Supplementary for Benchmarking Out-of-Distribution Detection in Visual Question Answering

1. In/Out-of-Distribution Samples of VQA Data

To further explore the attributes of different VQA datasets used in our VQA OOD benchmark, we started a game to guess the dataset from which a randomly collected VQA sample came. Through this game, we aimed to showcase the different distributions of data sampled from different datasets in an intuitive and clear way.

In Figure 1, we have listed a batch of randomly selected VQA samples from **VQAv2** [5], **GQA** [9], **CLEVR** [11], **VizWiz** [6], **VQA Abstract Scene** [2], and **QRPE** [15]. We have concealed the dataset name of each sample and will release them at the left-bottom of this page [1]. As discussed, some datasets may have strong biases in either visual or linguistic modalities, or both, such as CLEVR and VQA Abstract Scene, making them more distinguishable.

On the other hand, some of them may share some visual or linguistic similarity (GQA, VizWiz, and QRPE) with in-distribution data (VQAv2), making it difficult to determine their origin with the information from a single modality. The data from QRPE is more challenging since it has the same visual and linguistic distribution with VQAv2 but novel combinations.

2. Training Configuration of VQA Methods

Implementation Details of BUTD. We adapt our BUTD [1] implementation from [18]. A question is encoded via an LSTM [8] with GloVe word embeddings [16] into a 1024-dimension representation. N object features are extracted with VinVL-based [21] object detection model. A softmax score, $A_0 \in R^{N\times 1}$, is computed from the concatenation of visual and question features for each object. A multimodal representation is computed as the element-wised multiplication of the question-attended image and question representation and then projected to the answer domain. The model is trained with Adamax [12] for 13 epochs. The learning rate is $1e-4$

Implementation Details of MCAN. We adapt the MCAN [19] model from the same repository. Questions and images are pre-processed as in BUTD to word tokens and object features. The model is based on the Encoder-Decoder structure introduced in [19] where the encoder and decoder consist of 6 and 12 transformer [10] layers, respectively. Each attention layer has 8 heads, and the dimension of hidden states is 1024. The model is trained for 13 epochs with Adam [12] at an initial learning rate of $7e-5$.

Implementation Detail of X-VLM and X-VLM*. The X-VLM-based VQA model [20] consists of a pretrained X-VLM encoder and a randomly-initialized transformer-based answer decoder. The image module of X-VLM encoder is initialized with a Swin transformer with a window size of 7 trained on ImageNet-22K [3]. The question and cross-modality encoders are initialized with the first and the last 6 transformer layers of the base Bert model released by [4]. In Table 2 of the main paper, X-VLM is the pretrained model released by [20] and X-VLM* is the model trained from scratch only on VQAv2 with the same initialization. We finetune the model with AdamW [14] and a learning rate of $5e-5$ for 10 epochs. More details can be found in [20]. Note that in this paper, we *only* consider the attention maps and hidden states in the X-VLM encoders to compute OOD scores, e.g. MAP and Maha.

For all the VQA-based models, a total of 3129 answers are considered during the training and MSP computation.

Implementation Details of LangM, LangVAE, and I2Q. LangM, LangVAE, and I2Q models decode questions by predicting the probabilities of word tokens conditioned on different inputs. We build up these three models based on the proposed Encoder-Decoder transformer structure [17]. For LangVAE, inspired by [10], questions are tokenized by the tokenizer of Bert, and then sent to a transformer-based predictor to predict mean and variance vectors for each word token. Token-wised question features are sampled from a Gaussian function with predicted mean and variance vectors. A transformer-based question decoder takes the token-wised question features to reconstruct ques-

[1] Answer for the guessing game: VQA: A2, D1, G1, G4, B1, A4, H3, J1, K0, J2, I4; VizWiz: C4, D2, B0, E2, E0, N4, K1, K4, H2, N2, H1; VQA Abstract Scene: D4, B4, F0, G3, C1, F1, E4, M2, L0, I2, N0, N1; GQA: B2, F4, C0, E3, A3, G0, L2, L4, M1, L3, J0, L1; CLEVR: D3, C3, B3, D0, G2, F2, E1, I0, J4, M4, I3, M3, J3, H0; QRPE: A1, F3, C2, A0, H4, K3, M0, I1, K2, N3

Figure 1. Randomly sampled image-question pairs from the six datasets used in our benchmark – VQA, VizWiz, GQA, CLEVR, VQA_{ABS}, and QRPE. We encourage readers to try to identify where each sample came from and provide the answer key at the left-bottom of the first page. In our experience with this challenge, visual and linguistic clues are generally sufficient to separate the datasets.

	#	Method (Score)	Q	I	VIZWIZ	GQA	CLEVR	VQA_{ABS}	$\text{I}_{\text{In}}/\text{Q}_{\text{Out}}$	$\text{I}_{\text{Out}}/\text{Q}_{\text{In}}$	QRPE	Average
Density -based	1	LangM	✓		0.768	0.869	0.983	0.606	0.913	0.500	0.439	0.725
	2	I2Q	✓	✓	0.729	0.884	0.983	0.755	0.956	0.792	0.620	**0.817**
Reconst. -based	3	RIAD		✓	0.246	0.546	0.016	0.584	0.500	0.145	0.492	0.361
	4	LangVAE	✓		0.554	0.522	0.835	0.512	0.666	0.500	0.512	0.586
Prediction -based	5	BUTD (MSP)	✓	✓	0.775	0.512	0.700	0.608	0.580	0.529	0.698	0.629
	6	MCAN (MSP)	✓	✓	0.794	0.506	0.667	0.591	0.573	0.518	**0.739**	0.627
	7	X-VLM (MSP)	✓	✓	0.714	0.583	0.670	0.656	0.605	0.549	0.726	0.644
Feature -based	8	BUTD/MCAN (Maha-V)		✓	**0.974**	0.416	0.996	0.946	0.500	0.725	0.566	0.732
	9	X-VLM (Maha-V)		✓	0.967	0.442	0.988	**0.999**	0.500	0.732	0.592	0.746
	10	BUTD (Maha-L)	✓		0.653	0.641	0.784	0.464	0.710	0.500	0.496	0.607
	11	MCAN (Maha-L)	✓		0.628	0.660	0.729	0.506	0.690	0.500	0.540	0.602
	12	X-VLM (Maha-L)	✓		0.593	0.686	0.940	0.530	0.875	0.500	0.432	0.651
	13	BUTD (Maha-X)	✓	✓	0.824	0.394	0.412	0.365	0.638	0.468	0.700	0.543
	14	MCAN (Maha-X)	✓	✓	0.754	0.602	0.743	0.660	0.539	0.643	0.685	0.661
	15	X-VLM (Maha-X)	✓	✓	0.852	0.534	0.784	0.705	0.685	0.640	0.619	0.688
	16	Swin (Maha-V)		✓	0.933	0.488	0.997	0.983	0.500	0.756	0.561	0.745
	17	BERT (Maha-L)	✓		0.645	0.836	0.942	0.496	0.872	0.500	0.390	0.669
	18	Swin (MAP-V)		✓	0.323	0.623	0.178	0.452	0.500	0.396	0.493	0.424
	19	BERT (MAP-L)	✓		0.449	0.782	0.977	0.519	0.848	0.500	0.550	0.661
	20	X-VLM (MAP-V)		✓	0.849	0.332	0.985	0.495	0.500	0.671	0.542	0.625
	21	MCAN (MAP-L)	✓		0.809	0.497	0.544	0.475	0.541	0.500	0.415	0.552
	22	X-VLM (MAP-L)	✓		0.960	**0.916**	0.999	0.570	**0.999**	0.500	0.605	0.793
	23	BUTD (MAP-X)	✓	✓	0.465	0.542	0.681	0.600	0.521	0.583	0.518	0.559
	24	MCAN (MAP-X)	✓	✓	0.884	0.431	0.791	0.554	0.567	0.706	0.641	0.666
	25	X-VLM (MAP-X)	✓	✓	0.930	0.578	0.857	0.528	0.922	**0.816**	0.680	0.759
	26	MCAN (MAP-A)	✓	✓	0.861	0.479	0.614	0.495	0.580	0.560	0.463	0.579
	27	X-VLM (MAP-A)	✓	✓	0.953	0.880	**1.000**	0.562	0.998	0.630	0.652	0.811
	28	X-VLM* (MAP-A)	✓	✓	0.962	0.872	0.996	0.560	0.990	0.548	0.681	0.801

Table 1. AUCROC results of OOD detection on different OOD sets. BUTD/MCAN represents the object features share by BUTD and MCAN models. Single-modality results are grayed for off-modality OOD settings.

tions. For I2Q, grid embeddings of the image extracted by a pretrained ResNet101 [7] are encoded by a transformer encoder. Then the question decoder takes the encoded image features to decode the corresponding questions. Differently, in LangM, no prior information is provided. Thus the transformer encoder is not needed, and the question decoder takes zero vectors directly to decode the questions. Each transformer layer in the encoders and decoders has 8 heads and a dimension of 512. In this paper, we stack 4 layers for both the encoder and decoder. The models are trained with Adam [12] for 30 epochs with the learning rate of $5e-4$. LangM and I2Q are supervisedly trained with a Cross-Entropy loss, while LangVAE is optimized with ELBO-based VAE objective [13].

3. Computation of Feature-Based OOD Scores

In this paper, we compute the feature-based OOD scores based on the features captured from single and cross-modal modules. The single-modal modules, e.g. image and language-modal encoders, are defined as the modules that process the features from only one modality. The module will be treated as cross-modal only if it takes inputs from more than one modality.

Computation of BUTD-based Maha and MAP scores. Following the definition of MAP, a BUTD-based MAP score is represented as the maximum values of the single

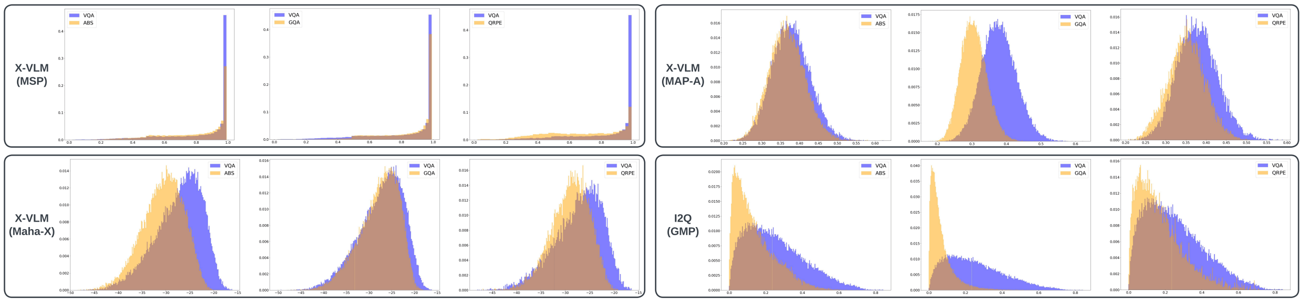

Figure 2. Histograms of four multimodal model-score combinations in VQA_{ABS}, GQA, and QRPE datasets.

softmax attention map A_0. Since the object features are not further processed by an image encoder before the cross attention, we average object features as the image representation for Maha-V for BUTD. Maha-X and Maha-L are computed with the multimodal and question representation.

Computation of MCAN-based Maha and MAP scores Considering that there is not a module to encode the object features alone in MCAN, like BUTD, we also take the average of the object features as the image representation. For question and cross-modal representations, we take the average of the hidden states outputted from the last layer of the text encoder and decoder respectively. In our MCAN model, there are 6 encoder and decoder blocks. Each encoder block contains 1 self-attention transformer layer, resulting in 48 attention maps for the question encoder. For the cross-modal decoder, each decoder block contains 1 cross-attention attention layer, providing also 48 attention maps. The MAP score for each modality is computed as the average of the maximum value of each softmax attention map.

Computation of X-VLM-based Maha and MAP scores. The question and cross-modality encoder contains 6 transformer layers with 12 heads in each, resulting in 72 attention maps for each modality. The image encoder contains 4 Swin Transformer Blocks. The number of heads of the transformer layers are 4, 8, 16, and, 32 separately in corresponding blocks. Due to Shifted Window mechanism, a total of 1984 attention maps are computed for a single image. We average the maximum softmax values of the maps as the MAP score of each modality. The MAP-A is computed as the mean of the MAP-V, MAP-L, and MAP-V. Similar to MCAN-based Maha, we capture the hidden states of the last transformer layers of each modality and perform a meanpool operation to compute representations for the score computation.

4. More Experiments on VQA OOD Methods

An unabridged accounting of our experimental results is shown in Table 1. This extends the results shown in the main paper by including Maha and MAP results for BUTD and MCAN. Interestingly, similar to Swin transformer, the Maha-V of object features of BUTD and MCAN (row `8`) performs well in the image modality, gaining the best performance in VizWiz. Compared to the variant of MAP-L, we find MCAN-based MAP-L has nearly random performance, especially in the GQA, CLEVR and $\mathrm{I_{In}}/\mathrm{Q_{Out}}$ (`21`), where X-VLM-based and Bert-based MAP-L still works well, suggesting that initializing with large-scale pretrained language model, e.g. Bert, can benefit the MAP-based OOD detection in Language modality. Checking the Maha and MAP score of cross-modality (row `13-15` vs `23-25`), we can find the MAP-X works better on the Average score than Maha-X with BUTD and X-VLM and achieve similar performance with MCAN, suggesting that the compared with the feature distance, the cross-modal matching could be a more reliable way to figure out the anomaly sources, especially in the case having both distinct image and questions, e.g. CLEVR.

Figure 2 shows histograms for selected multimodal scoring methods. From the figure, we can see that more than 40% of the MSP score of ID data points are concentrated at the high-score area ($f_{\text{MSP}} > 98\%$) and the scores of OOD data points also have the same trend. Comparing the histograms of Maha-X scores based on X-VLM, we can see that the method has less overlapping between the histograms of VQA and ABS, suggesting the features of cross-attention layers maintains more image information and have less ability to tell if there is a novel relationship of the image-question pair. However, checking the histograms of X-VLM (MAP-A) and I2Q (GMP), we can see that the question information is more influential in these 2 methods, resulting a more distinguishable ID and OOD scores.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In Proceedings of the IEEE

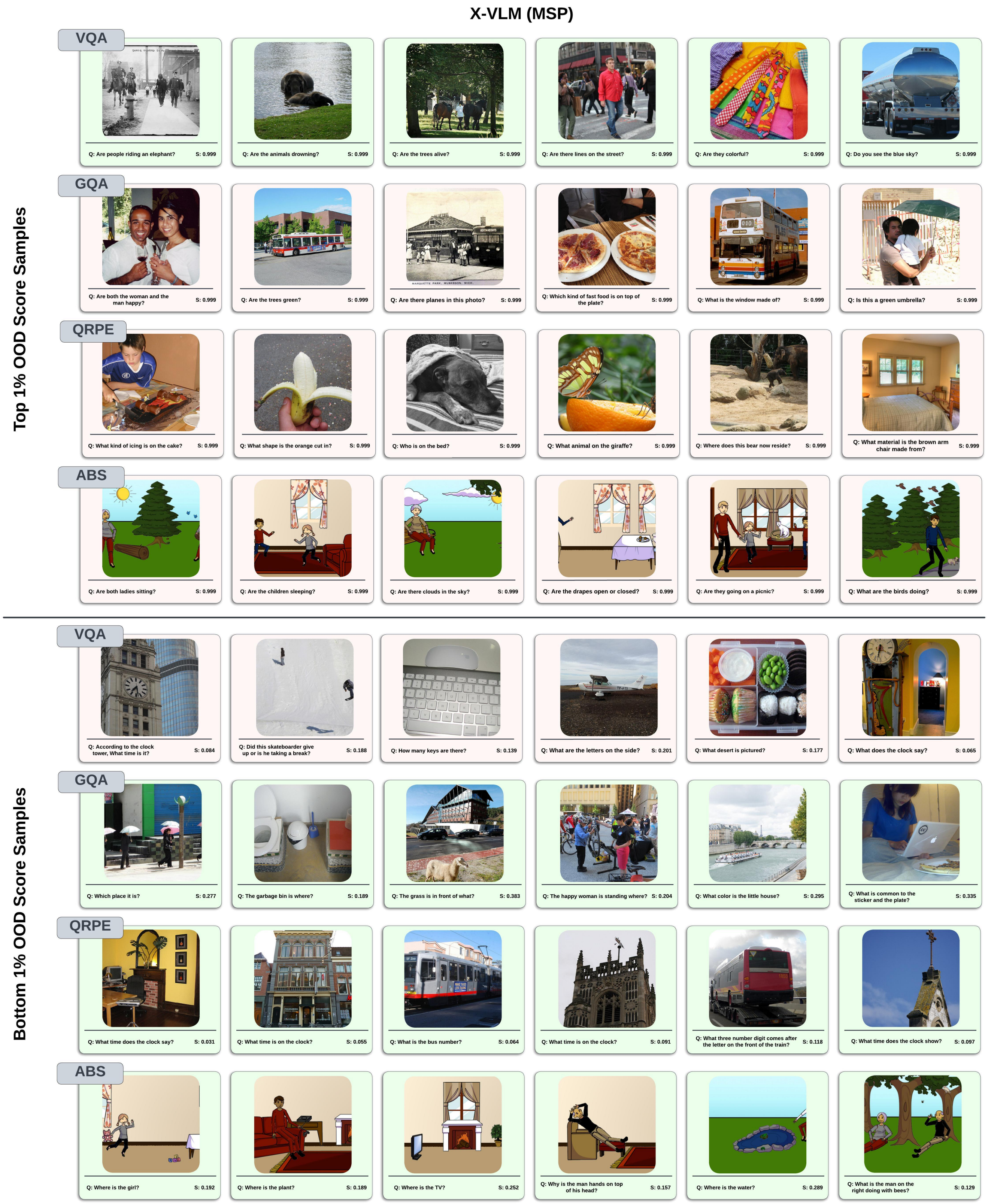

Figure 3. More visualization samples of VQA OOD detection with X-VLM MSP scores.

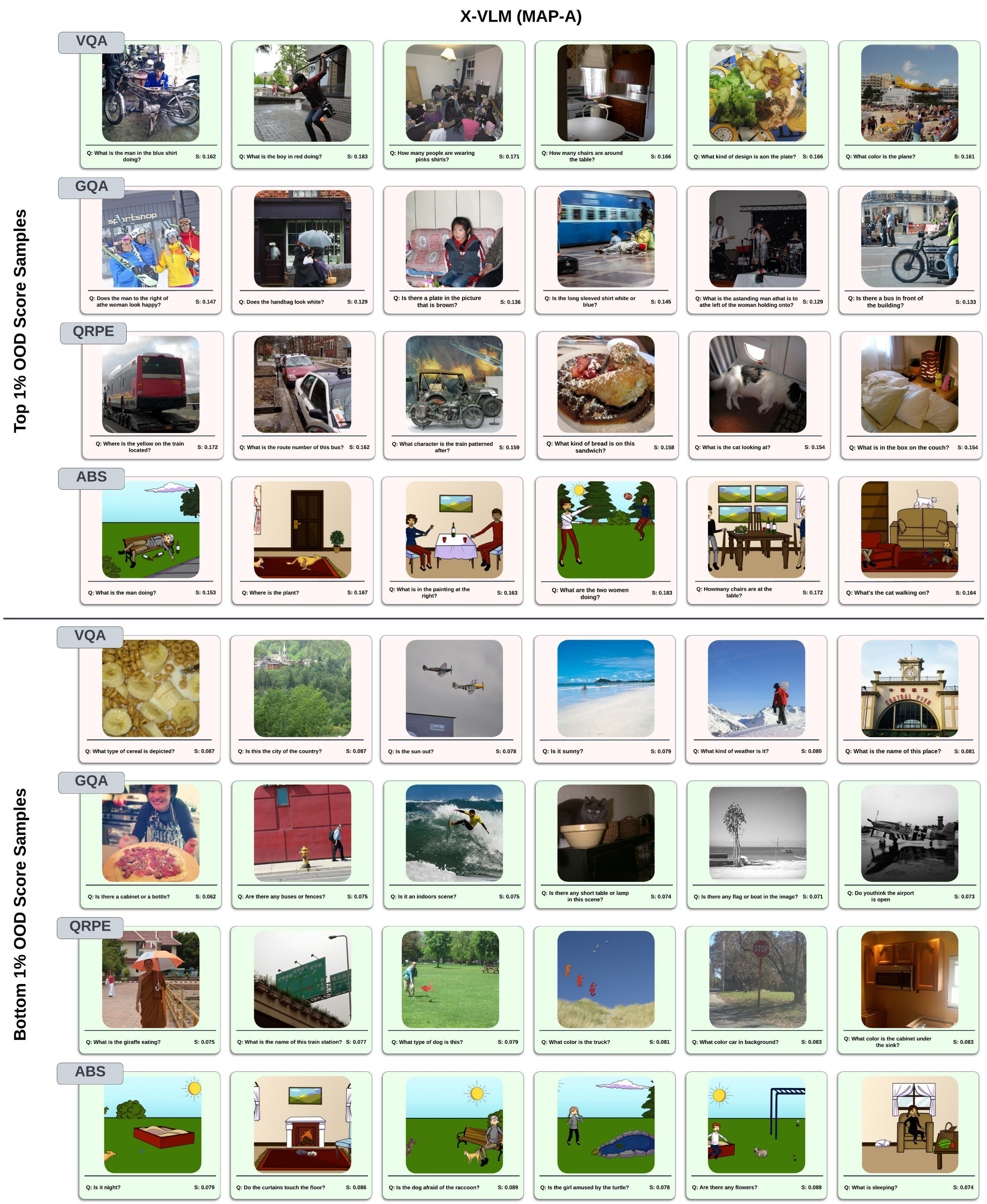

Figure 4. More visualization samples of VQA OOD detection with X-VLM MAP-A scores.

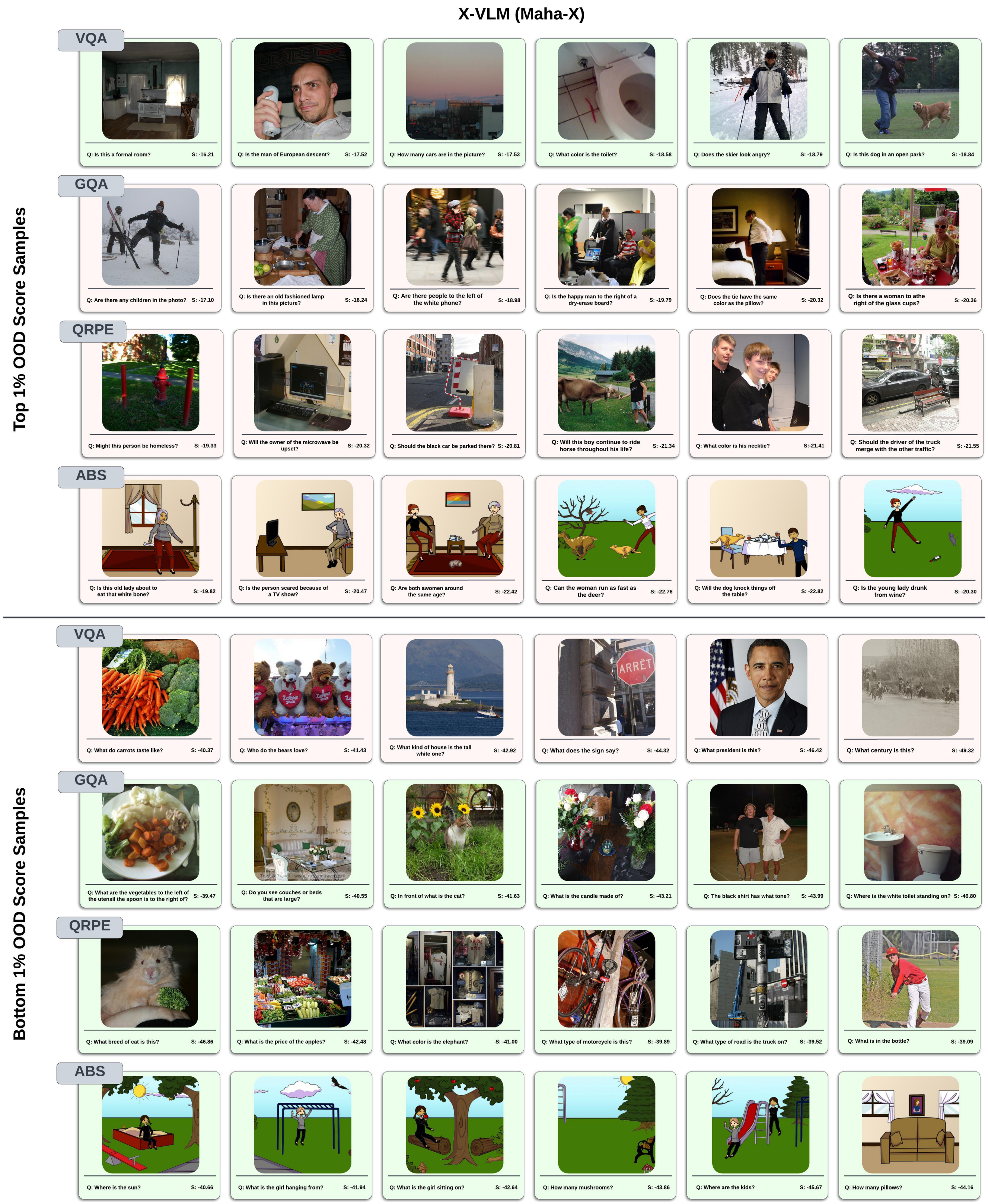

Figure 5. More visualization samples of VQA OOD detection with X-VLM Maha-X scores.

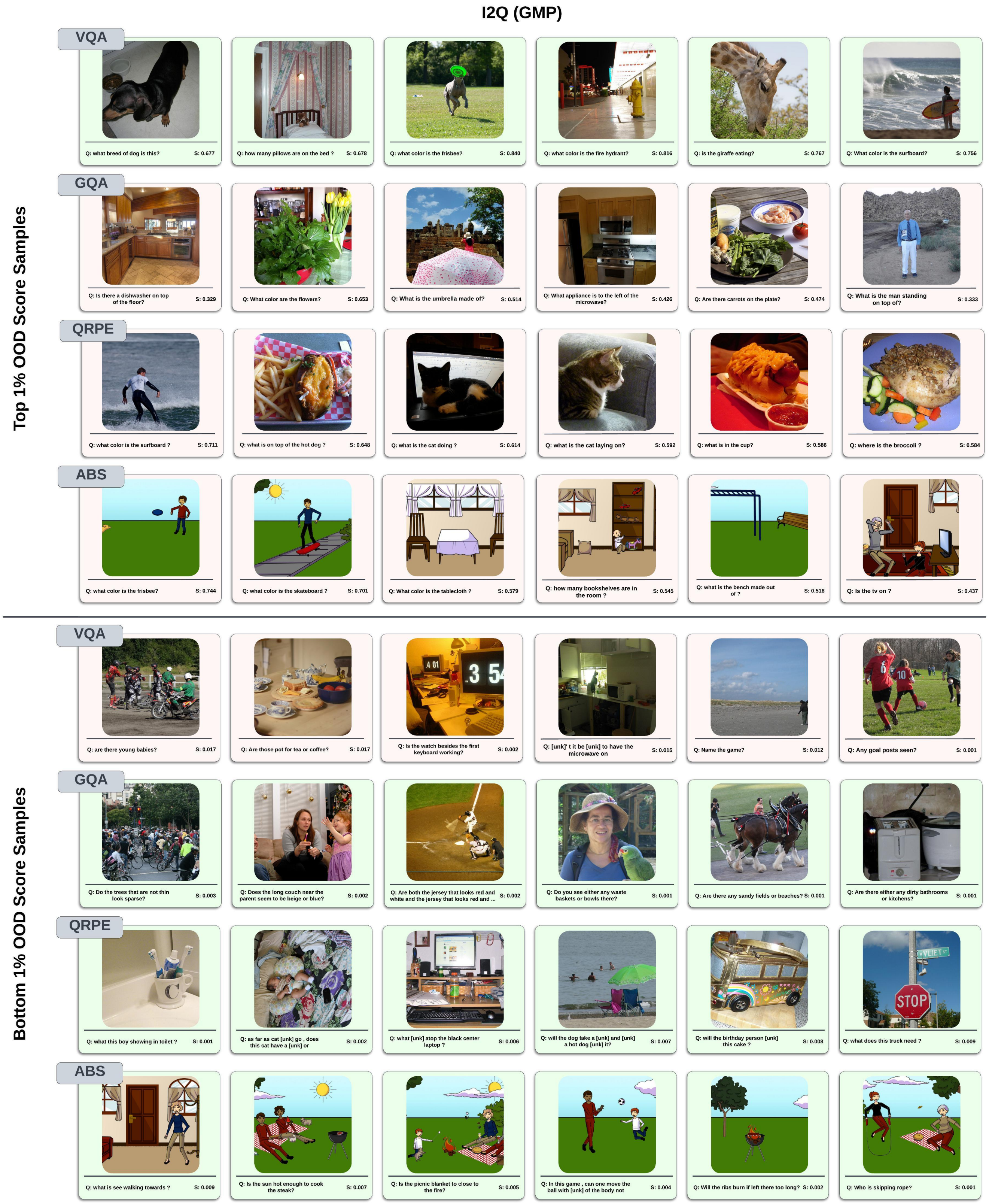

Figure 6. More visualization samples of VQA OOD detection with I2Q GMP scores.

conference on computer vision and pattern recognition, pages 6077–6086, 2018. 1

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In Proceedings of the IEEE international conference on computer vision, pages 2425–2433, 2015. 1

[3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition, pages 248–255. Ieee, 2009. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018. 1

[5] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA Matter: Elevating the role of image understanding in visual question answering. In CVPR, 2017. 1

[6] Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In CVPR, 2018. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016. 3

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. Neural computation, 9(8):1735–1780, 1997. 1

[9] Drew A Hudson and Christopher D Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In CVPR, 2019. 1

[10] Junyan Jiang, Gus G Xia, Dave B Carlton, Chris N Anderson, and Ryan H Miyakawa. Transformer vae: A hierarchical model for structure-aware and interpretable music representation learning. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 516–520. IEEE, 2020. 1

[11] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In CVPR, 2017. 1

[12] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. 1, 3

[13] Diederik P Kingma and Max Welling. An introduction to variational autoencoders. arXiv preprint arXiv:1906.02691, 2019. 3

[14] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017. 1

[15] Aroma Mahendru, Viraj Prabhu, Akrit Mohapatra, Dhruv Batra, and Stefan Lee. The promise of premise: Harnessing question premises in visual question answering. arXiv preprint arXiv:1705.00601, 2017. 1

[16] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, 2014. 1

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. Advances in neural information processing systems, 30, 2017. 1

[18] Zhou Yu, Yuhao Cui, Zhenwei Shao, Pengbing Gao, and Jun Yu. Openvqa. https://github.com/MILVLG/openvqa, 2019. 1

[19] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 6281–6290, 2019. 1

[20] Yan Zeng, Xinsong Zhang, and Hang Li. Multi-grained vision language pre-training: Aligning texts with visual concepts. arXiv preprint arXiv:2111.08276, 2021. 1

[21] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 5579–5588, 2021. 1