1. Extended LXMERT VQA visual results

In Fig. 1 we present extended results for Fig. 4 in the paper, *i.e*. we present the explanations extracted by each method for typical samples from the VQA dataset using the LXMERT model for question answering.

2. Preparing the DETR relevancy maps for the COCO segmentation evaluation code

In this section, we elaborate on the process of extracting segmentation masks from DETR's object detection results. The extracted segmentation masks are then used for our DETR tests, as presented in Sec. 5 of the paper.

DETR has been trained for object detection, *i.e*., producing a bounding box and a classification for each object in the input image. In order to evaluate the different explainability methods, we refer to the $\mathbf{R^{qi}}$ relevancy map, where the j-th row defines the relevance of each image feature to the j-th query, *i.e*. the j-th bounding box, as described in Sec. 3.2 of the paper. Our test uses each of the explainability methods on the $5,000$ samples of the MSCOCO validation set to produce segmentation masks, as described in Alg. 1. We first filter the queries to include only ones where the classification probability is higher than 50% (Alg. 1, L. 3). Then, for each query j that is left, we use the relevancy matrix $\mathbf{R^{qi}}$ in row j as a heatmap of the image features (Alg. 1, L. 6), noting the important pixels for the j-th predicted bounding box. Since most of our baselines, as well as our method, produce non-negative relevancies, we use Otsu's thresholding method to separate the foreground and the background of the segmentation mask (Alg. 1, L. 7). Then, the DETR segmentation evaluation code upsamples the masks to the target mask size, followed by a sigmoid operation, which only leaves the strictly positive values of the segmentation map (Alg. 1, L. 8-9). Finally, the DETR segmentation evaluation code upsamples the generated map back to the size of the original image (Alg. 1, L. 10).

Algorithm 1 Obtain Segmentation Masks from Heatmaps

Input : (i) input image (ii) $logits \in \mathbb{R}^{q \times c}$ obtained by the detection alg., where q is the number of queries (bounding boxes), and c is the number of object classes, (iii) $\mathbf{R^{qi}}$- relevancy matrix per query, from the explainability alg.
Output : *masks* $\in \mathbb{R}^{q \times h \times w}$ where q is the number of queries, and h, w are the spatial dimensions of the input image. *masks*$[j]$ is the segmentation map corresponding to the j-th bounding box.

1: $q \leftarrow queries$
2: *probabilities* $\leftarrow softmax(logits)$
3: *keep* $\leftarrow j \in q$, *where max(probabilities*$[j]) > 0.5$
4: *masks* $\leftarrow [[0, ..., 0], ..., [0, ..., 0]]$
5: *for* $j \in$ *keep*:
6: *masks*$[j] \leftarrow \mathbf{R^{qi}}[j]$
7: *masks*$[j] \leftarrow Otsu($*masks*$[j])$
8: *masks*$[j] \leftarrow Upsample($*masks*$[j]$, *size=targetMaskSize, method="bilinear"*$)$
9: *masks*$[j] \leftarrow$ *sigmoid(masks[j])* > 0.5
10: *masks*$[j] \leftarrow Upsample($*masks*$[j]$, *size=origImageSize, method="nearest"*$)$

3. Ablation Study

	Ours	w/o norm.	w/o aggregation	Eq.10 w/o self-att.
AP	**13.1**	11.7	0.1	11.5
AP_{medium}	**14.4**	13.9	0.0	13.8
AP_{large}	**24.6**	20.9	0.2	20.5
AR	**19.3**	18.0	0.5	17.8
AR_{medium}	**23.9**	23.9	0.0	23.8
AR_{large}	**33.2**	29.2	1.0	28.6

Table 1: Performance for different ablation variants of our method on the DETR experiments. Higher is better.

We present three variations of our method. Firstly, we verify the effectiveness of our normalization for the self-attention relevancies presented in Eq. 8,9. Since the normalization is applied to rule 10, we expect it to affect mostly bi-modal

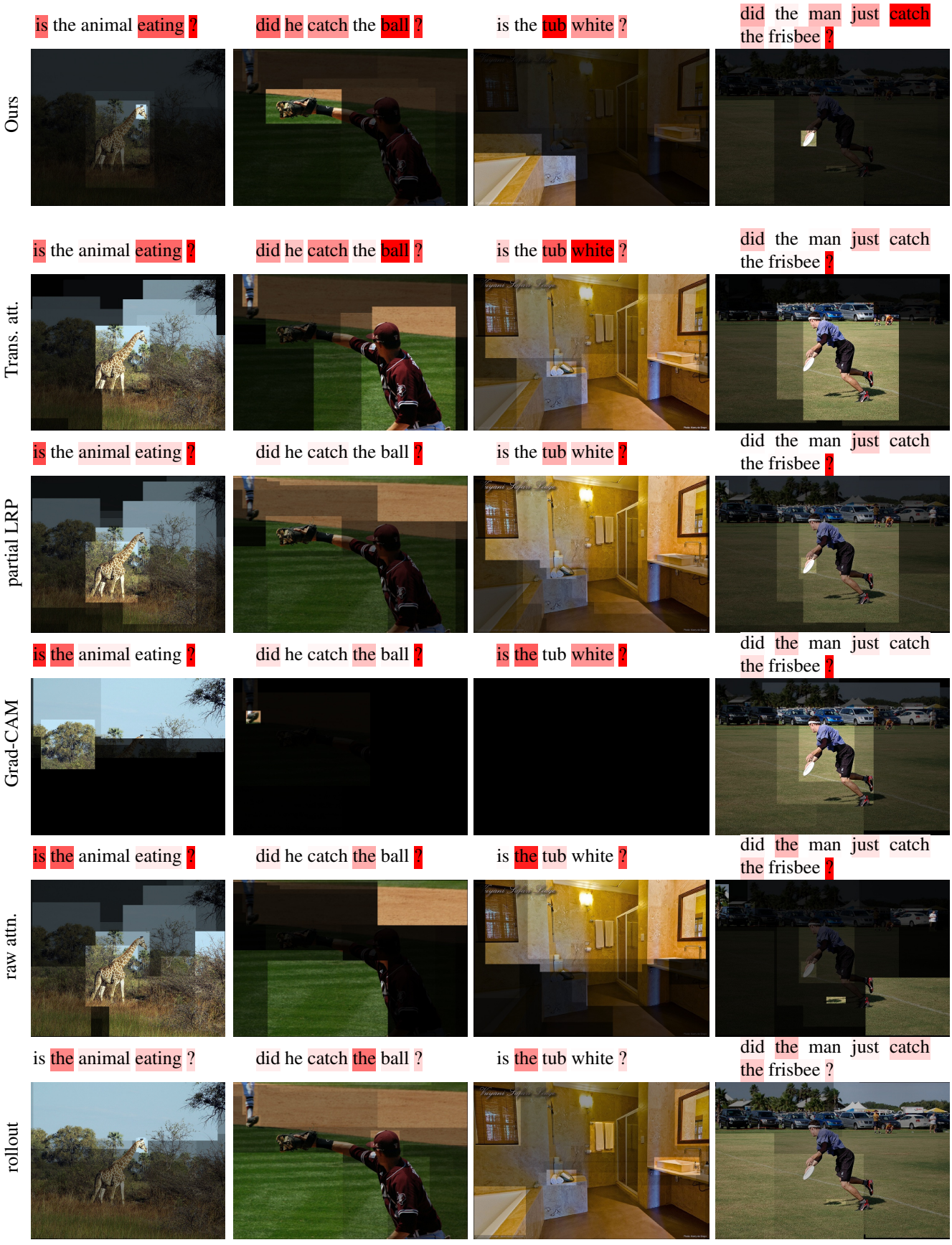

Figure 1: A comparison between our method (top) and the baselines for VQA with the LXMERT model. Relevancy for text is given as shades of red. Relevancy for images is given by multiplying each region by the relative relevancy. Notice that both for the images and the text our method achieves favorable results. Answers (left to right): no, yes, yes, no.

	Ours	w/o norm.	w/o aggregation	Eq.10 w/o self-att.
Neg. img	**63.24**	62.49	60.41	62.18
Pos. img	51.10	50.60	60.40	**50.57**
Neg. text	**48.70**	48.64	41.72	48.64
Pos. text	21.61	**21.59**	41.72	**21.59**

Table 2: Area-under-the-curve for different ablation variants of our method on the LXMERT experiments. For negative perturbation, larger AUC is better; for positive perturbation, smaller AUC is better.

relevancies, *i.e*. the image perturbation experiments for LXMERT, and the DETR tests. The second ablation we present studies the necessity of the aggregation in all our rules 6,7,10,11, *i.e*. instead of adding the former relevancy matrix to the newly constructed one, we only keep the new one, *e.g*. for rule 6 the update becomes: $\mathbf{R}^{ss} = \bar{\mathbf{A}} \cdot \mathbf{R}^{ss}$. Lastly, we explore the need for the self-attention updates to the bi-modal rule 10 by changing the update rule to: $\mathbf{R}^{sq} = \mathbf{R}^{sq} + \bar{\mathbf{A}}$. All our ablations are done on the LXMER, DETR experiments since, as mentioned several times, VisualBERT is based on pure self-attention, which yields similar results to the Transformer attribution baseline.

As can be seen from Tab. 1, all the components included in our method are crucial to its success on DETR, and the ablations cause a sizeable decrease in performance. It should be noted that for the reasonable ablations of not using normalization and not using self-attention in Eq.10, our ablations still outperform all other methods significantly for the DETR experiment.

For the image perturbation test on LXMERT, presented in Tab. 2, we observe relatively mild differences between our method and the ablations of no normalization and no self-attention, this can be attributed to the fact that in contrast to DETR, LXMERT only uses 36 image regions that had gone through Non-maximum Suppression (NMS), therefore the added context from the self-attention to the multi-modal attention is not as crucial, since usually the top-1 image region is identical to that of the ablations, and is sufficient to make the classification.

4. Using LRP with our method

We present the results for the LXMERT perturbation tests evaluated by the area-under-the-curve measure for our method with LRP for completeness, *i.e*. with head averaging as presented in the Transformer attribution method and in Eq. 13 instead of the head averaging in Eq. 5. The results in Tab. 3 support and substantiate the conclusions presented in the paper: for the image perturbation tests, whether or not LRP is used, our method's contributions lead to a large gap in performance over all baseline methods. LRP itself leads to a small degradation in performance. For the text perturbation tests which are, as mentioned in the paper, self-attention based, our method is similar in performance to the Transformer Attribution method. Here, too, the choice of whether or not to use LRP is insignificant. Given the complexity of implementing LRP (see Sec. 4 of the main text), we advocate to eliminate it.

	Ours	Ours w/ LRP	Transformer att.	raw attn.	partial LRP	Grad-CAM	rollout
Neg. img	**63.24**	62.41	61.46	61.34	60.90	60.08	58.64
Pos. img	**51.10**	51.10	52.75	54.17	52.82	59.21	57.23
Neg. text	**48.70**	48.25	48.24	38.32	45.15	37.99	32.05
Pos. text	**21.61**	21.68	21.68	32.56	24.22	34.14	39.29

Table 3: Area-under-the-curve for all the baselines and our method with and without LRP on the LXMERT experiments. For negative perturbation, larger AUC is better; for positive perturbation, smaller AUC is better.

5. Perturbation experiments graphs

In Fig. 2, 3, we present enlarged graphs corresponding to our perturbation experiments for better clarity.

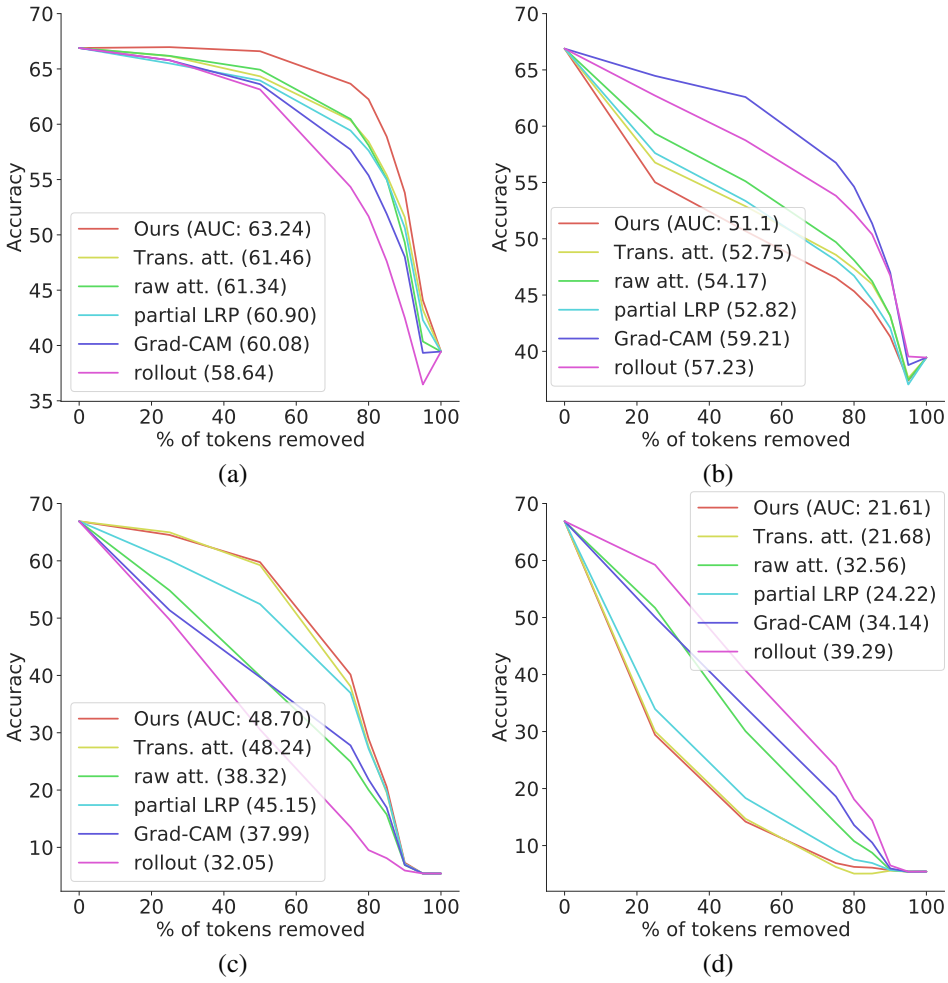

Figure 2: LXMERT perturbation test results. For negative perturbation, larger AUC is better; for positive perturbation, smaller AUC is better. (a) negative perturbation on image tokens, (b) positive perturbation on image tokens, (c) negative perturbation on text tokens, and (d) positive perturbation on text tokens.

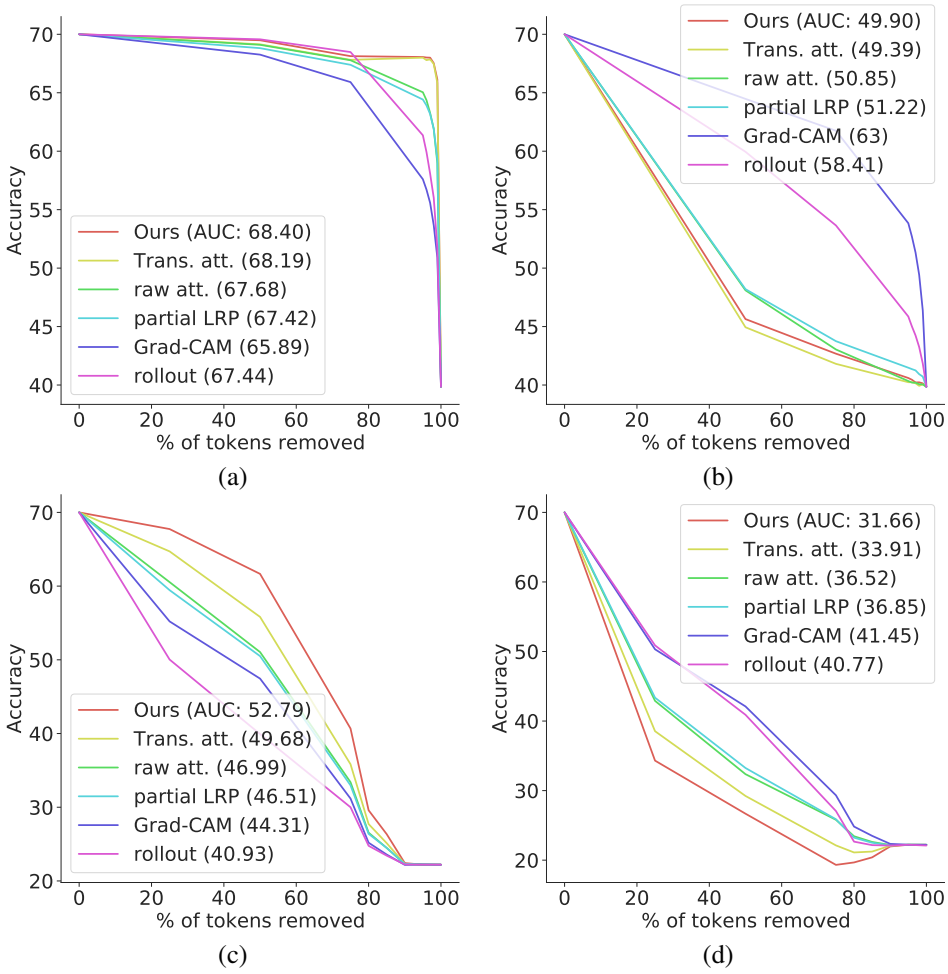

Figure 3: VisualBERT perturbation test results. For negative perturbation, larger AUC is better; for positive perturbation, smaller AUC is better. (a) negative perturbation on image tokens, (b) positive perturbation on image tokens, (c) negative perturbation on text tokens, and (d) positive perturbation on text tokens.