The Dialog Must Go On:
Improving Visual Dialog via Generative Self-Training

Gi-Cheon Kang | Sungdong Kim† | Jin-Hwa Kim† | Donghyun Kwak† | Byoung-Tak Zhang

(† equal contribution)

SEOUL NATIONAL UNIVERSITY

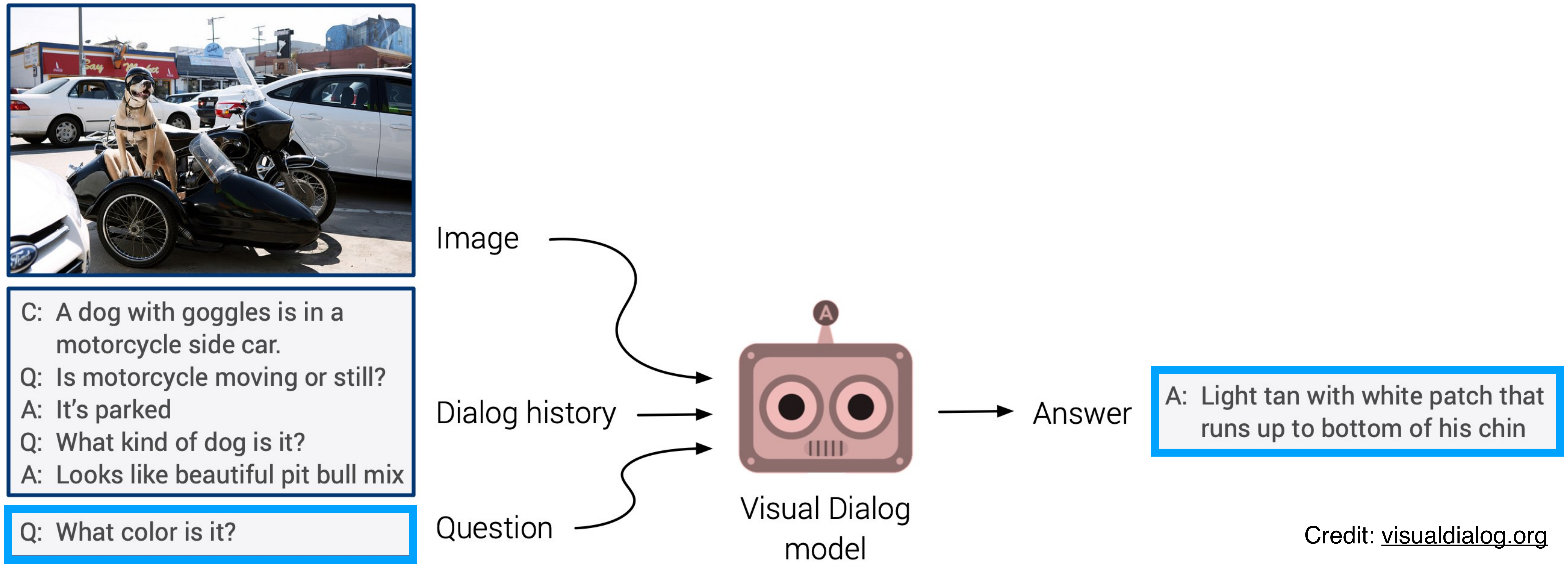
NAVER Cloud

CVPR 2023

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

What is Visual Dialog?

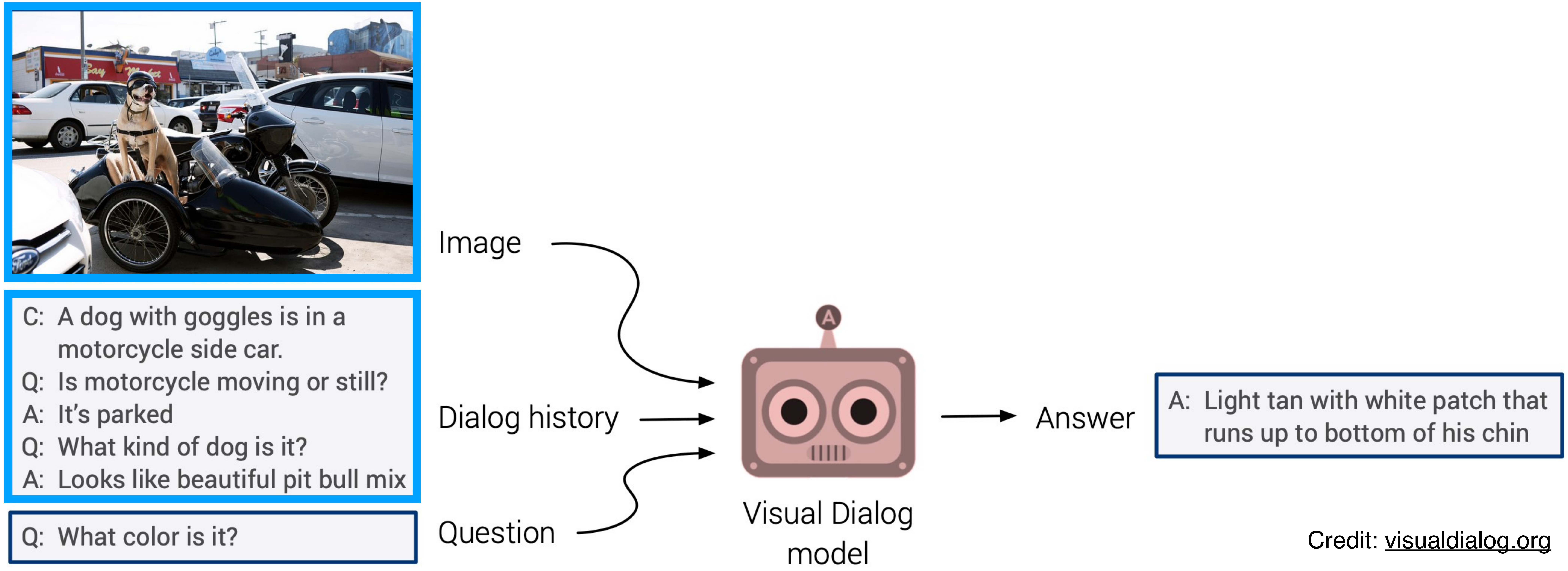
- **Answer a sequence of questions grounded in an image**
- Image and dialog history as a context



Credit: visualdialog.org

What is Visual Dialog?

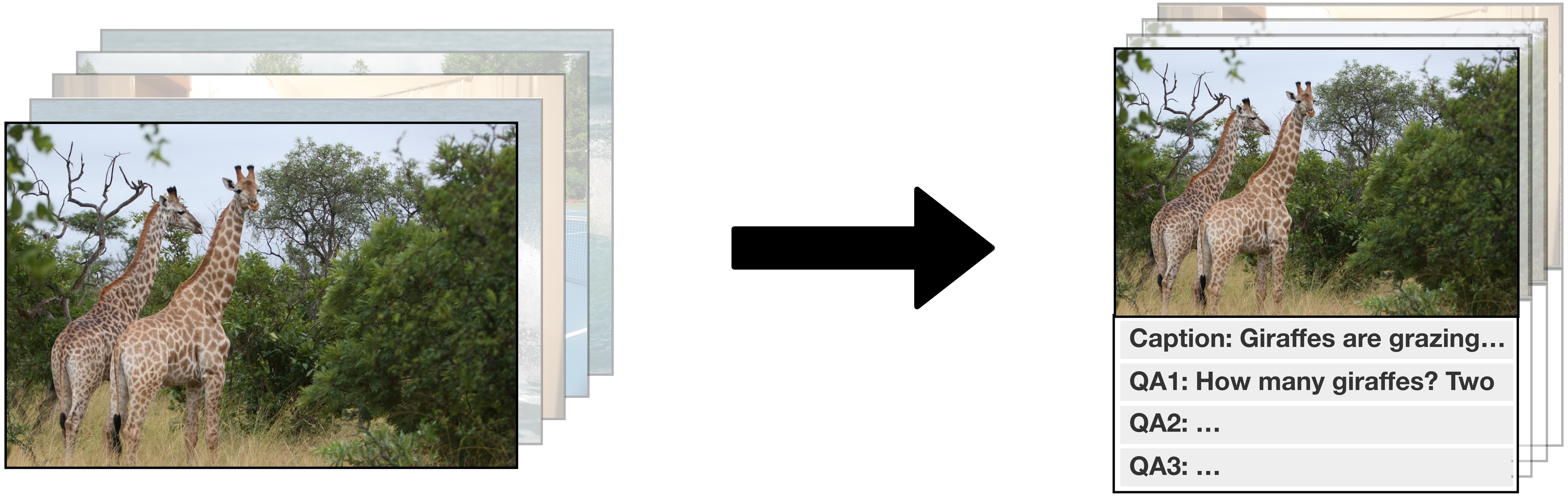
- Answer a sequence of questions grounded in an image
- **Image and dialog history as context**



Credit: visualdialog.org

Quick Preview

- Semi-supervised learning approach for Visual Dialog
- Generate visually-grounded dialog data for unlabeled Web images
- Leveraging the dialog data improves overall performance, adversarial robustness …



Unlabeled Images

Artificial Visual Dialog Dataset

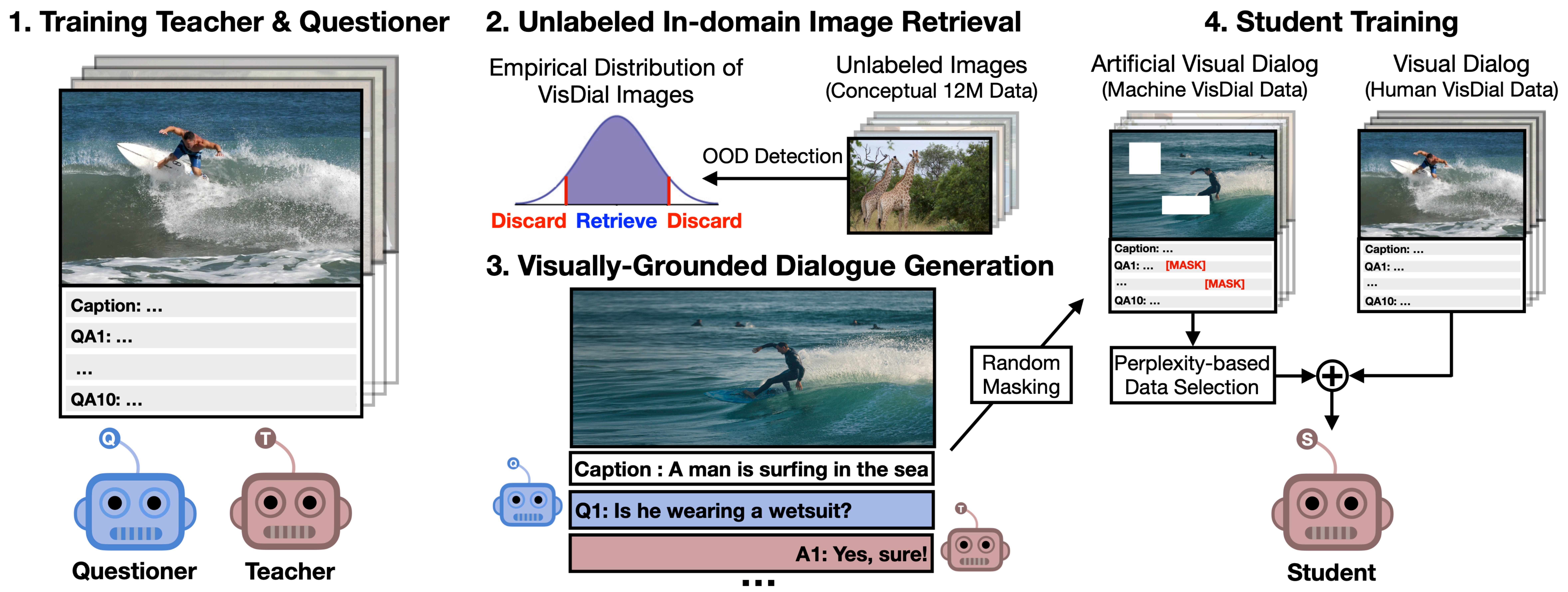
Motivation

- Prior work has trained the dialog agents solely on VisDial data via supervised learning or leveraged pre-training on related vision-and-language datasets.

- How can the dialog agent expand its knowledge beyond what it can acquire via supervised learning or self-supervised pre-training on the provided datasets?

- We propose a semi-supervised learning approach, called Generative Self-Training (GST), that artificially generates multi-turn visual QA data and utilizes the synthetic data for training.

Generative Self-Training (GST)



Teacher & Questioner Training

Given VisDial data $L = \{(v_n, d_n)\}_{n=1}^{N}$ $d_n = \{\underbrace{c_n}_{d_{n,0}}, \underbrace{(q_{n,1}, a_{n,1})}_{d_{n,1}}, \cdots, \underbrace{(q_{n,T}, a_{n,T})}_{d_{n,T}}\}$

① We first train teacher model $P_{\mathcal{T}}$ by minimizing the negative log likelihood of the ground-truth answers $a_{n,t} = (w_1, \cdots, w_S)$

$$\begin{aligned}\mathcal{L}_{teacher} &= -\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\log P_{\mathcal{T}}(a_{n,t}|v_n, d_{n,<t}, q_{n,t}) \\ &= -\frac{1}{NTS}\sum_{n=1}^{N}\sum_{t=1}^{T}\sum_{s=1}^{S}\log P_{\mathcal{T}}(w_s|v_n, d_{n,<t}, q_{n,t}, w_{<s})\end{aligned}$$

② Similarly, we train the question generation model $P_{\mathcal{Q}}$

Model Architecture of Teacher & Questioner

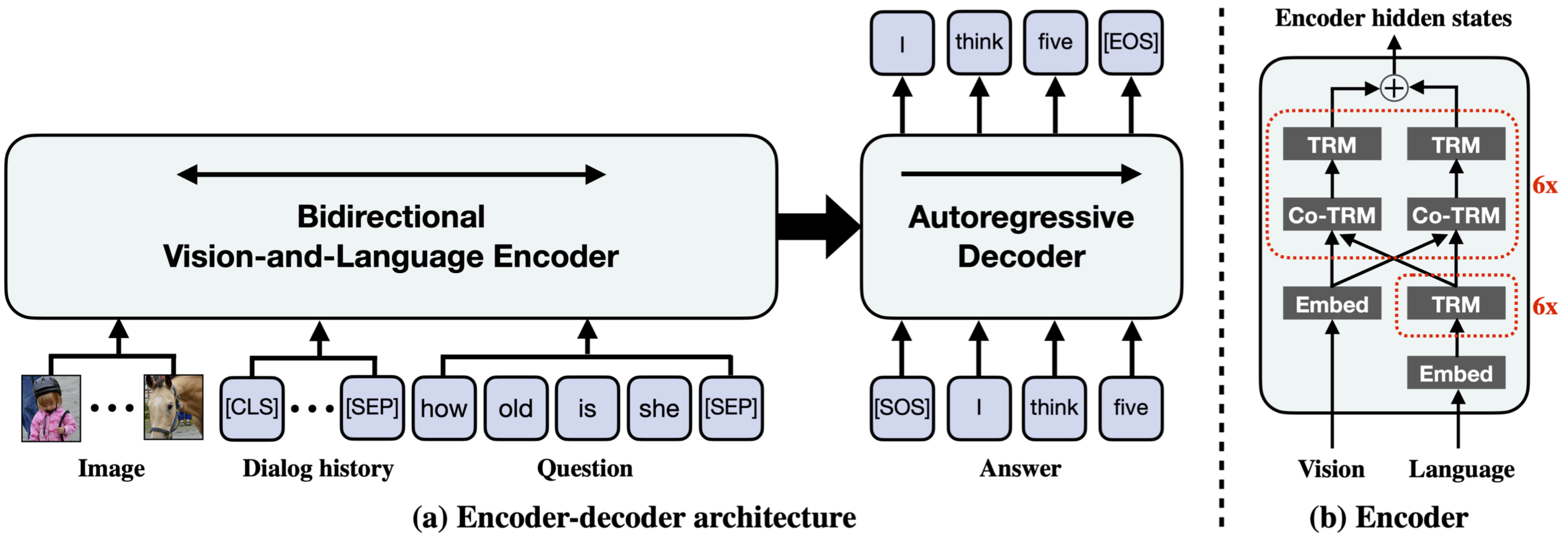


Figure 3: A detailed architecture of our proposed model. We propose the encoder-decoder model where the encoder aggregates the given multimodal context, and the decoder generates the target sentence. (b): a more detailed view of the encoder. TRM and Co-TRM denote the transformer module and the co-attentional transformer module, respectively. $\oplus$ denotes the concatenation operation.

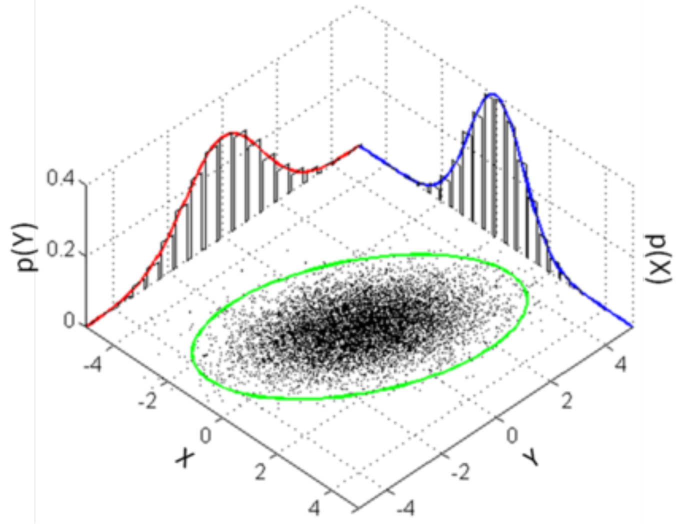
Unlabeled In-Domain Image Retrieval

Visual Dialog

Image Encoder

Feature vectors for 120k images

Multivariate Normal Distribution



Sorting by Probability

CC12M

Image Encoder

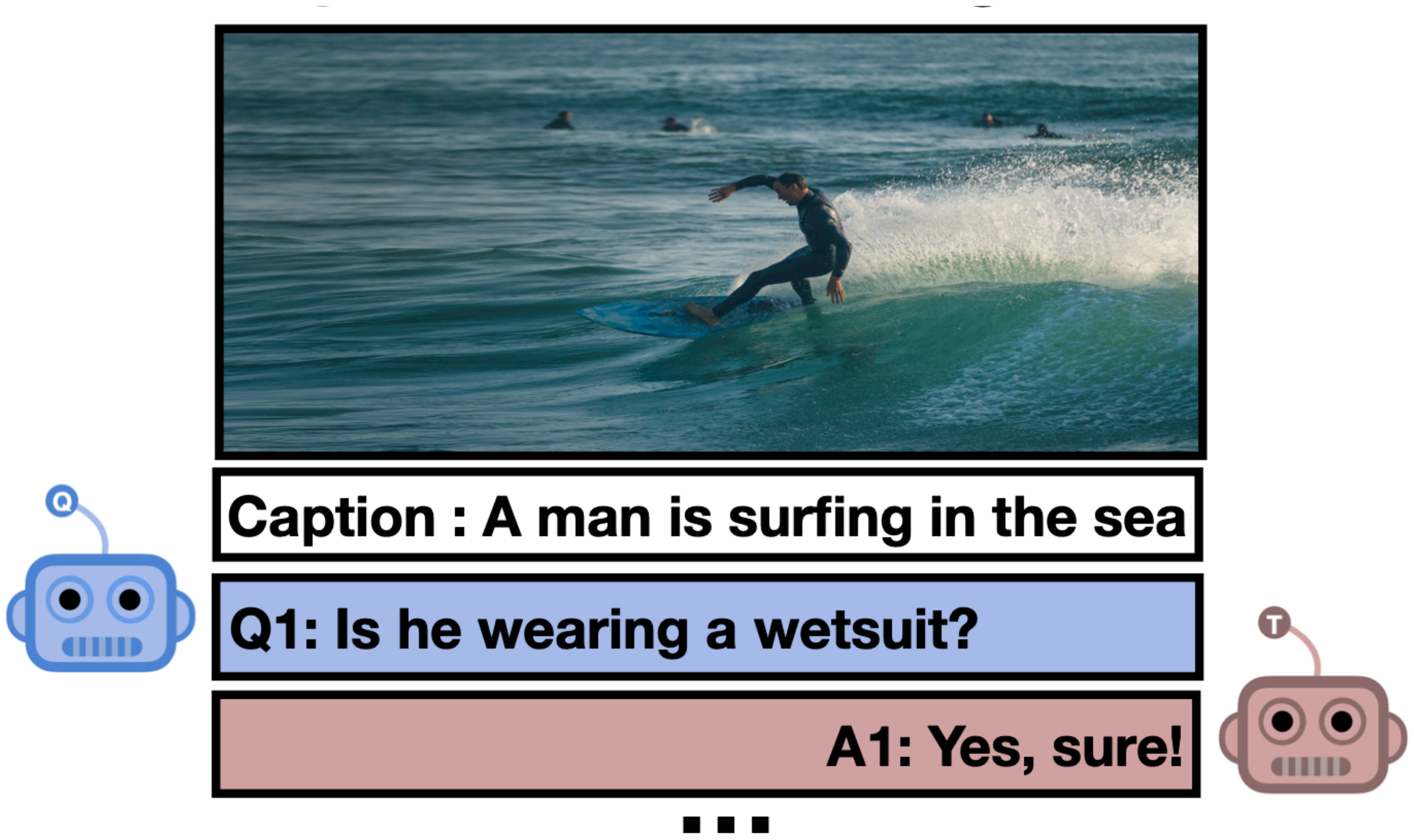
Feature vectors for 12M images

Visually-Grounded Dialogue Generation

Given unlabeled images and the captions, the questioner and the teacher generate the dialogs

For 3.6M images, 36M QA pairs are generated (1 image + 10 QA pairs)

Decoding strategy: Top-k sampling(k=7) with temperature 0.7



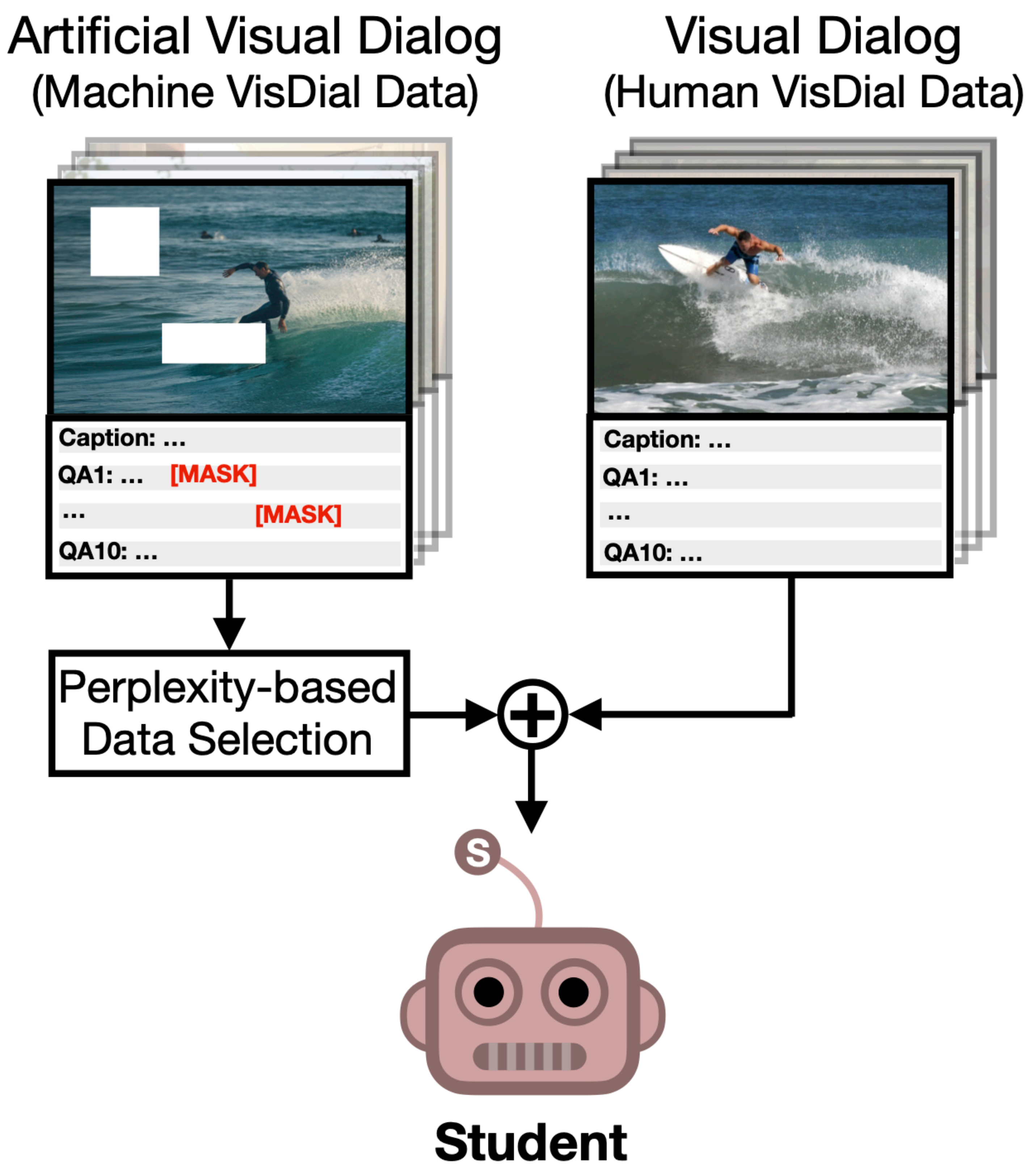
Student Training

We propose perplexity-based data selection (PPL) and multimodal consistency regularization (MCR) to effectively train the artificially generated dialog dataset

$$\mathcal{L}_{Student} = -\frac{1}{MT}\sum_{m=1}^{M}\sum_{t=1}^{T}\mathbb{1}(\mathrm{PPL}(\tilde{a}_{m,t}) < \tau)\log \underbrace{P_{\mathcal{S}}(\tilde{a}_{m,t}|\,\mathcal{M}(\tilde{v}_m, \tilde{d}_{m,<t}, \tilde{q}_{m,t}))}_{\mathrm{MCR}}$$

$$-\frac{1}{NT}\sum_{n=1}^{N}\sum_{t=1}^{T}\log P_{\mathcal{S}}(a_{n,t}|v_n, d_{n,<t}, q_{n,t})$$

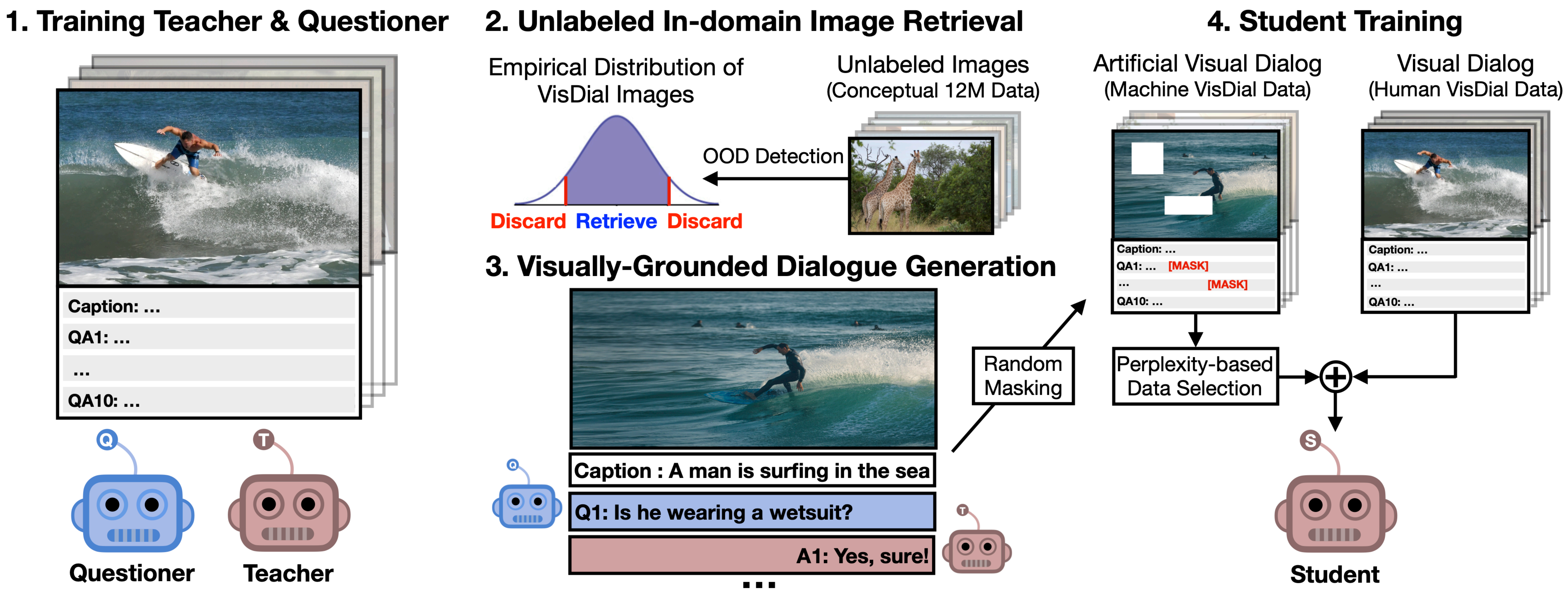
$$\text{where } \mathrm{PPL}(\tilde{a}_t) = \exp\left\{-\frac{1}{S}\sum_{s=1}^{S}\log P_{\mathcal{T}}(\tilde{w}_s|\tilde{v}, \tilde{d}_{<t}, \tilde{q}_t, \tilde{w}_{<s})\right\}$$



Iterative Training

The student model at i-th iteration as a teacher model at $(i + 1)$-th iteration

Repeats the third and fourth steps up to 3 times



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 !

Experimental Results

SOTA Comparison

	VisDial v0.9 (val)					VisDial v1.0 (val)					
Model	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓	NDCG↑	MRR↑	R@1↑	R@5↑	R@10↑	Mean↓
MN† [12]	52.59	42.29	62.85	68.88	17.06	51.86	47.99	38.18	57.54	64.32	18.60
HCIAE† [55]	53.86	44.06	63.55	69.24	16.01	59.70	49.07	39.72	58.23	64.73	18.43
CoAtt† [90]	55.78	46.10	65.69	71.74	14.43	59.24	49.64	40.09	59.37	65.92	17.86
CorefNMN [40]	53.50	43.66	63.54	69.93	15.69	-	-	-	-	-	-
RvA [61]	55.43	45.37	65.27	72.97	**10.71**	-	-	-	-	-	-
Primary [22]	-	-	-	-	-	-	49.01	38.54	59.82	66.94	16.60
DMRM [10]	55.96	46.20	66.02	72.43	13.15	-	50.16	40.15	60.02	67.21	15.19
ReDAN [19]	-	-	-	-	-	60.47	50.02	40.27	59.93	66.78	17.40
DAM [29]	-	-	-	-	-	60.93	50.51	40.53	60.84	67.94	16.65
KBGN [28]	-	-	-	-	-	60.42	50.05	40.40	60.11	66.82	17.54
LTMI [60]	-	-	-	-	-	63.58	50.74	40.44	61.61	69.71	14.93
VD-BERT [89]	55.95	46.83	65.43	72.05	13.18	-	-	-	-	-	-
MITVG [9]	56.83	47.14	67.19	73.72	11.95	61.47	51.14	41.03	61.25	68.49	14.37
UTC [8]	-	-	-	-	-	63.86	52.22	42.56	62.40	69.51	15.67
Student (ours)	**60.03**±.18	**50.40**±.15	**70.74**±.09	**77.15**±.13	12.13±.18	**65.47**±.14	**53.19**±.11	**43.08**±.10	**64.09**±.05	**71.51**±.13	**14.34**±.15

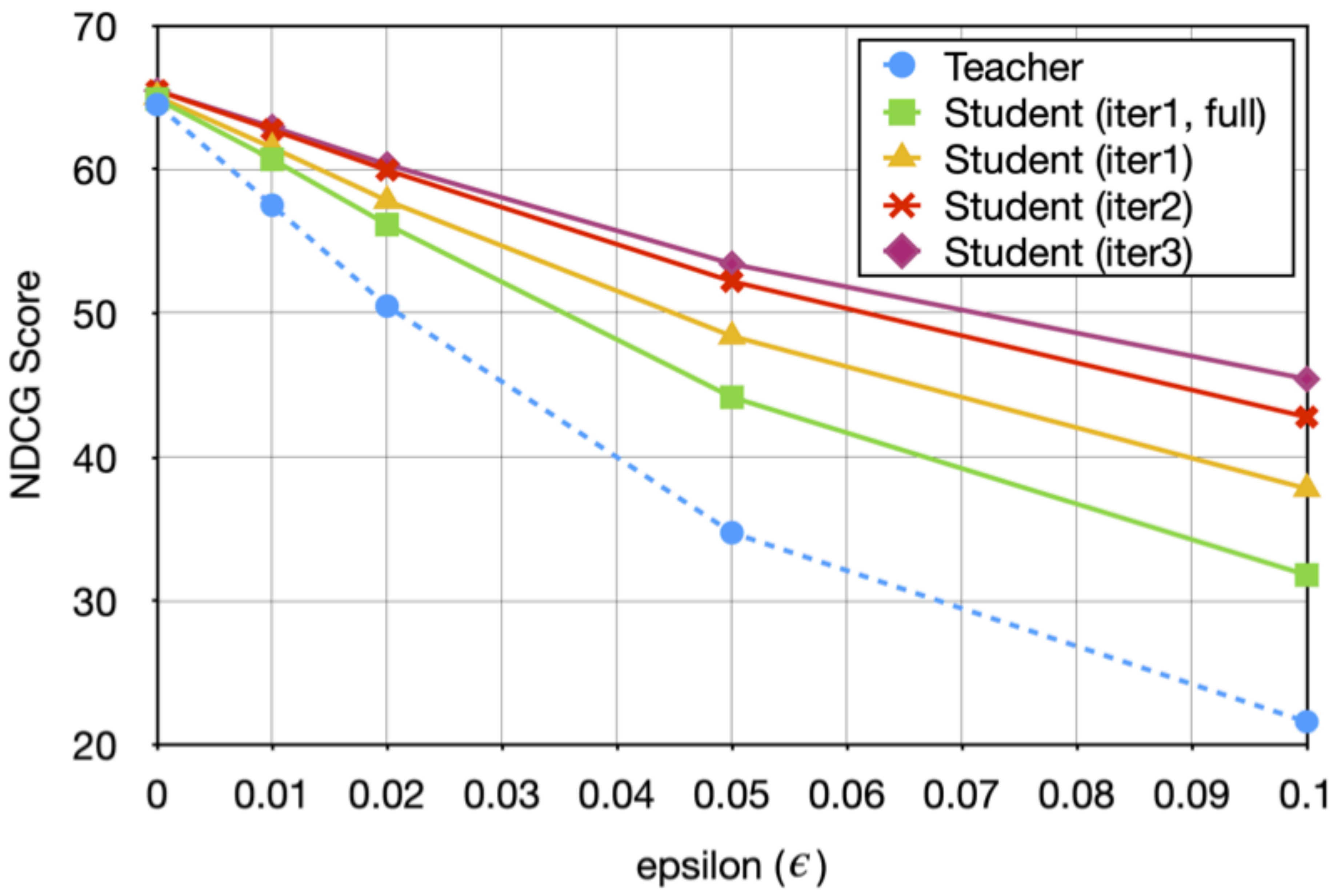
GST in the Low-data Regime

	NDCG				
Model	1%	5%	10%	20%	30%
Teacher	27.64	50.04	54.46	57.14	60.67
Student	**38.73** (**+11.09**)	**56.60** (**+6.56**)	**58.62** (**+4.16**)	**60.92** (**+3.78**)	**63.09** (**+2.42**)

N-gram Diversity of Generated Questions

Model	N-gram Diversity				No Match
	N=1	N=2	N=3	N=4	
Questioner	**28.06** ±0.14	**56.46** ±0.09	**76.98** ±0.08	**92.80** ±0.08	**95.38** ±0.15

Experimental Results



Adversarial Robustness
(Visual FGSM attack)

Model	No Attack	Coreference Attack	Random Token Attack			
			10%	20%	30%	40%
Teacher	56.55	52.60	54.69±1.12	52.86±0.79	49.41±2.09	45.04±2.28
Student (iter1, full)	58.53	54.26	56.59±1.37	54.55±1.15	50.98±2.06	46.56±1.96
Student (iter1)	58.63	54.34	55.59±0.88	54.26±1.54	51.04±2.39	47.04±2.03
Student (iter2)	56.92	52.69	55.59±0.88	53.57±1.40	49.95±1.91	46.82±2.02
Student (iter3)	**59.30**	**55.44**	**57.25**±0.91	**55.10**±1.50	**52.11**±2.75	**48.00**±2.90

Adversarial Robustness
(Textual attack)

Visualization of Human and Machine Conversation

Image & Caption	Human Conversation	Machine Conversation
 A man is talking on a cell phone in a parking lot	D1 : Is it daytime? Yes D2 : Is this is a busy full parking lot? No D3 : Is the man dressed for work? No D4 : What is he wearing? Polo shirt and shorts D5 : Are there other people around? No D6 : Do you see any cars? Yes D7 : Are there a lot of cars? No D8 : How many cars do you think there are? 3 D9 : What colors are the cars? White and black D10 : Is the man wearing sunglasses? Yes	D1 : Is the man wearing a hat? No D2 : Is the man dressed for cold or warm weather? Warm weather D3 : Is the man talking on a cell phone? Yes he is D4 : Is it a flip phone? Yes it is D5 : What color is the phone? Black and silver D6 : Is the man holding it in his hand? Yes it is D7 : Are there any cars in the parking lot? Yes D8 : Is he waiting for someone to get on the car? No just standing D9 : Is the man looking at the camera? No he is looking off to the side D10 : Are there any buildings in the background? \ There is a brick building behind him
 A person stand in front of a brown horse	D1 : Is the horse wearing a saddle? No D2 : Is it a large horse? Yes D3 : Is the person male or female? Female D4 : About what age is she? She is about 25 D5 : Is she wearing riding gear? Yes D6 : Is this a country setting? Yes it is D7 : Do you see a barn? No D8 : Does the horse look well cared for? Yes it do D9 : Is there a blanket on the horse? No D10 : Is the weather nice? Yes	D1 : Is the horse large? Yes D2 : Is it facing the camera? No D3 : Is the person a man or a woman? Woman D4 : Does the horse have a saddle? No she does not D5 : Is it sunny? It is hard to tell but it is daylight D6 : Is the woman wearing a hat? Yes she is D7 : What color is the woman's hair? It appears to be brown D8 : Does the woman have long hair? I cannot tell D9 : What color is her shirt? She is wearing a black jacket D10 : Are there any trees? There are no trees
 A woman rides her bike on a sidewalk next to a fence	D1 : Is it daytime? Yes D2 : Is the sun out? Yes D3 : How old is the woman? 24 D4 : Is she alone? Yes D5 : Is she wearing sunglasses? No D6 : Is she carrying anything else with her? No D7 : Is she wearing sneakers? Yes D8 : What color is her shirt? Wearing a black jacket D9 : Is she smiling? No D10 : Where is she going? West on the road	D1 : Is this a color photo? Yes, it is a color photo D2 : Is it daytime? Yes D3 : What is the weather like? Looks like sunny D4 : What color is the woman's hair? Hair is brown D5 : How old does she seem? Around 35 D6 : Is she wearing a helmet? No helmet seen D7 : Is she wearing jeans? No D8 : Is she wearing glasses? No D9 : What color is her bike? **Bike is blue in color** D10 : Are there any cars? Yes, i see 1 car

Thank You !

Code: https://github.com/gicheonkang/gst-visdial
Paper: https://arxiv.org/abs/2205.12502