Explaining CLIP Zero-shot Predictions Through Concepts

Supplementary Material

A. Implementation Details

This section provides additional details on concept construction, training procedures, and evaluation settings used for EZPC. We omit definitions and equations already introduced in the main paper and focus on implementation-specific clarifications.

A.1. Backbone and Embedding Extraction

We primarily use the CLIP RN50 backbone, with additional experiments using ViT-B/32, ViT-L/14 (via OpenAI CLIP), and SigLIP ViT-SO400M/14 (via OpenCLIP). For every dataset, we precompute:

- ℓ_2-normalized image embeddings for all train/validation images,
- text embeddings for all dataset class names,
- text embeddings for the corresponding concept vocabulary.

All embeddings remain frozen during the training of EZPC.

A.2. Concept Vocabulary per Dataset

We follow the LF-CBM [22] and use their GPT-3-generated concept sets.

- **ImageNet-1k & ImageNet-100:** We use the 4,751 ImageNet-derived GPT-3 concepts from LF-CBM. No additional concepts are merged.
- **CIFAR-100, CUB, Places365:** The original LF-CBM concept sets for these datasets are limited. Therefore, we merge the dataset's own LF-CBM concepts with the larger ImageNet concept set to obtain a sufficiently expressive concept space. After merging, duplicate concepts are removed.

All concepts are encoded once using the corresponding model's text encoder.

A.3. Training the Concept Projection Matrix A

The learnable matrix $A \in \mathbb{R}^{d\times m}$ maps CLIP/SigLIP embeddings into the m-dimensional concept space. The matrix is initialized as

$$A^{(0)} = \Phi, \tag{10}$$

where Φ is the CLIP concept embedding matrix.

The training objective consists of a matching term and a reconstruction term, as defined in the main paper. The scalar λ (typically $\lambda = 1$, and $\lambda = 5$ for CUB and Places365) controls the relative weight of the reconstruction loss. No orthogonality or sparsity regularizers are used.

We optimize only A (all backbone parameters remain frozen) using Adam with learning rate 10^{-2} for 10,000 iterations. After every epoch we renormalize all concept vectors

$$A_{:,j} \leftarrow \frac{A_{:,j}}{\|A_{:,j}\|_2}, \tag{11}$$

which stabilizes the concept geometry and prevents drift during training.

A.4. Zero-Shot and Generalized Zero-Shot Evaluation

Let $\mathcal{Y}_S$ and $\mathcal{Y}_U$ denote the seen and unseen class sets.

Zero-shot learning (ZSL). Evaluation is performed over unseen classes only. For each test image x, the predicted label is

$$\hat{y} = \underset{k:\, y_k\in\mathcal{Y}_U}{\arg\max} \langle c_x, c_k\rangle, \tag{12}$$

and accuracy is computed over the unseen test set $\mathcal{D}_U$ as

$$\mathrm{Acc}_U = \frac{1}{|\mathcal{D}_U|} \sum_{(x,y_k)\in\mathcal{D}_U} \mathbb{1}[\hat{y} = y_k]. \tag{13}$$

Generalized zero-shot learning (GZSL). Predictions are made over the combined label set $\mathcal{Y}_G = \mathcal{Y}_S \cup \mathcal{Y}_U$. We report accuracies on seen and unseen classes separately (Acc_S, Acc_U), along with the harmonic mean

$$H = \frac{2 \times \mathrm{Acc}_S \times \mathrm{Acc}_U}{\mathrm{Acc}_S + \mathrm{Acc}_U}. \tag{14}$$

A.5. Time Analysis Protocol

We measure inference latency on the ImageNet-100 validation set using a single NVIDIA H100 GPU. For each method, we report two quantities: (i) the *embedding time*, which measures only the concept decomposition step (excluding the shared CLIP forward pass), and (ii) the *full pipeline time*, which includes CLIP encoding plus the method-specific decomposition.

To obtain stable estimates, we first run warm-up iterations, then time each method over the full validation set and record per-image latencies. We report the median latency together with 95% confidence intervals.

For EZPC, the decomposition reduces to a single matrix multiplication $c_x = v_x A$, which adds negligible overhead to CLIP's forward pass. To verify this statistically, we apply a Wilcoxon signed-rank test comparing per-image latencies of CLIP and EZPC, yielding $p = 0.31$ (no significant difference). In contrast, SpLiCE requires iterative sparse coding

per image, and Z-CBM performs a retrieval-based regression over the concept bank, both of which involve iterative optimization and result in substantially higher latency.

A.6. Concept-Region Alignment

Generating spatial heatmaps. To produce concept-level spatial activation maps, we extract patch-level representations from the CLIP RN50 backbone. We register a forward hook on `layer4` of the ResNet visual encoder to capture the spatial feature map before attention pooling, yielding $N_p = 7 \times 7 = 49$ patch embeddings. Each patch is passed through CLIP's attention pooling projections (v_proj, c_proj) and ℓ_2-normalized to obtain patch embeddings $\{p_i\}_{i=1}^{N_p}$ in $\mathbb{R}^d$.

Each patch embedding is then projected into the concept space via A to obtain $z_i = p_i A \in \mathbb{R}^m$. We normalize each z_i by its maximum absolute value and mean-center across concepts to ensure comparable activation scales. For a given concept j, the spatial heatmap is formed by extracting the j-th coordinate of each normalized patch, applying ReLU, and reshaping to the 7×7 grid. The result is bilinearly upsampled to the original image resolution for visualization.

Quantitative evaluation metrics. We evaluate spatial alignment on CUB-200-2011 using ground-truth segmentation masks. For each class, we manually specify a *positive* concept (e.g., *a blue-gray body* for Indigo Bunting) and a *negative* concept (e.g., *a red face*), compute heatmaps for both across all images of that class, and compare against the binary segmentation mask M (resized to the patch grid). We report:

- **Pointing Accuracy**: fraction of images where the maximally activated patch falls inside M.
- **Inside Activation Ratio**: proportion of total activation mass falling inside M.
- **IoU@τ%**: intersection-over-union between M and the binary mask obtained by thresholding the heatmap at the $(100-\tau)$-th percentile. We report IoU@10% and IoU@20%.

For each metric, we report the mean and standard deviation across all images in the class. As shown in Table 5 of the main paper, positive concepts consistently localize on the object region, while negative concepts produce near-zero alignment scores.

B. Faithfulness & Causal Validation

A core requirement for concept-based interpretability is faithfulness: the degree to which the concepts identified as important are causally responsible for the model's predictions. In this section, we evaluate whether the learned concept space A discovered by EZPC produces faithful explanations of CLIP's zero-shot classifier.

All experiments are conducted on **ImageNet-100** using the **CLIP RN50** backbone. Because ImageNet-100 is large, we evaluate faithfulness on a randomly sampled subset of validation images, ensuring stable estimates while keeping computation tractable.

B.1. Concept Removal for Causal Testing

As defined in the main paper, the concept-wise interaction score between image x and class y_k is $s_{x,k} = c_x \odot c_k$, where the j-th entry $s_{x,k}^{(j)}$ measures how strongly concept j contributes to the prediction. Let $\mathcal{J}_n(x, y_k)$ denote the indices of the top-n concepts ranked by $s_{x,k}^{(j)}$. To test causal influence, we ablate the top-n influential concepts by removing their contribution from the scoring function

$$f'(x, y_k) = f(x, y_k) - \sum_{j \in \mathcal{J}_n(x, y_k)} s_{x,k}^{(j)}, \tag{15}$$

where $f(x, y_k) = \langle c_x, c_k \rangle$ is the concept-space logit reconstructed via EZPC.

Faithfulness is quantified as the expected logit drop

$$\Delta_n = \mathbb{E}_{(x,y_k)} \big[f(x, y_k) - f'(x, y_k) \big]. \tag{16}$$

Higher Δ_n values imply stronger causal reliance on the discovered concepts.

B.2. Faithfulness Results

Mean Logit Drops. Table 7 reports the mean logit drop and prediction flip rate after ablating the top-n most influential concepts, averaged across the sampled ImageNet-100 validation images.

Table 7. **Faithfulness results on ImageNet-100 (CLIP RN50).** Mean logit drop and prediction flip rate after ablating the top-n most influential concepts.

Top-n	Logit Drop	Flip Rate
1	0.0306	0.059
3	0.0816	0.099
5	0.1263	0.132
10	0.2256	0.169

Both metrics increase monotonically with n: ablating more high-ranked concepts results in larger logit drops and higher flip rates, confirming that the learned concept directions are causally involved in reconstructing CLIP's similarity structure.

Distributional Effects. Figure 7 shows the full empirical distributions of logit drops for $n = \{1, 3, 5, 10\}$.

As n increases, the distributions shift consistently to the right, indicating that the effect is not limited to a few outlier images but holds broadly across the dataset.

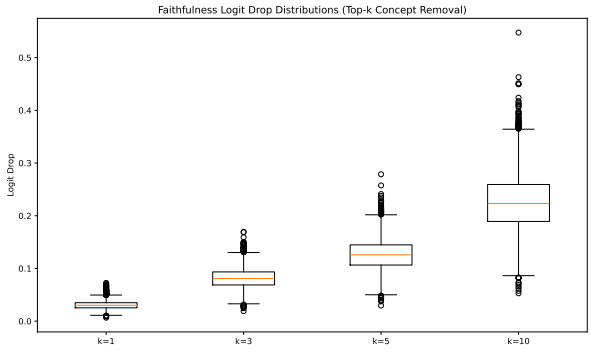


Figure 7. **Faithfulness distributions for n=1, 3, 5, 10.** Removing more highly ranked concepts yields consistently larger drops in model confidence.

B.3. Causal Intervention: Top-10 vs. Random-10

To distinguish causal influence from mere correlation, we compare removing the top-10 influential concepts with removing 10 random concepts

$$\Delta_{\text{top-10}} \quad \text{vs.} \quad \Delta_{\text{rand-10}}. \tag{17}$$

Table 8 reports prediction flip statistics when removing the top-10 most influential concepts compared to removing 10 random concepts. Removing the top-10 concepts changes the predicted class for 16.9% of the evaluated samples, whereas removing 10 random concepts causes flips in only 1.4%. The large gap between these two settings indicates that the discovered concepts correspond to directions that are causally used by CLIP for decision making rather than reflecting spurious correlations.

Table 8. **Top-10 vs Random-10 concept removal.** Prediction flip counts and flip rates on ImageNet-100 (5000 samples).

Removal type	Flip Count	Flip Rate
Top-10 concepts	845	0.169
Random-10 concepts	70	0.014

Figure 8 shows that top-10 removal induces a much larger logit decrease than random removal, whose distribution remains tightly centered near zero. This separation is a strong indicator of true causal involvement: if concept scores merely reflected correlations or dataset priors, random removal would produce similar changes, which it does not.

B.4. Discussion and Takeaways

Across all faithfulness and causal tests, the learned concept space A demonstrates:

- **Causal responsibility:** Removing top-ranked concepts reliably degrades classifier confidence.
- **Stable, monotonic attribution:** Logit drops increase smoothly with n, consistent with the additive structure of the concept-space logit.

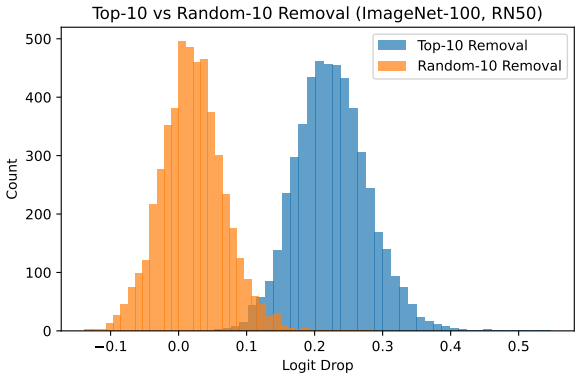


Figure 8. **Causal intervention analysis.** Removing the top-10 influential concepts produces a substantially larger drop in the predicted class logit than removing 10 random concepts.

- **Robustness to dataset variability:** Distributions shift consistently, indicating that effects are widespread and not image-specific.
- **No need for sparsity:** Despite dense activations, explanations remain faithful. Concept rankings are sufficient for causal attribution.

These results confirm that the discovered concept axes form a faithful and causally meaningful basis for explaining CLIP's zero-shot decisions. The concept directions are not merely descriptive; they directly mediate model predictions in a measurable and controllable manner.

C. Concept Space Structure Analysis

This section provides an extended analysis of the geometry and activation behavior of the learned concept space A compared to the original CLIP concept space Φ. All analyses are conducted on **ImageNet-100** using the **CLIP RN50** backbone, matching the quantitative setup of the main paper. Our goal is to understand how optimization alters the concept space, whether semantic identity is preserved, and whether the learned space is geometrically stable, coherent, and suitable for interpretation.

C.1. How the Concept Spaces Are Obtained

The original CLIP concept matrix Φ is created by encoding each concept name j using the CLIP text encoder:

$$\Phi_j = \text{normalize}\big(f_{\text{text}}(\text{“a photo of } j\text{”})\big). \tag{18}$$

The learned concept space A is obtained from our optimization objective, using the same concept vocabulary.

For an image x and class label y_k, we compute the image embedding $v_x \in \mathbb{R}^d$ and text embedding $t_k \in \mathbb{R}^d$

$$v_x = \text{normalize}(f_{\text{img}}(x)), \qquad t_k = \text{normalize}(f_{\text{text}}(y_k)). \tag{19}$$

Following the explanation model of the main paper, the image-side, label-side, and joint concept activations are

$$c_x = v_x A, \quad c_k = t_k A, \quad s_{x,k} = c_x \odot c_k. \tag{20}$$

These activations form the basis for all structure measurements in this section, ensuring full consistency with the interpretability mechanism.

C.2. Summary of Quantitative Results

Table 9 summarizes the key geometric properties of A and Φ. We provide detailed explanations in the subsections below.

Table 9. **Summary of concept space statistics.** All measurements are computed on ImageNet-100 using CLIP RN50. The learned space A preserves semantic identity while becoming more compact and uniformly correlated.

Metric	**CLIP** (Φ)	**Learned** (A)
Alignment Mean / Median	-	0.651 / 0.648
Alignment Std	-	0.036
Alignment Min / Max	-	0.559 / 0.822
Total PCA Variance	0.1396	0.1047
Top-10 Variance Fraction	0.4586	0.4364
Off-diagonal Mean ($\Phi^\top \Phi$ vs. $A^\top A$)	0.6920	0.7461
Off-diagonal Std	0.0705	0.0593

These results support the main claim that our learned concept space remains semantically meaningful and structurally well-behaved.

C.3. Alignment Between A and Φ

We evaluate how much each learned concept direction A_j deviates from its original CLIP counterpart Φ_j using cosine alignment

$$\text{align}(A_j, \Phi_j) = \left\langle \frac{A_j}{\|A_j\|}, \frac{\Phi_j}{\|\Phi_j\|} \right\rangle. \tag{21}$$

Figure 9 shows that alignment scores are tightly clustered around 0.65, with a standard deviation of only 0.036 and minimum/maximum values of 0.559 and 0.822. This distribution demonstrates two important facts:

- The learned concept directions preserve **semantic grounding**.
- Optimization introduces **controlled refinements** rather than large distortions.

Thus, each learned concept remains strongly related to its original semantic meaning, ensuring stable and interpretable explanations.

C.4. PCA Geometry

To assess global geometry, we perform PCA on A and Φ, treating concepts as data points in the embedding space.

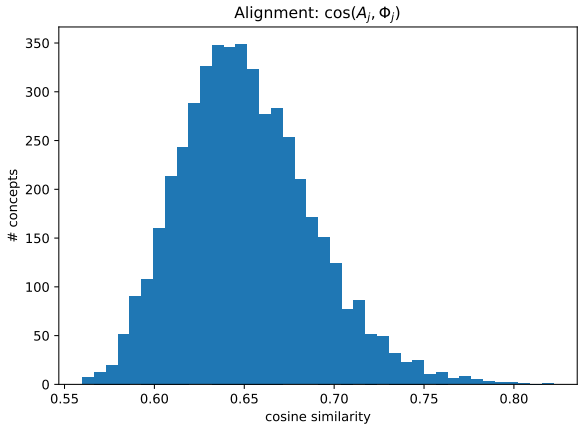


Figure 9. **Alignment distribution** between learned concept directions A_j and original CLIP directions Φ_j. High alignment values indicate semantic consistency between the learned space and CLIP.

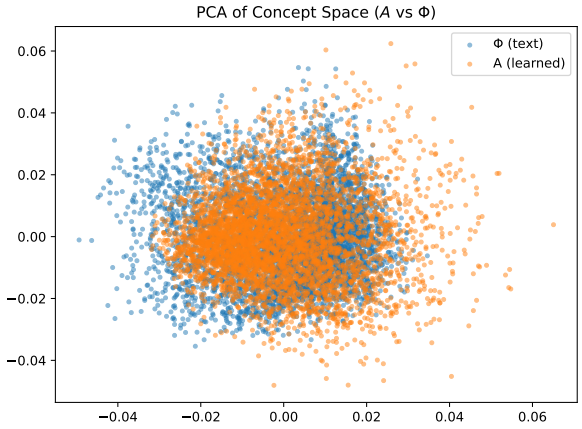


Figure 10. **PCA comparison** between CLIP concept space (Φ) and learned space (A). The learned space is more compact yet preserves the structural layout of concept groups.

The PCA results show:

- **Lower total variance**: A has 0.1047 vs. 0.1396 in Φ.
- **Comparable top-10 variance fraction**: 0.4364 vs. 0.4586.

This means that A is more compact, but not low-rank or collapsed. The PCA scatter in Figure 10 shows that concept clusters and their relative positions are preserved, indicating that the refined space maintains CLIP's semantic organization while smoothing its geometry.

C.5. Activation Density and Why Sparsity Is Not Required

We examine the density of concept activations using the joint interaction vector $s_{x,k} = c_x \odot c_k$. For each image, a concept j is counted as active if $s_{x,k}^{(j)} > \tau$, where $\tau = 0.01$. Figures 11 and 12 show the distribution of active concepts per image and the number of images activating each concept.

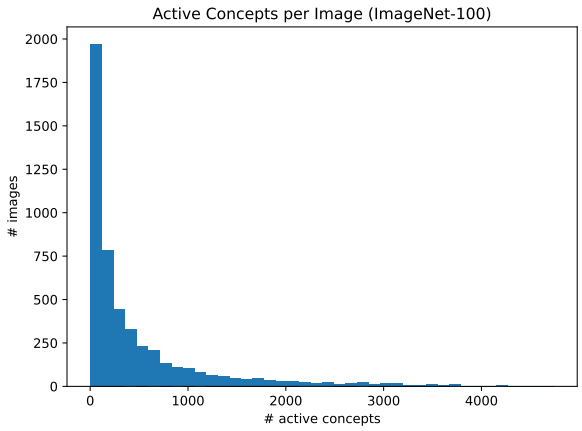


Figure 11. **Activation density**: number of active concepts per image.

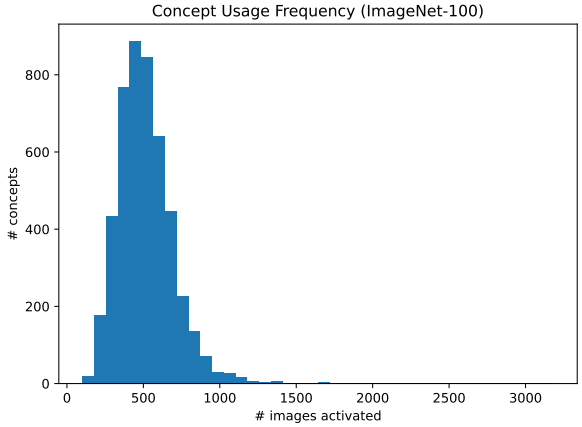


Figure 12. **Activation coverage**: number of images activating each concept.

Across ImageNet-100:

- Each image activates **490 concepts on average** (median 190).
- Each concept is activated for **516 images on average** (median 494).

These dense activation patterns are expected because CLIP embeddings are intrinsically distributed. Importantly, **sparsity is not required** for concept-based interpretability in our method. Explanations rely on identifying and ranking the most influential concept directions (top-n), not on enforcing a few active dimensions. Dense signals still yield clear, semantically aligned top concepts, as demonstrated in the qualitative results of the main paper.

D. CLIP-EZPC Fidelity

We evaluate how closely EZPC reproduces the predictions of the original CLIP model. This measures the faithfulness of the learned concept space to the teacher model. We report the following metrics: *Top-1 agreement*, *Spearman rank correlation*, *Kendall rank correlation*, and *KL divergence*. Results are shown in Table 10.

Table 10. **Prediction-level fidelity between CLIP and EZPC (RN50 backbone) with label consistency, ranking preservation, and KL.** Top-1 agreement measures label consistency. Spearman and Kendall correlations quantify ranking preservation. Kendall correlation is computed over the top-50 CLIP-ranked classes.

Dataset	Top-1 Agree. (%)	Spearman	Kendall	KL
CIFAR-100	80.26	0.959	0.733	6.79×10^{-6}
IN-100	92.92	0.994	0.904	6.10×10^{-6}
CUB	84.03	0.994	0.876	6.83×10^{-6}
ImageNet-1k	72.37	0.924	0.536	7.98×10^{-5}
Places365	79.92	0.972	0.715	1.46×10^{-5}

Fidelity remains high across datasets, indicating that EZPC preserves the ranking structure of CLIP predictions while enabling concept-based explanations. Lower agreement on larger datasets, such as ImageNet-1k, reflects the increased semantic ambiguity rather than the failure of the concept model.

E. Additional Ablation Studies

E.1. Impact of Concept Vocabulary Size

We analyze how the number of concepts affects both predictive performance and explanation fidelity. We randomly subsample the concept vocabulary and train EZPC with $m \in \{250, 500, 1000, 2000, 3000, 4751\}$ concepts. For each size, we repeat training with three random seeds. Results are shown in Table 11.

Table 11. **Effect of concept vocabulary size m on performance, tested on IN-100 under the generalized ZSL setting using the RN50 backbone.**

m	Top-1 Agree. (%)	Seen Acc.	Unseen Acc.	H
250	69.05 ± 0.34	0.5527 ± 0.0027	0.5637 ± 0.0105	0.5581 ± 0.0054
500	80.95 ± 1.35	0.6252 ± 0.0064	0.6423 ± 0.0090	0.6336 ± 0.0072
1000	89.09 ± 0.33	0.6622 ± 0.0028	0.6827 ± 0.0032	0.6723 ± 0.0030
2000	91.73 ± 0.17	0.6702 ± 0.0017	0.6890 ± 0.0026	0.6795 ± 0.0004
3000	92.58 ± 0.26	0.6742 ± 0.0010	0.6900 ± 0.0026	0.6820 ± 0.0010
4751 (full)	92.92	0.6745	0.6900	0.6821

Performance improves monotonically with the number of concepts, with diminishing returns beyond approximately 3000 concepts. This shows that EZPC is robust to moderate reductions in vocabulary size, while larger vocabularies mainly improve prediction fidelity.

E.2. Impact of Training Objectives

We analyze the contribution of each component of the training objective used to learn the projection matrix A. Our method optimizes a combination of a matching loss $\mathcal{L}_{\text{match}}$ and a reconstruction loss $\mathcal{L}_{\text{recon}}$.

We compare the following settings:

- No training ($A = \Phi$)
- Matching loss only
- Reconstruction loss only
- Full objective ($\mathcal{L}_{\text{match}} + \lambda\mathcal{L}_{\text{recon}}$)

Results calculated on ImageNet-100 using the generalized zero-shot setting are shown in Table 12.

Table 12. **Ablation study on the training objective of projection matrix A using RN50 on IN-100 (generalized zero-shot setting).**

Training Setting	Seen Acc.	Unseen Acc.	H
$A = \Phi$ (0-step, no training)	0.013	0.0	0.0
Matching loss only ($\mathcal{L}_{\text{match}}$)	0.013	0.0	0.0
Reconstruction loss only ($\mathcal{L}_{\text{recon}}$)	0.680	0.708	0.693
Full objective ($\mathcal{L}_{\text{match}} + \lambda\mathcal{L}_{\text{recon}}$)	0.674	0.690	0.682

Without any training ($A = \Phi$) or with the matching loss alone, the model achieves near-zero accuracy. This is expected as the raw CLIP concept embeddings do not form a basis that can reconstruct the original similarity structure, and the matching loss by itself only regularizes A toward Φ without learning to preserve predictive information. The reconstruction loss alone achieves the highest harmonic mean ($H = 0.693$), indicating that it is the primary driver of classification performance. Adding the matching loss in the full objective slightly reduces the harmonic mean to 0.682, but as shown in Figure 2, the matching loss plays a critical role in maintaining concept interpretability. Without it, learned concept directions drift from their original semantic meaning, producing less interpretable explanations despite higher quantitative scores. The full objective, therefore, represents a trade-off between predictive fidelity and explanation quality.

F. Additional Qualitative Visualizations

This section presents additional qualitative results that complement the analysis in the main paper, covering image-level, class-level, concept-clustering, and region-level alignment visualizations:

- **Image-level explanations (Figures 13 to 16)** Top activated concepts for individual predictions.
- **Class-level explanations (Figures 17 to 20)** Average concept activations across a random subset of images per class.
- **Concept clustering (Figures 21a to 21d)** Clusters of images that strongly activate the given concept.
- **Region-level concept alignment (Figures 22a to 22c)** Spatial localization of positive and negative concepts within individual images.

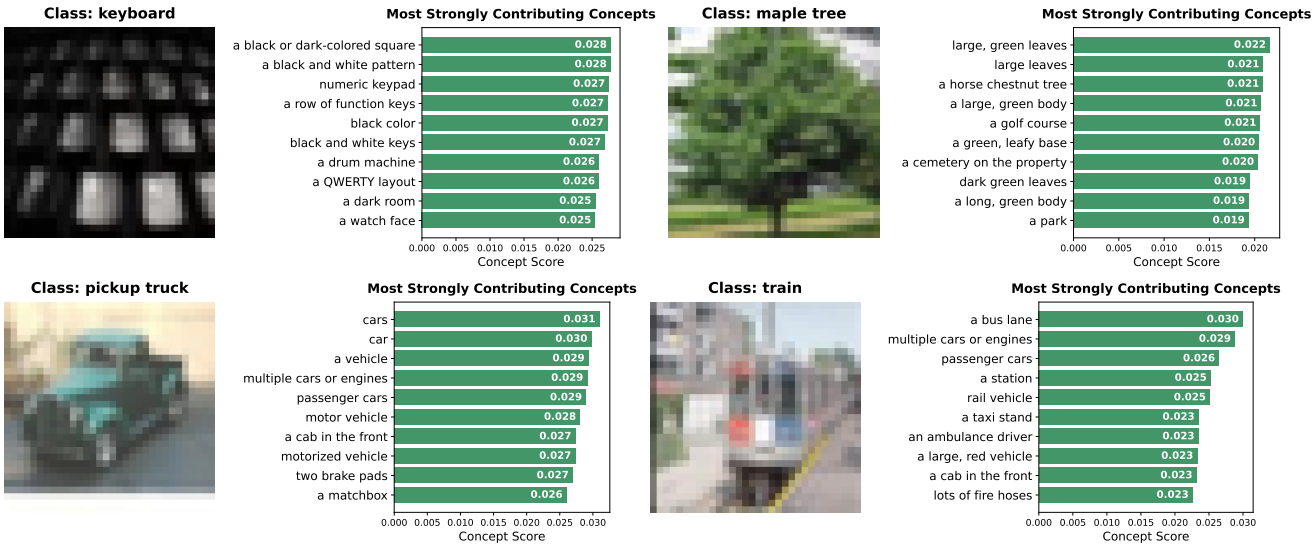


Figure 13. CIFAR-100 image-level explanations.

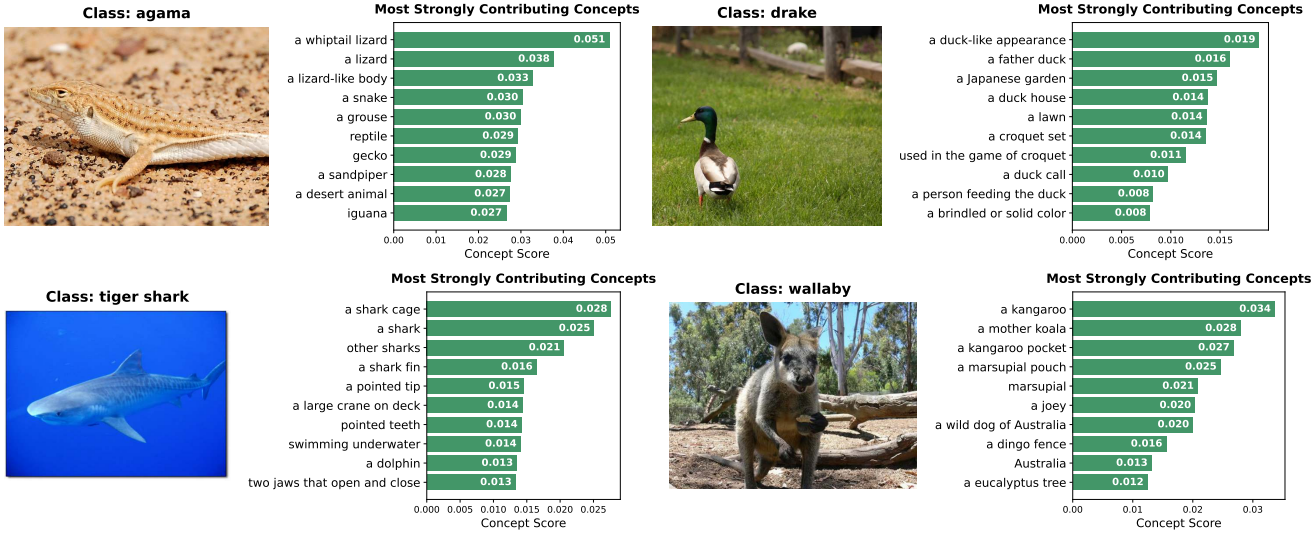


Figure 14. ImageNet-100 image-level explanations.

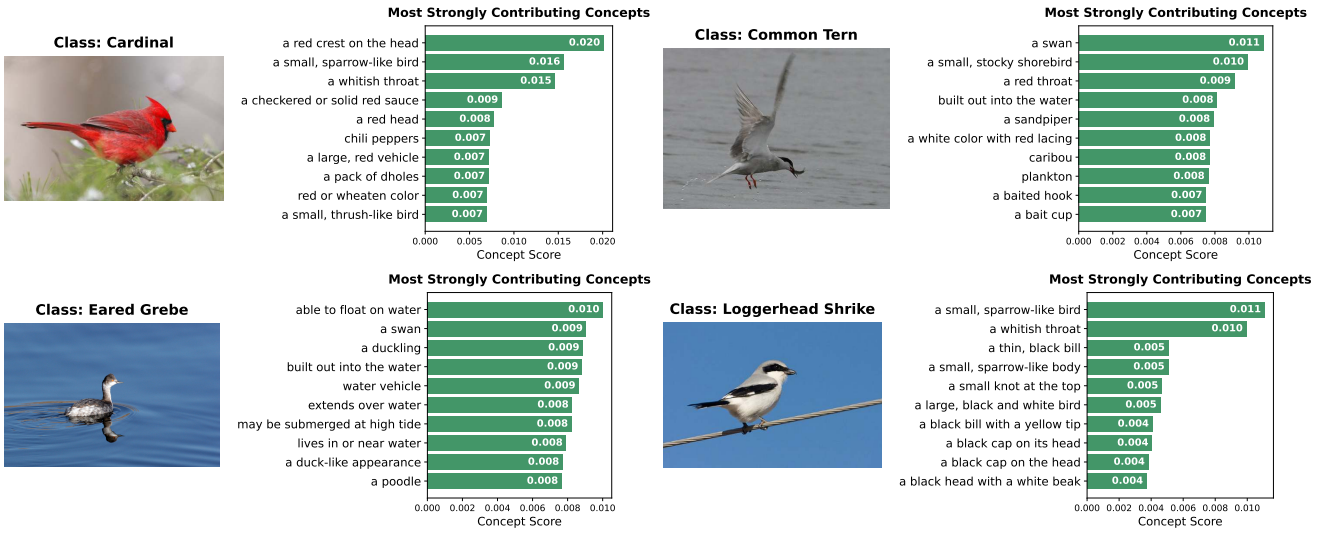


Figure 15. CUB image-level explanations.

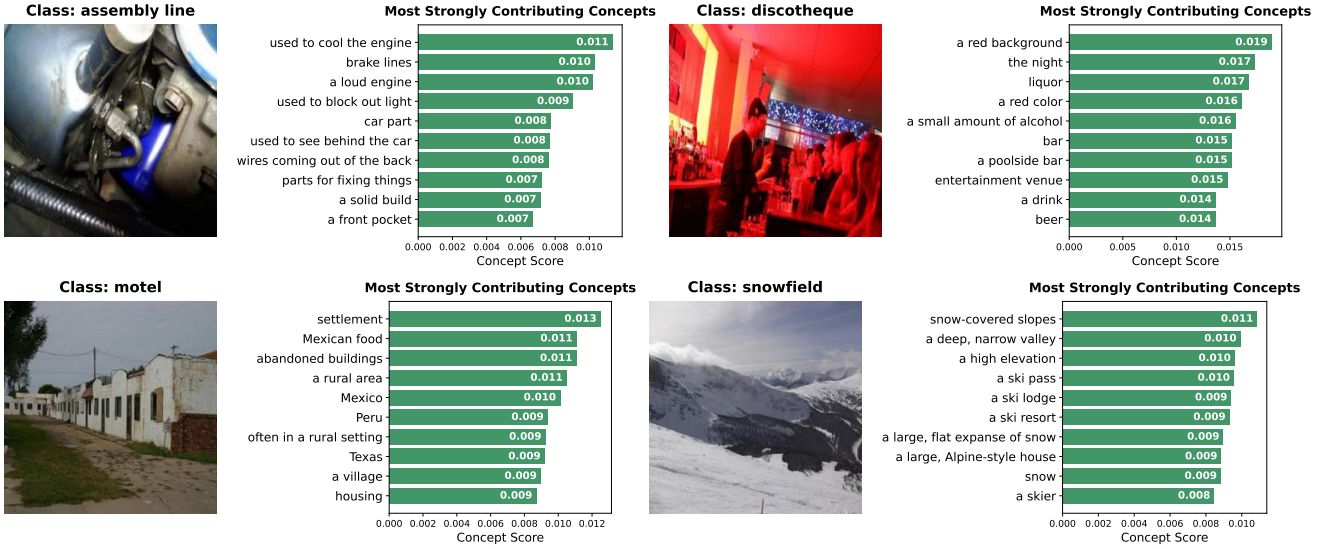


Figure 16. Places365 image-level explanations.

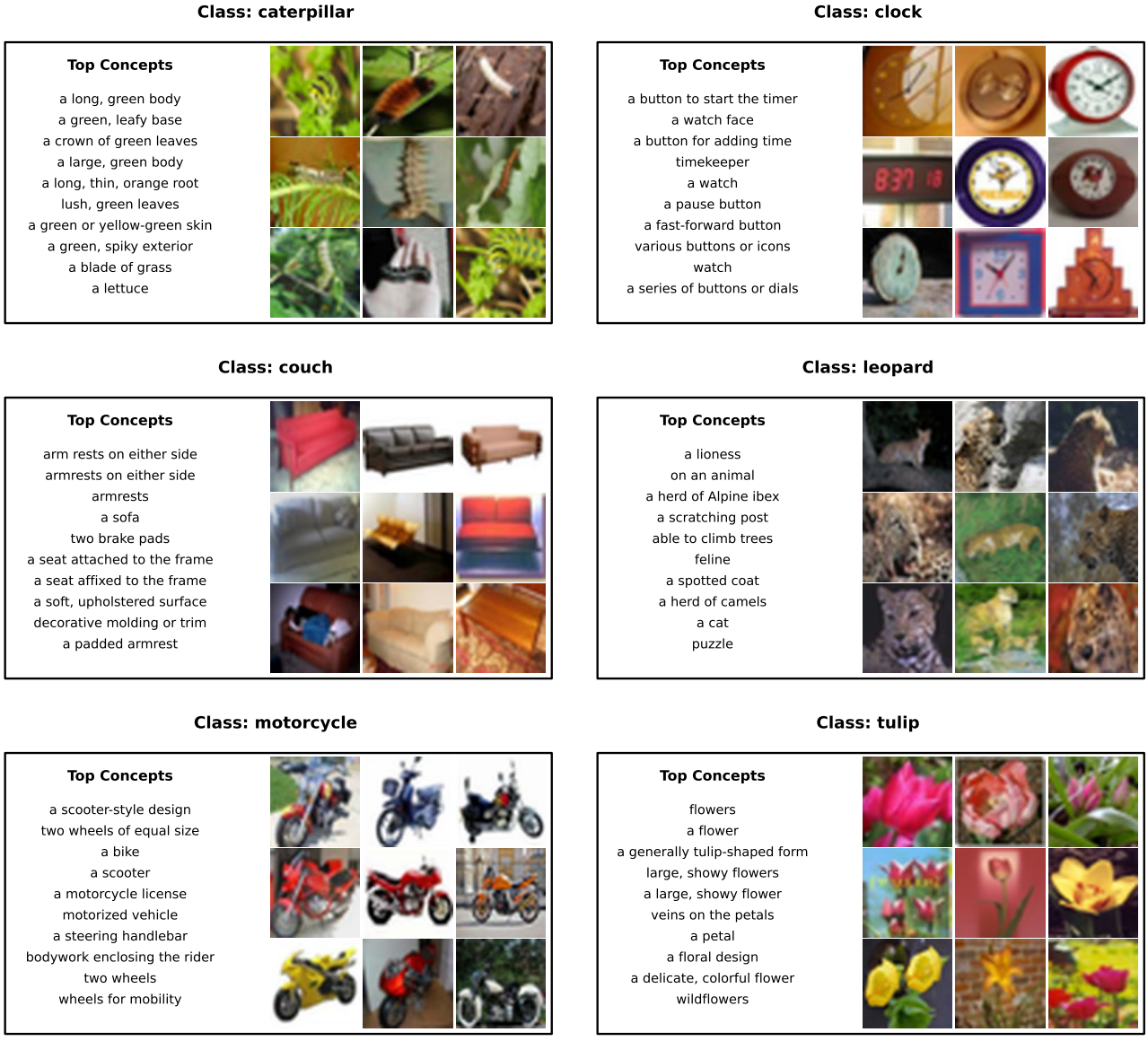


Figure 17. CIFAR-100 class-level visualizations.

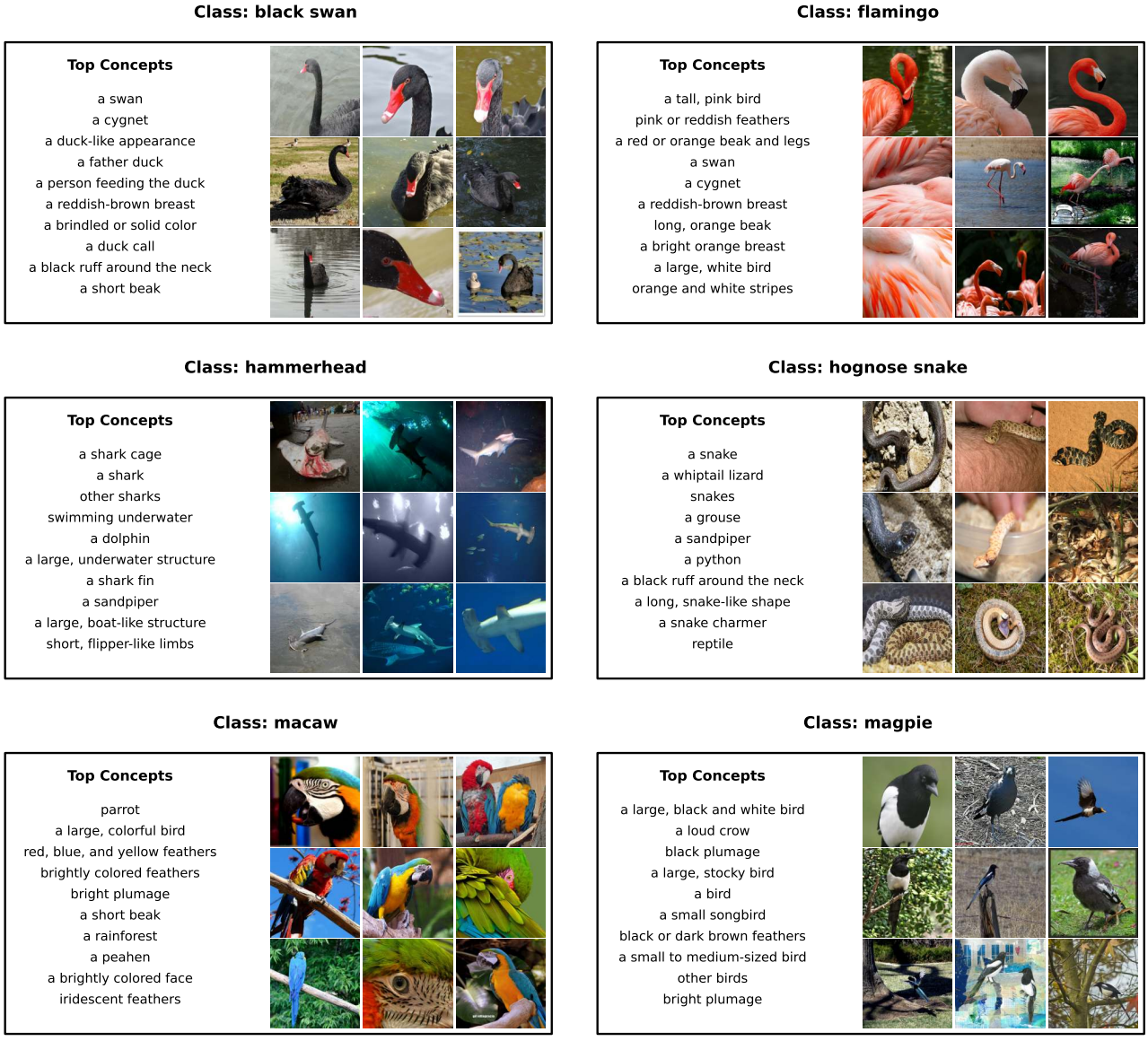


Figure 18. ImageNet-100 class-level visualizations.

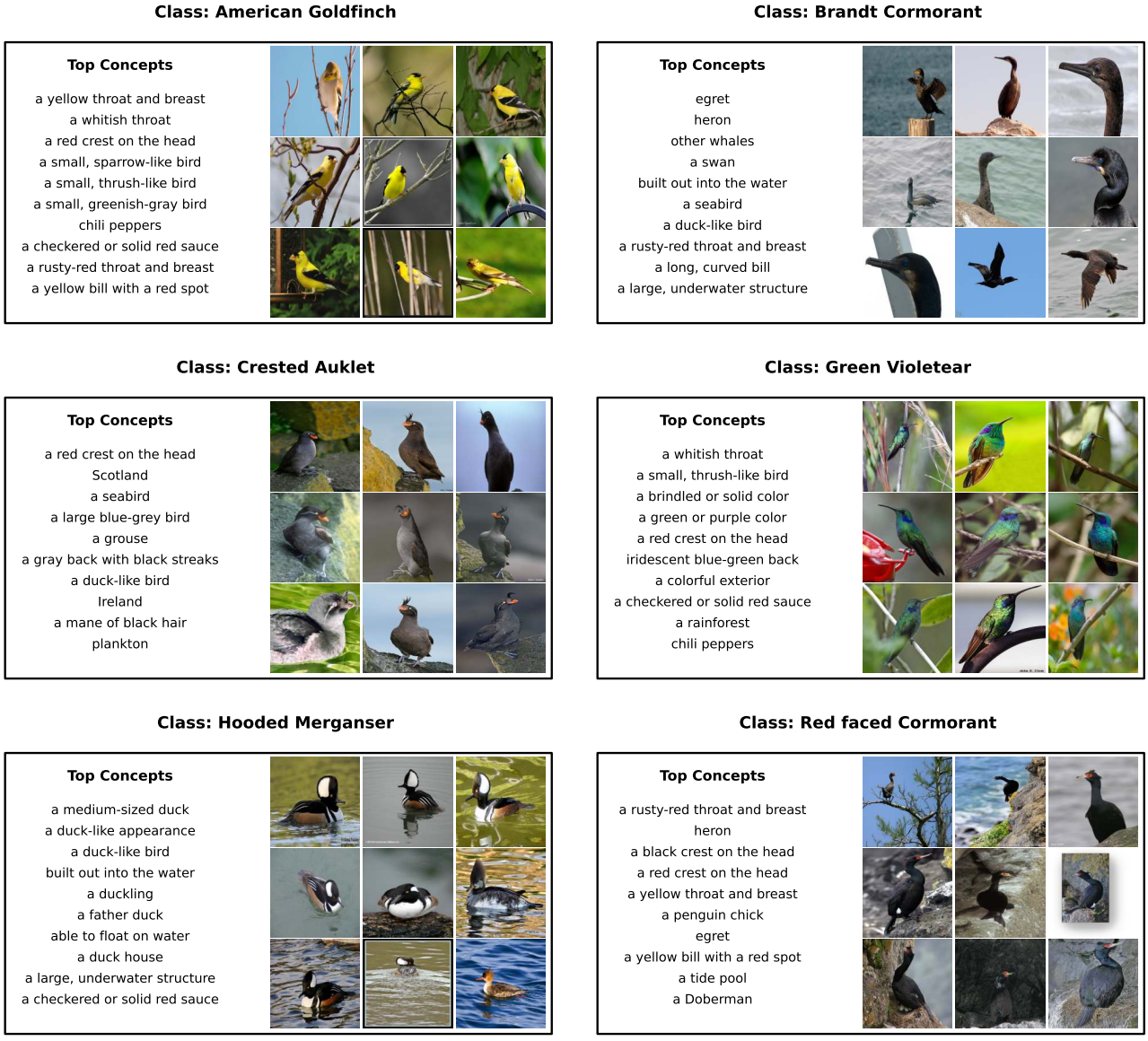


Figure 19. CUB class-level visualizations.

Class: amphitheater

Top Concepts

ancient architecture
ancient ruins
a pilgrimage site
a major pilgrimage site
a theater
Italian dish
lots of dirt and rubble
a scenic location
usually made of stone or brick
theater

Class: amusement arcade

Top Concepts

game
game center
a variety of games
a game
gambling
a place to play games
a video game console
a sports game
a slot on the top for coins
A toy

Class: bow window, indoor

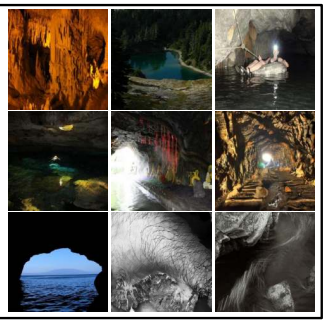
Top Concepts

a leather or fabric pouch
plants growing in the water
a room in a house or apartment
memorial plaque
large room with high ceilings
property
a flat, open area
indoor
a light, airy feel
paintings

Class: grotto

Top Concepts

a cave
a dark, hidden cave or recess
a canyon
a gorge
a deep, narrow valley
Ireland
a lightbulb at the top
a light at the top
a dark interior
under a rock



Class: skyscraper

Top Concepts

tall buildings
large buildings
a view of the cityscape
often surrounded by buildings
a city
a group of buildings
a skyline
china
high towers
place

Class: supermarket

Top Concepts

goods
stores
retailer
shopping
product displays
a brightly colored label
a checkout area
a place to store food
aisles between the shelves
a wide variety of merchandise

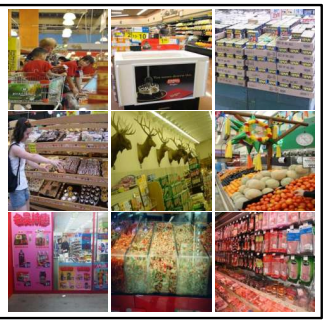
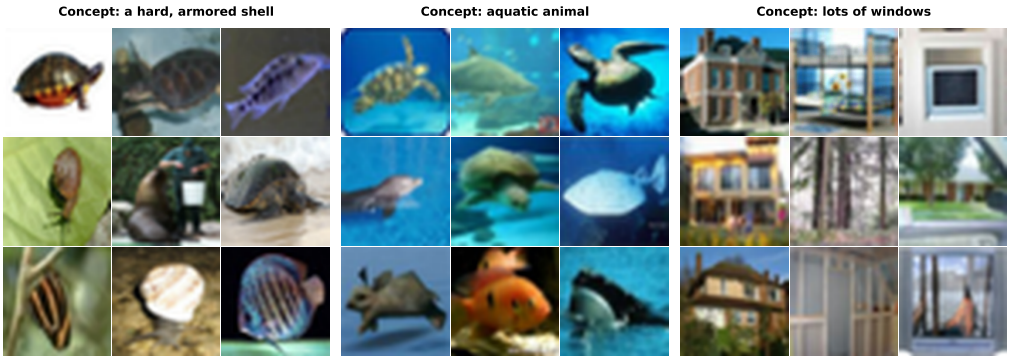


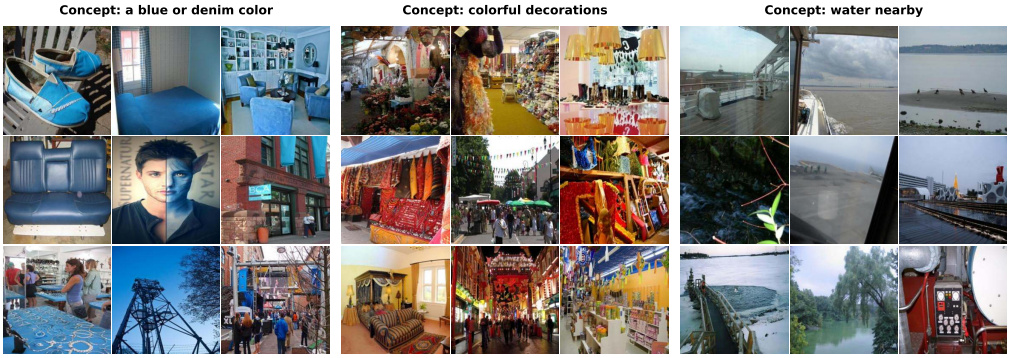
Figure 20. Places365 class-level visualizations.



(a) CIFAR-100 concept clustering examples.

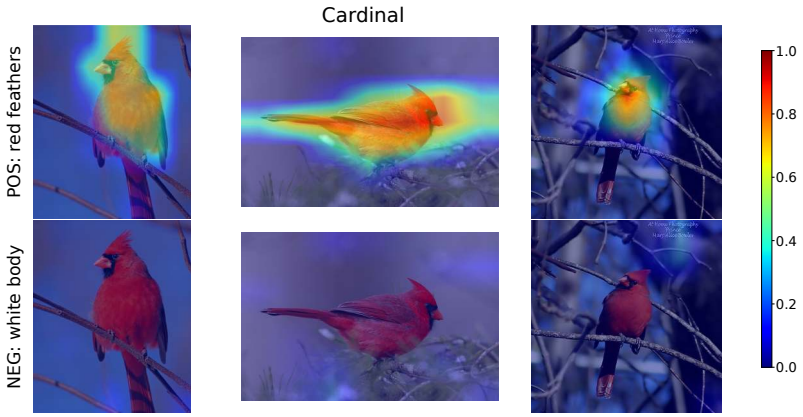
(b) ImageNet-100 concept clustering examples.

(c) CUB concept clustering examples.

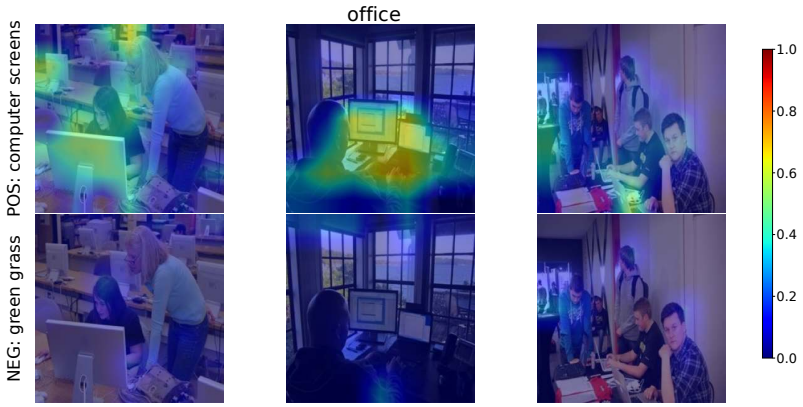


(d) Places365 concept clustering examples.

Figure 21. Concept clustering examples across datasets.



(a) **Region-level concept alignment for Cardinal class from CUB dataset.** Top row: positive concept (*red feathers*). Bottom row: negative concept (*white body*).



(b) **Region-level concept alignment for Office class from Places365 dataset.** Top row: positive concept (*computer screens*). Bottom row: negative concept (*green grass*).

(c) **Region-level concept alignment for Albatross class from ImageNet-100 dataset.** Top row: positive concept (*large wings*). Bottom row: negative concept (*yellow feathers*).

Figure 22. Region-level concept alignment examples across datasets.