# INTERPRETABLE VISUAL QUESTION ANSWERING VIA REASONING SUPERVISION

*Maria Parelli*[†‡]   *Dimitrios Mallis*[†]   *Markos Diomataris*[†‡*]   *Vassilis Pitsikalis*[†]

[†] DeepLab, Athens, Greece
[‡] ETH Zürich, Zürich, Switzerland

maryparelli@gmail.com, mallidimi1@gmail.com, mdiomataris@student.ethz.ch, vpitsik@deeplab.ai

## ABSTRACT

Transformer-based architectures have recently demonstrated remarkable performance in the Visual Question Answering (VQA) task. However, such models are likely to disregard crucial visual cues and often rely on multimodal shortcuts and inherent biases of the language modality to predict the correct answer, a phenomenon commonly referred to as *lack of visual grounding*. In this work, we alleviate this shortcoming through a novel architecture for visual question answering that leverages *common sense reasoning as a supervisory signal*. Reasoning supervision takes the form of a textual justification of the correct answer, with such annotations being already available on large-scale Visual Common Sense Reasoning (VCR) datasets. The model's visual attention is guided toward important elements of the scene through a similarity loss that aligns the learned attention distributions guided by the question and the correct reasoning. We demonstrate both quantitatively and qualitatively that the proposed approach can boost the model's visual perception capability and lead to performance increase, without requiring training on explicit grounding annotations.

***Index Terms***— Visual Question Answering, Visual Grounding, Interpretability, Attention Similarity

## 1. INTRODUCTION

Models for Visual Question Answering (VQA) provide answers to natural language questions about an image by perceiving both textual and image cues. VQA lies at the intersection of vision and language and has recently generated significant research interest. Existing methods aim to tackle the task via deep multi-layer transformer architectures, attending to linguistic and visual tokens [1, 2, 3, 4, 5]. However, despite their superior performance, attempts to diagnose these models' robustness and reasoning capability have revealed that they often rely on linguistic biases and shallow correlations to generate the correct answer [6, 7]. The language modality has been proven a strong signal that is easy to exploit, causing the model to overlook visual information and rely on shallow



**Fig. 1**. This work proposes a novel mechanism for leveraging *common sense reasoning as a supervisory signal*. Our VQA model, guided by the correct reasoning (R:[PERSON1] IS HOLDING THE CIGARETTE), is able to attend to the appropriate image regions and accurately select the right answer (A:[PERSON1] IS SMOKING).

patterns, such as correlations between words in the question [8]. It has been shown that the performance of recent models can clearly degrade under evaluation settings that penalize reliance on such spurious correlations [9, 10, 11].

This tendency of recent models to reason about the correct answer without attending to the relevant image areas has been referred to as *lack of visual grounding* [12, 13]. To alleviate this, a line of work explores techniques for training VQA models that are sensitive to the same image regions as human annotators, commonly by enforcing alignment with human attention maps [12, 14]. While such methods can reduce reliance on language biases, they also require explicit grounding supervision that is rarely available. In this work, we explore an alternative approach towards attending to informative image regions, that does not require explicit grounding supervision, but leverages instead *common sense reasoning* as a supervisory signal.

We take advantage of the fact that reasoning-level supervision in the form of textual justification of why an answer is true, is already available in large-scale Visual Common-

---

*Work was done while Markos Diomataris was with DeepLab.

sense Reasoning datasets like [15]. For example, in Fig. 1, to answer the **question** *'What is* [PERSON1] *doing?'*, the **reasoning** *'*[PERSON1] *is holding a cigarette and is leaned over it'* can accurately guide a model's visual attention towards predicting the correct **answer**, *'*[PERSON1] *is smoking'*. The correct reasoning often contains details of the scene and references to objects and people relevant to the right answer. Our VQA model is trained to utilize reasoning supervision as a proxy signal to generate interpretable attention maps that guide visual attention toward informative image regions.

Our proposed framework processes question/answer pairs using a multilayer BERT [16] transformer architecture. A separate visual attention stream is incorporated to generate two attention distributions, one conditioned on the question and the other on the correct reasoning. We distill knowledge from the reasoning attention to our VQA model through a similarity loss term, that encourages question and reasoning attention alignment. Our model can accurately capture the visual components required to find the correct answer and produce interpretable, human-like attention maps, thus boosting baseline performance. We evaluate our pipeline both quantitatively and qualitatively on the Visual Commonsense Reasoning dataset [15], a large-scale dataset for cognition-level visual understanding. To the best of our knowledge, we are one of the first works to employ implicit attention guidance, free from explicit grounding supervision in a vision-language transformer setting.

## 2. RELATED WORK

The main VQA paradigm is multi-layer transformers operating on joint image-text embeddings [3, 5, 1, 3]. These methods benefit from extensive pre-training on large-scale VL datasets, to extract meaningful image-text representations and align visio-linguistic clues. One notable example is VL-BERT [2], a model that is pre-trained on text-only corpora with standard Masked Language Modeling (MLM) as well as visual-linguistic corpora via predicting randomly masked words and Regions of Interest (RoIs) of the image.

Despite superior performance, state-of-the-art VQA models can often make decisions by relying on shortcuts and statistical regularities instead of comprehending the scene as demonstrated in [10]. Similarly, authors in [17] identify that VQA models exploit co-occurrences of words in the question and object segments in the image, which they define as multimodal shortcuts.

In an attempt to counter shortcuts and language priors, some methods encourage the model to effectively attend to visual components and infer visual relationships. The authors of [12] align gradient-based explanations with human attention annotations via a ranking loss to guide the network to focus on the correct image regions. The authors of [14] train an attention auxiliary model with ground truth human-labeled attention maps and consequently apply human-like attention

supervision to an attention-based VQA model. Another work in this direction [18] proposes a method that automatically selects region and object annotations from Visual Genome [19] that serve as labels for implementing visual grounding as an auxiliary task for VQA. In contrast to these approaches, this work mitigates over-reliance on language priors without requiring annotated attention maps. We train our network instead, to look at the image and attend to meaningful visual evidence through reasoning supervision.

## 3. METHODOLOGY

**Problem statement.** A VQA model is tasked with answering natural language questions from the visual content of a scene. Given a dataset $\mathcal{X} = \{u_i, q_i, a_i, r_i\}_{i=1}^N$ of $N$ images where $u_i \in V$ is the visual input with question $q_i \in Q$, reasoning $r_i \in R$ and groundtruth answer $a_i \in A$, our goal is to learn a function $f : Q \times V \to \mathbb{R}^A$ that predicts a distribution $P(A)$ over possible answers in $A$. Our proposed pipeline consists of two parallel streams, a *language stream* with model parameters $\theta_L$ and a *visual attention stream* with model parameters $\theta_{V_q}$ and $\theta_{V_r}$ (question and reasoning guided attention decoder that we will discuss next). During training, we will utilize reasoning supervision as an additional supervisory signal, thus modeling $P(A|u_i, q_i, r_i; (\theta_L, \theta_{V_q}, \theta_{V_r}))$ that simplifies to $P(A|u_i, q_i; (\theta_L, \theta_{V_q}))$ at test time.

**Language Stream.** The first stream is language-focused and aims to generate an informative representation of the input question and answer sentence pairs by modeling their relationship. The core of its architecture is a bi-directional 12-layer transformer initialized with weights from BERT [16]. It takes a sequence of word embeddings of the question and answer as input (separated by a separation element [SEP]) and adds a sequence positional embedding to each token. The final output feature $x_{[CLS]}$ of the [CLS] element is used to obtain the final pooled linguistic representation.

**Visual Attention Stream.** The visual attention stream consists of two 9-layer transformer decoders. The first one generates an attention vector over the image features guided by the question, and the second an attention vector over the image features guided by the correct reasoning. We take advantage of the cross-attention module to perceive multimodal information and capture relationships between image features and word embeddings. The process is as follows: The image is first processed via the backbone of a ResNet-50-FPN to extract visual appearance features. The output is a feature map $\mathcal{F} \in \mathbb{R}^{H \times W \times 256}$, which we treat as a sequence of 256-dimensional image features. Following [2], a visual geometric embedding is added to each input token to inject 2D awareness into the model. We also encode the question and correct reasoning language tokens via a pre-trained BERT model, which yields a 786-dim representation for each word.

**Fig. 2**. Proposed VAQ architecture: Our model comprises 2 main streams that operate in parallel, a *language* and a *visual* stream. The output of the 2 streams is fused via Hadamard multiplication to obtain the final prediction. During training, we utilize a *reasoning* attention decoder to distil reasoning information into the model, through a similarity loss between question and reasoning-guided attention maps. Reasoning supervision leads in the formation of interpretable attention maps.

Question and reasoning word embeddings are used as input to the corresponding question and reasoning transformer decoders (functioning as query tokens). The image visual features $\mathcal{F}$ are used to generate the keys and values. Then, the attention weights are calculated based on the pairwise similarity of the query and key elements. The output of each decoder is an attention distribution over the image regions $\alpha \in \mathbb{R}^{H \times W}$, conditioned on either the word embeddings of the question (referred to as $\alpha^Q$) or the word embeddings of the correct reasoning (referred to as $\alpha^R$). In practice, to obtain the final attention vectors $\alpha^Q$ and $\alpha^R$, we compute the average per-head attention of the last layer generated by the $[CLS]$ token over the image features.

The generated attention map $\alpha^Q$, is then used to take the *weighted sum* over the image features $\mathcal{F}$, which is passed through a linear layer to obtain the final *attended-by-the-question* representation of the image, $V_q$. The same operation is performed to obtain the *attended-by-the-reasoning* image representation $V_r$. Formally,

$$V_q = Linear(\alpha^Q \odot \mathcal{F})$$
$$V_r = Linear(\alpha^R \odot \mathcal{F}) \qquad (1)$$

**Combining Language and Visual Streams.** The model outputs two separate predictions, one conditioned on the question $y_q$ and the other on the correct reasoning $y_r$. These are produced by fusing the outputs of the *language stream* $x_{[CLS]}$ and the *visual attention stream*, via Hadamard multiplication and then passing them through a softmax classifier $s$, thus $y^q = s(x_{[CLS]} \odot V_q)$ and $y^r = s(x_{[CLS]} \odot V_r)$. At test time, $y^q$ is used to provide predictions over possible answers.

**Training.** Our training pipeline consists of 2 stages. In the first stage, we train with two cross-entropy loss terms $\mathcal{L}_q$ and $\mathcal{L}_r$ w.r.t the ground truth answer $[a_i]$, or

$$\mathcal{L}_{stage_1} = -\frac{1}{N} \sum_i^N \log(y_i^q)[a_i] - \frac{1}{N} \sum_i^N \log(y_i^r)[a_i] \quad (2)$$

In the second stage, we distill knowledge from the reasoning decoder by aligning the attention distributions conditioned on the question $\alpha^Q$ to the attention distributions conditioned on the correct reasoning $\alpha^R$. To that end, we freeze the weights of the reasoning attention decoder and only fine-tune the question attention decoder (through $\mathcal{L}_q$) while also utilizing an attention similarity loss, formulated as the forward Kullback-Leibler divergence between attention maps $\alpha^Q$ and $\alpha^R$, or $D_{KL}(\alpha^Q || \alpha^R)$. The complete $\mathcal{L}_{stage_2}$ loss is:

$$\mathcal{L}_{stage_2} = -\frac{1}{N} \sum_i^N \log(y_i^q)[a_i] + \frac{1}{N} \sum_i^N \alpha_i^Q \log(\frac{\alpha_i^Q}{\alpha_i^R}) \quad (3)$$

The whole process is illustrated in Figure 2. Our model is trained for 11 epochs for stage 1 and then finetuned for 5 more epochs in stage 2.

## 4. EXPERIMENTS

**Dataset.** We validate our VQA model on the Visual Commonsense Reasoning dataset [15], which consists of 290k QA problems derived from 110k movie scenes. Four possible answers and four rationales are provided for each question,

**Fig. 3**. Comparison of question-guided attention maps (only the question attention decoder) before *(first row)*, and after fine-tuning with reasoning supervision *(second row)*. We observe that the finetuned model is able to attend to informative regions.

| Model | Acc(%) |
|---|---|
| *Baseline model* | 61.2 |
| *Reasoning Supervision* | **63.9** |

**Table 1**. Accuracy of the baseline and our proposed model finetuned with *reasoning supervision* on VCR.

| Model | Acc(%) | (+*masking*) Acc(%) |
|---|---|---|
| *Baseline model* | 61.2 | 59.3 (−1.9) |
| *Reasoning Supervision* | 63.9 | 61.1 (−2.8) |

**Table 2**. Performance drop on the VCR validation set due to object masking.

but we use only the correct rationale/reasoning. Note that reasoning is only used during training as additional supervision.

**Quantitative Evaluation.** Results in terms of model accuracy are reported in Table 1. Our baseline model (only the question decoder) achieves 61.2% accuracy on the validation set. Finetuning by aligning question and reasoning attention distributions yields 63.9%, that is a 2.7% absolute improvement, thus demonstrating the benefit of reasoning supervision. We note that our main goal is to propose a novel training strategy for boosting a VQA model's visual explanatory strength by exploiting reasoning as an alternative supervisory signal. Thus, we do not directly compare to methods such as [2, 3, 20] that contain a larger number of parameters, leverage large-scale VL and video pretraining or ground-truth object bounding boxes. For comparison, the best performance reported in R2C [15] was 63.8%.

To further investigate our model's ability to leverage the visual modality, we perform an ablation study where we mask the visual features of the objects/people referenced by the question at test time and measure the effect on accuracy.

Results are reported in Table 2. We observe that the baseline VQA model (that does not fully alleviate the lack of visual grounding) suffers a lesser performance degradation of 1.9% compared to 2.8% for our finetuned model (on reasoning supervision). This is a different manifestation of the fact, that the baseline model is over-reliant on the language modality, thus performance is penalized less when visual information is not available due to object masking.

**Visual Results.** In Fig. 3, we visualize attention maps ($\alpha^Q$) for both the baseline model *(above)* and finetuned model (with reasoning supervision) *(below)*. The correct reasoning can intuitively provide important guidance during training. For example, for the question (Q: WHICH PERSON IS THE LEAD FOR THIS DANCE GROUP?), the reasoning (R: [1] IS IN THE MIDDLE, WHICH IS GENERALLY WHERE THE MAIN DANCER GOES)) clearly explains the dynamics of different elements of the scene. This information is distilled to our VQA model through our attention similarity loss. In Fig. 3, we observe that after fine-tuning, visual attention improves. Our method is able to produce interpretable, human-like attention maps, thus being able to predict the correct answer by perceiving relevant visual concepts.

## 5. CONCLUSION

In this work, we alleviate the lack of visual grounding through reasoning supervision. This additional supervision takes the form of textual justifications of the correct answer and it's already available for VCR datasets. We incorporate a similarity loss that encourages the alignment between the visual attention maps (guided by the question and correct reasoning) thus improving the model's visual perception capability. We demonstrate qualitatively and quantitatively that reasoning information can lead to interpretable attention maps and performance increase for visual question answering.

# 6. REFERENCES

[1] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao, "Oscar: Object-semantics aligned pre-training for vision-language tasks," *ECCV 2020*, 2020.

[2] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai, "Vl-bert: Pre-training of generic visual-linguistic representations," in *ICLR*, 2020.

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu, "Uniter: Universal image-text representation learning," in *ECCV*, 2020.

[4] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao, "Vinvl: Making visual representations matter in vision-language models," *CVPR*, 2021.

[5] Hao Tan and Mohit Bansal, "Lxmert: Learning cross-modality encoder representations from transformers," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

[6] Tejas Gokhale, Pratyay Banerjee, Chitta Baral, and Yezhou Yang, "MUTANT: A training paradigm for out-of-distribution generalization in visual question answering," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020, pp. 878–892.

[7] Qingyi Si, Yuanxin Liu, Fandong Meng, Zheng Lin, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou, "Towards robust visual question answering: Making the most of biased samples via contrastive learning," *ArXiv*, vol. abs/2210.04563, 2022.

[8] 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," *IJCV*, 2016.

[9] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi, "Don't just assume; look and answer: Overcoming priors for visual question answering," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980, 2017.

[10] Keren Ye and Adriana Kovashka, "A case study of the shortcut effects in visual commonsense reasoning," in *AAAI*, 2019.

[11] Vipul Gupta, Zhuowan Li, Adam Kortylewski, Chenyu Zhang, Yingwei Li, and Alan Yuille, "Swapmix: Diagnosing and regularizing the over-reliance on visual context in visual question answering," in *CVPR*, 2022.

[12] Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh, "Taking a hint: Leveraging explanations to make vision and language models more grounded," in *ICCV*, 2019.

[13] Chongyan Chen, Samreen Anjum, and Danna Gurari, "Grounding answers for visual questions asked by visually impaired people," in *CVPR*, 2022.

[14] Tingting Qiao, Jianfeng Dong, and Duanqing Xu, "Exploring human-like attention supervision in visual question answering," in *AAAI*, 2018.

[15] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi, "From recognition to cognition: Visual commonsense reasoning," in *CVPR*, June 2019.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *ArXiv*, vol. abs/1810.04805, 2019.

[17] Corentin Dancette, Remi Cadene, Damien Teney, and Matthieu Cord, "Beyond question-based biases: Assessing multimodal shortcut learning in visual question answering," 2021.

[18] Yundong Zhang, Juan Carlos Niebles, and Alvaro Soto, "Interpretable visual question answering by visual grounding from attention supervision mining," in *WACV*, 2019.

[19] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *IJCV*, 2017.

[20] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi, "Merlot: Multimodal neural script knowledge models," in *NeurIPS*, 2021.

[21] Ramprasaath R. Selvaraju, Purva Tendulkar, Devi Parikh, Eric Horvitz, Marco Tulio Ribeiro, Besmira Nushi, and Ece Kamar, "Squinting at vqa models: Introspecting vqa models with sub-questions," in *CVPR*, 2020.