A Appendix

A.1 Remarks on executed benchmarks

We executed all benchmarks faithfully and to the best of our knowledge. The selection of compared methods was made to be rather diverse and obtain a good overview in this field of research. In particular, with regards to the multi-modal transformer scaling behavior, as there are in fact no such studies for AR models yet to compare to. It is possible, for all methods, that there are still improvements we missed in quality as well as performance. However, we see the optimizations of other methods to multi-modal AR transformer models as a research direction on its own.

Chefer. The integration of Chefer was straightforward. As it can be derived by the visualizations, there are noticeable artifacts, particularly on the edges of images. In this work the underlying transformer model was MAGMA, which is finetuned using sequential adapters. It is possible that this, or the multi-modal AR nature itself, is the cause for these artifacts. We did not further investigate to what extent the adapters are to be particularly integrated in the attribute accumulation of Chefer. Also notice that ATMAN often has similar, but not as severe, artifacts.

IxG, IG and GradCAM. The methods IxG, IG, and (guided) GradCAM failed completely from the quality perspective in recall when applied to pixel granularity (c.f. Tab. 2). Those are the only ones directly applicable to the image domain, and thus also included the vision encoder in the backward pass. We did not further investigate or fine-tune evaluations to any method. For better comparison we therefore only present the token-based comparisons. All methods are evaluated with the same metrics and therewith give us a reasonable performance comparison without additional customization or configuration.

Details on Results. For a fair comparison, all experiments were executed on a single GPU, as scaling naturally extends all methods. We also want to highlight that we did not optimize the methods for performance further but rather adopted the repositories as they were. The memory inefficiency of gradient-based methods arises from the backward pass. A maximal memory performant representative is the Single-Layer-Attribution method IxG, which only computes partial derivatives on the input with respect to the loss. Even this approach increases the memory requirement beyond an additional 50% and fails for the scaling experiments up to 34B.

In Fig. 5 we ran Chefer with a full backward pass. We adopted this to the minimum amount of gradients (we saw) possible and plot the full scaling benchmark below in Fig 7[3]. The key message remains the same. With the given IxG argument, we do not see much potential in improving memory consumption further.

The methods IxG, IG and GradCam are integrated using the library Captum [15]. We expect them to be implemented as performant as possible. IntegratedGradients is a perturbation method on the input, integrating changes over the gradients. The implementation at hand vastly runs OOM. Finally GradCam is a method specialized on CNN networks and therefore does not work for text only (or varying sequence lengths). It requires the least amount of resources but also produces poor results, without further investigations.

ATMAN Parallelizability. As a final remark, we want to recall again that the runtime measured in the sequential execution can be drastically reduced due to its parallelizability, i.p., as it only requires forward passes. For sequence length 1024, we measured 1024 iterations in order to explain each token. However note that ATMAN can also be applied to only parts or chunks of the sequence (c.f. Sec. A.8), in contrast to gradient methods. Moreover, all tokens to explain can be computed entirely in parallel. In a cluster deployment, these can be distributed amongst all available workers. On top, it can be divided by the available batch size and true pipeline-parallelism (c.f. Fig. 6).

[3] Setting *requires_grad=False* to every but the attention tensors.

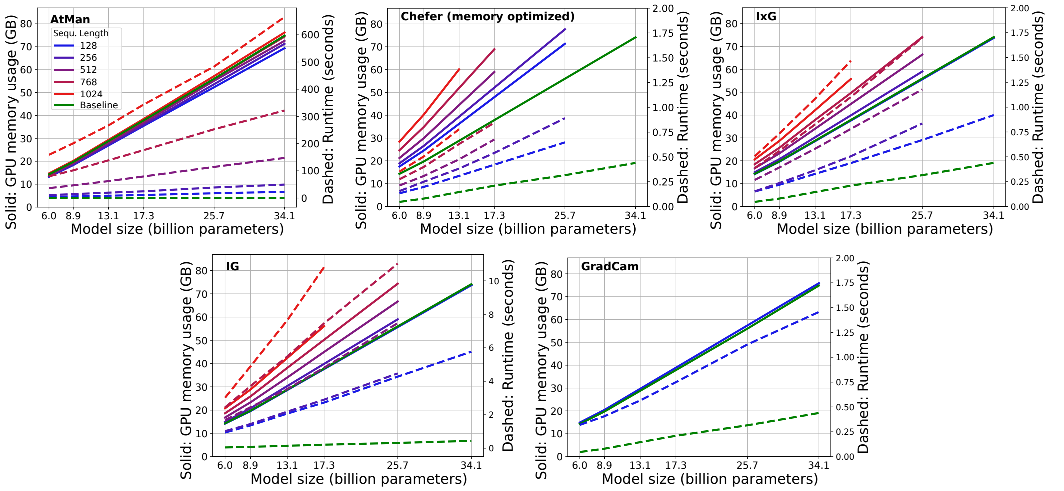
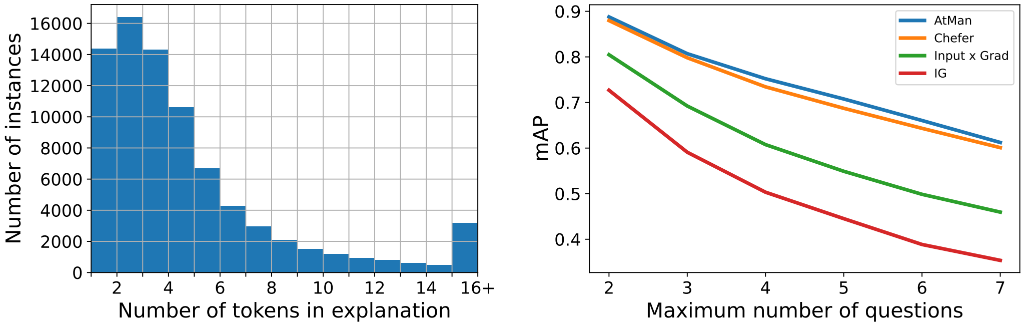


Figure 7: Performance comparison of the explainability methods over various model sizes (x-axis) executed on a single 80GB memory GPU, with fixed batch size 1. Solid lines refer to the GPU memory consumption in GB (left y-axis). Dashed lines refer to the runtime in seconds (right y-axis). Colors indicate experiments on varying input sequence lengths. As baseline (green) a plain forward pass with a sequence length of 1024 is measured. Note that GradCAM can only be applied to the vision domain, it is therefore fixed to 144 tokens. Note that it already consumes as much memory as a forward pass of 1024 tokens. (Best viewed in color.)



(a) Histogram of explanation token length.

(b) mAP for all methods, grouped by number of question/ answer pairs in the dataset.

Figure 8: Further evaluations on the SQuAD dataset. (Best viewed in color.)

A.2 Detailed SQuAD Evaluations

This sections gives more detailed statistics on the scores presented in Tab. 1. First Fig. 8a is the histogram of the token lengths of all explanations. Fig. 8b is the mAP score for all methods on the entire dataset, grouped by the number of questions occuring per instance.

A.3 Detailed OpenImages Evaluations

This section gives more detailed statistics on the scores presented in Tab. 1. Fig. 9 is the histogram of the fraction of label coverage on all images. Fig. 10a and 10b are boxplots for all methods on the entire dataset, for mean average precision as well as recall. Note that the methods IG, IxG and GradCam can be evaluated on image-pixel level, contrary to AtMan and Chefer. However their performances (i.p. wrt. recall) significantly degenerate, c.f. Tab. 2, Fig. 15.

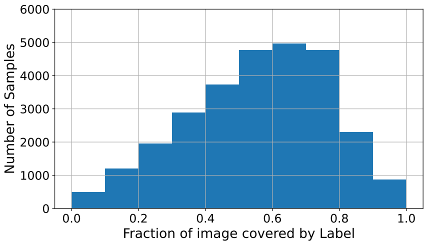
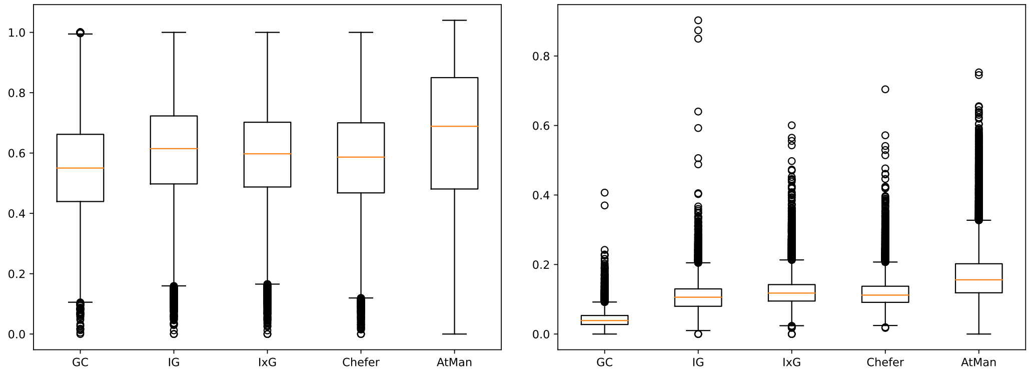


Figure 9: Histogram of percentage of label coverage of the images.



(a) mAP Boxplot for all methods of all images.

(b) mAR Boxplot for all methods of all images.

Figure 10: Further evaluations on the OpenImages dataset. (Best viewed in color.)

A.4 Discussion of Cosine Embedding Similarity and Parameter values

We fixed the parameter $\kappa = 0.7$ of Eq. 6 and $f = 0.9$ of Eq. 4 throughout this work. They were empirically concluded by running a line sweep on a randomly sampled subset of the OpenImages dataset once, c.f. Fig. 11. Note that pure token-masking ("mask") always leads to worse performance compared to our more subtle concept-suppression operations. We use the same parameters throughout this work, however, parameters may be optimizable depending on the use case and dataset.

In Fig. 12a and 12b we compare the mean average precision and recall scores for OpenImages for both variants, with and without correlated token suppression (to threshold κ). Clearly the latter outperforms single token suppression.

The following Fig. 13 shows visually the effect on weak image segmentation when correlated suppression of tokens is activated, or when using single token suppression only. Notice how single token only occasionally hits the label, and often marks a token at the edge. This gives us a reason to believe that entropy is accumulated around such edges during layer wise processing. This effect (on these images) completely vanishes with correlated suppression of tokens.

A.5 Variation Discussion of the method

Note that the results of Eq. 4 are directly passed to a softmax operation. The softmax of a vector $\mathbf{z}$ is defined as

$$\text{softmax}(\mathbf{z})_i = e^{z_i} / \sum_j e^{z_j}.$$

In particular, the entries $z_k = 0$ and $z_l = -\infty$ will yield to the results $\text{softmax}(\mathbf{z})_k = 1/\sum_j e^{z_j}$ and $\text{softmax}(\mathbf{z})_l = 0$. So one might argue as follows: If we intent to suppress the entropy of token i, we do not want to multiply it by a factor 0, but rather subtract $-\infty$ of it. I.e. we propose the modification

$$(H)_{u,*,*} = (H)_{u,*,*} + \log(f). \tag{7}$$

The only problem with this Eq. 7 is, that it skews the cosine neighborhood factors. While we experienced this working more naturally in principle, for hand-crafted factors, we did not get best

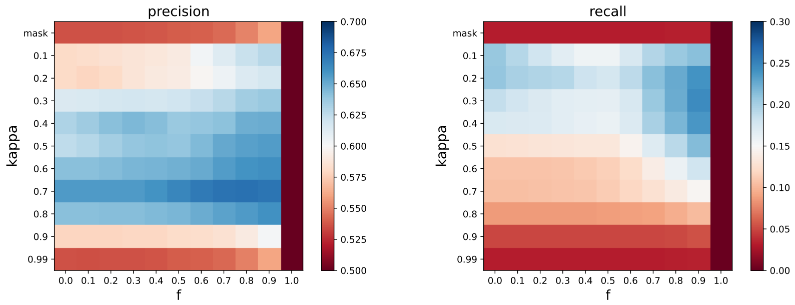
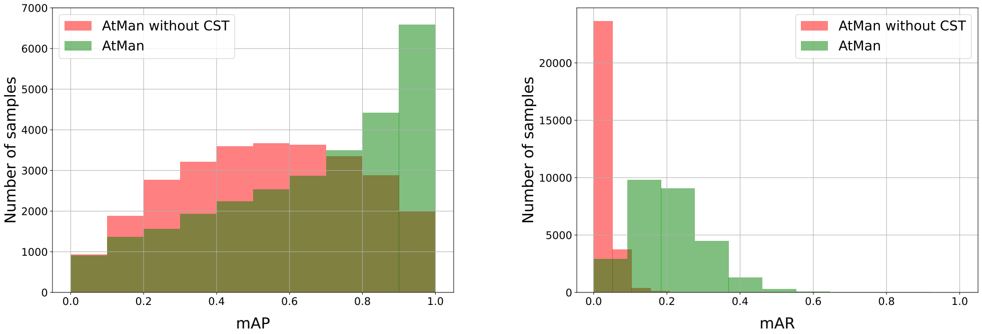


Figure 11: Parameter sweep for Eq. 6. It can be observed that indeed cosine similarity (y-axis) along with more subtle modification (x-axis) outperforms other configurations, i.p. row "mask" that entirely masks single tokens (i.e. multiplies $1 - f$). This result indicates that indeed conceptual entropy is scattered across tokens, that need to be suppressed at once.



(a)

(b)

Figure 12: Histogram metric evaluation on OpenImages for AtMan with (green) and without (orange) correlated suppression of tokens. (Best viewed in color.)

performance in combination with Eq. 6. In the following Fig. 14a and 14b, we show analogous evaluations to Fig. 12a and 12b. It is in particular interesting that the mode without correlated tokens slightly improves, while the one with slightly decreases in scores, for both metrics.

A.6 Artifacts and failure modes

Merged into A.12 (kept here for reference).

A.7 Qualitative comparison weak image segmentation

In the following Fig. 15 we give several examples for better comparison between the methods on the task of weak image segmentation. To generate the explanations, we prompt the model with "<Image> This is a picture of " and extract the scores towards the next target tokens as described with Eq. 5 for AtMan. For multiple target tokens, these results are averaged. In the same fashion, but with an additional backpropagation towards the next target token, we derive the explanations for Chefer and the other gradient methods.

A.8 Application to document q/a

In Fig. 16 we apply AtMan on a larger context of around 500 tokens paragraph wise. The Context is first split into chunks by the delimiter tokens of ".", ",", "\n" and " and". Then iteratively each chunk is evaluated by prompting in the fashion "$\{Context\}$ Q:$\{Question\}$ A: " and the cross entropy extracted towards the target tokens, suppressing the entire chunk at once, as described in Sec. 3. It can be observed that the correct paragraphs are highlighted for the given questions and expected targets.

Table 2: OpenImages (OI) evaluations on pixel granularity.

	IxG	IG	GC
mAP	38.0	46.1	56.7
mAP$_{IQ}$	34.1	45.2	60.4
mAR	0.2	0.3	0.1
mAR$_{IQ}$	0.1	0.1	0.1

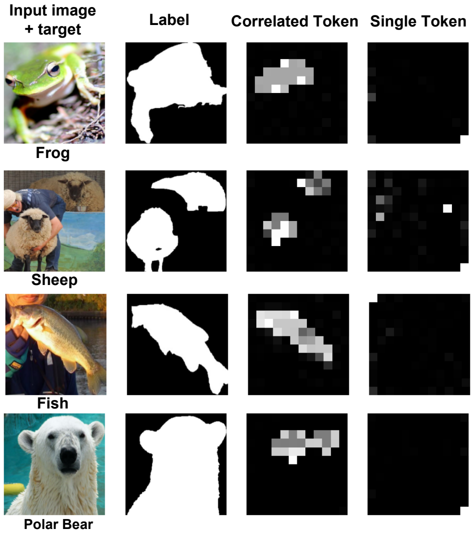


Figure 13: Example images showing the effect of correlated suppression as described in Correlated Token Attention Manipulation Sec. 3.3 and Single Token Attention Manipulation Sec. 3.2. (Best viewed in color.)

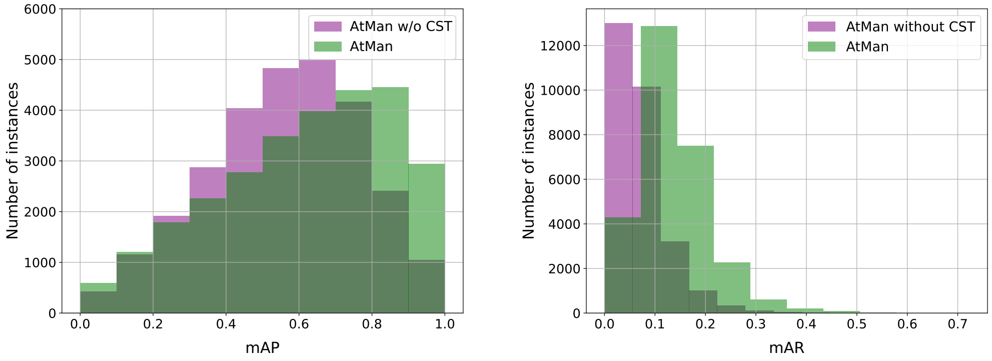
In particular, one can observe the models interpretation, like the mapping of formats or of states to countries. Note in particular that it is not fooled by questions not answered by the text (last row).

A.9 Attention Rollout, Flow, other Perturbation method

Attention Flow constructs a graph representing the flow of attention through the model. Unfortunately this approach is computationally too intense for evaluation on large language models; Chefer et al. (2021) already ran into issues on smaller BERT models. LIME [24] and SHAP [19] are perturbation methods closely related to ATMAN. However they result in several 100.000 forward trials which is inpractical for use with large language models.

Finally, we ran additional experiments for Attention Rollout [1]. Note that the authors stress its design for encoder classification architectures, while we explore the applicability to generative decoder models in this work. Further, Rollout is by default class-agnostic, in contrast to ATMAN or the studied gradient methods. We adopted Attention Rollout by appending the target to the prompt, and only aggregating attentions over those last target sequence positions. Note that in contrast to the other studied methods, Rollout disregards the output distribution.

We observed large artifacts on the explanations at the positions of the last 2 image patches; apparently most attention aggregation take place at those positions, which may be due to the causal mask. On OpenImages our adoption of Rollout achieves a mAP score of 43.6 and a mAR of 1.3 (c.f. AtMan with 65.5 and 13.7). Scores do increase when removing the last image patch, however, this also entirely erases explanation on this patch. On Squad we achieved mAP of 23.1 and a mAR of 53.4 (AtMan: 73.7, 93.4).



(a) (b)

Figure 14: Histogram of the evaluation metrics on OpenImages for AtMan with and without correlated token suppression. (Best viewed in color.)

Table 3: Evaluation of XAI on BLIP model.

	Chefer	AtMan
mAP	59.7	64.6
mAP$_{IQ}$	60.3	66.1
mAR	20.2	26.4
mAR$_{IQ}$	19.4	28.4

A.10 Demarcation to other architectures, BLIP experiments

To answer the question of *"why we think that AtMan could be generalized to other (generative) multimodal Transformers"*, consider the following arguments.

We conduct for the first time experiments on generative decoder models. The underlying MAGMA model extends the standard decoder transformer architecture, having causal attention masks in place, by additional sequential adapters that could potentially shift distributions. However, the proposed attention manipulation still leads to a robust perturbation method — in text-only as well as image-text modality. We thus conjecture that the skip connections lead to a somewhat consistent entropy flow throughout the network. OFA ($<$ 1B) and BLIP rely on standard encoder-decoder architectures —even without adapters— and the same skip connections. Furthermore, both models process the input image in a similar fashion to obtain "image tokens" forwarded throughout the transformer at the same positions.

BLIP, in contrast to OFA, separates the processing of the image, the prompt, and the answer generation into 3 different models —standard encoders, only the answer generation model being a decoder. The different outputs are cross-attended. We apply AtMan only on the vision-encoder model.

For comparison, we furthermore evaluate Chefer on BLIP, its attention aggregation being only applied to the vision encoder as well. Tab. 3 compares these two methods, showing that AtMan still outperforms Chefer. Note, however, that Chefer achieves better scores compared to being applied to the MAGMA model.

On language models, we found that "chat"-finetuned-versions are more applicable to this method, most likely as they are specifically trained to respect the given input. In general, prompt-tuning instructions may improve the results, such as "Answer with respect to the given context".

A.11 Metrics

All recall and precision metrics in this work are based on the normalized continuous values as follows: For relevancy scores x and binary segmentation label y, let $\widetilde{x} = x / \max_i(x_i)$. We compute $AP = (\sum_i x_i y_i) / \sum_i x_i$ for precision and for recall $AR = (\sum_i \widetilde{x}_i y_i) / \sum_i y_i$. The final mean scores

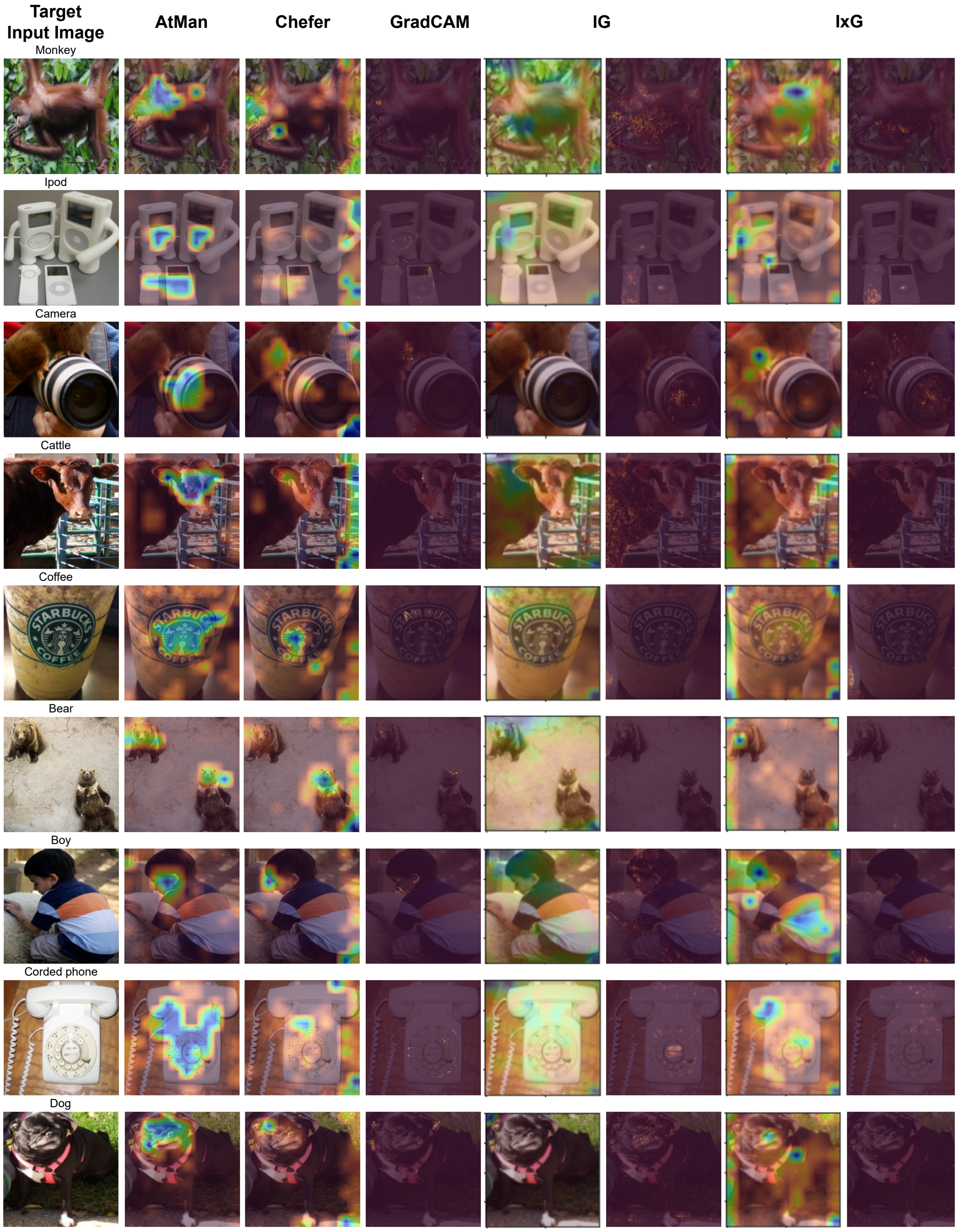
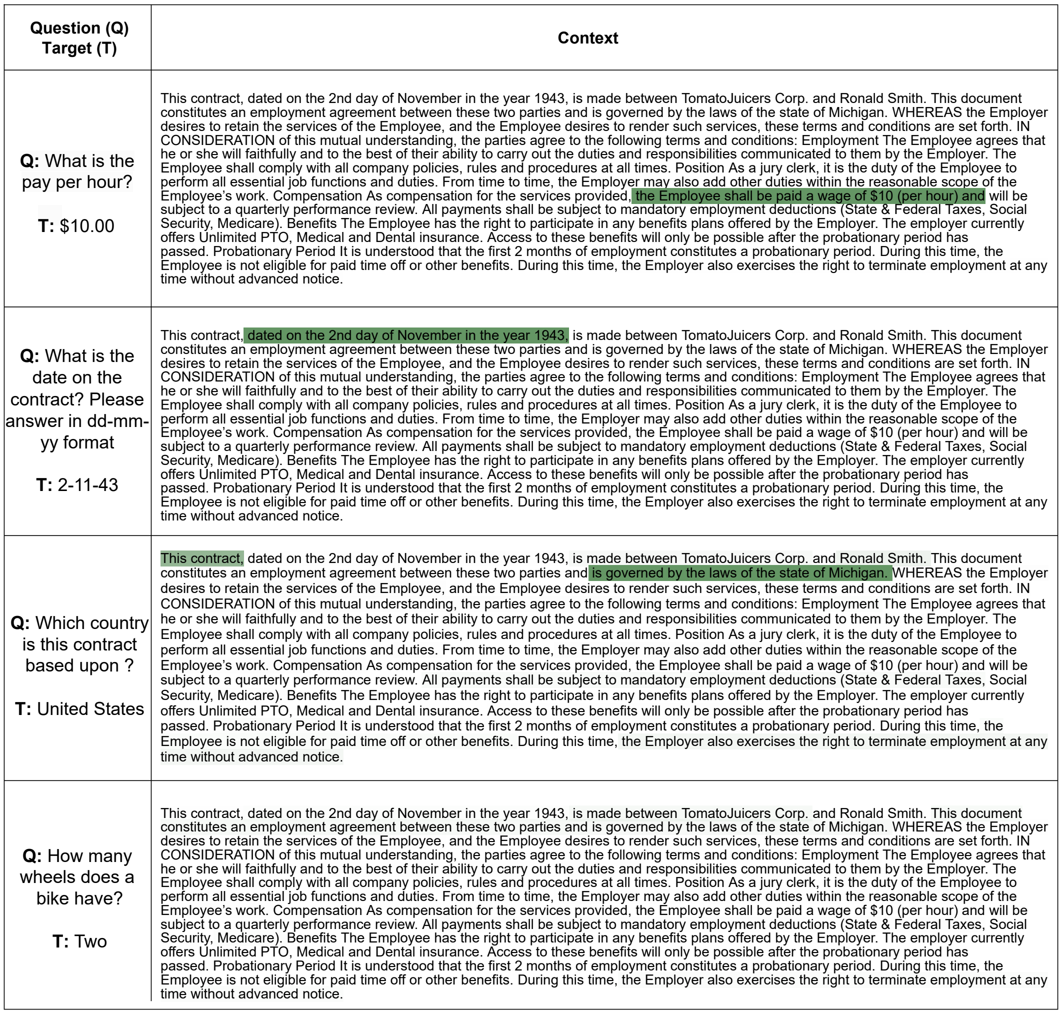


Figure 15: Weak image segmentation comparison of several images for all methods studied in this work. Note that IG and IxG are presented in token (left) and pixel-granularity (right). (Best viewed in color.)

are the average of all samples. This yields a comparable untuned metric also catching the robustness of the methods.

A.12 Interpretability of "complex facts" and failure modes

All experiments were executed in the generative setting, c.f. Eq. 1,2 and the remark to "global explanations" at the end of Sec. 3. Fig. 1a shows "real text generation" and the accumulation of explanations on the input image during the model's generation process. For comparison benchmarks,



Question (Q) **Target (T)**	**Context**
Q: What is the pay per hour? **T:** $10.00	This contract, dated on the 2nd day of November in the year 1943, is made between TomatoJuicers Corp. and Ronald Smith. This document constitutes an employment agreement between these two parties and is governed by the laws of the state of Michigan. WHEREAS the Employer desires to retain the services of the Employee, and the Employee desires to render such services, these terms and conditions are set forth. IN CONSIDERATION of this mutual understanding, the parties agree to the following terms and conditions: Employment The Employee agrees that he or she will faithfully and to the best of their ability to carry out the duties and responsibilities communicated to them by the Employer. The Employee shall comply with all company policies, rules and procedures at all times. Position As a jury clerk, it is the duty of the Employee to perform all essential job functions and duties. From time to time, the Employer may also add other duties within the reasonable scope of the Employee’s work. Compensation As compensation for the services provided, the Employee shall be paid a wage of $10 (per hour) and will be subject to a quarterly performance review. All payments shall be subject to mandatory employment deductions (State & Federal Taxes, Social Security, Medicare). Benefits The Employee has the right to participate in any benefits plans offered by the Employer. The employer currently offers Unlimited PTO, Medical and Dental insurance. Access to these benefits will only be possible after the probationary period has passed. Probationary Period It is understood that the first 2 months of employment constitutes a probationary period. During this time, the Employee is not eligible for paid time off or other benefits. During this time, the Employer also exercises the right to terminate employment at any time without advanced notice.
Q: What is the date on the contract? Please answer in dd-mm-yy format **T:** 2-11-43	This contract, dated on the 2nd day of November in the year 1943, is made between TomatoJuicers Corp. and Ronald Smith. This document constitutes an employment agreement between these two parties and is governed by the laws of the state of Michigan. WHEREAS the Employer desires to retain the services of the Employee, and the Employee desires to render such services, these terms and conditions are set forth. IN CONSIDERATION of this mutual understanding, the parties agree to the following terms and conditions: Employment The Employee agrees that he or she will faithfully and to the best of their ability to carry out the duties and responsibilities communicated to them by the Employer. The Employee shall comply with all company policies, rules and procedures at all times. Position As a jury clerk, it is the duty of the Employee to perform all essential job functions and duties. From time to time, the Employer may also add other duties within the reasonable scope of the Employee’s work. Compensation As compensation for the services provided, the Employee shall be paid a wage of $10 (per hour) and will be subject to a quarterly performance review. All payments shall be subject to mandatory employment deductions (State & Federal Taxes, Social Security, Medicare). Benefits The Employee has the right to participate in any benefits plans offered by the Employer. The employer currently offers Unlimited PTO, Medical and Dental insurance. Access to these benefits will only be possible after the probationary period has passed. Probationary Period It is understood that the first 2 months of employment constitutes a probationary period. During this time, the Employee is not eligible for paid time off or other benefits. During this time, the Employer also exercises the right to terminate employment at any time without advanced notice.
Q: Which country is this contract based upon ? **T:** United States	This contract, dated on the 2nd day of November in the year 1943, is made between TomatoJuicers Corp. and Ronald Smith. This document constitutes an employment agreement between these two parties and is governed by the laws of the state of Michigan. WHEREAS the Employer desires to retain the services of the Employee, and the Employee desires to render such services, these terms and conditions are set forth. IN CONSIDERATION of this mutual understanding, the parties agree to the following terms and conditions: Employment The Employee agrees that he or she will faithfully and to the best of their ability to carry out the duties and responsibilities communicated to them by the Employer. The Employee shall comply with all company policies, rules and procedures at all times. Position As a jury clerk, it is the duty of the Employee to perform all essential job functions and duties. From time to time, the Employer may also add other duties within the reasonable scope of the Employee’s work. Compensation As compensation for the services provided, the Employee shall be paid a wage of $10 (per hour) and will be subject to a quarterly performance review. All payments shall be subject to mandatory employment deductions (State & Federal Taxes, Social Security, Medicare). Benefits The Employee has the right to participate in any benefits plans offered by the Employer. The employer currently offers Unlimited PTO, Medical and Dental insurance. Access to these benefits will only be possible after the probationary period has passed. Probationary Period It is understood that the first 2 months of employment constitutes a probationary period. During this time, the Employee is not eligible for paid time off or other benefits. During this time, the Employer also exercises the right to terminate employment at any time without advanced notice.
Q: How many wheels does a bike have? **T:** Two	This contract, dated on the 2nd day of November in the year 1943, is made between TomatoJuicers Corp. and Ronald Smith. This document constitutes an employment agreement between these two parties and is governed by the laws of the state of Michigan. WHEREAS the Employer desires to retain the services of the Employee, and the Employee desires to render such services, these terms and conditions are set forth. IN CONSIDERATION of this mutual understanding, the parties agree to the following terms and conditions: Employment The Employee agrees that he or she will faithfully and to the best of their ability to carry out the duties and responsibilities communicated to them by the Employer. The Employee shall comply with all company policies, rules and procedures at all times. Position As a jury clerk, it is the duty of the Employee to perform all essential job functions and duties. From time to time, the Employer may also add other duties within the reasonable scope of the Employee’s work. Compensation As compensation for the services provided, the Employee shall be paid a wage of $10 (per hour) and will be subject to a quarterly performance review. All payments shall be subject to mandatory employment deductions (State & Federal Taxes, Social Security, Medicare). Benefits The Employee has the right to participate in any benefits plans offered by the Employer. The employer currently offers Unlimited PTO, Medical and Dental insurance. Access to these benefits will only be possible after the probationary period has passed. Probationary Period It is understood that the first 2 months of employment constitutes a probationary period. During this time, the Employee is not eligible for paid time off or other benefits. During this time, the Employer also exercises the right to terminate employment at any time without advanced notice.

Figure 16: Showing AtMan capabilities to highlight information in a document q/a setting. The model is prompted with “{Context} Q:{$Question$} A: ” and asked to extract the answer (target) of the given Explanation. Here, AtMan is run paragraph wise, as described in text, and correctly highlights the ones containing the information. **All Explanations were split into** ~ 50 **paragraphs (thus requiring only 50 AtMan forwad-passes)**. In particular it is shown in row 2 that the model can interpret, i.e. convert date-time formats. Row 3 shows that it can derive from world knowledge that Michigian is in the US. Row 4 shows that the method AtMan is robust against questions with non-including information. (Best viewed in color.)

Table 4: Performance of scaled MAGMA models.

	6b	13b	30b
VQA	60.0	62.6	64.2
OKVQA	37.6	38.2	43.3
GQA	47.4	43.7	45.6

our target was to separate generated explanations from the underlying model's interpretation, by retrieving "obvious facts". Fig. 4 demonstrates qualitatively target-class awareness.

More complex results are shown in Fig. 17. Fig. 17a gives an example where the model is VQA prompted and ATMAN explains both modalities at the same time for the completion token "white". It highlights the words "tub" and "color" of the text, and an edge of the tub in the picture as significant to produce the completion. Note that we find these "joint-modal explanations" to work best when the target token is not just a yes/no answer, c.f. Fig. a'). Further note how these artifacts remain amongst explainability methods such as Chefer, which may be due to the underlying models architecture. We leave these investigations for further research.

In Fig. 17b ATMAN highlights multiple heterogenous concepts to answer the counting task. Note that here the explanation for the completion "two", which actually highlights visually the animals, is found at the token position "animals", and not at the position of the predicted token. ATMAN returns at the same time explanations to all previous context positions, without additional forward passes (in contrast to gradient-based methods).

Finally, we ran evaluations on visual reasoning using the GQA dataset 17. Fig. 17c shows an instance of GQA, in which the model is prompted to answer "Does the man ride a horse?", while the man in the picture actually rides a bike. ATMAN highlights at two different sequence positions of the prompt the respective concepts in the image required to answer the question, once the man, and once the bike. Fig. 17d.1 and d.2 are the resulting IG explanations when applied to explain the same token positions; somewhat it captures parts of the concept but focuses on noise to the edge. Note that many samples of this dataset are rather difficult to answer, it is often not clear on what token position an explanation is supposed to be found. In particular, often ambiguous concepts occur multiple times in the image and are further restricted to the correct instances much later in the text description. This is, in particular, difficult to grasp for generative decoder models due to the causal masking. As the dataset contains annotations for areas as in example Fig. 17c, we ran preliminary quantitative results on 10k randomly chosen instances. ATMAN matches bounding boxes with mAP 52.4 and mAR 24.3. For IG we obtained mAP 33.3 and mAR 27.0.

A.13 Partial attention manipulation

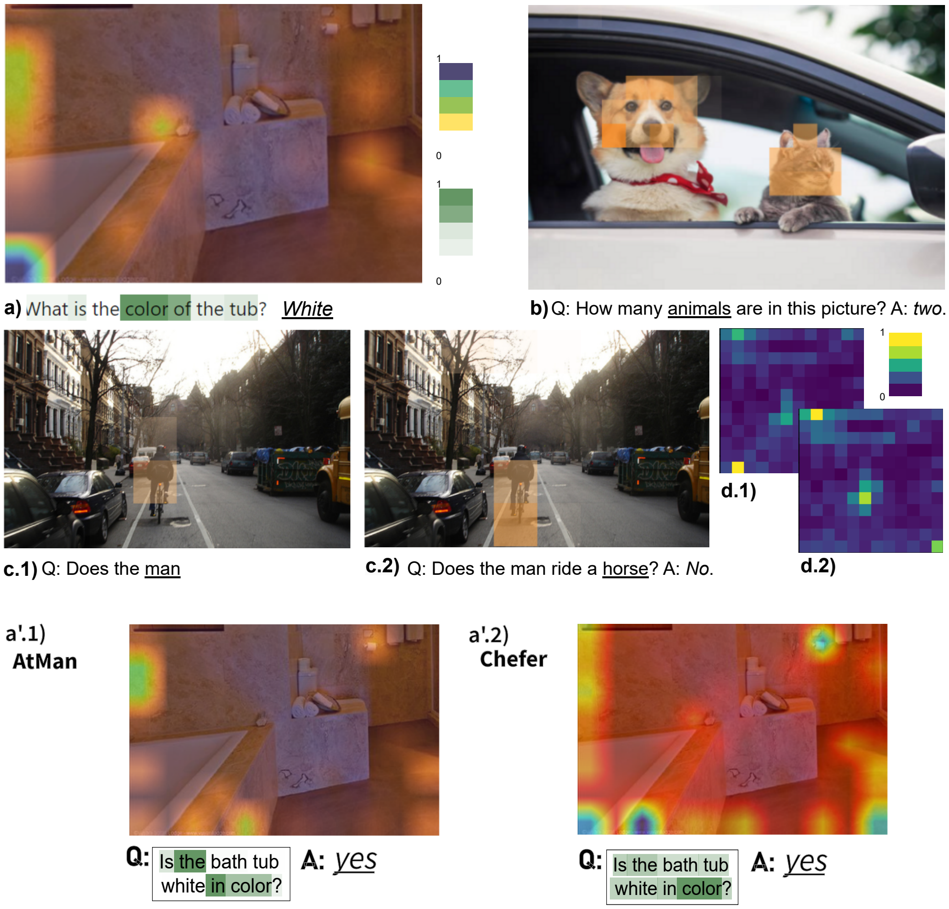
We apply the same manipulation in each layer. We hypothesised —and first experiments indicate— that one needs a "critical mass", to remove the entire concept entropy from the generation process. This is i.p. due to the skip connections. Only applying it to a single (arbitrary) layer did not yield comparable results. However we did not finalize a conclusion yet and leave this to future work. C.f. A.15, Fig. 18.

A.14 MAGMA scaling

MAGMA Model performances on standard text-image evaluation benchmarks are shown in Tab. 4.

A.15 Variants of ATMAN and final remarks

There are many promising variants of ATMAN not yet fully understood. In particular finding the relevant layers and other formulae to manipulate reliably the activations, or to normalize the output scores, could yield further insights into the transformer architecture. We showed how conceptual suppression improves overall performance. For pure NLP tasks, utilizing an embedding model such as BERT to compute similarities could be beneficial. Moreover, it should be possible to aggregate tokens and increase the level of details as needed, to save additional computational time.


1
0
1
0
1
0
a'.1)
AtMan
a'.2)
Chefer
Q: Is the bath tub white in color? A: yes
Q: Is the bath tub white in color? A: yes

a) What is the color of the tub? *White*

b) Q: How many animals are in this picture? A: *two*.

c.1) Q: Does the man

c.2) Q: Does the man ride a horse? A: *No*.

d.1)

d.2)

Figure 17: More complex examples showing generative multi-modal explanations (c.f. Sec. A.12 for detailed description). *italic*: model completion; underline: token position for explanation. (Best viewed in color.)

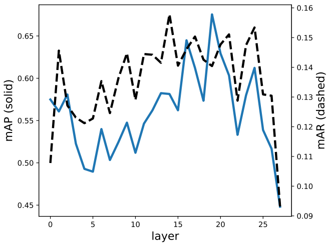


Figure 18: Measured mAP and mAR when applying AtMan only to a single layer, with similarity of the first embedding layer. Evaluated on a subset of openimages. Apparently, some layers contribute more to the explanation than others. "Full AtMan" as proposed yet outperforms on both metrics with scores mAP = 0.66 and mAR = 0.24.

Fig. 18 shows scores when applying ATMAN only to a single layer. I.p. the second half of layers seem more influential to the ATMAN scores, which is in accordance with recent research studying that higher-level concepts form later throughout the network.

In Fig. 19 we applied softmax on the embeddings (still only of the first layer) instead of the proposed thresholded cosine similarity Eq. 6. The proposed cosine similarity clearly outperforms its softmax variant in quality, however visible correlations can be found in the softmax-image as well.

As a final remark, we want to highlight again the rather undefined question of "what an explanation is". Afterall, we show what input leads to a diverging generation of the model faithfully, even if it would not match the expected outcome.

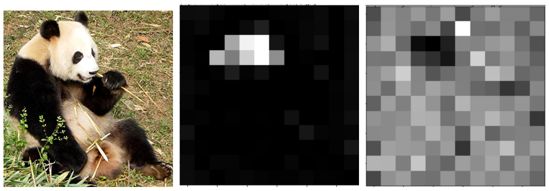


Figure 19: Qualitative example for an explanation of AtMan as proposed (middle) and AtMan when applying the softmax on the embedding instead of cosine similarity (right), as suggested by a reviewer. While some correlation can be found, it is much more noisy.