Supplementary Materials for Show and Tell: Visually Explainable Deep Neural Nets via Spatially-Aware Concept Bottleneck Models

A. Implementation details

A.1. Local image-concept similarities

The pseudo-algorithm for computing local image-concept similarities using visual prompts is provided below. The operation of drawing a red circle within training image x_n at location (h, w) with radius r is denoted by Circle$(x_n; (h, w, r))$. We use circles with a line width of 2 pixels.

Algorithm 1 Local image-concept similarities

Input: (i) training images $\{x_n\}_{n=1}^{N} \in \mathbb{R}^{3\times H\times W}$; (ii) concept list $\{t_m\}_{m=1}^{M}$; (iii) CLIP's image encoder E_I and text encoder E_T; (iv) circle radius r, grid dimensions $(\tilde{H}, \tilde{W})$.
Output: Spatial concept similarity matrix P for the entire training set.
Initialize: $P \leftarrow \mathbf{0}$.

$d_H \leftarrow \lfloor H/(\tilde{H}-1) \rfloor$, $d_W \leftarrow \lfloor W/(\tilde{W}-1) \rfloor$
$T \leftarrow E_T(\{t_m\}_{m=1}^{M})$ ▷ encode all concepts once
for $n \leftarrow 0$ **to** $N-1$ **do** ▷ iterate over training images
  $X_n^{aug} \leftarrow \emptyset$
  for $h \leftarrow r$ **to** $\tilde{H}-r$ **by** d_H **do** ▷ iterate over grid locations
    for $w \leftarrow r$ **to** $\tilde{W}-r$ **by** d_W **do**
      $X_n^{aug} \leftarrow X_n^{aug} \cup$ Circle$(x_n; (h, w, r))$
    end for
  end for
  $I_n \leftarrow E_I(X_n^{aug})$ ▷ encode entire augmented batch
  $P[n, m, h, w] \leftarrow \frac{I_n^{(h,w)} \cdot T_m}{\|I_n^{(h,w)}\| \|T_m\|}$
end for
return P

A.2. Choosing the grid parameters

We experiment with different settings of the visual prompting grid. In Table 1, we present the classification accuracy obtained using different values for the circle radius r and the grid size $\tilde{H} \times \tilde{W}$, on the ImageNet (left) and CUB-200 (right) datasets. In both cases, the best performance is achieved with $r = 32$ and a grid size of 7×7. We use the same values for Places365.

Grid size	$r = 27$	$r = 32$	$r = 37$
5×5	74.17%	74.37%	75.01%
7×7	74.67%	**75.32%**	75.31%
9×9	75.06%	75.22%	75.22%

(a) Results on ImageNet.

Grid size	$r = 27$	$r = 32$	$r = 37$
5×5	73.36%	73.59%	73.80%
7×7	73.42%	**74.35%**	74.01%
9×9	73.83%	74.12%	73.93%

(b) Results on CUB-200.

Table 1. Classification accuracy for different settings of the visual prompting grid, on the ImageNet (left) and CUB-200 (right) datasets.

A.3. Spatial concept bottleneck layer

Our spatial concept bottleneck layer is comprised of a single 1×1 convolution layer with M output channels and no bias, where M is the number of concepts. Therefore, it requires the same number of parameters as the fully-connected bottleneck layer typically used in non-spatial CBMs, i.e., $D \times M$ where D is the dimensionality of the "black-box" features.

B. Results with ViT backbone

B.1. Classification accuracy

We report the classification results of our SALF-CBM with a ViT-B/16 backbone pre-trained on ImageNet. We experiment with two variations: (1) Only patch tokens are used, reshaped into their original spatial formation; (2) Both patch tokens and the `CLS` token are used, by reshaping the patch tokens into their original spatial formation and concatenating them with the `CLS` token along the channels dimension. For each variation, the model is trained with both sparse and non-sparse classification layers. We compare its results to the corresponding standard model—i.e., using the same backbone model without a bottleneck layer and with a comparable classification layer (sparse or non-sparse). Results are shown in Figure 1. When using a sparse final layer, our model significantly outperforms the corresponding standard model for both backbone versions. With a non-sparse final layer, our model's performance is comparable to the standard model when using the `CLS` token, and is slightly lower when the `CLS` token is excluded.

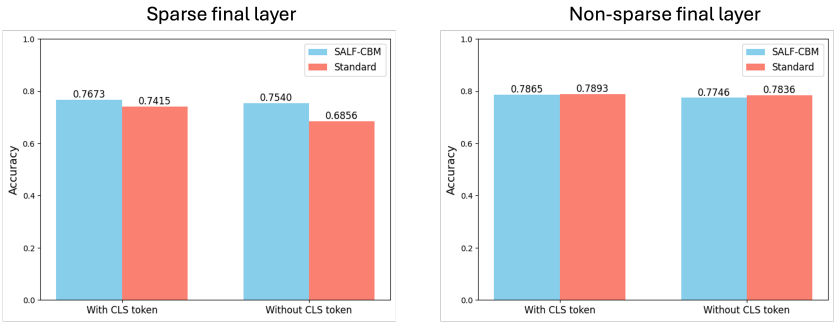

Figure 1. ImageNet classification results with ViT-B/16 backbone, when using a sparse final layer (left) and a dense final layer (right).

B.2. Spatial heatmaps

We present qualitative results of the heatmaps generated by our method when using a ViT-B/16 backbone pre-trained on ImageNet. Similar to section **??**, we train our model on ImageNet using a concept list of the form "An image of a {class}", where {class} refers to each of the ImageNet classes. In Figure 2, we show the heatmaps produced by our method compared to the raw attention maps of the ViT model, for different images from the ImageNet validation set. We observe that our SALF-CBM's heatmaps tend to be more exclusive, while the raw attention maps often include background areas outside the target class object.

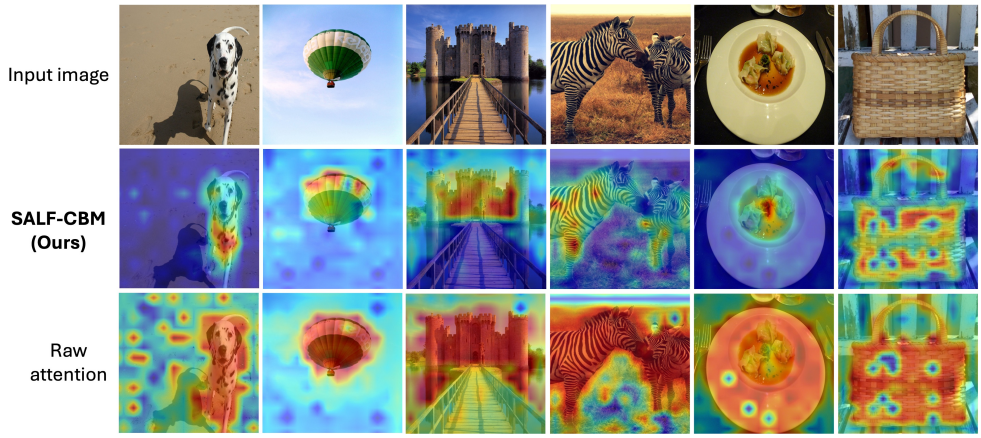

Figure 2. Heatmaps generated by our SALF-CBM with a ViT-B/16 backbone (middle row) for random images from the ImageNet validation set, compared to the raw attention maps of the standard ViT model (bottom row). The ground-truth class of the images (from left to right): "Dalmatian", "Balloon", "Castle", "Zebra", "Consomme" and "Hamper".

C. Bottleneck interpretability validation

C.1. User study questions examples

We show an example of the *semantic consistency* and *concept accuracy* questions used in our user study, as described in the main paper.

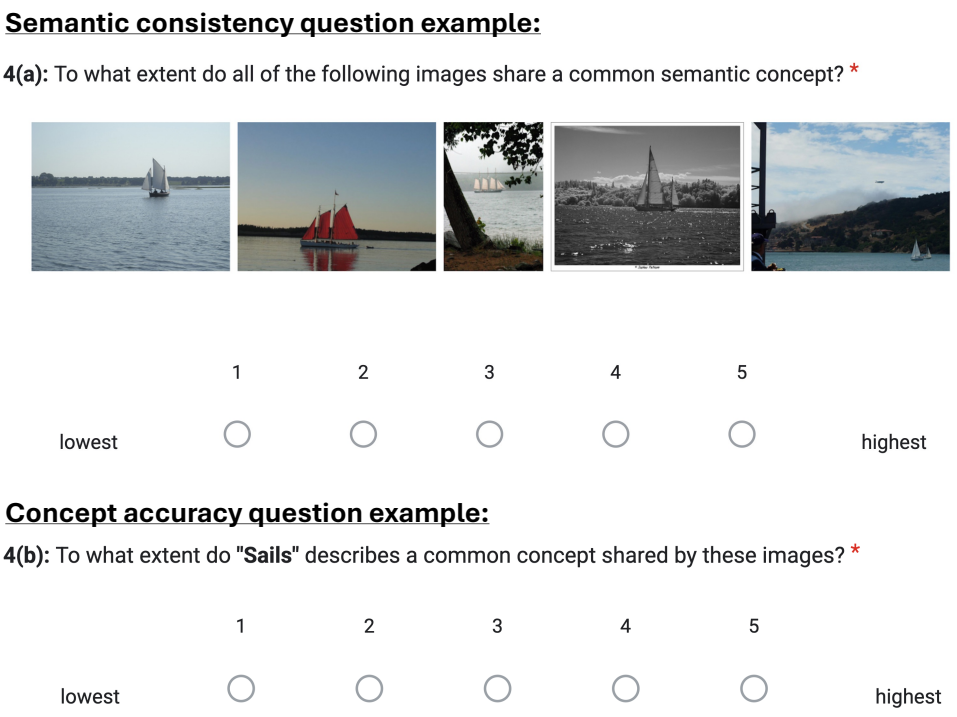

Figure 3. User study question example.

C.2. Concept neurons validation

In addition to the user study described in the main paper, we qualitatively validate that neurons in our concept bottleneck layer indeed correspond to their designated target concepts. We train a SALF-CBM on each dataset (ImageNet, Places365 and CUB-200) and randomly select 5 neurons from its concept bottleneck layer. For each neuron, we retrieve the top-3 images with the highest global concept activation c^* from the corresponding validation set. As shown in Figure 4, the target concept of each neuron highly corresponds to the retrieved images.

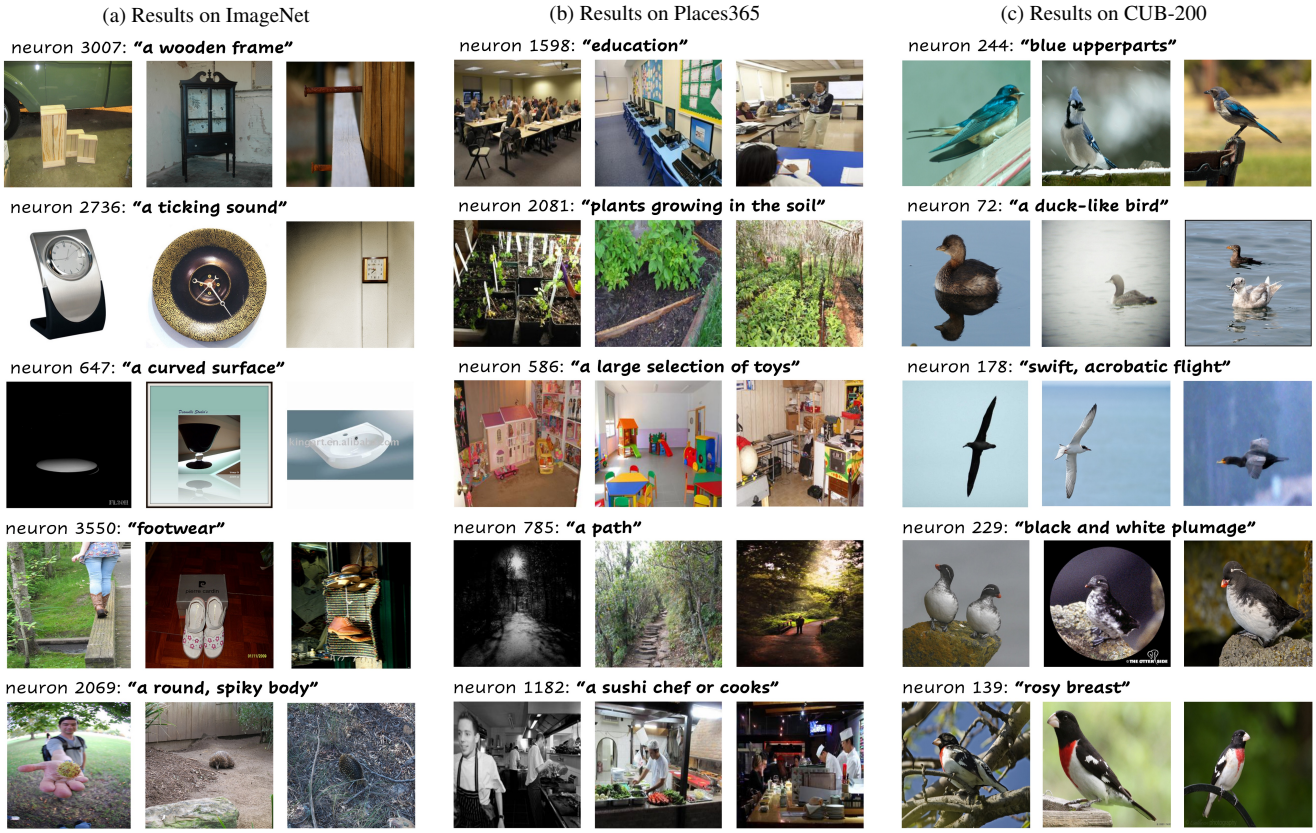

Figure 4. **Qualitative validation of concepts learned by CBL neurons.** Top 3 images with the highest concept activation c^*, for 5 randomly selected neurons in the CBL. The retrieved images are highly correlated with the neuron's target concept. Results are shown for ImageNet (left), Places365 (middle) and CUB-200 (right) datasets.

D. Additional experiments

D.1. Explanations on different datasets

We present qualitative results of concept-based and spatial explanations across images from different datasets: ImageNet (Figure 5), Places365 (Figure 6) and CUB-200 (Figure 7). For each image, we present the most important concepts used by our SALF-CBM to classify the image, along with a heatmap of one of these concepts. By offering both concept-based explanations and their visualizations on the input image, our model enables a comprehensive understanding of its decision-making process. For example, in the second row of Figure 6, we see that our model correctly classified the image as "athletic field, outdoor" by identifying and accurately localizing the track behind the athlete.

D.2. Explaning multi-class images

We demonstrate our method's ability to produce class-specific explanations in Figure 8. Given an image x with two possible classes, $\hat{y} = l_1$ and $\hat{y} = l_2$, we compute the concept contribution scores for predicting each class, i.e., $S(x, m, \hat{y} = l_1)$ and $S(x, m, \hat{y} = l_2)$, as described in Section **??**. For each image, we present the concepts with the highest contribution scores along with the heatmap of the most contributing concept.

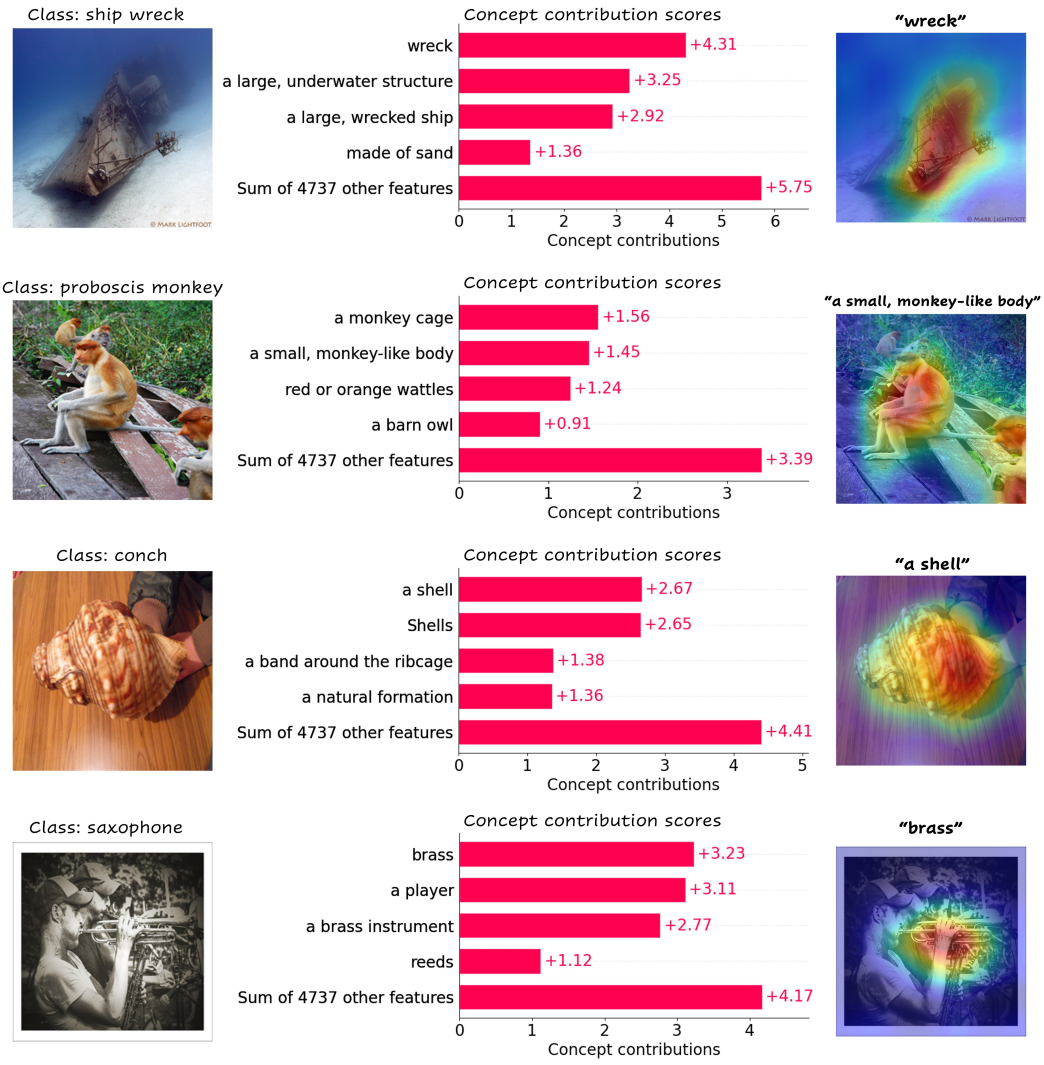

Figure 5. Concept-based and visual explanations on ImageNet.

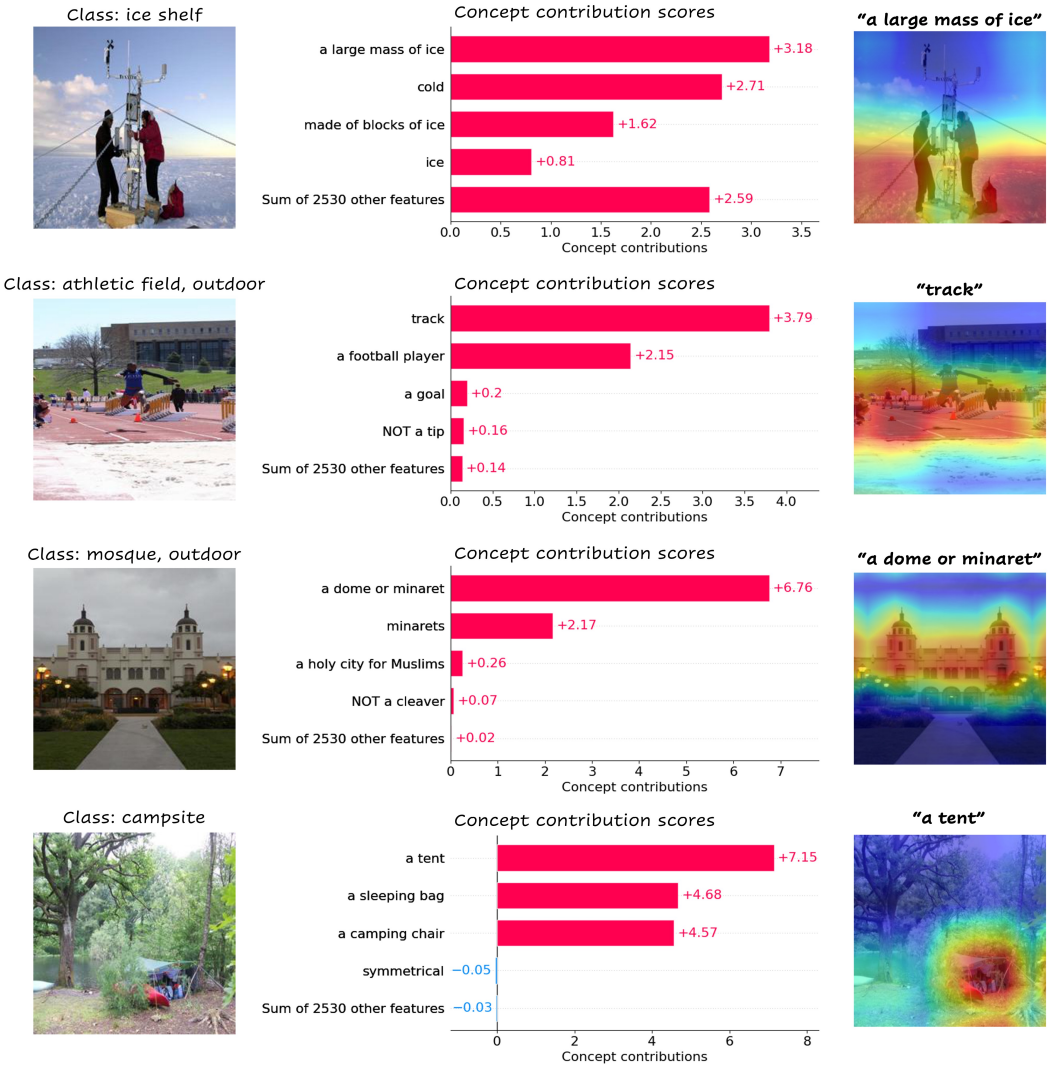

Figure 6. Concept-based and visual explanations on Places365.

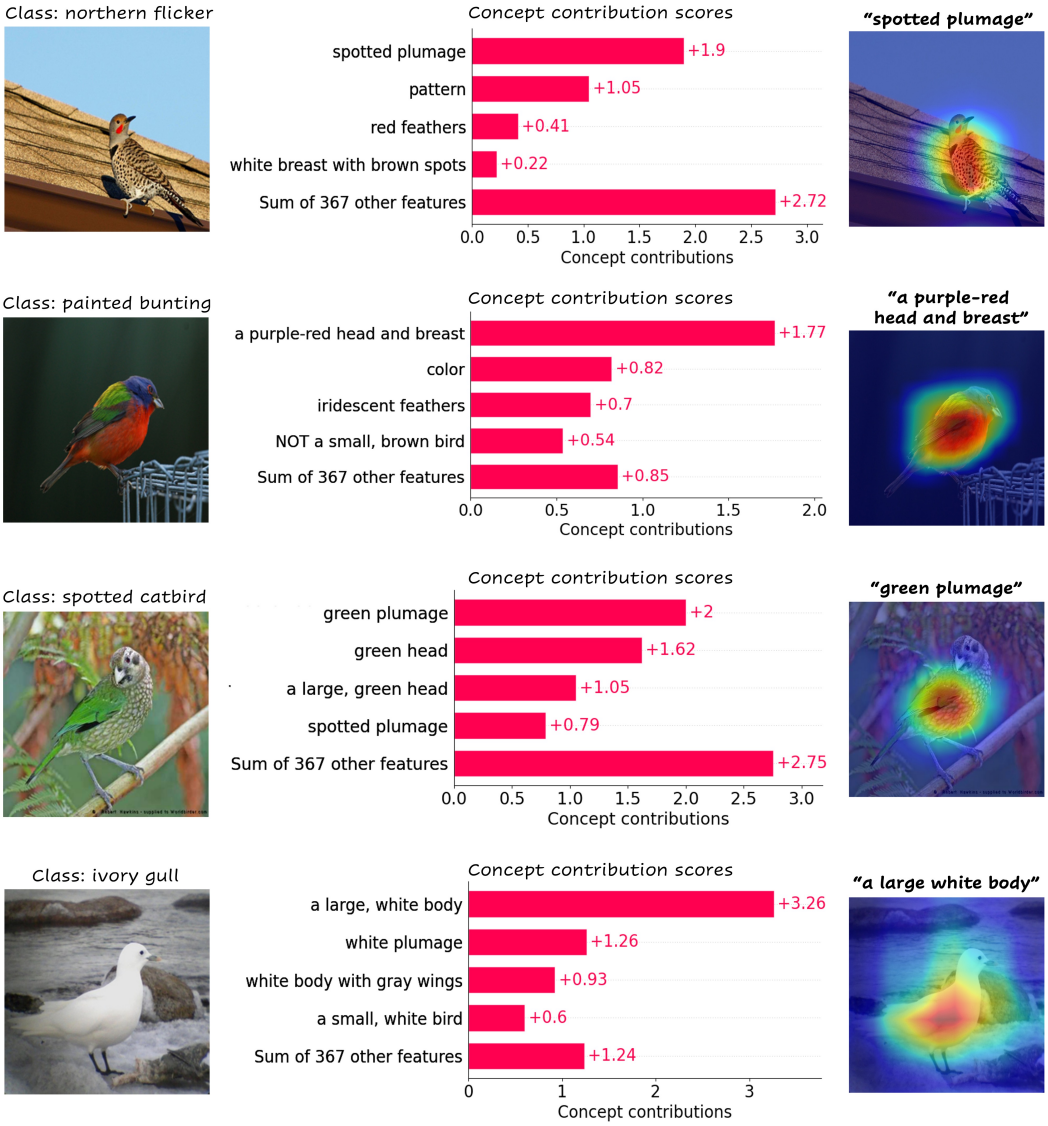

Figure 7. Concept-based and visual explanations on CUB-200.

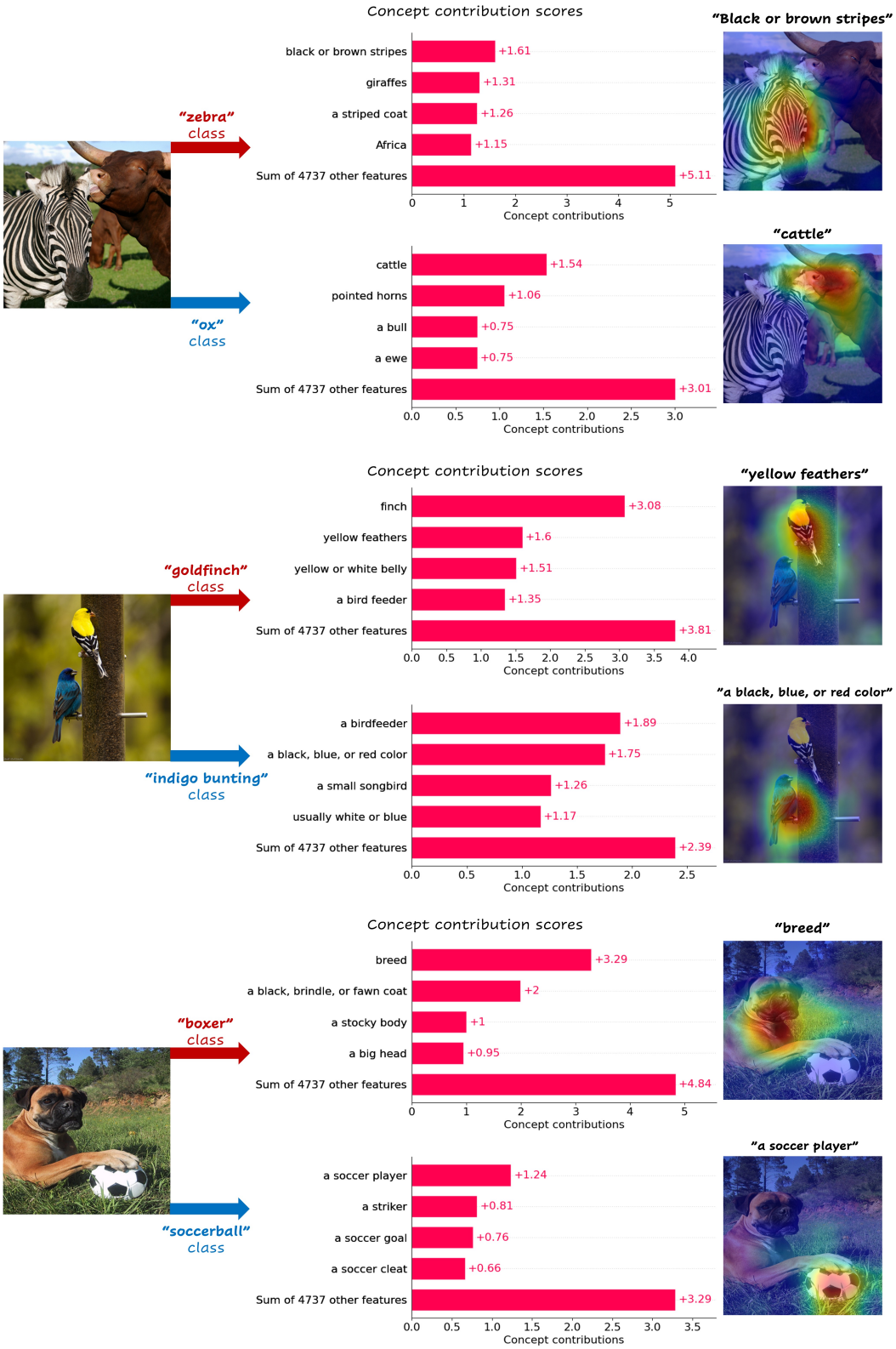

Figure 8. **Explaining predictions on multi-class images.** For each image, we present the most contributing concepts identified by SALF-CBM for explaining two different output classes that fit the image. We show the heatmap of the top concept for each class.

E. Additional heatmaps results

E.1. Visualizing multiple concepts

We demonstrate our method's ability to localize multiple concepts within a single image. In Figure 9, we present qualitative results on several images from the ImageNet validation set. For each image, we show three heatmaps generated by our SALF-CBM, each corresponding to a different visual concept.

E.2. Visualizing concepts in videos

By applying SALF-CBM to video sequences in a frame-by-frame manner, we achieve visual tracking of specific concepts. In Figure 10, we demonstrate this capability on several videos from the DAVIS 2017 dataset using a SALF-CBM trained on ImageNet. Despite being trained on a completely different dataset, our model successfully localizes various concepts throughout these videos. For example, in the "soccer ball" video at the top of the figure, the soccer ball is accurately highlighted, even when it is partially occluded in the last frame.

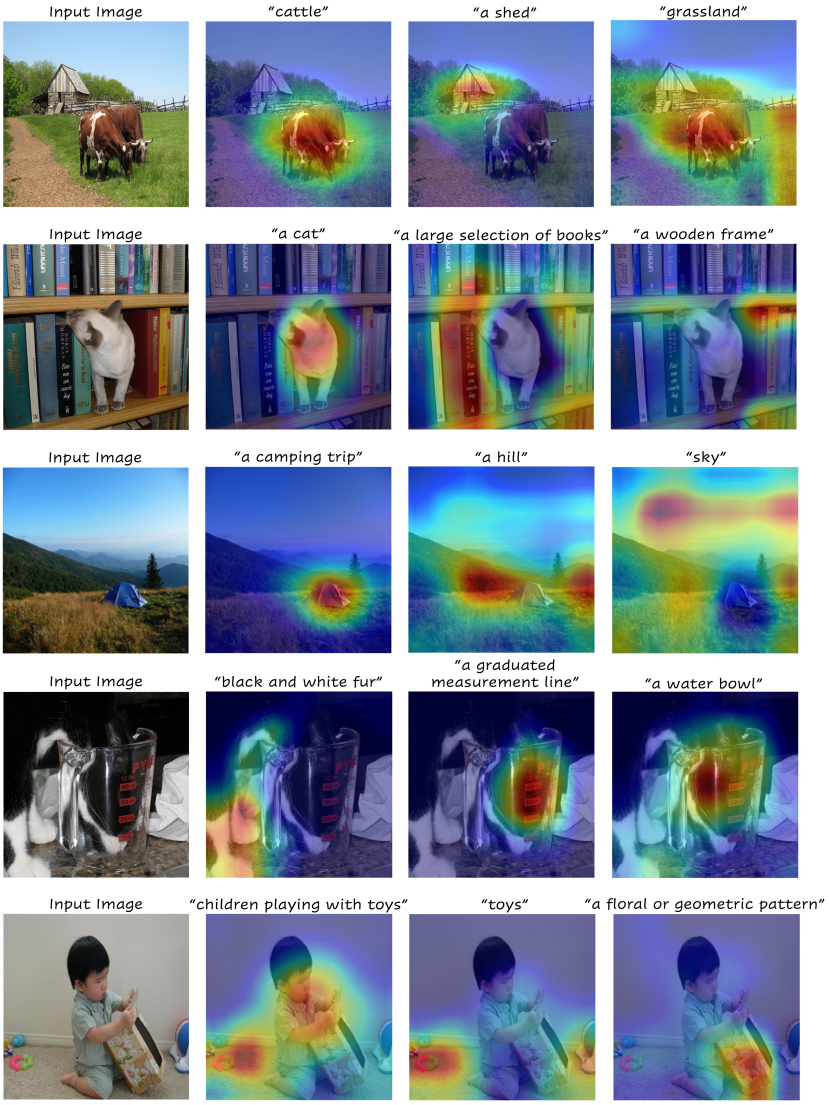

Figure 9. **Localizing multiple concepts in images.** For each image, we present three heatmaps, each corresponding to a different visual concepts.

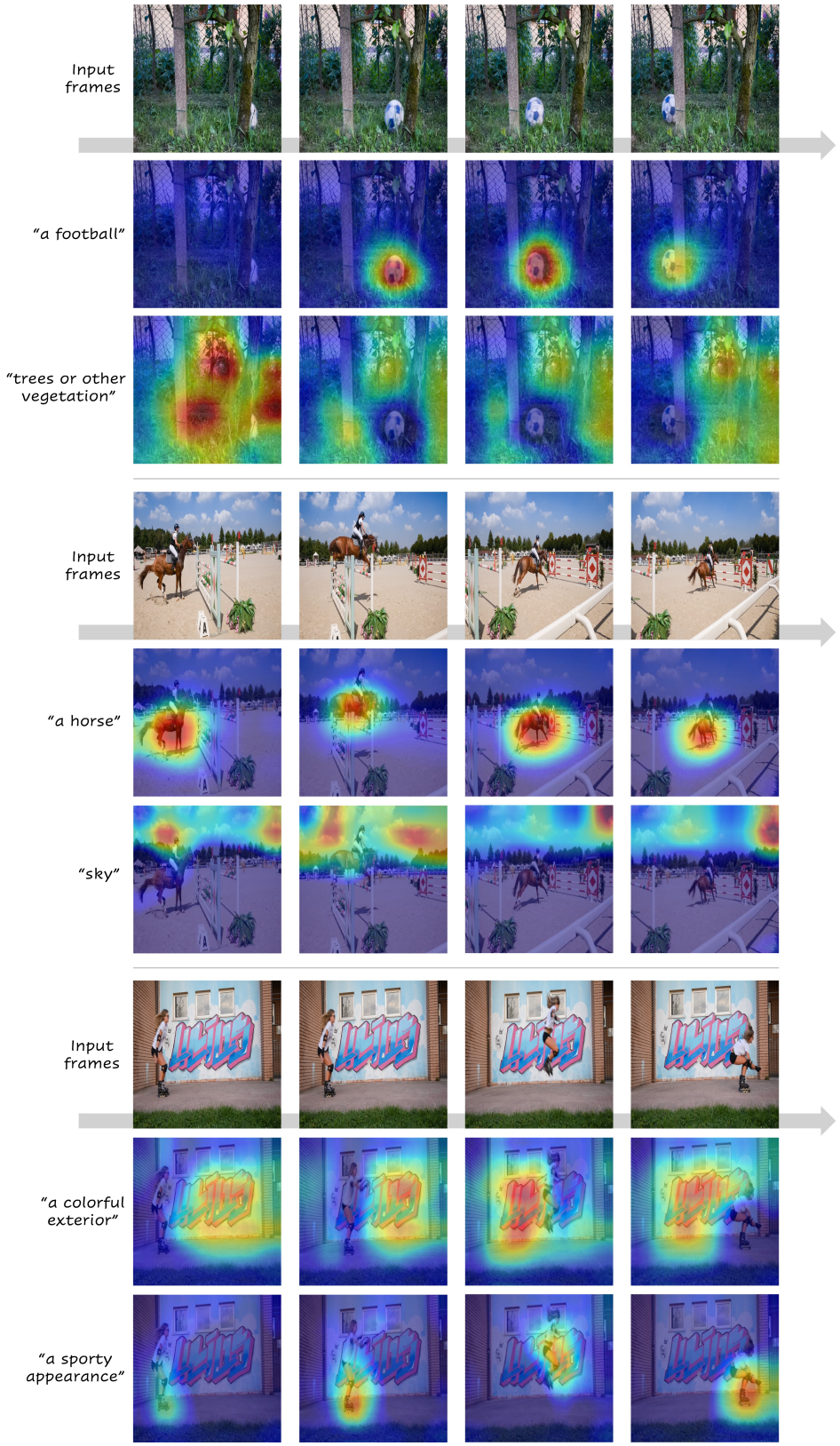

Figure 10. **Visualizing concepts in videos.** By applying SALF-CBM in a frame-by-frame manner, one can visually track concepts over time. Videos are from the DAVIS 2017 dataset (from top to bottom): "soccer ball", "horsejump-high" and "rollerblade".