𝕍𝔻-𝔾ℝ: Boosting 𝕍isual 𝔻ialog with Cascaded Spatial-Temporal Multi-Modal 𝔾ℝaphs
-Supplementary Material-

Appendix

1. Limitations

Although our model managed to outperform previous models on four challenging datasets, it is important to acknowledge some of its limitations: First, 𝕍𝔻-𝔾ℝ leverages extra data in the form of adjacency matrices of the multi-modal GNNs and relies on external models to acquire them. Although inferring these models on the VisDial data is cheap, this approach can lead to inaccurate predictions of adjacency matrices, especially for the question and history modalities. Thus, by keeping the graph structures constant, our model's performance might be limited by this introduced noise. This could be remedied in future work by jointly learning the graphs' parameters as well as refining their structures over time [2–4]. Second, similar to almost all previous methods on this task, we did not manage to achieve new state-of-the-art performance across all metrics of this challenging dataset. Finally, inline with previous works [1, 5, 7, 9–11], fine-tuning our model (both in the single model as well as the ensemble setting) on dense annotations improved the most relevant metric of the dataset, i.e. the NDCG score, at the expense of the other (sparse) ones. Although our model's performance dropped with respect to the sparse metrics, we managed to outperform previous works by achieving an NDCG score of 76.43, which is the main objective of dense annotation fine-tuning.

2. Graph Construction and Pruning

Image Modality. Given two object features $\mathbf{v}_i$ and $\mathbf{v}_j$, their bounding boxes and centre coordinates (x_i, y_i) and (x_j, y_j), we computed the value of their intersection over unions IoU_{ij} and relative angle ϕ_{ij}. As shown in Figure 4, there are two spacial cases: The first occurs when the bounding box of $\mathbf{v}_i$ completely includes the bounding box of $\mathbf{v}_j$ and this class is denoted as *inside* with index $i = 1$. The second occurs when the bounding box of $\mathbf{v}_i$ is entirely covered by the bounding box of $\mathbf{v}_j$. This class is denoted as *cover* with index $i = 2$. The remaining classes are solely determined by the value of IoU_{ij}. If $\text{IoU}_{ij} \geq 0.5$, then the relationship between the objects is denoted as *overlap* and has the index $i = 3$. Finally, if $\text{IoU}_{ij} < 0.5$, the class index is computed as

$$i = \lceil \frac{\phi_{ij}}{0.25\pi} \rceil + 3.$$

By construction, all classes of index $i \neq 3$ are pairwise symmetric as can be seen from Figure 1a where we plotted the distribution of the different image graph relationship classes over the training split of VisDial v1.0.

Method	VisDialBERT [7]	VD-BERT [9]	VD-PCR [11]	𝕍𝔻-𝔾ℝ
# Parameters	250M	250M	255M	260M
Tr. time / epoch	0.6h	0.6h	1.00h	1.05h

Table 1. Model complexity and runtime comparison with respect to VisDial v1.0 on our hardware setup.

Question Modality. The question graph relationship classes were determined by the dependency between the question words. To this end, we input each question to the Stanza dependency parser that output the classes between the different word pairs resulting in a total of 47 classes. As shown in Figure 1b, the distribution of these classes within the VisDial v1.0 training split is not uniform with `det` and `nsubj` being the most frequent. We illustrate a qualitative sample in Figure 2.

History Modality. We relied in coreference resolution to construct the history graph. Specifically, an edge exists between two rounds `i` and `j` (`i` > `j`) if and only if a word in round `j` was used to reference another word in round `i`. The only exception is the caption `C` that links to all upcoming rounds in the history even if there is no explicit coreference between them. We posit that the caption is complementary to the visual input and helps the model better understand the scene. We illustrate a qualitative sample in Figure 3.

3. Model Complexity

The overhead for constructing the multi-modal graphs only incurs once during a *cheap* offline pre-processing stage and therefore does not lead to crucial increase in compute complexity, i.e. number of trainable parameters and epoch training time, compared with previous seminal models, e.g. VisDial-BERT [7], VD-BERT [9], and VD-PCR [11], as can be seen in Table 1.

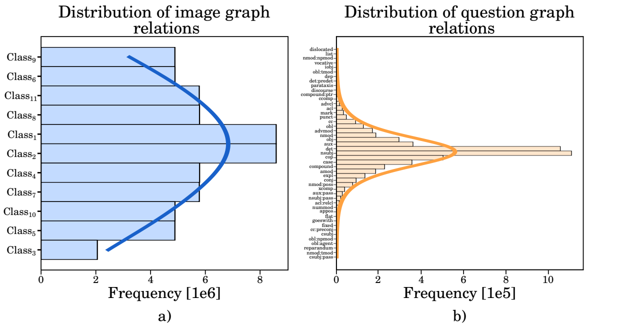


Figure 1. **Right:** The distribution of the image graph relationship classes within the training split of VisDial v1.0. **Left:** The distribution of the question graph relationship classes within the training split of VisDial v1.0.

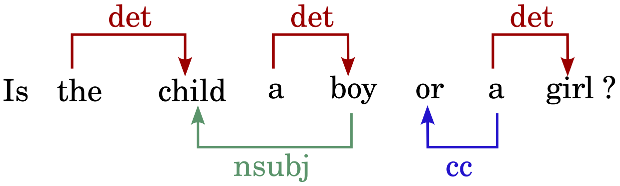


Figure 2. A qualitative sample of the dependency relationships between question word pairs.

4. Training Details

We implemented our model using PyTorch [8] and trained it on a server with 8 NVIDIA Tesla V100 GPUs using an effective batch size of 64 and Adam optimiser [6] with a linear decay learning rate schedule with warm-up. We set the initial learning rates of the BERT and GNN weights to 5×10^{-6} and 5×10^{-4}, respectively. Furthermore, we set the loss coefficients $\alpha_1 = \alpha_2 = 1$ and the residual connection coefficient $\lambda = 0.5$. We refer to Table 2 for a complete overview of our experimental setup.

5. Additional Qualitative Results

We present additional qualitative examples from the *val* split of VisDial v1.0 in Figure 5 and Figure 6. As in the main text, we compared the top-1 predictions of $\mathbb{VD}$-$\mathbb{GR}$ with the ground-truth and the predictions of VD-PCR since it achieved the second best performance on this split.

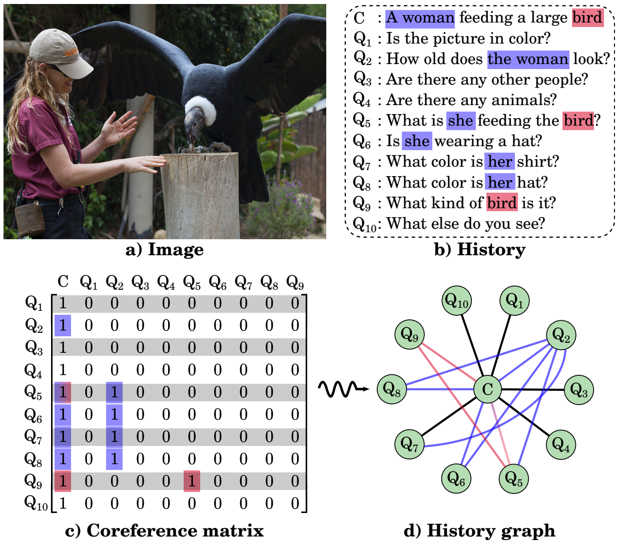

C : A woman feeding a large bird
Q1 : Is the picture in color?
Q2 : How old does the woman look?
Q3 : Are there any other people?
Q4 : Are there any animals?
Q5 : What is she feeding the bird?
Q6 : Is she wearing a hat?
Q7 : What color is her shirt?
Q8 : What color is her hat?
Q9 : What kind of bird is it?
Q10: What else do you see?
a) Image
b) History
C Q1 Q2 Q3 Q4 Q5 Q6 Q7 Q8 Q9
Q1 1 0 0 0 0 0 0 0 0 0
Q2 1 0 0 0 0 0 0 0 0 0
Q3 1 0 0 0 0 0 0 0 0 0
Q4 1 0 0 0 0 0 0 0 0 0
Q5 1 0 1 0 0 0 0 0 0 0
Q6 1 0 1 0 0 0 0 0 0 0
Q7 1 0 1 0 0 0 0 0 0 0
Q8 1 0 1 0 0 0 0 0 0 0
Q9 1 0 0 0 0 1 0 0 0 0
Q10 1 0 0 0 0 0 0 0 0 0
Q10
Q1
Q9
Q2
Q8
C
Q3
Q7
Q4
Q6
Q5
c) Coreference matrix
d) History graph

Figure 3. A qualitative sample of the coreference relationships between different dialog rounds. The hub-node was not visualised for clarity.

References

[1] Cheng Chen, Yudong Zhu, Zhenshan Tan, Qingrong Cheng, Xin Jiang, Qun Liu, and Xiaodong Gu. UTC: A Unified Transformer with Inter-Task Contrastive Learning for Visual Dialog. In *CVPR*, 2022. 1

[2] Yu Chen, Lingfei Wu, and Mohammed J. Zaki. Iterative deep graph learning for graph neural networks: Better and robust node embeddings. In *NeurIPS*, 2020. 1

[3] Pantelis Elinas, Edwin V. Bonilla, and Louis C. Tiao. Variational inference for graph convolutional networks in the absence of graph data and adversarial settings. In *NeurIPS*, 2020. 1

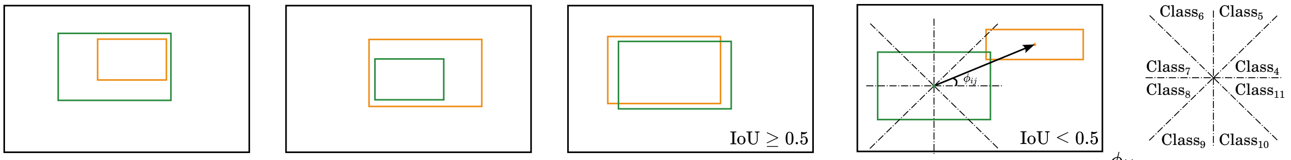
[4] Luca Franceschi, Mathias Niepert, Massimiliano Pontil, and Xiao He. Learning discrete structures for graph neural networks. In *ICML*, 2019. 1

[5] Gi-Cheon Kang, Junseok Park, Hwaran Lee, Byoung-Tak Zhang, and Jin-Hwa Kim. Reasoning visual dialog with sparse graph learning and knowledge transfer. In *Findings of EMNLP*, 2021. 1

[6] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[7] Vishvak Murahari, Dhruv Batra, Devi Parikh, and Abhishek Das. Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In *ECCV*, 2020. 1, 2

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In *NeurIPS*, 2019. 2



a) Class$_1$: Inside

b) Class$_2$: Cover

c) Class$_3$: Overlap

d) Class$_{4\text{-}11}$: $i = \lceil \frac{\phi_{ij}}{0.25\pi} \rceil + 3$

Figure 4. The different spatial relationships (without the hub-node relationship) used in constructing the image graph. The orange and green rectangles correspond to the bonding boxes of two objects within the scene.

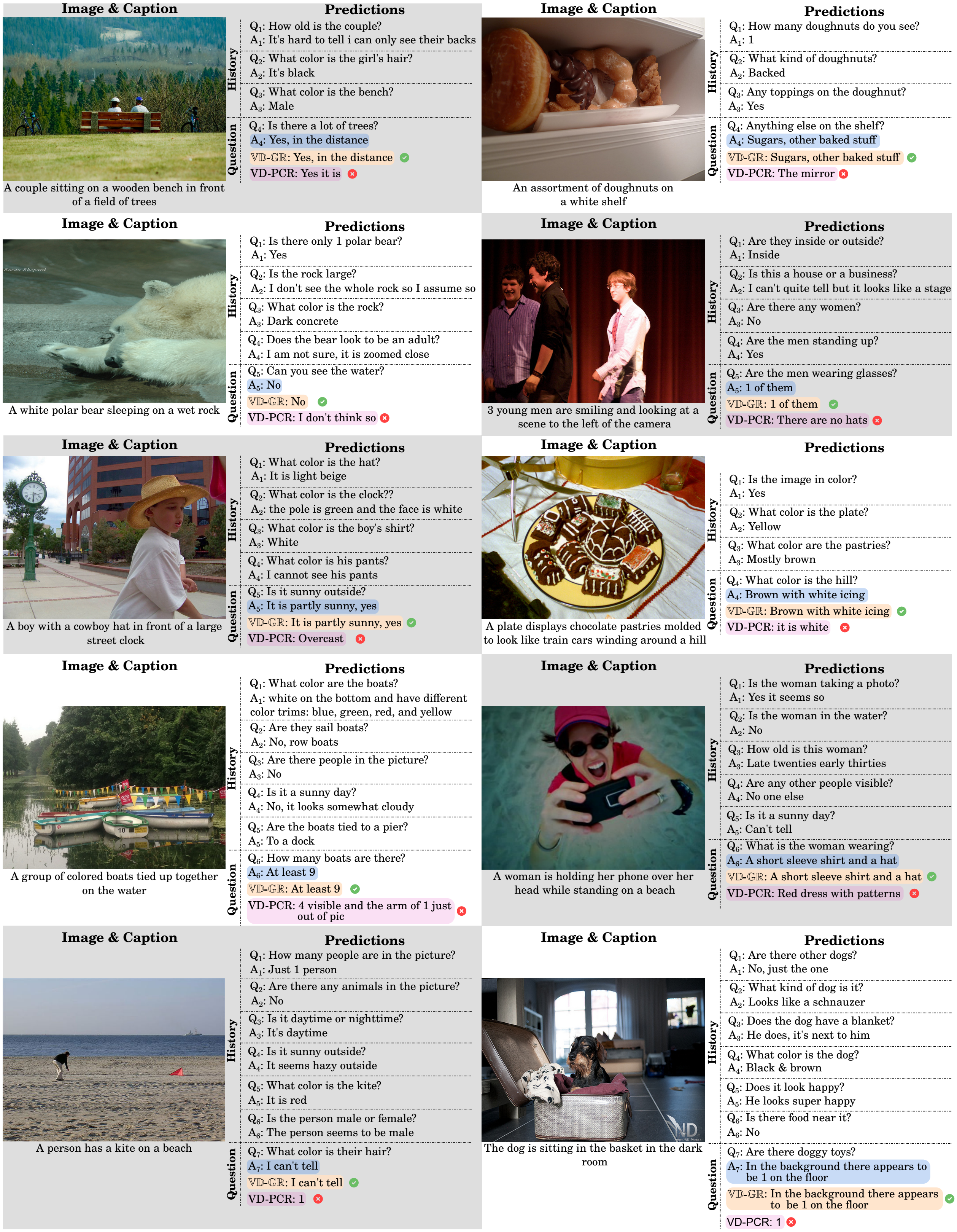
Hyper-parameter	**Training Stage**		
	Warm-up	**Sparse fine-tuning**	**Dense fine-tuning**
Number of GNN layers K	2	2	2
Number of GNN heads H	4	4	4
Residual connection coefficient λ	0.5	0.5	0.5
Dimension of GNN$_{\mathcal{I}}$ node features	1024	1024	1024
Dimension of GNN$_{\mathcal{Q}}$ node features	768	768	768
Dimension of GNN$_{\mathcal{H}}$ node features	768	768	768
Dimension of GNN$_{\mathcal{I}}$ edge features	12	12	12
Dimension of GNN$_{\mathcal{Q}}$ edge features	48	48	48
Dimension of GNN$_{\mathcal{H}}$ edge features	2	2	2
Dimension of Linear$_{\mathcal{I}\to\mathcal{H}}(.)$	$(1024, 768)$	$(1024, 768)$	$(1024, 768)$
Dimension of Linear$_{\mathcal{Q}\to\mathcal{I}}(.)$	$(768, 1024)$	$(768, 1024)$	$(768, 1024)$
Maximum number of text tokens	256	256	256
Maximum number of image regions	37	37	37
Text token mask probability	0.1	0.1	–
Image region mask probability	0.1	0.1	–
Graph edge mask probability	0.15	–	–
Optimiser	`Adam`	`Adam`	`Adam`
Minimum learning rate of BERT parameters	0	0	1×10^{-5}
Minimum learning rate of GNN parameters	0	0	1×10^{-5}
Maximum learning rate of BERT parameters	5×10^{-6}	5×10^{-6}	2×10^{-5}
Maximum learning rate of GNN parameters	5×10^{-4}	5×10^{-4}	1×10^{-4}
Learning rate warm-up of BERT parameters	`True`	`True`	`True`
Learning rate warm-up of GNN parameters	`True`	`True`	`True`
Learning rate schedule of BERT parameters	`Linear`	`Linear`	`Linear`
Learning rate schedule of GNN parameters	`Linear`	`Linear`	`Linear`
Training Loss	$\mathcal{L}_{\text{warm}}$	$\mathcal{L}_{\text{VD}}$	$\mathcal{L}_{\text{CE}}$ / $\mathcal{L}_{\text{ListNet}}$
Number of epochs	5	20	3
Effective batch size	64	64	100
GPU Model	`Tesla V100-32GB`	`Tesla V100-32GB`	`Tesla V100-32GB`
Number of GPUs	8	8	8
Distributed training	`Apex`	`Apex`	`PyTorch DP`

Table 2. Hyper-parameter settings of $\mathbb{VD}$-$\mathbb{GR}$ for the different stages of training. Linear$_{\mathcal{I}\to\mathcal{H}}(.)$ and Linear$_{\mathcal{Q}\to\mathcal{I}}(.)$ denote the linear layers that produce the history and image hub-node features, respectively.

[9] Yue Wang, Shafiq Joty, Michael R. Lyu, Irwin King, Caiming Xiong, and Steven C.H. Hoi. VD-BERT: A unified vision and dialog transformer with BERT. In *EMNLP*, 2020. 1, 2

[10] Zihao Wang, Junli Wang, and Changjun Jiang. Unified multimodal model with unlikelihood training for visual dialog. In *ACM MM*, 2022. 1

[11] Xintong Yu, Hongming Zhang, Ruixin Hong, Yangqiu Song, and Changshui Zhang. VD-PCR: Improving visual dialog with pronoun coreference resolution. *Pattern Recognition*, 2022. 1, 2



A couple sitting on a wooden bench in front of a field of trees

Q$_1$: How old is the couple?
A$_1$: It's hard to tell i can only see their backs
Q$_2$: What color is the girl's hair?
A$_2$: It's black
Q$_3$: What color is the bench?
A$_3$: Male
Q$_4$: Is there a lot of trees?
A$_4$: Yes, in the distance
VD-GR: Yes, in the distance ✓
VD-PCR: Yes it is ✗

An assortment of doughnuts on a white shelf

Q$_1$: How many doughnuts do you see?
A$_1$: 1
Q$_2$: What kind of doughnuts?
A$_2$: Backed
Q$_3$: Any toppings on the doughnut?
A$_3$: Yes
Q$_4$: Anything else on the shelf?
A$_4$: Sugars, other baked stuff
VD-GR: Sugars, other baked stuff ✓
VD-PCR: The mirror ✗

A white polar bear sleeping on a wet rock

Q$_1$: Is there only 1 polar bear?
A$_1$: Yes
Q$_2$: Is the rock large?
A$_2$: I don't see the whole rock so I assume so
Q$_3$: What color is the rock?
A$_3$: Dark concrete
Q$_4$: Does the bear look to be an adult?
A$_4$: I am not sure, it is zoomed close
Q$_5$: Can you see the water?
A$_5$: No
VD-GR: No ✓
VD-PCR: I don't think so ✗

3 young men are smiling and looking at a scene to the left of the camera

Q$_1$: Are they inside or outside?
A$_1$: Inside
Q$_2$: Is this a house or a business?
A$_2$: I can't quite tell but it looks like a stage
Q$_3$: Are there any women?
A$_3$: No
Q$_4$: Are the men standing up?
A$_4$: Yes
Q$_5$: Are the men wearing glasses?
A$_5$: 1 of them
VD-GR: 1 of them ✓
VD-PCR: There are no hats ✗

A boy with a cowboy hat in front of a large street clock

Q$_1$: What color is the hat?
A$_1$: It is light beige
Q$_2$: What color is the clock??
A$_2$: the pole is green and the face is white
Q$_3$: What color is the boy's shirt?
A$_3$: White
Q$_4$: What color is his pants?
A$_4$: I cannot see his pants
Q$_5$: Is it sunny outside?
A$_5$: It is partly sunny, yes
VD-GR: It is partly sunny, yes ✓
VD-PCR: Overcast ✗

A plate displays chocolate pastries molded to look like train cars winding around a hill

Q$_1$: Is the image in color?
A$_1$: Yes
Q$_2$: What color is the plate?
A$_2$: Yellow
Q$_3$: What color are the pastries?
A$_3$: Mostly brown
Q$_4$: What color is the hill?
A$_4$: Brown with white icing
VD-GR: Brown with white icing ✓
VD-PCR: it is white ✗

A group of colored boats tied up together on the water

Q$_1$: What color are the boats?
A$_1$: white on the bottom and have different color trims: blue, green, red, and yellow
Q$_2$: Are they sail boats?
A$_2$: No, row boats
Q$_3$: Are there people in the picture?
A$_3$: No
Q$_4$: Is it a sunny day?
A$_4$: No, it looks somewhat cloudy
Q$_5$: Are the boats tied to a pier?
A$_5$: To a dock
Q$_6$: How many boats are there?
A$_6$: At least 9
VD-GR: At least 9 ✓
VD-PCR: 4 visible and the arm of 1 just out of pic ✗

A woman is holding her phone over her head while standing on a beach

Q$_1$: Is the woman taking a photo?
A$_1$: Yes it seems so
Q$_2$: Is the woman in the water?
A$_2$: No
Q$_3$: How old is this woman?
A$_3$: Late twenties early thirties
Q$_4$: Are any other people visible?
A$_4$: No one else
Q$_5$: Is it a sunny day?
A$_5$: Can't tell
Q$_6$: What is the woman wearing?
A$_6$: A short sleeve shirt and a hat
VD-GR: A short sleeve shirt and a hat ✓
VD-PCR: Red dress with patterns ✗

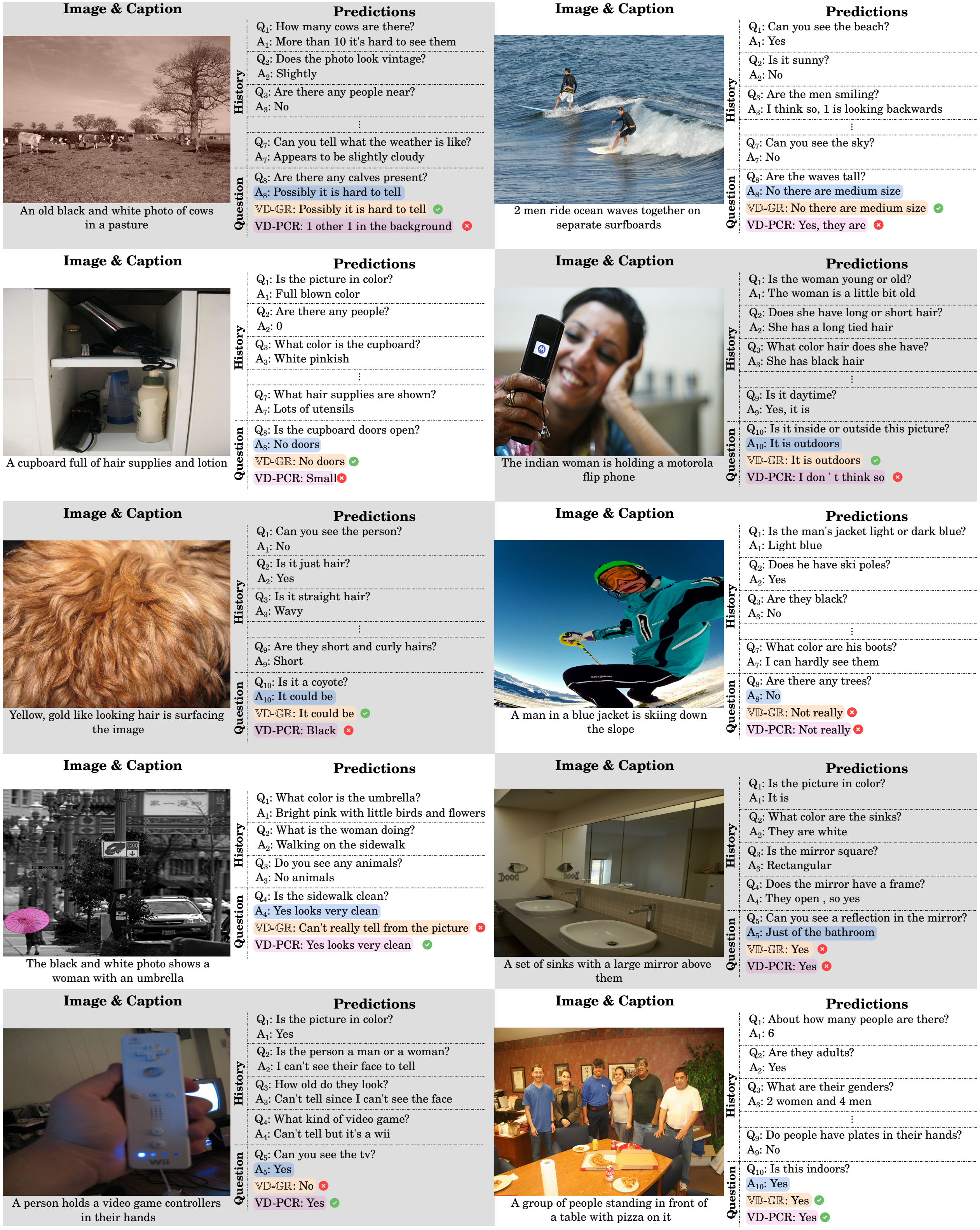
A person has a kite on a beach

Q$_1$: How many people are in the picture?
A$_1$: Just 1 person
Q$_2$: Are there any animals in the picture?
A$_2$: No
Q$_3$: Is it daytime or nighttime?
A$_3$: It's daytime
Q$_4$: Is it sunny outside?
A$_4$: It seems hazy outside
Q$_5$: What color is the kite?
A$_5$: It is red
Q$_6$: Is the person male or female?
A$_6$: The person seems to be male
Q$_7$: What color is their hair?
A$_7$: I can't tell
VD-GR: I can't tell ✓
VD-PCR: 1 ✗

The dog is sitting in the basket in the dark room

Q$_1$: Are there other dogs?
A$_1$: No, just the one
Q$_2$: What kind of dog is it?
A$_2$: Looks like a schnauzer
Q$_3$: Does the dog have a blanket?
A$_3$: He does, it's next to him
Q$_4$: What color is the dog?
A$_4$: Black & brown
Q$_5$: Does it look happy?
A$_5$: He looks super happy
Q$_6$: Is there food near it?
A$_6$: No
Q$_7$: Are there doggy toys?
A$_7$: In the background there appears to be 1 on the floor
VD-GR: In the background there appears to be 1 on the floor ✓
VD-PCR: 1 ✗

Figure 5. Qualitative samples from the *val* split of the VisDial v1.0 dataset. We compared the predictions of VD-GR (orange) with the ground truth answers (blue) and the predictions of VD-PCR (pink). The answers of both models correspond to the top-1 predictions.



An old black and white photo of cows in a pasture

Q_1: How many cows are there?
A_1: More than 10 it's hard to see them
Q_2: Does the photo look vintage?
A_2: Slightly
Q_3: Are there any people near?
A_3: No
⋮
Q_7: Can you tell what the weather is like?
A_7: Appears to be slightly cloudy
Q_8: Are there any calves present?
A_8: Possibly it is hard to tell
𝕍𝔻-𝔾ℝ: Possibly it is hard to tell ✓
VD-PCR: 1 other 1 in the background ✗

2 men ride ocean waves together on separate surfboards

Q_1: Can you see the beach?
A_1: Yes
Q_2: Is it sunny?
A_2: No
Q_3: Are the men smiling?
A_3: I think so, 1 is looking backwards
⋮
Q_7: Can you see the sky?
A_7: No
Q_8: Are the waves tall?
A_8: No there are medium size
𝕍𝔻-𝔾ℝ: No there are medium size ✓
VD-PCR: Yes, they are ✗

A cupboard full of hair supplies and lotion

Q_1: Is the picture in color?
A_1: Full blown color
Q_2: Are there any people?
A_2: 0
Q_3: What color is the cupboard?
A_3: White pinkish
⋮
Q_7: What hair supplies are shown?
A_7: Lots of utensils
Q_8: Is the cupboard doors open?
A_8: No doors
𝕍𝔻-𝔾ℝ: No doors ✓
VD-PCR: Small✗

The indian woman is holding a motorola flip phone

Q_1: Is the woman young or old?
A_1: The woman is a little bit old
Q_2: Does she have long or short hair?
A_2: She has a long tied hair
Q_3: What color hair does she have?
A_3: She has black hair
⋮
Q_9: Is it daytime?
A_9: Yes, it is
Q_{10}: Is it inside or outside this picture?
A_{10}: It is outdoors
𝕍𝔻-𝔾ℝ: It is outdoors ✓
VD-PCR: I don ' t think so ✗

Yellow, gold like looking hair is surfacing the image

Q_1: Can you see the person?
A_1: No
Q_2: Is it just hair?
A_2: Yes
Q_3: Is it straight hair?
A_3: Wavy
⋮
Q_9: Are they short and curly hairs?
A_9: Short
Q_{10}: Is it a coyote?
A_{10}: It could be
𝕍𝔻-𝔾ℝ: It could be ✓
VD-PCR: Black ✗

A man in a blue jacket is skiing down the slope

Q_1: Is the man's jacket light or dark blue?
A_1: Light blue
Q_2: Does he have ski poles?
A_2: Yes
Q_3: Are they black?
A_3: No
⋮
Q_7: What color are his boots?
A_7: I can hardly see them
Q_8: Are there any trees?
A_8: No
𝕍𝔻-𝔾ℝ: Not really ✗
VD-PCR: Not really ✗

The black and white photo shows a woman with an umbrella

Q_1: What color is the umbrella?
A_1: Bright pink with little birds and flowers
Q_2: What is the woman doing?
A_2: Walking on the sidewalk
Q_3: Do you see any animals?
A_3: No animals
Q_4: Is the sidewalk clean?
A_4: Yes looks very clean
𝕍𝔻-𝔾ℝ: Can't really tell from the picture ✗
VD-PCR: Yes looks very clean ✓

A set of sinks with a large mirror above them

Q_1: Is the picture in color?
A_1: It is
Q_2: What color are the sinks?
A_2: They are white
Q_3: Is the mirror square?
A_3: Rectangular
Q_4: Does the mirror have a frame?
A_4: They open , so yes
Q_5: Can you see a reflection in the mirror?
A_5: Just of the bathroom
𝕍𝔻-𝔾ℝ: Yes ✗
VD-PCR: Yes ✗

A person holds a video game controllers in their hands

Q_1: Is the picture in color?
A_1: Yes
Q_2: Is the person a man or a woman?
A_2: I can't see their face to tell
Q_3: How old do they look?
A_3: Can't tell since I can't see the face
Q_4: What kind of video game?
A_4: Can't tell but it's a wii
Q_5: Can you see the tv?
A_5: Yes
𝕍𝔻-𝔾ℝ: No ✗
VD-PCR: Yes ✓

A group of people standing in front of a table with pizza on it

Q_1: About how many people are there?
A_1: 6
Q_2: Are they adults?
A_2: Yes
Q_3: What are their genders?
A_3: 2 women and 4 men
⋮
Q_9: Do people have plates in their hands?
A_9: No
Q_{10}: Is this indoors?
A_{10}: Yes
𝕍𝔻-𝔾ℝ: Yes ✓
VD-PCR: Yes ✓

Figure 6. Qualitative samples from the *val* split of the VisDial v1.0 dataset. We compared the predictions of 𝕍𝔻-𝔾ℝ (orange) with the ground truth answers (blue) and the predictions of VD-PCR (pink). The answers of both models correspond to the top-1 predictions.