Caption Bottleneck Models

Seref Baris Cagliyan[1], Umut Ozdemir[1], Merve Tapli[1], and Emre Akbas[1,2]

[1] Dept. of Computer Eng., Middle East Technical University (METU)
[2] Robotics & AI Center (ROMER), METU

Abstract. Concept Bottleneck Models (CBMs) provide interpretability by routing predictions through a layer of human-understandable concepts. However, defining an optimal concept set for a specific dataset remains an open challenge. Existing approaches rely on expensive expert annotations or LLM-generated lists based solely on class names. Even "open-vocabulary" variants typically depend on static concept sets, which restrict discovery and introduce label bias. Furthermore, traditional CBMs often suffer from information leakage, where unmodeled visual features bypass the bottleneck and compromise the integrity of the explanations. To overcome these limitations, we propose Caption Bottleneck Models (CaBM), a framework that circumvents the need for predefined concept sets by replacing rigid concept layers with free-form natural language. By representing images via LMM-generated captions and training a classifier strictly on this text, CaBM ensures a leakage-free architecture by construction. Additionally, by analyzing the text classifier post-training, CaBM autonomously discovers high-quality, dataset-specific concepts. Our results across fine- and coarse-grained benchmarks demonstrate that CaBM achieves competitive accuracy while preserving interpretability without the constraints of external dictionaries or manual labeling. Our code is available at `https://github.com/bariscagliyan/CaptionBottleneckModels`.

Keywords: Concept Bottleneck Models · Concept Discovery

1 Introduction

Deep Neural Networks (DNNs) have achieved remarkable success across a wide range of domains, including computer vision and natural language processing. Their high accuracy and representational capacity have made them ubiquitous in real-world systems. However, a fundamental limitation remains: the mechanisms underlying their decisions are often opaque to human users. While high-dimensional intermediate representations enable state-of-the-art performance, they provide limited insight into *how* and *why* a particular prediction is produced. This "black-box" behavior hinders reliable deployment in high-stakes settings where interpretability, accountability, and user trust are essential alongside predictive performance.

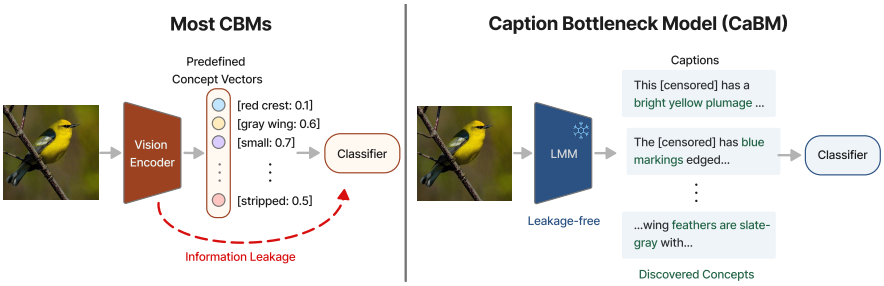

Fig. 1: CaBM performs recognition through a caption bottleneck: a frozen LMM produces captions, a text classifier predicts labels from these captions, and open-vocabulary concepts are extracted via post-hoc analysis.

Concept Bottleneck Models (CBMs) provide a principled approach for *interpretable by design* predictors: they first map an input to a set of human-interpretable concepts, and then predict the label from these concepts, enabling inspection and test-time interventions on the concept layer [8]. Subsequent works broadened the CBM paradigm in several directions, including post-hoc explanation of pretrained models with concept bottlenecks [32] and interactive variants that query humans for concept labels at inference time [4]. More recently, foundation models have been used to reduce the cost of building and annotating concept interfaces by automatically proposing candidate attributes and weakly grounding them with vision–language embeddings [17, 23, 29, 30]. This trend has also motivated concept-based explanations for zero-shot vision–language predictions [18], alongside systematic analyses of modern CBMs' design choices, evaluation pitfalls, and performance trade-offs [26].

Despite rapid progress, current CBM pipelines still face two open problems:

(i) Obtaining a concept set. A central practical bottleneck in developing CBMs is specifying a concept vocabulary that is both *human-meaningful* and *task-sufficient*. Early CBMs relied on image-level concept annotations provided by domain experts [8], which yields high-quality concept labels but is costly and hard to scale. To reduce supervision, recent methods increasingly *outsource concept definition* to foundation models. Typically, LLMs propose candidate concepts (often from class names), which are then grounded via zero-shot VLM alignment (*e.g.* using CLIP [21]) [11, 17, 31]. Although scalable, this approach often generates noisy or uncertain labels that undermine explanation fidelity and intervention reliability [19]. Even when additional filtering or selection is applied, these methods still begin from *externally proposed*, typically static LLM/VLM-mediated concept sets that may be incomplete, biased toward label semantics, or poorly grounded in the dataset's actual visual evidence [11, 25, 29, 30].

(ii) Information leakage undermining interpretability. The presence of a concept layer does not guarantee that internal representations capture *only* the intended concept information. Multiple studies show that CBMs, particularly "soft" CBMs trained jointly or sequentially, can encode task-relevant signals

that bypass the intended concept bottleneck layer, yielding high accuracy but unfaithful explanations [7, 13, 14]. This tension is often framed as an accuracy-faithfulness trade-off: passing *continuous concept probabilities* boosts accuracy but creates a high-bandwidth channel for information leakage that can corrupt concept predictions and undermine interventions [7]. Conversely, independently training the concept and classification layers of CBMs on hard (binarized) concepts makes them substantially more resilient to leakage, but significantly degrades their accuracy [7].

We propose **CaBM** (*Caption Bottleneck Models*), a new paradigm that addresses both challenges by *moving the bottleneck from structured concept vectors to natural language* and *enforcing independence* of the two stages (concept layer and classifier) *by design* (Fig. 1). Instead of predicting a fixed concept set, CaBM first translates an image into multiple *diverse* and *attribute-centric captions* using a *frozen* off-the-shelf Large Multimodal Model (LMM) guided by a structured prompt. We then train a text classifier on these captions to predict the original image label, turning recognition into language-space classification through a caption bottleneck. Importantly, this two-stage design mirrors the leakage-resistance principle of *independent training of CBM with hard concepts* [7]: the downstream predictor cannot shape the bottleneck through end-to-end gradients, and it never observes pixels – only the discrete text produced upstream. To further eliminate trivial shortcuts, we apply deterministic *taxonomy censoring* that removes explicit class names and higher-level category terms from the captions, ensuring that prediction relies on descriptive visual evidence rather than label-name cues.

Operating in free-form language space also enables *concept discovery*: after training the text classifier, we identify class-discriminative phrases directly from its decision process and aggregate them into per-class concept vocabularies. These concepts are *open-vocabulary by construction*, *data- and model-induced*, and naturally interpretable because they correspond to literal spans in the generated captions rather than opaque latent variables.

Our work makes the following contributions: (i) We introduce **CaBM**, a model-agnostic framework that decouples perception and recognition by performing classification *entirely in language space* through a caption bottleneck. (ii) We propose a practical **leakage-free caption pipeline** with structured multi-caption generation and deterministic taxonomy censoring, enabling robust caption-based supervision without label-name shortcuts. (iii) We introduce an automated **open-vocabulary concept extraction** approach that recovers per-class concepts directly from the trained classifier's behavior, eliminating predefined concept sets while yielding human-meaningful explanations.

2 Related Work

Open-vocabulary CBMs and concept-set construction. A major line of work aims to reduce the reliance on expert concept annotations by constructing concept sets automatically using foundation models. Approaches such as LaBo [30] and

label-free CBMs [17, 23, 29] query LLMs for candidate attributes and use CLIP-style vision–language alignment to associate images with concept text, enabling scalable concept supervision without concept-labeled datasets. Similarly, EZPC [18] projects CLIP's joint image–text embeddings into a predefined concept space via a learned linear projection, providing concept-level explanations of zero-shot predictions. OpenCBM [25] further relaxes the *fixed* concepts by allowing users to specify arbitrary textual concepts *after* training via CLIP-space reconstruction, enabling flexible add/remove/replace operations. However, these pipelines still fundamentally operate over an *externally specified* concept set (LLM-proposed or user-provided text concepts) whose coverage and bias are determined outside the target dataset; consequently, they can miss dataset-specific cues or over-emphasize label semantics rather than grounded evidence.

Automated concept discovery. Rather than defining concepts *a priori*, DN-CBM [22] inverts the usual workflow by first discovering a sparse set of latent "concepts" by training a sparse auto-encoder on a pretrained backbone and then naming them post-hoc by matching to a large vocabulary in a text-embedding space. This task-agnostic discovery improves reusability across datasets, but the semantic interface remains mediated by the choice of an external vocabulary and embedding space used for naming (*e.g.*, nearest-neighbor matching), which can constrain expressiveness and yield brittle or overly generic names. Extending this discovery line, DCBM [20] defines concepts as image regions obtained from segmentation/detection foundation models, improving data efficiency while keeping concepts *visual* rather than textual, and OCB [24] couples a CBM with a pre-trained object-centric model to ground concepts at the object level. Both remain in the image domain and rely on visual concepts. HybridCBM [11] addresses incomplete predefined concept sets by complementing them with a learned *dynamic* concept set and a concept translator that maps learned vectors back to text. While effective, HybridCBM still assumes a static set seeded by LLM-generated concepts and learns additional concepts as continuous embeddings aligned to that set, so the interpretation is ultimately tied to the external concept set and the translator's alignment quality. The same work also introduces CaptionCBM, a variant that replaces learned dynamic concepts with generated captions embedded as textual concepts; however, prediction still relies on CLIP image features at inference.

Language bottlenecks and caption-based explanations. A related direction replaces discrete concept lists with free-form textual explanations. XBMs [28] train a vision encoder and explanation decoder end-to-end to generate task-relevant natural-language explanations from image embeddings, then predict the task label using a classifier that conditions on both the generated text and image embeddings via cross-attention – image features are thus present at every stage of the pipeline, from explanation generation to final prediction. In contrast, CaBM enforces a *text-only* recognition channel by construction: a frozen LMM produces (censored) captions and the downstream classifier never observes pixels or image embeddings.

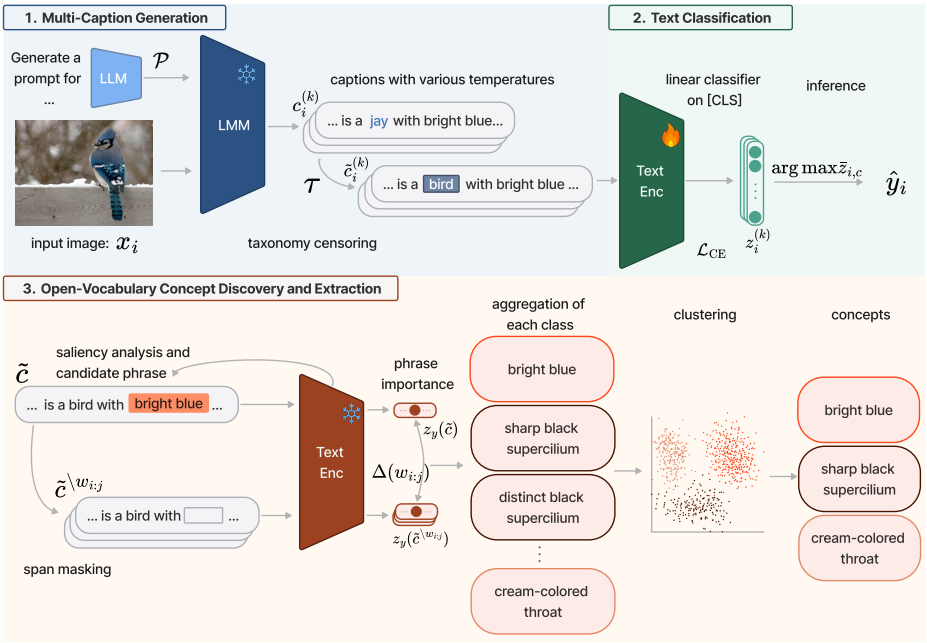

Fig. 2: Overview of CaBM. (1) A frozen LMM generates K diverse captions per image using a structured prompt $\mathcal{P}$ and multiple decoding temperatures; captions are deterministically censored by τ to remove class names and taxonomy terms. **(2)** A text encoder predicts per-caption logits, which are averaged for image-level inference. **(3)** Concepts are obtained post-hoc by proposing salient spans (gradient×embedding), scoring them with span masking (erasure), and semantically clustering phrases into compact per-class concept sets.

Positioning CaBM. We position CaBM as an alternative to both families: unlike open-vocabulary CBMs, it does not require an externally proposed concept set to be specified *before* or *after* training; instead, it performs recognition through a *caption bottleneck* and discovers concepts post-hoc from the trained text classifier's evidence. Unlike concept discovery methods that name latent vectors via dictionary lookup, CaBM's concepts are literal spans drawn from the model's own caption evidence, yielding true open-vocabulary, dataset-specific phrases while keeping the predictor strictly in the text domain.

3 Method

CaBM decouples perception from recognition by translating images into text and performing classification entirely on text. Our pipeline (Fig. 2) has three stages: (1) **multi-caption generation** with taxonomy censoring, (2) **text classification** on the resulting caption dataset, and (3) **open-vocabulary concept discovery and extraction** from the trained classifier and captions.

3.1 Multi-Caption Generation

Given a labeled image dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$ with $y_i \in \{1, \ldots, C\}$, we generate a set of diverse textual descriptions for each image using a Large Multimodal Model (LMM). For each image $\mathbf{x}_i$, we sample K captions $\mathcal{C}_i = \{c_i^{(k)}\}_{k=1}^{K}$:

$$c_i^{(k)} \sim p_\theta\Big(\cdot \mid \mathbf{x}_i, \mathcal{P}; T^{(k)}\Big), \qquad k = 1, \ldots, K, \tag{1}$$

where $\mathcal{P}$ is a fixed structured prompt and $T^{(k)}$ denotes the sampling temperature.

Structured prompting. To make captions discriminative across diverse datasets, we construct a single prompt template $\mathcal{P}$ using an LLM-assisted design step: given only the dataset domain and the nature of class granularity in the dataset (coarse or fine-grained), the LLM proposes an attribute checklist (*e.g.*, color, shape, texture/pattern, part-level cues, and spatial arrangement) and a constrained output format. The resulting prompt is frozen and used for all images. Importantly, $\mathcal{P}$ explicitly discourages taxonomic naming and encourages grounded descriptions.

Temperature-driven diversity. We generate K complementary captions per image by keeping $\mathcal{P}$ fixed and varying the sampling temperature. Lower temperatures favor concise, prototypical descriptions, while higher temperatures promote lexical and compositional diversity, yielding multiple textual "views" of the same image.

Decoding and normalization. We use standard stochastic decoding with conservative controls to stabilize generation and reduce degenerate repetition, and apply lightweight text normalization.

Taxonomy censoring. Even with constrained prompting, LMM outputs may contain label leakage via exact class names or higher-level taxonomic/group terms. We therefore apply a deterministic censoring function $\tau(\cdot)$ to each caption:

$$\tilde{c}_i^{(k)} = \tau\Big(c_i^{(k)}\Big), \tag{2}$$

which replaces any matched class string and any matched group term with a dataset-generic referent (*e.g.*, "the object"). This ensures downstream classification cannot rely on trivial lexical cues and must instead use descriptive visual attributes.

3.2 Text Classification

Given K censored captions per image $\mathcal{C}_i$ (Sec. 3.1), we learn a caption-to-label classifier and predict image labels by aggregating caption-level predictions. To prevent leakage between training and validation splits, we partition the data *at the image level* with class stratification and place all K captions of each image in the same split.

Model. Let Enc_ϕ be a pretrained transformer encoder. A caption is tokenized and truncated to a maximum length L, producing a sequence of $T \leq L$ tokens. The encoder outputs $\mathbf{H} = \mathrm{Enc}_\phi(\tilde{c}) \in \mathbb{R}^{T \times d}$. We use the final-layer `[CLS]`

representation $\mathbf{h} \in \mathbb{R}^d$ and apply a linear classifier:

$$\mathbf{z} = \mathbf{W}\mathbf{h} + \mathbf{b}, \qquad p_\phi(y \mid \tilde{c}) = \mathrm{softmax}(\mathbf{z}), \tag{3}$$

where $\mathbf{W} \in \mathbb{R}^{C \times d}$ and $\mathbf{b} \in \mathbb{R}^C$.

Training objective. We fine-tune Enc_ϕ end-to-end by minimizing cross-entropy over caption samples:

$$\mathcal{L}_{\mathrm{CE}} = -\log p_\phi(y \mid \tilde{c}). \tag{4}$$

Optimization details are provided in Sec. 4.1.

Multi-caption inference (logit voting). At test time, we compute logits $\mathbf{z}_i^{(k)}$ for each caption and predict at the image level by averaging logits across captions:

$$\bar{\mathbf{z}}_i = \frac{1}{K} \sum_{k=1}^{K} \mathbf{z}_i^{(k)}, \qquad \hat{y}_i = \underset{c \in \{1,\ldots,C\}}{\arg\max}\, \bar{z}_{i,c}. \tag{5}$$

This aggregation acts as a textual multi-view ensemble and reduces sensitivity to any single caption.

3.3 Open-Vocabulary Concept Discovery and Extraction

A central design choice in CaBM is that classification is carried out entirely in the text space, enabling the model's evidence to be decomposed into meaningful phrases. We propose an automated, open-vocabulary concept extraction pipeline that (i) proposes candidate phrases from correctly classified samples using gradient-based attribution, (ii) rescores them via span-masking (erasure) to obtain faithful importance scores, and (iii) consolidates and ranks them into compact, per-class concept vocabularies through semantic clustering.

Correctness filtering. To ensure that extracted phrases reflect features genuinely used for robust discrimination, we restrict extraction to correctly classified training images.

Gradient × embedding attribution. For each caption $\tilde{c}_i^{(k)}$ of a retained image, we compute token-level attribution scores using gradient × embedding [1]. Let $\mathbf{e}_t$ denote the embedding-layer representation of token t, and let z_{y_i} be the logit for the target class (ground-truth label during extraction). We define the raw saliency of token t as

$$s_t \;=\; \left\| \frac{\partial z_{y_i}}{\partial \mathbf{e}_t} \odot \mathbf{e}_t \right\|_1, \tag{6}$$

where $\odot$ is the Hadamard product. We zero the saliency of special tokens and normalize $\{s_t\}$ to sum to one over non-padding tokens, and sum subword scores to obtain word-level scores.

Candidate phrase proposal (saliency n-grams). Given the word-level saliency map, we propose candidate phrases using variable-length n-grams. For each span of words $w_{i:j}$, we compute a lightweight *proposal score* as the sum of its word saliencies. We remove generic filler terms (*e.g.*, "image", "appears") and edge stopwords, and select the top P candidates per caption using greedy non-maximum

suppression with a maximum of 50% span overlap. This yields up to KP candidate phrases per image.

Span-masking (erasure) scoring. Gradient attribution is used only to propose spans; *phrase importance* is computed via an erasure test that measures how much masking a candidate span decreases the target-class logit [10]. For a candidate phrase $w_{i:j}$ occurring in caption $\tilde{c}$, let $\tilde{c}^{\setminus w_{i:j}}$ denote the caption where the corresponding token span is erased (by dropping them via padding and zeroing the attention mask). We define the *phrase importance* as the (clipped) logit drop:

$$\Delta(w_{i:j}) \;=\; \max\Big(0,\; z_y(\tilde{c}) - z_y(\tilde{c}^{\setminus w_{i:j}})\Big). \tag{7}$$

Semantic clustering via encoder embeddings. We treat extracted phrases as short texts and map them to a semantic embedding space using the trained CaBM encoder's `[CLS]`-style pooled representation, which is standard in transformer encoders for classification and sentence-level representations [6]. Within each class, we cluster phrase embeddings with HDBSCAN [15], which is well-suited to discovering variable-density groups while marking outliers as noise. To reduce redundancy, we apply a lightweight post-processing merge based on centroid similarity and string-level normalization (substring/word-overlap), so that near-duplicate surface forms are consolidated into a single concept. Each cluster $\mathcal{G}$ is summarized by a representative phrase:

$$p^*(\mathcal{G}) \;=\; \arg\max_{p \in \mathcal{G}} \cos\big(\mathbf{e}(p), \boldsymbol{\mu}_{\mathcal{G}}\big). \tag{8}$$

where $\mathbf{e}(p)$ is the L2-normalized phrase embedding and $\boldsymbol{\mu}_{\mathcal{G}}$ is the (L2-normalized) cluster centroid. Finally, we assign each concept a *concept importance* score by aggregating phrase-level erasure evidence:

$$S(\mathcal{G}) \;=\; \sum_{m=1}^{|\mathcal{G}|} \Delta(p_m), \tag{9}$$

and rank concepts in descending order of $S(\mathcal{G})$.

4 Experiments

We comprehensively evaluate CaBM across six coarse- and fine-grained recognition benchmarks to validate its effectiveness, interpretability, and predictive performance. Our evaluation is structured around three core pillars. First, we assess the quality of our open-vocabulary concept discovery both quantitatively and qualitatively (Secs. 4.2 and 4.3), demonstrating that CaBM extracts semantically valid and highly visually grounded concepts compared to state-of-the-art baselines. Second, we evaluate the faithfulness of our model through test-time human interventions (Sec. 4.4), illustrating that modifying the caption bottleneck directly and predictably alters classification outcomes. Finally, we benchmark overall classification accuracy (Sec. 4.5), showing that despite operating strictly in the text domain during inference, CaBM achieves highly competitive predictive performance against traditional image-domain CBMs.

4.1 Experimental Setup

Datasets. We evaluate CaBM on six benchmarks: CIFAR-10, CIFAR-100 [9], CUB-200-2011 [27], Food-101 [3], Flowers-102 [16], ImageNet-1K [5]. For datasets without an official validation split, we create a stratified validation set by holding out 10% of training images; all K captions of an image remain in the same split.

Caption generation. We generate K=5 captions per image using structured prompting (Sec. 3.1) at temperatures $\{0.7, 0.9, 1.1, 1.3, 1.5\}$. For ImageNet-1K, we use K=3 captions for scalability. We employ a frozen LMM, Qwen3-VL-2B-Instruct [2], and apply deterministic taxonomy censoring (τ).

Text classifier. We use RoBERTa-base [12] for all datasets and train a linear classification head on the final-layer `[CLS]` representation, as described in Sec. 3.2. We fine-tune with AdamW (wd=0.05), a cosine learning-rate schedule with 6% warm-up, label smoothing (ϵ=0.1), batch size 16 with gradient accumulation (effective batch size 48), maximum sequence length 512, and early stopping with patience 6 for up to 30 epochs. We additionally use layer-wise learning-rate decay with factor 0.85, freeze the embeddings and the two lowest encoder layers, and apply light caption-level augmentation during training only (word dropout p=0.1 and short span dropout p=0.03, applied with probability 0.3). No augmentation is used at validation or test time.

At test time, we predict each image's label by averaging logits across its K captions (Eq. (5)). Details about compute and running time can be found in the supplementary material.

4.2 Open-Vocabulary Concept Discovery: Quantitative Evaluation

Concept quality. Beyond qualitative inspection, we evaluate the semantic quality of per-class concept sets using the metrics introduced in HybridCBM [11]. Given a concept set that provides textual concepts for each class $k \in \{1, \ldots, C\}$, we embed both concept phrases and class names with a CLIP text encoder [21], and report:

(i) Purity: the cosine similarity between the class-wise mean concept embedding (L2-normalized) and the CLIP text embedding of the corresponding class name, averaged over classes.

(ii) Separation: inter-class distinctiveness measured as the average pairwise cosine distance between class-mean concept embeddings; larger distances indicate more differentiated concept sets.

(iii) Semantics: an LLM-based validity rate (GPT-3.5) obtained from binary yes/no judgments of whether a concept phrase is associated with its intended class, reflecting interpretability and class relevance.

Tab. 1 reports results on CUB-200. CaBM attains high semantic validity with competitive purity and improved separation. Importantly, CaBM concepts are *discovered*, rather than selected from an externally proposed, *LLM-generated* vocabulary as in VLG-CBM. Despite this, CaBM surpasses VLG-CBM in Semantics and Separation, while matching purity.

Table 1: Quantitative comparison of concept sets on CUB200. Bold values indicate the best performance.

Method (CUB)	Purity	Separation	Semantics
VLG-CBM (NeurIPS'24) [23]	**0.52**	0.07	0.86
HybridCBM (CVPR'25) [11]	0.40	**0.80**	0.46
CaBM	**0.52**	0.16	**0.92**

Table 2: Downstream utility under NEC-controlled VLG-CBM. We keep the VLG-CBM evaluation protocol fixed and replace only the concept list: the original LLM-generated concepts vs. CaBM-discovered concepts. Top-1 accuracy is reported as ANEC-5 (NEC=5) and ANEC-avg (NEC $\in \{5, 10, 15, 20, 25, 30\}$) following [23].

	C10		C100		CUB	
	ANEC-5	ANEC-avg	ANEC-5	ANEC-avg	ANEC-5	ANEC-avg
VLG-CBM	88.55	88.63	65.73	66.48	75.79	75.82
CaBM	**88.87**	**88.97**	**66.13**	**66.66**	**76.25**	**76.34**

HybridCBM shows the highest Separation, but this is partly driven by its orthogonality regularization, which increases inter-class cosine distance by construction. In our evaluation, this comes with lower Purity and lower Semantics. CaBM instead yields image-grounded concepts, preserving strong Purity while substantially improving semantic validity at moderate Separation.

Downstream utility under controlled effective concepts. Complementary to CLIP-based concept-quality metrics (Purity/Separation) and LLM Semantics, we test whether the discovered concepts form a discriminative vocabulary under the NEC-controlled evaluation of VLG-CBM [23]. We keep the VLG-CBM evaluation pipeline fixed and swap *only* the concept set (LLM-generated vs. CaBM-discovered). Tab. 2 shows consistent gains on CIFAR-10, CIFAR-100, and CUB, indicating that CaBM concepts capture decision-relevant evidence even when constrained to a small, inspectable set of effective concepts.

4.3 Concept Discovery: Qualitative Comparisons

Cross-method concept inspection. Fig. 3 contrasts the types of concepts produced by each method. *For fairness, both the image instances and the listed concepts are taken directly from the experiments of the corresponding baseline papers; they are not selected or curated by us.* We run CaBM on these identical instances to enable a *like-for-like comparison.* Across examples, CaBM surfaces attribute-centric phrases that correspond to directly inspectable visual cues (*e.g.*, part color/shape and distinctive markings), yielding compact, instance-level explanations.

In comparison, baselines that rely on externally proposed concept sets or post-hoc translation exhibit a mixture of visually grounded evidence and state-

Fig. 3: Qualitative concept comparison across methods on CUB200. For each class, we list top-ranked concepts from CaBM and baselines. Concepts are manually tagged on the shown instance: OpenCBM marks *(new)* expansions; HybridCBM reports *Static* (S) and *Dynamic* (D). For a strictly fair comparison, we used the specific test images originally selected by the baseline papers.

ments that are either mismatched to the shown instance (red) or not visually verifiable (gray). For example, LaBo [30] may include geographic breeding information; OpenCBM's [25] expansion can increase coverage but also introduces relational/encyclopedic concepts (*e.g.*, diet or relatedness); and HybridCBM's [11] translated concepts can be diverse yet sometimes describe non-visual properties or mismatched attributes. LF-CBM [17] often produces semantically overlapping variants around a small number of cues, leading to redundancy. Overall, the tagging highlights that CaBM's concepts are typically concise and visually grounded with reduced redundancy relative to the compared baselines.

Concept importance across instances. To assess whether the discovered concepts provide consistent, instance-level evidence rather than isolated artifacts, we visualize concept importance across multiple images from the same class. Fig. 4 shows four randomly selected classes (one per dataset) and nine randomly sampled test images per class, together with the top-5 concepts and their importance scores (as used for ranking in Sec. 3.3). Across datasets, the highest-ranked concepts correspond to recurring, visually diagnostic attributes that reappear across diverse views and backgrounds within the class, while lower ranked concepts capture complementary cues. This supports that the concept sets are compact and reusable across instances, enabling inspection of which evidence is most influential for a given prediction.

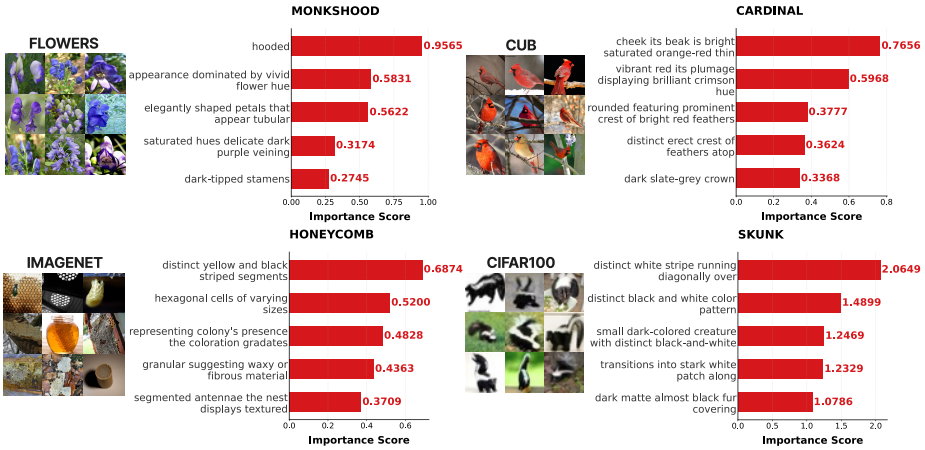

Fig. 4: Concept importance across images. For four datasets, we select one random class and sample nine random test images. For each class, we show the top-5 concepts and their importance scores (higher indicates stronger influence under the ranking criterion in Sec. 3.3).

Table 3: Human intervention via concept injection. Accuracy on the intervention subset (test images from the 20 most misclassified classes). We inject one top-ranked concept of the ground-truth class to each caption and re-evaluate the frozen classifier.

Dataset	Baseline Acc. (%)	Injected Acc. (%)	Δ
Flowers-102	60.6	72.8	+12.2
CIFAR-100	62.3	75.2	+13.0
CUB-200	41.7	55.5	+13.8

4.4 Faithfulness via Human Intervention

A key advantage of concept bottleneck models is *human intervention*: users can edit concepts and observe predictable changes in the output. We probe this property in CaBM by intervening directly at the caption bottleneck. For each dataset, we form an *intervention subset* consisting of test images from the 20 classes with the highest misclassification frequency. For each test instance, we inject the top-ranked concept from its ground-truth class to the caption and re-evaluate the same frozen text classifier. If the discovered concepts function as actionable evidence, such minimal edits should systematically improve accuracy.

Results. Tab. 3 reports accuracy on the intervention subset. Concept injection yields consistent gains across datasets (+12.2 to +13.8 points). Fig. 5 shows a representative case where a single injected concept flips an incorrect prediction to the correct class.

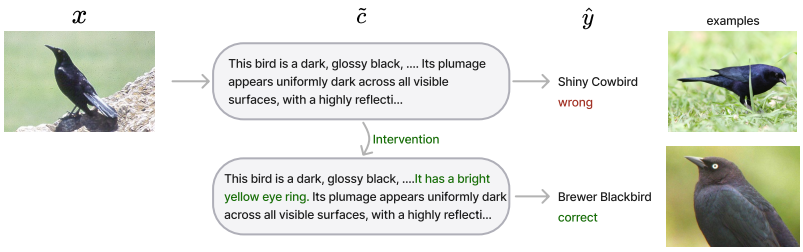

Fig. 5: Human intervention example. Prepending one concept to the caption flips the prediction under the same frozen classifier.

4.5 Classification Performance

Tab. 4 reports Top-1 test accuracy across six benchmarks. Despite operating *entirely* in the text domain at inference time, CaBM achieves competitive performance on both coarse- and fine-grained recognition tasks.

We compare against representative CBM baselines: PCBM [32], LF-CBM [17], LaBo [30], VLG-CBM [23] and PS-CBM [33] which operate in the *image* domain (typically via a CLIP-style image encoder and a concept bottleneck). We further include XBM [28], which uses a *multimodal* classifier over image and generated explanation text, retaining visual features at inference—fundamentally different from CaBM's strictly text-only pathway.

Discussion. Tab. 4 shows that CaBM attains competitive Top-1 accuracy under a strict *text-only* inference constraint. Direct accuracy comparison across CBMs is inherently limited because the methods differ in their visual backbones, concept set construction, and whether the classifier has access to image features. The purpose of this table is therefore not to claim that CaBM is the strongest vision classifier, but to show that a strictly text-only caption bottleneck remains competitive while providing faithful and inspectable concept-based explanations. Stronger vision-backbone CBMs such as HybridCBM and CBM-Suite achieve higher accuracy on several datasets, whereas CaBM trades some visual information for a leakage-resistant, open-vocabulary bottleneck whose concepts are mined directly from the captions.

5 Conclusion, Discussion & Limitations

In this work, we introduced Caption Bottleneck Models (CaBM), an interpretable-by-design framework that overcomes the reliance on predefined concept vocabularies and the vulnerability to information leakage inherent in traditional Concept Bottleneck Models (CBMs). By shifting the information bottleneck entirely into the natural language domain, CaBM strictly decouples perception from recognition. Using a frozen Large Multimodal Model (LMM) we generate diverse, taxonomy-censored captions, from which an independent text classifier predicts the target label. This discrete text interface ensures a leakage-free architecture

Table 4: Classification accuracy (%) comparison. Methods are grouped by bottleneck/protocol. CaBM predicts from LMM captions only, whereas other CBMs use visual backbones or image-derived concept representations. Baseline results are taken from the cited papers; LF-CBM, LaBo, VLG-CBM, and PS-CBM follow [33]. –: not reported.

Method	C10	C100	CUB	Food	Flowers	IN-1K
CBMs with soft concepts (CNN)						
PCBM (ICLR'23) [32]	77.70	52.00	58.80	–	–	–
LF-CBM (ICLR'23) [17]	87.30	68.80	58.60	77.70	94.40	67.50
LaBo (CVPR'23) [30]	87.50	68.10	66.80	82.20	96.60	72.10
VLG-CBM (NeurIPS'24) [23]	89.60	68.30	68.00	81.60	97.10	65.70
PS-CBM (AAAI'26) [33]	89.80	72.10	70.10	83.00	97.90	74.00
CBMs with soft concepts (ViT)						
HybridCBM (CVPR'25) [11]	97.93	86.22	84.25	92.62	99.23	83.67
CBM-Suite (CVPR'26) [26]	–	92.50	86.73	–	–	81.56
Explanation bottleneck (text + image classifier)						
XBM (AAAI'25) [28]	–	–	80.99	–	–	67.83
Strictly text bottleneck (text-only classifier)						
CaBM	96.42	81.14	76.92	93.68	86.57	75.12

by construction. Furthermore, our automated extraction pipeline successfully discovers dataset-specific, open-vocabulary concepts via gradient attributions and erasure signals. Evaluations across diverse benchmarks confirm that CaBM achieves competitive accuracy while yielding highly faithful, human-interpretable explanations that support targeted test-time interventions.

The evolution of CBMs reflects a broader trajectory toward scalable interpretability. Early CBMs relied on costly manual annotations, while subsequent generations leveraged LLMs to propose concepts and Vision-Language Models (VLMs) to ground them. However, these methods remained constrained by externally proposed, static concept sets. The CaBM paradigm represents the next logical step, enabled by recent advancements in powerful LMMs (*e.g.*, Qwen-VL) and robust text encoders (*e.g.*, RoBERTa). By harnessing these models to generate and classify rich, accurate visual descriptions, CaBM eliminates the need for predefined concept sets altogether, allowing the data's true visual evidence to drive concept discovery.

Limitations. While CaBM offers robust, leakage-free interpretability, it introduces a distinct architectural trade-off. A primary goal of certain post-hoc CBM variants (*e.g.*, PCBM) is the ability to retroactively convert an arbitrary, pre-trained black-box vision encoder into an explainable model. Because CaBM is inherently interpretable-by-design, bypassing traditional vision encoders in favor of a specialized LMM-to-text pipeline, it cannot be attached to existing black-box models to explain their internal representations. Consequently, CaBM

is best suited for scenarios where deploying a novel, transparent pipeline is preferred over auditing a legacy vision backbone.

Acknowledgements

The numerical calculations reported in this paper were fully performed at TUBITAK ULAKBIM, High Performance and Grid Computing Center (TRUBA resources). We also gratefully acknowledge the computational resources provided by METU Center for Robotics and Artificial Intelligence (METU-ROMER).

References

1. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. In: International Conference on Learning Representations (2018)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bossard, L., Guillaumin, M., Van Gool, L.: Food-101 – mining discriminative components with random forests. In: European Conference on Computer Vision. pp. 446–461 (2014)
4. Chauhan, K., Tiwari, R., Freyberg, J., Shenoy, P., Dvijotham, K.: Interactive concept bottleneck models. In: AAAI Conference on Artificial Intelligence. pp. 5948–5955 (2023)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
7. Havasi, M., Parbhoo, S., Doshi-Velez, F.: Addressing leakage in concept bottleneck models. In: Advances in Neural Information Processing Systems. vol. 35, pp. 23386–23397 (2022)
8. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International Conference on Machine Learning. pp. 5338–5348 (2020)
9. Krizhevsky, A., Hinton, G.: Learning multiple layers of features from tiny images. University of Toronto (2009)
10. Li, J., Monroe, W., Jurafsky, D.: Understanding neural networks through representation erasure. arXiv preprint arXiv:1612.08220 (2016)
11. Liu, Y., Zhang, T., Gu, S.: Hybrid concept bottleneck models. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 20179–20189 (2025)
12. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
13. Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., Pan, W.: Promises and pitfalls of black-box concept learning models. arXiv preprint arXiv:2106.13314 (2021)

14. Margeloiu, A., Ashman, M., Bhatt, U., Chen, Y., Jamnik, M., Weller, A.: Do concept bottleneck models learn as intended? arXiv preprint arXiv:2105.04289 (2021)
15. McInnes, L., Healy, J., Astels, S., et al.: hdbscan: Hierarchical density based clustering. J. Open Source Softw. **2**(11), 205 (2017)
16. Nilsback, M., Zisserman, A.: Automated flower classification over a large number of classes. In: Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729 (2008)
17. Oikarinen, T., Das, S., Nguyen, L.M., Weng, T.W.: Label-free concept bottleneck models. In: International Conference on Learning Representations (2023)
18. Ozdemir, O., Christensen, A., Alaniz, S., Akata, Z., Akbas, E.: Explaining CLIP zero-shot predictions through concepts. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 31336–31345 (2026)
19. Park, S., Mun, J., Oh, D., Lee, N.: An analysis of concept bottleneck models: Measuring, understanding, and mitigating the impact of noisy annotations. In: Advances in Neural Information Processing Systems (2025)
20. Prasse, K., Knab, P., Marton, S., Bartelt, C., Keuper, M.: Dcbm: Data-efficient visual concept bottleneck models. In: International Conference on Machine Learning (2025)
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning. pp. 8748–8763 (2021)
22. Rao, S., Mahajan, S., Böhle, M., Schiele, B.: Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In: European Conference on Computer Vision. pp. 444–461 (2024)
23. Srivastava, D., Yan, G., Weng, T.W.: Vlg-cbm: Training concept bottleneck models with vision-language guidance. In: Advances in Neural Information Processing Systems. vol. 37, pp. 79057–79094 (2024)
24. Steinmann, D., Stammer, W., Wüst, A., Kersting, K.: Object-centric concept-bottlenecks. In: Advances in Neural Information Processing Systems (2025)
25. Tan, A., Zhou, F., Chen, H.: Explain via any concept: Concept bottleneck model with open vocabulary concepts. In: European Conference on Computer Vision. pp. 123–138 (2024)
26. Tapli, M., Bouniot, Q., Stammer, W., Akata, Z., Akbas, E.: Rethinking concept bottleneck models: From pitfalls to solutions. In: IEEE Conference on Computer Vision and Pattern Recognition (2026)
27. Wah, C., Branson, S., Welinder, P., Perona, P., Belongie, S.: The Caltech-UCSD birds-200-2011 dataset. California Institute of Technology (2011)
28. Yamaguchi, S., Nishida, K.: Explanation bottleneck models. In: AAAI Conference on Artificial Intelligence. pp. 21886–21894 (2025)
29. Yan, A., Wang, Y., Zhong, Y., Dong, C., He, Z., Lu, Y., Wang, W.Y., Shang, J., McAuley, J.: Learning concise and descriptive attributes for visual recognition. In: International Conference on Computer Vision. pp. 3090–3100 (2023)
30. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 19187–19197 (2023)
31. Yu, L., Han, H., Tao, Z., Yao, H., Xu, C.: Language guided concept bottleneck models for interpretable continual learning. In: IEEE Conference on Computer Vision and Pattern Recognition. pp. 14976–14986 (2025)

32. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: International Conference on Learning Representations (2023)
33. Zhao, D., Huang, Q., Yan, D., Sun, Y., Yu, J.: Partially shared concept bottleneck models. In: AAAI Conference on Artificial Intelligence (2026)