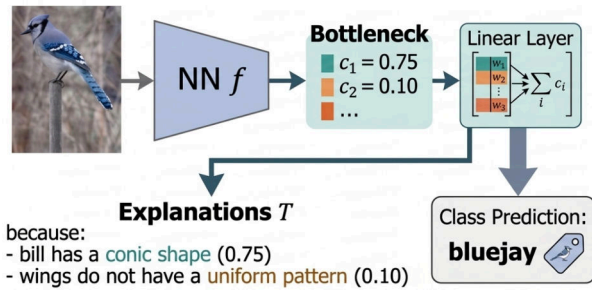


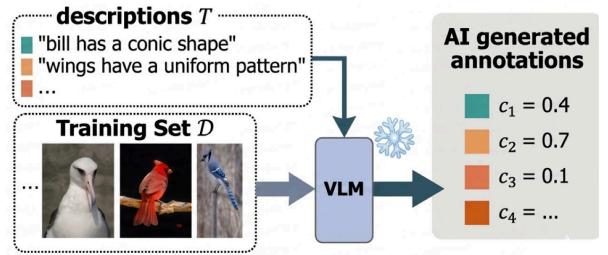
Rethinking Interpretable Concept-Based Models

VISUAL-LANGUAGE CONCEPT BOTTLENECK MODELS (VLM-CBM)

Use an interpretable CBM...



...but replace expensive human annotations with AI generated ones.



INTERPRETABILITY ISSUES OF VLM-CBMs



If Concept Bottlenecks are the Question, are Foundation Models the Answer?

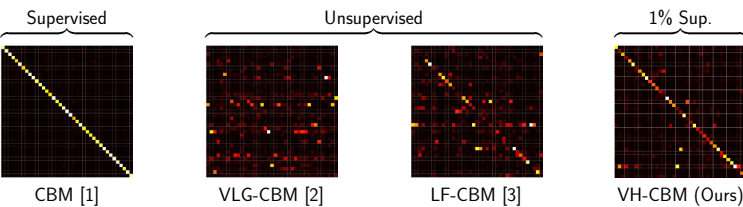
N. Debole, P. Barbiero, F. Giannini, A. Passerini, S. Teso, E. Marconato

Accepted at Machine Learning, Springer, 2025.

SCAN ME

IDENTIFIED PROBLEMS

- VLMs struggle with fine grained concepts. CLIP, GroundingDINO and Llava all achieve low concept accuracy.
- Concepts are **not disentangled**. The learned concepts often depends on unrelated ground-truth factors.



- Concepts suffer from **leakage**. Extra **non-concept-relevant** information is encoded, aiding task prediction but hindering interpretability.

FIXING VLM-CBMs

Improving VLM-guided Concept Bottleneck Models with Few Annotations

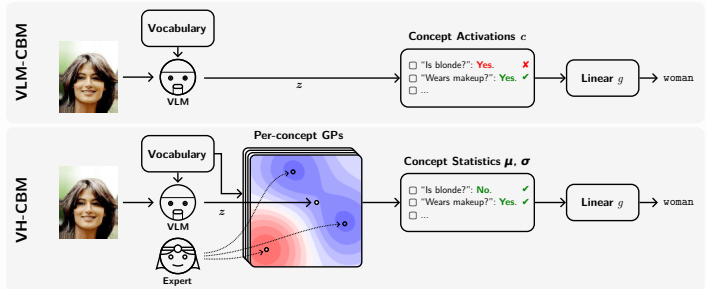
N. Debole, E. Marconato, A. Passerini, A. Pugnana, S. Teso

Under revision.

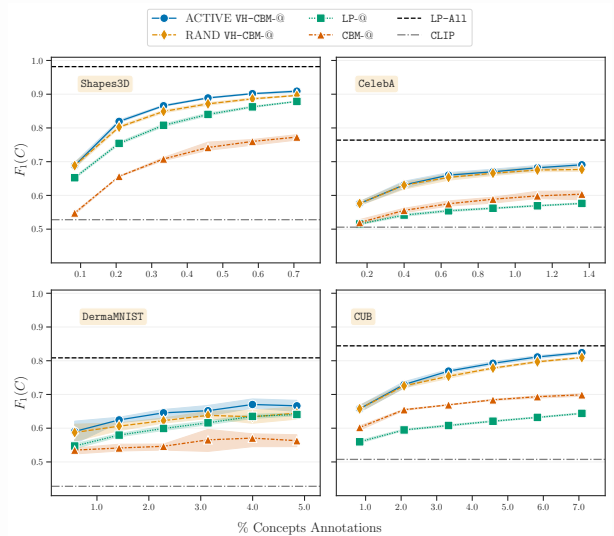
1 CONCEPT PREDICTION

- Each concept predictor is a Gaussian Process (GP).
- Training follows [4]: class labels are mapped to a Dirichlet-induced latent space, enabling efficient GP regression with calibrated uncertainty.
- The calibrated predictive uncertainty drives active learning, selecting the most informative samples for expert annotation.

2 VH-CBM



3 RESULTS



4 BENEFITS

- High concept quality is achieved with only a small number of expert annotations.
- The calibrated GP outputs may drive the task predictor to down-weight uncertain concept predictions, leading to improved task accuracy.
- VH-CBM is compatible with any VLM that produces latent embeddings.

REFERENCES

- [1] Koh *et al.*, Concept bottleneck models, ICML (2020).
- [2] Srivastava *et al.*, Vlg-cbm, NeurIPS (2024).
- [3] Oikarinen *et al.*, Label-free concept bottleneck models, ICLR Poster (2023).
- [4] Milios *et al.*, Dirichlet-based Gaussian Processes for Large-scale Calibrated Classification, NIPS (2018).