Concept Bottleneck Models in the Wild: Assessing Interpretability under Image Corruptions

Kexin Zhang [1] Katharina Prasse [1] Mishal Fatima [1] Margret Keuper [1,2]

[1]Chair of Machine Learning, Data and Web Science Group, University of Mannheim
[2]Max-Planck-Institute for Informatics, Saarland Informatics Campus

Workshop @ NeurIPS 2026

Introduction

Concept Bottleneck Models (CBMs) route every prediction through **named, human-auditable concepts** [6]. When deployed, they face **sensor noise, blur, weather and dead pixels**, yet CBMs have so far been evaluated only on clean data. A corrupted CBM loses accuracy *and* interpretability **without warning**, since it still outputs a fluent concept explanation.

Take-aways

1. **Corruption hurts all three CBMs:** top-1 accuracy drops by 4-66 pp across corruptions and severities (severity 5 worst); Gaussian noise is harshest, patch masking mildest.
2. **Corruption reorders concepts.** The CLIP embedding direction is largely preserved (cosine ≥ 0.59), yet the concept activation vector is reordered with the top-1 activated concept changes in 70–100% of images.
3. **Concepts differ in robustness, and the robust ones can be found.** A per-concept shift score d_k ranks concepts consistently across severities and moderately across corruption types. Keeping the most stable ones beats a size-matched random subset by +1.5 to +3.3 pp on corrupted images, at ≤ 0.25 pp clean cost with the head refit.

Setup: A Deployed CBM in the Wild

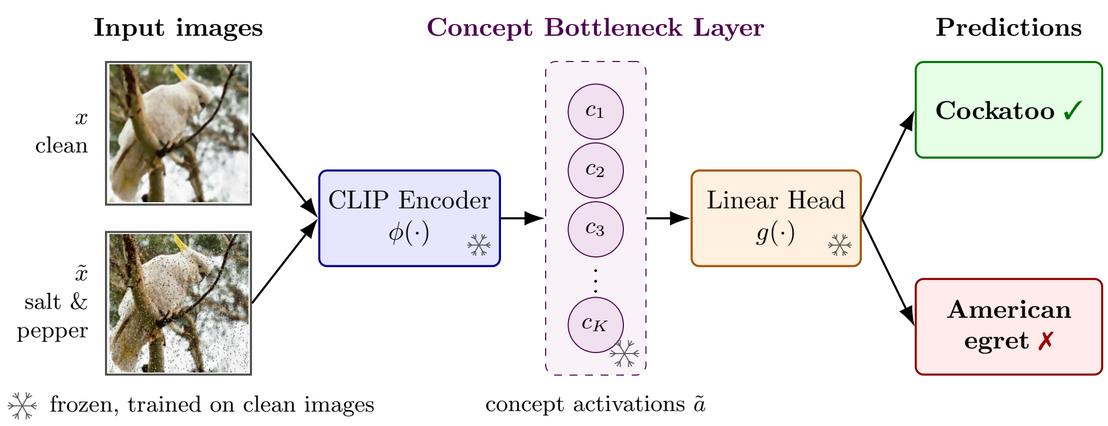


Figure 1. The same image, clean and with salt-and-pepper noise, passes through the frozen pipeline (encoder ϕ, concept layer, linear head g; trained on clean images only). The clean image is classified as cockatoo, the corrupted one as American egret, and both predictions come with a concept explanation.

- **CLIP encoder** ϕ (ViT-B/16) [9] embeds the image.
- **Concept bottleneck layer**: projects the image embedding onto K concept embeddings $c_1, \ldots, c_K$, giving scores $\bar{a}_k(x) \propto c_k^\top \phi(x)$. The head reads them standardized as concept activations $\tilde{a}_k$ (DCBM applies a softmax first).
  - **DCBM** [8]: 2048 image concepts, clustered from SAM2 [10] crops.
  - **LF-CBM** [7], **PCBM** [11]: 4589 text concepts.
- **Linear head** g: class scores $g(\tilde{\mathbf{a}}) = W\tilde{\mathbf{a}}$, so $W_{y,k}\tilde{a}_k$ is the contribution of concept c_k to class y.

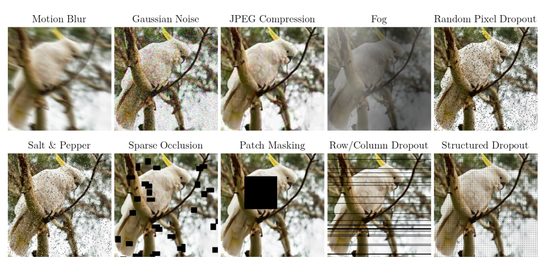


Figure 2. 10 corruptions (4 ImageNet-C [3], 6 sensor-level), severity 3 of 5.

Stage-Wise Analysis: Where Does the Damage Happen?

Table 1. ImageNet, severity 5, mean ± std over the 10 corruptions (cosine min 0.59).

Stage	Metric	DCBM	LF-CBM	PCBM
Reference	Clean top-1 ↑	0.6977	0.7548	0.7204
Embedding	Cosine sim. ↑	0.72 ±0.06	0.72 ±0.06	0.72 ±0.06
Concept activation	Argmax shift (top-1 concept changed) ↓	0.91 ±0.05	0.83 ±0.07	0.82 ±0.08
	Top-5 Jaccard ↑	0.08 ±0.04	0.16 ±0.06	0.16 ±0.07
	Spearman ρ ↑	0.41 ±0.12	0.48 ±0.12	0.48 ±0.12
Prediction	Top-1 accuracy ↑	0.27 ±0.12	0.35 ±0.13	0.32 ±0.13

- **Reordering.** The embedding stays close to clean (cosine 0.72) [1], but the concept ranking changes: top-5 Jaccard ≤ 0.16 and Spearman ≤ 0.48.
- **DCBM changes most** (Jaccard 0.08, Spearman 0.41): its softmax gives near-equal concept activations, so small shifts reorder them.

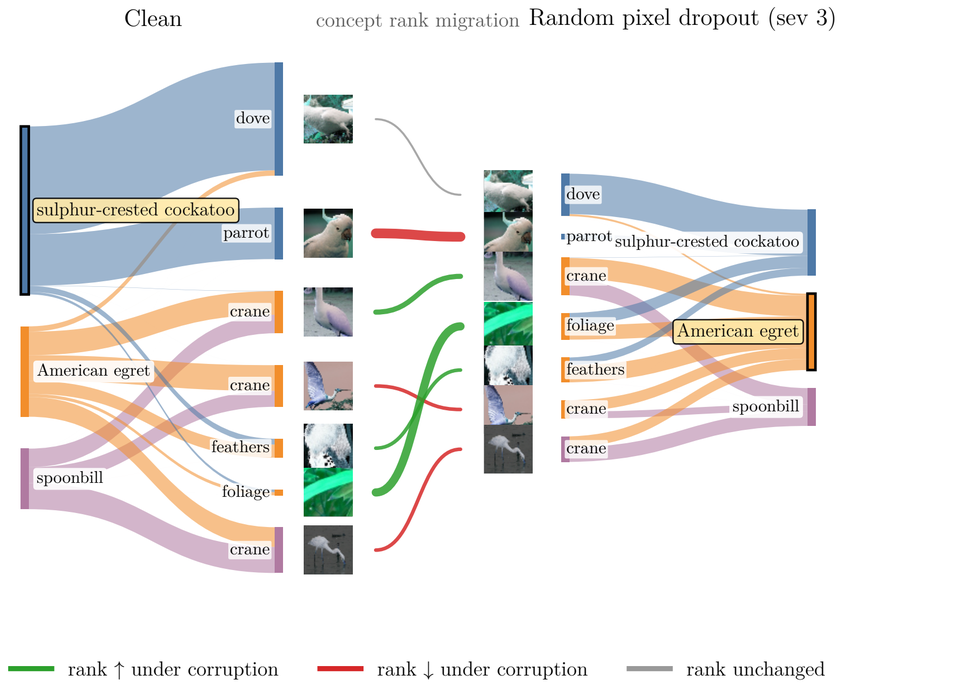


Figure 3. One image (DCBM): clean (left) vs. random pixel dropout, severity 3 (right). **Ribbon width**: contribution $W_{y,k}\tilde{a}_k$ of concept c_k to class y. **Bridge**: each concept's rank change, up, down or unchanged. The correct class receives less contribution as its supporting concepts fall in rank, and the prediction changes from *cockatoo* to *American egret*. Concept names are assigned automatically, so several concepts can share a name.

Concept Selection: Which Concepts Stay Stable?

- Concepts shift **unevenly**. The shift of the score $\bar{a}_k$ depends on corruption type f and severity s; we average it over both, since the concept ranking is consistent across severities and moderately shared across types (Fig. 4):

$$d_k = \frac{1}{|F|\,|S|} \sum_{f \in F} \sum_{s \in S} \mathbb{E}_x \left| \bar{a}_k(\tilde{x}^{f,s}) - \bar{a}_k(x) \right|$$

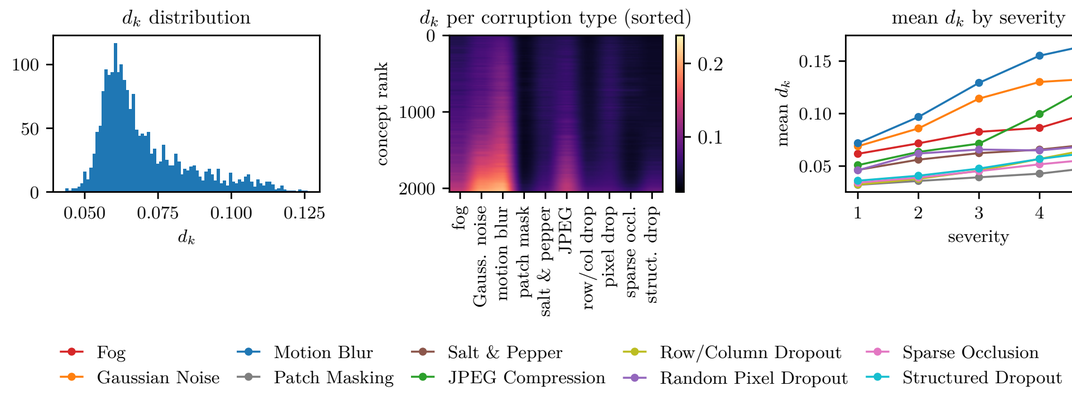


Figure 4. d_k for DCBM (K=2048). Left: unimodal with a fragile right tail. Middle: d_k per corruption type, rows sorted by overall d_k; the ordering is moderately shared across types (leave-one-out Spearman 0.60 on average, from -0.02 for sparse occlusion to 0.88 for motion blur). Right: mean d_k grows with severity.

- **Robust concept selection (proposed)**: keep the m lowest-d_k concepts [2] and refit only the head [4, 5]. Against a size-matched random subset, corrupted accuracy rises by +1.5 to +3.3 pp at comparable clean accuracy.

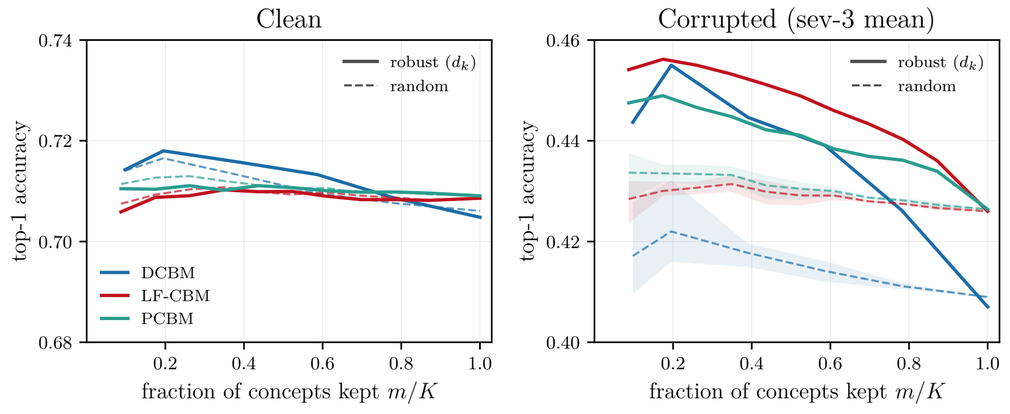
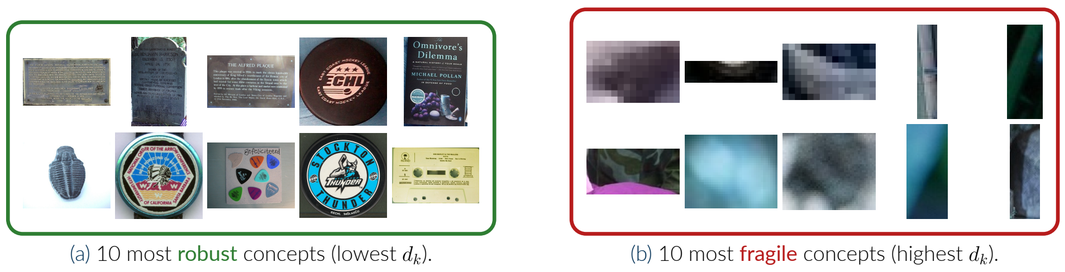


Figure 5. Top-1 accuracy vs. fraction of concepts kept: robust d_k (solid) vs. random (dashed; shaded: range of the 9 random fits).

Table 2. Accuracy at each model's best kept size m (corrupted = severity-3 mean). Selection changes the head's input, so every refit row, the full bank included, trains a new head with one shared recipe (L_2 logistic regression on 128k images) instead of each model's sparse deployed head. At the full bank this refit alone is 3.5–9.4 pp below the deployed head on corrupted images, so d_k is compared with random subsets and the full bank under the same refit, not with the deployed model. Robust-subset clean accuracy (refit): 71.80 / 70.88 / 71.04 costing at 0.25pp clean accuracy.

	DCBM	LF-CBM	PCBM
Best m / K	400 / 2048	800 / 4589	800 / 4589
Deployed reference, no refit (%)			
Clean / corrupted	69.77 / 44.20	75.48 / 51.95	72.04 / 48.68
Corrupted accuracy, refit head (%)			
Full bank	40.70	42.60	42.64
Random, best m	42.20	43.00	43.35
Robust d_k (val), best m	**45.50**	**45.62**	**44.89**
Robust d_k (disjoint), best m	45.30	45.70	44.74
Corrupted margin over random, pp			
Robust d_k (val) − random	**+3.3**	**+2.6**	**+1.5**
Robust d_k (disjoint) − random	+3.1	+2.7	+1.4

Case Study: What Makes a Concept Stable in DCBM?



(a) 10 most robust concepts (lowest d_k). (b) 10 most fragile concepts (highest d_k).

Figure 6. DCBM medoid crops ranked by d_k: (a) large, sharp whole-object crops; (b) small, thin, low-resolution slivers.

- **Robust concepts tend to be large crops**: d_k falls as the medoid crop's short side grows ($\rho = -0.65$).
- **Crop resolution explains part of the gain**: at m=400, +2.3 of the +3.3 pp margin over random; d_k adds +1.0 pp beyond it.

Outlook

- The robust subset is a **5× smaller set of concepts with known stability**, not a full repair.
- Next: more backbones and datasets, and measures of explanation quality beyond the rank statistics used here.

References

[1] Fang et al. Data determines distributional robustness in contrastive language image pre-training (CLIP). In *ICML*, 2022.
[2] Guyon and Elisseeff. An introduction to variable and feature selection. *JMLR*, 2003.
[3] Hendrycks and Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *ICLR*, 2019.
[4] Kang et al. Decoupling representation and classifier for long-tailed recognition. In *ICLR*, 2020.
[5] Kirichenko et al. Last layer re-training is sufficient for robustness to spurious correlations. In *ICLR*, 2023.
[6] Koh et al. Concept bottleneck models. In *ICML*, 2020.
[7] Oikarinen et al. Label-free concept bottleneck models. In *ICLR*, 2023.
[8] Prasse et al. DCBM: Data-efficient visual concept bottleneck models. In *ICML*, 2025.
[9] Radford et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.
[10] Ravi et al. SAM 2: Segment anything in images and videos. In *ICLR*, 2025.
[11] Yuksekgonul et al. Post-hoc concept bottleneck models. In *ICLR*, 2023.

Paper & Code

Paper Code