Beyond Semantics: Disentangling Information Scope in Sparse Autoencoders for CLIP

Supplementary Material

A. SAE Configuration Across Vision Encoders

A.1. Details for CLIP SAE Configurations

This section provides additional statistics for the CLIP SAE configurations used in the main paper. We follow the same BatchTopK SAE setup for CLIP-ViT-B/16, CLIP-ViT-L/14, and CLIP-ViT-L/14-336px, and report here the resulting dictionary sizes and CDS-based partition statistics.

The resulting SAE dictionaries contain 24,576 features for CLIP-ViT-B/16 and 32,768 features for both CLIP-ViT-L/14 and CLIP-ViT-L/14-336px. To separate low-CDS and high-CDS features, we use the same threshold γ as defined in the main paper. The resulting partition sizes are summarized in Table 4. Across all three CLIP backbones, the high-CDS set forms a relatively small subset of the overall SAE dictionary.

Model	Dictionary Size	γ	Low-CDS	High-CDS
CLIP-B/16	24,576	0.14	17,329	7,247
CLIP-L/14	32,768	0.20	27,887	4,881
CLIP-L/14-336px	32,768	0.13	22,323	10,445

Table 4. SAE dictionary sizes and CDS-based partition statistics for the CLIP vision encoders used in the main paper. The high-CDS set forms a relatively small subset of the full SAE dictionary across all three CLIP backbones.

A.2. SAE Configuration for Other ViTs

To assess whether our findings generalize beyond CLIP, we additionally train BatchTopK SAEs on three representative ViT families: DINOv2 [25], SigLIP2 [34], and DeiT3 [33]. These experiments are intended to test whether the main phenomena studied in this paper, including outlier instability under SCC and the CDS-based separation of feature groups, also arise in other vision transformer architectures.

For each architecture, we train SAEs at three model scales for the scaling analysis in linear-probe classification. For all other mechanistic analyses, including the SCC-based analysis of outlier instability, we use the Base-scale models to enable a controlled comparison across architectures.

All SAEs are trained on the ImageNet-1K training set. During SAE training, we randomly sample four visual patch tokens per image from a designated transformer layer of each ViT, and all subsequent analyses are performed on that layer. For all models, we use an expansion factor of $\epsilon = 32$ and set $K = 30$ in BatchTopK.

Architecture	Model	Layer Used	τ
DINOv2	Base	Penultimate (11th)	100
	Large	Penultimate (23rd)	100
	Giant	Last (40th)	150
DeiT3	Small	10th	500
	Base	9th	500
	Large	18th	1000
SigLIP2	Base	Last (12th)	150
	Large	Last (24th)	150
	Giant	Last (27th)	300

Table 5. Model-specific SAE configurations for the additional ViT architectures. We report the transformer layer used for SAE training and analysis, together with the empirically selected outlier threshold τ for each model.

We determine the outlier threshold τ for each model empirically from the distribution of patch-token norms. Table 5 summarizes the transformer layer used for SAE training and analysis, together with the corresponding outlier threshold for each model.

B. SAE Evaluation Across Vision Encoders

We have included a comprehensive evaluation of our trained SAEs in the Table 6, measuring Fraction of Variance Unexplained (FVU) as a scale-invariant measure of reconstruction error, along with L_0 sparsity (L_0), and Cosine Similarity (CS). These metrics were computed on the ImageNet-1K validation set, covering the entire token set and the outlier subset to directly address concerns regarding high-norm tokens.

Model	$\text{FVU}_A \downarrow$	$\text{FVU}_O \downarrow$	L_{0A}	L_{0O}	$\text{CS}_A \uparrow$	$\text{CS}_O \uparrow$
CLIP-B/16	.099	.021	30.201	13.064	.923	.998
CLIP-L/14	.176	.042	29.307	14.170	.873	.999
CLIP-L/14-336px	.115	.028	29.654	16.391	.913	.999
DINOv2-B	.039	.001	29.425	7.298	.911	.999
DINOv2-L	.115	.001	30.416	15.334	.917	.999
DINOv2-G	.062	.002	30.664	28.709	.912	.999
DeiT3-S	.123	.081	32.067	20.318	.941	.956
DeiT3-B	.215	.101	32.293	10.208	.900	.940
DeiT3-L	.170	.096	31.533	16.113	.930	.953
SigLIP2-B	.030	.001	29.982	13.183	.948	.999
SigLIP2-L	.042	.001	29.628	13.573	.925	.999
SigLIP2-G	.116	.001	30.144	10.202	.918	.999

Table 6. SAE reconstruction statistics for all tokens (A) and outlier tokens (O).

C. Task-Specific Linear Probe Setup

In this section, we provide detailed experimental setups for the downstream tasks used in Section 5.4 to evaluate the functional roles of the low-CDS and high-CDS feature sets.

Probing Method. For ImageNet classification, we first applied global average pooling over all visual tokens after feature-group removal to obtain a single image-level embedding. A linear classifier was then trained on these pooled embeddings to predict class labels. For ADE20K semantic segmentation, we trained a linear probe on the feature-group-removed embeddings, following the protocol outlined in [25].

For monocular depth estimation on the NYUd dataset, we first preprocessed the ground-truth (GT) depth maps. We applied a center crop to the GT maps to match the spatial resolution of the ViT input. From the final layer of the ViT, we extracted both the CLS token and the patch tokens. The corresponding feature-group removal was applied only to the patch tokens, while the CLS token remained unmodified. We then broadcast the CLS token and concatenated it to every patch token along the feature dimension. These augmented patch tokens were reshaped into their 2D spatial grid and bilinearly upsampled to the full input resolution. We formulated depth estimation as a classification task by discretizing the depth range into 256 uniformly distributed bins. A linear probe was trained on the upsampled features to predict the logits for these bins. The final scalar depth value for each pixel was computed as a linear combination of the bin centers, weighted by the predicted softmax probabilities [3]. The probe was trained end-to-end using a scale-invariant loss [12] on the estimated depth.

Training Settings. All linear probes were trained for 100 epochs using SGD with momentum 0.9 and a cosine annealing scheduler. The batch size was set to 1024 for ImageNet classification. For the dense prediction tasks, we used smaller batch sizes of 64 for ADE20K segmentation and 8 for NYUd depth estimation due to GPU memory constraints. We conducted a random hyperparameter search to determine the optimal learning rate and weight decay. Specifically, the learning rate was sampled from $[10^{-3}, 10^{-2}]$ for classification and segmentation, while a higher range of $[10^{-1}, 1]$ was used for depth estimation to accommodate pixel-wise optimization. The weight decay was sampled from $[10^{-5}, 10^{-4}]$ across all tasks. For each embedding variant, we performed 5 independent trials and reported the result achieving the best performance on the validation set in Table 3.

D. Experiments on other ViTs

D.1. CDS Distributions and Instability Across ViTs

We first examine whether the CDS-based feature partition observed in CLIP also emerges in other ViT families. Figure 6 shows the CDS distributions of the Base-scale DINOv2, SigLIP2, and DeiT3 models. Across all three architectures, most SAE features cluster into a dominant low-CDS region, while additional mass appears in a separated high-CDS region. This qualitatively mirrors the trend observed in CLIP and supports the use of a CDS-based partition into low-CDS and high-CDS feature groups beyond CLIP.

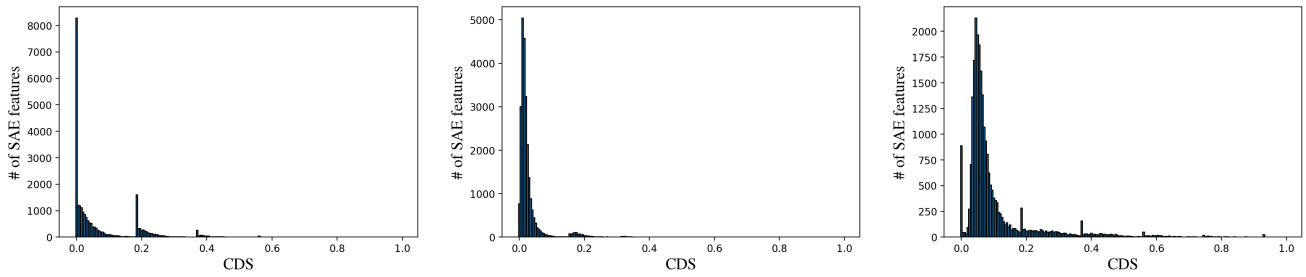
We next ask whether outlier tokens in these architectures likewise exhibit strong contextual instability. To empirically quantify the spatial instability of outlier tokens, we apply the analysis detailed in Section 4.2. We conduct our analysis using 1,000 images, randomly sampled one-per-class from the ImageNet-1K validation set. As reported in Table 7, outlier tokens exhibit substantially higher EMD than non-outliers across all evaluated ViTs. This drastic discrepancy confirms that outlier emergence is not an intrinsic property of the underlying visual patch, but rather a highly context-dependent phenomenon. Consequently, even a subtle spatial shift in the surrounding context can trigger a complete reassignment of outlier locations, severely disrupting the model's attention patterns.

To connect this feature-level CDS separation with token-level outlier behavior, we next investigate representational instability by quantifying contextual dependency at the token level.

We define the Activation-Weighted CDS (awCDS) of a token v_t as the activation-weighted average of the CDS values of the SAE features activated by that token:

$$\mathrm{awCDS}(v_t) = \frac{\sum_{j=1}^{q} a_{t,j}\,\mathrm{CDS}_j}{\sum_{j=1}^{q} a_{t,j}},$$

where $a_{t,j}$ denotes the activation of the j-th SAE feature for token v_t. To systematically analyze this metric across the token spectrum, we sort the patch tokens extracted from the same set of 1,000 images by their norm, divide them into percentile-based bins, and compute the average awCDS for each bin. As illustrated in Figure 7, awCDS spikes exponentially specifically for outlier tokens across all architectures. This structural correlation serves two purposes. First, it shows that the spatial instability observed in Table 7 is a direct manifestation of outlier tokens being densely composed of highly context-sensitive representations. Second, together with the CDS distributions in Figure 6, it further supports the interpretation that CDS isolates feature groups with systematically different degrees of contextual dependency across architectures.



(a) DINOv2 (b) DeiT3 (c) SigLIP2

Figure 6. **Distribution of Contextual Dependency Score (CDS).** Histogram of CDS values for SAE features across DINOv2, SigLIP2, and DeiT3.

Table 7. **Contextual instability of attention maps for outlier vs. non-outlier tokens.** Average EMD scores are computed on the designated layer's attention maps using our SCC method ($s = 1$). Across all models, outlier tokens exhibit significantly higher EMD scores, demonstrating severe sensitivity to contextual shifts.

Model	non-outlier ($\bar{D}_{\text{non}}$)	outlier ($\bar{D}_{\text{out}}$)
DINOv2	0.61	6.96
DeiT3	0.38	5.97
SigLIP2	3.75	7.18

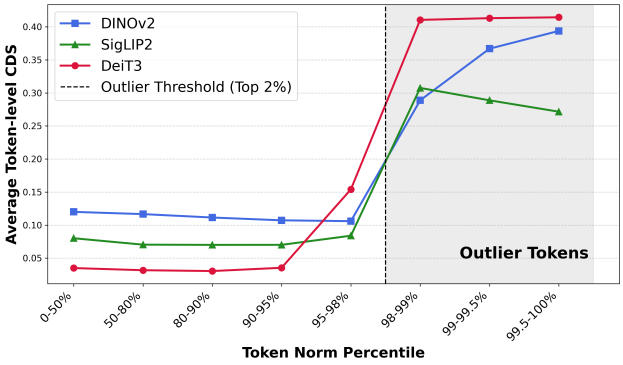


Figure 7. **Contextual dependency across the norm spectrum.** We plot the Activation-Weighted CDS (awCDS) across the token norm spectrum. Across all architectures, awCDS spikes exponentially for outlier tokens.

D.2. Feature-Group Removal Analysis Across ViTs

Having established in D.1 that other ViT families also exhibit a meaningful CDS-based feature separation, and that outlier tokens remain spatially and representationally unstable across architectures, we next ask whether the downstream functional roles of low-CDS and high-CDS features also generalize beyond CLIP. To this end, we repeat the feature-group removal analysis of Section 5.4 on DINOv2, SigLIP2, and DeiT3, using the same CDS-based partitioning procedure described in Section 5.3. For ImageNet-1K classification, we report results across the three model scales described in Appendix A.2. For ADE20K semantic segmentation and NYUd monocular depth estimation, we use the Base-scale models to keep the dense-prediction setting controlled across architectures.

Analysis of Linear Probing on Image Classification. Figure 8 summarizes the scaling behavior across the three ViT families. Across all three architectures, removing the high-CDS set consistently improves linear-probe accuracy over the original embedding at every scale, while removing the low-CDS set causes a clear performance drop that becomes progressively smaller as model size increases. These results extend the trend observed in CLIP: low-CDS features provide the signals that are most directly exploitable by a linear classifier, while the contribution of high-CDS features becomes more competitive as model scale increases. The weaker gap between the original and low-CDS-removed embeddings in larger models suggests that larger models encode broader contextual information in a form that is increasingly useful for classification.

Analysis of Linear Probing on Dense Prediction Tasks. Table 8 reports the dense-prediction results on ADE20K and NYUd. Across all three architectures, removing the low-CDS set causes a substantial degradation on both tasks, confirming that low-CDS features carry the localized information required for dense prediction beyond CLIP. In contrast, removing the high-CDS set leaves performance close to the original embedding, with only minor architecture-dependent changes. On ADE20K, high-CDS removal improves performance for DINOv2 while causing only slight drops for DeiT3 and SigLIP2. On NYUd, its effect is likewise modest but mixed, slightly degrading performance for DINOv2 while improving it for DeiT3 and SigLIP2. This consistent asymmetry indicates that high-CDS features provide complementary broader context, whereas low-CDS features remain the primary source of the fine-grained spatial information needed for dense prediction. Taken together with the instability results in Table 7 and Figure 7, these findings show that the CDS-based functional distinction is not specific to CLIP, but recurs across self-supervised, vision-language, and supervised ViTs.

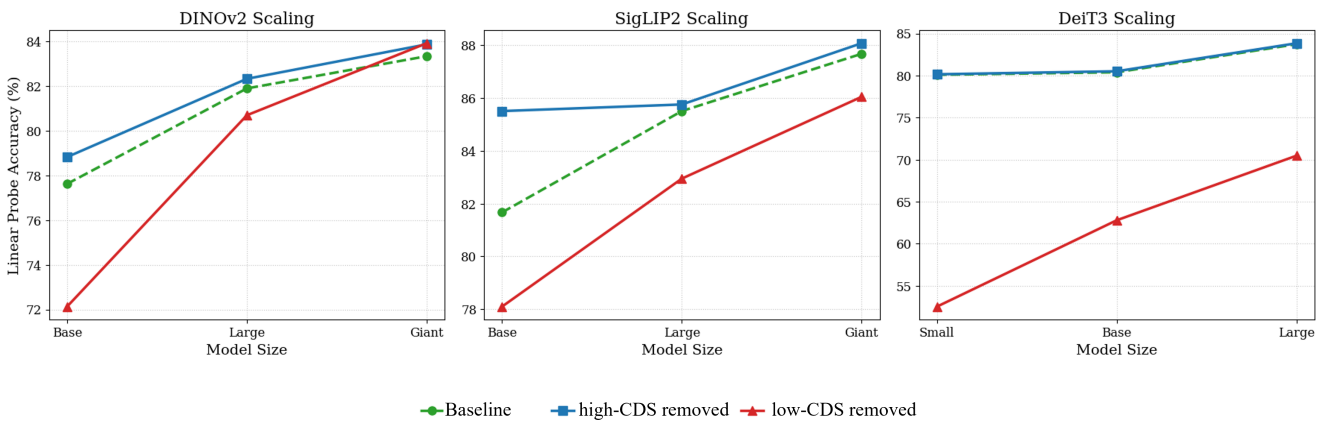


Figure 8. **Linear probe accuracy across ViT scales on ImageNet-1K.** The high-CDS-removed embedding (blue) consistently outperforms the Baseline (green) across all architectures, indicating that local signals are more discriminative for classification. Conversely, the performance gap of the low-CDS-removed embedding (red) generally narrows as model size increases, suggesting that larger models encode broader context more effectively.

Table 8. **Linear probing performance of ablated feature sets on other ViT architectures.** We report linear probing results (ADE20K mIoU, NYUd RMSE) for the Base-scale DINOv2, DeiT3, and SigLIP2 models. We compare the original embedding against two ablations: removing low-CDS features and removing high-CDS features. **BOLD** denotes the best result, while UNDERLINE denotes the worst result.

Model	Embedding Type	ADE20K mIoU↑	NYUd rmse↓
DINOv2	Original	**25.26**	**0.7500**
	high-CDS-removed	26.11	0.7539
	low-CDS-removed	15.47	0.8402
DeiT3	Original	**26.48**	0.8444
	high-CDS-removed	26.26	**0.8391**
	low-CDS-removed	8.38	1.0607
SigLIP2	Original	**35.62**	0.8178
	high-CDS-removed	35.39	**0.8123**
	low-CDS-removed	16.93	0.9205

Table 9. **Sensitivity of non-outlier and outlier tokens under varying shifting factors in SCC.** Outlier tokens consistently exhibit higher instability than non-outlier tokens across all values of s, the shifting factor in Shifted Context Crop (SCC).

Shifting factor s	1	2	3	4	5	6
Non-outlier	1.490	1.745	1.842	1.855	1.915	1.958
Outlier	5.179	6.752	6.869	6.876	5.811	5.533

E. Outlier Tokens' Sensitivity to SCC

In Section 5.2, we used the Shifted Context Crop (SCC) method to empirically quantify the sensitivity of outlier tokens to contextual shifts. While the main experiments used a fixed shifting factor of $s = 1$ to maximize spatial overlap between paired images, it is important to verify whether the observed instability is robust to the magnitude of the spatial shift. To this end, we conducted an ablation study varying s from 1 to 6 using CLIP-L/14-336px.

E.1. Experimental Setup

We followed the exact procedure described in Algorithm 1 and Section 4.2. The shifting factor s determines the displacement magnitude between the two generated views I_1 and I_2. A larger s implies a greater difference in absolute positional embeddings for corresponding patches and a more substantial alteration in global context due to the reduced overlap. We measured the average Earth Mover's Distance (EMD) for both non-outlier and outlier tokens across all heads in the target layer.

E.2. Results and Analysis

The results are summarized in Table 9. We observe two distinct trends. The EMD scores for non-outlier tokens exhibit a monotonic increase as s grows, from 1.490 at $s = 1$ to 1.958 at $s = 6$. This is expected: as the spatial shift increases, the discrepancy in positional information and the divergence of the surrounding context become more pronounced, leading to naturally larger variations in attention patterns. However, the instability of outlier tokens remains consistently high across all values of s. Notably, the EMD scores for outlier tokens are substantially larger than those of non-outlier tokens under every setting. The score peaks around $s = 3$ and $s = 4$, then decreases slightly at larger shifts. This mild decline at higher shifts may be due to the reduced size of the overlapping region, which limits the effective context available for the attention mechanism to

form high-confidence outliers.

This experiment confirms that contextual instability is an intrinsic property of outlier tokens rather than an artifact of a particular shifting factor. Even at the minimal shift of $s = 1$, the disparity between non-outlier and outlier tokens remains stark. We therefore adopted $s = 1$ in the main experiments to preserve the largest possible evaluation area while still exposing the instability of outlier tokens.

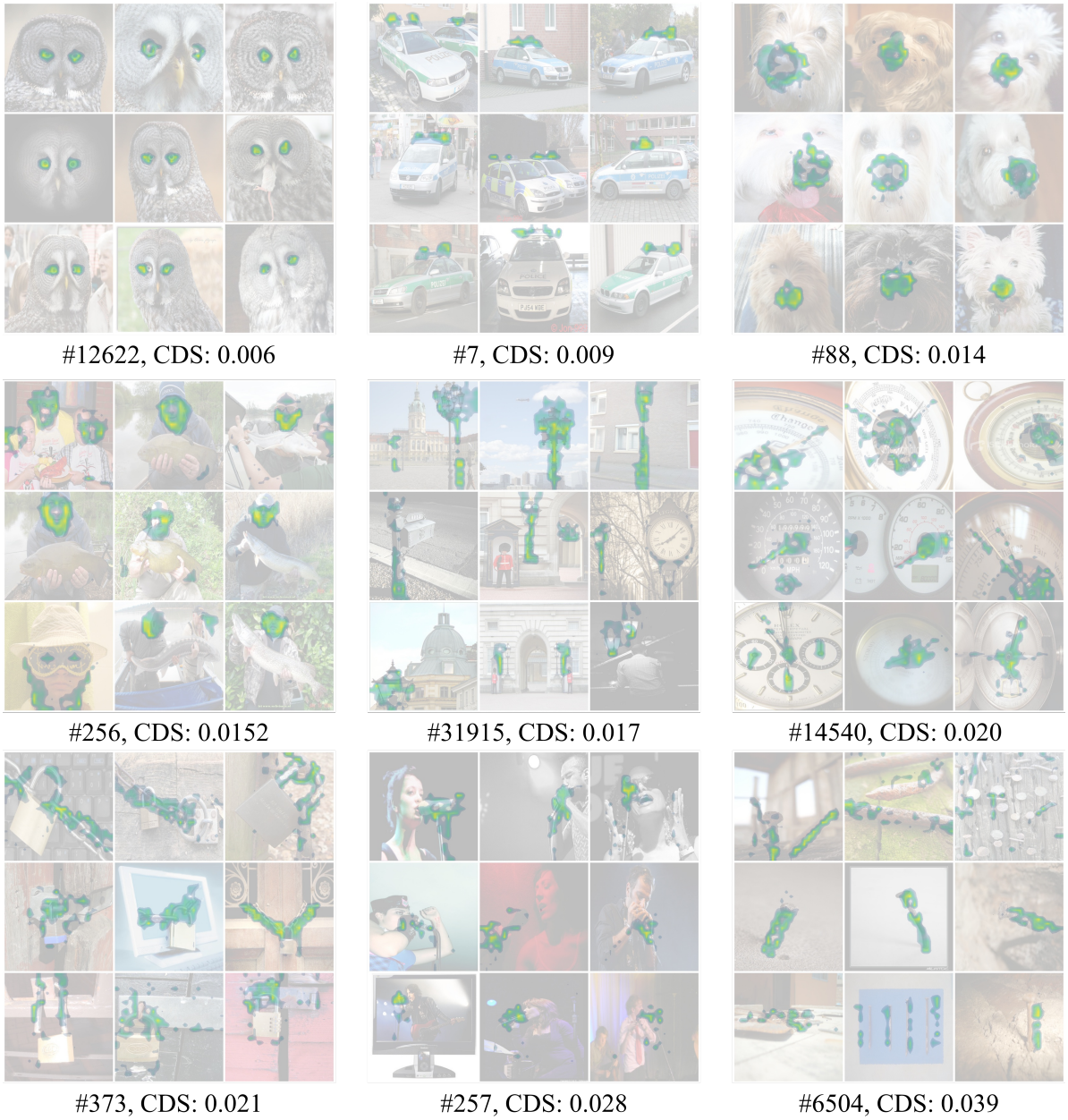


Figure 9. **Visualization of low-CDS features.** These features exhibit strong spatial grounding, consistently localizing specific visual concepts across different images.

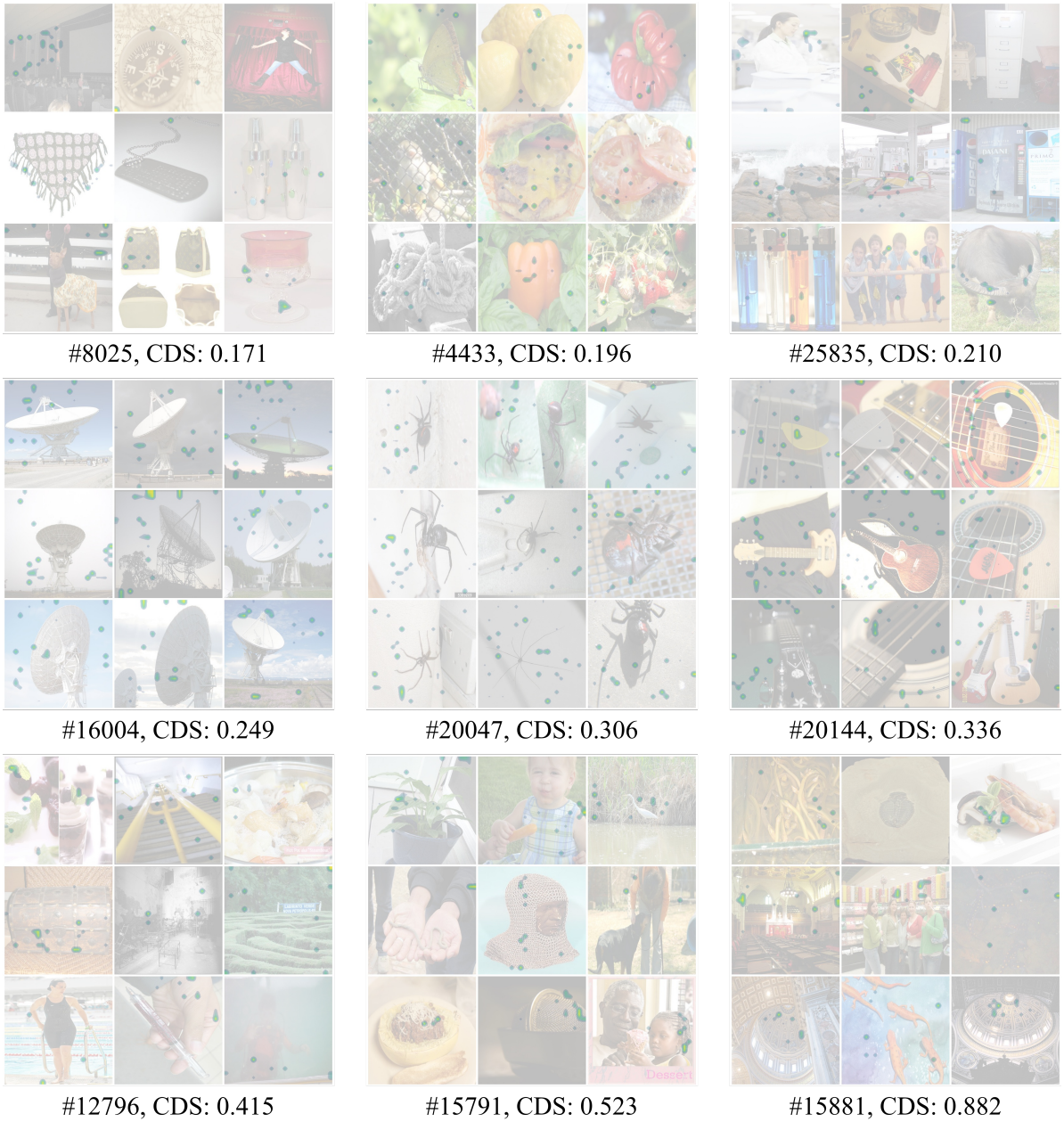


Figure 10. **Visualization of high-CDS features.** These features exhibit diffuse activation patterns and capture broader contextual information rather than localized visual details.