Zero-shot Translation of Attention Patterns in VQA Models to Natural Language

Leonard Salewski[1][0000−0001−8531−3011], A. Sophia Koepke[1][0000−0002−5807−0576],
Hendrik P. A. Lensch[1][0000−0003−3616−8668], and
Zeynep Akata[1,2][0000−0002−1432−7747]

[1] University of Tübingen
[2] MPI for Intelligent Systems
{leonard.salewski, a-sophia.koepke, hendrik.lensch,
zeynep.akata}@uni-tuebingen.de

Abstract. Converting a model's internals to text can yield human-understandable insights about the model. Inspired by the recent success of training-free approaches for image captioning, we propose ZS-A2T, a zero-shot framework that translates the transformer attention of a given model into natural language without requiring any training. We consider this in the context of Visual Question Answering (VQA). ZS-A2T builds on a pre-trained large language model (LLM), which receives a task prompt, question, and predicted answer, as inputs. The LLM is guided to select tokens which describe the regions in the input image that the VQA model attended to. Crucially, we determine this similarity by exploiting the text-image matching capabilities of the underlying VQA model. Our framework does not require any training and allows the drop-in replacement of different guiding sources (e.g. attribution instead of attention maps), or language models. We evaluate this novel task on textual explanation datasets for VQA, giving state-of-the-art performances for the zero-shot setting on GQA-REX and VQA-X. Our code is available here.

Keywords: Zero-Shot Translation of Attention Patterns · VQA.

1 Introduction

Deep learning systems have become an integral part of our society, both in non-obvious applications, e.g. customer credit scoring, as well as in prominent applications, e.g. ChatGPT [25]. Their impact on the lives of millions of people establishes the need to make these algorithms more transparent and accessible. In the context of Visual Question Answering (VQA) [2,13], methods for attribution [33] or attention visualization [1] aim to highlight the input image regions that are most relevant for the final decision. However, the actual visual concepts that the model "saw" in the salient regions can remain obscure to the user. In contrast, a natural language description of those visual concepts can be a more intuitive format for a human user.

There is a wide variety of approaches to determine image regions that were relevant for a model's output. Many of those methods do not require training and can be directly applied to a given model to generate attribution or attention visualizations. As it is

infeasible to train a dedicated model for translating method-specific visual explanations to natural language, we try to address the question: Can we convert the output of any attention or attribution method to natural language without any supervisory data?

Inspired by the impressive capabilities of pre-trained LLMs, we propose a *zero-shot attention-to-text* (ZS-A2T) framework which translates the internal attention of a transformer-based VQA model into natural language without requiring training (see Figure 1). In particular, ZS-A2T uses a LLM that is steered by visual inputs corresponding to model attribution or attention visualizations. We judge the visual relevance of the LLM's proposals with an image-text matching framework. In contrast to related zero-shot image captioning methods like [37,41,40], ZS-A2T does not exploit
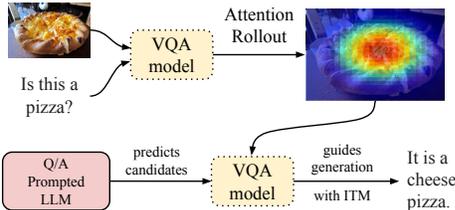

Fig. 1: Our ZS-A2T framework performs training-free translation of attention patterns from VQA models into natural language by combining a pre-trained large language model with language guiding with the VQA model's attention.

CLIP [29], whose image-text understanding would be different from the original task model. Instead, we re-use the encoders of the underlying VQA model for quantifying the agreement between the visual evidence and the candidate word. This guides the text generation without introducing external input into the translation process.

We hypothesize that the content of the verbalizations of the attention patterns for VQA should capture the visual evidence that was used for VQA. This visual evidence should also be described in the corresponding textual explanations. Therefore, we evaluate the quality of our generated attention pattern translations on the VQA-X and GQA-REX datasets. Whilst naturally giving weaker results than methods that were trained to generate explanations in a fully supervised manner along with solving the VQA task, our proposed ZS-A2T outperforms all related methods in this novel zero-shot attention pattern translation setting. Additionally, the training-free setup of our method allows our approach to use different language models without any adaption or training. Similarly, our framework works for any attention aggregation or visual attribution method that can be applied to the underlying VQA model.

To summarize, our contributions are: 1) We introduce the task of converting the internal structures of VQA models to natural language. 2) Our proposed zero-shot ZS-A2T framework is a simple, yet effective method for generating textual outputs with the guidance of the aggregated internal attention and text-image matching capabilities of the VQA model. 3) ZS-A2T can be utilized in conjunction with any pre-trained language model or visual attribution technique, and achieves state-of-the-art results on the GQA-REX and VQA-X textual explanation benchmarks in the zero-shot setting.

2   Related Work

Attention visualizations and visual attribution. Visual attribution methods commonly use backpropagation to trace a model's output to its input, e.g. by visualizing the

(slightly modified) gradients on the corresponding input image [34,36,46,35,48,4]. In addition to that, CAM [50], Grad-CAM [33,32], and variants thereof [9,12,24,28] use model activations in the attribution visualizations. Different to visual attribution methods, perturbation-based methods slightly alter the input and record the resulting changes in the output [11,10]. Unfortunately, it is hard to quantify the quality of attribution visualizations [18]. For modern transformer-based models, a number of studies [16,45,1,7] have investigated the extraction and visualization of attention scores. In particular, attention rollout [1] determines the relevance of input tokens by following the flow of information that is captured in the attention scores and residual connections of the transformer network. However, it is not clear if the resulting visualizations are an intuitive way of explaining deep learning models to human users. Our proposed ZS-A2T framework offers a way of translating such outputs into natural language.

Visual conditioning for text generation with LLMs. Several works have used LLMs for zero-shot image captioning [41,37,40,47,44]. ZeroCap [41] uses CLIP [29] for updating the language model's hidden activations to match the image input. Similarly, MAGIC [37] combines the prediction of the language model, CLIPs rating and a degeneration penalty [38]. EPT [40] only optimizes the hidden activations of a few selected pseudo-tokens for zero-shot video captioning. In contrast to these approaches, Socratic Models [47] generates captions by conditioning the language model on CLIP-detected class names. Another creative approach [44] finetunes a language model to process CLIP text embeddings which at test time can be replaced by CLIP image embeddings. Similarly, [22] finetunes with caption data and addresses the modality gap in a training-free manner. Different to the two aforementioned methods [44,22], ZS-A2T does not require any training (beyond the pre-trained models used) or external models for guiding the language generation, such as CLIP. In particular, the text generation in ZS-A2T is controlled using the VQA model's internal attention along with exploiting the image-text matching capabilities of the same model.

Textual explanations for VQA. Several VQA datasets with textual explanations have driven research on explaining outputs of models trained for VQA. In particular, the VQA-X [15] dataset extends a subset of VQA v2 [13] with textual explanations from obtained from humans. In contrast, the VQA-E dataset [21] automatically sources explanations from image captions. Recently, the CLEVR-X [30] and GQA-REX [8] extended the CLEVR [17] and GQA [14] datasets with textual explanations by exploiting the corresponding scene graphs. In this work, we evaluate the verbalization of a model's internal attention on textual explanation datasets in the context of VQA, since textual explanations and visual attribution visualizations should both indicate the key features that influenced a VQA model's output.

3   The ZS-A2T Framework

In this section, we explain our proposed framework which is visualized in Figure 2. We propose the ZS-A2T (**Z**ero-**S**hot **A**ttention **to T**ranslation) framework which converts the internal attention in a VQA model into natural language by translating the visualization of the aggregated attention in the transformer model into natural language without any supervision. In particular, we prompt a pre-trained large language model for the
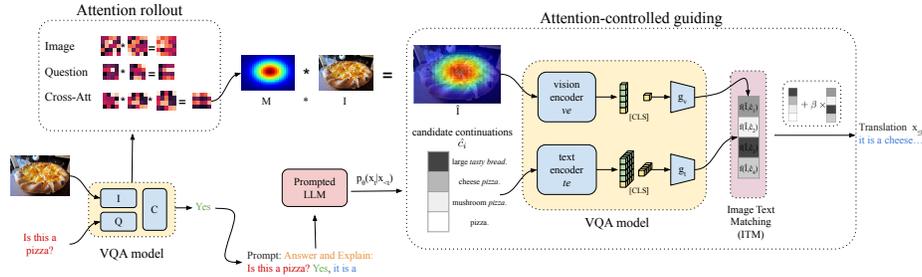
Fig. 2: Our ZS-A2T framework translates the attention rollout [1] of a VQA model into natural language. A pre-trained large language model is prompted with the task description, the question $\mathcal{Q}$, and the predicted answer a. The candidates for the next token x_t along with their continuations (*italic*) are then re-ranked by re-using the image-text matching capabilities of the VQA model. We match the full sentences consisting of already generated tokens, the candidates and their respective continuations with the attention-masked image $\hat{I}$. This information is used to steer the output of a pre-trained language model without requiring any training.

translation task, and additionally provide the input question and the answer prediction from the VQA model. The text generation is guided by the attention rollout result through the help of the VQA model itself. Token by token, this approach converts visual explanations, i.e. visual attention patterns, into natural language.

Pre-trained language model. Our ZS-A2T framework exploits a pre-trained large language model as its text generator. We condition the language model on two different inputs: (i) a task description, (ii) sample specific inputs (i.e. the question $\mathcal{Q}$, and the predicted answer a), and the already generated tokens. We then generate the translation in an autoregressive setup by feeding each predicted token back to the language model. It models the probability

$$p_\theta\big(x_t \mid \underbrace{x_0, \ldots, x_{l-1}}_{\text{task}}, \underbrace{x_l, \ldots, x_{l+s-1}}_{\text{sample } (\mathcal{Q}, a)}, \underbrace{x_{l+s}, \ldots, x_{t-1}}_{\text{autoregressive}}\big) \tag{1}$$

for a token x_t at step t, where l is the number of task tokens and s is the number of sample tokens. θ are the parameters of the pre-trained language model.

The probability of the next token x_t defined in Equation (1) is independent of the input image. As the pre-trained language model is conditioned on the question and answer, we expect it to predict tokens that fit the question and answer. However, candidate tokens that are likely from a grammatical or statistical point of view, e.g. those that occur frequently in the training text corpus, will be ranked highly. Thus, we argue that the language model predicts the correct tokens for the given question, image, and answer triplet, but not necessarily ranked correctly. To accurately describe the visual concepts, we rely on attention-controlled guiding of the VQA model.

Attention-controlled guiding. To generate natural language that actually corresponds to the prediction of the VQA model, we condition the text generation on a masked version of the input image. We obtain an attention-masked image by removing the image

information that was not relevant for generating the VQA model's answer prediction a. The removal of information from the image is determined by the internal attention of the VQA model. For this, we use attention rollout which will be described in the following.

Attention rollout. The attention of the given VQA model is aggregated based on attention rollout [1]. Here, we describe the rollout process for an encoder-decoder VQA model which interlinks the modalities with cross-attention. However, different techniques for obtaining an attention-masked image could be used.

To determine the answer relevant image parts, we record the self-attention scores A_l^{qq} (for the question tokens) and A_l^{ii} (for the image tokens). We also trace the cross-attention scores A_l^{qi}, which model the attention to the image tokens by the question tokens in layer l. The attention scores are saved for each layer l in the VQA model during inference. As attention heads differ in importance [43] we only use the maximum activation $\bar{A}_l = \max(A_l)$ across all heads. Following [1,7], we model how the attention flows throughout the network. We successively multiply the attention scores of a layer with the rolled out values R_l of the previous layer to obtain a final attention rollout map. Additionally, we add the map of the previous layer $\tilde{R}_{l-i}$ to model the residual connection. For self-attention, R_0 is initialized as a diagonal matrix since each token only contains its own information. For cross-attention, R_0^{qi} is initialized to zero, as the tokens have not been contextualized yet.

To roll out the self-attention for the question and visual tokens, we compute the following values for all layers which leverage self-attention before applying normalization,

$$R_l^{qq} = R_{l-i}^{qq} + \bar{A}_l^{qq} \cdot R_{l-i}^{qq}, \qquad R_l^{ii} = R_{l-i}^{ii} + \bar{A}_l^{ii} \cdot R_{l-i}^{ii}. \tag{2}$$

Additionally, we unroll the cross-attention to capture how the VQA model incorporates the visual information into the question representation:

$$R_l^{qi} = R_{l-i}^{qi} + {R_{l-i}^{ii}}^{\mathsf{T}} \cdot \hat{A}_l \cdot R_{l-i}^{qq}. \tag{3}$$

After modelling the residual connection, this captures how the attention rollout of both the question and image tokens is incorporated into the cross-attention. For layers that mix self- and cross-attention we also account for the cross-modal information mixed in at the previous layer:

$$R_l^{qi} = R_{l-i}^{qi} + \hat{A}_l^{qq} \cdot R_{l-i}^{qi}. \tag{4}$$

In the final layer, R_L^{qi} models the importance of each input question token and each visual token for the VQA model's prediction. We average R_L^{qi} over all question tokens to obtain the final aggregated attention map $\bar{R}_L^{qi}$. The threshold τ is applied to get a binary mask which is rescaled to the input image size to obtain M,

$$M = \begin{cases} 1, & \text{if } \tau < \bar{R}_L^{qi}, \\ 0, & \text{otherwise.} \end{cases} \tag{5}$$

The attention-masked image $\hat{I}$ is computed by eliminating irrelevant image parts $\hat{I} = M * I$, where $*$ is the element-wise multiplication.

Visually guided text generation. In the following, we describe the attention-controlled visual guiding of the language decoding process in detail. The language generation at

step t starts with the pre-trained language model predicting the probability distribution over all tokens in the vocabulary. We only consider the top-k predictions $c_1, \cdots, c_k$ for the token x_t and refer to those as *candidates*. This subset can be selected for two reasons: a) as argued above it is plausible that a sufficiently large language model conditioned on the question and answer ranks plausible words highly, and b) the weighted sum described in Equation (7) does not change for sufficiently large k as p_θ is a result of a softmax operation, yielding small values for non-top activations.

In contrast to previous works [41,40,37] which score incomplete sentences with an external image-text matching model (e.g. CLIP), we score a completed sentence with the same task model whose attention patterns we aim to convert to text. Our approach relies on the fact that image-text matching losses used by VQA models are commonly supervised with complete sentences instead of sentence fractions.

We let the language model complete the sentence for each candidate c_i separately based only on the predictions of the language model conditioned on the already generated tokens $x_{\leq t}$ and the candidate c_i. Each continuation is terminated at the first ".", such that the previously generated tokens, the candidate, and the continuation form a complete sentence $\hat{c}_i$. We do not include the question or answer as parts of the candidates. We use top-p sampling to filter the predictions of the language model to get a plausible continuation. Next, the completed sentences $\hat{c}_i$ are ranked by performing image-text matching.

The attention controlled guiding is executed by feeding the text input $\hat{c}_i$ and the masked image $\hat{I}$ as inputs to the VQA model's uni-modal encoders. In particular, we proceed with the contextualized embedding of the [CLS] token $ve(\hat{I})$ of the vision encoder and the corresponding [CLS] token $te(\hat{c}_i)$ of the text encoder. Both tokens are projected into a joint space using the linear maps g_v and g_t provided by the VQA model. This allows to measure the quality of the matching of the image-text pairs.

The matching quality $f(\hat{I}, \hat{c}_i)$ of a candidate sentence $\hat{c}_i$ with respect to the attention rollout-masked image $\hat{I}$ is determined by first computing the cosine similarity $\text{CosSim}(\cdot, \cdot)$ of all possible image-sentence matches,

$$f(\hat{I}, \hat{c}_i) = \frac{e^{\kappa \cdot \text{CosSim}\left(g_v(ve(\hat{I})), g_t(te(\hat{c}_i))\right)}}{\sum_{j \in 1, \ldots, k} e^{\kappa \cdot \text{CosSim}\left(g_v(ve(\hat{I})), g_t(te(\hat{c}_j))\right)}}. \tag{6}$$

Then, we apply a temperature κ in the softmax operation which affects the sharpness of the distribution. Very small values of κ would approximate a uniform distribution, largely disabling the influence of the visual guiding component. On the other hand, too large values may overemphasize highly salient visual concepts, at the cost of grammatical issues (due to overruling the language model). Thus, κ is manually chosen, such that the distribution of the predictions f roughly matches the language model p_θ.

To determine the next token, we compute a weighted sum of the prediction of the language model p_θ and the matching quality f obtained with the VQA model. Thus, at time step t, the next token x_t is computed according to:

$$x_t = \underset{i \in 1, \ldots, k}{\arg\max} \left\{ p_\theta \left(c_i \mid x_0, \ldots, x_{<t} \right) + \beta \cdot f\left(\hat{I}, \hat{c}_i\right) \right\}, \tag{7}$$

Setting	Framework↓	GQA-REX [14,8]					VQA-X [15]				
		B4	M	RL	C	S	B4	M	RL	C	S
Zero-shot	ZeroCap GPT-2 [41]*	1.4	4.6	12.3	16.9	5.3	0.7	4.7	14.0	5.8	2.0
	EPT GPT-2 [40]*	0.0	3.3	3.2	2.6	2.8	0.9	6.5	14.9	6.7	2.9
	MAGIC GPT-2 [37]*	2.3	10.8	18.8	41.1	18.8	1.0	8.8	19.3	10.6	7.1
	MAGIC OPT 6.7B [37]*	3.3	11.6	22.2	48.8	21.4	1.9	9.5	20.5	14.7	8.9
	Socratic Models OPT 6.7B [47]*	3.3	14.1	22.8	40.5	19.3	3.6	12.8	25.7	19.9	10.1
	ZS-A2T OPT 6.7B (ours)	**10.2**	**18.2**	**35.0**	**113.5**	**31.4**	**8.5**	**13.8**	**34.2**	**38.1**	**10.5**
Supervised	NLX-GPT [31] GPT-2	-	-	-	-	-	23.8	20.3	47.2	89.2	18.3
	VisualBert-REX [8] LSTM	54.6	39.2	78.6	464.2	46.8	-	-	-	-	-

Table 1: Zero-shot model attention to natural language translation results in the context of VQA evaluated on GQA-REX and VQA-X. We report Bleu-4 (B4), Meteor (M), Rouge-L (RL), Cider (C), and Spice (S). Higher is better for all reported metrics. State-of-the-art performances with textual explanation generation methods in the context of VQA are included for reference (supervised). *We adapted the related frameworks to the attention to natural language translation task, and also provided those methods with privileged access to the VQA ground-truth answers.

where β is a scalar weighting factor. After selecting the next token x_t, we append it to the original language model prompt and repeat the above process for generating tokens at step $t + 1$ until reaching a stopping criterion (either an [EOS] token, or a period).

4 Experiments

In this section, we describe our experimental setup, including the pre-trained models used in ZS-A2T, the datasets used, and the evaluation metrics. Experimental results generated by ZS-A2T for the zero-shot textual explanation task in the context of VQA are compared to four related training-free methods on the VQA-X and GQA-REX datasets. Although we do not claim that our generated texts are explanations, we do believe that these datasets are well suited for evaluation. Finally, we investigate individual components in our framework in detail, such as the attention-controlled guiding, and the language model prompting before providing qualitative results for ZS-A2T.

Experimental setup. We used the OPT [49] pre-trained language models for language generation. The transformer-based ALBEF [20] finetuned for VQA, served as our underlying VQA model. The weights for the projection layers are loaded from the non-finetuned, pre-trained ALBEF model, which was trained on an image-text matching objective. Both ALBEF variants were *not* trained to generate texts from the datasets that we test on (VQA-X and GQA-REX). The threshold for the attention rollout scores was set to $\tau = {}^{200}/_{256}$. For guiding, we determine the matching quality for the top-k candidate tokens as $k = 45$ and our continuations are sampled with $p = 0.15$. The guiding temperature is set to $\kappa = 100$. We set the scalar weighting factor between the

language model and the attention-controlled guiding to $\beta = 0.7$. To maintain a clean zero-shot protocol, we selected the hyperparameters on the validation split of VQA-X, which is disjoint from the subsets used for the final evaluation. We applied the same settings for GQA-REX without any further tuning.

Datasets. To evaluate all methods, we used the VQA-X [15] and GQA-REX [14,8] datasets for textual explanations in the context of VQA. VQA-X[3] extends a subset of non-trivial VQAv2 [13] questions with human-generated natural language explanations for the ground-truth answer. GQA-REX[4] contains explanations for a subset of the real-world visual reasoning question answering task posed in the GQA dataset. GQA-REX contains one explanation per question-image pair. As two of the related approaches have slow inference speeds (≈ 25s and ≈ 70s per sample [41,40] on an NVIDIA v100 GPU), we evaluate all methods on a subset of the test set containing 2000 samples. This is comparable in size to VQA-X's test set. Our approach only needs 7.2s per sample, and we compare inference speeds of all methods in Section C of the supplementary material.

Evaluation metrics. Unless stated otherwise, all models are evaluated in the zero-shot setting, i.e. without any training, on the test sets of the respective datasets using commonly reported natural language generation metrics similar to [15]. In particular, we report Bleu-4 (B4) [26], Meteor (M) [5], Rouge-L (RL) [23], Cider (C) [42], and Spice (S) [3] scores. These metrics aim to capture the semantic overlap between sentences by measuring (modified) precision [26] and recall [26,5,23] of n-grams. Additionally, generalizations of n-grams like stemming [5,42], measures of sentence fragmentation [5] or tf-idf weighting [42] are applied to better match human judgement of sentence similarity.

4.1   Comparing to Related Frameworks

To evaluate the quality of the generated translations, we show experimental results on VQA-X and GQA-REX. As there are, to the best of our knowledge, no related works that translate VQA attention patterns into natural language, we adapted a number of zero-shot image captioning methods. For fair comparison, we modified all related works by prompting them with the question and ground-truth answer, as they do not have a dedicated VQA module. This favors the related works, as our framework may base its translation on a wrongly predicted answer. Additionally, we show other common evaluation schemes in Section B of the supplementary material.

First, we compare to the zero-shot image captioning works [41,37,40,47]. For a fair comparison, we adapted the two stronger models (MAGIC and Socratic Models) to operate with the same language model as the one used in our framework (OPT 6.7B). ZS-A2T outperforms all four related approaches for zero-shot translation of attention patterns into natural language by wide margins (see Table 1). Interestingly, the optimization-based approaches ZeroCap and EPT show relatively weak performances (on VQA-X more so than on GQA-REX). This suggests that longer, more complicated prompts (in contrast to the ones used for image captioning) make it hard to optimize helpful starting parameters for the next token prediction (ZeroCap) or next sentence prediction (EPT).

---

[3] Licensed under the BSD-2 license.

[4] Licensed under the MIT license.

MAGIC gives stronger results, but it is still largely outperformed by our ZS-A2T framework. Similarly, Socratic Models (SMs), the strongest of the adapted related works, is outperformed in all metrics. The generated sentences by ZS-A2T exhibit greater word-by-word overlap with the ground-truth references than those of the related approaches. In particular, this is indicated by the n-gram-based metric Bleu-4 for which ZS-A2T obtains a score of 8.5 compared to 3.6 for SMs on VQA-X. Similarly, ZS-A2T is stronger than SMs in terms of Meteor with 18.2 compared to 14.1 on GQA-REX and 13.8 compared to 12.8 on VQA-X. The same pattern holds true for all other metrics.

For context, we also list results with the recently published supervised models NLX-GPT[31] for VQA-X and VisualBert-REX [8] for GQA-REX. They both employ joint multi-modal transformer models to predict the sentences of the respective datasets. Whilst not strictly comparable, since we translate only the question relevant image regions, it is still interesting to note that our 5-shot variant (c.f. Table 5) significantly shrinks the gap to the supervised models which used 31k VQA-X and 128k GQA-REX training samples respectively, whereas our model does not require any training or just a few in-context examples.

4.2   Ablation Studies on Guiding Inputs

In this section, we study the impact of using different input images in our attention-controlled guiding (see Section 3) as well as the effect of guiding with completed sentences (in contrast to incomplete sentences). In addition to this, we provide results for using different visual explanation methods in ZS-A2T, i.e. for visual attribution and perturbation methods. We show additional ablations for the attention thresholding parameter θ and the guiding temperature κ in Section D of the supplementary material.

Influence of attention masking. Table 2 shows the effect of using an attention-masked image in the visual guiding. Restricting the guiding of the language generation to the attention-masked image improves the language generation in terms of the Bleu-4, Meteor, Rouge-L, and Cider metrics. Interestingly, the Spice metric is slightly higher (10.8 vs. 10.5) when using the full image for guiding.

Guiding Input	B4	M	RL	C	S
Full Image	8.1	13.7	34.1	37.6	**10.8**
No Continuation	6.2	12.5	31.1	28.2	9.4
ZS-A2T (Rel. Masking + Cont.)	**8.5**	**13.8**	**34.2**	**38.1**	10.5

Table 2: Ablating the guiding input and text continuation on VQA-X [15]. Our ZS-A2T model uses an attention-masked image, obtained from attention rollout [1], and text continuation.

Influence of using text continuations. In Table 2, we also investigate the effect of using the language model to generate text continuations, so that the guiding component can judge completed sentences. Using the continuations increases all metrics, e.g. Cider from 28.2 to 38.1. We hypothesize that this happens for two reasons. First, it reduces the distribution shift between the contrastive image-text matching training of the VQA model. Second, it allows the guiding to judge whether a greedy selection of the visually grounded token at step t may lead to a completed sentence that is not visually supported.

Different visual explanation methods. In addition to using attention rollout to determine relevant image parts (described in Section 3), we use five other attribution methods

Attribution Method	B4	M	RL	C	S
Att. GradCAM [33,20]	7.5	13.3	32.8	33.0	9.2
EigenGradCAM [24]	7.5	13.1	32.8	32.3	9.2
XGradCAM [12]	6.7	12.4	32.1	30.1	9.1
GradCAMElementwise [28]	7.2	12.8	32.5	30.7	9.0
HiResCAM [9]	7.0	12.7	32.5	31.0	9.0
RISE [27]	7.5	13.3	33.1	33.5	9.9
ZS-A2T (attention rollout [1])	**8.5**	**13.8**	**34.2**	**38.1**	**10.5**

LM (#Params)	B4	M	RL	C	S
GPT-2 (125M)	3.6	11.0	26.6	19.8	7.7
OPT (125M)	3.4	10.7	26.5	18.5	7.1
OPT (350M)	3.9	11.6	27.9	20.2	7.9
OPT (1.3B)	7.1	13.0	32.8	28.9	9.2
OPT (2.7B)	7.1	13.3	32.5	31.0	10.1
OPT (6.7B)	**8.5**	**13.8**	**34.2**	**38.1**	**10.5**

Table 3: Ablating different attribution and perturbation methods (left) and type and size of the pre-trained language models (right) in our ZS-A2T framework on VQA-X [15].

and the perturbation-based visual explanation method RISE [27]. Table 3 (left) showcases that our approach can handle conceptually different visual explanation methods.

The backpropagation-based techniques from the GradCAM family build on the attention probabilities and their respective gradients [33,20] from the same layer of the VQA model ($L = 11$). Thus, the generated sentences using the different methods Eigen-GradCAM [24], XGradCAM [12], GradCAMElementwise [28] and HiResCAM [9] are of very similar quality in terms of the NLG metrics. Overall, their scores are lower than the scores we obtain when using attention rollout [1]

We additionally show results with the input perturbation-based method RISE [27]. It filters the images applied to the given VQA model and evaluates the VQA model multiple times whilst randomly masking parts of the image. The final importance map is obtained by summing the random masks weighted by the predicted class probability. This indicates the parts in the input that are salient for the VQA models' prediction. We find that the texts generated for RISE are slightly worse than those obtained with attention rollout based attribution method (e.g. Bleu-4 7.5 vs. 8.5 (ours)).

Overall, the visual explanation methods' different abilities to identify the correct relevant image regions is reflected in the quality of the translations to natural language. Attention rollout [7] generally outperforms GradCAM [32] in identifying relevant image regions and the same pattern is found in the quality of the translations. Thus, we conclude that our framework is not tied to a specific visual explanation method, and allows the drop-in replacement of different visual explanation methods for guiding the language generation.

4.3   Language Models

Our framework can be used with different language models without any changes in the setup (c.f. Table 3 (right)). Here, we analyze the performance of ZS-A2T for different language models. We demonstrate that our framework even outperforms other related approaches when using a pre-trained GPT-2 language model. For example, it achieves a Rouge-L value of 26.6 on VQA-X, whereas the previous best model with a GPT-2 backbone (MAGIC) only achieves 19.3 with the same language model. Furthermore, we clearly outperform MAGIC on the n-gram metric Bleu-4 (3.6 vs. 1.0), as well as

on Meteor, Rouge-L, Cider, and SPICE (see Table 1). This can be attributed to our temperature κ, which allows for better balancing of the two terms in Equation (7).

Next, we analyze the impact of the size of the pre-trained language models. Our ZS-A2T framework does benefit from larger, more powerful language models. The performance increase with larger size is consistent across all metrics, e.g. Spice goes from 7.1 to 9.2 when using the OPT model with 125M vs. 1.3B parameters. Using the 6.7B model boosts the Spice performance to 10.5. This suggests that high-quality candidate proposals are beneficial for the generated sentences. Additionally, it also showcases a benefit of our training-free approach: Large or newer language models can be swapped in without additional cost.

4.4   Prompt Ablations

Here, we analyze the effect of different input prompts for the pre-trained language model on the generated text outputs (c.f. Table 4).

Unsurprisingly, providing no task description (row 1) in the input prompt gives the worst outputs. Removing the task description ("Answer and explain:") yields significantly worse results with Bleu-4 decreasing from 8.5 to 3.7, suggesting that the task description is crucial for performance.

Inputting the task in a more structured way ("Question: $\langle q \rangle$? Answer: $\langle a \rangle$. Explanation:") increases the language quality only slightly compared to using no task description. However, the same structured prompt extended by

Prompt	B4	M	RL	C	S
$\langle q \rangle$? the answer is $\langle a \rangle$ because	3.7	10.3	25.7	21.6	8.4
Q: $\langle q \rangle$? A: $\langle a \rangle$. E:	3.5	11.0	23.2	22.6	9.6
Q: $\langle q \rangle$\n? A: $\langle x \rangle$\n. E:	4.2	11.2	23.7	23.0	9.6
Explain the A: $\langle q \rangle$? The A is $\langle a \rangle$ because	6.9	13.0	32.3	31.7	9.9
A and Explain: $\langle q \rangle$?\n The A is $\langle a \rangle$ because	7.6	13.5	32.9	35.8	**10.6**
A and Explain: $\langle q \rangle$? The A is $\langle a \rangle$ because	**8.5**	**13.8**	**34.2**	**38.1**	10.5

Table 4: Ablating different input prompts on VQA-X [15]. $\langle q \rangle$ and $\langle a \rangle$ are placeholders for the question and answer. Q denotes "Question", A "Answer", and E "Explanation". \n denotes a new line. We show the full prompt templates in Section A of the supplementary material.

newline characters \n shows increased or similar metric values for all metrics. A further improvement is achieved by using a meaningful task description ("Explain the answer:") in the input prompt. We used the best prompt ("Answer and explain:") in ZS-A2T.

Impact of in-context generation (n-shot prompting). We analyze the effect of prefixing the context with complete examples of questions, answers, and their respective explanations. This enables the language model to better understand the task, as it can see some examples before generating text [6].

We experiment with up to $n = 1$ and $n = 5$ randomly sampled examples from the training set of the respective dataset which are prepended to the context (see Table 5). The full prompts for this setup are included in Section E of the supplementary material.

Using a single example improves the language quality already. For $n = 1$, the natural language generation scores are on average 7.2% higher than

Model	n	B4	M	RL	C	S
ZS-A2T	0	8.5	13.8	34.2	38.1	10.5
ZS-A2T	1	9.8	14.5	34.8	42.7	11.7
ZS-A2T	5	11.9	15.3	37.5	49.6	12.4

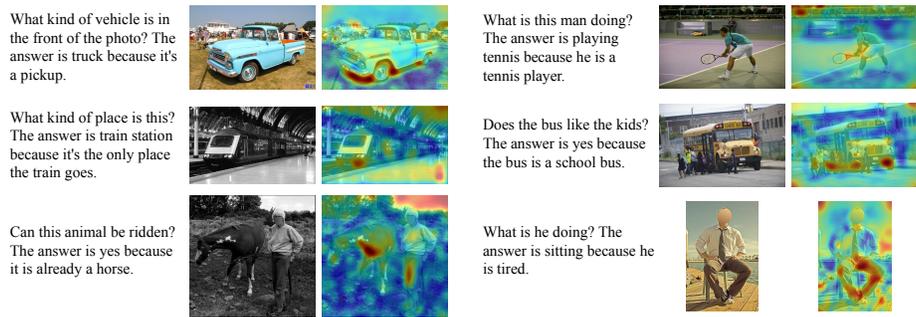Table 5: In-context learning with n examples on the VQA-X dataset.

Fig. 3: Qualitative results for ZS-A2T on VQA-X, showing the input image, question, attention rollout [1] output, and generated text. Red indicates higher relevance. The generated text indeed mentions the visual concepts detected by the VQA model. The face regions have been deliberately occluded in this figure.

for $n = 0$. By just prefixing five in-context learning examples, the generation quality increases on average by 19.3% over providing no examples. Qualitatively, we find that the language model does not merely copy or modify the texts of the examples when the question and/or answer match, but instead allows to adjust the candidate predictions for the generated sentences accordingly. Additionally, the generated texts better match the language biases in the datasets (e.g. on VQA-X many samples start with "the" or "there").

4.5   Qualitative Results

We provide qualitative examples for natural language translations generated with ZS-A2T for VQA-X in Figure 3. The attention rollout map is superimposed onto the original image. For each of the examples we show the question, predicted answer and generated sentence that translates the attention patters into natural language. We can observe that attention rollout, used for the attention-controlled guiding, selects relevant image regions that plausibly correspond to the input question. In the bottom left example the main attention is directed towards the *horse*. The generated sentences are fluent (due to the pre-trained language model) and they refer to visual elements (due to the image-text matching). Moreover, the framework can argue with common sense, i.e. it states that *a train station is the only place the train goes*. It combines this prior knowledge from the pre-trained language model with visual concepts detected by the VQA model, such as identifying that the masked image shows *a school bus*. The capability to mention visual elements stems from the attention controlled guiding of the language generation. Additionally, the translations accurately describe the part of the image that the VQA model used to answer the question (e.g. the pickup truck in the top left example).

An observed failure pattern can be seen in the bottom-right example in Figure 3. It seems very plausible that the person would be *tired*. However, this sentence does not refer to any visual information but instead uses common sense to explain the answer

to the input question. In conclusion, the qualitative results in Figure 3 show that ZS-A2T indeed generates text which mentions visual information contained in the attention patterns that are extracted from the VQA model. Additionally, the generated sentences are overall grammatically correct and fluent.

5 Limitations

Our proposed framework translates the internal attention of a VQA model into natural language. As no datasets exist specifically for this task, we chose to automatically evaluate our text translations of attention maps on textual explanation datasets. Due to the inherent task differences, we do not expect our attention translations to perfectly match the ground-truth explanations (e.g. in terms of writing style), explaining part of the performance gap compared to supervised upper bounds in Table 1. Further research into zero-shot translation methods and the creation of attention translation datasets will be important to better understand attention-based models using natural language.

Our *translation* approach relies on (pre-)trained language models. As a result, the faithfulness of the generated text with respect to the task model is hard to quantify. To address this, we use an attention-controlled visual guiding component for aligning the text generation with the VQA model.

Furthermore, we have only considered the ALBEF VQA model. However, our approach could easily be extended to other models such as LXMERT [39] or ViLT [19].

Our attention-controlled guiding outperforms guiding with the full image by only a slight margin (c.f. Table 2). This could be due to wrong internal reasoning of the task model and attention rollout not identifying the areas causing this. This should be addressed in future work, e.g. by using attention-perturbation to understand the importance of image patches.

Lastly, we hypothesize that the relatively small changes in the ablation studies might be due to the LLM already predicting a common sense translation for the question and answer.

6 Conclusion

In this work, we introduce ZS-A2T, a zero-shot framework for translating the aggregated attention in a VQA model to natural language. In particular, the language generation is guided using the VQA model itself, by means of its internal attention combined with its image-text matching capabilities for selecting word candidates in the language generation. Our proposed method does not require any training and can be flexibly used together with any language model to translate the visual attribution output for an attribution method of choice. Our framework outperforms zero-shot image captioning baselines on textual explanation datasets in the context of VQA.

References

1. Abnar, S., Zuidema, W.: Quantifying attention flow in transformers. In: ACL (2020)
2. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C.L., Parikh, D., Batra, D.: Vqa: Visual question answering. International Journal of Computer Vision (2015)
3. Anderson, P., Fernando, B., Johnson, M., Gould, S.: Spice: Semantic propositional image caption evaluation. In: ECCV (2016)
4. Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.R., Samek, W.: On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PloS one (2015)
5. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: IEEvaluation@ACL (2005)
6. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. Advances in neural information processing systems **33**, 1877–1901 (2020)
7. Chefer, H., Gur, S., Wolf, L.: Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In: ICCV (2021)
8. Chen, S., Zhao, Q.: Rex: Reasoning-aware and grounded explanation. In: CVPR (2022)
9. Draelos, R.L., Carin, L.: Use hirescam instead of grad-cam for faithful explanations of convolutional neural networks (2020)
10. Fong, R., Patrick, M., Vedaldi, A.: Understanding deep networks via extremal perturbations and smooth masks. In: ICCV (2019)
11. Fong, R.C., Vedaldi, A.: Interpretable explanations of black boxes by meaningful perturbation. In: ICCV (2017)
12. Fu, R., Hu, Q., Dong, X., Guo, Y., Gao, Y., Li, B.: Axiom-based grad-cam: Towards accurate visualization and explanation of cnns. In: BMVC (2020)
13. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., Parikh, D.: Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. In: CVPR (2017)
14. Hudson, D.A., Manning, C.D.: GQA: A new dataset for real-world visual reasoning and compositional question answering. In: CVPR (2019)
15. Huk Park, D., Anne Hendricks, L., Akata, Z., Rohrbach, A., Schiele, B., Darrell, T., Rohrbach, M.: Multimodal explanations: Justifying decisions and pointing to the evidence. In: CVPR (2018)
16. Jain, S., Wallace, B.C.: Attention is not explanation. In: North American Chapter of the Association for Computational Linguistics (2019)
17. Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Zitnick, C.L., Girshick, R.B.: Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In: CVPR (2017)
18. Kim, S.S., Meister, N., Ramaswamy, V.V., Fong, R., Russakovsky, O.: Hive: evaluating the human interpretability of visual explanations. In: ECCV (2022)
19. Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning (2021)
20. Li, J., Selvaraju, R.R., Gotmare, A.D., Joty, S.R., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021)
21. Li, Q., Tao, Q., Joty, S., Cai, J., Luo, J.: VQA-E: Explaining, elaborating, and enhancing your answers for visual questions. In: ECCV (2018)
22. Li, W., Zhu, L., Wen, L., Yang, Y.: Decap: Decoding CLIP latents for zero-shot captioning via text-only training. In: The Eleventh International Conference on Learning Representations (2023)
23. Lin, C.Y.: Rouge: A package for automatic evaluation of summaries. In: ACL (2004)

24. Muhammad, M.B., Yeasin, M.: Eigen-cam: Class activation map using principal components. In: IJCNN (2020)
25. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. NeurIPS **35**, 27730–27744 (2022)
26. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method and for automatic and evaluation of machine and translation. In: ACL (2002)
27. Petsiuk, V., Das, A., Saenko, K.: Rise: Randomized input sampling for explanation of black-box models. In: BMVC (2018)
28. Pillai, V., Pirsiavash, H.: Explainable models with consistent interpretations. In: AAAI (2021)
29. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML (2021)
30. Salewski, L., Koepke, A.S., Lensch, H., Akata, Z.: Clevr-x: A visual reasoning dataset for natural language explanations. In: International Workshop on Extending Explainable AI Beyond Deep Models and Classifiers (2022)
31. Sammani, F., Mukherjee, T., Deligiannis, N.: Nlx-gpt: A model for natural language explanations in vision and vision-language tasks. In: CVPR (2022)
32. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad-CAM: Visual explanations from deep networks via gradient-based localization. International Journal of Computer Vision (2019)
33. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: ICCV (2017)
34. Simonyan, K., Vedaldi, A., Zisserman, A.: Deep inside convolutional networks: Visualising image classification models and saliency maps. In: ICLR workshop (2013)
35. Smilkov, D., Thorat, N., Kim, B., Viégas, F., Wattenberg, M.: Smoothgrad: removing noise by adding noise. arXiv:1706.03825 (2017)
36. Springenberg, J.T., Dosovitskiy, A., Brox, T., Riedmiller, M.: Striving for simplicity: The all convolutional net. In: ICLR (workshop track) (2015)
37. Su, Y., Lan, T., Liu, Y., Liu, F., Yogatama, D., Wang, Y., Kong, L., Collier, N.: Language models can see: Plugging visual controls in text generation. arXiv:2205.02655 (2022)
38. Su, Y., Lan, T., Wang, Y., Yogatama, D., Kong, L., Collier, N.: A contrastive framework for neural text generation. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022)
39. Tan, H., Bansal, M.: Lxmert: Learning cross-modality encoder representations from transformers. arXiv:1908.07490 (2019)
40. Tewel, Y., Shalev, Y., Nadler, R., Schwartz, I., Wolf, L.: Zero-shot video captioning with evolving pseudo-tokens. arXiv:2207.11100 (2022)
41. Tewel, Y., Shalev, Y., Schwartz, I., Wolf, L.: Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In: CVPR (2022)
42. Vedantam, R., Zitnick, C.L., Parikh, D.: Cider: Consensus-based image description evaluation. CVPR (2015)
43. Voita, E., Talbot, D., Moiseev, F., Sennrich, R., Titov, I.: Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In: ACL (2019)
44. Wang, J., Zhang, Y., Yan, M., Zhang, J.C., Sang, J.: Zero-shot image captioning by anchor-augmented vision-language space alignment. arXiv:2211.07275 (2022)
45. Wiegreffe, S., Pinter, Y.: Attention is not not explanation. In: Conference on Empirical Methods in Natural Language Processing (2019)
46. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: ECCV (2014)

47. Zeng, A., Attarian, M., brian ichter, Choromanski, K.M., Wong, A., Welker, S., Tombari, F., Purohit, A., Ryoo, M.S., Sindhwani, V., Lee, J., Vanhoucke, V., Florence, P.: Socratic models: Composing zero-shot multimodal reasoning with language. In: ICLR (2023)
48. Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S.: Top-down neural attention by excitation backprop. International Journal of Computer Vision (2018)
49. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P.S., Sridhar, A., Wang, T., Zettlemoyer, L.: Opt: Open pre-trained transformer language models. arXiv:2205.01068 (2022)
50. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., Torralba, A.: Learning deep features for discriminative localization. In: CVPR (2016)