FLIPDIAL: A Generative Model for Two-Way Visual Dialogue Supplementary Material

Daniela Massiceti
University of Oxford, UK
daniela@robots.ox.ac.uk

N. Siddharth
University of Oxford, UK
nsid@robots.ox.ac.uk

Puneet K. Dokania
University of Oxford, UK
puneet@robots.ox.ac.uk

Philip H.S. Torr
University of Oxford, UK
phst@robots.ox.ac.uk

A. Glossary

block dialogue/architecture Models **B**/**B**$_{\mathbf{AR}}$ are built and trained for the task of two-way visual dialogue (2VD) with data $\boldsymbol{x} = \boldsymbol{d}$ and condition variable $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}\}$. Since $\boldsymbol{d}$ refers to the whole dialogue sequence/block $\langle(\boldsymbol{q}_t, \boldsymbol{a}_t)\rangle_{t=1}^{T}$ we refer to **B**/**B**$_{\mathbf{AR}}$ as *block* architectures.

generation This represents the scenario when only the condition variable $\boldsymbol{y}$ is available at test time. In this case, the decoder network receives a sample $\boldsymbol{z} \sim p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$, a multivariate Gaussian parametrised by $\boldsymbol{\mu}_p$ and exponentiated $\log \boldsymbol{\sigma}_p^2$ learned using the *prior* network. We call the decoded output $\hat{\boldsymbol{d}}$ a *generation*.

reconstruction Differing from a generation, both $\boldsymbol{y}$ *and* $\boldsymbol{d}$ are available. The decoder network receives a sample $\boldsymbol{z} \sim q_\phi(\boldsymbol{z} \mid \boldsymbol{d}, \boldsymbol{y})$, a multivariate Gaussian parametrised by $\boldsymbol{\mu}_q$ and exponentiated $\log \boldsymbol{\sigma}_q^2$ learned using the *encoder* network. We call the decoded output $\hat{\boldsymbol{d}}$ a *reconstruction*. The reconstruction pipeline is used during training when the input $\boldsymbol{d}$ *and* the condition variable $\boldsymbol{y}$ are available. Note, this pipeline is also used when **B**/**B**$_{\mathbf{AR}}$ are evaluated iteratively (see §4.2).

B. Extended Quantitative Results on 1VD task

Tab. 3 in the main paper evaluates **A** and **B**/**B**$_{\mathbf{AR}}$ in the task of one-way visual dialogue (1VD). Here we shed light on these numbers and the metrics used to obtain them. We also present a more extensive quantitative analysis of **B**/**B**$_{\mathbf{AR}}$ in the 1VD task (see Tab. 6).

Evaluating B/B$_{\mathbf{AR}}$ on 1VD We extend Tab. 3 with-Tab. 6, which further compares **B**/**B**$_{\mathbf{AR}}$ under the iterative evaluation settings of $\boldsymbol{d}$–$\boldsymbol{qa}$ and $\boldsymbol{d}$–$\boldsymbol{q}\hat{\boldsymbol{a}}$, using the cross entropy (CE) and Kullback-Leibler (KL) terms of the evidence lower bound (ELBO) and our two new metrics, sim$_{\boldsymbol{c},\boldsymbol{q}}$ and sim$_{\circlearrowleft}$. We observe that **B**/**B**$_{\mathbf{AR}}$ ($\boldsymbol{d}$–$\boldsymbol{qa}$) shows superior performance of around 7-10 points in MR over **B**/**B**$_{\mathbf{AR}}$ ($\boldsymbol{d}$–$\boldsymbol{q}\hat{\boldsymbol{a}}$), and also improves in MRR and recall rates. This is expected since the ground-truth rather than predicted answers are included in the dialogue history (along with the ground-truth questions). The metrics sim$_{cap,q}$ and sim$_{\circlearrowleft}$, on the other hand, show very little performance difference across the two evaluation settings. We also note that ranking performance is worse when both image $\boldsymbol{i}$ and caption $\boldsymbol{c}$ are excluded from condition variable. This does not, however, correlate with the CE and KL terms of the loss which are lower for a condition-less setting. We attribute this to the model being transformed from a CVAE to a VAE, hence lifting the burden of capturing the conditional posterior distribution (i.e. the KL is now between an unconditional $q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ and $\mathcal{N}(0, 1)$). Interestingly, however, excluding either the image or the caption achieves similar performance to when both are included, indicating that the caption acts as a good textual proxy of the image (a reassurance of our sim$_{\boldsymbol{c},\boldsymbol{q}}$ metric).

C. Extended Quantitative Results on 2VD task

Extending Tab. 4 in the main paper, Tab. 7 here shows results for **B**/**B**$_{\mathbf{AR}}$ trained with permutations of the image $\boldsymbol{i}$ and caption $\boldsymbol{c}$) (denoted by + if included in the condition, and – otherwise). We note the decrease in CE and KL as conditions ($\boldsymbol{i}, \boldsymbol{c}$) are excluded from the model. This is expected since the task of dialogue generation is made simpler without the constrains of an explicit visual/textual condition.

D. Network architectures and training

The following section provides detailed descriptions of the architectures of our models **A**, **B** and **B**$_{\mathbf{AR}}$. The descriptions are dense but thorough. We also include further details of our training procedure. Where not explicitly noted, each convolutional layer is proceeded by a batch normalisation

Table 6: 1VD evaluation of **B/B$_{\mathbf{AR}}$** on *VisDial* (v0.9) test set. Results show ranking of answer candidates based on the $\mathcal{S}_{w2v}$ scoring function. Note that ***d–qa*** indicates the iterative evaluation method when the *ground-truth* dialogue history is provided, while ***d–q$\hat{a}$***, the iterative evaluation method when the ground-truth question and *predicted* answer history is provided (see §4.2). The + and − indicate models *trained* with and without respective conditions, image $\boldsymbol{i}$ and caption $\boldsymbol{c}$.

Method	$\boldsymbol{i}$	$\boldsymbol{c}$		CE	KLD	MR	MRR	R@1	R@5	R@10	sim$_{cap,q}$	sim$_{\circlearrowleft}$
B	+	+	***d–qa***	18.87	4.36	28.45	0.2927	23.50	29.11	42.29	0.4374	2.68
	+	+	***d–q$\hat{a}$***	25.10	4.02	30.57	0.2188	16.06	20.88	35.37	0.4118	2.42
	−	+		16.80	3.13	27.76	0.3243	26.59	33.21	47.65	0.4491	4.48
	+	−		21.02	4.71	29.82	0.2144	15.25	21.07	34.96	0.4551	5.44
	−	−		19.35	13.34	29.00	0.3026	24.36	30.70	47.62	0.4638	6.17
B$_{\mathbf{AR}}$8	+	+	***d–qa***	15.11	2.53	25.87	0.3553	29.40	36.79	51.19	0.4703	4.30
	+	+	***d–q$\hat{a}$***	25.70	2.21	29.10	0.2864	22.52	29.01	48.43	0.3885	3.47
	−	+		16.19	2.80	26.04	0.3566	29.62	36.75	50.62	0.4626	4.17
	+	−		20.39	2.89	28.99	0.3024	24.33	30.74	47.17	0.4461	8.16
	−	−		20.92	2.84	28.79	0.3045	24.46	30.99	48.10	0.4442	0.18
B$_{\mathbf{AR}}$10	+	+	***d–qa***	16.04	1.89	26.30	0.3422	28.00	35.34	50.54	0.4708	4.84
	+	+	***d–q$\hat{a}$***	24.77	1.81	29.15	0.2869	22.68	28.97	46.98	0.4058	2.85
	−	+		19.97	2.58	26.84	0.3212	25.90	32.92	47.68	0.4424	5.95
	+	−		20.39	2.79	27.27	0.3157	25.45	32.26	47.87	0.4707	13.22
	−	−		19.17	0.00	29.00	0.3026	24.36	30.70	47.62	0.4614	0.00

layer (with momentum $= 0.001$ and learnable parameters) and a *ReLU* activation.

Prior network The prior neural network, parametrised by θ, takes as input the image $\boldsymbol{i}$, the caption $\boldsymbol{c}$ and the dialogue context. For the model **A**, this context is $\boldsymbol{h}_t^+$, containing the dialogue history up to t-1 and the current question $\boldsymbol{q}_t$. For models **B/B$_{\mathbf{AR}}$**, the dialogue context is the null set ($\boldsymbol{h} = \emptyset$). To obtain the image representation, we scale and centre-cropped each image to $3 \times 224 \times 224$ and feed it through *VGG-16* [23]. The output of the penultimate layer is extracted and ℓ_2-normalised (as in [6]) to obtain a 4096-dimensional image feature vector. For the caption, we pass $\boldsymbol{c}$ through a pre-trained *word2vec* [18] model (we do not learn these word embeddings) to obtain $\mathring{\boldsymbol{c}} \in \mathbb{R}^{300 \times L}$ where L is the maximum sentence length ($L = 64$). For the dialogue context (relevant only in the case of **A**) we pass the one-hot encoding of each word through a learnable word embedding module. We stack these embeddings as described in §3.1 of the main paper to obtain $\mathring{\boldsymbol{h}}_t^+ \in \mathbb{R}^{E \times L \times K}$, where E is the word embedding dimension ($E = 256$), L is the maximum sentence length ($L = 64$) and K is the number of dialogue entries at time t. We encode these inputs convolutionally to obtain $\boldsymbol{y}$ (the encoded condition) as follows: $\mathring{\boldsymbol{c}}$ is passed through a convolutional block (output size $64 \times 8 \times 8$) and concatenated with the image feature vector (reshaped to $64 \times 8 \times 8$). The concatenated output is passed through a convolutional block to obtain the jointly encoded image-caption (output size $64 \times 8 \times 8$). If $\boldsymbol{h} \neq \emptyset$, then the context is passed through a convolutional block (output size $64 \times 8 \times 8$) and is concatenated with the encoded image-caption and passed through yet another convolutional block to get the encoded image-caption-context (output size $64 \times 8 \times 8$). We call this the encoded condition $\boldsymbol{y}$. The encoded condition $\boldsymbol{y}$ is then passed through a further convolutional block (output size $256 \times 4 \times 4$) followed by two final convolutional layers (in parallel) to obtain $\boldsymbol{\mu}_p$ and $\log \boldsymbol{\sigma}_p^2$, respectively, the parameters of the conditional prior $p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$. At this stage, $\boldsymbol{\mu}_p$ and $\log \boldsymbol{\sigma}_p^2$ are both of size $512 \times 1 \times 1$ (the latent dimensionality). At test time, a sample is obtained via $\boldsymbol{z} \sim \mathcal{N}(\mathbf{z}; \boldsymbol{\mu}_p, \boldsymbol{\sigma}_p^2)$ and is passed to the decoder in order to generate a sample $\hat{\boldsymbol{a}}_t$ (for **A**) or $\hat{\boldsymbol{d}}$ (for **B/B$_{\mathbf{AR}}$**).

Encoder network The encoder network, parametrised by ϕ, takes $\boldsymbol{x}$ as input along with the encoded condition, $\boldsymbol{y}$, obtained from the prior network. For model **A**, $\boldsymbol{x} = \boldsymbol{a}_t$ and $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}, \boldsymbol{h}_t^+\}$. For models **B/B$_{\mathbf{AR}}$**, $\boldsymbol{x} = \boldsymbol{d} = \langle (\boldsymbol{q}_t, \boldsymbol{a}_t) \rangle_{t=1}^{T}$ and $\boldsymbol{y} = \{\boldsymbol{i}, \boldsymbol{c}\}$. In all models, $\boldsymbol{x}$ is passed through a learnable word embedding module, and the word embeddings stacked (see §3.1 in the main paper) to obtain $\mathring{\boldsymbol{x}} \in \mathbb{R}^{E \times L \times M}$, where $E = 256$, $L = 64$ and M is the number of entries in $\boldsymbol{x}$ (for **A**, $M = 1$ and for **B/B$_{\mathbf{AR}}$** $M = 2T$). In this way, we transform $\boldsymbol{x}$ into a single-channel answer 'image' in the case of **A**, and a multi-channel image of alternating questions and answers in the case of **B/B$_{\mathbf{AR}}$**. $\mathring{\boldsymbol{x}}$ is then passed through a convolutional block (output size $64 \times 8 \times 8$), the output of which is concatenated with $\boldsymbol{y}$ and forwarded through another convolutional block (output size

Table 7: 2VD evaluation on *VisDial* (v0.9) test set for **B/B$_{\mathbf{AR}}$** models. Note that $\boldsymbol{d}$ left blank indicates the block evaluation method, when a whole dialogue is generated given only an image and its caption, while $\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$ indicates the iterative evaluation method when previously generated questions *and* answers are included in the dialogue history (see Section 4.2). The + and – indicate models *trained* with and without respective conditions, image $\boldsymbol{i}$ and caption $\boldsymbol{c}$.

Method	$\boldsymbol{i}$	$\boldsymbol{c}$	$\boldsymbol{d}$	CE	KLD	sim$_{\boldsymbol{c},q}$	sim$_{\circlearrowleft}$
B	+	+	$\emptyset$	31.18	4.34	0.4931	14.20
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	25.40	4.01	0.4091	1.86
	–	+	$\emptyset$	29.09	3.26	0.4889	11.23
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	24.59	3.05	0.3877	3.45
	+	–	$\emptyset$	28.60	4.26	0.4634	15.56
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	29.85	4.66	0.4221	3.54
	–	–	$\emptyset$	19.92	6.42	0.4590	6.34
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	19.34	0.00	0.4638	0.00
B$_{\mathbf{AR}}$8	+	+	$\emptyset$	28.81	2.54	0.4878	31.50
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	26.60	2.29	0.3884	2.39
	–	+	$\emptyset$	30.59	2.72	0.4889	43.17
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	26.15	2.77	0.3758	3.57
	+	–	$\emptyset$	31.51	2.91	0.4602	24.75
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	21.41	2.68	0.4453	5.49
	–	–	$\emptyset$	20.32	2.77	0.4464	0.26
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	21.53	2.99	0.4419	0.10
B$_{\mathbf{AR}}$10	+	+	$\emptyset$	28.49	1.89	0.4927	44.34
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	24.93	1.80	0.4101	2.35
	–	+	$\emptyset$	30.83	2.53	0.4951	38.60
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	28.59	2.52	0.3903	1.91
	+	–	$\emptyset$	30.18	2.89	0.4592	100.81
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	28.32	2.44	0.4334	6.73
	–	–	$\emptyset$	19.60	0.00	0.4585	0.00
			$\boldsymbol{d}$–$\hat{\boldsymbol{q}}\hat{\boldsymbol{a}}$	19.17	0.00	0.4614	0.00

$256 \times 4 \times 4$). This output is forwarded through two final convolutional layers (in parallel) to obtain $\boldsymbol{\mu}_q$ and $\log \boldsymbol{\sigma}_q^2$, the parameters of the conditional latent posterior $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$. Here $\boldsymbol{\mu}_q$ and $\log \boldsymbol{\sigma}_q^2$ are both of size $512 \times 1 \times 1$.

At train time, the KL divergence term of the ELBO is computed using $\{\boldsymbol{\mu}_q, \boldsymbol{\sigma}_q\}$ (from the encoder network) and $\{\boldsymbol{\mu}_p, \boldsymbol{\sigma}_p\}$ (from the prior network).

Decoder network The decoder network (for simplicity, the parameters of the prior and decoder network are subsumed into θ) takes as input a latent $\boldsymbol{z}$ and the encoded condition $\boldsymbol{y}$. During training, $\boldsymbol{z}$ is sampled from a Gaussian parametrised by the $\boldsymbol{\mu}_q$ and exponentiated $\log \boldsymbol{\sigma}_q^2$ outputs of the *encoder* network. This distribution is $q_\phi(\boldsymbol{z} \mid \boldsymbol{x}, \boldsymbol{y})$. At test time, $\boldsymbol{z}$ is sampled from a Gaussian parametrised by the $\boldsymbol{\mu}_p$ and exponentiated $\log \boldsymbol{\sigma}_p^2$ outputs of the *prior* network. This distribution is $p_\theta(\boldsymbol{z} \mid \boldsymbol{y})$. At both train and test time, we employ the commonly-used 're-parametrisation trick' [17] to compute the latent sample as $\boldsymbol{z} = \mu + \epsilon\sigma$ where $\epsilon \sim \mathcal{N}(0, 1)$ and μ and σ correspond to those derived from the encoder or prior network as described above.

The sample $\boldsymbol{z}$ is then transformed through a transpose-convolutional block (output size $64 \times 8 \times 8$), concatenated with $\boldsymbol{y}$ and forwarded through a convolutional block (output size $64 \times 8 \times 8$). This output is forwarded through a second transpose-convolutional block, producing an intermediate output volume of dimension $M \times E \times L$ which we permute to match the size of $\mathring{\boldsymbol{x}}$. As before, $E = 256$, $L = 64$ and $M = 1$ (for **A**) or $M = 2T$ (for **B/B$_{\mathbf{AR}}$**).

Following this, our models diverge in architecture: **A** and **B** employ a standard linear layer which projects the E dimension of the intermediate output to the vocabulary size V. The **B$_{\mathbf{AR}}$** model instead employs an autoregressive module (detailed below) followed by this standard linear layer. At train time, the V-dimensional network output is *softmax*ed and used in the computation of the CE term of the ELBO. At test time, the $argmax$ of the (*softmax*-ed) output is taken to be the index of the word token predicted. We share the weight matrices of the decoder's final linear layer and the encoder and prior's learnable word embedding module (which are the same size by virtue of our network architecture) with the motivation that language encoders and decoders should share common word representations.

Autoregressive block The autoregressive (AR) block (AR-N in Fig. 4 - bottom) in **B$_{\mathbf{AR}}$**'s decoder is inspired by *PixelCNN* [26] which sequentially predicts the pixels in an image along the two spatial dimensions. In the same fashion, we use an autoregressive approach to sequentially predict the next sentence (question or answer) in a dialogue. Since our framework is convolutional with sentences viewable as 'images', our approach can similarly be adapted from that of [26, 9]. We first reshape the intermediate output of the decoder to $E \times L{*}M$ (essentially 'unravelling' the dialogue sequentially into a stack of its word embeddings). We then apply a size-preserving masked convolution to the reshaped output (followed by a learnable batch normalisation and a *ReLU* activation). We call this triplet an AR layer. The masked convolution of the AR layer ensures that future rows (i.e. future E-dimensional word embedding) are hidden in the prediction of the current row/word embedding. We apply N AR layers in this way with each layer taking in the output of the previous AR layer. Following the AR-N block, a linear layer projects the final output's E dimension to the vocabulary size V. We report numbers for $N = \{8, 10\}$. We base our implementation of the AR block on a publicly-available implementation of *PixelCNN*.

E. Dialogue preprocessing

The word vocabulary is constructed from the *VisDial* v0.9 [6] training dialogues (not including the candidate an-

swers). The dialogues are preprocessed as follows: apostrophes are removed, numbers are converted to their worded equivalents, and all exchanges are made lower-case and either padded or truncated to a maximum sequence length ($L = 64$). The vocabulary is also filtered such that words with a frequency of <5 are removed and replaced with the UNK token. After pre-processing and filtering, the vocabulary size is $V = 9710$.

F. Extended Qualitative Results

We present additional qualitative results for the **A** model in Figs. 4 and 5 (1VD task) and for the $\mathbf{B_{AR}}10$ model (under the block evaluation setting) in Figs. 6 and 7 (2VD task) . Note that for both, different colours indicate generations ($\hat{\boldsymbol{a}}_t$ for **A** and $\hat{\boldsymbol{d}}$ for **B**/$\mathbf{B_{AR}}$) from different samples of $\boldsymbol{z}$. In Figs. 6 and 7, whole generated dialogue blocks are shown with coloured sections indicating subsets exhibiting coherent question-answering and white sections indicating subsets that are not entirely coherent.

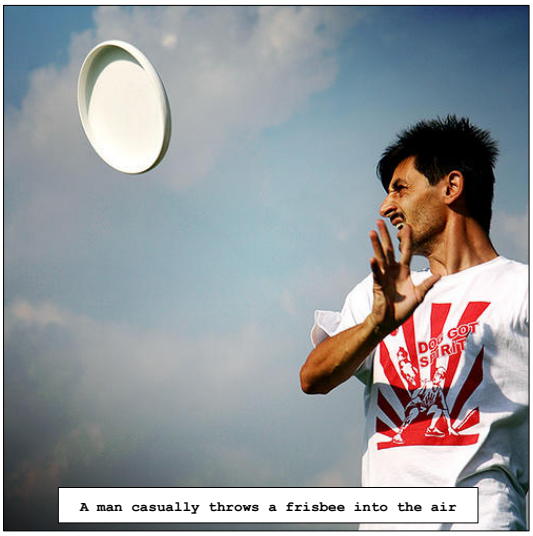
Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
How old is the girl?	Maybe three	I can't tell	Looks about six	She is maybe	I can't tell	Yes
What race is the girl?	White	All	Yes	White	White	Caucasian
Is she outside?	Yes	Yes	No	Yes	Yes	Yes
Is her hair long or short?	Short	Short	Short	Yes	Short	Short
What color is her hair?	Blonde	Blonde	Blonde	Blonde	Brown	Brown
Is her hair curly or straight?	It's straight	Straight	Straight	Short	Straight	Straight
What is she wearing?	Pink shirt, white pants	Clothes, shirt and top	T-shirt and jacket and pants	T-shirt and pants	Shirt and pants	Jeans like a pajamas
Is the teddy bear in her lap?	Yes	There	No	Yes	Yes	Yes
What color is the teddy bear?	White	Brown	Brown	Tan	Tan	Tan and white
Is it nice outside?	Yes sunny	Yes	Yes	It looks sunny	It looks lovely	Yes

Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
How old does she look?	Around seven or eight	I cannot tell about her	Looks about six	She is about teenager around	I can't tell	I can't tell
Any other people?	No	No	Yes	Yes	No	I don't see anyone
Any buildings?	No	Yes	No	No	Secret can't tell	Yes
Is it day or night?	Day	It looks like it in image is in so	Daytime	Day	Day	Yes
Is it raining?	No	Yes	Yes	No	No	No
What color umbrella?	Pink and clear	Dark colored color	White	White	Grey	White
Is it open or closed?	Open	Yes	Open	Open	Open	It is open
Is it sunny?	I can't tell	Yes	No	Yes	Yes	No
What color is her hair?	Dark brown	Brown	Brown	Brown	Brown	Brown
Is it long or short?	Long	Short	Short	I'd say long	Short	Yes

Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
Are there any animals present?	No animals	No	No	Yes	No	No
Are there any kids?	Two kids	Yes	No	Yes	Yes	I
Boys or girls?	Boys	Boy	Boys	I	Boys	Yes
Are there any adults?	One adult	Yes	Yes	Yes	Yes	Yes
What are they doing?	Adult is flying kite and boys are watching	Playing	I can't see their faces	She is watching parasols	Spectating	They are playing watching
Is it sunny?	Yes	Yes	No	I can't tell	Yes	Yes
Are there trees?	Yes	Yes	Yes	Yes	Yes	There are no background
Is kite colorful?	Yes	No	Yes	Yes	Yes	Yes
Is kite in air?	Yes	Yes	Yes	No	Yes	No
Are boys smiling?	Yes	Yes	Yes	No	Yes	Yes

Figure 4: Examples of diverse answer generations from the **A** model for the 1VD task.

Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
Is the man in a park?	I can't tell	It appears so	Yes	Yes	Yes	He is in
How old is the man?	Thirtys maybe	Locate	Maybe late thirtys	Thirtys	Thirty	Twentys
Has the frisbee left his hand?	Yes	Yes	Yes	Yes	Yes	Yes
Can you tell what he's throwing it towards?	No	No	I can't tell, is ball	I can't tell	No	Yes
What does 'casually thrown' mean?	It looks like he just tossed it	Casually	At least relief casually	I	He has in a skateboard shirt	Casually
What color is the frisbee?	White	White	White	Blue	It is white	White
What is the man wearing?	A white screen print tee	T-shirt	T-shirt and shorts	He looks casual	Shorts	T-shirt and shorts shorts
Is he wearing jeans?	I can't see his pants	No	Shorts	Yes	No	Yes
Are there other people around?	No	No	No	No	No	No
Do you see any animals?	No	No	No	No	No	No



Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
Is it a tall wall?	Yes, it goes underwater	Yes	Yes	Yes can't tell	No	No
Is there only one polar bear?	Yes, just one	Yes	One	Yes	One	Yes
Any people there?	No	Yes	I can't see any	No	No	Yes
Do you see quite a lot of water?	Yes the polar bear is underwater	Yes	I can't see the ground	No	Yes can't see any	Yes can't tell
Can you see the sky?	No	No	Yes	Yes	Yes	Yes
Is this picture in color?	Yes it is	Yes	Yes	Yes	Yes	Yes
Is this a professional picture?	I don't think it is	Looks like relaxed	I	Yes	No	Yes
Do you see any buildings?	No	No	Yes	There are no buildings	I	No
Any trees?	No	No	There see zero	Yes	Yes	No
Is this a close-up type of photo?	No	Yes	Yes	Yes	Yes	It looks like it is in frame

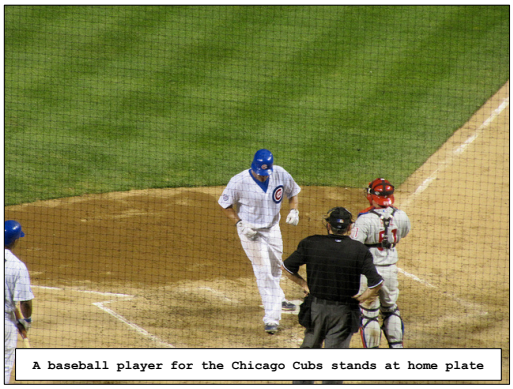
Question	Ground-truth answer	z_1	z_2	z_3	z_4	z_5
How many children are there?	Many	There are six	One	Three	There	More than ten ten
All boys or girls?	Both	I can't tell	Boy	Girl and each	It appears boy girl	They are so
How old do they look?	Six	Five	Ten-twelve	Twenty	Yes	Five looks six
Are they outside?	Yes	I	Yes	Yes	Yes	Yes
Is it sunny out?	Maybe	Yes	Yes	Yes	Yes	Yes
Are they in a field?	Yes	Yes	No	No	Yes	No
Are there many adults around?	Zero	No	Yes	No	Yes	No
Is there only one ball?	Yes	Yes	Yes	Yes	Yes	Yes
Are there soccer nets?	Not visible	No	No	Yes	I can't say	Yes
What's in the background?	Fence	Trees	I see the grass	There are numerous snow trees	Lots	I see some grass

Figure 5: Examples of diverse answer generations from the **A** model for the 1VD task – continued.

How many sheep are there?	Four
Are you any any water?	No
What color are they?	Brown
Any people?	No
What are the the?	Grazing
Are there people?	No
Is it sunny?	Yes
Can any animals?	No
Are any any?	No
Is any sunny?	Yes

Is this a color?	Yes
Can you see any?	No
Can you see trees?	No
Any people?	No
Is it sunny?	No
Are it sunny?	Yes
Any people?	No
Any people?	No
What time the day?	Can't tell
Are they eating?	Yes

Can you see any people?	No
Is the photo in color?	Yes
What is the fence made?	I is
Can you see any?	No
How many sheep are there?	I are
What color is the shacks?	It
Does the grass have?	Yes
Can you see any sky?	No
Is the grass green?	Yes
Do you see any sky?	No



Is the people male?	Yes
Is the person wearing?	Yes
Can you see any sky?	No
What color is the uniforms?	White
Is there grass in?	Yes
How many people are there?	Two
Is it sunny sunny?	Yes
Can you see any?	Yes
What color is the sky?	The
Is it sunny?	Yes

Can you see the ball?	Yes, I
Is the player a?	Yes, is is
Is it a a a?	Yes, is is
Can you see the baseball?	Yes, is is
What is the the wearing?	It is wearing
How is the the bat?	I is is
Is there any other people?	Yes
Can you see the pitcher?	No, the see
Can you see the scoreboard?	No
Can you see the sky?	No

Is this a professional game?	Yes
Any many players are?	One
Is the a?	Yes
Can you see the pitcher?	No
Is the a in?	Yes
Is it a in the?	Yes
What color is the batters?	White
Is there people in the?	No
Is the pitcher in?	Yes
Is it sunny?	Yes

What color is?	White
Is there a fridge?	Yes
Is there any people?	No
Is it a?	Yes
Is there a window?	No
Is there a?	No
Is there a?	Yes
Is the fridge on?	Yes
Can you see the window?	No
Is it a?	Yes

What color is the fridge?	White
What color is?	White
Can you see the sink?	Yes
Is there a window?	No
Any people?	No
Is there?	No
How pics the?	No
What color is the walls?	White
Any windows?	No
How about on?	Two

Can you see the?	No
What appliances?	Microwave
Is the fridge on?	Yes
Is it a?	Yes
Is the fridge have magnets?	Yes
Can you see what time?	Colgate
Is there a on the?	No
Can you tell what time?	No
Is there a?	Yes
Are there any people?	No

Is the photo in color?	Yes
Is it a professional photo?	Yes
Are the men wearing or?	Yes
Is any of wearing?	Yes
Are there any people in the?	Yes
How many men?	Two
Are the men wearing wearing uniforms?	Yes
Are there see other people?	No
What color are the uniforms?	White
Do the have the the?	Yes

What color are the uniforms?	One is white
Are they wearing?	Yes
Are there a?	Yes
Is there a?	Yes
Is there trees?	Yes
Is it sunny?	Yes
Are the wearing wearing?	Yes they does
Are they wearing cleats?	Yes
Are there a fence?	Yes
Can you see the?	No

What is the the of?	One
Are there see other	Yes
Are they people uniforms?	Yes
Is they wearing uniforms?	Yes
What color are the uniforms?	White
Does the have wearing hair	Yes
What color the wearing?	A
Can you see any ball?	No
Can you see any coaches?	No
Do you see the color?	Yes

Figure 6: Diverse two-way dialogue generations from the $\mathbf{B}_{\mathbf{AR}}10$ model (block evaluation) for the 2VD task.

Is the picture in color?	Yes
Is the photo in?	Yes
What color is the frisbee?	White
Can there?	No
Is the grass visible?	Yes
Is the sunny?	Yes
Is it sunny?	Yes
Is there other people?	No
What color is the frisbee?	White
Is the man green?	Yes

How many people are there?	Two
What are they?	I
Can you see the frisbee?	Yes
What color of the frisbee?	Blue
Is there people?	Yes
Is it sunny?	Yes
What color is the?	I
Is there sunny?	Yes
How many people are there?	I
Is they having fun?	Yes

What color is the man's?	White
What is the man man?	I
How old is the man?	I
Can you see any ball?	Yes
Is there any other people?	No
Does the man have a?	No
Is the man wearing a hat?	No
Does he man have a?	No
Are there see any other people?	No
Can you see any sky?	No

Is it big?	Yes
How is of the is?	I
Can you see the breed?	No
Are there any people?	No
Is the dog on?	No
Can you tell the breed of it?	No
What is the color?	It is
Any there see the driver?	No
Is it a Harley?	Yes
Any other?	No

Are there only one?	Yes
What kind of dog is it?	I is like
Does it look like be?	It, it
Is there any people in the boat?	No
What time of day is it?	It is light
What color is the boat?	White
How it is like is?	I is
Is it a or or?	It Is
What is of day is?	It is
What color is the boat?	It

What kind of dog is it?	It looks like a mutt
Does he have a collar?	Yes
Can you see what breed?	No
Is it people boat the the boat?	No
Is there oars?	No
How the boat the the?	I
Is it sunny?	Yes
How many barges?	No
Any people?	No
Any?	No

What color is the man?	White
What color is his hair?	Black
How old is he?	Thirtys
Is he have other?	Yes
Is it sunny?	Yes
Can you see any?	Yes
Is it court?	Yes
Is it sunny?	Yes
Is it sunny?	Yes
Can you see sky?	No

Is this a color photo?	Yes
What color is the tennis?	It
Are there any other people in the?	No
Can you see any tennis?	No
Can you see any people in the?	No
What time of day is it?	It
Is the appear like be be?	It
Are there any other in the the?	No
Can you see the time?	No
Can you see the season?	No

What is the color?	Is
Is the player?	Yes
Is it a?	Yes
Is it outdoors?	Yes
Is there people?	No
Is he playing?	Yes
Is it a?	Yes
Is it a?	Yes
Can you see the net?	No
Is it man?	Yes

Figure 7: Diverse two-way dialogue generations from the $\mathbf{B}_{\mathbf{AR}}10$ model (block evaluations) for the 2VD task – continued