Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Survey Track

Concept Bottleneck Models for Explainable Decision Making: A Survey of Progress, Taxonomy, and Future Directions

Chunjiang Wang[1,2], **Fan Li**[1,2], **Wenbo Hu**[1,2], **Rui Yan**[1,2], **Kun Zhang**[1,2*], **Shaohua Kevin Zhou**[1,2,3,4*]

[1]School of Biomedical Engineering, University of Science and Technology of China, Hefei, China
[2]Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China
[3]Jiangsu Provincial Key Laboratory of Multimodal Digital Twin Technology, Suzhou, China
[4]State Key Laboratory of Precision and Intelligent Chemistry, USTC
{chunjiang_wang, lf_2002, huwenbo}@mail.ustc.edu.cn, {yanrui, kkzhang, skevinzhou}@ustc.edu.cn

Abstract

Deep neural networks deliver strong performance but remain opaque, limiting their use in high-stakes domains that require transparency and human oversight. Concept Bottleneck Models (CBMs) address this gap by introducing a human-interpretable concept layer that mediates inputs and decisions, enabling semantic explanations and test-time intervention. This survey provides a unified review of CBMs organized along four dimensions: concept acquisition, concept-based decision making, concept intervention, and concept evaluation. We summarize the evolution of concept construction from manual annotation to lexicon-based mining, LLM/VLM-guided generation, and visually grounded discovery via prototypes and diffusion models; review emerging CBM architectures beyond strict bottlenecks; and consolidate evaluation and intervention protocols emphasizing faithfulness, sparsity, and intervenability, with particular relevance to high-stakes domains such as healthcare. We synthesize fragmented literature and outline key challenges and future directions for concept-based interpretable decision making.

1 Introduction

Deep neural networks have achieved strong performance in vision, language, and multimodal learning, enabling broad real-world adoption in healthcare, medicine, and so on [Mpinda *et al.*, 2026]. However, their decision-making is often opaque, creating risks in high-stakes settings that require trust, accountability, and human oversight. This performance-interpretability gap has driven explainable AI (XAI), which aims to make model reasoning understandable, verifiable, and correctable by humans.

Concept Bottleneck Models (CBMs) have emerged as a principled and influential paradigm to address this challenge [Koh *et al.*, 2020]. Rather than mapping inputs directly to outputs, CBMs explicitly factorize the prediction process by introducing an intermediate layer of human-understandable concepts between inputs and decisions. This structured decomposition enables semantic explanations, facilitates expert inspection of intermediate reasoning, and supports test-time intervention through concept correction. These properties distinguish CBMs from post-hoc explanation techniques [Selvaraju *et al.*, 2017] and position them as a unifying framework for interpretable decision-making.

Early concept bottleneck models relied on manually defined and annotated concepts (e.g., visual attributes or clinical findings) to provide a transparent intermediate interface, but were limited by annotation cost, incomplete coverage, and label noise [Koh *et al.*, 2020]. Recent progress can be organized into four stages of concept-based reasoning: **concept acquisition** is moving from manual curation to scalable lexicon mining, LLM/VLM-guided generation, and visually grounded discovery via prototypes or diffusion models, enabled by large-scale foundation models [Radford *et al.*, 2021; Yang *et al.*, 2023; Tan *et al.*, 2024b]; **concept-based decision making** is evolving from strict bottlenecks toward soft, hybrid, probabilistic, and energy-based designs that preserve a concept interface while improving predictive capacity [Espinosa Zarlenga *et al.*, 2022; Liu *et al.*, 2025; Kim *et al.*, 2023; Xu *et al.*, 2024]; **concept intervention** increasingly supports structured and dependency-aware corrections to better propagate human feedback through the concept layer [Shin *et al.*, 2023; Jeon *et al.*, 2024a]; and **concept evaluation** is expanding beyond accuracy to interpretability-centric metrics (e.g., faithfulness, consistency under intervention, robustness to noisy or missing concepts), with ongoing challenges in visual grounding, semantic stability, and alignment with human reasoning [Yang *et al.*, 2023].

Despite rapid progress, the CBM literature remains fragmented. Existing surveys [Hou *et al.*, 2024] and reviews typically focus on attribute learning, prototype-based explanations, or general interpretability methods, but they do not capture the broader evolution of CBMs as an integrated framework spanning concept acquisition, decision-making architectures, intervention mechanisms, and evaluation protocols. In particular, the recent convergence of CBMs with foundation models, editable learning, and multimodal interaction has not yet been systematically synthesized.

*Corresponding authors.

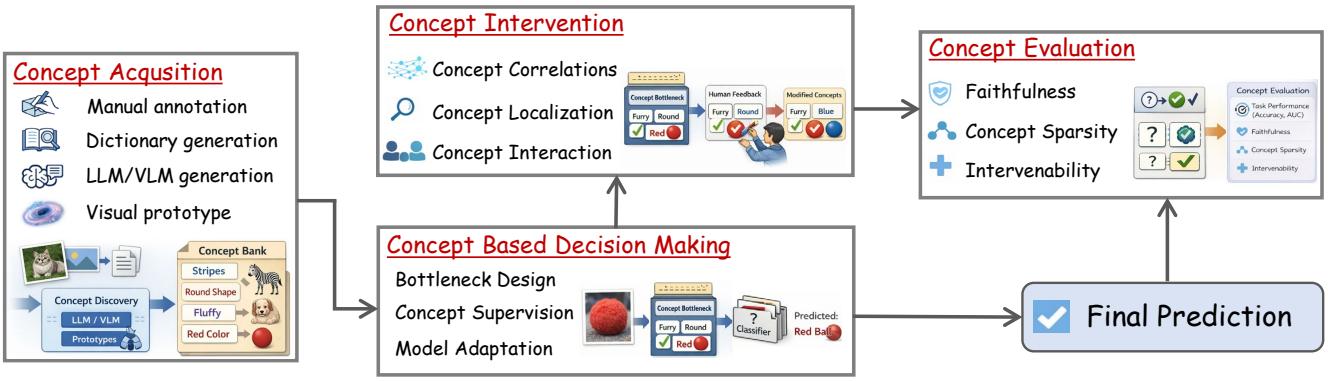

Figure 1: Overview of the CBM pipeline. The figure illustrates how explicit concepts are acquired, integrated into decision making, intervened upon, and evaluated to produce interpretable predictions.

The main contributions of this survey can be summarized as follows. **(1)** A unifying taxonomy of CBMs across concept acquisition, decision making, intervention, and evaluation. **(2)** A systematic review of concept construction methods from manual annotation to LLM-guided and prototype-based discovery. **(3)** A synthesis of emerging CBM architectures with different bottleneck designs and degrees of editability. **(4)** A consolidation of evaluation protocols and benchmarks, with emphasis on faithfulness, sparsity, intervenability, and medical applications. We maintain an updated repository of representative CBM papers, code, and benchmarks to support reproducibility and community use at Awesome CBMs.

2 Preliminary

2.1 Concept Bottleneck Model

Concept Bottleneck Models (CBMs) [Koh *et al.*, 2020] are a class of interpretable neural architectures that explicitly decompose prediction into two stages through human-understandable intermediate variables, referred to as *concepts*. Formally, a CBM factorizes the mapping from an input $x \in \mathcal{X}$ to a target $y \in \mathcal{Y}$ as

$$x \rightarrow c \rightarrow y, \tag{1}$$

where $c = (c_1, \ldots, c_K) \in \mathcal{C}^K$ denotes a set of interpretable concepts, such as visual attributes, anatomical structures, pathological findings, or semantic descriptions.

A canonical CBM consists of two components: a *concept predictor* $f_\theta : \mathcal{X} \rightarrow \mathcal{C}^K$, which maps raw inputs to concept activations, and a *task predictor* $g_\phi : \mathcal{C}^K \rightarrow \mathcal{Y}$, which produces the final prediction based solely on the predicted concepts. The overall model can be written as:

$$\hat{y} = g_\phi(f_\theta(x)). \tag{2}$$

This structure enables two defining capabilities of CBMs: (*i*) *semantic interpretability*, as each concept dimension aligns with a human-recognizable property; and (*ii*) *intervenability*, since users may inspect or modify c at test time to correct reasoning errors. These properties motivate a range of architectural variants discussed later in this survey, including models that relax the strict bottleneck constraint, incorporate structured dependencies among concepts, or leverage external knowledge such as multimodal foundation models.

2.2 Taxonomy

Figure 1 presents a four-part taxonomy of Concept Bottleneck Models, while Table 1 summarizes representative methods across these dimensions. We distinguish four primary dimensions: (1) *Concept Acquisition*, which concerns how human-interpretable concepts are constructed, ranging from expert annotation to automated discovery and language-induced vocabularies; (2) *Concept-based Decision Making*, which specifies how the concept layer mediates prediction, including strict bottlenecks and relaxed or hybrid designs; (3) *Concept Intervention*, which studies how humans or external systems interact with concepts at inference time to correct or guide reasoning; and (4) *Concept Evaluation*, which assesses interpretability-oriented properties such as faithfulness, sparsity, and robustness under intervention.

Importantly, a growing body of work cuts across these dimensions, such as jointly integrating concept acquisition, decision making, and intervention within a single architecture. We therefore review each dimension in turn and highlight representative methods.

3 Concept Acquisition

Concept construction is the foundation of Concept Bottleneck Models, defining a human-interpretable interface between inputs and predictions. Concept construction has evolved from fully manual annotation to dictionary-based mining, LLM/VLM-guided generation, and visual prototype discovery. The overarching trend is to reduce manual supervision while improving concept quality, scalability, and visual grounding. This section reviews four representative paradigms and their core mechanisms.

3.1 Manual Annotation-based Concepts

Manual annotation provides the strongest semantic grounding by relying on expert-defined and explicitly labeled concepts, such as visual attributes or clinical findings. This paradigm ensures high interpretability and clear alignment with domain knowledge, but is fundamentally constrained by annotation cost, incomplete concept coverage, and label noise.

Recent work within this paradigm focuses on improving robustness and flexibility under imperfect supervision. One

line of research addresses noisy or missing annotations by modifying the training objective, exemplified by CPO [Penaloza *et al.*, 2025], which introduces preference-based optimization to jointly stabilize concept prediction and downstream performance. A complementary direction models dependencies among annotated concepts to mitigate leakage and limited expressiveness. Havasi et al. [Havasi *et al.*, 2022] augment manually annotated concepts with latent side-channel variables and adopt autoregressive predictors to capture inter-concept structure. To support practical deployment, editable CBMs [Hu *et al.*, 2025] enable post-hoc correction of concept definitions and dataset labels using influence-function–based updates, without requiring full retraining. Despite these advances, manual annotation remains difficult to scale to complex domains, and its reliance on predefined concept sets limits adaptability to evolving tasks.

3.2 Dictionary-Based Concept Construction

Dictionary-based approaches replace task-specific manual definition with scalable discovery from predefined vocabularies, such as adjective lists or medical descriptors. By constraining the concept space to linguistically meaningful units, these methods offer improved controllability and transparency compared to fully automated generation.

Representative methods focus on filtering visually grounded concepts from large lexicons. V2C-CBM [He *et al.*, 2025c] constructs a candidate pool of n-grams and uses CLIP-based vision–language similarity to remove non-visual or weakly grounded entries, yielding compact and interpretable concept sets. OpenCBM [Tan *et al.*, 2024b] further aligns trainable visual features with CLIP embeddings and introduces mechanisms for discovering missing concepts, improving coverage beyond the initial dictionary.

While dictionary-based construction reduces annotation cost and limits hallucination risk, its expressiveness remains bounded by the predefined vocabulary, which may fail to capture fine-grained or task-specific semantics.

3.3 LLM/VLM-Guided Concept Generation

LLM- and VLM-guided concept generation substantially expands the concept space by leveraging large language models to produce diverse, high-level, and often hierarchical semantics. This paradigm offers strong scalability and flexibility, but introduces new challenges related to hallucination, weak visual grounding, and semantic instability.

Recent work addresses these issues through structure, alignment, and statistical regularization. Med-MICN [Hu *et al.*, 2024] organizes concepts into multi-level hierarchies and employs gating mechanisms to control abstraction. Vision–language consistency constraints [Lai *et al.*, 2023; Rao *et al.*, 2024; Schrodi *et al.*, 2025] and prompt-level alignment strategies [Bie *et al.*, 2024] further improve grounding reliability. From a supervision perspective, Label-free CBMs and FCBM [Oikarinen *et al.*, 2023; Du *et al.*, 2025] demonstrate that usable concepts can be induced without explicit annotations by coupling LLMs with CLIP, provided that strong visual regularization is imposed.

Overall, LLM/VLM-guided concept acquisition significantly lowers human supervision requirements, but its interpretability ultimately depends on the reliability of language priors and the effectiveness of grounding constraints.

3.4 Visual Prototype-based Concept Discovery

Visual prototype-based methods learn concepts directly from the image space by extracting representative regions, parts, or disentangled components, yielding strong visual grounding and spatial localization. This paradigm is particularly appealing for medical imaging and other domains where morphological evidence is central to decision making.

Prototype-based discovery faces inherent challenges, including unstable semantic naming, saliency-driven spurious correlations, and trade-offs between prototype granularity and coverage. Recent work explores several design patterns to mitigate these limitations. Disentanglement-oriented approaches, such as HU-MCD, LCBM [Grobrügge *et al.*, 2025; Jeon *et al.*, 2024b], learn multi-dimensional concept factors to reduce semantic underspecification. Deployment-oriented methods [Tan *et al.*, 2024a; Choukali *et al.*, 2024] focus on extracting prototype dictionaries from pretrained models to enable post-hoc interpretability. Interaction-oriented designs [Huang *et al.*, 2024; Huy *et al.*, 2025] explicitly associate prototypes with spatial regions, supporting region-level inspection and intervention.

Despite improved grounding fidelity, achieving stable semantic alignment and scalable reuse of visual prototypes across datasets remains an open challenge.

Summary. Concept acquisition in CBMs spans a spectrum from expert-defined semantics to automated discovery, reflecting trade-offs between semantic precision, supervision cost, and robustness, and motivating hybrid designs that integrate language and visual cues.

4 Concept-based Decision Making

Concept Bottleneck Models (CBMs) improve interpretability by mediating predictions through a layer of interpretable concepts, enabling validation and intervention. Beyond concept construction, CBM architectures have evolved to enhance flexibility and deployability. This section reviews three key directions: Bottleneck Design, Concept Supervision, and Model Adaptation, highlighting a shift from rigid pipelines toward adaptive, human-aligned decision making.

4.1 Bottleneck Design

A core innovation of CBMs lies in how the concept bottleneck governs decision making by controlling the information flow between inputs and outputs. Early CBMs adopted hard bottlenecks, enforcing a strict pipeline where all predictive signals pass through a discrete, interpretable concept layer. Such designs provide strong transparency and intervenability, but their rigidity may limit capacity and degrade performance under imperfect or incomplete concept supervision.

To address these limitations, soft bottlenecks were introduced [Mahinpei *et al.*, 2021; Margeloiu *et al.*, 2021], enabling continuous concept activations and joint optimization of concept and task predictors. This relaxation improves flexibility and predictive capacity, but makes the semantic boundary of each concept less explicit. Subsequent variants relaxed the strict bottleneck through residual pathways [Shang

Category	Method	Concept			Decision		Intervention		Code
		Source	Gran.	Sup.	Bottleneck	Corr.	Mechanism	Policy	Link
Acquisition	FVLC [Lai *et al.*, 2023]	LLM	global	weak	soft	indep.	–	–	Link
	OpenCBM [Tan *et al.*, 2024b]	dict.	local	none	hybrid	indep.	param.	vanilla	–
	Med-MICN [Hu *et al.*, 2024]	LLM	global	weak	hybrid	joint	–	–	–
	XCoOp [Bie *et al.*, 2024]	manual/LLM	global + local	strong/weak	sotf	indep.	–	–	Link
	PCPPN[Choukali *et al.*, 2024]	proto.	local	unsup.	soft	indep.	–	–	Link
	Tan et al. [2024a]	proto.	local	unsup.	soft	indep.	scalar	vanilla	–
	HU-MCD[Grobrügge *et al.*, 2025]	proto.	local	unsup.	–	indep.	–	–	Link
	CPO [Penaloza *et al.*, 2025]	manual	global	strong	soft	indep.	scalar	active	Link
	V2C-CBM [He *et al.*, 2025c]	dict.	global	weak	soft	indep.	scalar	vanilla	Link
Decision	CBMs[Koh *et al.*, 2020]	manual	global	strong	hard	indep.	scalar	vanilla	Link
	PCBMs [Yuksekgonul *et al.*, 2023]	manual/dict.	global	weak	hybrid	indep.	scalar/param.	vanilla	Link
	CLIP-QDA [Kazmierczak *et al.*, 2024]	dict.	global	weak	soft	joint	scalar	–	–
	Label-Free CBM [Oikarinen *et al.*, 2023]	LLM	global	weak	soft	indep.	scalar	vanilla	Link
	Explicd [Gao *et al.*, 2024b]	LLM	global	weak	soft	indep.	–	–	Link
	ECBMs [Hu *et al.*, 2025]	manual	global	strong	soft	indep.	param.	vanilla	Link
	AdaCBM [Chowdhury *et al.*, 2024]	LLM	global	weak	soft	indep.	–	–	Link
	Graph CBM [Xu *et al.*, 2025]	manual/dict.	global	strong/weak	soft	graph	scalar	prop.	Link
	C^2BMs [De Felice *et al.*, 2025]	LLM	global	weak	soft	causal	scalar	prop.	Link
	MVP-CBM [Wang *et al.*, 2025]	LLM	global	weak	soft	indep.	–	–	Link
Intervention	Havasi et al. [2022]	manual	global	strong	hard	joint	scalar	prop.	Link
	SCBMs [Vandenhirtz *et al.*, 2024]	manual/LLM	global	weak	hard	joint	scalar	active + prop.	Link
	ECBMs [Xu *et al.*, 2024]	manual	global	strong	hybrid	joint	scalar	prop.	Link
	evi-CEM [Gao *et al.*, 2024a]	manual/LLM	global	strong/weak	hybrid	indep.	scalar	active	Link
	CSR [Huy *et al.*, 2025]	manual + proto.	local	strong	soft	indep.	scalar/spatial	vanilla	Link
	Concept Realignment[Singhi *et al.*, 2024]	manual	global	strong	soft/hybrid	joint	scalar	active + prop.	Link
	CB2M [Steinmann *et al.*, 2024]	manual	global	strong	soft	indep.	scalar	active	Link
	Chat-CBM [He *et al.*, 2025b]	LLM	local	none	hard	indep.	lang.	vanilla	–
	SALF-CBM [Benou and Raviv, 2025]	LLM	local	weak	soft	indep.	spatial	vanilla	Link
Evaluation	CEM [Espinosa Zarlenga *et al.*, 2022]	manual	global	strong	soft	indep.	scalar	vanilla	Link
	IntCEMs [Espinosa Zarlenga *et al.*, 2023]	manual	global	strong	soft	indep.	scalar	active	Link
	CF-CBM [Panousis *et al.*, 2024]	LLM	global + local	weak	hard	hier.	scalar	prop.	Link
	Res-CBM [Shang *et al.*, 2024]	manual/LLM	global	weak	soft	indep.	scalar	vanilla	Link
	ProtoCBM[Huang *et al.*, 2024]	manual + proto.	local	strong	soft	indep.	–	–	Link
	VLG-CBM [Srivastava *et al.*, 2024]	LLM	local	weak	soft	indep.	–	–	Link
	PS-CBM[Zhao *et al.*, 2026]	LLM	global	weak	soft	joint	scalar	vanilla	Link
	HybridCBM [Liu *et al.*, 2025]	LLM + proto.	global	weak	soft	indep.	–	–	Link

Table 1: A unified taxonomy of CBM methods spanning concept acquisition, decision making, intervention, and evaluation. Concept Source: manual annotation(manual), dictionary (dict.), visual prototype(proto.), Large Language Model (LLM). Concept Spatial Granularity (Gran.): the spatial level at which concepts align with visual evidence: global, local, or both. Concept Supervision (Sup.): strong, weak, unsupervised (unsup.). Concept Bottleneck Layer (Bottleneck): hard, soft, hybrid. Concept Correlation Modeling (corr.): independent (indep.), joint, graph, hierarchical (hier.) causal. Intervention Mechanism: numerical adjustment (scalar), parametric (param.) mask/attention editing (spatial), natural language dialog (lang.). Intervention Policy: random/manual selection (vanilla), model-driven suggestion (active), correction propagation (prop.). Code: Link (code available), – (not released).

et al., 2024] and structured dependencies, including autoregressive sequencing [Havasi *et al.*, 2022], graphical priors [Xu *et al.*, 2025; Singhi *et al.*, 2024], and stochastic latent variables [Vandenhirtz *et al.*, 2024]. These designs capture inter-concept correlations and uncertainty difficult to represent with independent concept predictions. Energy-based CBMs [Xu *et al.*, 2024] unify prediction and intervention within a joint probabilistic framework, enabling decisions to be revised by modifying concept states. More recent work injects higher-level structure into the bottleneck, including causal graphs for robust intervention [De Felice *et al.*, 2025], concept decoupling to mitigate concept–label distortion [Zhang *et al.*, 2024], and differentiable logical composition to move beyond linear aggregation [Vemuri *et al.*, 2026; Gao *et al.*, 2025].

Together, these advances reflect a gradual relaxation of strict bottlenecks toward more expressive and structured architectures, enabling CBMs to better balance interpretability and predictive capacity in complex real-world settings, while preserving the concept layer as an aligned semantic interface for explanation and intervention.

4.2 Concept Supervision

Concept supervision defines how the intermediate semantic layer of CBMs is constructed and grounded. While early models relied on manually annotated concepts with strong domain semantics [Koh *et al.*, 2020], such supervision is often costly, sparse, or domain-specific. Consequently, recent research has pursued alternatives that scale to broader tasks while retaining semantic interpretability.

Post-hoc CBMs [Yuksekgonul *et al.*, 2023] bypass architectural constraints by learning concept predictors atop frozen encoders, recovering concept interpretability without retraining. Probing-based approaches [Semenov *et al.*, 2024] and prototype discovery methods [Huang *et al.*, 2024; Tan *et al.*, 2024a] extract interpretable units from pretrained features using sparsity regularization or part-based decomposition. Leveraging vision-language models, CLIP-based pipelines [Kazmierczak *et al.*, 2024; Oikarinen *et al.*, 2023; Lin *et al.*, 2025] align concept labels with textual prompts, enabling open-vocabulary supervision. LLM-guided strategies [Hu *et al.*, 2024; Lai *et al.*, 2023; Gao *et al.*, 2024b; Wang *et al.*, 2025] go further by generating hierarchical or

task-specific concepts, though challenges in visual grounding and hallucination remain.

These approaches shift concept supervision from direct annotation to alignment- and prompt-based induction, enabling scalable and adaptive concept construction. This evolution preserves the human-aligned semantics of CBMs while improving flexibility across tasks and domains.

4.3 Model Adaptation

For CBMs to be practically useful in real-world and evolving environments, they must support dynamic adaptation. Static pipelines with fixed concepts are difficult to update when task requirements shift, new concepts emerge, or user corrections are needed. As a result, model adaptation has become a critical frontier for operational CBMs.

Editable CBMs [Hu *et al.*, 2025] introduce influence function–based parameter updates, enabling localized changes to concept, instance, or label predictions without full retraining. This allows for fine-grained corrections and dataset updates while preserving the model's overall structure. Probabilistic energy-based models [Xu *et al.*, 2024] support test-time correction by formulating inference as energy minimization over inputs, concepts, and outputs. More interactively, recent methods such as Chat-CBM [He *et al.*, 2025b] and SALF [Benou and Raviv, 2025] expand the model interface to support natural language and spatial-region–based interventions, making adaptation accessible to non-technical users. Training-free test-time adaptation [Chowdhury *et al.*, 2024; He *et al.*, 2025a]further enables CBMs to handle concept-level distribution shifts using only a few image-level labels, without retraining or concept supervision.

These advances mark a shift from static interpretability toward dynamic, human-in-the-loop decision-making. Adaptable CBMs retain their core semantic bottleneck while allowing post-deployment editing, correction, and collaboration, extending their viability in real-world, high-stakes systems.

Summary. Concept-based decision making in CBMs has evolved from strict bottleneck pipelines to more flexible architectures that better balance interpretability and performance. Advances in bottleneck design, supervision, and model adaptation position the concept layer as a dynamic interface for prediction and intervention.

5 Concept Intervention

Concept intervention enables human and AI collaboration by correcting intermediate concepts, transforming CBMs from passive interpretable models into interactive systems. In the seminal work of Koh et al. [2020], intervention was formulated as a test-time edit that allows users to modify concept activation values. While foundational, this paradigm relies on strong independence assumptions, incurs high human effort, and supports limited interaction modalities.

5.1 Concept Correlations

Early intervention mechanisms typically assume that concepts are independent, implying that modifying one concept does not affect others. This assumption is often unrealistic in real-world settings, where concepts exhibit strong semantic and statistical dependencies in practice across complex decision tasks. To address this limitation, correlation-aware intervention methods explicitly model inter-concept relationships and propagate corrections accordingly.

Havasi et al. [2022] proposed an autoregressive concept predictor, allowing interventions on earlier concepts to influence subsequent predictions through sequential dependency modeling. Stochastic CBMs [Vandenhirtz *et al.*, 2024] further improve efficiency by modeling concept logits as a multivariate normal distribution, enabling instantaneous propagation of a single intervention through the learned covariance structure. Energy-Based CBMs [Xu *et al.*, 2024] adopt a joint energy formulation over inputs, concepts, and outputs, allowing concept corrections to propagate via energy minimization. Singhi et al. [2024] introduced a post-hoc Concept Realignment Module (CIRM), which updates non-intervened concepts based on learned statistical correlations.

While these approaches significantly improve intervention consistency and efficiency, they critically depend on the quality of learned correlations. If spurious dependencies are captured, interventions may propagate errors rather than corrections, highlighting an important open challenge.

5.2 Concept Localization

Identifying which concepts to intervene remains a major bottleneck in practical deployment, as manual selection is labor-intensive and cognitively demanding. To alleviate this burden, recent work has shifted toward model-guided localization strategies that prioritize high-impact intervention targets.

Some approaches internalize localization into training or model design. Espinosa Zarlenga et al. [2022] introduced intervention-aware training that simulates test-time intervention during learning, enabling the model to recommend concepts for correction. Evidential CEM (evi-CEM) [Gao *et al.*, 2024a] models concept predictions as Beta distributions, using epistemic uncertainty to guide intervention toward ambiguous concepts. Other methods rely on post-hoc analysis or auxiliary structures. Shin et al. [2023] proposed metrics such as Uncertainty-based Concept Potential (UCP) and Loss on Concept Prediction (LCP) to rank intervention priorities. Shen et al. [2025] further introduced FIGS-BD, which distills predictors into greedy sum-trees to identify concept groups with high contribution variance.

5.3 Concept Interaction

Beyond selecting and correcting concept values, recent work increasingly explores richer interaction modalities that are more intuitive for human users [Steinmann *et al.*, 2024]. Chat-CBM [He *et al.*, 2025b] replaces the concept-based predictor with large language models, allowing users to intervene through natural language dialogue. While this improves usability in practice, it may weaken the strict interpretability guarantees of traditional CBMs.

In the visual domain, Concept-based Similarity Reasoning (CSR) [Huy *et al.*, 2025] enables spatial interaction by allowing users to draw bounding boxes that guide model attention. However, such spatial guidance does not explicitly decouple interventions across distinct concepts. Spatially-Aware and Label-Free (SALF) CBM [Benou and Raviv, 2025] addresses

this limitation by leveraging spatial concept maps, enabling users to specify a region of interest and selectively modulate the activation of a single concept within that region. Beyond discriminative models, Concept Bottleneck Generative Models extend concept-level interaction to GANs, VAEs, and diffusion models, enabling interpretable steering and debugging of generation [Ismail *et al.*, 2024].

Summary. Concept intervention transforms CBMs from static explanatory models into interactive human-in-the-loop systems. By modeling concept correlations, enabling efficient localization of intervention targets, and supporting richer interaction modalities, recent advances improve the effectiveness and usability of concept-level correction.

6 Evaluation for Concept Bottleneck Models

Concept Bottleneck Models introduce interpretable and intervenable intermediate concepts, requiring evaluation protocols that go beyond downstream task performance. Existing evaluation frameworks therefore assess three core dimensions: concept faithfulness, concept sparsity, and intervenability, capturing semantic validity, cognitive tractability, and human-in-the-loop controllability.

6.1 Concept Faithfulness

Concept faithfulness measures whether learned bottleneck representations align with human-understandable semantics, serving as a prerequisite for reliable explanation and effective intervention in practice. Evaluation strategies have evolved from direct label matching to more rigorous assessment of intrinsic structure and external grounding.

Label Alignment. A common approach evaluates agreement between predicted concepts and ground-truth annotations using standard metrics such as Accuracy and F1 Score. However, in high-dimensional and sparse concept spaces, these metrics can be dominated by true negatives. To mitigate this bias, recent work emphasizes active concept evaluation using metrics such as Jaccard Similarity [Panousis *et al.*, 2024] and the Concept Existence Metric (CEM) [AYSEL *et al.*, 2025], which focus on positively activated concepts.

Intrinsic Representation Quality. Beyond label matching, valid concept representations should possess coherent topological structures and distinct separability. The Concept Alignment Score (CAS) [Espinosa Zarlenga *et al.*, 2022] addresses the former in embedding-based models, verifying whether local neighborhoods in the high-dimensional space maintain semantic homogeneity. Complementarily, to address the latter in scalar-based bottlenecks, Gap [Midavaine *et al.*, 2024; Park *et al.*, 2025] evaluates the latent space separability by measuring the divergence between positive and negative activation distributions, thereby quantifying the intrinsic discriminative quality of the representation.

Semantic and Visual Grounding. To detect spurious correlations, further evaluation aligns learned concepts with external semantic or visual references. Semantic grounding metrics, including CGIM [AYSEL *et al.*, 2025], Similarity [Shang *et al.*, 2024], and Concept Purity [Liu *et al.*, 2025], assess consistency with human-defined importance or canonical embeddings. Visual grounding metrics such as the Concept Trustworthiness Score [Huang *et al.*, 2024] and CLM [AYSEL *et al.*, 2025] quantify spatial alignment between concept activations and annotated regions. While effective in practice, these approaches often require costly annotations; VLM-based alternatives such as Concept-Image Relevance [Liu *et al.*, 2025] reduce supervision but introduce potential hallucination risks.

6.2 Concept Sparsity

Concept sparsity evaluates whether CBMs rely on a minimal and cognitively manageable set of concepts. Excessively dense bottlenecks undermine interpretability and increase the risk of information leakage. Metrics such as Sparsity [Panousis *et al.*, 2024] and the Number of Effective Concepts (NEC) [Srivastava *et al.*, 2024] quantify activation density and reasoning complexity. However, sparsity must be considered jointly with task performance. To capture this trade-off, composite metrics such as Concept Utilization Efficiency (CUE) [Shang *et al.*, 2024] and Concept-Efficient Accuracy (CEA) [Zhao *et al.*, 2026] balance predictive accuracy against concept usage, with CEA offering a bounded, text-invariant formulation for robust comparison.

6.3 Intervenability

Intervenability reflects the ability of CBMs to support effective human correction at inference time. Evaluation protocols in this category assess both the responsiveness of model predictions to concept-level corrections and the efficiency of intervention strategies. Test-Time Intervention (TTI) curves [Koh *et al.*, 2020] and their area-under-curve variants [Espinosa Zarlenga *et al.*, 2023] are widely used to summarize performance gains under increasing intervention budgets. Extending this analysis, recent work examines correction propagation among correlated concepts [Vandenhirtz *et al.*, 2024] and explicitly models the cost of intervention through empirical user time [Midavaine *et al.*, 2024] or theoretical complexity [Shin *et al.*, 2023; Pugnana *et al.*, 2025].

Summary. Evaluating CBMs requires a holistic protocol beyond downstream accuracy. Concept faithfulness ensures semantic validity, concept sparsity maintains cognitive tractability, and intervenability quantifies the effectiveness and cost of human-in-the-loop correction. Together, these dimensions form a principled basis for assessing the interpretability, reliability, and practical utility of CBMs.

7 Benchmark

Table 2 summarizes representative benchmarks used to evaluate Concept Bottleneck Models (CBMs) across general and medical domains. These datasets differ substantially in concept annotation source, supervision granularity, and support for concept-level intervention, which in turn determines which aspects of CBMs, such as faithfulness, scalability, or intervenability, can be reliably assessed. Unlike deep models, CBMs depend on benchmark design to validate concept faithfulness and intervention effectiveness.

Domain	Dataset	#Classes	#Concepts	#Samples	Task	Concept Annotation	Concept Type	Annotation Granularity	Intervention Support	Dataset Link
General	CUB-200-2011	200	312	11788	Fine-grained Cls.	Human-Annotated	Binary	Instance-level	✓	Link
	CelebA	10k	40	200000	Cls.	Human-Annotated	Binary	Instance-level	✓	Link
	Places365	365	102	∼1.8M	Scene Rec.	Weakly Supervised	Binary	Instance-level	×	Link
	AWA2	50	85	37322	Cls./Zero-shot Learning.	Human-Annotated	Binary	Class-level	✓	Link
	RIVAL10	10	18	26484	Cls.	Human-Annotated	Binary	Instance-level	✓	Link
	SUN Attribute	717	102	14340	Scene Cls.	Human-Annotated	Binary	Instance-level	✓	Link
	CEBaB	5	4	5113	Sentiment Cls.	Human-Annotated	Multi-class	Instance-level	✓	Link
	IMDB-C	2	4	200	Sentiment Cls.	Human-Annotated	Multi-class	Instance-level	✓	Link
	CIFAR-10/100	10/100	–	60000	Cls.	LLM/VLM-Generated	–	–	×	Link
	COCO-Stuff	171	–	164000	Scene Cls.	LLM/VLM-Generated	–	–	×	Link
	ImageNet	1000	–	∼1.2M	Cls.	LLM/VLM-Generated	–	–	×	Link
	DTD	47	–	5640	Texture Cls.	LLM/VLM-Generated	–	–	×	Link
	UCF101	101	–	13320	Action Cls.	LLM/VLM-Generated	–	–	×	Link
	Resisc45	45	–	31500	Scene Cls.	LLM/VLM-Generated	–	–	×	Link
	Food-101	101	–	101000	Fine-grained Cls.	LLM/VLM-Generated	–	–	×	Link
	Flower102	102	–	8189	Fine-grained Cls.	LLM/VLM-Generated	–	–	×	Link
	AGNews	4	–	127600	Text Cls.	LLM-Generated	–	–	×	Link
Medical	Derm7pt	5	7	1011	Skin Lesion Cls.	Expert-Annotated	Multi-class	Instance-level	✓	Link
	OAI	5	10	36369	Kellgren-Lawrence Grading	Expert-Annotated	Ordinal	Instance-level	✓	Link
	Fitzpatrick17k	114	48	16577	Skin Disease Cls.	From SkinCon	Binary	Instance-level	✓	Link
	SkinCon	4	48	3230	Skin Disease Cls.	Expert-Annotated	Binary	Instance-level	✓	Link
	WBCAtt	5	31	10298	Fine-grained WBC Cls.	Expert-Annotated	Continuous	Instance-level	✓	Link
	VinDr-CXR	6	22	18000	Thoracic Disease Cls.	Expert-Annotated	Binary	Instance-level	✓	Link
	HAM10000	7	–	10015	Skin Lesion Cls.	LLM/VLM-Generated	–	–	×	Link
	MIMIC-III EWS	16	–	796250	EWS Prediction	Rule-based	–	–	✓	Link
	NCT-CRC-HE	9	–	100k	Tissue Cls.	LLM/VLM-Generated	–	–	×	Link
	IDRiD	5	–	455	Diabetic Retinopathy Grading	LLM/VLM-Generated	–	–	×	Link
	BUSI	3	–	780	Breast Cancer Cls.	LLM/VLM-Generated	–	–	×	Link
	CMMD	2	–	1775	Breast Cancer Cls.	LLM/VLM-Generated	–	–	×	Link
	TBX11K	4	–	11200	Tuberculosis Cls.	LLM/VLM-Generated	–	–	×	Link
	CTI	3	–	750	Choroid Neoplasia Cls.	LLM-Generated	–	–	×	Link
	ISIC2018	7	–	10015	Skin Lesion Cls.	LLM/VLM-Generated	–	–	×	Link
	MiniDDSM	3	–	9684	Breast Cancer Cls.	LLM/VLM-Generated	–	–	×	Link

Table 2: Representative benchmarks for evaluating Concept Bottleneck Models (CBMs) across general and medical domains.

General Domain Datasets. General-domain benchmarks evaluate CBMs on visual and textual recognition tasks, including fine-grained classification, zero-shot learning, and NLP applications. A key distinction lies in concept supervision. Datasets with human-annotated concepts, such as CUB-200-2011 and AWA2, support concept evaluation and test-time intervention, enabling causal analysis of concept corrections. In contrast, large-scale benchmarks relying on LLM- or VLM-generated concepts assess scalability and open-vocabulary capability, but lack verified ground truth and therefore cannot support intervention-based evaluation.

Medical Domain Datasets. Medical benchmarks emphasize clinically meaningful concepts and safety-critical interpretability, making them particularly suitable for evaluating concept faithfulness and intervention effectiveness. Datasets with expert-annotated, ordinal, or continuous concepts explicitly support intervention experiments and enable validation of human-in-the-loop correction in realistic clinical settings. Larger medical datasets that rely on automatically generated concepts improve coverage and scalability, but the absence of validated concept annotations generally limits their suitability for rigorous intervention and causal analysis.

8 Conclusion and Future Directions

Concept Bottleneck Models provide a principled framework for interpretable and controllable decision making by grounding predictions in human-understandable concepts. This survey unified CBM research across the full pipeline of concept acquisition, decision making, intervention, and evaluation, with a particular emphasis on high-stakes medical applications. Despite substantial progress, several fundamental challenges remain, including balancing predictive performance with interpretability, ensuring the semantic faithfulness of learned concepts, enabling reliable and scalable interventions, and maintaining robustness under data scarcity and distribution shift in real-world practice.

Future research directions of CBMs include: **(1) Adaptive Concept Acquisition.** Future CBMs should move beyond fixed bottleneck designs toward adaptive and structured concept representations, such as hierarchical, compositional, or sparsely activated bottlenecks. Such designs better reflect the multi-level and task-dependent nature of human semantics, while improving efficiency. **(2) Faithful and Reliable Concepts.** Ensuring that learned concepts faithfully correspond to their intended semantics remains a central challenge, particularly under weak supervision or when concepts are derived from LLMs or VLMs. Promising directions include uncertainty-aware modeling, causal representation learning to mitigate spurious correlations, and mechanisms for detecting or correcting unstable concepts. **(3) Structured Concept Intervention.** Beyond direct concept replacement, future work should explore structured intervention strategies that account for dependencies among concepts and identify decision-critical subsets. Supporting partial or soft interventions is essential for reducing effort and enabling scalable expert-in-the-loop deployment in safety-critical domains. **(4) Robust Concept Evaluation.** Robust evaluation remains necessary for reliable deployment. Future research should examine how concept representations and intervention effects generalize across domains and shifts, and how concept-based reasoning supports uncertainty estimation, out-of-distribution detection, and intervention-aware evaluation protocols.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (62271465, 62502490), the National Key R&D Program of China (2025YFC3408300), the Suzhou Basic Research Program (SYG202338), the Natural Science Foundation of Jiangsu Province (BK20250496), the China Postdoctoral Science Foundation (2024M763178), and the Jiangsu Funding Program for Excellent Postdoctoral Talent.

Contribution Statement

C. Wang, F. Li, and W. Hu contributed equally to this work.

References

[AYSEL *et al.*, 2025] Halil Ibrahim AYSEL, Xiaohao Cai, et al. Concept-based explainable artificial intelligence: Metrics and benchmarks. In *Available at SSRN 5210908*, 2025.

[Benou and Raviv, 2025] Itay Benou and Tammy Riklin Raviv. Show and tell: Visually explainable deep neural nets via spatially-aware concept bottleneck models. In *CVPR*, 2025.

[Bie *et al.*, 2024] Yequan Bie, Luyang Luo, et al. Xcoop: Explainable prompt learning for computer-aided diagnosis via concept-guided context optimization. In *MICCAI*, 2024.

[Choukali *et al.*, 2024] Mohammad Amin Choukali, Mehdi Chehel Amirani, et al. Pseudo-class part prototype networks for interpretable breast cancer classification. In *Sci. Rep*, 2024.

[Chowdhury *et al.*, 2024] Townim F Chowdhury, Vu Minh Hieu Phan, et al. Adacbm: An adaptive concept bottleneck model for explainable and accurate diagnosis. In *MICCAI*, 2024.

[De Felice *et al.*, 2025] Giovanni De Felice, Arianna Casanova, et al. Causally reliable concept bottleneck models. In *ICLR 2025 Workshop: XAI4Science: From Understanding Model Behavior to Discovering New Scientific Knowledge*, 2025.

[Du *et al.*, 2025] Xingbo Du, Qiantong Dou, et al. Flexible concept bottleneck model. In *AAAI*, 2025.

[Espinosa Zarlenga *et al.*, 2022] Mateo Espinosa Zarlenga, Pietro Barbiero, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. In *NeurIPS*, 2022.

[Espinosa Zarlenga *et al.*, 2023] Mateo Espinosa Zarlenga, Katie Collins, et al. Learning to receive help: Intervention-aware concept embedding models. In *NeurIPS*, 2023.

[Gao *et al.*, 2024a] Yibo Gao, Zheyao Gao, et al. Evidential concept embedding models: Towards reliable concept explanations for skin disease diagnosis. In *MICCAI*, 2024.

[Gao *et al.*, 2024b] Yunhe Gao, Difei Gu, et al. Aligning human knowledge with visual concepts towards explainable medical image classification. In *MICCAI*, 2024.

[Gao *et al.*, 2025] Yibo Gao, Hangqi Zhou, et al. Learning concept-driven logical rules for interpretable and generalizable medical image classification. In *MICCAI*, 2025.

[Grobrügge *et al.*, 2025] Arne Grobrügge, Niklas Kühl, et al. Towards human-understandable multi-dimensional concept discovery. In *CVPR*, 2025.

[Havasi *et al.*, 2022] Marton Havasi, Sonali Parbhoo, et al. Addressing leakage in concept bottleneck models. In *NeurIPS*, 2022.

[He *et al.*, 2025a] Hangzhou He, Jiachen Tang, et al. Training-free test-time improvement for explainable medical image classification. In *MICCAI*, 2025.

[He *et al.*, 2025b] Hangzhou He, Lei Zhu, et al. Chatcbm: Towards interactive concept bottleneck models with frozen large language models. In *arXiv preprint arXiv:2509.17522*, 2025.

[He *et al.*, 2025c] Hangzhou He, Lei Zhu, et al. V2c-cbm: Building concept bottlenecks with vision-to-concept tokenizer. In *AAAI*, 2025.

[Hou *et al.*, 2024] Junlin Hou, Sicen Liu, et al. Self-explainable ai for medical image analysis: A survey and new outlooks. In *arXiv preprint arXiv:2410.02331*, 2024.

[Hu *et al.*, 2024] Lijie Hu, Songning Lai, et al. Towards multi-dimensional explanation alignment for medical classification. In *NeurIPS*, 2024.

[Hu *et al.*, 2025] Lijie Hu, Chenyang Ren, et al. Editable concept bottleneck models. In *ICML*, 2025.

[Huang *et al.*, 2024] Qihan Huang, Jie Song, et al. On the concept trustworthiness in concept bottleneck models. In *AAAI*, 2024.

[Huy *et al.*, 2025] Ta Duc Huy, Sen Kim Tran, et al. Interactive medical image analysis with concept-based similarity reasoning. In *CVPR*, 2025.

[Ismail *et al.*, 2024] Aya Abdelsalam Ismail, Julius Adebayo, et al. Concept bottleneck generative models. In *ICLR*, 2024.

[Jeon *et al.*, 2024a] Sujin Jeon, Inwoo Hwang, et al. Locality-aware concept bottleneck model. In *UniReps*, 2024.

[Jeon *et al.*, 2024b] Sujin Jeon, Inwoo Hwang, et al. Locality-aware concept bottleneck model. In *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024.

[Kazmierczak *et al.*, 2024] Rémi Kazmierczak, Eloïse Berthier, et al. Clip-qda: An explainable concept bottleneck model. In *TMLR*, 2024.

[Kim *et al.*, 2023] Eunji Kim, Dahuin Jung, et al. Probabilistic concept bottleneck models. In *ICML*, 2023.

[Koh *et al.*, 2020] Pang Wei Koh, Thao Nguyen, et al. Concept bottleneck models. In *ICML*, 2020.

[Lai *et al.*, 2023] Songning Lai, Lijie Hu, et al. Faithful vision-language interpretation via concept bottleneck models. In *ICLR*, 2023.

[Lin *et al.*, 2025] Jiakai Lin, Jinchang Zhang, et al. Graph integrated multimodal concept bottleneck model. In *arXiv preprint arXiv:2510.00701*, 2025.

[Liu *et al.*, 2025] Yang Liu, Tianwei Zhang, et al. Hybrid concept bottleneck models. In *CVPR*, 2025.

[Mahinpei *et al.*, 2021] Anita Mahinpei, Justin Clark, et al. Promises and pitfalls of black-box concept learning models. In *arXiv preprint arXiv:2106.13314*, 2021.

[Margeloiu *et al.*, 2021] Andrei Margeloiu, Matthew Ashman, et al. Do concept bottleneck models learn as intended? In *ICLR Workshop on Responsible AI*, 2021.

[Midavaine *et al.*, 2024] Nesta Midavaine, Gregory Hok Tjoan Go, et al. [re] on the reproducibility of post-hoc concept bottleneck models. In *TMLR*, 2024.

[Mpinda *et al.*, 2026] Berthine Nyunga Mpinda, Mehran Hosseinzadeh, et al. Towards multi-label concept bottleneck models in medical imaging: An exploratory survey. In *Medical Imaging with Deep Learning-Validation Papers*, 2026.

[Oikarinen *et al.*, 2023] Tuomas Oikarinen, Subhro Das, et al. Label-free concept bottleneck models. In *ICLR*, 2023.

[Panousis *et al.*, 2024] Konstantinos P Panousis, Dino Ienco, et al. Coarse-to-fine concept bottleneck models. In *NeurIPS*, 2024.

[Park *et al.*, 2025] Seonghwan Park, Jueun Mun, et al. An analysis of concept bottleneck models: Measuring, understanding, and mitigating the impact of noisy annotations. In *NeurIPS*, 2025.

[Penaloza *et al.*, 2025] Emiliano Penaloza, Tianyue H Zhang, et al. Addressing concept mislabeling in concept bottleneck models through preference optimization. In *ICML*, 2025.

[Pugnana *et al.*, 2025] Andrea Pugnana, Riccardo Massidda, et al. Deferring concept bottleneck models: Learning to defer interventions to inaccurate experts. In *NeurIPS*, 2025.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

[Rao *et al.*, 2024] Sukrut Rao, Sweta Mahajan, et al. Discover-then-name: Task-agnostic concept bottlenecks via automated concept discovery. In *ECCV*, 2024.

[Schrodi *et al.*, 2025] Simon Schrodi, Julian Schur, et al. Selective concept bottlenecks without predefined concepts. In *TMLR*, 2025.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, et al. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *ICCV*, 2017.

[Semenov *et al.*, 2024] Andrei Semenov, Vladimir Ivanov, et al. Sparse concept bottleneck models: Gumbel tricks in contrastive learning. In *arXiv preprint arXiv:2404.03323*, 2024.

[Shang *et al.*, 2024] Chenming Shang, Shiji Zhou, et al. Incremental residual concept bottleneck models. In *CVPR*, 2024.

[Shen *et al.*, 2025] Matthew Shen, Aliyah R Hsu, et al. Adaptive test-time intervention for concept bottleneck models. In *ICLR 2025 Workshop BuildingTrust*, 2025.

[Shin *et al.*, 2023] Sungbin Shin, Yohan Jo, et al. A closer look at the intervention procedure of concept bottleneck models. In *ICML*, 2023.

[Singhi *et al.*, 2024] Nishad Singhi, Jae Myung Kim, et al. Improving intervention efficacy via concept realignment in concept bottleneck models. In *ECCV*, 2024.

[Srivastava *et al.*, 2024] Divyansh Srivastava, Ge Yan, et al. Vlg-cbm: Training concept bottleneck models with vision-language guidance. In *NeurIPS*, 2024.

[Steinmann *et al.*, 2024] David Steinmann, Wolfgang Stammer, et al. Learning to intervene on concept bottlenecks. In *ICML*, 2024.

[Tan *et al.*, 2024a] Andong Tan, ZHOU Fengtao, et al. Post-hoc part-prototype networks. In *ICML*, 2024.

[Tan *et al.*, 2024b] Andong Tan, Fengtao Zhou, et al. Explain via any concept: Concept bottleneck model with open vocabulary concepts. In *ECCV*, 2024.

[Vandenhirtz *et al.*, 2024] Moritz Vandenhirtz, Sonia Laguna, et al. Stochastic concept bottleneck models. In *NeurIPS*, 2024.

[Vemuri *et al.*, 2026] Deepika SN Vemuri, Gautham Bellamkonda, et al. Logiccbms: Logic-enhanced concept-based learning. In *WACV*, 2026.

[Wang *et al.*, 2025] Chunjiang Wang, Kun Zhang, et al. Mvp-cbm: Multi-layer visual preference-enhanced concept bottleneck model for explainable medical image classification. In *IJCAI*, 2025.

[Xu *et al.*, 2024] Xinyue Xu, Yi Qin, et al. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *ICLR*, 2024.

[Xu *et al.*, 2025] Haotian Xu, Tsui-Wei Weng, et al. Graph concept bottleneck models. In *arXiv preprint arXiv:2508.14255*, 2025.

[Yang *et al.*, 2023] Yue Yang, Artemis Panagopoulou, et al. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *CVPR*, 2023.

[Yuksekgonul *et al.*, 2023] Mert Yuksekgonul, Maggie Wang, et al. Post-hoc concept bottleneck models. In *ICLR*, 2023.

[Zhang *et al.*, 2024] Rui Zhang, Xingbo Du, et al. The decoupling concept bottleneck model. In *TPAMI*, 2024.

[Zhao *et al.*, 2026] Delong Zhao, Qiang Huang, et al. Partially shared concept bottleneck models. In *AAAI*, 2026.