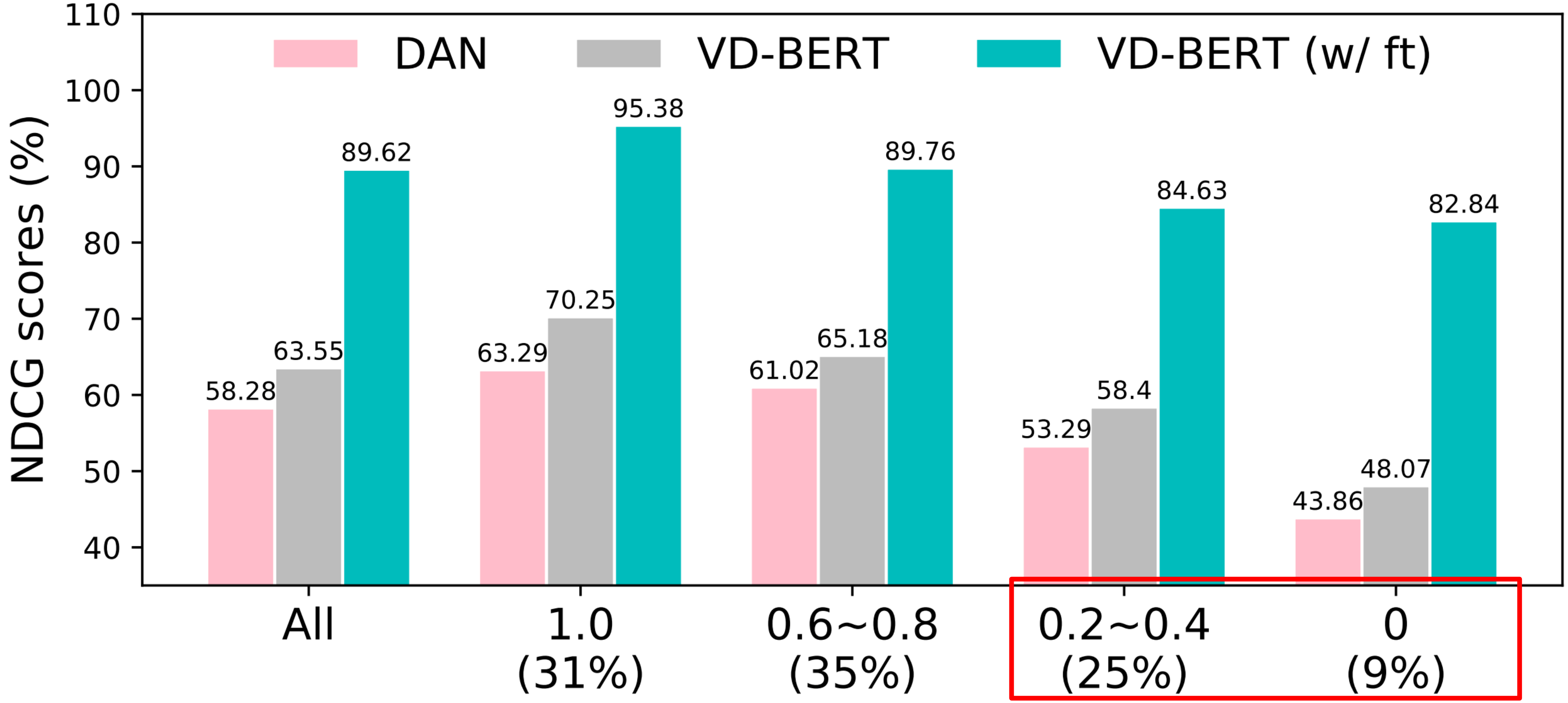
Base Model (NDCG=42.19)	W/ Fine-tuning (NDCG=91.80)
1. yes (0.0)	1. i cannot tell (0.8)
2. yes people (0.0)	2. i can't tell (0.8)
3. no it's empty (0.4)	3. can't tell (0.8)
4. i cannot tell (0.8)	4. not sure (0.8)
5. yes a few (0.0)	5. i don't know (0.8)
6. yes there are (0.0)	6. i cannot see any (0.8)
7. no (0.4)	7. not visible (0.6)
8. yes for sure (0.0)	8. not that i can see (0.6)

Base Model NDCG=42.19 → W/ Fine-tuning NDCG=91.80

Sparse and dense annotation mismatch!

Experiments

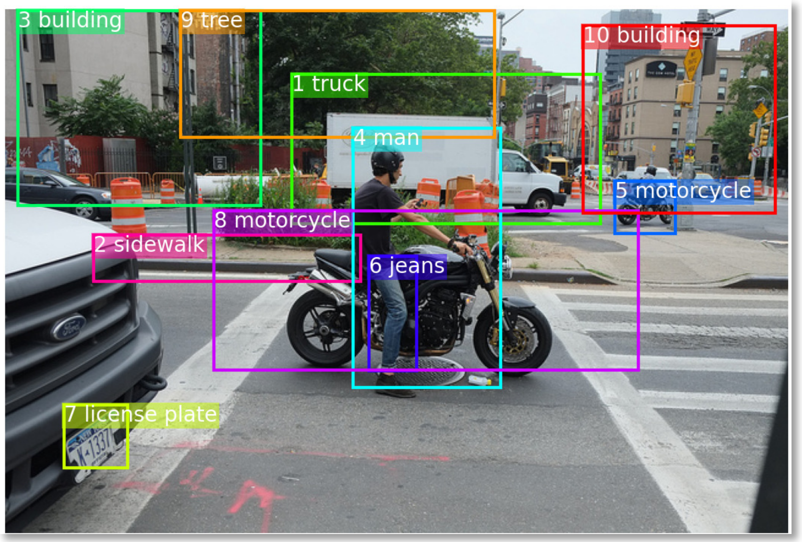
Relevance Score Analysis



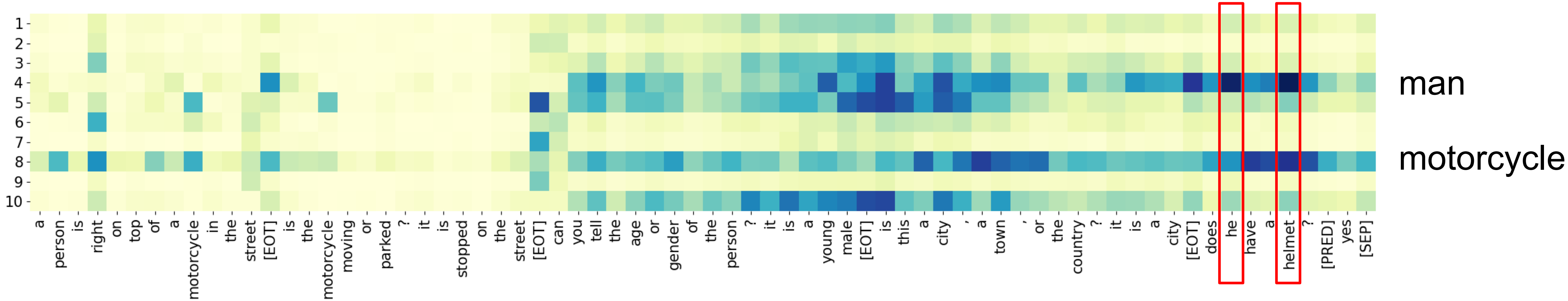
DAN is the model from (Kang et al., EMNLP 2019)

Experiments

Interpretability



- Entity grounding ("helmet")
- Visual pronoun coreference ("he")



Conclusion

- We propose a unified VD-BERT that extends BERT for effective vision and dialog fusion
- VD-BERT achieves a new state-of-the-art result on the VisDial challenge
- Extensive experiments provide insights for future transfer learning research in visual dialog tasks

Thanks!

香港中文大學
The Chinese University of Hong Kong

salesforce

Yue Wang

Shafiq Joty

Michael R. Lyu

Irwin King

Caiming Xiong

Steven C.H. Hoi

Code & Models: https://github.com/salesforce/VD-BERT