1. Details of the various Baselines

GradCAM As mentioned in Sec. 4, we consider the last attention layer (closest to the output) - namely $\mathbf{A}^{(1)}$. This results in a feature-map of size $h \times s \times s$. Following the process described in Sec. 3.4, we take only the `[CLS]` token's row (without the `[CLS]` token's column), and reshape to the patches grid size $h_p \times w_p$. This results in a feature-map similar to the 2D feature-map used for GradCAM, where the number of channels, in this case, is h, and the height and width are h_p and w_p. The reason we use the last attention layer is because of the sparse gradients issue described in Sec. 4.

raw-attention The raw-attention method visualizes the last attention layer (closest to the output) - namely $\mathbf{A}^{(1)}$. It follows the process described in Sec. 3.4 to extract the final output.

LRP In this method, we propagate relevance up to the input image, following the propagation rules of LRP (not our modified rules and normalizations).

partial-LRP Following [42], we visualize an intermediate relevance map, more specifically, we visualize the last attention-map's relevance, namely $R^{(n_1)}$, using LRP propagation rules.

rollout We follow Eq. 16.

2. Proofs for Lemmas

Given two tensors u and v, we compute the relevance propagation of binary operators (*i.e.*, operators that process two operands) as follows:

$$\begin{aligned} R_j^{u^{(n)}} &= \mathcal{G}(u, v, R^{(n-1)}) \\ R_k^{v^{(n)}} &= \mathcal{G}(v, u, R^{(n-1)}) \end{aligned} \tag{1}$$

where $R_j^{u^{(n)}}$ and $R_k^{v^{(n)}}$ are the relevances for u and v respectively.

The following lemma shows that for the case of addition, the conservation rule is preserved, i.e.,

$$\sum_j R_j^{u^{(n)}} + \sum_k R_k^{v^{(n)}} = \sum_i R_i^{(n-1)}. \tag{2}$$

However, this is not the case for matrix multiplication.

Lemma 1. *Given two tensors u and v, consider the relevances that are computed according to Eq. 1. Then, (i) if layer $L^{(n)}$ adds the two tensors, i.e., $L^{(n)}(u, v) = u + v$ then the conservation rule of Eq. 2 is maintained. (ii) if the layer performs matrix multiplication $L^{(n)}(u, v) = uv$, then Eq. 2 does not hold in general.*

Proof. For part (i), we note that the number of elements in u equals the number of elements in v, therefore $k = j$, and we can write Eq. 2 following the definition of $\mathcal{G}$:

$$\begin{aligned} &\sum_j \sum_i u_j \frac{\partial(u_i + v_i)}{\partial u_j} \frac{R_i^{(n-1)}}{u_i + v_i} + \sum_j \sum_i v_j \frac{\partial(u_i + v_i)}{\partial v_j} \frac{R_i^{(n-1)}}{u_i + v_i} \\ &= \sum_j \frac{u_j}{u_j + v_j} R_j^{(n-1)} + \sum_j \frac{v_j}{u_j + v_j} R_j^{(n-1)} \\ &= \sum_j \frac{u_j + v_j}{u_j + v_j} R_j^{(n-1)} = \sum_j R_j^{(n-1)} \end{aligned} \tag{3}$$

note that, in this case, it is possible that $\sum_j R_j^{u^{(n)}} \neq \sum_j R_j^{v^{(n)}}$.

As shown in the main text, while the sum of two tensors maintains the conservation rule, their values may explode. Consider $u = \begin{pmatrix} e^a \\ e^b \end{pmatrix}$, $v = \begin{pmatrix} 1 - e^a \\ 1 - e^b \end{pmatrix}$ and $R = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$, following the definition of $\mathcal{G}$ we have:

$$\begin{aligned} R_j^{u^{(n)}} &= \sum_i u_j \frac{\partial(u_i + v_i)}{\partial u_j} \frac{R_i^{(n-1)}}{u_i + v_i} = \frac{u_j}{u_j + v_j} R_j^{(n-1)}, \quad R_j^{v^{(n)}} = \frac{v_j}{u_j + v_j} R_j^{(n-1)} \\ R^u &= \begin{pmatrix} \frac{e^a}{e^a - e^a + 1} 1 \\ \frac{e^b}{e^b - e^b + 1} 1 \end{pmatrix} = \begin{pmatrix} e^a \\ e^b \end{pmatrix}, \quad R^v = \begin{pmatrix} 1 - e^a \\ 1 - e^b \end{pmatrix} \end{aligned} \tag{4}$$

which causes numerical instability.

For part (ii), in the case of matrix multiplication between u and v, where $u \in \mathbb{R}^{k,m}$, $v \in \mathbb{R}^{m,l}$, we will show that:

$\sum_k \sum_m R^{u^{(n)}}_{k,m} = \sum_m \sum_l R^{v^{(n)}}_{m,l} = \sum_l \sum_k R^{(n-1)}_{k,l}$, which invalidates the conservation rule:

$$R^{u^{(n)}}_{k,m} = \sum_l u_{k,m} \frac{\partial (uv)_{k,l}}{\partial u_{k,m}} \frac{R^{(n-1)}_{k,l}}{\sum_{m'} u_{k,m'} v_{m'l}} = \sum_l \frac{u_{k,m} v_{m,l}}{\sum_{m'} u_{k,m'} v_{m'l}} R^{(n-1)}_{k,l} \tag{5}$$

$$R^{v^{(n)}}_{m,l} = \sum_k v_{m,l} \frac{\partial (uv)_{k,l}}{\partial v_{m,l}} \frac{R^{(n-1)}_{k,l}}{\sum_{m'} u_{k,m'} v_{m'l}} = \sum_k \frac{u_{k,m} v_{m,l}}{\sum_{m'} u_{k,m'} v_{m'l}} R^{(n-1)}_{k,l} \tag{6}$$

$$\begin{aligned} \sum_k \sum_m R^{u^{(n)}}_{k,m} + \sum_m \sum_l R^{v^{(n)}}_{m,l} &= \sum_k \sum_m \sum_l \frac{u_{k,m} v_{m,l}}{\sum_{m'} u_{k,m'} v_{m'l}} R^{(n-1)}_{k,l} + \sum_m \sum_l \sum_k \frac{u_{k,m} v_{m,l}}{\sum_{m'} u_{k,m'} v_{m'l}} R^{(n-1)}_{k,l} \\ &= \sum_k \sum_l \frac{\sum_m u_{k,m} v_{m,l}}{\sum_{m'} u_{k,m'} v_{m'l}} R^{(n-1)}_{k,l} + \sum_l \sum_k \frac{\sum_m u_{k,m} v_{m,l}}{\sum_{m'} u_{k,m'} v_{m'l}} R^{(n-1)}_{k,l} \qquad (7) \\ &= 2 \sum_l \sum_k R^{(n-1)}_{k,l} \qquad (8) \end{aligned}$$

□

To address the lack of conservation in the attention mechanism, which employs multiplication, and the numerical issues of the skip connections, our method applies a normalization to $R^{u^{(n)}}_j$ and $R^{v^{(n)}}_k$:

$$\begin{aligned} \bar{R}^{u^{(n)}}_j &= R^{u^{(n)}}_j \frac{\left|\sum_j R^{u^{(n)}}_j\right|}{\left|\sum_j R^{u^{(n)}}_j\right| + \left|\sum_k R^{v^{(n)}}_k\right|} \cdot \frac{\sum_i R^{(n-1)}_i}{\sum_j R^{u^{(n)}}_j} \\ \bar{R}^{v^{(n)}}_k &= R^{v^{(n)}}_k \frac{\left|\sum_k R^{v^{(n)}}_k\right|}{\left|\sum_j R^{u^{(n)}}_j\right| + \left|\sum_k R^{v^{(n)}}_k\right|} \cdot \frac{\sum_i R^{(n-1)}_i}{\sum_k R^{v^{(n)}}_k} \end{aligned} \tag{9}$$

Lemma 2. *The normalization technique upholds the following properties: (i) it maintains the conservation rule, i.e.:* $\sum_j \bar{R}^{u^{(n)}}_j + \sum_k \bar{R}^{v^{(n)}}_k = \sum_i R^{(n-1)}_i$*, (ii) it bounds the relevance sum of each tensor such that:*

$$0 \leq \sum_j \bar{R}^{u^{(n)}}_j, \sum_k \bar{R}^{v^{(n)}}_k \leq \sum_i R^{(n-1)}_i \tag{10}$$

Proof. For part (i), it holds that:

$$\sum_j \bar{R}^{u^{(n)}}_j + \sum_k \bar{R}^{v^{(n)}}_k \tag{11}$$

$$= \sum_j R^{u^{(n)}}_j \frac{\left|\sum_j R^{u^{(n)}}_j\right|}{\left|\sum_j R^{u^{(n)}}_j\right| + \left|\sum_k R^{v^{(n)}}_k\right|} \cdot \frac{\sum_i R^{(n-1)}_i}{\sum_j R^{u^{(n)}}_j}$$

$$+ \sum_k R^{v^{(n)}}_k \frac{\left|\sum_k R^{v^{(n)}}_k\right|}{\left|\sum_j R^{u^{(n)}}_j\right| + \left|\sum_k R^{v^{(n)}}_k\right|} \cdot \frac{\sum_i R^{(n-1)}_i}{\sum_k R^{v^{(n)}}_k} \tag{12}$$

$$= \frac{\left|\sum_j R^{u^{(n)}}_j\right|}{\left|\sum_j R^{u^{(n)}}_j\right| + \left|\sum_k R^{v^{(n)}}_k\right|} \cdot \sum_i R^{(n-1)}_i + \frac{\left|\sum_k R^{v^{(n)}}_k\right|}{\left|\sum_j R^{u^{(n)}}_j\right| + \left|\sum_k R^{v^{(n)}}_k\right|} \cdot \sum_i R^{(n-1)}_i \tag{13}$$

$$= \frac{\left|\sum_j R^{u^{(n)}}_j\right| + \left|\sum_k R^{v^{(n)}}_k\right|}{\left|\sum_j R^{u^{(n)}}_j\right| + \left|\sum_k R^{v^{(n)}}_k\right|} \cdot \sum_i R^{(n-1)}_i = \sum_i R^{(n-1)}_i \tag{14}$$

For part (ii) it is trivial to see that we weigh each tensor according to its relative absolute-value contribution:

$$\sum_j \bar{R}_j^{u^{(n)}} = \sum_j R_j^{u^{(n)}} \frac{\left|\sum_j R_j^{u^{(n)}}\right|}{\left|\sum_j R_j^{u^{(n)}}\right| + \left|\sum_k R_k^{v^{(n)}}\right|} \cdot \frac{\sum_i R_i^{(n-1)}}{\sum_j R_j^{u^{(n)}}} \tag{15}$$

$$= \frac{\left|\sum_j R_j^{u^{(n)}}\right|}{\left|\sum_j R_j^{u^{(n)}}\right| + \left|\sum_k R_k^{v^{(n)}}\right|} \cdot \sum_i R_i^{(n-1)} \tag{16}$$

we see that:

$$0 \leq \frac{\left|\sum_j R_j^{u^{(n)}}\right|}{\left|\sum_j R_j^{u^{(n)}}\right| + \left|\sum_k R_k^{v^{(n)}}\right|} \leq 1 \tag{17}$$

therefore:

$$0 \leq \sum_j \bar{R}_j^{u^{(n)}}, \sum_k \bar{R}_k^{v^{(n)}} \leq \sum_i R_i^{(n-1)} \tag{18}$$

□

3. Visualizations - Multiple-class Images

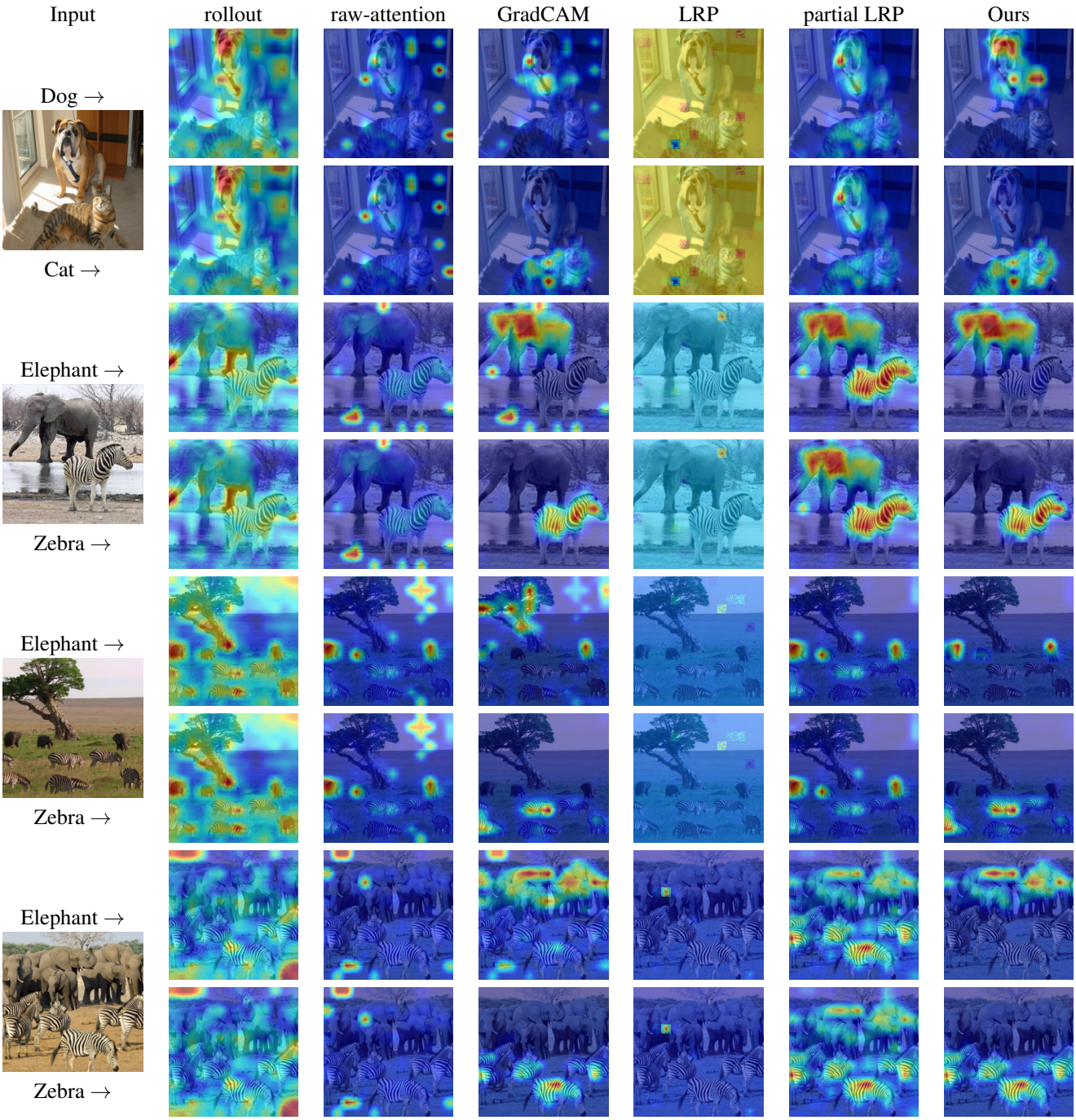


Figure 1: Multiple-class visualization. For each input image, we visualize two different classes. As can be seen, only our method and GradCAM produce class-specific visualisations, where our method has fewer artifacts, and captures the objects more completely.

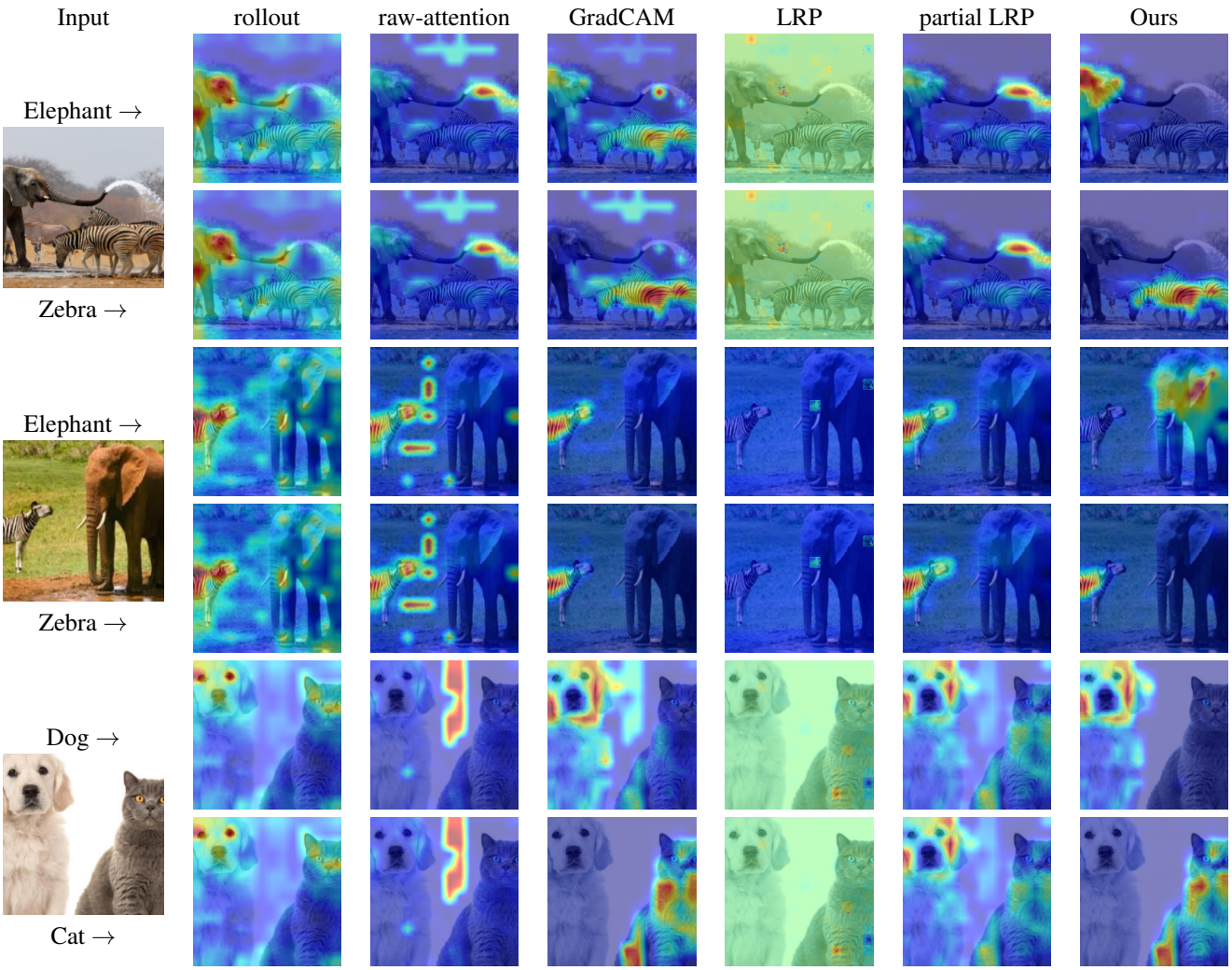


Figure 2: Multiple-class visualization. For each input image, we visualize two different classes. As can be seen, only our method and GradCAM produce class-specific visualisations, where our method has fewer artifacts, and captures the objects more completely.

4. Visualizations - Single-class Images

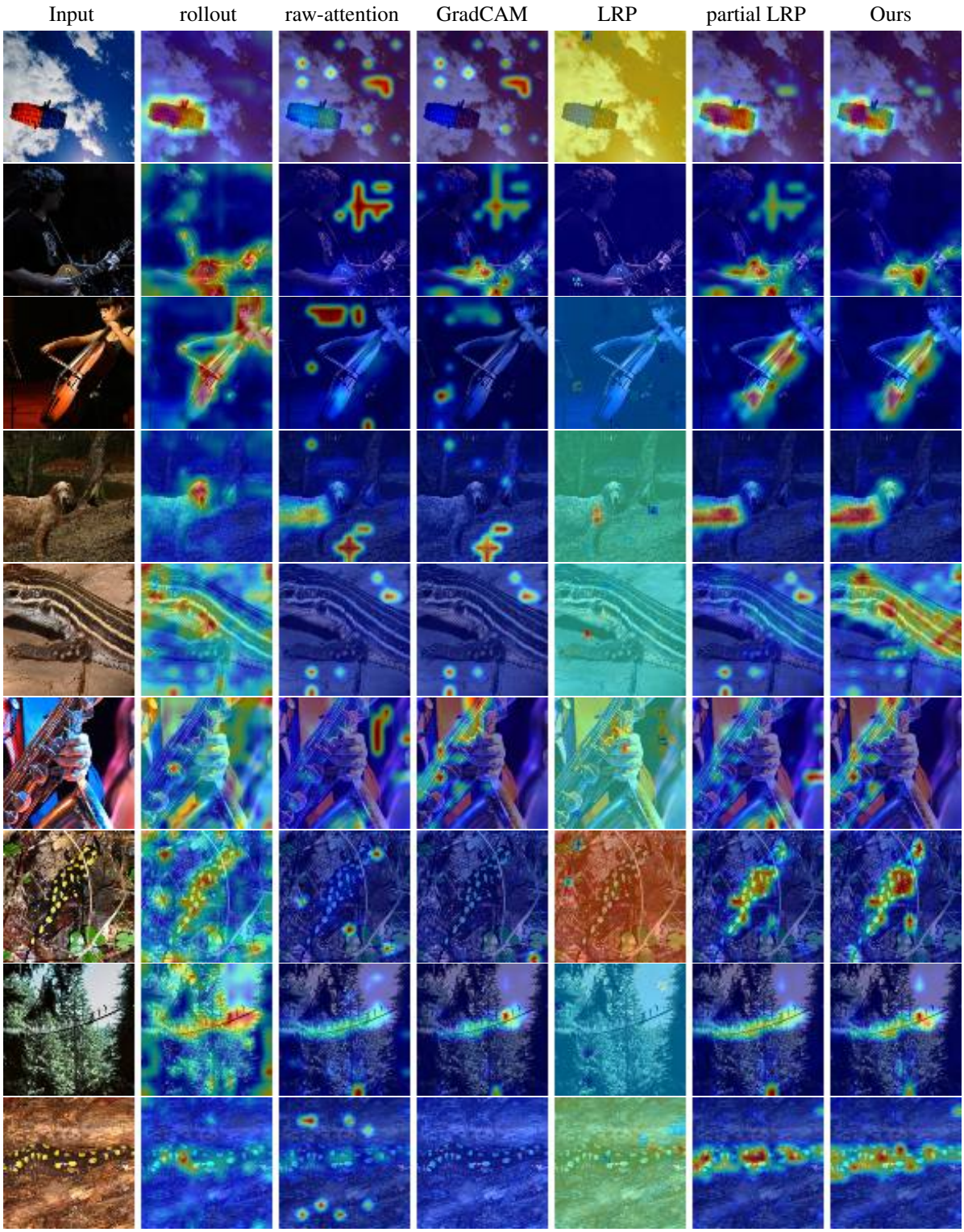


Figure 3: Sample images from ImageNet val-set.

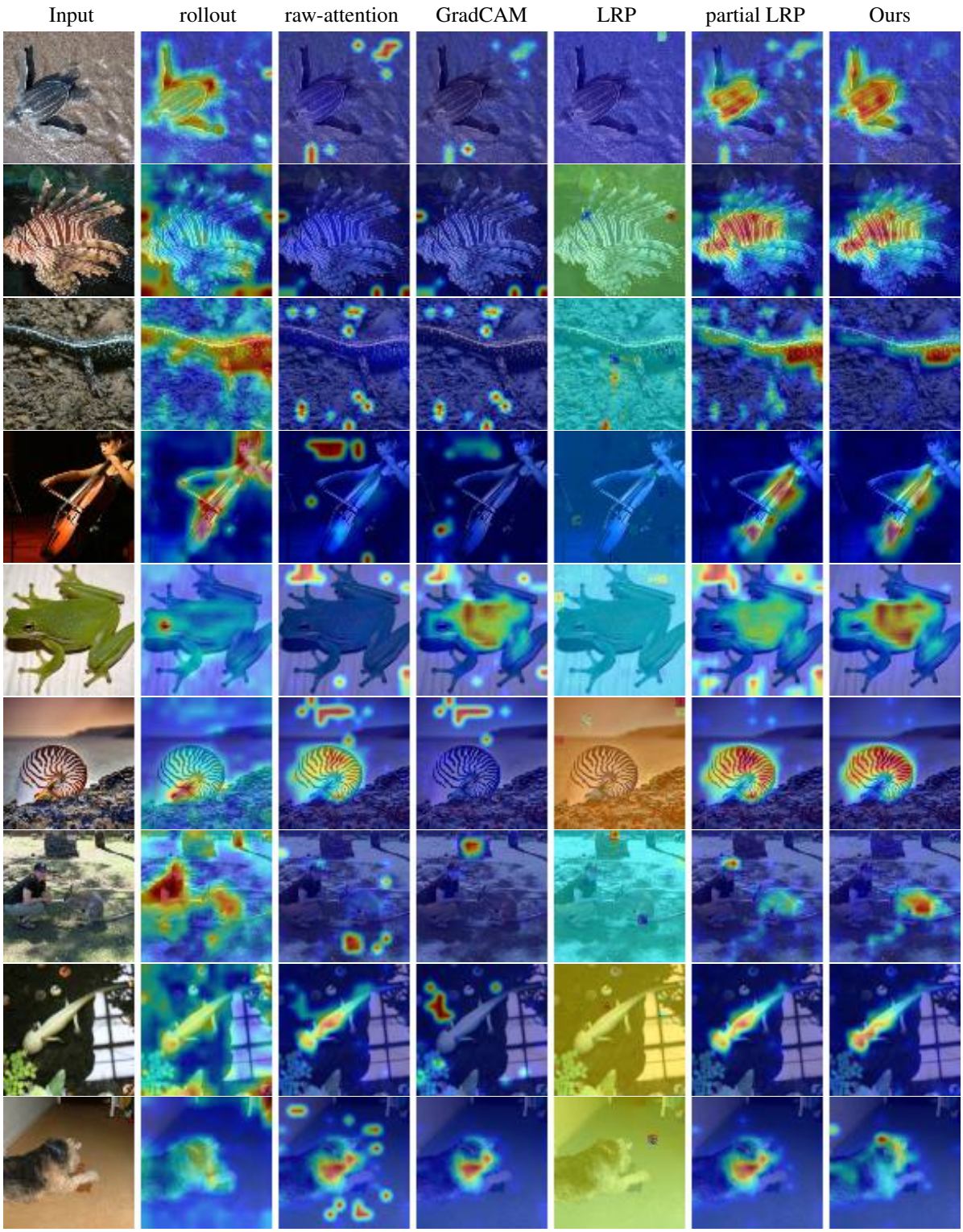


Figure 4: Sample images from ImageNet val-set.

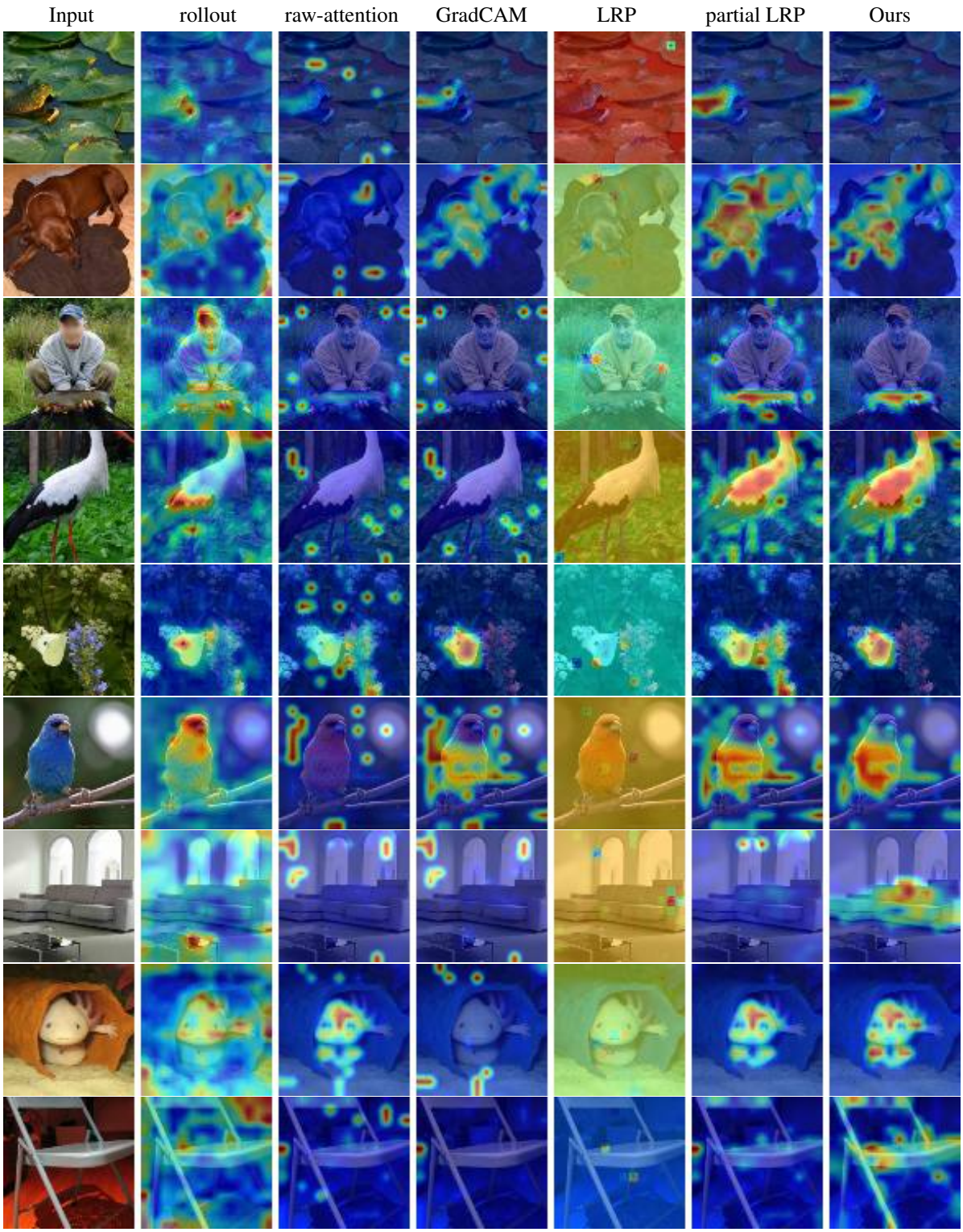


Figure 5: Sample images from ImageNet val-set.

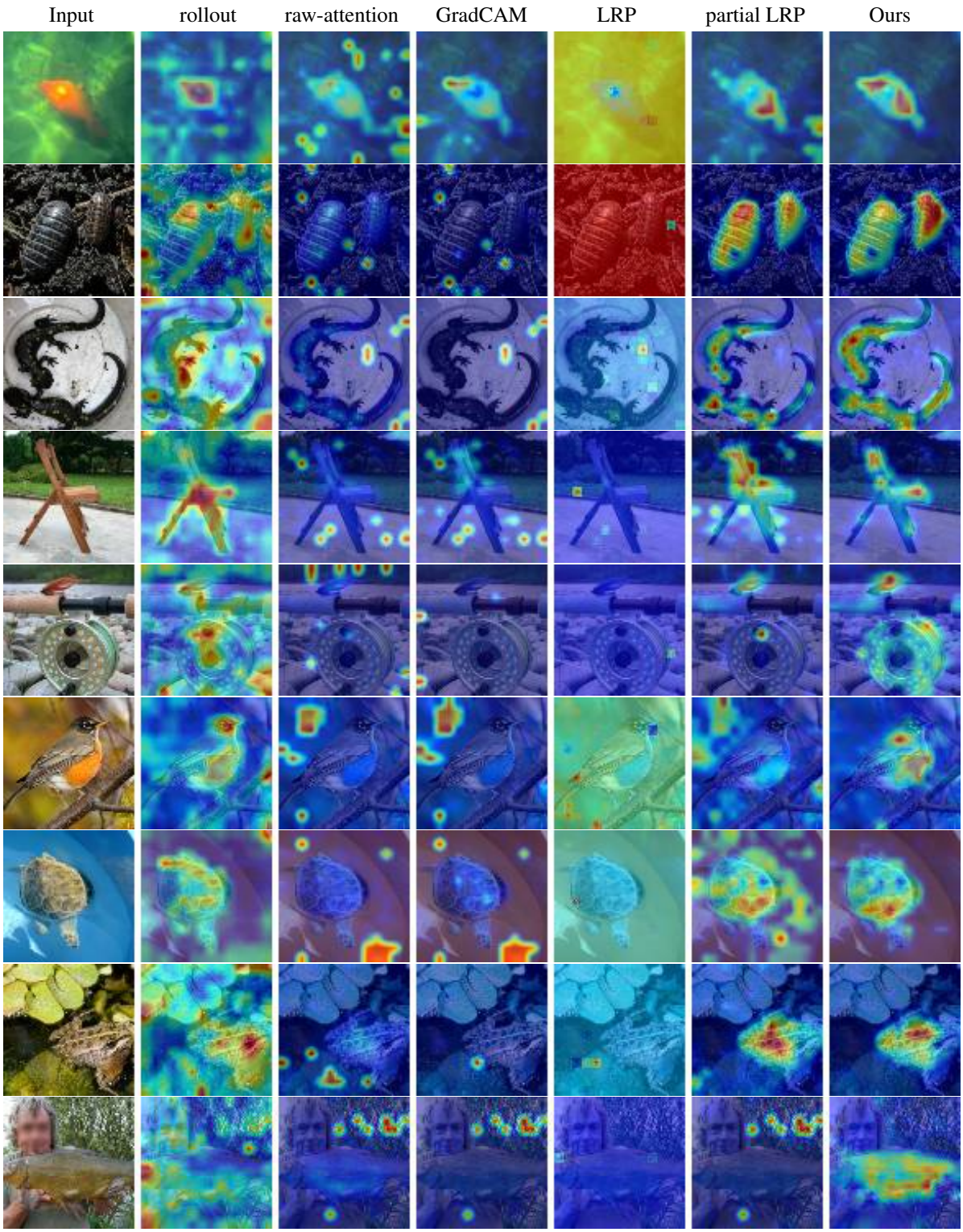


Figure 6: Sample images from ImageNet val-set.

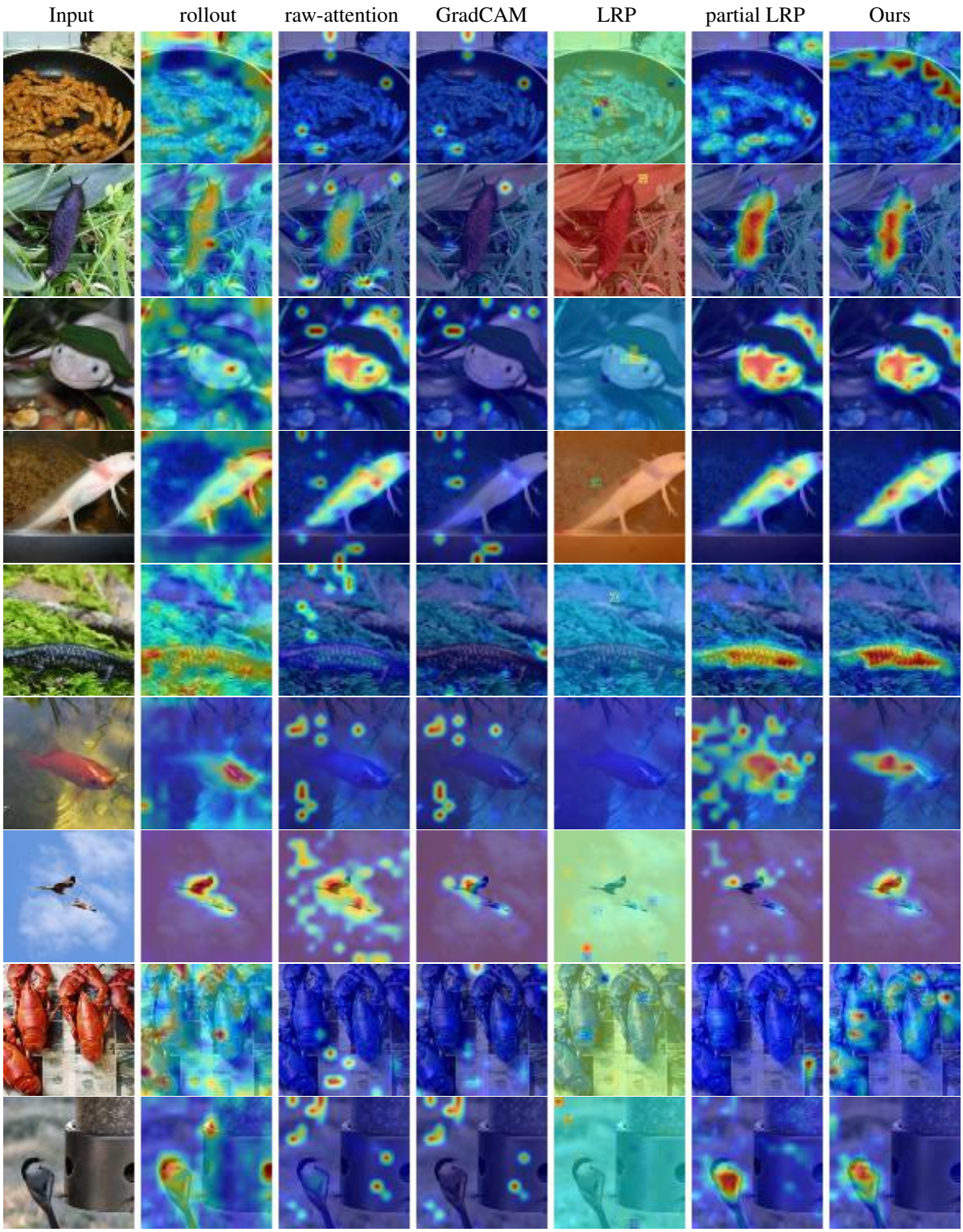


Figure 7: Sample images from ImageNet val-set.

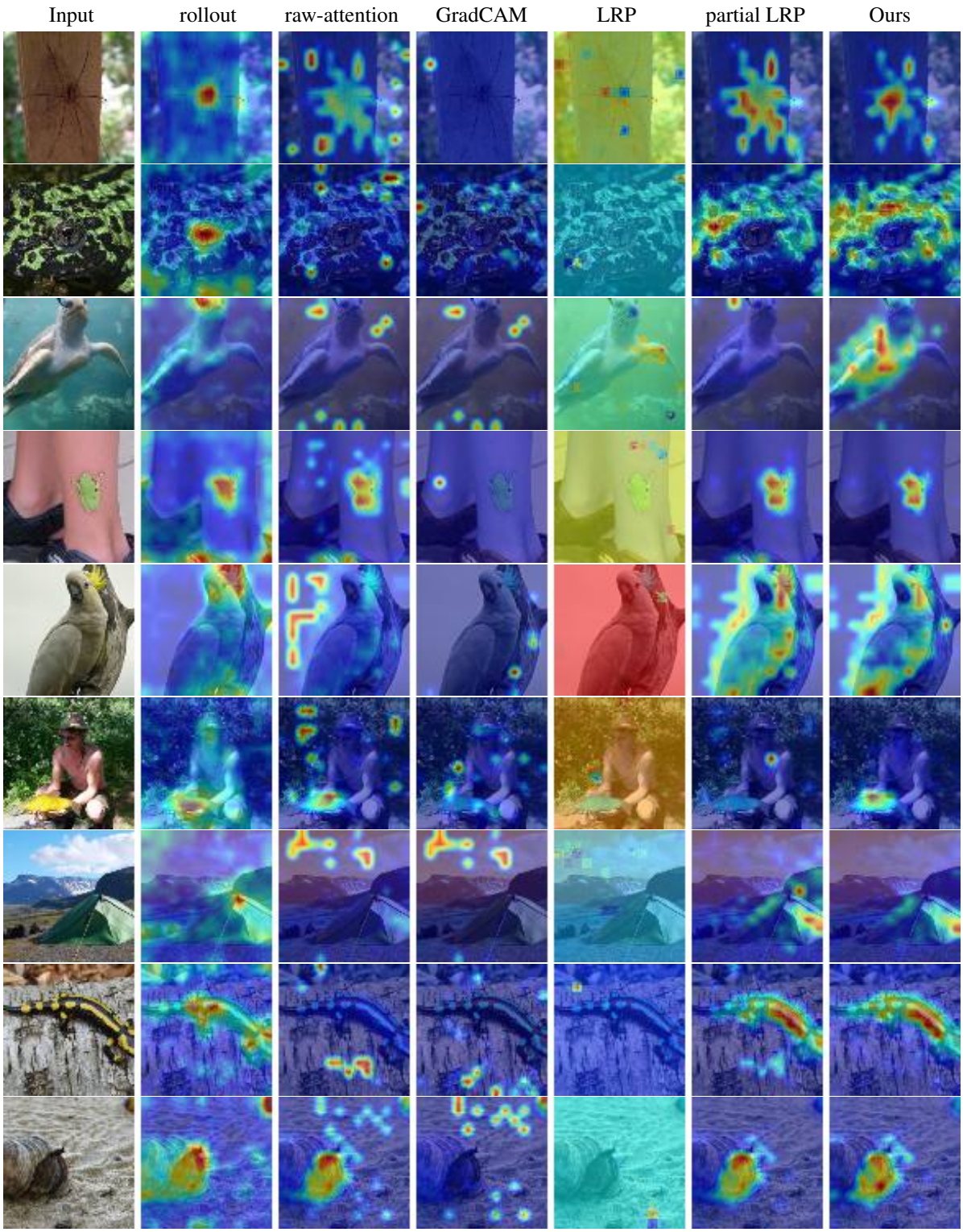


Figure 8: Sample images from ImageNet val-set.

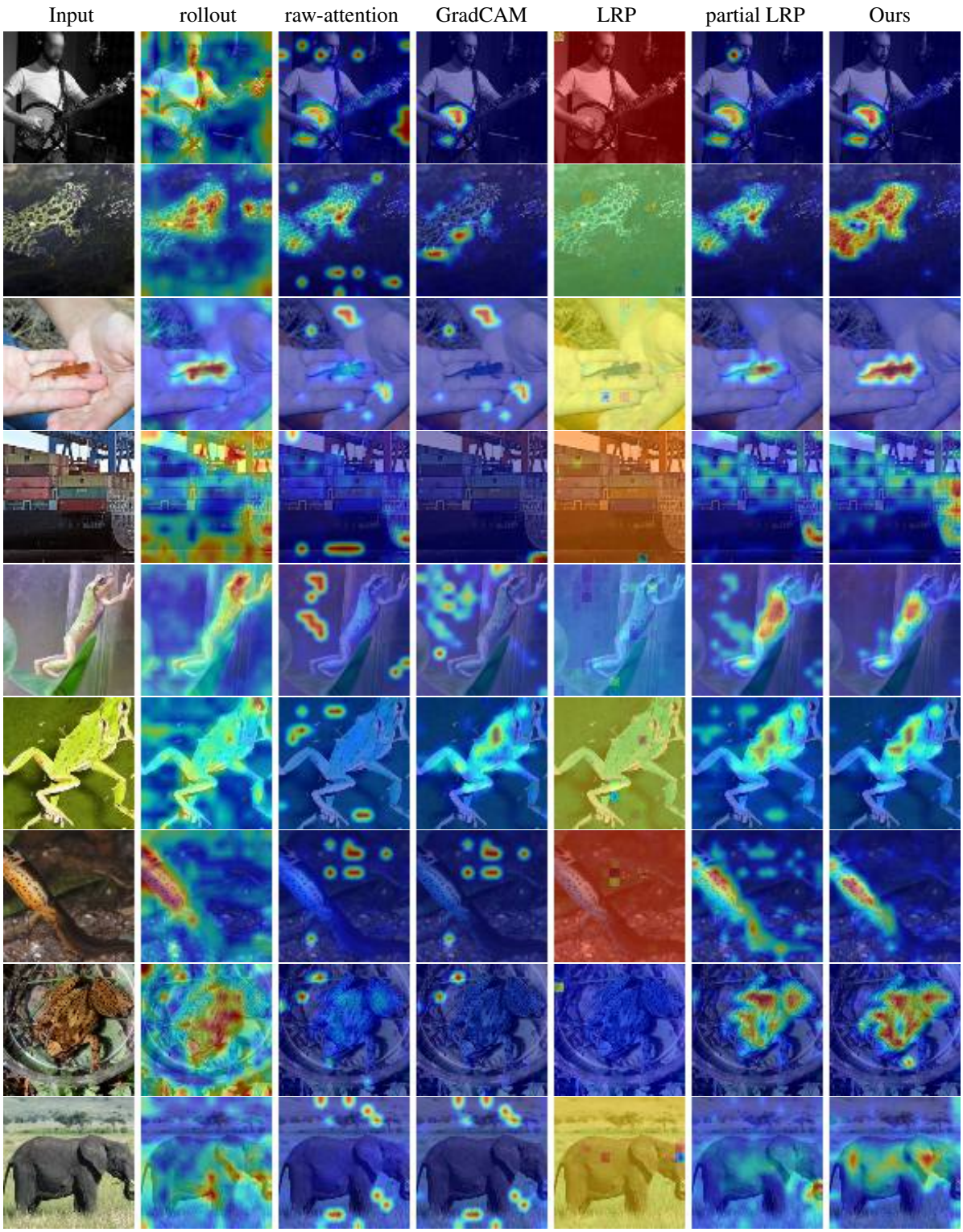


Figure 9: Sample images from ImageNet val-set.

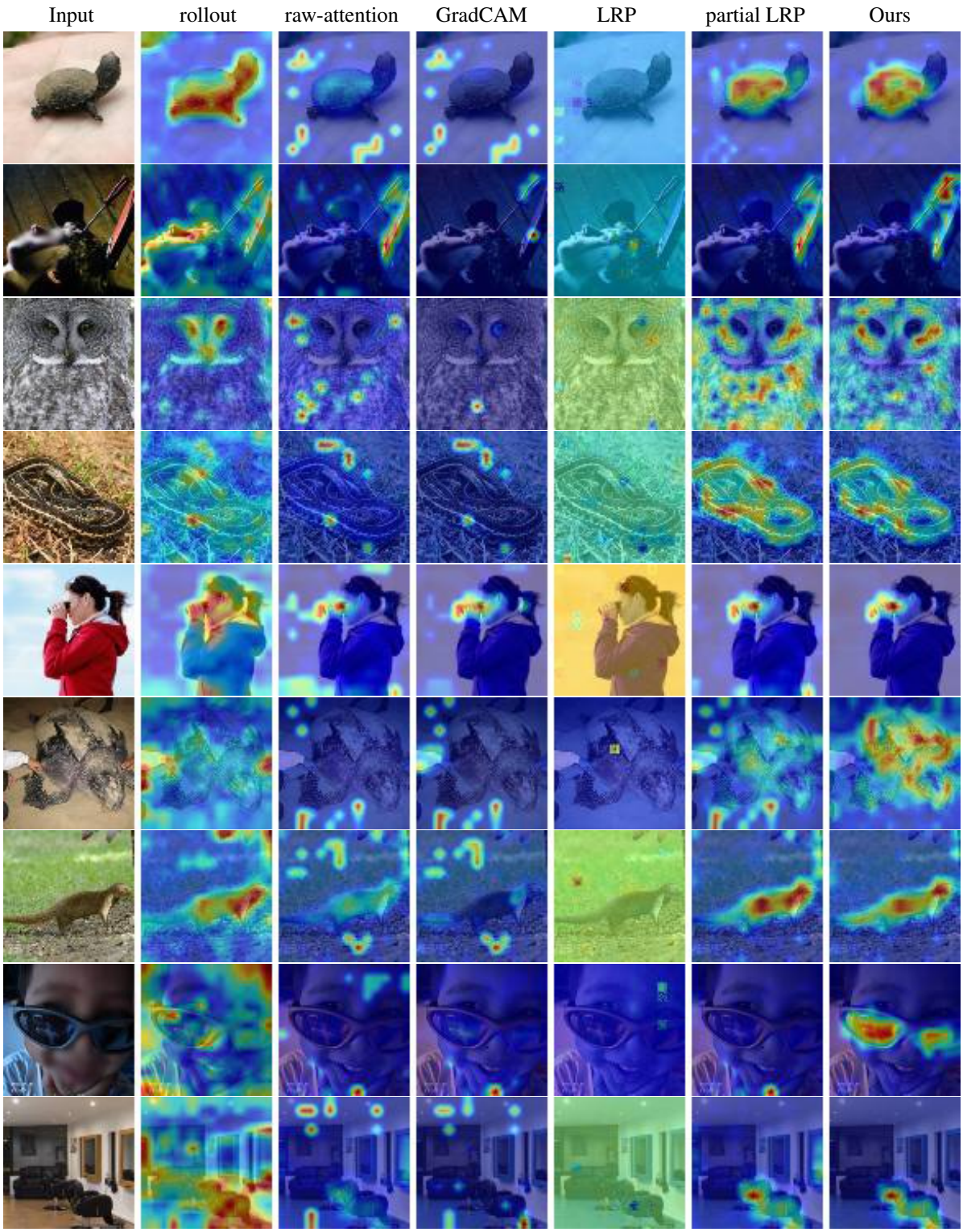


Figure 10: Sample images from ImageNet val-set.

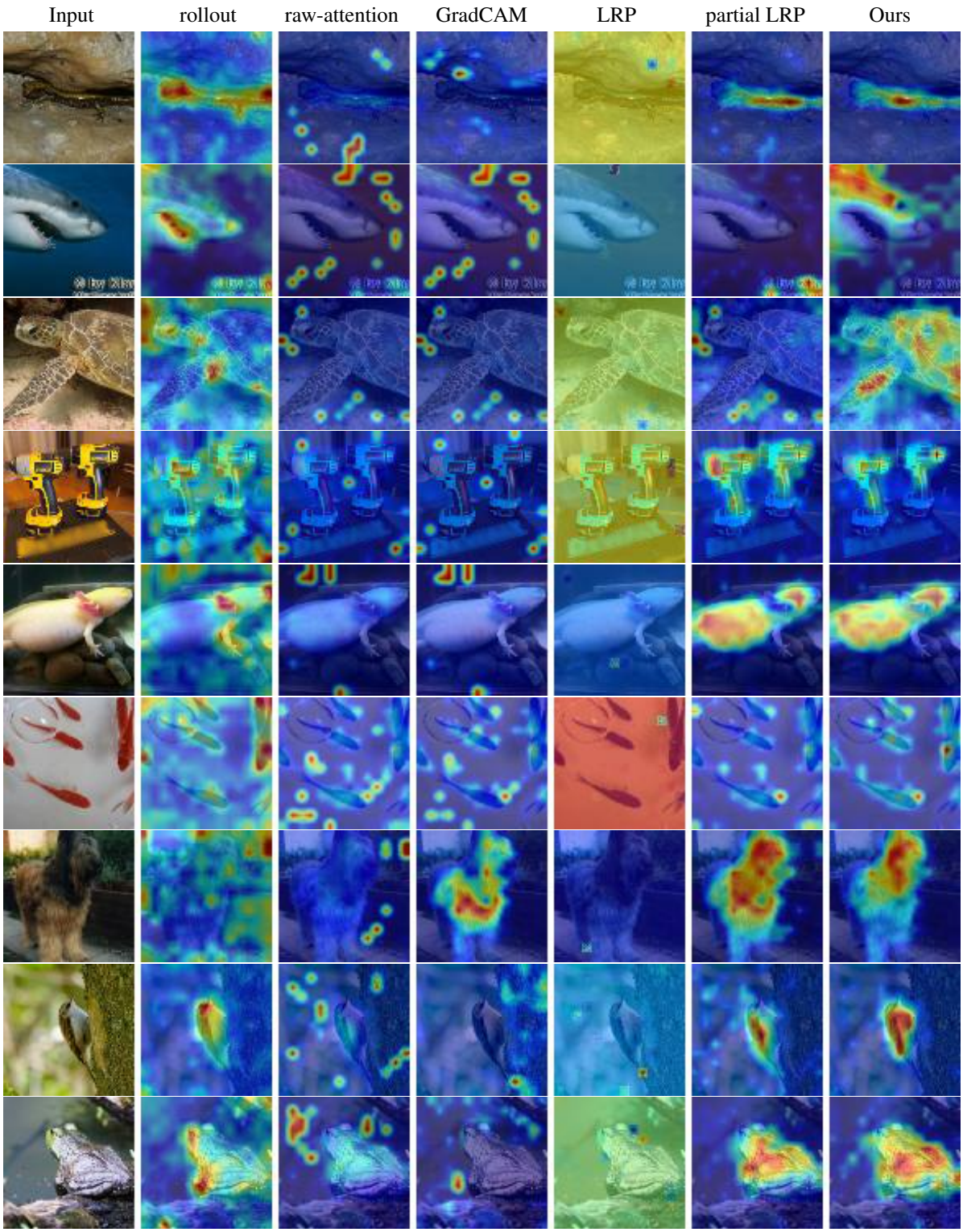


Figure 11: Sample images from ImageNet val-set.

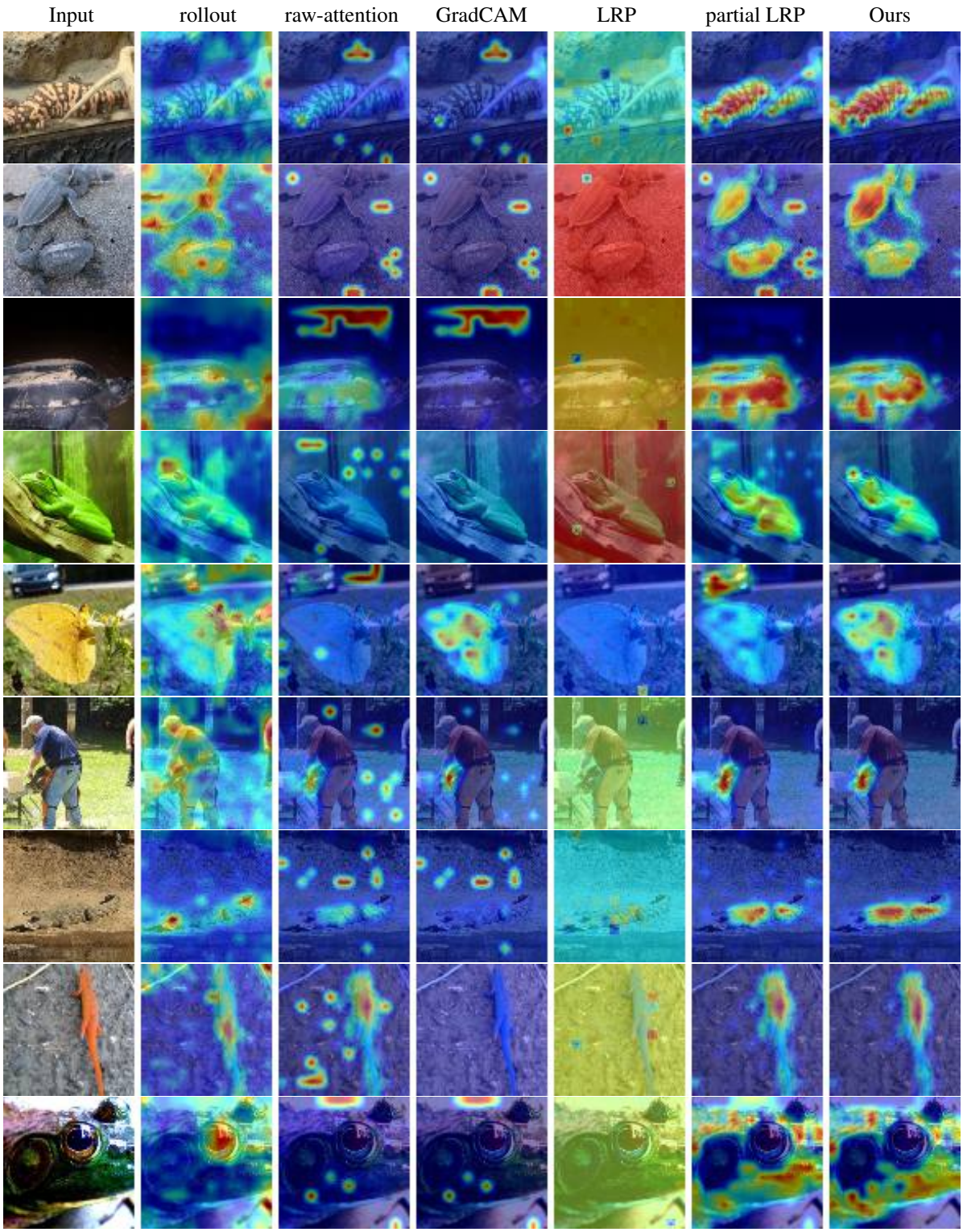


Figure 12: Sample images from ImageNet val-set.

5. Visualizations - Text

In the following visualizations, we use the TAHV heatmap generator for text (https://github.com/jiesutd/Text-Attention-Heatmap-Visualization) to present the relevancy scores for each method, as well as the excerpts marked by humans. For methods that are class-dependent, we present the attributions obtained both for the ground truth class and the counter-factual class.

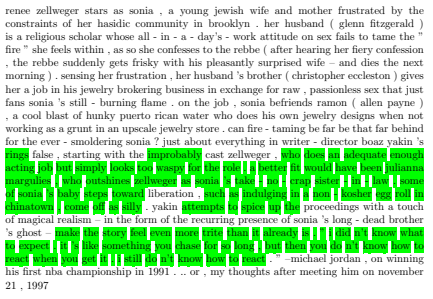
Evidently, our method is the only one that is able to present support for both sides, see panels (b,c) of each image. GradCAM often suffers from highlighting the evidence in the opposite direction (sign reversal), e.g., Fig. 13(g), in which the counter-factual explanation of GradCAM supports the negative, ground truth, sentiment and not the positive one.

Partial LRP (panels d,e) is not class-specific in practice. This provides it with an advantage in the quantitative experiments: Partial LRP highlights words with both positive and negative connotations from the same sentence, which better matches the behavior of the human annotators who are asked to mark complete sentences.

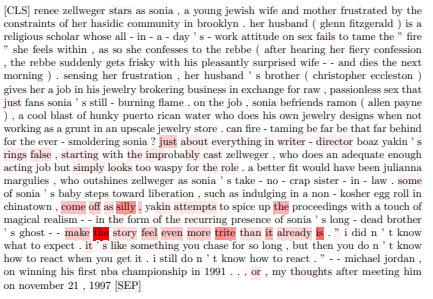
Notice that in most visualizations, it seems that the rollout method focuses mostly on the separation token [SEP], and fails to generate meaningful visualizations. This corresponds to the results presented in the quantitative experiments.

It seems from our results, e.g., Fig. 13(b,c) that the BERT tokenizer leads to unintuitive results. For example, “joyless” is broken down into “joy” and “less”, each supporting different sides of the decision.

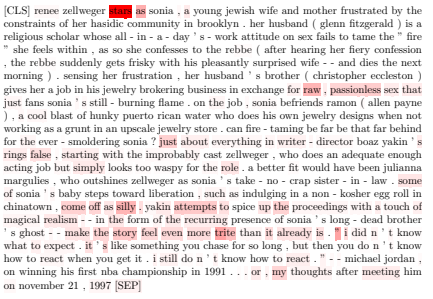
Figure 13: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.



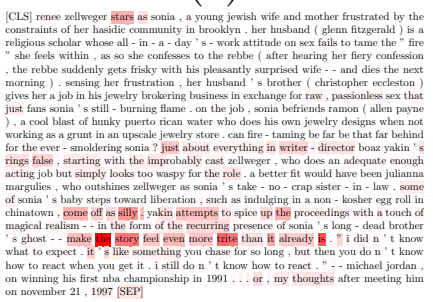
(a)



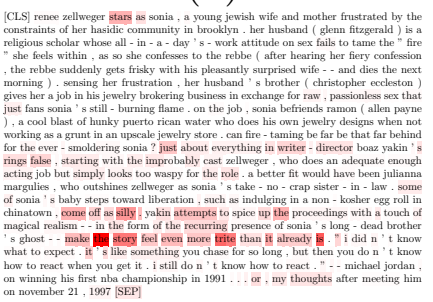
(b)



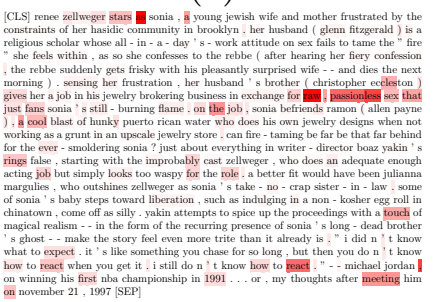
(c)



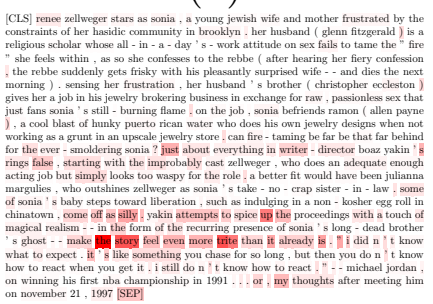
(d)



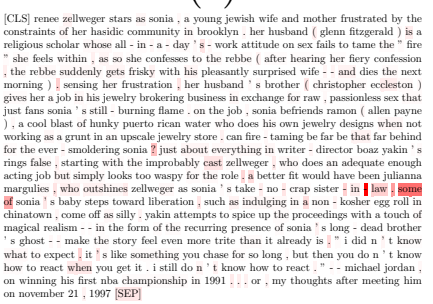
(e)



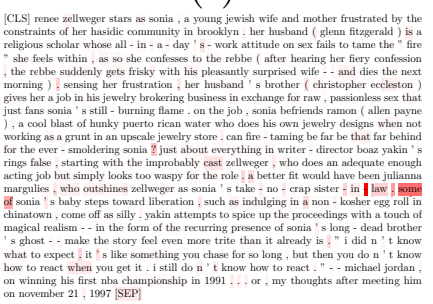
(f)



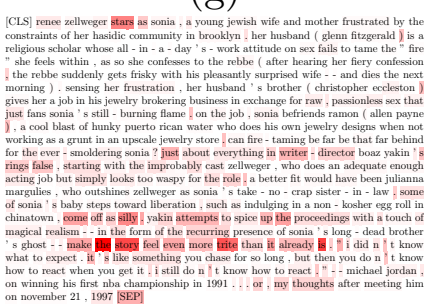
(g)



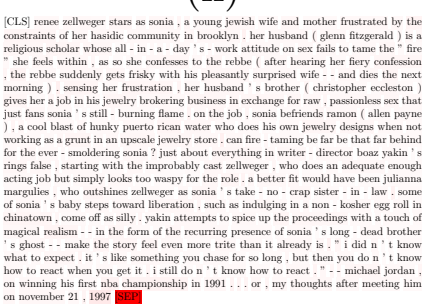
(h)



(i)



(j)



(k)

Figure 14: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 15: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 16: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (negative sentiment). Note that the BERT prediction on this sample was accurate. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 17: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 18: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 19: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 20: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 21: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

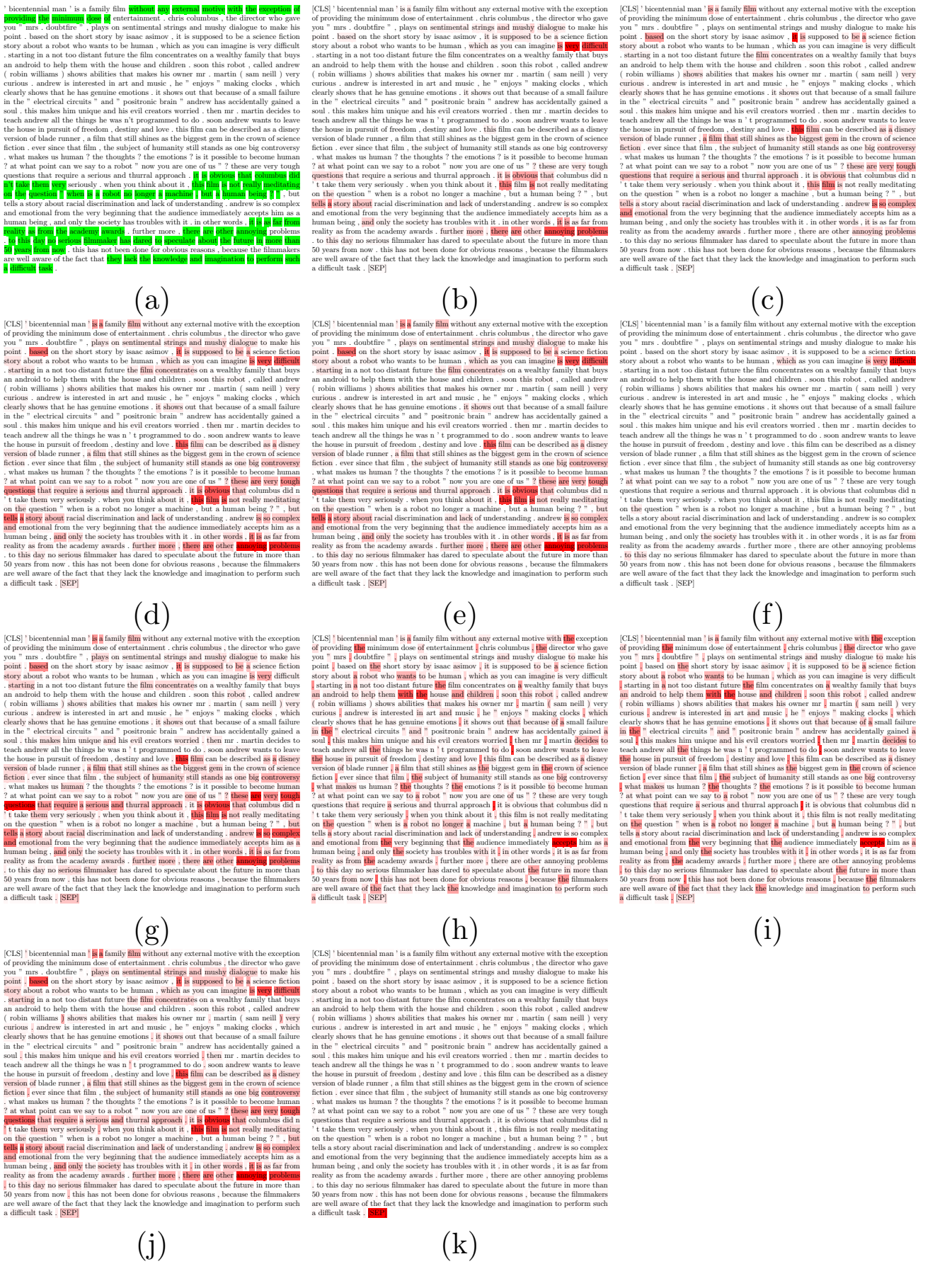


Figure 22: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**negative** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

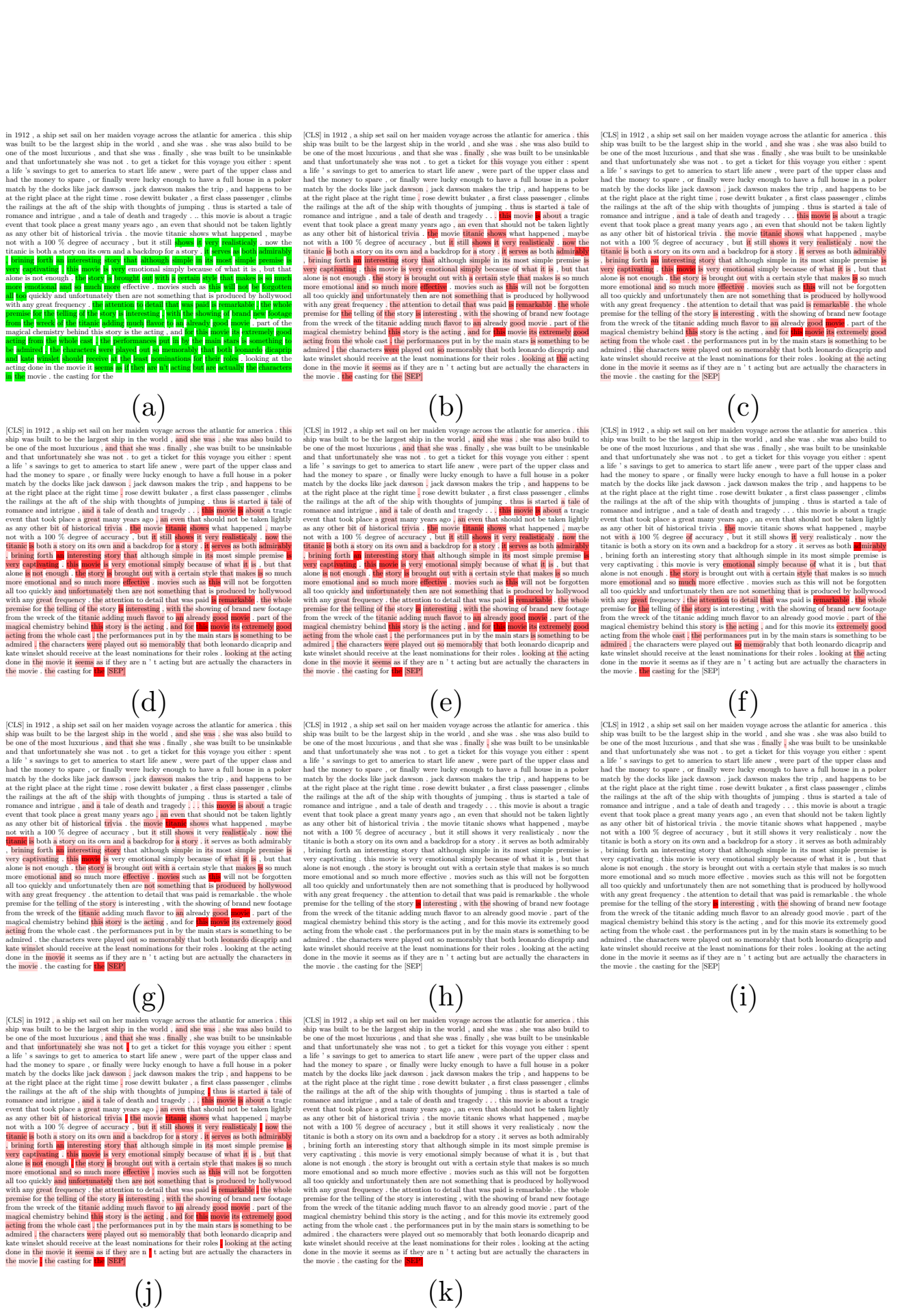


Figure 23: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 24: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

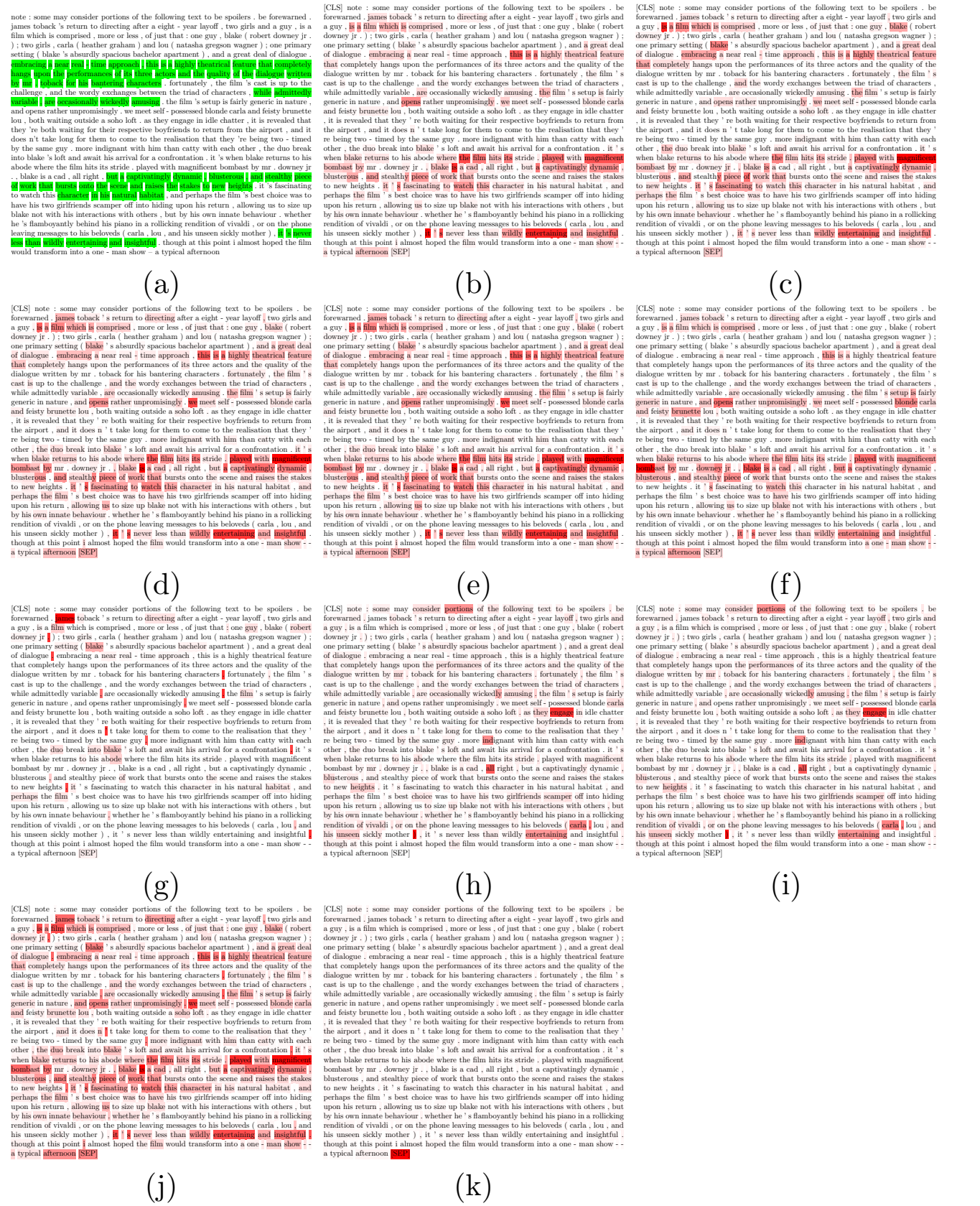


Figure 25: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 26: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 27: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **accurate**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 28: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

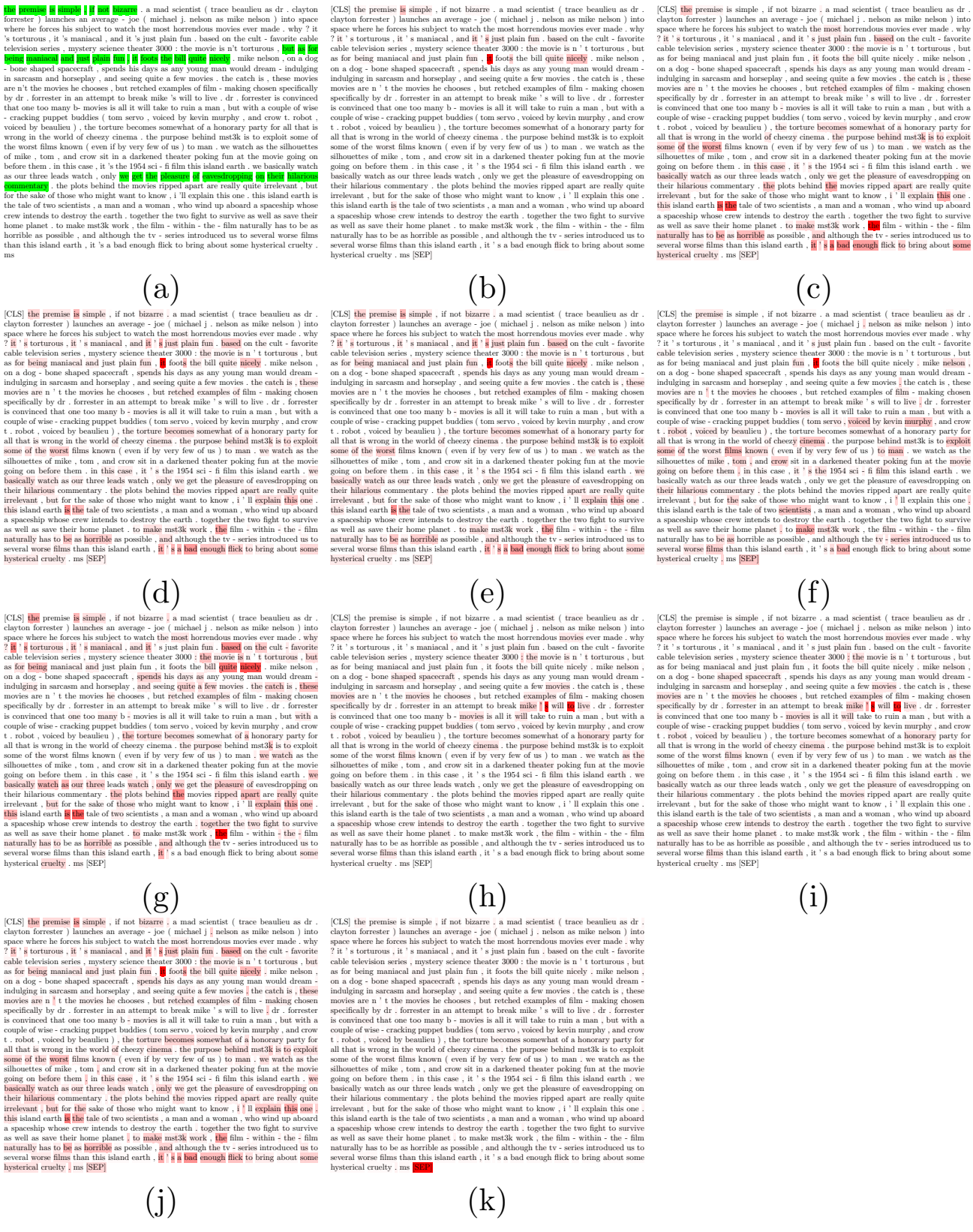


Figure 29: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 30: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.

Figure 31: A visualization of the results. For methods that rely on a specific class for propagation, we present both the ground-truth and counter-factual results. The rollout method, as well as the raw attention methods, are class-agnostic. Some words are split into multiple tokens by the BERT tokenizer. (a) Ground truth (**positive** sentiment). Note that the BERT prediction on this sample was **mistaken**. (b) Our method for the ground truth [GT] class. (c) Our method for the counter-factual [CF] class. (d) Partial LRP for the GT class. (e) Partial LRP for the CF class. (f) GradCAM for the GT class. (g) GradCAM for the CF class. (h) LRP for the GT class. (i) LRP for the CF class. (j) raw-attention. (k) rollout.