

ConceptLens: Interactive Concept Bottlenecks for Black-Box Models

Hasan Md Tusfiqur Alam

Interactive Machine Learning Department
German Research Center for Artificial Intelligence (DFKI GmbH)
Saarbrücken, Germany
hasan.alam@dfki.de

Abdulrahman Mohamed Selim

Interactive Machine Learning Department (IML)
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
University of Oldenburg
Oldenburg, Germany
abdulrahman.mohamed@dfki.de

László Kopácsi

Interactive Machine Learning
German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
laszlo.kopacsi@dfki.de

Daniel Sonntag

German Research Center for Artificial Intelligence (DFKI)
Saarbrücken, Germany
University of Oldenburg
Oldenburg, Germany
daniel.sonntag@dfki.de

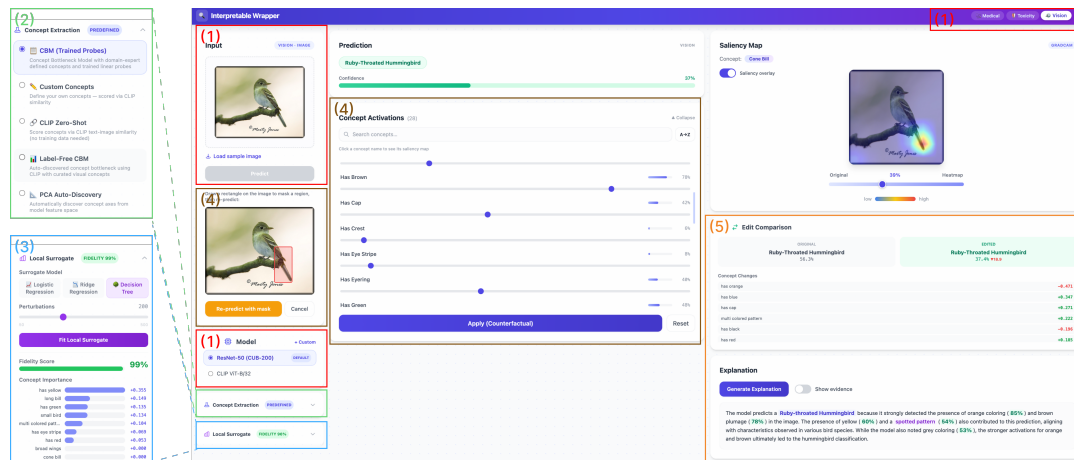


Figure 1: Overview of the demonstration workflow: (1) select task, input, and black-box model; (2) construct a concept representation using a chosen strategy; (3) fit a local surrogate in concept space and report fidelity; (4) interact via concept interventions and counterfactuals; (5) generate evidence-grounded explanations via RAG.

Abstract

Most interpretability tools explain model decisions post-hoc but offer limited support for user action. We present CONCEPTLENS, a model-agnostic framework that transforms explanations into an interactive interface for probing pre-trained classifiers across domains and modalities. By extracting concept vectors and fitting local approximations, the system enables users to evaluate local fidelity, perform what-if analysis, conduct counterfactual searches, and generate evidence-grounded explanations. Unlike traditional

feature-level tools, CONCEPTLENS shifts interaction to semantic concepts rather than isolated pixels or tokens. We demonstrate the system across three tasks spanning different domains and modalities: chest X-ray classification, bird species recognition, and toxic comment detection. The framework enables users to inspect black-box decisions and also to actively refine the underlying concept probes through interactive feedback. [Code available on GitHub](#).

CCS Concepts

• Computing methodologies → Causal reasoning and diagnostics; • Human-centered computing → Visualization toolkits.

ACM Reference Format:

Hasan Md Tusfiqur Alam, Abdulrahman Mohamed Selim, László Kopácsi, and Daniel Sonntag. 2026. ConceptLens: Interactive Concept Bottlenecks for Black-Box Models. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '26)*, October 05–09, 2026, Napoli, Italy. ACM, New York, NY, USA, 3 pages. <https://doi.org/10.1145/3776574.3832528>



This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License.

ICMI '26, Napoli, Italy

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2318-6/26/10

<https://doi.org/10.1145/3776574.3832528>

1 Introduction

Deep neural networks are highly optimised for predictive performance, but remain hard to interpret in deployment; this limits effective human-AI collaboration, which relies on understanding model decisions. Existing explanation methods address this in part, e.g., post-hoc explainers approximate local model behaviour (e.g., LIME, SHAP), while saliency-based methods highlight predictive visual regions (e.g., Grad-CAM) [3]. However, they operate at low-level pixel or token layers and provide limited support for structured semantic explanations or interaction with the model’s reasoning. Recent studies further argue that interpretability must be *actionable*, enabling concrete decisions and interventions through both *concreteness* and *validation* [12]. Overall, these limitations motivate interactive systems in which users do not just read explanations but actively manipulate them to verify their effects.

Concept-based interpretability offers a more human-aligned way to explain model behaviour by using higher-level abstractions that better match human reasoning. Concept Bottleneck Models (CBMs) make predictions using concept variables [8], creating a shared layer for inspection and intervention, whereas post-hoc CBMs [16] learn concept bottlenecks over frozen pre-trained representations, making this approach more practical. To make such models accessible in practice, interactive explainability systems support model probing and visual diagnosis [6], while newer tools provide concept-centric interaction (e.g., ConceptExplainer [7]) or domain-specific manual counterfactual editing (e.g., CLARUS [9]). However, they still lack a unified, model-agnostic post-hoc framework that combines visual evidence, textual grounding, concept interventions, local concept-space surrogates with fidelity reporting, aligned natural-language explanations, and conversational follow-up in a single workflow.

In this work, we present CONCEPTLENS, a post-hoc interactive framework, which builds a concept bottleneck interface around pre-trained black-box models. It integrates multiple concept construction strategies, including Vision Language Model (VLM)-based zero-shot scoring and label-free concepts [11], and learns interpretable local approximations in concept space with explicit fidelity reporting. This concept layer acts as an editable interface that lets users inspect concept evidence, perform what-if analysis, generate counterfactuals [10], and receive evidence-grounded natural-language explanations via retrieval-augmented generation (RAG) [2], with conversational follow-up. CONCEPTLENS supports actionable interpretability [12] by enabling concrete concept-level interventions and input edits, showing how faithfully the local approximation reflects the black-box model, and capturing user feedback for later concept-probe refinement ¹.

2 ConceptLens Overview

Let f be a pre-trained classifier and x an input instance. CONCEPTLENS builds a concept representation $c(x) \in [0, 1]^K$ using a selected concept construction strategy, and fits a local interpretable model h such that $h(c(x)) \approx f(x)$ (Figure 2(a)). This creates an interactive, model-agnostic bottleneck around f without changing its weights, enabling users to explore model behaviour across both input and concept spaces. The framework fits h to local perturbations and reports fidelity to ensure local alignment. The initial label

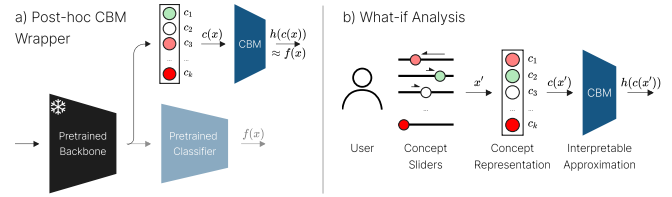


Figure 2: CONCEPTLENS Overview: (a) a post-hoc CBM wraps f , yielding $h(c(x)) \approx f(x)$; (b) users adjust concept sliders to inspect and intervene on model behaviour.

and confidence always come from the fixed classifier f , while h is used to estimate the effects of user interventions (Figure 2(b)) so that the original decision and the response to edits can be compared side by side.

2.1 Concept Extraction

We implemented several semi-supervised and unsupervised concept extraction strategies, allowing users to balance supervision, flexibility, and automation. **(1) Trained probes (post-hoc CBM).** We train lightweight probes (e.g., logistic regression) on features $\phi(x)$ from an internal layer of f , yielding concept activations $c_k(x) = \sigma(w_k^T \phi(x) + b_k)$. **(2) Zero shot concepts.** When CLIP embeddings are available, concepts are scored zero-shot via image-text cosine similarity using a curated label-free vocabulary [11], eliminating the need for labelled training data. **(3) Unsupervised discovery.** Applying PCA or K-means clustering over a feature bank identifies emergent concept directions or clusters, which are named with CLIP text prompts when available, following ACE and TCAV [5]. Unlike recent segmentation-based concepts [13], this strategy derives concepts directly within the black-box feature space. **(4) Text concepts.** For text models, we aggregate attention-based token saliency [14] into semantic categories (e.g., *insult*, *threat*), enabling users to reason over linguistic phenomena rather than single tokens.

2.2 Local Concept Approximation

We fit an interpretable model h in concept space to approximate f locally. Given perturbations around x (e.g., using input modifications or direct concept edits), we obtain training pairs $\{(c(x_i), f(x_i))\}$ and fit h using logistic regression, a ridge classifier, or a shallow decision tree. This model clarifies how changes in concepts affect the black-box prediction near x , enabling what-if analysis and counterfactual search in concept space.

We measure the agreement between the interpretable model and the black box in the sampled neighbourhood using a local fidelity score, $\text{fidelity}(x) = \frac{1}{N} \sum_{i=1}^N \mathbb{I}[h(c(x_i)) = f(x_i)]$. High fidelity indicates that h is a reliable local proxy for f under the chosen perturbation scheme, while low fidelity suggests insufficient concept coverage or non-linear interactions that the selected model family cannot capture.

3 Interactive Demonstration Workflow

The demo presents an interactive pipeline which enables users to inspect and interact with models based on the extracted concepts

¹ See Demo: <https://youtu.be/XLxfSSFhgug>

(Figure 1). In all scenarios, the external concept bottleneck c and its local approximation h are constructed and used for interventions.

3.1 Use Cases

CONCEPTLENS is designed as a general-purpose interface that can be adapted to different modalities, domains, and pre-trained deep learning models. We demonstrate its flexibility using three classification tasks: (1) **Medical imaging**: Chest X-ray classification on the COVID-QU dataset [4] using radiological concepts and Grad-CAM evidence maps; (2) **Animal identification**: Bird species classification on the CUB dataset [15] using visual attributes and concept attribution maps; and (3) **Natural language**: Toxic comment detection using label-free concepts [11]. Across these tasks, users interact through synchronised visual evidence, textual explanations, concept-level interventions, and conversational queries.

3.2 Setup and Interaction

Users configure the interface by selecting a task, input instance, and a pre-trained model (Figure 1, Step 1), with support for custom models from HuggingFace². Selecting a concept extraction strategy (Step 2) populates the interface with concept activations and modality-specific evidence, such as image heatmaps or text highlights. Choosing a local model family then fits h , displaying its weights and local fidelity score (Step 3).

Once configured, users perform **what-if analysis** by adjusting concept sliders to observe how edits affect the local model prediction in real time. The system also supports counterfactual search (Step 4) to find the minimal set of concept edits needed to flip a prediction and allows direct input modifications, such as masking image regions or altering text. For validation, the interface uses RAG [1] to generate natural-language explanations from retrieved domain documents (Step 5), and integrates a chat agent for follow-up conversations. Users can then refine concept probes and validate the impact using the updated fidelity scores.

4 Implementation

We used a Python backend with a React frontend dashboard. Concept probes and local approximations rely on lightweight classifiers to minimise latency, while explanations are generated via retrieval over an indexed corpus, optionally augmented by an LLM.

5 Conclusion, Limitations, and Future Work

We presented CONCEPTLENS, a model-agnostic framework that provides a concept interface adaptable across various modalities and domains. By extracting concept vectors and fitting local approximations, the system supports inspection, what-if analysis, counterfactual search, and evidence-grounded explanations. **Limitations.** Because the framework is post-hoc, its effectiveness depends on the coverage and quality of the extracted concepts. Since the interpretable model h is trained locally, its alignment with the black box holds near the sampled neighbourhood rather than globally. Furthermore, post-hoc concept attributions provide correlational evidence and should not be treated as definitive causal explanations.

Future work. We plan to explore richer text concepts, such as phrase-level taxonomies, alongside segmentation-aware visual concepts to improve explanation coherence. We also intend to extend interactive concept-probe refinement, allowing users to incrementally retrain and improve the underlying probes across repeated sessions. Finally, we aim to conduct a user study to evaluate the utility and deployment of CONCEPTLENS in practice.

Acknowledgement This work was funded by the Federal Ministry of Research, Technology and Space (BMFTR) under grant numbers 16IW23002 (No-IDLE), 16IW24006 (NoIDLEChatGPT), and the Lower Saxony Ministry of Science and Culture (MWK) in the zukunfft.niedersachsen program; and the Endowed Chair of AAI at the University of Oldenburg.

References

- [1] Hasan Md Tufiqur Alam et al. 2025. CBM-RAG: Demonstrating Enhanced Interpretability in Radiology Report Generation with Multi-Agent RAG and Concept Bottleneck Models. In *Companion Proceedings of EICS '25*. Association for Computing Machinery, New York, NY, USA. doi:10.1145/3731406.3731970
- [2] Hasan Md Tufiqur Alam et al. 2025. Towards Interpretable Radiology Report Generation via Concept Bottlenecks Using a Multi-agent RAG. In *Advances in Information Retrieval*. Springer Nature Switzerland, Cham, 201–209. doi:10.1007/978-3-031-88714-7_18
- [3] Alejandro Barredo Arrieta et al. 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* 58 (June 2020), 82–115. doi:10.1016/j.inffus.2019.12.012
- [4] Muhammad E. H. Chowdhury et al. 2020. Can AI Help in Screening Viral and COVID-19 Pneumonia? *IEEE Access* 8 (2020), 132665–132676. doi:10.1109/ACCESS.2020.3010287
- [5] Amirata Ghorbani et al. 2019. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems*, Vol. 32. Curran Associates, Inc., Red Hook, NY, USA. https://proceedings.neurips.cc/paper_files/paper/2019/hash/77d2afcb31f6493e350fca61764efb9a-Abstract.html
- [6] Fred Hohman et al. 2019. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Transactions on Visualization and Computer Graphics* 25, 8 (Aug. 2019), 2674–2693. doi:10.1109/TVCG.2018.2843369
- [7] Jinbin Huang et al. 2023. ConceptExplainer: Interactive Explanation for Deep Neural Networks from a Concept Perspective. *IEEE Transactions on Visualization and Computer Graphics* 29, 1 (2023), 831–841. doi:10.1109/TVCG.2022.3209384
- [8] Pang Wei Koh et al. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning*, Vol. 119. PMLR, 5338–5348. <https://proceedings.mlr.press/v119/koh20a.html>
- [9] Jacqueline Michelle Metsch et al. 2024. CLARUS: An Interactive Explainable AI Platform for Manual Counterfactuals in Graph Neural Networks. *Journal of Biomedical Informatics* 150 (2024), 104600. doi:10.1016/j.jbi.2024.104600
- [10] Ramaravind K. Mothilal et al. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 607–617. doi:10.1145/3351095.3372850
- [11] Tuomas Oikarinen et al. 2023. Label-Free Concept Bottleneck Models. In *International Conference on Learning Representations (ICLR)*. <https://openreview.net/forum?id=FLCg47MNvBA>
- [12] Hadas Orgad et al. 2026. Position: Interpretability Can Be Actionable. In *Mechanistic Interpretability Workshop at ICML 2026*. <https://openreview.net/forum?id=Sn6Cjs515>
- [13] Katharina Prasse et al. 2025. DCBM: Data-Efficient Visual Concept Bottleneck Models. In *International Conference on Machine Learning*. <https://openreview.net/forum?id=BdO4R6XxUH>
- [14] Ashish Vaswani et al. 2017. Attention is all you need. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS '17)*. Curran Associates Inc., Red Hook, NY, USA, 6000–6010. <https://dl.acm.org/doi/10.5555/3295222.3295349>
- [15] Catherine Wah et al. 2011. *The Caltech-UCSD Birds-200-2011 Dataset*. Technical Report CNS-TR-2011-001. California Institute of Technology.
- [16] Mert Yuksekgonul et al. 2022. Post-hoc Concept Bottleneck Models. In *ICLR 2022 Workshop on PAIR^2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data*. https://openreview.net/forum?id=HAMEOIRD_g9

²<https://huggingface.co/> (Accessed June 15, 2026)