Neural Concept Verifier: Scaling Prover-Verifier Games via Concept Encodings

Berkant Turan [1] **Suhrab Asadulla** [1] **David Steinmann** [2] **Kristian Kersting** [2 3 4 5] **Wolfgang Stammer** [6] **Sebastian Pokutta** [1 7]

Abstract

While *Prover-Verifier Games* (PVGs) offer a promising path toward verifiability in nonlinear classification models, they have not yet been applied to complex inputs such as high-dimensional images. Conversely, expressive *concept encodings* effectively allow to translate such data into interpretable concepts but are often utilised in the context of low-capacity linear predictors. In this work, we push towards real-world verifiability by combining the strengths of both approaches. We introduce *Neural Concept Verifier (NCV)*, a unified framework combining PVGs for formal verifiability with concept encodings to handle complex, high-dimensional inputs in an interpretable way. NCV achieves this by utilizing recent minimally supervised concept discovery models to extract structured concept encodings from raw inputs. A *prover* then selects a subset of these encodings, which a *verifier*, implemented as a nonlinear predictor, uses exclusively for decision-making. Our evaluations show that NCV outperforms classic concept-based models and pixel-based PVG classifier baselines on high-dimensional, logically complex datasets and helps mitigate shortcut behavior. Overall, we demonstrate NCV as a promising step toward concept-level, verifiable AI.

1. Introduction

Deep learning has achieved remarkable predictive performances, but often at the expense of *interpretability and trustworthiness* (Rudin, 2019). However, particularly in high-stakes applications, it is critical that models provide *verifiable justifications* for their decisions (Irving et al., 2018; Fok & Weld, 2023). *Prover-Verifier Games* (PVGs), introduced by Anil et al. (2021), formalize such justifications via a game-theoretic approach: a prover provides evidence to convince a verifier, who accepts only verifiable proofs. A prominent instantiation of PVGs is the *Merlin-Arthur Classifier* (Wäldchen et al., 2024), a classifier guided by cooperative and adversarial provers, offering formal interpretability guarantees through information-theoretic bounds. Specifically, identified features are provably relevant for a given output, rather than mere correlations as in typical feature explanation methods (Adebayo et al., 2018; Zhou et al., 2022). However, the Merlin-Arthur framework faces significant scalability challenges when applied to high-dimensional real-world data, as explanations based on raw pixels are both computationally difficult to optimize and offer limited human understandability (*cf.* Fig. 1; Wäldchen et al. (2024)).

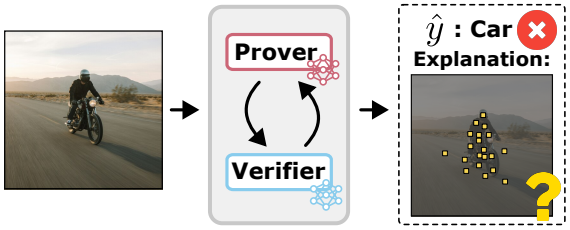


Figure 1. Challenges of Prover-Verifier Games (PVGs) in image classification: (i) It is non-trivial to scale up for high-dimensional data. (ii) Furthermore, the learned explanation masks on the pixel level remain difficult for humans to understand.

Concurrently, *Concept Bottleneck Models* (CBMs) have emerged as a powerful framework for interpretable machine learning, structuring predictions through intermediate interpretable concept encodings (Koh et al., 2020; Stammer et al., 2021). Despite their advantages, CBMs typically employ *linear classifiers* on top of concept encoding layers, thereby potentially restricting expressivity and failing on tasks requiring nonlinear interactions among concepts (e.g., XOR problems, counting or permutation invariance, *cf.* Suppl. B) (Mahinpei et al., 2021; Kimura et al., 2024; Lee et al., 2019).

In this work, we combine the best of both worlds by in-

[1]Zuse Institute Berlin, Germany [2]AI & ML Lab, Computer Science Department, TU Darmstadt [3]Hessian Center for AI (hessian.AI) [4]Lab1141 [5]German Research Center for AI (DFKI) [6]Max Planck Institute for Informatics, SIC [7]Institute of Mathematics, Technische Universität Berlin, Germany. Correspondence to: Berkant Turan <turan@zib.de>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

troducing the **Neural Concept Verifier (NCV)**, a novel framework integrating concept-based representations into PVGs in the form of Merlin-Arthur Classifiers. NCV shifts the prover–verifier interaction from the image level to a structured, symbolic concept level, overcoming both the scalability limitations encountered by Merlin-Arthur Classifiers in high-dimensional settings and the expressivity constraints inherent to linear CBMs. We demonstrate through extensive evaluations that NCV successfully scales PVGs to complex, high-dimensional classification tasks. At the same time, NCV enables verifiable, performant nonlinear classifiers on top of concept extractors, effectively narrowing the interpretability–accuracy gap in linear CBMs. Lastly, our framework improves robustness to shortcut learning, thereby enhancing the generalizability and trustworthiness of predictions, particularly in high-stakes applications.

Our summarized contributions are: (i) We propose *Neural Concept Verifier* (NCV), a framework combining concept-based models with Prover–Verifier Games (PVGs). (ii) NCV scales PVGs to high-dimensional image data by operating on compact concept encodings. (iii) It enables expressive yet interpretable classification via sparse, nonlinear reasoning over concepts. (iv) We validate NCV on synthetic and real-world benchmarks, demonstrating strong accuracy and verifiability. (v) We highlight that NCV improves generalization under spurious correlations. The remainder of the paper is structured as follows. We begin with a review of related work, followed by an introduction and formal description of NCV. We then provide a comprehensive evaluation and, finally, discuss our findings and conclude the paper.

2. Background

Prover-Verifier Games. PVGs were introduced by Anil et al. (2021) as a game-theoretic framework to encourage learning agents to produce *testable* justifications through interactions between an untrusted prover and a trusted verifier. Their work showed that, under suitable conditions, the verifier can learn robust decision rules even when the prover actively attempts to persuade it of arbitrary outputs. Wäldchen et al. (2024) extended this idea to the *Merlin-Arthur Classifier* (MAC), which provides formal interpretability guarantees by bounding the mutual information between the selected features and the ground-truth label. Recently, Kirchner et al. (2024) applied a PVG-inspired approach to improve legibility of Large Language Model (LLM) outputs and Amit et al. (2024) introduced *self-proving models* that leverage interactive proofs to formally verify the correctness of model outputs. PVG-style setups have also been explored in safety-focused learning protocols (Irving et al., 2018; Brown-Cohen et al., 2024; Głuch et al., 2024). These developments reflect a broader trend of utilising multi-agent learning (Pruthi et al., 2022; Schneider & Vlachos, 2024; Du et al., 2024; Nair et al., 2023; Stammer et al., 2024a). Our work builds on this line of research by embedding PVGs into a concept-based classification framework to address the dimensionality bottleneck that limits their performance in high-dimensional settings.

Concept Representations for Interpretability. The introduction of Concept Bottleneck Models (CBMs) (Koh et al., 2020; Zarlenga et al., 2022) and concept-based explanations (Kim et al., 2018; Crabbé & van der Schaar, 2022; Poeta et al., 2023; Lee et al., 2025) was an important moment in the growing interest in AI interpretability research. The particular appeal of CBMs lies in the promise of interpretable predictions and a controllable, structured interface for human interactions (Stammer et al., 2021). While initial CBMs relied on fully supervised concept annotations, subsequent research has relaxed this requirement by leveraging pretrained vision-language models like CLIP for concept extraction (Bhalla et al., 2024; Yang et al., 2023; Oikarinen et al., 2023; Panousis et al., 2024; Steinmann et al., 2025), or employing fully unsupervised concept discovery methods (Ghorbani et al., 2019; Stammer et al., 2024b; Schut et al., 2025). Post-hoc variants instead attach concept-based components on top of pretrained black-box predictors (Yuksekgonul et al., 2023; Wang et al., 2025), while complementary methods statistically test the importance of semantic concepts for an existing classifier post hoc (Teneggi & Sulam, 2024).

While most works have focused on the bottleneck itself by reducing supervision requirements, mitigating concept leakage, or dynamically expanding the concept space, several recent approaches also enrich the classifier component using nonlinear or symbolic predictors, including concept-based memory reasoning (Debot et al., 2024), neural-symbolic reasoning (Barbiero et al., 2023), and causal concept models (Dominici et al., 2024; De Felice et al., 2025). A complementary information-theoretic perspective is taken by Variational Information Pursuit (Chattopadhyay et al., 2023), which sequentially selects queries to maximise mutual information rather than producing a single sparse selection over a fixed concept vocabulary (see Sec. C.5 for a detailed comparison). NCV is complementary to these methods: rather than proposing a new classifier architecture, it wraps *any* concept-level predictor in a prover–verifier game that enforces sparse, per-sample concept selection and evaluates predictions under competing subsets. In contrast to Sparse Autoencoders (Cunningham et al., 2023), which decompose internal activations of black-box models into interpretable features post hoc, NCV operates over pre-defined, grounded concept vocabularies and enforces sparse selection through adversarial training during learning.

Shortcut Learning. Independent of their interpretability, training deep models can often lead to unwanted artifacts

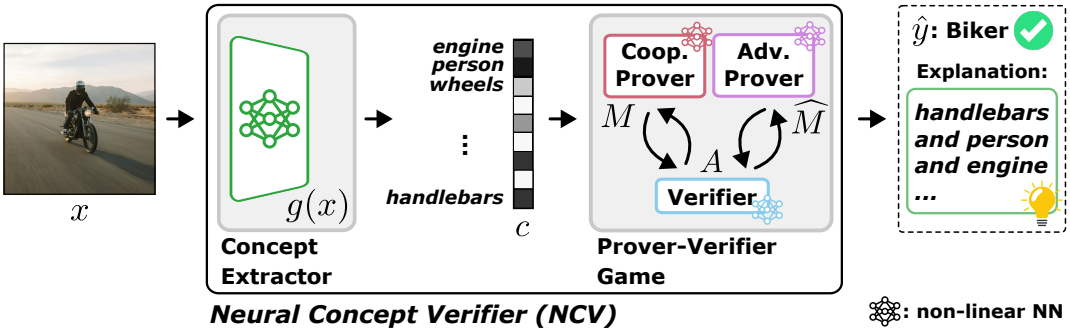


Figure 2. Overview of the Neural Concept Verifier (NCV). The input image is first processed by a concept extractor to produce symbolic concept encodings. A prover–verifier game is then played over these encodings: a cooperative prover selects a sparse concept subset supporting the true class, while an adversarial prover selects misleading concepts. Finally, the nonlinear verifier makes a prediction based only on these selected concepts, ensuring verifiable and robust classification.

and side effects. Shortcut learning describes the problem of models learning to rely on unwanted and unintended features to resolve a task (Geirhos et al., 2020). This is a common problem when training purely deep learning models (Lapuschkin et al., 2019; Schramowski et al., 2020) or neuro-symbolic models (Marconato et al., 2023; Bortolotti et al., 2025), and, if not taken care of, can lead to predictions being right for the wrong reasons (Ross et al., 2017). There have been various approaches to tackle this problem, from careful dataset curation (Ahmed et al., 2021) to modified model training (Friedrich et al., 2023), *cf.* Steinmann et al. (2024) for a comprehensive overview. While NCV is not specifically designed to mitigate shortcuts, we show that our setup can intrinsically mitigate their impact on the data.

3. Neural Concept Verifier (NCV)

In this section, we introduce the *Neural Concept Verifier* (NCV), a framework that combines concept-based representations with the Merlin-Arthur prover-verifier paradigm (Wäldchen et al., 2024). NCV trains nonlinear classifiers whose predictions provably rely on sparse subsets of high-level concepts, with information-theoretic guarantees formalized through completeness and soundness criteria (*cf.* Suppl. A). In contrast to the original Merlin-Arthur classifier (MAC), which operates on raw pixel features and struggles to scale to high-dimensional inputs, NCV performs the prover–verifier interaction directly in concept space. This shift enables optimization on complex datasets and grounds PVG explanations in human-interpretable concepts.

NCV consists of two main components (*cf.* Fig. 2): (i) a weakly supervised *concept extractor* that transforms input data into interpretable concept encodings, and (ii) a nonlinear MAC that selects sparse concepts for the final predictions. After providing background notations and a high-level overview of NCV, we present detailed descriptions of each main component as well as training and inference details.

3.1. Problem Setup and Notation

Let $\mathcal{X} \in \mathbb{R}^{N \times D}$ denote a dataset of N inputs (e.g., images), each of dimension D, and let $\mathcal{Y} \in \{1, \ldots, K\}^N$ be the corresponding class labels for K classes. We assume that the pairs (x, y) are drawn *i.i.d.* from an unknown data distribution $\mathcal{D}$ over $\mathbb{R}^D \times \{1, \ldots, K\}$, and that $(\mathcal{X}, \mathcal{Y})$ correspond to a finite sample from $\mathcal{D}$. For the verifier, we consider an extended prediction space $\{1, \ldots, K, \perp\}$ with an additional *rejection class* $\perp$, allowing the classifier to abstain from a decision when uncertain, which is crucial for enforcing interpretability guarantees in adversarial setups (Wäldchen et al., 2024). Our overall goal is to learn a model $f : \mathbb{R}^D \to \{1, \ldots, K, \perp\}$ that maps an input $x \in \mathbb{R}^D$ to a prediction $\hat{y} \in \{1, \ldots, K, \perp\}$ that is based on a small, interpretable subset of high-level concepts.

3.2. The NCV Framework

NCV decomposes f into the following components (*cf.* Fig. 2 for an illustrative overview):

1. A **concept extractor** $g : \mathcal{X} \to \mathcal{C}$, which maps each input x to a high-level concept encoding $\mathbf{c} \in \mathbb{R}^C$, where $C \in \mathbb{N}$ is the number of discovered concepts.

2. A pair of **provers** $M, \widehat{M} : \mathcal{C} \to \{0, 1\}^C$ that produce sparse binary masks selecting m concepts each. M (Merlin, *cooperative prover*) aims to help classification; $\widehat{M}$ (Morgana, *adversarial prover*) aims to mislead, which is crucial for overall robustness.

3. A nonlinear **verifier** (Arthur) $A : \mathbb{R}^C \to \mathcal{Y}$, which predicts a label based only on the masked concepts.

The three modules are trained jointly, where their interaction encourages the verifier to rely only on robust, informative concept features. Let us now provide details on these individual components.

3.3. Concept Extraction

The concept extractor $g : \mathcal{X} \to \mathbb{R}^C$ transforms raw input data into interpretable, high-level concept representations, where each dimension corresponds to a semantically meaningful concept. The resulting concept encoding $\mathbf{c} \in \mathbb{R}^C$ serves as the input to the PVG.

While conceptually simple, a careful combination of PVGs and concept-based models is necessary. The concept extractor must satisfy three key requirements: interpretability, expressiveness, and modularity. Concepts should correspond to human-understandable features that can serve as meaningful explanations, the concept space should capture sufficient information for the target task without creating information bottlenecks, and the extractor should operate independently of the prover-verifier components. Unlike traditional concept extractor approaches that enforce sparsity constraints directly on $\mathbf{c}$, our framework delegates sparsity to the prover-verifier interaction. This allows the concept space to remain dense and expressive while achieving interpretable sparsity through downstream concept selection. Further, NCV can accommodate different supervision paradigms for concept extraction: supervised methods that leverage predefined concept vocabularies, self-supervised methods that exploit multi-modal correspondences (e.g., vision-language alignment), or unsupervised methods that discover latent conceptual structures can all be used as concept extractors.

Overall, NCV requires that g produces consistent, interpretable encodings while maintaining sufficient information for accurate classification. In our evaluations, we instantiate NCV's concept extractor via the recent unsupervised, object-centric NCB framework (Stammer et al., 2024b) and the multi-modal, CLIP-based SpLiCE (Bhalla et al., 2024). Operating in concept space rather than raw input space provides: (i) scalability through dimensionality reduction and (ii) explanations based on human-interpretable concepts.

3.4. Verifiable Classification via Merlin-Arthur Classifiers

The second core component of NCV is a interpretable classifier of the Merlin-Arthur setup (Wäldchen et al., 2024) originally inspired by Interactive Proof Systems (Goldwasser et al., 1985). This setup generally formalizes the idea of proving that a classification decision is supported by a sparse and informative set of features.

In NCV, specifically, two competing provers, **Merlin** and **Morgana**, select concept subsets either to support or mislead classification, respectively. The verifier, Arthur, then makes predictions based solely on these masked concepts without knowledge of the prover's intent. Formally, each prover outputs a sparse binary mask with m active entries, producing Merlin's cooperative subset $S = M(\mathbf{c}) \odot \mathbf{c}$ and Morgana's adversarial subset $\widehat{S} = \widehat{M}(\mathbf{c}) \odot \mathbf{c}$, where $\mathbf{c}$ is the concept encoding from extractor g, and $\odot$ denotes element-wise masking. Notably, all three modules are differentiable nonlinear models.

This interactive setup enables two key metrics: *completeness* and *soundness*. With Merlin's subset S and Morgana's subset $\widehat{S}$, we define

$$\text{Completeness} = \mathbb{P}_{(x,y)\sim\mathcal{D}}\big[A(S) = y\big], \tag{1}$$

$$\text{Soundness} = \mathbb{P}_{(x,y)\sim\mathcal{D}}\big[A(\widehat{S}) \in \{y, \perp\}\big], \tag{2}$$

where $\perp$ denotes the rejection class. Intuitively, completeness measures how often Arthur can recover the true label from Merlin's sparse, helpful concepts (similar to standard accuracy), while soundness measures how often Arthur can avoid committing to a wrong label under Morgana's misleading subset, either by staying correct or abstaining (*cf.* Suppl. A for the theoretical interpretation).

Furthermore, sparsity has recently become central in concept-based models, as large concept spaces require sparse predictions for interpretability. Prior work (Bhalla et al., 2024; De Santis et al., 2025) achieves this by regularizing the concept space itself, restricting the number of active concepts before training the classifier. In contrast, NCV keeps the concept space fully expressive and enforces sparsity only in the concepts passed to the classifier.

3.5. Information-Theoretic Interpretation

The completeness and soundness defined in Sec. 3.4 are not just empirical metrics; they carry an information-theoretic interpretation grounded in the Merlin-Arthur classifier theory of Wäldchen et al. (2024), instantiated class-wise in concept space. We summarize the key statements here and defer all formal derivations, full precision bounds, and a detailed discussion of limitations to Suppl. A.

Concept-space certificates. Since the prover M from Sec. 3.4 acts on the concept encoding, we slightly abuse notation and write $M(x) := M(g(x)) \in \{0,1\}^C$. Following the partial-input formalism of Wäldchen et al. (2024), we write $M(x') \subseteq x$ for the event that $g(x)_i = g(x')_i$ at every index i with $M(x')_i = 1$, i.e., x and x' agree on every concept that Merlin selects from x'. For a fixed class $k \in \{1, \ldots, K\}$, define the one-vs-rest *labeling function* $Y_k : \{1, \ldots, K\} \to \{-1, +1\}$ by $Y_k(y) := +1$ if $y = k$ and -1 otherwise. We assume the label y is determined by the input x. Imperfect precision then reflects insufficiency of the selected concepts rather than intrinsic label noise.

Class-wise precision and mutual information. For a fixed class k, the *average precision* of M is

$$\overset{(k)}{\Pr_{\mathcal{D}}}(M) := \mathbb{E}_{(x',y')}\big[\mathbb{P}_{(x,y)}\big[Y_k(y) = Y_k(y') \mid M(x') \subseteq x\big]\big].$$

Applying the binary mutual-information bound of Wäldchen et al. (2024) with $Y_k(y)$ as the binary label, we obtain a class-wise lower bound on the mutual information between the class-k label and Merlin's certificate:

$$\mathbb{E}_{(x',y')\sim\mathcal{D}}\big[I_{(x,y)\sim\mathcal{D}}\big(Y_k(y);\, M(x') \subseteq x\big)\big] \geq H_{y\sim\mathcal{D}}(Y_k(y)) - H_b\big(\Pr_{\mathcal{D}}^{(k)}(M)\big), \quad (3)$$

where $H_b : [0,1] \to [0,1]$ is the binary entropy and $M(x') \subseteq x$ in the mutual information denotes the binary containment variable for fixed x'.

Observable completeness and soundness. A single trained triple $(M, \widehat{M}, A)$ induces a pair of class-wise errors for every k:

$$\varepsilon_c^{(k)} := \mathbb{P}_{(x,y)\sim\mathcal{D}}\big[A\big(M(g(x)) \odot g(x)\big) \neq k \mid y = k\big],$$
$$\varepsilon_s^{(k)} := \mathbb{P}_{(x,y)\sim\mathcal{D}}\big[A\big(\widehat{M}(g(x)) \odot g(x)\big) = k \mid y \neq k\big],$$

measuring how often Arthur fails to recognise class k from Merlin's selection and how often Morgana flips Arthur into predicting k on a non-k input; these refine the aggregate completeness and soundness of Sec. 3.4, and Suppl. A gives the worst-class variant of Wäldchen et al. (2024) that enters the bound. Under three further conditions, namely a bounded asymmetric feature correlation $\kappa_k \geq 1$ and a relative success rate $\alpha_k > 0$ comparing Morgana's strength to Merlin's, together with the class imbalance $B_k \geq 1$ (directly estimable from label frequencies), $\Pr_{\mathcal{D}}^{(k)}(M)$ is lower-bounded in $\varepsilon_c^{(k)}$ and $\varepsilon_s^{(k)}$ (see Sec. A.2). When both errors are small, Merlin's sparse concept selection almost determines the class.

Faithfulness and scope. We call an explanation *faithful* (see Sec. A.3) when Arthur, with high probability, classifies correctly from Merlin's sparse selection and is not driven to a wrong prediction by Morgana; under the precision-bound conditions this yields the guarantee of Eq. 3. The bound holds under two assumptions, κ_k and α_k, that cannot in general be estimated from data; in exchange, it replaces the unverifiable question of whether the selected concepts are informative with the measurable errors $\varepsilon_c^{(k)}, \varepsilon_s^{(k)}$. While $\Pr_{\mathcal{D}}^{(k)}(M)$ is well-defined for any concept space, estimating it from data requires selected concepts to recur across inputs. This holds for discrete or low-dimensional encodings, but not for dense continuous ones such as CLIP, so the bound is most readily evaluated in the discrete, moderate-C regime. Accordingly, we treat the theory as the motivation for our faithfulness criterion rather than as a quantity measured directly in our experiments.

3.6. Training and Inference

The training step of NCV updates only the parameters of M, $\widehat{M}$, and A, since g is a pretrained model that is kept frozen. The three modules are jointly trained by optimizing a three-module game, which encourages Arthur to rely on concepts selected by Merlin, while being robust to potentially misleading concepts selected by Morgana. Given a concept encoding $\mathbf{c} \in \mathbb{R}^C$, label $y \in \mathcal{Y}$ and cross-entropy function $CE(\cdot,\cdot)$, we define:

- **Merlin's loss:** $L_M = CE(A(\mathcal{S}), y)$, where $\mathcal{S}$ is the sparse concept subset selected by Merlin. This loss encourages Arthur to classify correctly based on Merlin's input.
- **Morgana's loss:** $L_{\widehat{M}} = CE(A(\widehat{\mathcal{S}}), y)$, where $\widehat{\mathcal{S}}$ is Morgana's adversarial concept subset. Here, the loss is interpreted as the classifier's inability to be misled by deceptive inputs[1].

Overall, Arthur's loss combines both objectives with a hyperparameter $\gamma \in \mathbb{R}_{\geq 0}$, controlling the emphasis on predictive performance (*completeness*) versus robustness (*soundness*):

$$L_A = (1-\gamma)\, L_M + \gamma\, L_{\widehat{M}}, \quad (4)$$

In detail, the three modules are updated jointly in a two-phase min-max optimization. First, the provers are updated where Merlin minimizes L_M and Morgana maximizes $L_{\widehat{M}}$; then Arthur is updated by minimizing L_A on the sparse selected concepts chosen by the provers. This scheme incentivizes Arthur to base its predictions on informative, task-relevant, and verifiably robust concept subsets.

At inference time, only the cooperative prover M is used to select a sparse subset of concepts, based on the input's concept encoding. The verifier A then predicts a label or rejects based solely on this selected subset.

Overall, by integrating concept-extractor modules and leveraging the Merlin–Arthur framework, NCV emphasizes faithfulness and interpretability while preserving nonlinear modeling capabilities, and shifts the min–max optimization into a lower-dimensional concept space, improving efficiency, scalability, and stability in high-dimensional settings, a common challenge in min–max optimization for deep learning (Mescheder et al., 2018; Nagarajan & Kolter, 2017).

4. Experimental Evaluations

In this section, we present a comprehensive evaluation of Neural Concept Verifier (NCV) on both synthetic and real-world high-dimensional image datasets. We evaluate based

[1] In practice, the CE loss of Morgana is a slightly modified CE loss (*cf.* Sec. A.4).

on two instantiations of NCV that utilize different concept extractors: a CLIP-based extractor and the Neural Concept Binder (NCB) (Stammer et al., 2024b). We assess predictive performance and interpretability across multiple datasets, compare against several baselines and examine scalability as well as robustness against shortcut learning.

Our evaluation is structured around the following research questions: **(Q1)** Does shifting Prover-Verifier Games (PVGs) to concept-encodings via NCV lead to performative classifiers (*i.e.*, high completeness and soundness) on high-dimensional synthetic and real-world images? **(Q2)** Does NCV reduce the "interpretability-accuracy gap" in the context of CBMs? **(Q3)** Does NCV allow for more detailed explanations over pixel-based PVGs? Finally, **(Q4)** Can training via NCV reduce shortcut learning?

4.1. Experimental Setup

Datasets. We investigate NCV on CLEVR-Hans3 and CLEVR-Hans7 (Stammer et al., 2021), synthetic benchmarks derived from CLEVR (Johnson et al., 2017) that capture complex object compositions and include visual shortcuts. CLEVR-Hans3 features three compositional classes, while CLEVR-Hans7 increases the complexity to seven, with all images rendered at 128×128 pixels. The training and validation sets contain spurious correlations between attributes and labels (e.g., gray cubes linked to a specific class), which are absent in the test set, making the datasets well-suited for studying shortcut behavior. Models that exploit such correlations often fail under the decorrelated test distribution. We first report results on non-confounded versions of these datasets, where feature distributions are consistent across splits, and later return to the confounded versions for shortcut mitigation. To assess scalability and generalization to natural images, we additionally evaluate on ImageNet-1k (Deng et al., 2009) with 1.2M high-resolution images across 1,000 classes (resized to 224×224 pixels), and on CIFAR-100 (Krizhevsky, 2009) with 60,000 low-resolution 32×32 images across 100 fine-grained categories. Lastly, we perform experiments on COCOLogic (Steinmann et al., 2025), a recent benchmark combining real-world images with complex, compositional class rules.

Baseline Models. We compare our framework against several representative baselines, with training details provided in Suppl. C. As a strong but non-interpretable baseline, we use a standard ResNet-18 (He et al., 2016) for evaluations on CLEVR-Hans, and a ResNet-50 for CIFAR-100, COCOLogic and ImageNet-1k, each trained end-to-end on raw images. We further evaluate a pixel-based MAC (Wäldchen et al., 2024) (denoted as *Pixel-MAC*), an instantiation of the Prover-Verifier Game in which the verifier is initialized from a pretrained ResNet-18, while both provers (Merlin and Morgana) are U-Net models (Ronneberger et al., 2015) that output continuous feature-importance masks over the input image. These masks are discretized using Top-k selection to define the features visible to the verifier, and all agents are jointly fine-tuned; the resulting explanations (*i.e.*, certificates) correspond to masks in pixel space (see (Wäldchen et al., 2024) for further details). We also compare to a vanilla Concept Bottleneck Model (Koh et al., 2020) (denoted as *CBM*), where a linear classifier predicts from concept features extracted by either NCB (Stammer et al., 2024b) for CLEVR-Hans or SpLiCE (Bhalla et al., 2024) for CIFAR-100, ImageNet-1k and COCOLogic. In addition, we include a nonlinear CBM variant that replaces the linear classifier head with a nonlinear, non-interpretable head while operating on the same concept encodings, isolating the effect of a more expressive concept-level predictor.

NCV Instantiations. NCV is agnostic to the choice of concept extractor, and we exploit this flexibility to instantiate it across two complementary supervision paradigms: NCB for CLEVR-Hans, where its pretrained slot-based concept encodings match the object-centric structure of the synthetic dataset, and a CLIP-based extractor for the real-world datasets, where a general-purpose vocabulary derived from natural image captions is more appropriate. Within each dataset, we compare NCV against baselines that share the same concept extractor, so that performance differences isolate the effect of the NCV framework rather than the choice of extractor.

For CLEVR-Hans3 and CLEVR-Hans7, we instantiate NCV with NCB (Stammer et al., 2024b) as the concept extractor, using models pretrained on CLEVR (Johnson et al., 2017). A permutation-invariant Set Transformer (Lee et al., 2019) serves as the verifier (Arthur) to process the unordered NCB encodings. The provers (Merlin and Morgana) are independent Set Transformers that take the full concept-slot encodings as input and output a sparse mask of 12 active concepts for the verifier. All modules are jointly trained with the Adam optimizer (Kingma & Ba, 2014). Further details and ablations are provided in Suppl. D. For ImageNet-1k, CIFAR-100 and COCOLogic, we use a CLIP-based concept extractor (Radford et al., 2021), following the approach of SpLiCE (Bhalla et al., 2024) to compute image–text similarity scores. All concepts are drawn from a fixed textual vocabulary derived from LAION captions (Schuhmann et al., 2021) and embedded using CLIP (see Suppl. E). Unlike SpLiCE, which performs per-sample optimization, our method (denoted as *CLIP-Sim*) retains the full activation vector and delegates concept selection to the provers, avoiding expensive inference-time optimization and enabling scalability. Here, the verifier and both provers are two-layer MLPs; the provers output sparse masks of 32 concepts per example. All modules are trained with the Adam optimizer. Additional details and ablations are provided in Suppl. D and Suppl. E including the effect of varying the mask size

and the weighting parameter γ.

Metrics. All methods are evaluated for completeness and, where applicable, soundness (*cf.* Sec. 3.4). We use 20 random seeds for CLEVR-Hans and 10 for ImageNet-1k, CIFAR-100 and COCOLogic, reporting mean and standard deviation across all seeds. For shortcut learning evaluations with CLEVR-Hans, we additionally report a separate *shortcut robustness* metric: the validation-test gap, i.e., the difference between validation accuracy on data with shortcuts and test accuracy on a split without these shortcuts.

4.2. Evaluations

Scaling PVGs to High Dimensions (Q1). In our first evaluation, we examine whether shifting the Prover–Verifier Game (PVG) to concept encodings enables NCV to scale to high-dimensional image domains while achieving strong performance in terms of completeness and soundness. We hereby compare NCV against two key baselines: (1) a black-box ResNet classifier (ResNet-18 for CLEVR-Hans and ResNet-50 for CIFAR-100, ImageNet-1k and COCOLogic), and (2) Pixel-MAC, a nonlinear PVG model operating in raw pixel space. Tab. 1 summarizes results across synthetic (CLEVR-Hans3, CLEVR-Hans7) and real-world (CIFAR-100, ImageNet-1k, COCOLogic) benchmarks. Each model's feature space is indicated for clarity. On the synthetic CLEVR-Hans benchmarks, we observe that NCV consistently achieves the highest completeness scores, surpassing Pixel-MAC and even ResNet-18 on CLEVR-Hans3, while also attaining perfect soundness. This demonstrates that NCV not only matches or exceeds the performance of strong black-box classifiers but also certifiable decision-making. Pixel-MAC performs well in these settings but falls slightly short in completeness and cannot match NCV's zero-error soundness.

On the more challenging real-world datasets, Pixel-MAC either fails entirely or performs poorly. In contrast, NCV successfully scales to these datasets, achieving superior completeness and near-perfect soundness. Notably, NCV surpasses ResNet-50 even in raw accuracy for CIFAR-100 and COCOLogic, providing both higher predictive performance while retaining interpretability. In summary, NCV generalizes well across domains: it scales beyond the limitations of pixel-based PVGs, delivers competitive accuracy even on large-scale and complex datasets, and retains soundness throughout. Overall, these findings affirm Q1: shifting PVGs to concept space yields interpretable classifiers whose decisions can be evaluated via completeness and soundness, while remaining performant and scalable on high-dimensional synthetic and real-world images.

Narrowing the Interpretability–Accuracy Gap (Q2). In Tab. 1, we further examine whether NCV can overcome a central limitation of a majority of Concept Bottleneck Models (CBMs): the interpretability–accuracy gap resulting from their use of constrained linear classifiers (*cf.* Suppl. B for a discussion). We observe that NCV consistently narrows or closes this gap while maintaining high completeness and soundness. Specifically, on CLEVR-Hans3, NCV ($\sim$99%) exceeds the linear CBM ($\sim$95%) and both the opaque ResNet-18 and nonlinear CBM baselines ($\sim$98%) with perfect soundness. On CLEVR-Hans7, NCV narrows the 10-point gap left by linear CBMs to match the performance of its nonlinear counterparts. This trend persists on real-world datasets. On CIFAR-100, NCV outperforms both the linear and nonlinear CBMs and even slightly exceeds ResNet-50's performance. This is even more pronounced on COCOLogic, where NCV outperforms all other models by a large margin. As the additional benefits of a nonlinear classifier are quite small on ImageNet-1k, the additional training complexity of NCV results in a slightly worse performance compared to the CBMs there. Overall, NCV improves over linear CBMs in accuracy, especially on tasks requiring complex concept reasoning such as COCOLogic. At the same time, NCV can match or even surpass the accuracy of opaque ResNet models, demonstrating that interpretable, concept-level classification via Prover–Verifier Games can deliver competitive performance. Additional comparisons with Deep Concept Reasoning (DCR; Barbiero et al. 2023) using the same concept vocabulary are reported in the supplementary material, highlighting that NCV can also outperform CBMs with nonlinear, yet interpretable classifiers on large-scale datasets (*cf.* Tab. 4 in Suppl. C). We therefore answer Q2 affirmatively.

More Detailed Explanations (Q3) We next investigate the resulting explanations produced by NCV, with a focus on explanatory clarity. Since our goal is to improve over classic vision-based Prover–Verifier Games, we compare against pixel-level MAC explanations. Fig. 3 illustrates a qualitative example from both the CLEVR-Hans3 (*cf.* Fig. 4 for more examples) and ImageNet-1k datasets. Notably, under Pixel-MAC, the Prover–Verifier setup operates directly on pixels, yielding broad, diffuse explanation masks that often cover entire objects or irrelevant background regions, arguably providing limited insight regarding which exact features drive the verifier's final decision. In contrast, NCV leverages its internal concept encodings to isolate sparse, high-level concepts that are consistently associated with a class decision under the prover–verifier interaction, recovering the class rule for CLEVR-Hans (i.e., small cube and small metal sphere) and providing a meaningful concept explanation for the coffee-mug class of ImageNet-1k.[2] For clarity, we only show the five most frequent concepts (from

[2]The availability of object-level concepts in NCV depends on the underlying concept extractor. For CLEVR-Hans, we use NCB, which provides such object-based explanations.

Table 1. NCV delivers high predictive performance and soundness through verifiable, concept-based reasoning evaluated via completeness and soundness. We report completeness and soundness scores for ResNet, Pixel-MAC, nonlinear and linear CBMs, and NCV across synthetic (CLEVR-Hans3, CLEVR-Hans7) and real-world (CIFAR-100, ImageNet-1k, COCOLogic) datasets. NCV matches or outperforms baselines in completeness in most settings, while offering strong soundness guarantees.

Model	Feature Space	Completeness (Accuracy)	Soundness (Robustness)	Completeness (Accuracy)	Soundness (Robustness)	Completeness (Accuracy)	Soundness (Robustness)
		CIFAR-100		**ImageNet-1k**		**COCOLogic**	
ResNet-50	pixel space	81.45± 0.60	n/a	76.01± 0.02	n/a	65.80± 3.41	n/a
CBM (nonlin.)	SpLiCE	79.29± 0.42	n/a	69.02± 0.38	n/a	70.09± 0.56	n/a
Pixel-MAC	pixel space	15.27± 4.78	96.31± 4.12	35.06± 3.20	99.65± 0.26	42.57± 3.13	97.70± 0.61
CBM	SpLiCE	75.42± 0.04	n/a	**68.59± 0.01**	n/a	58.84± 0.09	n/a
NCV (ours)	CLIP-Sim	**83.32± 0.28**	**99.99± 0.01**	67.04± 0.16	**99.94± 0.02**	**75.42± 3.21**	**97.87± 0.47**
		CLEVR-Hans3		**CLEVR-Hans7**			
ResNet-18	pixel space	97.87± 0.24	n/a	98.71± 0.24	n/a		
CBM (nonlin.)	NCB	98.13± 0.37	n/a	97.83± 0.25	n/a		
Pixel-MAC	pixel space	96.59± 0.72	99.99± 0.01	97.61± 0.38	99.88± 0.28		
CBM	NCB	95.44± 0.08	n/a	89.12± 0.12	n/a		
NCV (ours)	NCB	**98.92± 0.32**	**100.00± 0.00**	**97.89± 0.31**	**100.00± 0.00**		

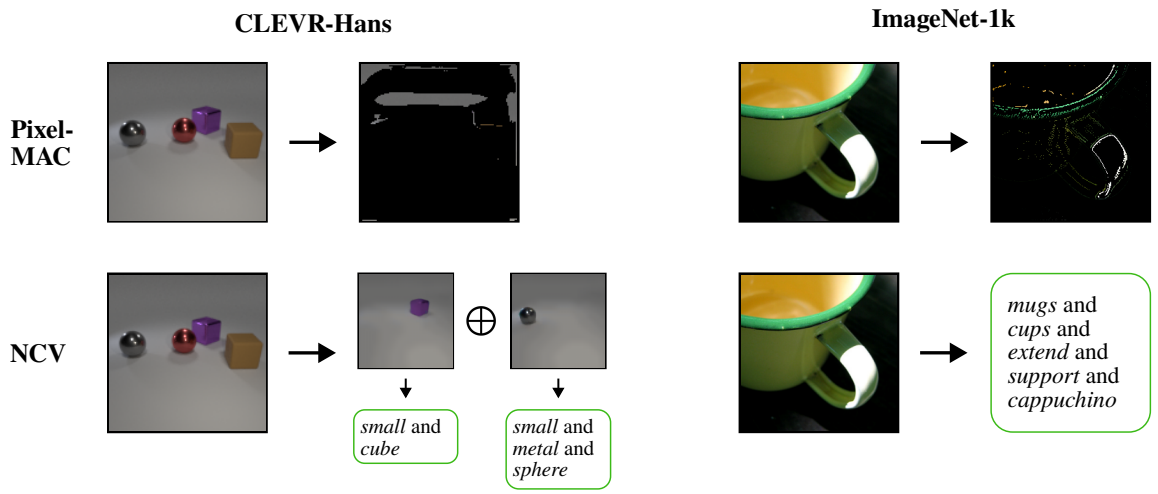


Figure 3. Comparison of explanations from NCV *vs.* Pixel-MAC. **(top)** Merlin–Arthur training on pixel space yields uninformative masks. **(bottom)** MAC on a concept-space via NCV translates into combinations of high-level concepts and, in turn, in an interpretable prediction.

a total mask size of 32) across 32 samples for ImageNet-1k. All concepts are part of the CLIP/SpLiCE-based concept vocabulary that is derived from LAION captions (Schuhmann et al., 2021) (see Suppl. E for more details).

Overall, these examples highlight that NCV offers higher-level, semantically meaningful explanations rather than fine-grained pixel masks, and that concept-level PVGs yield interpretable decisions whose supporting concept subsets can be evaluated even for complex, high-dimensional data. This leads us to answer Q3 affirmatively.

Mitigating Shortcut Learning (Q4) Lastly, to assess whether NCV can mitigate shortcut learning in image classification, we train models on different versions of CLEVR-Hans3 and CLEVR-Hans7 with varying ratios of clean samples (*i.e.*, without shortcut) in the training and validation sets. We then measure validation accuracy with shortcuts and test accuracy on a held-out, clean data split. This setup allows us to track both predictive performance and robustness to shortcut learning. Tab. 2 reports results for three model types: a linear CBM, a nonlinear CBM, and our instantiation of NCV using NCB as concept extractor. We observe that while NCV achieves the highest test accuracy among all models in the 0% clean data setting, it still exhibits a sizeable validation-test gap, indicating a strong influence of the underlying shortcuts. As the amount of clean samples is progressively increased, test accuracy and test-validation gap improves across all models. However, NCV consistently achieves the highest test accuracy in every setting, and its validation–test gap decreases more rapidly than for either

Table 2. Shortcut robustness on CLEVR-Hans3 and CLEVR-Hans7. We report validation accuracy on a shortcut-confounded split, test accuracy on a clean split, and the resulting validation–test gap ("shortcut robustness", lower is better) across models trained with varying amounts of clean data. When any amount of clean data is available, NCV shows improvements compared to training without PVGs.

Ratio Non-Conf. (Samples)	Model	CLEVR-Hans3 Val Acc (w/ shortcut)	CLEVR-Hans3 Test Acc (w/o shortcut)	CLEVR-Hans3 Val-Test Gap (↓)	CLEVR-Hans7 Val Acc (w/ shortcut)	CLEVR-Hans7 Test Acc (w/o shortcut)	CLEVR-Hans7 Val-Test Gap (↓)
0%	CBM (lin.)	95.65± 0.09	90.54± 0.09	5.11	90.37± 0.10	85.27± 0.15	**5.10**
	CBM (non-lin.)	98.70± 0.32	**95.04**± 0.96	**3.66**	98.09± 0.24	90.69± 1.17	7.40
	NCV	**99.44**± 0.15	94.21± 1.41	5.23	**98.38**± 0.18	**92.23**± 0.67	6.15
1% (105)	CBM (lin.)	96.28± 0.16	91.03± 0.31	5.25	90.74± 0.12	85.41± 0.17	5.33
	CBM (non-lin.)	99.10± 0.27	94.84± 0.98	4.26	98.17± 0.17	92.65± 1.31	5.52
	NCV	**99.37**± 0.18	**97.11**± 0.98	**2.26**	**98.19**± 0.24	**94.68**± 0.64	**3.51**
5% (525)	CBM (lin.)	95.38± 0.37	93.34± 0.51	2.04	90.37± 0.15	86.37± 0.18	4.00
	CBM (non-lin.)	98.41± 0.55	96.13± 0.71	2.28	98.32± 0.22	95.19± 0.80	3.13
	NCV	**99.59**± 0.19	**98.88**± 0.37	**0.71**	**98.47**± 0.24	**96.24**± 0.71	**2.23**
20% (2100)	CBM (lin.)	95.67± 0.28	93.46± 0.23	2.21	89.93± 0.29	87.21± 0.31	2.72
	CBM (non-lin.)	99.15± 0.21	98.09± 0.51	1.06	98.21± 0.29	97.00± 0.49	1.21
	NCV	**99.37**± 0.28	**98.82**± 0.67	**0.55**	**98.63**± 0.13	**97.74**± 0.28	**0.89**

CBM variant. This trend indicates that NCV is not only better at leveraging clean supervision when available, but also exhibits a faster reduction of shortcut reliance as clean data increases. Together, these results demonstrate that concept-level prover–verifier interaction in NCV encourage models to rely on more robust, task-relevant features, making NCV more resilient to shortcut learning, even with limited amounts of clean data.

5. Discussion

Overall, our results show that shifting Prover–Verifier Games (PVGs) to the concept level yields a powerful and scalable framework for verifiable, interpretable classification. By operating on symbolic concept embeddings, NCV avoids the computational cost of per-sample inference in pixel space, yet matches or surpasses pixel-based baselines in both completeness and soundness. It reduces the performance gap typical of Concept Bottleneck Models (CBMs), achieving parity with opaque models on synthetic tasks and even surpassing them on natural images. Concept-level outputs offer concise, human-readable explanations. Finally, NCV exhibits higher resilience to spurious correlations, generalizing from confounded training splits and closing the generalization gap with minimal available clean data.

That said, NCV has several limitations. Its effectiveness depends on the quality of the underlying concept extractor: noisy or entangled concept spaces can reduce both accuracy and human understandability. The increased training complexity introduced by the three-module PVG setup also results in greater computational cost and training instability, e.g., compared to linear CBMs. Moreover, when using pretrained models like CLIP for concept discovery, NCV inherits their biases and inconsistencies to some extent (Birhane et al., 2021; Gehman et al., 2020; Bhalla et al., 2024). Finally, recent work (Debole et al., 2025) shows that such concept spaces can diverge from expert semantics, even when yielding strong downstream performance.

6. Conclusion

In this work, we have introduced the Neural Concept Verifier (NCV), a unified framework that combines Prover–Verifier Games (PVGs) and concept-level representations for interpretable classification at scale. Through extensive experiments on CLEVR-Hans, CIFAR-100, ImageNet-1k, and COCOLogic, we have shown that NCV scales well to high-dimensional datasets by achieving high completeness and soundness, reduces the interpretability–accuracy gap of concept bottleneck models, delivers detailed concept-based explanations, and helps mitigate shortcut learning when clean supervision is available. Thus, NCV provides a promising pathway for deploying trustworthy and transparent models in domains where both predictive performance and verifiabile explanations are essential.

Future work should explore how concept encodings can be integrated into alternative PVG-style setups, where structured representations may improve performance or reduce communication overhead. It is also promising to investigate applications such as natural language processing and structured data, where interpretable verification may be equally valuable. Finally, at the optimization level, developing more stable schemes for training Merlin and Morgana end-to-end on discrete, binarized masks could further strengthen both the framework and the provers themselves.

Impact Statement

This work aims to advance methods for interpretable and verifiable machine learning models. While improved interpretability may support safer deployment in high-stakes settings, we do not foresee any negative societal impacts beyond those generally associated with machine learning research.

Acknowledgements

This research was partially supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – The Berlin Mathematics Research Center MATH+ (EXC-2046/1, EXC-2046/2, project ID: 390685689). This work was further supported by the DFG under Germany's Excellence Strategy (EXC 3066/1 "The Adaptive Mind", Project No. 533717223) and by the "ML2MT" project from the Volkswagen Stiftung. It has further benefited from the HMWK projects "The Third Wave of Artificial Intelligence - 3AI", and Hessian.AI, the EU-funded "TANGO" project (EU Horizon 2023, GA No 57100431), from early stages of the Cluster of Excellence "Reasonable AI" funded by the DFG under Germany's Excellence Strategy – EXC-3057 and the research collaboration between TU Darmstadt and Aleph Alpha Research through Lab1141. It has also benefited from the DFG project GRK 2853 "Neuroexplicit Models of Language, Vision, and Action" (project number 471607914).

Software and Data

The code and data used in this work are publicly available at https://github.com/ZIB-IOL/Neural-Concept-Verifier to facilitate reproducibility.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *Advances in neural information processing systems*, 31, 2018.

Ahmed, K. B., Goldgof, G. M., Paul, R., Goldgof, D. B., and Hall, L. O. Discovery of a generalization gap of convolutional neural networks on COVID-19 x-rays classification. *IEEE Access*, 9:72970–72979, 2021.

Amit, N., Goldwasser, S., Paradise, O., and Rothblum, G. Models that prove their own correctness. *arXiv preprint arXiv:2405.15722*, 2024.

Anil, C., Zhang, G., Wu, Y., and Grosse, R. Learning to give checkable answers with prover-verifier games. *arXiv preprint arXiv:2108.12099*, 2021.

Barbiero, P., Ciravegna, G., Giannini, F., Zarlenga, M. E., Magister, L. C., Tonda, A., Lió, P., Precioso, F., Jamnik, M., and Marra, G. Interpretable neural-symbolic concept reasoning. In *International Conference on Machine Learning*, pp. 1801–1825. PMLR, 2023.

Bhalla, U., Oesterling, A., Srinivas, S., Calmon, F., and Lakkaraju, H. Interpreting CLIP with sparse linear concept embeddings (SPLICE). *Advances in Neural Information Processing Systems*, 37:84298–84328, 2024.

Birhane, A., Prabhu, V. U., and Kahembwe, E. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021.

Bortolotti, S., Marconato, E., Morettin, P., Passerini, A., and Teso, S. Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. *arXiv preprint arXiv:2502.11245*, 2025.

Brown-Cohen, J., Irving, G., and Piliouras, G. Scalable AI safety via doubly-efficient debate. In *International Conference on Machine Learning*. JMLR.org, 2024.

Chattopadhyay, A., Chan, K. H. R., Haeffele, B. D., Geman, D., and Vidal, R. Variational information pursuit for interpretable predictions. *arXiv preprint arXiv:2302.02876*, 2023.

Crabbé, J. and van der Schaar, M. Concept activation regions: A generalized framework for concept-based explanations. *Advances in Neural Information Processing Systems*, pp. 2590–2607, 2022.

Cunningham, H., Ewart, A., Riggs, L., Huben, R., and Sharkey, L. Sparse autoencoders find highly interpretable features in language models. *arXiv preprint arXiv:2309.08600*, 2023.

De Felice, G., Flores, A. C., De Santis, F., Santini, S., Schneider, J., Barbiero, P., and Termine, A. Causally reliable concept bottleneck models. *arXiv preprint arXiv:2503.04363*, 2025.

De Santis, F., Bich, P., Ciravegna, G., Barbiero, P., Giordano, D., and Cerquitelli, T. Towards better generalization and interpretability in unsupervised concept-based models. *arXiv preprint arXiv:2506.02092*, 2025.

Debole, N., Barbiero, P., Giannini, F., Passeggini, A., Teso, S., and Marconato, E. If concept bottlenecks are the question, are foundation models the answer? *arXiv preprint arXiv:2504.19774*, 2025.

Debot, D., Barbiero, P., Giannini, F., Ciravegna, G., Diligenti, M., and Marra, G. Interpretable concept-based memory reasoning. *Advances in Neural Information Processing Systems*, 37:19254–19287, 2024.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255, 2009.

Dominici, G., Barbiero, P., Zarlenga, M. E., Termine, A., Gjoreski, M., Marra, G., and Langheinrich, M. Causal concept graph models: Beyond causal opacity in deep learning. *arXiv preprint arXiv:2405.16507*, 2024.

Du, Y., Li, S., Torralba, A., Tenenbaum, J. B., and Mordatch, I. Improving factuality and reasoning in language models through multiagent debate. In *International Conference on Machine Learning*, volume 235, pp. 11733–11763. PMLR, 2024.

Fok, R. and Weld, D. S. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine*, 2023.

Friedrich, F., Stammer, W., Schramowski, P., and Kersting, K. A typology for exploring the mitigation of shortcut behaviour. *Nature Machine Intelligence*, 5(3):319–330, 2023.

Gehman, S., Gururangan, S., Sap, M., Choi, Y., and Smith, N. A. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. *arXiv preprint arXiv:2009.11462*, 2020.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., and Wichmann, F. A. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Ghorbani, A., Wexler, J., Zou, J. Y., and Kim, B. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 2019.

Głuch, G., Turan, B., Nagarajan, S. G., and Pokutta, S. The Good, the Bad and the Ugly: Meta-analysis of watermarks, Transferable Attacks and Adversarial Defenses. *arXiv preprint arXiv:2410.08864*, 2024.

Goldwasser, S., Micali, S., and Rackoff, C. The knowledge complexity of interactive proof-systems. In *ACM Symposium on Theory of Computing*, pp. 291–304. ACM, 1985.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Hendrycks, D. and Gimpel, K. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016.

Irving, G., Christiano, P., and Amodei, D. AI safety via debate. *arXiv preprint arXiv:1805.00899*, 2018.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., and Girshick, R. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., and Viegas, F. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, pp. 2668–2677. PMLR, 2018.

Kimura, M., Shimizu, R., Hirakawa, Y., Goto, R., and Saito, Y. On permutation-invariant neural networks. *arXiv preprint arXiv:2403.17410*, 2024.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kirchner, J. H., Chen, Y., Edwards, H., Leike, J., McAleese, N., and Burda, Y. Prover-verifier games improve legibility of LLM outputs. *arXiv preprint arXiv:2407.13692*, 2024.

Koh, P. W., Nguyen, T., Tang, Y. S., Mussmann, S., Pierson, E., Kim, B., and Liang, P. Concept bottleneck models. In *International Conference on Machine Learning*, pp. 5338–5348. PMLR, 2020.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lapuschkin, S., Wäldchen, S., Binder, A., Montavon, G., Samek, W., and Müller, K.-R. Unmasking clever hans predictors and assessing what machines really learn. *Nature Communications*, 10(1):1–8, 2019.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3744–3753. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/lee19d.html.

Lee, J. H., Lanza, S., and Wermter, S. From neural activations to concepts: A survey on explaining concepts in neural networks. *Neurosymbolic Artificial Intelligence*, 1: NAI–240743, 2025.

Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., and Pan, W. Promises and pitfalls of black-box concept learning models. *arXiv preprint arXiv:2106.13314*, 2021.

Marconato, E., Teso, S., Vergari, A., and Passerini, A. Not all neuro-symbolic concepts are created equal: Analysis

and mitigation of reasoning shortcuts. *Advances in Neural Information Processing Systems*, 36:72507–72539, 2023.

Mescheder, L., Geiger, A., and Nowozin, S. Which training methods for GANs do actually converge? In *International Conference on Machine Learning*, pp. 3481–3490. PMLR, 2018.

Nagarajan, V. and Kolter, J. Z. Gradient descent GAN optimization is locally stable. In *Advances in Neural Information Processing Systems*, pp. 5591–5600. Curran Associates Inc., 2017.

Nair, V., Schumacher, E., Tso, G., and Kannan, A. DERA: enhancing large language model completions with dialog-enabled resolving agents. *arXiv preprint arXiv:2303.17071*, 2023.

Oikarinen, T., Das, S., Nguyen, L. M., and Weng, T.-W. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Panousis, K. P., Ienco, D., and Marcos, D. Coarse-to-fine concept bottleneck models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., and Antiga, L. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32, 2019.

Poeta, E., Ciravegna, G., Pastor, E., Cerquitelli, T., and Baralis, E. Concept-based explainable artificial intelligence: A survey. *arXiv preprint arXiv:2312.12936*, 2023.

Pruthi, D., Bansal, R., Dhingra, B., Baldini Soares, L., Collins, M., Lipton, Z. C., Neubig, G., and Cohen, W. W. Evaluating explanations: How much do explanations from the teacher aid students? *Transactions of the Association for Computational Linguistics*, 10:359–375, 2022. doi: 10.1162/tacl_a_00465. URL https://aclanthology.org/2022.tacl-1.21/.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., and Clark, J. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241. Springer, 2015.

Ross, A. S., Hughes, M. C., and Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. In *International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2017.

Rudin, C. Stop explaining black box models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

Schneider, J. and Vlachos, M. Reflective-net: Learning from explanations. *Data Mining and Knowledge Discovery*, 38(5):2975–2996, 2024.

Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.-G., Mahlein, A.-K., and Kersting, K. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8):476–486, 2020.

Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., and Komatsuzaki, A. LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*, 2021.

Schut, L., Tomašev, N., McGrath, T., Hassabis, D., Paquet, U., and Kim, B. Bridging the human–AI knowledge gap through concept discovery and transfer in AlphaZero. *Proceedings of the National Academy of Sciences*, 122 (13), 2025.

Stammer, W., Schramowski, P., and Kersting, K. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3619–3629, 2021.

Stammer, W., Friedrich, F., Steinmann, D., Brack, M., Shindo, H., and Kersting, K. Learning by self-explaining. *Transactions on Machine Learning Research*, 2024a.

Stammer, W., Wüst, A., Steinmann, D., and Kersting, K. Neural concept binder. In *Advances in Neural Information Processing Systems*, pp. 71792–71830. Curran Associates, Inc., 2024b.

Steinmann, D., Divo, F., Kraus, M., Wüst, A., Struppek, L., Friedrich, F., and Kersting, K. Navigating shortcuts, spurious correlations, and confounders: From origins via detection to mitigation. *arXiv preprint arXiv:2412.05152*, 2024.

Steinmann, D., Stammer, W., Wüst, A., and Kersting, K. Object centric concept bottlenecks. *arXiv preprint arXiv:2505.24492*, 2025.

Teneggi, J. and Sulam, J. Testing semantic importance via betting. In Globerson, A., Mackey, L., Belgrave, D.,

Fan, A., Paquet, U., Tomczak, J., and Zhang, C. (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 76450–76499. Curran Associates, Inc., 2024. doi: 10.52202/079017-2435.

Wäldchen, S., Sharma, K., Turan, B., Zimmer, M., and Pokutta, S. Interpretability guarantees with Merlin-Arthur classifiers. In *International Conference on Artificial Intelligence and Statistics*, volume 238, pp. 1963–1971. PMLR, 2024.

Wang, Z., Popel, A., and Sulam, J. Concept bottleneck model with zero performance loss. In Chen, B., Liu, S., Pilanci, M., Su, W., Sulam, J., Wang, Y., and Zhu, Z. (eds.), *Conference on Parsimony and Learning*, volume 280 of *Proceedings of Machine Learning Research*, pp. 433–461. PMLR, 24–27 Mar 2025. URL https://proceedings.mlr.press/v280/wang25b.html.

Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., and Yatskar, M. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19187–19197, 2023.

Yuksekgonul, M., Wang, M., and Zou, J. Post-hoc concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Zarlenga, M. E., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lio, P., and Jamnik, M. Concept embedding models: beyond the accuracy-explainability tradeoff. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

Zhou, Y., Booth, S., Ribeiro, M. T., and Shah, J. Do feature attribution methods correctly attribute features? In *Proceedings of the AAAI conference on artificial intelligence*, volume 36, pp. 9623–9633, 2022.

Supplementary Materials

A. Theoretical Guarantees and Relation to Merlin–Arthur Classifiers

Neural Concept Verifier (NCV) builds on the Merlin–Arthur classifier (MAC) framework of (Wäldchen et al., 2024), which provides information-theoretic interpretability guarantees in a binary classification setting. In this appendix, we briefly recall these guarantees and explain how they apply in our concept-based setting. We do *not* prove new theorems here; rather, we instantiate existing results and make explicit which assumptions are required and where our claims remain empirical.

A.1. Merlin–Arthur guarantees in the binary case

We briefly recall the guarantees for Merlin–Arthur classifiers in the original binary setting of (Wäldchen et al., 2024), focusing on the quantities that appear in our NCV discussion: average precision, mutual information, completeness, soundness, asymmetric feature correlation, and relative success rate.

Setup. Throughout this subsection, the notation is local to the binary Merlin–Arthur setting of (Wäldchen et al., 2024) and should not be conflated with the concept-space notation of the main text; in particular, $\mathcal{D}$ here denotes a distribution over data points (with a deterministic label map), rather than the joint distribution over input–label pairs used in the main paper. We consider a two-class data space $\mathfrak{D} = (\Omega, \mathcal{D}, \ell)$, where Ω is the set of data points, $X \sim \mathcal{D}$, and $\ell : \Omega \to \{-1, 1\}$ is a deterministic label map with $L = \ell(X)$. The class-conditional distributions are $\mathcal{D}_s := \mathcal{D}(\cdot \mid L = s)$ for $s \in \{-1, 1\}$. Following (Wäldchen et al., 2024), each data point $x \in \Omega$ is identified with the set of its indexed entries $\{(1, x_1), \ldots, (d, x_d)\}$, and a *feature* is a partial object $z \in \Omega_p$ (*e.g.*, a subset of pixels); we write $z \subseteq x$ when z is contained in this set, *i.e.* when x exhibits the feature z. A feature selector maps a data point to one of its features, $M(x) \subseteq x$; we write M for Merlin and $\widehat{M}$ for Morgana. Arthur is a classifier A predicting in $\{-1, \perp, 1\}$, where $\perp$ denotes abstention (the "Don't know" option).

Average precision and mutual information. Given a feature z and a reference point x' that exhibits it ($z \subseteq x'$), the *precision* of z is the probability that an independently drawn point $x \sim \mathcal{D}$ which also exhibits z shares the label of x':

$$\Pr(z; x') := \mathbb{P}_{x \sim \mathcal{D}}\big[\ell(x) = \ell(x') \mid z \subseteq x\big] .$$

For a feature selector M, the *average precision* is the expected precision of the features it selects:

$$\Pr_{\mathcal{D}}(M) := \mathbb{E}_{x' \sim \mathcal{D}}\Big[\mathbb{P}_{x \sim \mathcal{D}}\big[\ell(x) = \ell(x') \mid M(x') \subseteq x\big]\Big]. \tag{5}$$

This quantity bounds the average conditional entropy of the class given Merlin's features, and hence the mutual information. Writing H_b for the binary entropy and $H(\cdot)$, $I(\cdot\,;\cdot)$ for entropy and mutual information, (Wäldchen et al., 2024) show

$$\mathbb{E}_{x' \sim \mathcal{D}}\big[I_{x \sim \mathcal{D}}\big(\ell(x);\, M(x') \subseteq x\big)\big] \;\geq\; H_{x \sim \mathcal{D}}\big(\ell(x)\big) - H_b\big(\Pr_{\mathcal{D}}(M)\big), \tag{6}$$

where, for fixed x', $M(x') \subseteq x$ denotes the binary containment variable. As $\Pr_{\mathcal{D}}(M) \to 1$, the term $H_b(\Pr_{\mathcal{D}}(M)) \to 0$ and Merlin's features carry almost all available label information in this binary setting.

Idealised min–max guarantee (optimal players). In the idealised setting, Arthur and Morgana are assumed to play optimally against a fixed Merlin. Define the error set

$$E_{M,\widehat{M},A} := \Big\{x \in \Omega \,\Big|\, A(M(x)) \neq \ell(x) \;\vee\; A(\widehat{M}(x)) = -\ell(x)\Big\},$$

i.e., the points where Merlin fails to convince Arthur of the correct label or Morgana forces an incorrect (non-abstaining) prediction. The min–max error of Merlin is

$$\varepsilon_M := \min_A \max_{\widehat{M}} \mathbb{P}_{x \sim \mathcal{D}}\big[x \in E_{M,\widehat{M},A}\big].$$

Theorem 2.7 of (Wäldchen et al., 2024) states that if ε_M is small, there exists a subset $\Omega' \subseteq \Omega$ of mass at least $1 - \varepsilon_M$ such that, on the induced data space $\mathfrak{D}' = (\Omega', \mathcal{D}', \ell)$, Merlin achieves perfect precision, and hence zero conditional class entropy:

$$\Pr_{\mathcal{D}'}(M) = 1 \quad \text{and thus} \quad H_{x', x \sim \mathcal{D}'}\big(\ell(x') \mid M(x') \subseteq x\big) = 0.$$

In words: if an optimal Arthur–Morgana pair can almost never disagree with Merlin, then on almost all of the data space Merlin's features determine the label uniquely.

Realistic players: completeness, soundness, AFC, and relative strength. For high-dimensional data, exhaustive search for an optimal Morgana is not feasible. The analysis is therefore relaxed to *realistic* (*e.g.*, neural) players and expressed in terms of:

- **Completeness**
$$\min_{s\in\{-1,1\}} \mathbb{P}_{x\sim\mathcal{D}_s}\big[A(M(x)) = \ell(x)\big] \;\geq\; 1-\varepsilon_c,$$
i.e., Merlin's features let Arthur classify correctly with high probability in each class.

- **Soundness**
$$\max_{s\in\{-1,1\}} \mathbb{P}_{x\sim\mathcal{D}_s}\big[A(\widehat{M}(x)) = -\ell(x)\big] \;\leq\; \varepsilon_s,$$
i.e., Morgana almost never forces a wrong prediction; on her subsets, Arthur either stays correct or abstains (predicts $\perp$).

- **Class imbalance** B, bounding how skewed the class prior can be,
$$B := \max_{s\in\{-1,1\}} \frac{\mathbb{P}_{x\sim\mathcal{D}}[\ell(x)=s]}{\mathbb{P}_{x\sim\mathcal{D}}[\ell(x)=-s]}.$$

- **Asymmetric Feature Correlation (AFC)** κ, which measures how strongly a set of features can be *concentrated* in a few points of one class while *spread out* across many points of the other. A large κ lets Merlin use globally uninformative features (appearing equally often in both classes) while still attaining high completeness and soundness. We use κ only through the assumption $\kappa = O(1)$ and refer to Definition 2.8 of (Wäldchen et al., 2024) for its formal definition.

- **Relative success rate** α of Morgana, comparing how often she finds a convincing feature in the *wrong* class to how often Merlin does so in the *correct* class, restricted to points where such a feature exists:
$$\alpha := \min_{s\in\{-1,1\}} \frac{\mathbb{P}_{x\sim\mathcal{D}_{-s}}\big[A(\widehat{M}(x)) = s \mid x\in F_s^*\big]}{\mathbb{P}_{x\sim\mathcal{D}_s}\big[A(M(x)) = s \mid x\in F_s^*\big]},$$
where $F_s^* := \{x\in\Omega \mid \exists\, z\subseteq x : z\in M(\mathcal{D}_s),\ A(z)=s\}$ is the set of points containing a feature that Merlin selects on some class-s point and that Arthur accepts as class s. Intuitively, α is large when Morgana's search is at least as powerful as Merlin's.

Under these conditions, (Wäldchen et al., 2024) prove that completeness, soundness, AFC, class imbalance, and relative success rate jointly lower-bound the average precision:

$$\Pr_{\mathcal{D}}(M) \;\geq\; 1-\varepsilon_c - \frac{\kappa\alpha^{-1}\varepsilon_s}{1-\varepsilon_c+\kappa\alpha^{-1}B^{-1}\varepsilon_s}. \tag{7}$$

Furthermore, the AFC is bounded by the maximum number of features per data point: if each point contains at most $F_{\max}$ features of the type considered (*e.g.*, patches of a fixed size), then $\kappa \leq F_{\max}$, which gives a concrete regime in which the assumption $\kappa = O(1)$ holds.

Combining the precision bound in Eq. 7 with the mutual-information bound in Eq. 6 yields

$$\mathbb{E}_{x'\sim\mathcal{D}}\big[I_{x\sim\mathcal{D}}\big(\ell(x);\, M(x')\subseteq x\big)\big] \;\geq\; H_{x\sim\mathcal{D}}\big(\ell(x)\big) - H_b\!\left(1-\varepsilon_c - \frac{\kappa\alpha^{-1}\varepsilon_s}{1-\varepsilon_c+\kappa\alpha^{-1}B^{-1}\varepsilon_s}\right),$$

with $M(x')\subseteq x$ understood as above. For *balanced* datasets ($B\approx 1$), bounded AFC ($\kappa = O(1)$), and a reasonably strong Morgana ($\alpha = O(1)$), high completeness ($\varepsilon_c \ll 1$) and soundness ($\varepsilon_s \ll 1$) therefore imply that Merlin's features carry almost all label information in this binary setting.

A.2. Instantiation for concept-based models

NCV applies the same prover–verifier game as Merlin–Arthur classifiers, but in a *concept space* rather than pixel space, and in a *multi-class* rather than binary setup. We return here to the notation of the main text (Sec. 3.1): for an input $x \in \mathbb{R}^D$, a concept extractor g produces an encoding $\mathbf{c} = g(x) \in \mathbb{R}^C$. In analogy to the index–value features of Sec. A.1, a feature here is a concept index j together with its value $g(x)_j$; Merlin selects up to m such (concept, value) pairs via the mask $M(g(x))$. Merlin and Morgana are implemented as provers $M, \widehat{M} : \mathbb{R}^C \to \{0,1\}^C$ that output sparse binary masks $M(\mathbf{c}), \widehat{M}(\mathbf{c})$ with at most m active entries. Arthur then predicts using the masked encodings

$$S = M(\mathbf{c}) \odot \mathbf{c} \quad \text{and} \quad \widehat{S} = \widehat{M}(\mathbf{c}) \odot \mathbf{c}.$$

We treat the selected partial encoding $M(g(x')) \odot g(x')$ as Merlin's feature, and define the containment relation $M(x') \subseteq x$ (writing $M(x') := M(g(x'))$ as in the main text) to hold when $g(x)_i = g(x')_i$ at every index i with $M(g(x'))_i = 1$, *i.e.* when x and x' agree on every concept Merlin selects from x'. Under this instantiation, the binary results of Sec. A.1 apply to each one-vs-rest subproblem, with concept-space parameters $(\kappa_k, \alpha_k, B_k)$, subject to the evaluability caveat discussed below.

To connect this to the binary setting of Sec. A.1, we apply a one-vs-rest reduction: for a fixed class $k \in \{1, \dots, K\}$ we form the binary task with label $Y_k(y) = +1$ if $y = k$ and -1 otherwise (class k vs. all others). The selector M, the feature $M(g(x)) \odot g(x)$, and Arthur (now predicting in $\{-1, \perp, 1\}$) play the roles of the corresponding objects in Sec. A.1, with the containment relation as defined above.

Let $\Pr_{\mathcal{D}}^{(k)}(M)$ denote the *average precision* of Merlin's concept features on this one-vs-rest task. As in Sec. 3.5 we work over input–label pairs $(x, y), (x', y') \sim \mathcal{D}$, and set, analogously to Eq. 5,

$$\Pr_{\mathcal{D}}^{(k)}(M) := \mathbb{E}_{(x',y')}\Big[\mathbb{P}_{(x,y)}\big[Y_k(y) = Y_k(y') \mid M(x') \subseteq x\big]\Big].$$

Applying the mutual-information bound of Eq. 6 to the binary label $Y_k(y)$ and selector M yields

$$\mathbb{E}_{(x',y')\sim\mathcal{D}}\big[I_{(x,y)\sim\mathcal{D}}\big(Y_k(y);\, M(x') \subseteq x\big)\big] \;\geq\; H_{y\sim\mathcal{D}}\big(Y_k(y)\big) - H_b\big(\Pr_{\mathcal{D}}^{(k)}(M)\big), \tag{8}$$

where, for fixed x', $M(x') \subseteq x$ denotes the binary containment variable. Thus, if $\Pr_{\mathcal{D}}^{(k)}(M)$ is close to 1, Merlin's sparse concept subsets for class k carry almost all information about $Y_k(y)$.

The analysis of Sec. A.1 further relates $\Pr_{\mathcal{D}}^{(k)}(M)$ to *observable* completeness and soundness on this binary subproblem, under three additional assumptions:

- a bounded *Asymmetric Feature Correlation* (AFC) κ_k for the one-vs-rest task of class k, measured in concept space. When a single constant is convenient we use the uniform bound $\kappa := \max_k \kappa_k$. If each input has at most $m_{\max}$ active concepts (*e.g.*, under a fixed sparsity level or top-m selection), then $\kappa_k \leq m_{\max}$, giving a concrete regime in which the assumption $\kappa_k = O(1)$ holds.
- a bounded *class imbalance* B_k for the one-vs-rest task;
- a *relative success rate* $\alpha_k > 0$ of Morgana, *i.e.* her search over concept subsets is roughly as powerful as Merlin's. In the ideal full-search setting $\alpha_k = 1$; in NCV we use symmetric neural architectures for Merlin and Morgana as heuristic evidence that α_k is close to 1, but we do not estimate it explicitly.

Under these conditions, the precision bound Eq. 7 applies class-wise: for each k one obtains

$$\Pr_{\mathcal{D}}^{(k)}(M) \;\geq\; 1 - \varepsilon_c^{(k)} - \frac{\kappa_k \alpha_k^{-1} \varepsilon_s^{(k)}}{1 - \varepsilon_c^{(k)} + \kappa_k \alpha_k^{-1} B_k^{-1} \varepsilon_s^{(k)}},$$

where $\varepsilon_c^{(k)}$ and $\varepsilon_s^{(k)}$ are the completeness and soundness errors of NCV on the one-vs-rest problem for class k. Combining this with Eq. 8 yields a lower bound on the mutual information between the class-k label and Merlin's concept subsets.

In NCV we view these results as an *idealised* description of the concept-level prover–verifier game:

- On each one-vs-rest task, and under the AFC and relative-strength assumptions above, high completeness and soundness imply that Merlin's sparse concept subsets are highly informative about the label in the sense of Eq. 8.
- The precision $\Pr_{\mathcal{D}}^{(k)}(M)$ is well-defined for any concept extractor, but estimating it empirically requires the selected concepts to recur across inputs. This holds for discrete or low-dimensional encodings (*e.g.*, our NCB instantiation), but for dense continuous encodings such as CLIP exact recurrence is vanishingly rare, and the precision is then meaningful only as a population quantity. The bound is therefore most directly evaluable in the discrete, moderate-C regime.
- In our high-dimensional, multi-class experiments we do not estimate the class-wise parameters $(\varepsilon_c^{(k)}, \varepsilon_s^{(k)}, \alpha_k, B_k, \kappa_k)$ explicitly. We therefore interpret the Merlin–Arthur theory as a *theoretical lens* for NCV, rather than a quantitative certification for each dataset; empirically reported completeness and soundness should be read in this light (*cf.* the discussion in Sec. 3.5).

A.3. Notion of faithfulness

Definition (Faithfulness of NCV). For a class $k \in \{1, \ldots, K\}$, we say that NCV produces *faithful* explanations in the information-theoretic sense inherited from Merlin–Arthur (and adapted to concept space) when the following three conditions hold:

1. **Completeness.** For an input x with true class k, Merlin produces a sparse concept subset $S(x) = M(\mathbf{c}) \odot \mathbf{c}$ (with $\mathbf{c} = g(x)$) such that, with high probability over $x \sim \mathcal{D}_k$, Arthur can correctly predict k from $S(x)$ alone; *i.e.*, the completeness error $\varepsilon_c^{(k)}$ for the corresponding one-vs-rest task is small.
2. **Soundness.** Morgana cannot find alternative concept subsets that force Arthur into a wrong prediction; at worst, Arthur abstains via the rejection class $\perp$. Equivalently, the soundness error $\varepsilon_s^{(k)}$ in the sense of Sec. A.1 is small.
3. **Precision-bound assumptions.** The concept-space AFC parameter κ_k, the class imbalance B_k, and Morgana's relative success rate α_k are well-behaved (bounded, with α_k not too small), so that the precision bound of Sec. A.1 links the errors $\varepsilon_c^{(k)}, \varepsilon_s^{(k)}$ to the class-wise average precision $\Pr_{\mathcal{D}}^{(k)}(M)$.

When all three hold, the precision bound guarantees that $\Pr_{\mathcal{D}}^{(k)}(M)$ is close to 1, and the mutual-information bound Eq. 8 then implies that Merlin's sparse concept subsets for class k carry near-maximal information about $Y_k(y)$.

In our experiments we report completeness and soundness empirically, but we do not attempt to estimate the resulting lower bounds on $\Pr_{\mathcal{D}}^{(k)}(M)$, for the reasons detailed in Sec. A.5. This is explicitly *not* a causal guarantee: NCV does not show that the concepts are causally sufficient for the task. It ensures only that, relative to the given concept representation and under the assumptions above, the sparse subsets Merlin selects are informative and robust in the sense made precise by the bounds above.

A.4. Use of the rejection class and training objective

Finally, we clarify how the rejection class is used during training, since it is crucial for enforcing soundness in the sense of Sec. A.1. Arthur outputs logits $z = A(S) \in \mathbb{R}^{K+1}$ over K classes plus a rejection class $\perp$, and we write $p = \mathrm{softmax}(z)$ for the predictive probabilities, with p_y and $p_\perp$ the probabilities of the true class y and the rejection class. Given a concept encoding $\mathbf{c} = g(x)$ and the Merlin/Morgana subsets $S = M(\mathbf{c}) \odot \mathbf{c}$ and $\widehat{S} = \widehat{M}(\mathbf{c}) \odot \mathbf{c}$, we use the following losses.

Merlin loss.

$$L_M = -\log p_y,$$

i.e., standard cross-entropy with respect to the true class on Merlin's subset.

Morgana loss (soundness). Morgana's loss operates directly on the logits for the true class and the rejection class. Let $\widehat{z} = A(\widehat{S}) \in \mathbb{R}^{K+1}$ be Arthur's logits on Morgana's subset, with components $\widehat{z}_y$ and $\widehat{z}_\perp$. We first define a modified target $\tilde{y}$ that switches to the rejection class whenever Arthur already prefers $\perp$ over y on $\widehat{S}$:

$$\tilde{y} = \begin{cases} y, & \text{if } \widehat{z}_y \geq \widehat{z}_\perp, \\ \perp, & \text{if } \widehat{z}_\perp > \widehat{z}_y, \end{cases}$$

and set

$$L_{\widehat{M}} = \mathrm{CE}\big(\widehat{z}, \tilde{y}\big) \;+\; \frac{1}{N}\sum_{i=1}^{N}\Big(-\log\big(1+\exp(-|\widehat{z}_y^{(i)} - \widehat{z}_\perp^{(i)}|)\big)\Big),$$

where N is the batch size. Arthur minimises this loss while Morgana maximises it. The cross-entropy term targets the correct-or-abstain label $\tilde{y}$, and the second term is a softplus penalty on the gap $|\widehat{z}_y - \widehat{z}_\perp|$. Through the $\perp$ branch of $\tilde{y}$, abstention counts as sound, consistent with Sec. A.1.

Arthur's objective and optimisation. Arthur's overall loss is the convex combination

$$L_A = (1-\gamma)\, L_M + \gamma\, L_{\widehat{M}}, \qquad \gamma \in [0, 0.5],$$

trading off completeness and soundness; all losses use mean reduction over the batch. Although the concept selections are ultimately discrete, both provers are trained through continuous masks. Each prover outputs real-valued scores $m_{\text{cont}}, \widehat{m}_{\text{cont}} \in \mathbb{R}^C$, used as soft masks to form

$$S_{\text{soft}} = m_{\text{cont}} \odot \mathbf{c}, \qquad \widehat{S}_{\text{soft}} = \widehat{m}_{\text{cont}} \odot \mathbf{c}.$$

When updating the provers (Merlin by gradient descent on L_M, Morgana by gradient ascent on $L_{\widehat{M}}$), Arthur is frozen and fed $S_{\text{soft}}, \widehat{S}_{\text{soft}}$, so that gradients flow back into the continuous mask parameters. After the prover updates we discretise the masks with a top-m operator, setting the m entries of largest magnitude in m_{cont} (respectively $\widehat{m}_{\text{cont}}$) to 1 and the rest to 0, yielding hard masks $M(\mathbf{c}), \widehat{M}(\mathbf{c}) \in \{0,1\}^C$. Arthur is then updated by gradient descent on L_A using the hard-masked encodings S and $\widehat{S}$. This alternating scheme, namely continuous masks for gradient flow in the provers and hard top-m masks for Arthur's update, implements a practical min-max training procedure for NCV. The explicit rejection class $\perp$ ensures that soundness measures robustness to *adversarial* concept selections, in that Arthur is never driven to be confidently wrong, aligning with the Merlin–Arthur framework instantiated in concept space.

A.5. Limitations of the theoretical guarantees

For completeness, we also spell out the main limitations of the information-theoretic guarantees when applied to NCV's high-dimensional, multi-class, concept-based setting.

First, the original theory is formulated for binary classification. Our instantiation in Sec. A.2 uses a one-vs-rest reduction to obtain class-wise guarantees on the mutual information $I(Y_k(y);\, M(x') \subseteq x)$ for each k, but it does not provide a direct statement about the joint K-class decision or about the final $\arg\max$ prediction over all classes.

High-dimensional sparsity and feature reuse. A central limitation arises from the extreme sparsity of Merlin's and Morgana's selections in our high-dimensional concept spaces. In the original Merlin–Arthur setting, features are typically small, localized structures (*e.g.*, image patches) and the experiments are conducted on relatively low-dimensional, small-scale datasets. In this regime, many inputs share the same features, so the event $\{M(x') \subseteq x\}$ has non-negligible probability and the average precision $\mathrm{Pr}_{\mathcal{D}}(M)$ in Eq. 5 can be meaningfully interpreted and estimated from finite samples. In NCV, Merlin and Morgana select small subsets of a large concept vocabulary (*e.g.*, $m \ll C$ with C in the thousands), and in practice the exact subsets $M(x')$ and $\widehat{M}(x')$ are often highly specific to each input. As a result, the precise event $\{M(x') \subseteq x\}$ may have very low probability under $\mathcal{D}$, and empirical estimates of $\mathrm{Pr}_{\mathcal{D}}^{(k)}(M)$ become unstable in finite samples. For this reason, we do not attempt to estimate average precision or the resulting mutual-information lower bounds numerically in our experiments. Instead, we use completeness and soundness as observable proxies and treat the precision/MI guarantees as an idealised, distribution-level description of the behaviour that NCV is designed to promote, rather than as directly calibrated finite-sample certificates.

Class imbalance in one-vs-rest reductions. In large-K settings such as ImageNet-1k, the one-vs-rest subproblems for each class k are highly imbalanced (the positive prior is approximately $1/K$). In the precision bound of Sec. A.2, this is reflected in the imbalance parameter B_k, whose large value makes the resulting lower bound on $\mathrm{Pr}_{\mathcal{D}}^{(k)}(M)$ more conservative: small soundness errors $\varepsilon_s^{(k)}$ are penalised more strongly relative to completeness $\varepsilon_c^{(k)}$. This provides an additional reason why we do not attempt to compute numerical mutual-information lower bounds in our experiments, and instead interpret the Merlin–Arthur theory qualitatively as a guiding framework.

Relative success and adversary class. Finally, the robustness interpretation of soundness in our setting is tied to the particular adversarial prover class we train in practice (a neural network $\widehat{M}$ with a specific architecture and loss). In the Merlin–Arthur framework, this dependence is captured abstractly via the relative success rate α, which measures how powerful Morgana is compared to Merlin on those points where Merlin can provide convincing evidence. In NCV we use symmetric neural architectures for M and $\widehat{M}$ as heuristic evidence that α_k is not tiny, but we do not attempt to verify or optimise α_k beyond this design choice. As a result, our empirical soundness estimates should be interpreted as robustness against this trained adversary class, rather than against an arbitrary worst-case adversary over all admissible subsets. This is fully in line with the practical use of the Merlin–Arthur framework in (Wäldchen et al., 2024), where the general theory is instantiated and evaluated with specific neural implementations of Merlin, Morgana, and Arthur.

Overall, we see the Merlin–Arthur framework as providing a rigorous *idealised model* for the kind of behaviour NCV is designed to encourage: high completeness, robustness to adversarial concept subsets, and information-rich sparse explanations in concept space. In our experiments, we use completeness and soundness as observable proxies for these properties, but we do not claim fully certified guarantees beyond the stated assumptions.

B. Why Linear Classifiers Fall Short in CBMs

While linear classifiers are generally considered to be interpretable, these models are not suited to solve arbitrarily complex problems. A linear classifier is only able to capture linear relationships between inputs and output features and cannot model complex, non-linear relationships. In the context of CBMs, this problem is usually tackled by utilizing a linear classifier to predict the output based on the detected concepts. The concepts themselves can be detected using non-linear models, and only the classification based on these concepts is done with a linear model. However, this is not always sufficient, as there are also simple examples where non-linear relationships between concepts and the output exist, for example thresholds detection (three out of five symptoms need to be present to indicate an illness) or multiplicative effects (crop yield is the result of a multiplicative relationship between rain and fertility).

To illustrate the problem in a simple experimental setup, let us assume we have a dataset of simple shapes and every image contains between one and four of these shapes. The shapes are either a square or a circle and either orange or blue. We consider two simple classification scenarios for this dataset.

- **XOR:** This setting classification follows the traditional XOR problem: We want to classify images that contain either an orange square or a blue circle as class one and all other images as class two.
- **Counting:** This setting includes object counting and illustrates that even for classification based on a single attribute, a linear layer can be insufficient. Here, we want to classify all images with exactly one blue shape as class one, and all other images as class two.

We evaluate a linear layer and a simple MLP on this toy dataset. To further simplify things, we assume that our concept encoder is able to perfectly detect the concepts in the image, thus providing for each element the information whether there is an object and if so, its shape and color. We randomly generate 5000 samples of the dataset and train the models on a train split of 80% and evaluate on the remaining 20%. The MLP has one hidden layer of size 16 and uses ReLU activation functions.

The results of this evaluation are shown in Tab. 3. In both scenarios, the linear classification layer is not able to solve the task, despite the deceptively simple relationship between concepts and output classes. On the other hand, the MLP achieves close to perfect accuracy on both settings.

Table 3. A linear prediction layer cannot solve *XOR* or *counting*. Even with the assumption of a perfect concept encoder, the linear layer fails.

Model	XOR (Acc)	Counting (Acc)
Linear Layer	0.766 ± 0.011	0.677 ± 0.006
MLP	0.953 ± 0.053	0.982 ± 0.015

So far, we have argued that not every task can be solved with a CBM and a linear classification layer. However, this is not entirely accurate. In principle, any task can be solved linearly—provided that we define the right *linear-sufficient* concepts. For instance, in the XOR setting, detecting the concepts "orange square and no blue circle" and "blue circle and no orange square" would allow a linear classifier to solve the task. Similarly, in the counting task, introducing a concept such as "exactly one blue object" would make linear classification trivial.

That said, the assumption that such sufficient concepts are always available is not realistic. First, designing or discovering these concepts often makes concept detection considerably more difficult. Second, as concepts become increasingly specific and compositional, they tend to lose interpretability. Finally, requiring tailored concepts for every individual task does not scale. Returning to our example, the concept "exactly one blue object" might help with task two but is essentially useless for task one.

Taken together, this illustrates why relying solely on linear classifiers in CBMs is often impractical. To address such cases, non-linear classifiers should also be considered.

C. Experimental Details: Baselines

In this section, we provide training details for the considered baselines: *ResNet-18*, *ResNet-50*, *Pixel-MAC*, *CBM* (linear and nonlinear variants), and *DCR*.

C.1. ResNet-18 and ResNet-50

We initialize the framework with a pretrained ResNet-18 model and employ the Adam optimizer across all experiments. On the CLEVR datasets, the model is trained with a batch size of 128 for 30 epochs using a learning rate of 10^{-4} and weight decay of 10^{-4}, repeated across 20 random seeds with early stopping based on validation loss.

On CIFAR-100, we use a ResNet-50 trained for 100 epochs with a learning rate of 10^{-4}, weight decay of 10^{-5}, and a batch size of 128, averaged over 10 random seeds with early stopping. The ResNet-50 baseline on ImageNet is evaluated directly using pretrained PyTorch (Paszke et al., 2019) weights without further finetuning. For COCOLogic, a ResNet-50 is trained for 300 epochs with a batch size of 256, learning rate of 10^{-4}, and weight decay of 10^{-2}, again averaged over 10 random seeds with early stopping.

In the Pixel-MAC setup, a separate ResNet-18 is trained under the same configuration as above but with a reduced learning rate of 10^{-5}, while keeping the batch size, weight decay, and early stopping criterion unchanged. All Pixel-MAC results are obtained from these ResNet-18 checkpoints.

C.2. Pixel-MAC

In this setup, we apply Merlin-Arthur training on pixel space by utilizing the pretrained ResNet-18 models as classifiers and U-Net architectures for both Merlin and Morgana. Throughout all experiments, we employ the Adam optimizer for both classifier and feature selector optimization, with $\gamma = 0.5$ to ensure high soundness.

For the CLEVR datasets, we train with a batch size of 128 for 40 epochs, using a learning rate of 10^{-5} and weight decay of 10^{-6} for the classifier optimization. The U-Net architectures are trained with a learning rate of 10^{-4}, weight decay of 10^{-5} and an L1 penalty coefficient of 0.1. We set the mask size to 1500, meaning that the U-Nets select a subset of 1500 pixels per sample (out of 128×128 pixels).

For CIFAR-100, we reduce the batch size to 64 and train for 100 epochs. The classifier is optimized with a learning rate of 10^{-5} and weight decay of 10^{-4}. Both Merlin and Morgana are trained using a learning rate of 10^{-3} and a reduced mask size of 32 pixels. We use an L1 penalty of 0.01 and apply early stopping across 10 random seeds.

For ImageNet and COCOLogic, we further reduce the batch size to 32 due to memory constraints and train for 80 epochs using the same learning rates and hyperparameters as in the CIFAR-100 setting. The mask size is set to 1000 pixels per image. As with CIFAR-100, early stopping is applied across 10 random seeds, and training is initialized from the pretrained ResNet-18 backbone.

C.3. CBM with Linear Classifier

Next, we present the implementation details for the CBM baseline, where a linear classifier operates on concept features obtained from the concept extractor.

For the CLEVR datasets, we train a linear classifier on concepts extracted by the Neural Concept Binder. The training process employs a batch size of 128, a learning rate of 10^{-3}, and weight decay of 10^{-4}. The model is trained for 60 epochs on CLEVR-Hans3 and 30 epochs on CLEVR-Hans7, using early stopping based on validation loss, repeated across 20

Table 4. Comparison between DCR and NCV at 10k-concept vocabulary size on CLIP-Sim. NCV metrics correspond to completeness (accuracy). For COCOLogic-10, we report balanced accuracy due to class imbalance.

Dataset	DCR (acc. %)	**NCV** (ours) (completeness, %)
CIFAR-100	51.76	**83.32**
ImageNet-1k	n/a (timeout)	**67.04**
COCOLogic-10 (bal.)	38.61	**75.42**

different random seeds.

For CIFAR-100 and ImageNet, we train a linear classifier on sparse SpLiCE encodings using a dictionary size of 10,000. In both cases, we use a batch size of 4096 and train for 250 epochs with early stopping, a learning rate of 10^{-3}, and no weight decay. A hidden layer with 512 units is used, and an L1 penalty of 0.2 is applied within SpLiCE to encourage sparsity in the concept representations. All results are averaged over 10 random seeds.

C.4. DCR Baseline Implementation

For the additional nonlinear CBM comparison in Sec. 4, we implemented Deep Concept Reasoning (DCR; Barbiero et al. 2023) following the official `torch-explain` library. Our implementation uses the same experimental setup as NCV to ensure a fair comparison.

Architecture. We employ a ResNet-18 backbone (pretrained weights disabled for consistency) followed by two DCR-specific modules: (1) a `ConceptEmbedding` layer that maps backbone features to concept predictions $\mathbf{c}_{\text{pred}} \in [0,1]^C$ and learned concept embeddings $\mathbf{c}_{\text{emb}} \in \mathbb{R}^{C\times d}$, and (2) a `ConceptReasoningLayer` that learns class-specific logic rules in Disjunctive Normal Form (DNF) over the concept embeddings to produce class predictions.

Concept Supervision. We supervise the concept predictions using the same CLIP-Sim similarity vectors as NCV. Since these are cosine similarities in $[-1, 1]$, we apply a linear normalization $c_{\text{truth}} = (\text{sim} + 1)/2$ to map them to $[0, 1]$ for BCE loss compatibility.

Training. Following the official DCR tutorial, we optimize a joint loss $\mathcal{L} = \mathcal{L}_{\text{concept}} + \lambda \cdot \mathcal{L}_{\text{task}}$, where both terms use binary cross-entropy (BCE). The task loss uses one-hot encoded labels, as the `ConceptReasoningLayer` outputs per-class probabilities (not logits). We train with Adam optimizer for 100 epochs.

Hyperparameter Sweep. For each dataset, we performed a grid search over:

- Vocabulary size: $n \in \{1000, 3000, 10000\}$
- Concept embedding dimension: $d \in \{8, 16, 32\}$
- Learning rate: $\eta \in \{10^{-3}, 10^{-4}\}$
- Task loss weight: $\lambda \in \{0.5, 1.0\}$

This resulted in $3 \times 3 \times 2 \times 2 = 36$ configurations per dataset, totaling over 100 training runs. For COCOLogic-10, we report balanced accuracy due to class imbalance.

Results. Tab. 4 summarizes the best accuracies achieved by DCR compared to NCV at 10k vocabulary size. DCR successfully trains on CIFAR-100 and COCOLogic-10 but fails to complete training on ImageNet-1k within reasonable time ($>$ 3 days per epoch). In contrast, NCV's sparsity-inducing mechanism enables efficient scaling to all three datasets under the same 10k-concept space.

C.5. Comparison with Variational Information Pursuit (V-IP)

Variational Information Pursuit (V-IP) (Chattopadhyay et al., 2023) is closely related to NCV in its information-theoretic motivation: both methods aim to identify a small set of features that are maximally informative about the label. In this section we discuss the methodological differences and provide a representative numerical comparison on CIFAR-100.

Method. V-IP starts from an empty history and iteratively selects the next query whose answer is expected to be most informative about the label, conditional on the history collected so far. The query strategy and a classifier are parameterised by neural networks and trained end-to-end to minimise the expected cross-entropy between the true and predicted label. At inference time, V-IP issues queries sequentially until a confident prediction can be made, yielding an interpretable decision chain of variable length per input. In its original form, queries are realised as image patches (*e.g.*, 8×8 patches with stride 4 on CIFAR, yielding 49 candidate queries). Subsequent extensions also consider semantic queries from task-specific vocabularies.

NCV differs in two main ways. First, NCV is *single-shot*: Merlin selects a sparse concept subset in one step and Arthur predicts from this subset, rather than building up a sequential history. Second, NCV operates on a fixed, general-purpose concept vocabulary derived from CLIP/LAION captions or extracted unsupervised by NCB, rather than on raw image patches or hand-crafted semantic queries. These differences make the two methods complementary rather than directly substitutable: V-IP produces an interpretable decision chain whose length adapts per input, while NCV produces a fixed-size, inspectable concept subset.

CIFAR-100 comparison. On CIFAR-100, which overlaps with our experiments, V-IP attains approximately 60% classification accuracy after 7 patch queries on the 8×8 stride-4 vocabulary (cf. Figure 5b in Chattopadhyay et al. (2023)). The CLIP-based instantiation of NCV on CIFAR-100 reaches 83.32% completeness with a single-shot concept selection of fixed size 32 (Tab. 4). We stress that this is not a controlled head-to-head experiment: V-IP and NCV operate on different feature spaces (pixel patches vs. concept similarities) and emit different forms of explanation (sequential queries vs. a sparse concept subset).

Scaling considerations. The size of V-IP's query set scales with the number of candidate features. For pixel-patch queries on 224×224 images with $1{,}000$ classes (*e.g.*, ImageNet-1k), the candidate set grows substantially and individual patches carry less label information, which makes pure pixel-level V-IP challenging in this regime. Semantic-query variants of V-IP avoid this issue but typically rely on a hand-crafted, task-specific query set. NCV instead reuses the same general-purpose concept vocabulary across datasets and delegates the difficult selection problem to the Merlin–Morgana–Arthur game, which scales to ImageNet-1k under the same setup used for CIFAR-100.

D. NCB-based Neural Concept Verifier Experiments

In the following, we provide details on NCB-based NCV, experimental evaluations as well as additional evaluations.

D.1. Pretraining

Before training NCV, we first pretrain the models without the feature selectors. The corresponding results for the pretraining are shown in Tab. 5, where we evaluate on 20 random seeds. These pretrained models are then used as initialization for the subsequent NCV training. For the pretraining, we use a Set Transformer with two stacked multi-head attention blocks, a hidden dimension of 128 and four attention heads. We use a batch size of 128, 30 epochs and the Adam optimizer with a learning rate of 10^{-3} for both datasets, applying early stopping based on validation loss.

Table 5. Pretraining results on the CLEVR-Hans3 and CLEVR-Hans7 datasets without shortcuts

CLEVR-Hans3		**CLEVR-Hans7**	
Val. Accuracy	Test Accuracy	Val. Accuracy	Test Accuracy
99.02 ± 0.31	98.13 ± 0.37	98.08 ± 0.24	97.83 ± 0.25

D.2. NCV Training

For the experiments presented in our main results in Tab. 1, the experimental details for both datasets are as follows:

Model Architecture. The verifier is implemented as a pretrained Set Transformer consisting of two stacked multi-head attention blocks with hidden dimension 128, four attention heads, and layer normalization. Merlin and Morgana are implemented as independent neural networks, each parameterized by a Set Transformer with two stacked attention blocks with hidden dimensions 256, four attention heads, and layer normalization. The provers receive the full concept slot matrix

Table 6. Completeness and soundness on the CLEVR-Hans3 dataset without shortcuts for different mask sizes and feature selector architectures. The highlighted values are used for Table 1.

		Validation		Test	
Mask Size	**Feature Selector**	**Completeness**	**Soundness**	**Completeness**	**Soundness**
4	Set Transformer	98.35 ± 0.31	99.88 ± 0.28	97.69 ± 0.63	99.82 ± 0.23
	MLP	96.51 ± 1.18	99.81 ± 0.27	95.54 ± 1.37	99.85 ± 0.28
6	Set Transformer	98.71 ± 0.52	99.87 ± 0.13	98.11 ± 0.62	99.88 ± 0.19
	MLP	96.21 ± 0.89	99.96 ± 0.03	94.78 ± 1.24	99.97 ± 0.06
12	Set Transformer	99.20 ± 0.11	100.00 ± 0.00	$\mathbf{98.92 \pm 0.32}$	$\mathbf{100.00 \pm 0.00}$
	MLP	99.28 ± 0.11	99.98 ± 0.06	98.89 ± 0.21	99.99 ± 0.07

Table 7. Completeness and soundness on the CLEVR-Hans7 dataset without shortcuts for different mask sizes and feature selector architectures. The highlighted values are used for Table 1.

		Validation		Test	
Mask Size	**Feature Selector**	**Completeness**	**Soundness**	**Completeness**	**Soundness**
4	Set Transformer	96.69 ± 1.28	99.93 ± 0.09	96.71 ± 1.37	99.91 ± 0.09
	MLP	92.63 ± 1.24	99.89 ± 0.12	92.71 ± 1.31	99.87 ± 0.13
6	Set Transformer	97.32 ± 0.42	99.98 ± 0.02	97.14 ± 0.51	99.98 ± 0.02
	MLP	95.43 ± 1.48	99.88 ± 0.13	95.12 ± 1.48	99.86 ± 0.14
12	Set Transformer	98.13 ± 0.11	100.00 ± 0.00	$\mathbf{97.89 \pm 0.31}$	$\mathbf{100.00 \pm 0.00}$
	MLP	97.41 ± 1.07	99.99 ± 0.03	97.01 ± 0.93	99.99 ± 0.04

as input and output a sparse selection mask with exactly 12 nonzero entries (out of 64 total features), indicating the active blocks provided to the verifier.

Training Details. All components are jointly trained using the Adam optimizer with a learning rate of 10^{-3} and weight decay of 10^{-4}. Models are trained for 50 epochs and a batch size of 512 is used throughout. For the Merlin and Morgana provers, a hard selection constraint is enforced, limiting the number of selected concepts to a fixed budget of 12 block-encodings per sample. To ensure high soundness, we set $\gamma = 0.5$, giving equal weight to both feature selector losses in the total loss computation. We train our models using 20 random seeds.

Extended Results. Additionally, we evaluated the NCV framework with varying mask sizes and an alternative model architecture for the feature selectors. The results are presented in Tab. 6 for the CLEVR-Hans3 dataset and Tab. 7 for the CLEVR-Hans7 dataset, where we evaluate both the validation set and the test set. The alternative architecture implements a MLP with two hidden layers and ReLU activation functions for the feature selectors, while maintaining a pretrained Set Transformer as the classifier across all experiments. Our results reveal that the Set Transformer feature selector consistently outperforms the MLP feature selector on the test set, particularly with smaller mask sizes such as 4 and 6. Furthermore, this configuration maintains high completeness (>96%) and soundness (>99%), even with a reduced number of selected features.

D.3. Explanations

Here, we present supplementary examples of explanations generated by both Pixel-MAC and NCV on the CLEVR-Hans3 dataset in Fig. 4. These results further substantiate our claim that NCV provides significantly more transparent and interpretable explanations compared to the pixel-based PVG baseline.

E. Experimental details for CLIP-based NCV

In the following section, we present the implementation details of CLIP-based NCV training.

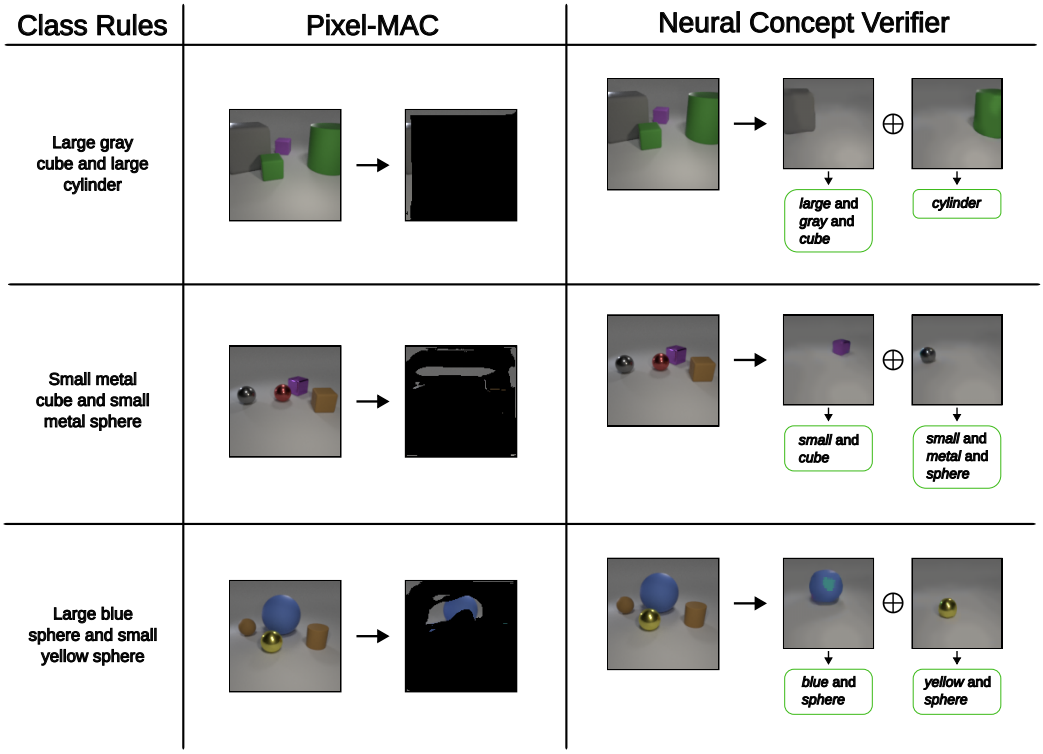


Figure 4. Comparison of explanations from NCV *vs.* Pixel-MAC for CLEVR-Hans3 images of all three classes. **(a)** Merlin–Arthur training on pixel space yields uninformative masks. **(b)** NCV provides clear explanations by highlighting object features corresponding to the class rule. The single-object images are reconstructions from the respective slots selected by Merlin (prover).

Table 8. Pretraining accuracy of the verifier (without provers) for CLIP-based NCV on CIFAR-100, COCOLogic and ImageNet-1k.

Dataset	Accuracy (%)
CIFAR-100	85.96
COCOLogic	81.39
ImageNet-1k	77.07

E.1. Pretraining

Once more, before starting with the actual NCV training, we first pretrain the models without the provers (Merlin and Morgana). The corresponding results for the pretraining are shown in Tab. 8. As textual concept descriptions T, we used the top 10,000 most frequent one- and two-word phrases from LAION (Schuhmann et al., 2021) captions, following the setup of Bhalla et al. (2024). For pretraining the verifier, we use a two-layer multilayer perceptron (MLP) with a hidden dimension of 512 and GELU activations (Hendrycks & Gimpel, 2016) on CIFAR-100, ImageNet-1k, and COCOLogic. On CIFAR-100 and ImageNet, training uses a batch size of 4096 and a learning rate of 10^{-4}, with dropout (0.3), weight decay of 10^{-4}, and early stopping (patience 10). On COCOLogic, we instead train for 100 epochs with a batch size of 512, learning rate of 10^{-4}, and weight decay of 10^{-2}, using a learning-rate scheduler (plateau, patience 5, factor 10^{-3}, minimum learning rate 10^{-6}) and no early stopping. All pretraining is conducted without provers, and the resulting verifiers are used to initialize the CLIP-based NCV training.

E.2. NCV Training

For the experiments presented in our main results in Tab. 1, we detail the training setup separately for CIFAR-100 and ImageNet.

Model Architecture. The verifier (Arthur) is initialized as the pretrained two-layer multilayer perceptron (MLP) described

Table 9. Ablation results for our method on CIFAR-100 with varying mask sizes. We report mean ± std over 10 seeds.

Mask Size	Completeness Train	Completeness Validation	Soundness Train	Soundness Validation
4	$74.97_{\pm 4.45}$	$69.32_{\pm 2.66}$	$99.85_{\pm 0.07}$	$99.85_{\pm 0.09}$
8	$87.42_{\pm 2.60}$	$71.20_{\pm 22.43}$	$99.95_{\pm 0.03}$	$99.95_{\pm 0.03}$
16	$94.08_{\pm 1.46}$	$81.82_{\pm 0.50}$	$99.96_{\pm 0.03}$	$99.97_{\pm 0.02}$
64	$97.65_{\pm 0.47}$	$84.01_{\pm 0.31}$	$100.00_{\pm 0.00}$	$100.00_{\pm 0.00}$

Table 10. Ablation results for our method on ImageNet with varying mask sizes. We report mean ± std over 10 seeds.

Mask Size	Completeness Train	Completeness Validation	Soundness Train	Soundness Validation
4	$59.30_{\pm 0.34}$	$55.96_{\pm 0.30}$	$99.81_{\pm 0.03}$	$99.83_{\pm 0.06}$
8	$64.84_{\pm 0.41}$	$60.98_{\pm 0.30}$	$99.94_{\pm 0.01}$	$99.94_{\pm 0.03}$
16	$68.85_{\pm 0.34}$	$64.60_{\pm 0.12}$	$99.96_{\pm 0.03}$	$99.96_{\pm 0.03}$
64	$73.35_{\pm 0.39}$	$69.03_{\pm 0.18}$	$99.97_{\pm 0.01}$	$99.97_{\pm 0.02}$

above. Merlin and Morgana are implemented as independent neural networks, each parameterized by a two-layer MLP with hidden dimension 512 and ReLU activations. Both provers receive the full concept activation vector as input and output a sparse selection mask indicating the active concepts that are passed to the verifier.

CIFAR-100. For CIFAR-100, all components are trained jointly for 100 epochs using the Adam optimizer. We set the verifier learning rate to 10^{-4}, and use 5×10^{-4} for both Merlin and Morgana. A batch size of 256 is used throughout. Weight decay is set to 0.1, and a hard mask size of 32 concepts is enforced per input. To incentivize sparse masks, an L1 penalty of 0.1 is applied to the provers. A learning rate scheduler (plateau-based) is employed with a patience of 5, minimum learning rate of 10^{-6}, and decay factor of 0.001. Early stopping is disabled, and all results are averaged over 10 random seeds.

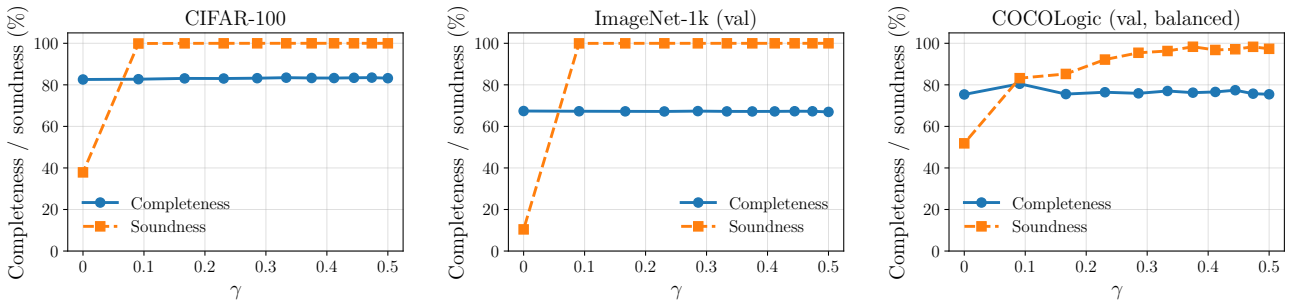
ImageNet. The ImageNet setup mirrors CIFAR-100 in most aspects. We again train for 100 epochs with a batch size of 256, using the same learning rates for verifier (10^{-4}) and provers (5×10^{-4}), mask size of 32 features, and L1 penalty (0.1). Weight decay is reduced to 0.005 to improve generalization. The same learning rate scheduler and seed setup are used as in the CIFAR-100 experiments.

COCOLogic. Training on COCOLogic follows the CIFAR-100 configuration with minor adjustments: models are trained for 100 epochs with a batch size of 512, verifier learning rate of 10^{-4}, and prover learning rates of 5×10^{-4}. We use a weight decay of 0.01, a mask size of 32, and an L1 penalty of 0.1 on the provers. As with CIFAR-100, early stopping is disabled, and learning rate scheduling and seed averaging remain unchanged.

E.3. Prover and Verifier Architectures

In all experiments, the choice of prover and verifier architectures is guided by the structure of the concept representation. For unordered, slot-based concept encodings (as in NCB on CLEVR-Hans), we use permutation-invariant Set Transformers for Merlin, Morgana, and Arthur. For fixed-size concept vectors (as in CLIP/SpLiCE on CIFAR-100, ImageNet-1k, and COCOLogic-10), we use shallow nonlinear MLPs.

In preliminary experiments, we observed that performance is stable across a range of nonlinear architectures (varying depth, width, and activations), whereas purely linear models consistently underperformed by collapsing to simple correlation tests. Based on this, we recommend using permutation-invariant architectures for slot-based concepts and nonlinear MLPs for vector-based concepts when applying NCV to new datasets.



(a) CIFAR-100 (b) ImageNet-1k (validation) (c) COCOLogic (validation, balanced metrics)

Figure 5. Effect of the weighting parameter γ on completeness and soundness for (a) CIFAR-100, (b) ImageNet-1k validation, and (c) COCOLogic validation (balanced metrics). All curves show means over 3 random seeds.

E.4. Effect of the weighting parameter γ

The weighting parameter γ controls the trade-off between Merlin's cooperative objective and Morgana's adversarial objective when training Arthur. Recall that Arthur's loss is given by

$$L_A = (1-\gamma)\, L_M + \gamma\, L_{\widehat{M}},$$

so that $\gamma = 0$ corresponds to training Arthur only on Merlin's loss, and larger γ increases the relative weight of Morgana's adversarial objective. To assess the influence of this trade-off on NCV, we sweep γ over a range from 0 (Merlin-only) up to 0.5 (substantial weight on Morgana) on CIFAR-100, ImageNet-1k, and COCOLogic, keeping all other hyperparameters fixed. For each setting, we train three models with different random seeds and report the mean completeness and soundness; for COCOLogic, we report balanced metrics due to class imbalance.

Across all datasets the same qualitative behavior emerges. When $\gamma = 0$, i.e., the verifier is trained without adversarial pressure from Morgana, completeness remains high but soundness collapses (e.g., 37.9% on CIFAR-100, 10.4% on ImageNet-1k, and 51.8% balanced soundness on COCOLogic). As soon as $\gamma > 0$, soundness rapidly recovers to values close to those reported in Tab. 1, while completeness remains essentially unchanged throughout the range of γ we consider. These results demonstrate that incorporating adversarial concept selection is crucial for learning verifiers that remain reliable under misleading concept subsets, without sacrificing standard predictive performance in the regimes we study. For a detailed discussion of why an adversarial prover and its relative strength are essential for the Merlin–Arthur mutual-information guarantees underlying NCV, see Suppl. A.

F. Computational Cost and Hardware Setup

To complement the main results, we report approximate training times for all models and datasets considered in Sec. 4. All runs were executed on a single NVIDIA A100 or A40 GPU; times are reported as rounded wall-clock estimates and are meant to convey relative cost rather than exact benchmarks. For CBM baselines, we *exclude* the per-sample SpLiCE optimization step at inference time, which would add substantial overhead and further increase their deployment cost.

Table 11. Approximate training time comparison across models and datasets. Times are rounded wall-clock estimates on a single NVIDIA A100 or A40 GPU. For CBM baselines, the per-sample SpLiCE optimization at inference time is omitted here, but would incur significant additional cost. (ImgNet = ImageNet-1k, COCO-10 = COCOLogic-10, CL-H3/CL-H7 = CLEVR-Hans3/7.)

Model	**CIFAR**	**ImgNet**	**COCO-10**	**CL-H3**	**CL-H7**
ResNet (baseline)	∼25–30m	∼1.5d	∼50m	$<$ 10m	$<$ 10m
Pixel-MAC	∼1.3d	∼3d	∼4h	$<$ 2h	$<$ 2h
CBM	∼10–15m	∼4h	∼2m	∼2m	∼2m
NCV (ours)	∼20m (11m A + 7m PVG)	∼5h (2.5h A + 2.5h PVG)	∼5m	∼1.5m	∼5m

Overall, NCV is 1–3 orders of magnitude cheaper to train than Pixel-MAC across all datasets, while remaining comparable

to standard CBMs in runtime. Since concept encodings are precomputed once and reused across models, NCV scales similarly to a conventional classifier without additional architectural overhead, and remains practical even in large-scale settings such as ImageNet-1k.

G. Use of Large Language Models

Large language models were used to support this work by assisting with text refinement, implementation of code components (including methods and plot generation), and by providing input during idea development and approach refinement.