Towards a Unified Model for Generating Answers and Explanations in Visual Question Answering

Chenxi Whitehouse, Tillman Weyde, Pranava Madhyastha
City, University of London
{chenxi.whitehouse, t.e.weyde, pranava.madhyastha}@city.ac.uk

Abstract

The field of visual question answering (VQA) has recently seen a surge in research focused on providing explanations for predicted answers. However, current systems mostly rely on separate models to predict answers and generate explanations, leading to less grounded and frequently inconsistent results. To address this, we propose a multitask learning approach towards a **U**nified **M**odel for **A**nswer and **E**xplanation generation (UMAE). Our approach involves the addition of artificial prompt tokens to training data and fine-tuning a multimodal encoder-decoder model on a variety of VQA-related tasks. In our experiments, UMAE models surpass the prior state-of-the-art answer accuracy on A-OKVQA by 10∼15%, show competitive results on OK-VQA, achieve new state-of-the-art explanation scores on A-OKVQA and VCR, and demonstrate promising out-of-domain performance on VQA-X.[1]

1 Introduction

Contemporary models for visual question answering (VQA) and commonsense reasoning are typically trained discriminatively to select the best answers from Multiple-Choice questions or to classify single-word answers to a predetermined vocabulary (e.g. Anderson et al., 2018). Such settings often lead to limitations such as encouraging models to find superficial correlations (Ye and Kovashka, 2021) or penalising model performance even when the answers are plausible (e.g. synonyms and multi-word expressions, and morphological variations are not considered correct). Most current explanation generation models are trained independently of the QA model and the explanations are usually generated after the QA model has provided an answer. As a result, these explanation models lack access to the process that generated the answer and thus the grounding of the explanation is limited to the answer text.

We posit that a unified model that simultaneously performs answer prediction and explanation generation is a more effective and consistent approach for VQA. Generative models, such as GPT-3 (Brown et al., 2020), T5 (Raffel et al., 2020), or OFA (Wang et al., 2022a), have been shown to be successful at rapidly adapting to downstream tasks and generating high-quality open-ended text, and hence are suitable candidates for this unified approach.

We propose a multitask learning approach for multimodal transformer-based encoder-decoder models, towards a United Model for Answer and Explanation generation (UMAE). In addition to the current trend of separate answer prediction and explanation generation based on the answers, our approach adds the capability of jointly generating answers and explanations together. Inspired by the success of artificial prompt tokens in Neural Machine Translation (NMT) (Johnson et al., 2017), we extend and demonstrate the efficacy of the artificial prompt-based method for VQA in a multitask setup. We augment training instances with artificial prompt tokens, enabling the model to distinguish different tasks while learning shared semantic features. Experiments on a combination of three knowledge-intensive VQA datasets, OK-VQA (Marino et al., 2019), A-OKVQA (Schwenk et al., 2022), and VCR (Zellers et al., 2019), show that the UMAE models achieve a new state-of-the-art (SOTA) answer accuracy on A-OKVQA, new SOTA explanation score on VCR, and competitive out-of-domain performance on VQA-X (Park et al., 2018). UMAE supports the generation of the answer to a question, the explanation for a given question and answer, and both together jointly, making the model efficient and flexible. An illustration of the training setup is shown in Figure 1.

In summary, our main contributions are as follows: (1) the UMAE framework where answers and

[1] Code is available at: https://github.com/chenxwh/UMAE.

Findings of the Association for Computational Linguistics: EACL 2023, pages 1693–1705
May 2-6, 2023 ©2023 Association for Computational Linguistics

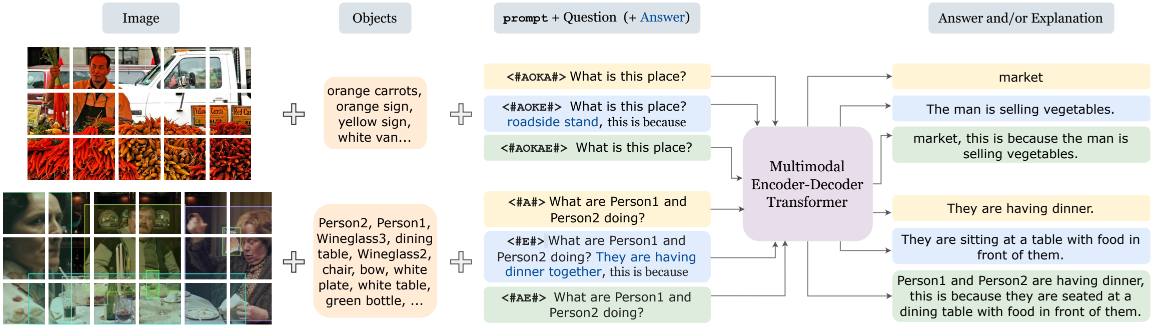


Image
Objects
prompt + Question (+ Answer)
Answer and/or Explanation
orange carrots, orange sign, yellow sign, white van…
<#AOKA#> What is this place?
<#AOKE#> What is this place? roadside stand, this is because
<#AOKAE#> What is this place?
Multimodal Encoder-Decoder Transformer
market
The man is selling vegetables.
market, this is because the man is selling vegetables.
Person2, Person1, Wineglass3, dining table, Wineglass2, chair, bow, white plate, white table, green bottle, …
<#A#> What are Person1 and Person2 doing?
<#E#> What are Person1 and Person2 doing? They are having dinner together, this is because
<#AE#> What are Person1 and Person2 doing?
They are having dinner.
They are sitting at a table with food in front of them.
Person1 and Person2 are having dinner, this is because they are seated at a dining table with food in front of them.

Figure 1: Illustration of UMAE: we train a multimodal encoder-decoder model on the mix of VQA tasks for jointly optimising answer and explanation, where we distinguish the training instances and target output with artificial prompt tokens (e.g. <#AOKA#>). The top and bottom examples are from A-OKVQA and VCR, respectively.

explanations can be generated by a single unified model (§3.1); (2) a simple and efficient training approach that uses multitask learning with artificial prompts and demonstrates its ability to generalise across domains (§4); (3) a method to map generated answers to Multiple-Choice options via evaluating the perplexity of the generation (§3.2); (4) new SOTA results by UMAE, particularly for explanation generation and promising out-of-domain performance (§5).

2 Related Work

Multimodal Transformer-based Models achieve SOTA performance on various vision-language tasks (Chen et al., 2020; Li et al., 2020; Cho et al., 2021; Wang et al., 2022c; Zhang et al., 2021). They showcase the possibility of capturing richer multimodal semantic coherence than discriminatively trained models and are further capable of generating self-explanations. Pretrained on multitask settings with natural language instructions, e.g. *"what does the region describe?"*, models like OFA (Wang et al., 2022a) are claimed to have the capability to transfer to unseen tasks and domains via similar instructions. However, contrary to these claims, we observe that pretrained OFA is incapable of generating valid explanations through simple natural language instructions (§5).

Artificial Prompt Tokens have previously been explored for NMT by Johnson et al. (2017); Mitzalis et al. (2021). They propose a single model with the traditional NMT model architecture (usually for one language pair) and jointly train on different language pairs with added artificial prompts, e.g. `2es` to distinguish the target language. This approach has been found to foster implicit cross-lingual bridging and exhibit zero-shot translation capability. In this paper, we exploit a similar approach with artificial prompts for answer and explanation generation in VQA with a united model. This enables the model to learn shared features among tasks and datasets in various domains.

Explanation Generation for VQA has gained growing interest in research. However, most recent approaches use separate models to predict answers and generate explanations (Dua et al., 2021). Wu and Mooney (2019) generate explanations with an object detector and a GRU unit for text embedding, then train on a subset of VQA-X in which the explanations contain the objects most attended to by the model. Kayser et al. (2021) develop an e-UG model combining UNITER (Chen et al., 2020) for processing multimodal input and GPT-2 (Radford et al., 2019) for generation. In contrast, in this paper, we propose using a single united model for more grounded answer and explanation generation.

3 Methodology

3.1 Multitask Learning with Artificial Prompt

We formulate three generation settings: Q→A: answer prediction; QA→E: explanation generation conditioned on the answer; and Q→AE: *joint* answer and explanation generation for a given question. We hypothesise that by training the model to generate both the answer and its explanation *simultaneously*, the result answer and explanation will be more grounded and consistent.

We use a pretrained multimodal encoder-decoder transformer as our base model (here we build on the openly released version of OFA as a strong baseline), and finetune the model on a mix of VQA datasets from different domains.

Different from OFA, for each image in the VQA datasets, we first extract objects and attributes us-

MODEL	OK-VQA	A-OKVQA					VCR	
	direct answer	*multiple choice*			*direct answer*		*multiple choice*	BERTSCORE
	TEST	VAL (*ppl*)	VAL (*GloVe*)	TEST	VAL	TEST	VAL (*ppl*)	VAL
OFA*	40.40	24.54	56.19	47.40	48.09	39.77	33.55	64.55
OFA$_{\text{Q->A}}$	49.93	74.32	65.30	61.71	63.00	53.91	54.89	83.85
UMAE$_{\text{ALL}}$	**51.77**	**74.59**	**65.67**	**63.26**	**63.29**	**56.14**	**56.66**	**85.97**
PRIOR-BEST	54.41	–	60.30	53.70	48.60	40.70	(77.10)†	–

Table 1: Performance of models for answer generation. Better results are in bold. OFA* refers to the pretrained OFA. Prior-best results for the three datasets are from Gui et al. (2022), Schwenk et al. (2022), Wang et al. (2022b), respectively. † is from a discriminative model and thus not comparable (see Ye and Kovashka, 2021).

DATASET	MODEL	e-ViL SCORES			N-GRAM SCORES					LEARNT SCORE
		S_O	S_T	S_E	BLEU4	ROUGE-L	METEOR	CIDEr	SPICE	BERTSCORE
A-OKVQA	OFA*	4.44	56.19	7.90	0.30	4.45	3.26	4.82	4.62	68.64
	OFA$_{\text{Q->A}}$+OFA$_{\text{QA->E}}$	35.82	74.32	48.29	22.18	48.51	23.56	86.76	22.46	85.96
	UMAE$_{\text{A-OKVQA}}$	37.10	73.97	50.15	**27.61**	52.23	24.06	**104.39**	22.88	87.86
	UMAE$_{\text{ALL}}$	**37.91**	**74.59**	**50.82**	27.35	**52.56**	**24.83**	101.09	**23.33**	**88.21**
VCR	e-UG	19.30	**69.80**	27.60	4.30	22.50	11.80	32.70	12.60	79.00
	UMAE$_{\text{VCR}}$	22.57	**56.68**	39.82	12.25	28.87	16.67	**48.14**	27.36	81.77
	UMAE$_{\text{ALL}}$	**22.82**	56.66	**40.27**	**13.44**	**29.53**	**17.54**	47.33	**26.45**	**81.91**
VQA-X	e-UG	36.50	80.50	45.40	23.20	45.70	22.10	74.10	20.10	87.00
	UMAE$_{\text{ALL}}$	31.58	77.65	40.67	14.63	35.12	20.29	50.35	19.13	85.40

Table 2: Explanation Scores. OFA* is the pretrained OFA, showing the transferability of OFA for generating explanations with natural language instructions. Results with e-UG are from Kayser et al. (2021). We show the best results of A-OKVQA and VCR in bold. The last row in blue shade shows *out-of-domain* performance.

ing a bottom-up top-down attention-based model, which is crucial for open-domain VQA tasks (Anderson et al., 2018). We then add artificial prompt tokens at the beginning of the textual input to signal the generation task (answer, explanation, or both) and the dataset[2]. For `Q→AE`, we concatenate answers and explanations with a separator in between. Finally, we mix all training instances, each consisting of an image (processed in patches), objects and attributes, and textual input with artificial prompts.

3.2 Perplexity as Multiple Choice Metric

To map the generated output to Multiple-Choice options, in previous work the predictions are loosely matched with options or gold answers using embedding-based methods, such as GloVe embedding similarity (Schwenk et al., 2022). In contrast to these approaches, we propose to evaluate each option as a *text generation* task, by feeding the model the information that was used to generate the answer as a prompt, and calculating the likelihood of each option being generated. Formally, given an option $Y = (y_1, y_2, ..., y_t)$ with t tokens, we calculate the probability of each token y_i being generated by feeding the image, objects, and question, as well as the first $i-1$ tokens from Y to the model p_θ. The perplexity is then calculated with: $\mathrm{PPL}(Y) = \exp\left\{-\frac{1}{t}\sum_i^t \log p_\theta\left(y_i|y_{<i}\right)\right\}$, which reflects the probability of option Y being generated by the model. Finally, the option with the lowest perplexity is chosen as the answer.

We also compare the performance of our approach, using perplexity as the metric, with GloVe embedding similarity for A-OKVQA (see Table 1).

4 Experimental Setup

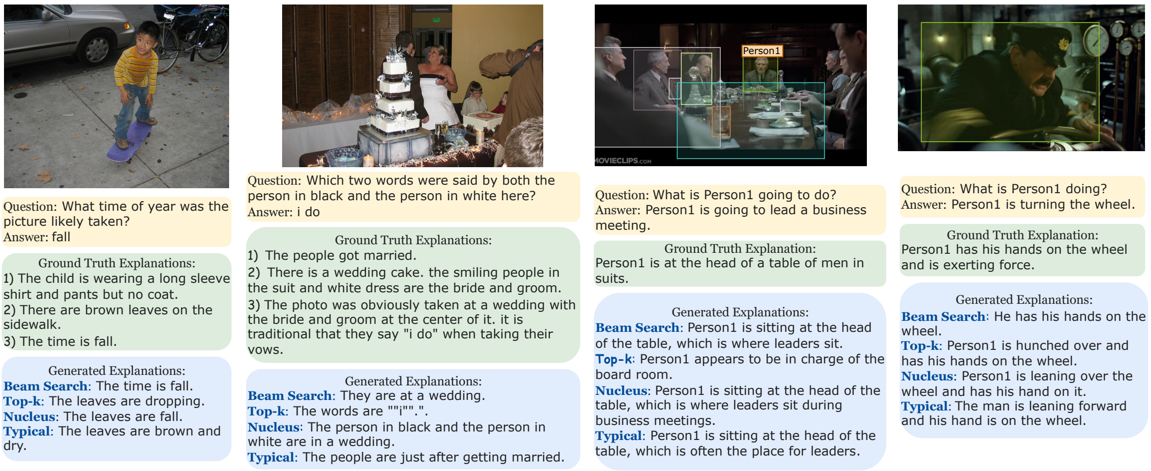
We primarily evaluated our proposed UMAE approach using pretrained OFA[3] as the base model on three knowledge-intensive VQA datasets: OK-VQA, A-OKVQA and VCR[4]. We split the original train set into train and validation set (95%-5%) for all three datasets. Since the test set is not publicly available for A-OKVQA and VCR, we use the original validation set for experimental analyses. We prepare training instances[5] as introduced

[2] Artificial prompt tokens are added as special tokens to the tokenizer to avoid bias in the pretrained embeddings. However, we note that these tokens may be biased w.r.t their association with specific tasks after training, which is an intended effect.

[3] `https://github.com/OFA-Sys/OFA`

[4] See Appendix A for datasets details.

[5] Specifically, we add `<#OKA#>` for OK-VQA (only answers are available), `<#A#>`, `<#E#>`, `<#AE#>` for VCR, and `<#AOKA#>`, `<#AOKE#>`, `<#AOKAE#>` for A-OKVQA.



Question: What time of year was the picture likely taken?
Answer: fall

Ground Truth Explanations:
1) The child is wearing a long sleeve shirt and pants but no coat.
2) There are brown leaves on the sidewalk.
3) The time is fall.

Generated Explanations:
Beam Search: The time is fall.
Top-k: The leaves are dropping.
Nucleus: The leaves are fall.
Typical: The leaves are brown and dry.

Question: Which two words were said by both the person in black and the person in white here?
Answer: i do

Ground Truth Explanations:
1) The people got married.
2) There is a wedding cake. the smiling people in the suit and white dress are the bride and groom.
3) The photo was obviously taken at a wedding with the bride and groom at the center of it. it is traditional that they say "i do" when taking their vows.

Generated Explanations:
Beam Search: They are at a wedding.
Top-k: The words are ""i"".".
Nucleus: The person in black and the person in white are in a wedding.
Typical: The people are just after getting married.

Question: What is Person1 going to do?
Answer: Person1 is going to lead a business meeting.

Ground Truth Explanation:
Person1 is at the head of a table of men in suits.

Generated Explanations:
Beam Search: Person1 is sitting at the head of the table, which is where leaders sit.
Top-k: Person1 appears to be in charge of the board room.
Nucleus: Person1 is sitting at the head of the table, which is where leaders sit during business meetings.
Typical: Person1 is sitting at the head of the table, which is often the place for leaders.

Question: What is Person1 doing?
Answer: Person1 is turning the wheel.

Ground Truth Explanation:
Person1 has his hands on the wheel and is exerting force.

Generated Explanations:
Beam Search: He has his hands on the wheel.
Top-k: Person1 is hunched over and has his hands on the wheel.
Nucleus: Person1 is leaning over the wheel and has his hand on it.
Typical: The man is leaning forward and his hand is on the wheel.

Figure 2: Examples of generated explanations from UMAE_{ALL} model with different decoding strategies. The two examples on the left are from A-OKVQA and the two on the right are from VCR.

in §3.1. Additionally, for VCR, we draw coloured highlights around the referenced entity on the images, following Zellers et al. (2021) (Appendix A). To account for the imbalance in size among the datasets, we up-sample instances in OK-VQA and A-OKVQA, and shuffle all instances to train the UMAE_{ALL} model.

For ablation studies, we finetune OFA for separate answer prediction ($\text{OFA}_{\text{Q->A}}$) and explanation generation conditioned on answers ($\text{OFA}_{\text{QA->E}}$). To better understand the impact of mixing datasets from different domains, we also train $\text{UMAE}_{\text{A-OKVQA}}$ and UMAE_{VCR}, focusing on all three answer and explanation generation tasks but only using data from a single dataset: either with A-OKVQA or with VCR. Details of training parameters are included in Appendix B.

We use beam search for generating answers and additionally experiment with different decoding methods including top-k sampling, Nucleus sampling (Holtzman et al., 2020), and Typical sampling (Meister et al., 2022), for generating explanations. We evaluate answer accuracy as well as explanation quality with automatic NLG metrics and e-ViL scores (Kayser et al., 2021). e-ViL scores consist of S_{T} (task/answer accuracy), S_{E} (explanation score), and overall S_{O} (product of S_{T} and S_{E}), where S_{E} is the harmonic mean of NGRAMScore (the harmonic mean of n-gram scores ROUGE-L (Lin and Och, 2004), METEOR (Banerjee and Lavie, 2005), CIDEr (Vedantam et al., 2015), and SPICE (Anderson et al., 2016)) and additionally the BERTScore (Zhang et al., 2020), a learned similarity metric over contextual representations of sentences.

5 Results and Discussion

5.1 Answer Accuracy

Table 1 presents our observations for answer accuracy on Q->A task over the three datasets. We also evaluate VCR answers using BERTScore as the answers for VCR are usually sentences. We observe that UMAE_{ALL} outperforms $\text{OFA}_{\text{Q->A}}$ on all datasets, improves the prior SOTA on A-OKVQA by 10∼15%, and achieves competitive results on OK-VQA. For models that are finetuned on A-OKVQA, we also see a salient improvement (+9%) with the proposed mapping of options by perplexity in Multiple-Choice, instead of GloVe embeddings similarity[6]. We conducted several ablation studies on the dependency of the modality for the answer accuracy in A-OKVQA, where we find the visual encoder is crucial for performance. Details are included in Appendix C.

5.2 Explanation Evaluation

Table 2 shows e-ViL sores (§4) for explanations using automatic NLG metrics[7]. Following the same setup as in Kayser et al. (2021), an explanation is evaluated only if the answer predicted by the system is correct[8]. We observe that pretrained OFA with natural language prompts, e.g. *"what is the explanation for the answer?"* or *"this is*

[6] Preliminary experiments with NLG metrics (BERTScore and BLEU) for selecting the options given generation were suboptimal.

[7] Nucleus sampling shows best results and is reported. Detailed scores with different decoding methods are shown in Appendix D.

[8] A limitation of evaluating all explanations is that explanations of wrong answers may get high scores with n-gram metrics, even though they are justifying wrong answers and should be penalised.

MODEL	S_{E}	BLEU4	R-L	MET.	CIDEr	SPICE	BERTSc.
OFA$_{\text{Q->A}}$+OFA$_{\text{QA->E}}$	42.4	20.0	44.2	19.3	66.7	19.1	85.1
UMAE$_{\text{A-OKVQA}}$	45.8	23.6	47.9	21.7	78.0	20.5	86.9
UMAE$_{\text{ALL}}$	**46.8**	**24.9**	**49.5**	**22.3**	**84.1**	**20.8**	**87.3**

Table 3: Explanation scores on the same subset of A-OKVQA.

because" performs poorly, as most generated explanations are words (*"yes/no"*) or short-phrases[9]. We compare UMAE models (on all and individual datasets) with prior best results from e-UG (see §2), and standard separated trained baselines (OFA$_{\text{Q->A}}$+OFA$_{\text{QA->E}}$). UMAE$_{\text{ALL}}$ achieves better results across all datasets, showing the advantage of mixing tasks and datasets in different domains. For out-of-domain evaluation on VQA-X, UMAE$_{\text{ALL}}$ also shows mostly competitive results. Examples of explanation generation are shown in Figure 2 and Appendix E.

Since e-ViL only evaluates an explanation if a model generates the correct answer, the subset of explanations evaluated varies by model. To *fairly* compare explanations on the same subset, we propose only using the subset of samples where all models provide correct answers for explanation prediction. Table 3 shows the results on A-OKVQA with such a subset of 770 candidates, where UMAE$_{\text{ALL}}$ shows an even higher explanation score. This highlights that UMAE$_{\text{ALL}}$ generates explanations that overlap significantly better with gold explanations.

In summary, our experiments demonstrate that the UMAE model leads to improved answer and explanation generation and allows for the flexibility to generate different types of outputs, including answers, explanations, or both. We observe that UMAE exhibits promising results in jointly generating both the answer and explanation. We further provide a comparative evaluation in Appendix F as a first step towards comparison as there is currently no standard evaluation setup for the joint answer and explanation evaluation.

5.3 Error Analysis

To better understand the generated answers and errors, we randomly sample 50 errors in OK-VQA and A-OKVQA. Our analysis reveals the following main error types, where the first three are related to model performance: (1) *Knowledge*: the implicit knowledge learned by the model is insufficient for answering some of the knowledge-intensive questions, such as questions asking *when* a certain sport was invented; (2) *Visual*: the model fails to identify the visual attributes correctly, such as questions about *recognising object shape or material*; (3) *Semantic disassociation*: the model misinterprets questions or fails to match the intended semantic meaning. For example, it may answer what *an object is* instead of a more complex question such as *what is commonly packed in it* (e.g. answering "suitcase" instead of "clothes"); (4) *Metric*: the evaluation metric may penalise some of the plausible answers, especially when searching for exact match answers (mostly due to the difference of singular/plural or phrases with/without space in between); and (5) *Dataset*: errors due to issues in the datasets themselves. We discuss prominent issues in dataset quality briefly in Appendix G and further present the distribution of error types in Figure 3.

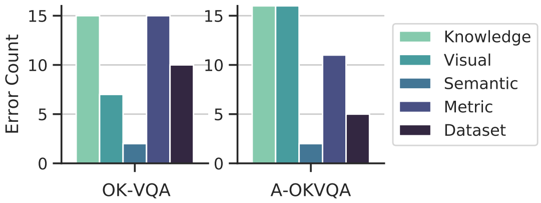


Figure 3: Error type distribution in 100 random samples from A-OKVQA and OK-VQA.

6 Conclusions

In this work, we propose UMAE, a unified model that generates answers and explanations in VQA using a multitask learning approach for multimodal encoder-decoder models, where artificial prompt tokens are added to distinguish different tasks while learning shared semantics. Evaluation of our approach on various VQA tasks shows that UMAE outperforms prior best models and separately trained baselines in both answer and explanation scores, where we also demonstrate the benefit of using perplexity as the metric for mapping generated answers to Multiple-Choice options. Additionally, UMAE offers flexibility in output and can generate explanations for datasets without explanations for training, e.g. OK-VQA, while also improving answer quality. Through case studies and error analysis, we identify potential areas for future improvement, including dataset quality.

[9] BERTScore in not representative of the validity of outputs from OFA*. We refer the reader to an exposition of the problems associated with NLG metrics in Caglayan et al. (2020).

Limitations

We discuss the limitations of our work in the following two aspects. Firstly, the experiments with our proposed framework and finetuning approach are primarily on the OFA model. We believe our approach applies to any multimodal generative model, however, it would also provide insights to experiment with more models. Secondly, regarding the evaluation of our proposed joint framework, to better evaluate the generated explanation quality, especially to evaluate the difference between explanations generated jointly with answers and generated conditioned on the answers, human judgement would be an important criterion compared to automatic NLG metrics.

Acknowledgements

We acknowledge the support of Apoorv Khandelwal from AI2 for providing us with results for the evaluation of our model predictions over a hidden test set. This was valuable for our earlier draft of the paper. We would like to thank the anonymous reviewers who provided valuable feedback on the previous draft of our paper.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic Propositional Image Caption Evaluation. In *European conference on computer vision*, pages 382–398. Springer.

Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. 2018. Bottom-up and Top-down Attention for Image Captioning and Cisual Question Answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086.

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

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-shot Learners. *Advances in neural information processing systems*, 33:1877–1901.

Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2020. Curious case of language generation evaluation metrics: A cautionary tale. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2322–2328, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal Image-Text Representation Learning. In *European conference on computer vision*, pages 104–120. Springer.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying Vision-and-Language Tasks via Text Generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR.

Radhika Dua, Sai Srinivas Kancheti, and Vineeth N Balasubramanian. 2021. Beyond vqa: Generating Multi-Word Answers and Rationales to Visual Questions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1623–1632.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, Seattle, United States. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The Curious Case of Neural Text Degeneration. In *International Conference on Learning Representations*.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Maxime Kayser, Oana-Maria Camburu, Leonard Salewski, Cornelius Emde, Virginie Do, Zeynep Akata, and Thomas Lukasiewicz. 2021. e-ViL: A Dataset and Benchmark for Natural Language Explanations in Vision-Language Tasks. In *Proceedings* *of the IEEE/CVF International Conference on Computer Vision*, pages 1244–1254.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In *European Conference on Computer Vision*, pages 121–137. Springer.

Chin-Yew Lin and Franz Josef Och. 2004. Automatic evaluation of machine translation quality using longest common subsequence and skip-bigram statistics. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, pages 605–612, Barcelona, Spain.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft Coco: Common Objects in Context. In *European conference on computer vision*, pages 740–755. Springer.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. 2022. Typical Decoding for Natural Language Generation. *arXiv preprint arXiv:2202.00666*.

Faidon Mitzalis, Ozan Caglayan, Pranava Madhyastha, and Lucia Specia. 2021. BERTGen: Multi-task generation through BERT. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6440–6455, Online. Association for Computational Linguistics.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie Description. *International Journal of Computer Vision*, 123(1):94–120.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A Benchmark for Visual Question Answering using World Knowledge. *ArXiv*, abs/2206.01718.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based Image Description Evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022a. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR.

Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. 2022b. VQA-GNN: Reasoning with Multimodal Semantic Graph for Visual Question Answering. *arXiv preprint arXiv:2205.11501*.

Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2022c. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*.

Jialin Wu and Raymond Mooney. 2019. Faithful multimodal explanation for visual question answering. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 103–112, Florence, Italy. Association for Computational Linguistics.

Keren Ye and Adriana Kovashka. 2021. A Case study of the Shortcut Effects in Visual Commonsense Reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 3181–3189.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From Recognition to Cognition: Visual Commonsense Reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition*.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. MERLOT: Multimodal Neural Script Knowledge Models. In *Advances in Neural Information Processing Systems*, volume 34, pages 23634–23651. Curran Associates, Inc.

Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting Visual Representations in Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5579–5588.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

A Datasets

The datasets used in the paper are as follows:

OK-VQA (Marino et al., 2019) is a knowledge-based VQA dataset that requires outside knowledge beyond the images to answer the questions. It has train and test splits of size 9,009 and 5,046. Each question is provided answers by five annotators. To use the VQA (Antol et al., 2015) metric, each annotated answer is then repeated twice to form a gold answer set with 10 answers. Since no explanation is provided, we only train `Q→A` task on OK-VQA.

A-OKVQA (Schwenk et al., 2022) is currently the largest knowledge-based VQA dataset split into 17.1K, 1.1K, and 6.7K for train, validation, and test, respectively. The questions cover four knowledge types: visual, commonsense, knowledge bases, and physical. For each question, it provides both multiple-choice answers and 10 free-form answers (annotated by 10 different people), as well as three explanations. Images in both OK-VQA and A-OKVQA are from MSCOCO (Lin et al., 2014), and answers in both datasets are in single words or short phrases.

VCR (Zellers et al., 2019) is a large multiple-choice dataset for Visual Commonsense Reasoning. The train, validation, and test splits have 191.6k, 21.3k, and 26.5k instances, respectively. Each question has four answer options in sentences, and the correct answer is further provided with four explanation options. Images in VCR are from movie clips (Rohrbach et al., 2017). Bounding boxes of entities are provided associated with mentions such as `Person1` in questions, answers and explanations. We follow Zellers et al. (2021) and draw coloured highlights around the referenced entity on the images, where entity names and the coloured highlights are consistent in the entire dataset, expecting the model to learn the association between the coloured bounding box and the entity.

VQA-X (Park et al., 2018) contains a subset from the VQAv2 (Goyal et al., 2017) dataset and further provides three explanations for each question. The image-question pairs are split into train, validation, and test with 29.5k, 1.5k, and 2k instances, respectively. We only use the original test set to evaluate the zero-shot performance of the trained models.

Question	Objects	Images	Accuracy
✓	✓	original	50.39
✓	✗	✗	39.16
✓	✗	random	33.48
✓	✓	✗	33.28

Table 4: Ablation on the modality dependency for answer accuracy of A-OKVQA.

B Hyper-Parameters and Training

We begin with the pretrained weights from the original OFA-large[10], which is trained on vision-only tasks including Image Classification, language-only tasks including Sentence Classification, Text Summarisation, as well as various vision-language tasks including Image Captioning, Visual Question Answering and Visual Entailment. Adam is used as the optimizer and cross-entropy is the loss function. We set the learning rate to 10^{-5}, the warm-up ratio to 0.4, and the patch image size to 480. We shuffle all the training examples and use batch size 16. Due to the large size of VCR, we train for 30 epochs on models involving VCR ($\text{OFA}_{\text{Q->A}}$ for VCR, UMAE_{VCR} and UMAE_{ALL}), and up to 100 epochs for other models. We report the empirical performance with checkpoints that perform best on the validation set (the 5% split from the original train set). For A-OKVQA, we additionally report the answer accuracy on the original test set.

C Ablations on Modality Dependency

We conduct several ablation studies to investigate the dependency of object features and images on the performance of our model UMAE_{ALL} for answer accuracy of A-OKVQA, where we removed images, replaced them with random images, and removed extracted attributes and features. Results in Table 4 show that the visual encoder is crucial for performance and that visual objects alone are not sufficient for answer prediction. Using a random image would introduce noise and therefore performs worse than not including the image at all. We did not test removing the question because we believe the model needs the questions to be able to provide answers.

[10] `https://github.com/OFA-Sys/OFA`

Dataset	Decoding	e-ViL	N-gram Scores								Learnt Sc.
		S_E	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	METEOR	CIDEr	SPICE	BERTScore
A-OKVQA	beamsearch	44.71	52.01	36.69	26.72	19.88	40.39	22.06	68.48	20.94	86.05
	top-k ($k = 100$)	44.34	52.56	37.06	27.06	19.72	44.45	21.58	73.44	19.38	86.27
	nucleus ($p = 0.4$)	**50.82**	58.92	44.66	35.06	27.35	52.56	24.83	101.09	23.33	88.21
	typical ($p = 0.6$)	47.27	54.18	39.39	29.82	22.18	47.78	22.79	84.43	21.47	86.95
VCR	beamsearch	40.23	26.41	20.15	15.95	12.47	29.13	16.82	49.72	27.70	81.84
	top-k ($k = 50$)	33.19	20.98	14.89	11.18	8.33	23.65	13.72	32.73	21.99	80.31
	nucleus ($p = 0.1$)	**40.27**	31.42	22.95	17.62	13.44	29.53	17.54	47.33	26.45	81.91
	typical ($p = 0.4$)	35.12	23.42	16.88	12.83	9.64	25.36	14.70	35.85	23.32	80.70
VQA-X	beamsearch	35.88	37.84	24.91	16.67	10.97	31.32	17.90	38.23	16.23	84.39
	top-k ($k = 50$)	33.28	38.35	23.11	14.21	8.45	29.15	17.05	32.89	15.26	83.41
	nucleus ($p = 0.1$)	**40.67**	47.56	31.44	21.47	14.63	35.12	20.29	50.35	19.13	85.40
	typical ($p = 0.5$)	36.31	40.85	25.57	16.82	11.14	31.08	18.15	39.71	16.62	83.93

Table 5: Explanation scores with automatic NLG for generated explanations (QA→E) from UMAE$_{\text{ALL}}$ model with different decoding strategies. The last two rows (with blue shadow) indicate out-of-domain performance.

Dataset	Decoding	e-ViL	N-gram Scores								Learnt Sc.
		S_E	BLEU1	BLEU2	BLEU3	BLEU4	ROUGE-L	METEOR	CIDEr	SPICE	BERTScore
A-OKVQA	beamsearch	**47.01**	54.75	41.39	**32.08**	**24.25**	**49.75**	**22.54**	**86.28**	**20.68**	**87.39**
	nucleus ($p = 0.5$)	46.72	**55.53**	**41.63**	31.91	23.67	49.16	22.48	82.37	20.67	87.18
VCR	beamsearch	**37.02**	25.00	18.90	**14.87**	**11.54**	**27.07**	**15.66**	**38.77**	**25.03**	**80.68**
	nucleus ($p = 0.1$)	35.10	**27.41**	**19.36**	14.50	10.73	26.18	15.21	34.99	21.88	80.52
VQA-X	beamsearch	38.13	39.91	26.30	17.99	12.46	31.69	19.11	42.10	18.15	84.95
	nucleus ($p = 0.1$)	**39.67**	**44.92**	**28.88**	**19.04**	**12.55**	**33.08**	**20.07**	**44.28**	**19.19**	**85.21**

Table 6: Explanation scores with automatic NLG for generated explanations from Q→AE with UMAE$_{\text{ALL}}$ model. The last two rows (with blue shadow) indicate out-of-domain performance.

D More Explanation Scores

For decoding, we evaluate the performance of beam search with the size of 5, top-k sampling with k from $\{50, 100, 200, ..., 1000\}$, and Nucleus and Typical (Meister et al., 2022) sampling, both with p from $\{0.1, 0.2, ..., 0.9\}$. We show the details of the NLG scores using different decoding strategies for explanations generated from QA→E in Table 5, and Q→AE in Table 6.

E Examples of Generated Explanations

Examples of the explanations generated with beam search and Nucleus sampling for A-OKVQA are shown in Figure 4, and VCR in Figure 5.

F Joint Generation Performance

We present the results of the proposed Q→AE task where answers and explanations are jointly generated. We parse the generated sequence to the answer and the explanation and use the same sets of metrics as the separate generation for evaluation. Results for answers in Table 7 and explanations in Table 8. For answers, since the perplexity metric does not directly compare the generation, we show the Multiple-Choice accuracy using the Glove metric for A-OKVQA and BERTScore for VCR answer sentences.

Task	A-OKVQA	VCR	VQA-X
	MC (GOLVE)	BERTScore	DA
Q->A	65.67	81.91	77.65
Q->AE	65.67	82.30	69.60

Table 7: Evaluation of answers generated given questions (Q->A) and jointly generated with explanations (Q->AE). MC stands for Multiple Choice, DA for Direct Answer. The last column with a blue shadow indicates out-of-domain performance.

Dataset	S_E		NgramScore		BERTScore	
	QA->E	Q->AE	QA->E	Q->AE	QA->E	Q->AE
A-OKVQA	50.82	47.01	35.69	32.15	88.21	87.39
VCR	40.27	37.02	26.70	24.02	81.91	80.68
VQA-X	40.67	39.67	26.69	25.85	85.40	85.21

Table 8: Scores of explanations generated given answers (QA->E) and jointly generated with answers (Q->AE). The last row with a blue shadow indicates out-of-domain performance.

	OK-VQA	A-OKVQA	
	DA	MC (GLOVE)	DA
BEST	80.94	80.74	66.20
AVERAGE	54.98	71.53	57.29
WORST	16.37	59.35	41.46

Table 9: Human performance on OK-VQA and A-OKVQA measured from the ground truth answers.

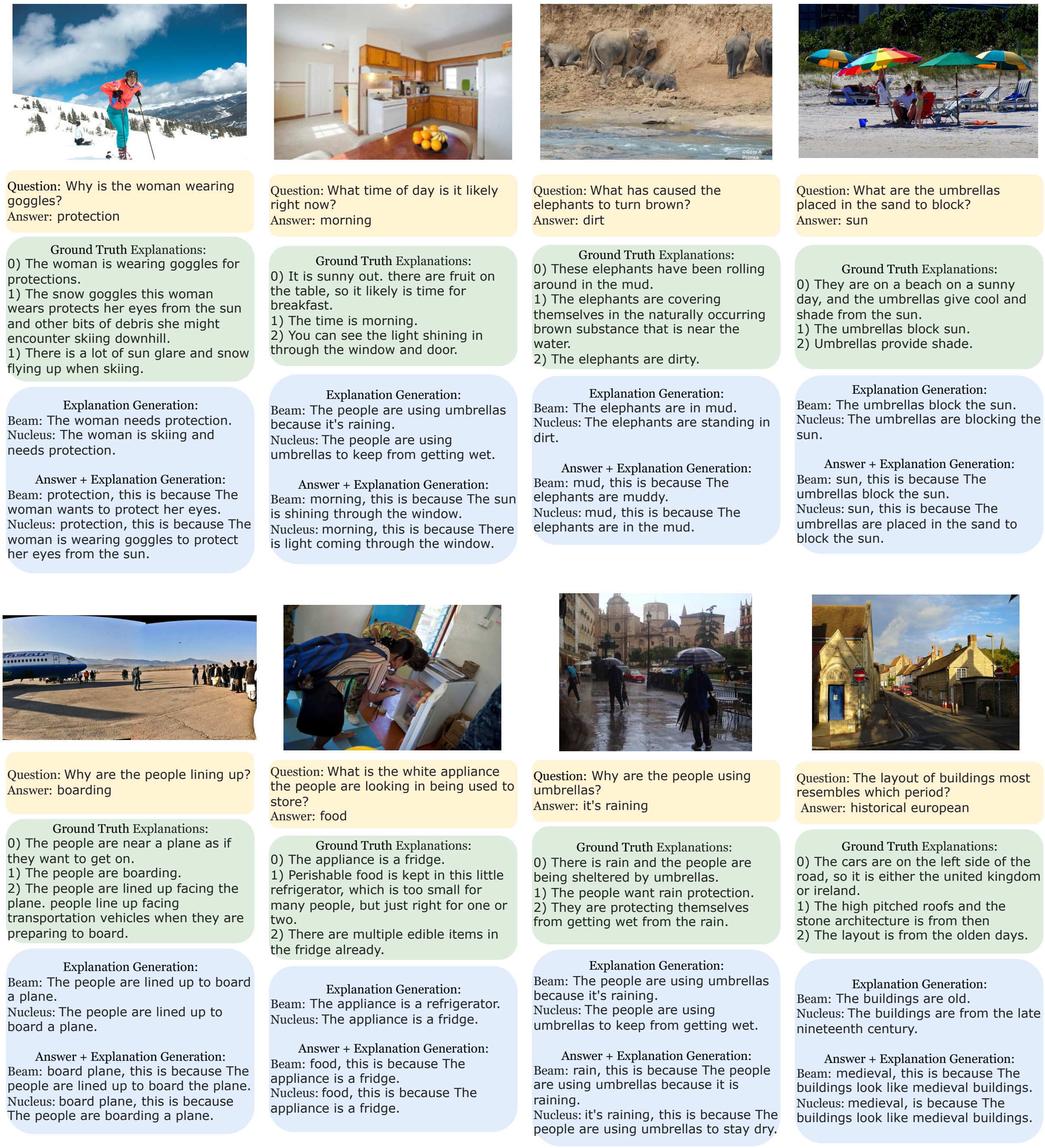
G Datasets Quality and Issues

As mentioned in subsection 5.3, during error analysis we found that many errors are due to the issue in the dataset itself. Concretely, we observe the following issues in the existing datasets: (1) wrong answers (2) subjective or unanswerable questions (3) typos or unclear expressions (4) not requiring images or knowledge to answer the question as designed.

Furthermore, since the answer and explanation for a question in VCR are obtained from the same person who authored the question, this may result in severe subjectivity in the answers or explanations. For example, we find that many questions in VCR require knowledge of the *movie plot* from which the image is extracted, rather than *common-sense reasoning* to answer the questions. While human annotators have an implicit understanding of the movies, the dataset itself does not contain relevant contextual information.

We show some of the issues in the datasets below. Figure 6 shows examples from VCR that require an understanding of the movie plot to generate answers. Figure 7 shows examples from OK-VQA where questions and answers are subjective or ambiguous. Figure 8 shows examples from A-OKVQA and VQA-X that either contain wrong answers, questions that do not need visual input or typos which severely impact the model generation ("house" should be "horse").

To understand the inter-annotator agreement for the datasets, we further measure the best, average and worst human performance on OK-VQA and A-OKVQA by selecting the most common answer, a random answer, and the least common answer, respectively, from the 10 ground truth answers for each question. We calculate the performance using the VQA metric for direct answers, and the GloVe metric for Multiple Choice for simplicity. Note that we also remove the answer selected from the ground truth answers when measuring human performance. From the results in Table 9 we can see that the average performance on both datasets is relatively poor, which indicates the noise in the datasets. The quality of the datasets needs to be more carefully inspected so that the model performance evaluated on these datasets can be more meaningful.



Question: Why is the woman wearing goggles?
Answer: protection

Ground Truth Explanations:
0) The woman is wearing goggles for protections.
1) The snow goggles this woman wears protects her eyes from the sun and other bits of debris she might encounter skiing downhill.
1) There is a lot of sun glare and snow flying up when skiing.

Explanation Generation:
Beam: The woman needs protection.
Nucleus: The woman is skiing and needs protection.

Answer + Explanation Generation:
Beam: protection, this is because The woman wants to protect her eyes.
Nucleus: protection, this is because The woman is wearing goggles to protect her eyes from the sun.

Question: What time of day is it likely right now?
Answer: morning

Ground Truth Explanations:
0) It is sunny out. there are fruit on the table, so it likely is time for breakfast.
1) The time is morning.
2) You can see the light shining in through the window and door.

Explanation Generation:
Beam: The people are using umbrellas because it's raining.
Nucleus: The people are using umbrellas to keep from getting wet.

Answer + Explanation Generation:
Beam: morning, this is because The sun is shining through the window.
Nucleus: morning, this is because There is light coming through the window.

Question: What has caused the elephants to turn brown?
Answer: dirt

Ground Truth Explanations:
0) These elephants have been rolling around in the mud.
1) The elephants are covering themselves in the naturally occurring brown substance that is near the water.
2) The elephants are dirty.

Explanation Generation:
Beam: The elephants are in mud.
Nucleus: The elephants are standing in dirt.

Answer + Explanation Generation:
Beam: mud, this is because The elephants are muddy.
Nucleus: mud, this is because The elephants are in the mud.

Question: What are the umbrellas placed in the sand to block?
Answer: sun

Ground Truth Explanations:
0) They are on a beach on a sunny day, and the umbrellas give cool and shade from the sun.
1) The umbrellas block sun.
2) Umbrellas provide shade.

Explanation Generation:
Beam: The umbrellas block the sun.
Nucleus: The umbrellas are blocking the sun.

Answer + Explanation Generation:
Beam: sun, this is because The umbrellas block the sun.
Nucleus: sun, this is because The umbrellas are placed in the sand to block the sun.

Question: Why are the people lining up?
Answer: boarding

Ground Truth Explanations:
0) The people are near a plane as if they want to get on.
1) The people are boarding.
2) The people are lined up facing the plane. people line up facing transportation vehicles when they are preparing to board.

Explanation Generation:
Beam: The people are lined up to board a plane.
Nucleus: The people are lined up to board a plane.

Answer + Explanation Generation:
Beam: board plane, this is because The people are lined up to board the plane.
Nucleus: board plane, this is because The people are boarding a plane.

Question: What is the white appliance the people are looking in being used to store?
Answer: food

Ground Truth Explanations:
0) The appliance is a fridge.
1) Perishable food is kept in this little refrigerator, which is too small for many people, but just right for one or two.
2) There are multiple edible items in the fridge already.

Explanation Generation:
Beam: The appliance is a refrigerator.
Nucleus: The appliance is a fridge.

Answer + Explanation Generation:
Beam: food, this is because The appliance is a fridge.
Nucleus: food, this is because The appliance is a fridge.

Question: Why are the people using umbrellas?
Answer: it's raining

Ground Truth Explanations:
0) There is rain and the people are being sheltered by umbrellas.
1) The people want rain protection.
2) They are protecting themselves from getting wet from the rain.

Explanation Generation:
Beam: The people are using umbrellas because it's raining.
Nucleus: The people are using umbrellas to keep from getting wet.

Answer + Explanation Generation:
Beam: rain, this is because The people are using umbrellas because it is raining.
Nucleus: it's raining, this is because The people are using umbrellas to stay dry.

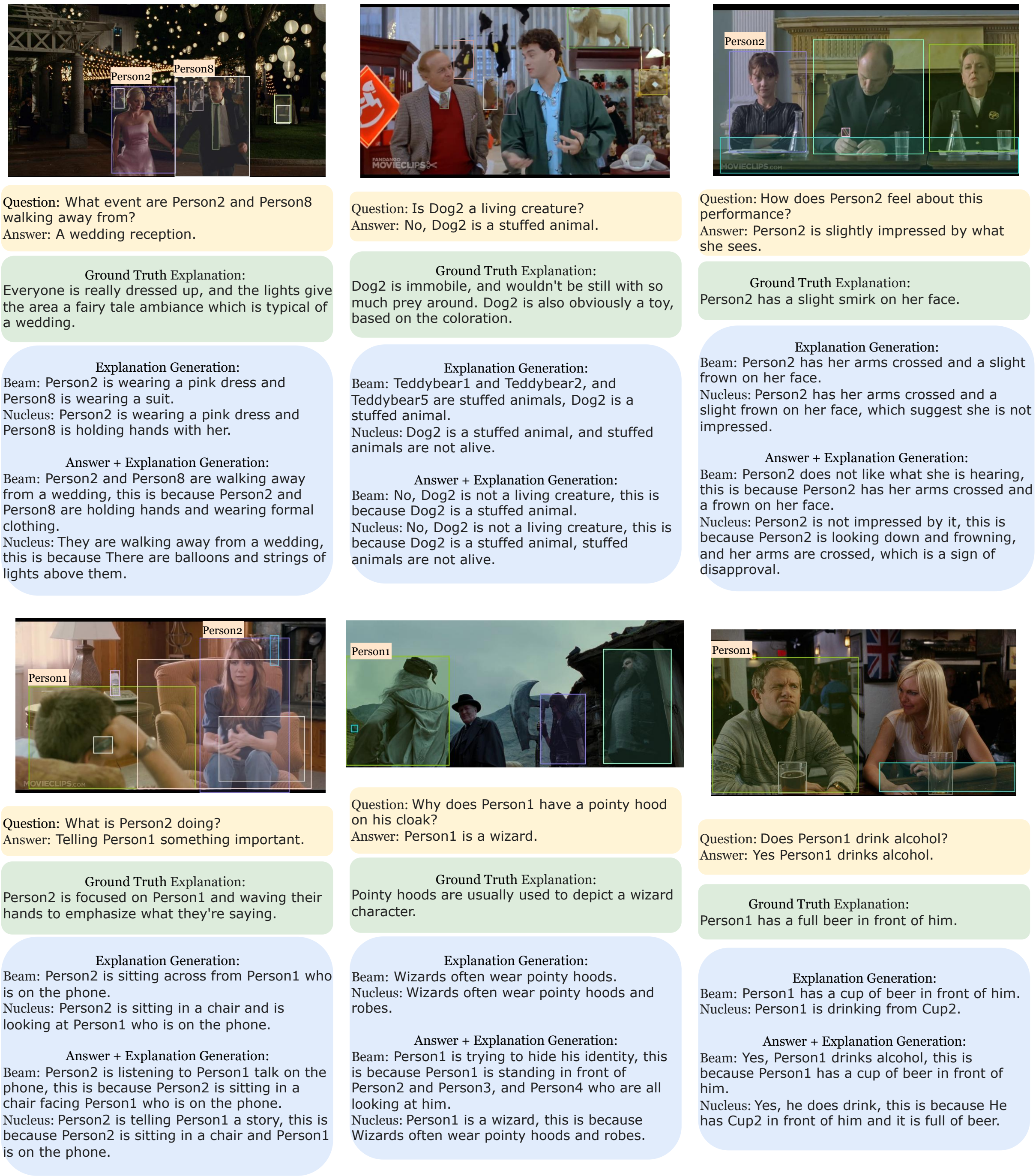
Question: The layout of buildings most resembles which period?
Answer: historical european

Ground Truth Explanations:
0) The cars are on the left side of the road, so it is either the united kingdom or ireland.
1) The high pitched roofs and the stone architecture is from then
2) The layout is from the olden days.

Explanation Generation:
Beam: The buildings are old.
Nucleus: The buildings are from the late nineteenth century.

Answer + Explanation Generation:
Beam: medieval, this is because The buildings look like medieval buildings.
Nucleus: medieval, is because The buildings look like medieval buildings.

Figure 4: Examples of generated answers and explanations for A-OKVQA.



Question: What event are Person2 and Person8 walking away from?
Answer: A wedding reception.

Ground Truth Explanation:
Everyone is really dressed up, and the lights give the area a fairy tale ambiance which is typical of a wedding.

Explanation Generation:
Beam: Person2 is wearing a pink dress and Person8 is wearing a suit.
Nucleus: Person2 is wearing a pink dress and Person8 is holding hands with her.

Answer + Explanation Generation:
Beam: Person2 and Person8 are walking away from a wedding, this is because Person2 and Person8 are holding hands and wearing formal clothing.
Nucleus: They are walking away from a wedding, this is because There are balloons and strings of lights above them.

Question: Is Dog2 a living creature?
Answer: No, Dog2 is a stuffed animal.

Ground Truth Explanation:
Dog2 is immobile, and wouldn't be still with so much prey around. Dog2 is also obviously a toy, based on the coloration.

Explanation Generation:
Beam: Teddybear1 and Teddybear2, and Teddybear5 are stuffed animals, Dog2 is a stuffed animal.
Nucleus: Dog2 is a stuffed animal, and stuffed animals are not alive.

Answer + Explanation Generation:
Beam: No, Dog2 is not a living creature, this is because Dog2 is a stuffed animal.
Nucleus: No, Dog2 is not a living creature, this is because Dog2 is a stuffed animal, stuffed animals are not alive.

Question: How does Person2 feel about this performance?
Answer: Person2 is slightly impressed by what she sees.

Ground Truth Explanation:
Person2 has a slight smirk on her face.

Explanation Generation:
Beam: Person2 has her arms crossed and a slight frown on her face.
Nucleus: Person2 has her arms crossed and a slight frown on her face, which suggest she is not impressed.

Answer + Explanation Generation:
Beam: Person2 does not like what she is hearing, this is because Person2 has her arms crossed and a frown on her face.
Nucleus: Person2 is not impressed by it, this is because Person2 is looking down and frowning, and her arms are crossed, which is a sign of disapproval.

Question: What is Person2 doing?
Answer: Telling Person1 something important.

Ground Truth Explanation:
Person2 is focused on Person1 and waving their hands to emphasize what they're saying.

Explanation Generation:
Beam: Person2 is sitting across from Person1 who is on the phone.
Nucleus: Person2 is sitting in a chair and is looking at Person1 who is on the phone.

Answer + Explanation Generation:
Beam: Person2 is listening to Person1 talk on the phone, this is because Person2 is sitting in a chair facing Person1 who is on the phone.
Nucleus: Person2 is telling Person1 a story, this is because Person2 is sitting in a chair and Person1 is on the phone.

Question: Why does Person1 have a pointy hood on his cloak?
Answer: Person1 is a wizard.

Ground Truth Explanation:
Pointy hoods are usually used to depict a wizard character.

Explanation Generation:
Beam: Wizards often wear pointy hoods.
Nucleus: Wizards often wear pointy hoods and robes.

Answer + Explanation Generation:
Beam: Person1 is trying to hide his identity, this is because Person1 is standing in front of Person2 and Person3, and Person4 who are all looking at him.
Nucleus: Person1 is a wizard, this is because Wizards often wear pointy hoods and robes.

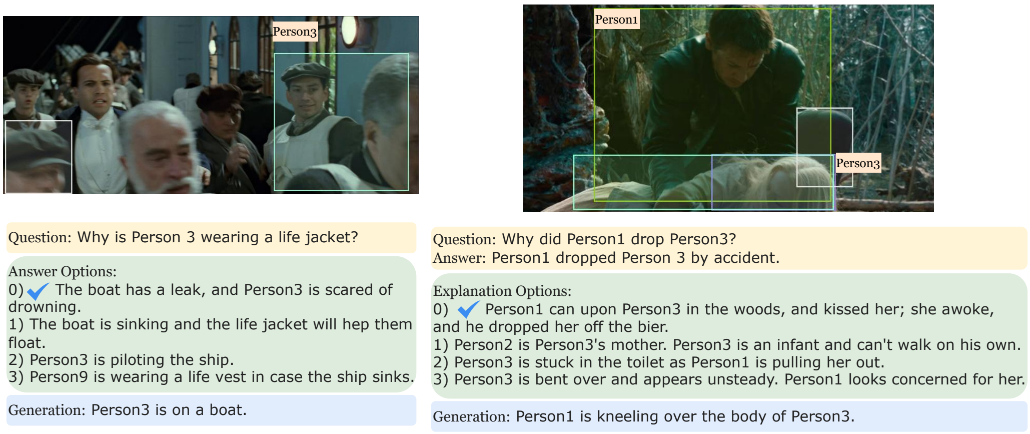
Question: Does Person1 drink alcohol?
Answer: Yes Person1 drinks alcohol.

Ground Truth Explanation:
Person1 has a full beer in front of him.

Explanation Generation:
Beam: Person1 has a cup of beer in front of him.
Nucleus: Person1 is drinking from Cup2.

Answer + Explanation Generation:
Beam: Yes, Person1 drinks alcohol, this is because Person1 has a cup of beer in front of him.
Nucleus: Yes, he does drink, this is because He has Cup2 in front of him and it is full of beer.

Figure 5: Examples of generated answers and explanations generation for VCR.



Question: Why is Person 3 wearing a life jacket?

Answer Options:
0) ✓ The boat has a leak, and Person3 is scared of drowning.
1) The boat is sinking and the life jacket will hep them float.
2) Person3 is piloting the ship.
3) Person9 is wearing a life vest in case the ship sinks.

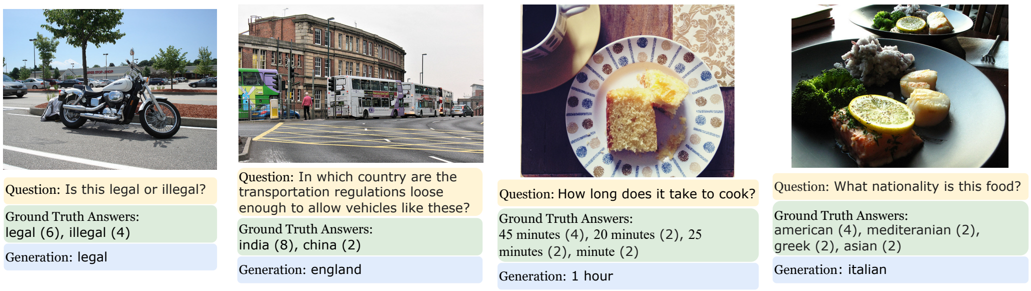
Generation: Person3 is on a boat.

Question: Why did Person1 drop Person3?
Answer: Person1 dropped Person 3 by accident.

Explanation Options:
0) ✓ Person1 can upon Person3 in the woods, and kissed her; she awoke, and he dropped her off the bier.
1) Person2 is Person3's mother. Person3 is an infant and can't walk on his own.
2) Person3 is stuck in the toilet as Person1 is pulling her out.
3) Person3 is bent over and appears unsteady. Person1 looks concerned for her.

Generation: Person1 is kneeling over the body of Person3.

Figure 6: Questions that require knowledge of the movie plots to generate the answers from VCR.



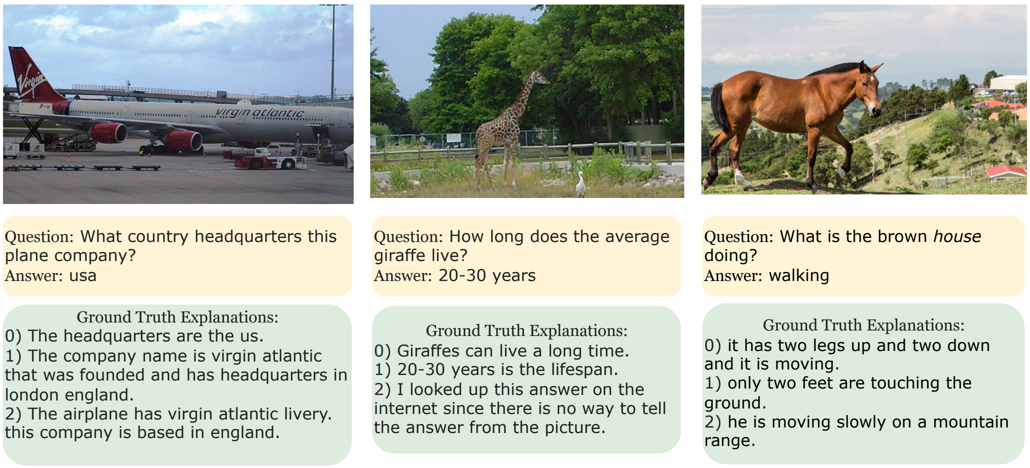
Question: Is this legal or illegal?
Ground Truth Answers:
legal (6), illegal (4)
Generation: legal

Question: In which country are the transportation regulations loose enough to allow vehicles like these?
Ground Truth Answers:
india (8), china (2)
Generation: england

Question: How long does it take to cook?
Ground Truth Answers:
45 minutes (4), 20 minutes (2), 25 minutes (2), minute (2)
Generation: 1 hour

Question: What nationality is this food?
Ground Truth Answers:
american (4), mediteranian (2), greek (2), asian (2)
Generation: italian

Figure 7: Examples of subjective questions from OK-VQA.



Question: What country headquarters this plane company?
Answer: usa

Ground Truth Explanations:
0) The headquarters are the us.
1) The company name is virgin atlantic that was founded and has headquarters in london england.
2) The airplane has virgin atlantic livery. this company is based in england.

Question: How long does the average giraffe live?
Answer: 20-30 years

Ground Truth Explanations:
0) Giraffes can live a long time.
1) 20-30 years is the lifespan.
2) I looked up this answer on the internet since there is no way to tell the answer from the picture.

Question: What is the brown *house* doing?
Answer: walking

Ground Truth Explanations:
0) it has two legs up and two down and it is moving.
1) only two feet are touching the ground.
2) he is moving slowly on a mountain range.

Figure 8: Issues in the datasets that severely impact the model generation: wrong answers (left, from A-OKVQA), questions do not need visual input to answer (middle, from A-OKVQA), and typo (right, from VQA-X).