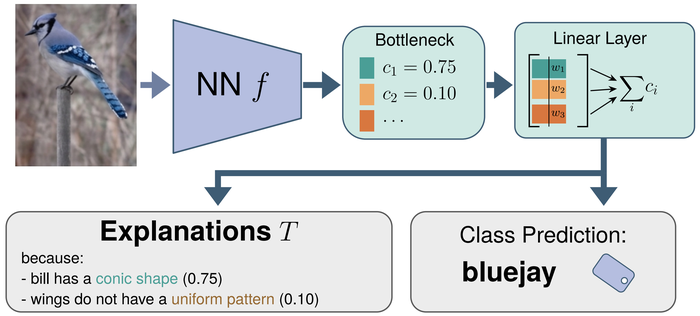
Concepts Worth Having: Refining VLM-Guided Concept Bottleneck Models with Minimal Annotations

N. Debole[1], A. Passerini[1], S. Teso[1], A. Pugnana[1], E. Marconato[2]*
[1]University of Trento [2]University of Copenhagen
*Work conducted while at the University of Trento

NEURAL INFORMATION PROCESSING SYSTEMS

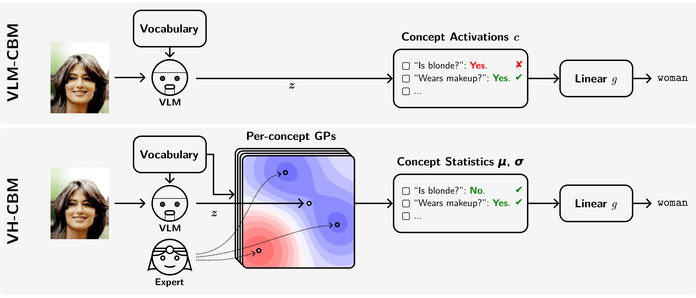
Standard CBM

Concepts are the explanation



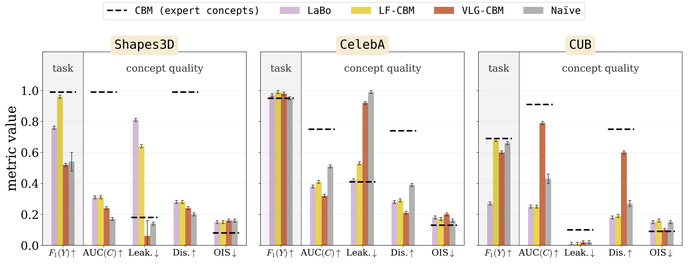
Overview of VH-CBM

Top: VLM-guided CBM. Bottom: ours



Concept annotation problem in VLM-CBMs

Good task accuracy, poor concepts



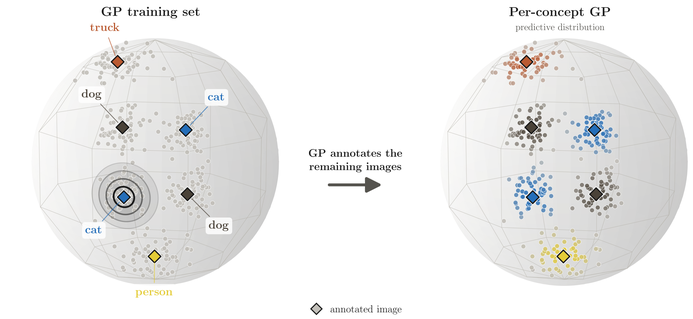
↑ higher is better, ↓ lower is better (CBM with expert concepts dashed)

- VLM-CBMs achieve good task accuracy, but what about their concepts?
- VLMs struggle to accurately detect fine-grained concepts [5].
- Most concepts learned by VLM-CBMs are thus low-quality: **Inaccurate**, **Entangled**, **Leaky** [5]. This makes them less interpretable.

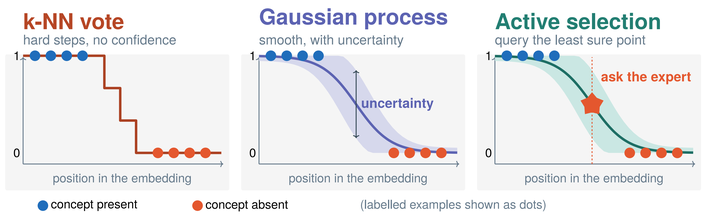
The idea: GP on the VLM geometry

Keep the embedding, drop the VLM labels

We discard the VLM labels (generated via the text encoder), keep the embedding geometry, and fit a Gaussian process on it.

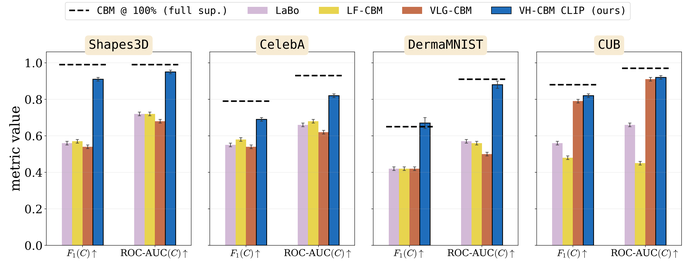


- The GP kernel measures similarity in the embedding space.
- The calibrated uncertainty drives active selection.



More accurate concepts

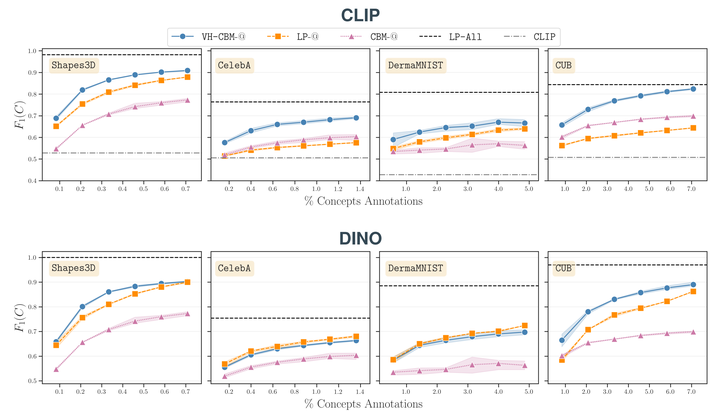
Concept F_1 and ROC-AUC, same annotation budget



↑ higher is better, dashed: CBM with full concept supervision

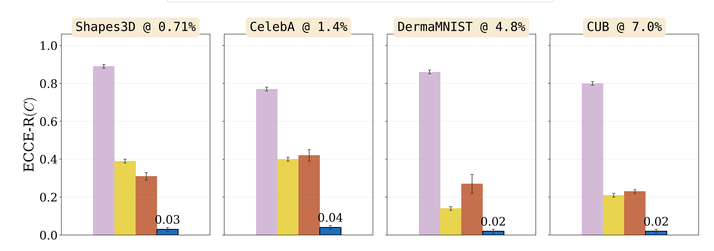
Works with any backbone

Concept F_1 vs. annotation budget



Better calibrated predictions

ECCE-R: lower is better



↓ lower is better: calibration error of the concept probabilities

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 (2023).
[4] Milios *et al.*, Dirichlet-based Gaussian processes for large-scale calibrated classification, NeurIPS (2018).
[5] Debole *et al.*, If concept bottlenecks are the question, are foundation models the answer?, Machine Learning (2026).

Paper, code & contact

Scan the QR code or get in touch

debryu/vhcbm
nicola.debole@unitn.it

TANGO