VOCABULARY ALIGNED EXPLANATION FOR COMPUTER VISION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Optimizing vision models for accuracy does not force the model to learn and utilize human-understandable and domain-specific concepts and rules. We propose explaining image classifications through domain-specific concepts, visual evidence, and expert knowledge expressed as logical rules. We introduce FLUENT (Framework for Localized, Uncertainty-Aware Explanations in Normative Terminology), a post-hoc framework that ranks expert-authored rules combining jointly required, alternative, and absent concepts. To account for prediction errors, FLUENT combines predicted labels and concept states with empirically estimated likelihoods and concept priors. It scores each rule by calculating support over satisfying assignments to the concepts mentioned in that rule, then normalizes these scores to express relative confidence among candidate explanations. The selected rule is illustrated through occlusion-based highlights for positive concept mentions in the query image and localized examples in retrieved reference images for negative mentions. These illustrations are intended to show end users unfamiliar with the domain what the concepts named in an explanation look like. FLUENT combines domain knowledge and model accuracy, and delivers visual concept highlights and explicit rule confidence to explain an image in the vocabulary of an application domain.

1 INTRODUCTION

Deep vision models can achieve high classification accuracy in visual tasks (He et al., 2016; Woo et al., 2023), but label scores do not explain why a specific class was chosen. Optimizing classification accuracy does not require a model to express its predictions through concepts familiar to its users. This has led to a proliferation of highly capable but effectively black-box vision models.

Explainable artificial intelligence (XAI) methods were developed to explicitly address this shortcoming. Methods such as Grad-CAM or Integrated Gradients identify influential image regions (Selvaraju et al., 2017; Sundararajan et al., 2017), while LIME fits local, interpretable approximations of prediction behavior (Ribeiro et al., 2016). These methods can be used by AI experts to inspect where a model looks, but a highlighted region can contain several attributes; highlights alone do not specify what matters, whether the presence or absence of a feature supports the class label, or how different features interact with each other. As such, the usefulness of these methods for end-users have been under debate (Kim et al., 2018).

To illustrate how these methods differ from human-generated explanations, consider an entomologist trying to explain a bee's identity to a non-expert (Figure 1 provides a visual overview). They set out with a set of domain-specific vocabulary or concepts, denoting relevant attributes (body hair, color, stripes, etc.). The identity of the bee itself is determined by rules defined in terms of the absence or presence of these individual concepts. For example, the picture in Figure 1 corresponds to a green metallic bee because it has a green head and thorax, has pollen in its hind leg, doesn't have a fuzzy body, and doesn't have dense hair in its hind legs. As part of the explanation, the expert would present the rule to the non-expert, along with examples that would teach them the individual concepts. For positive concepts (*e.g.* green head and thorax), the expert could rely on the current image, and for negative ones they might have to rely on images from alternative species that do exhibit that attribute.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

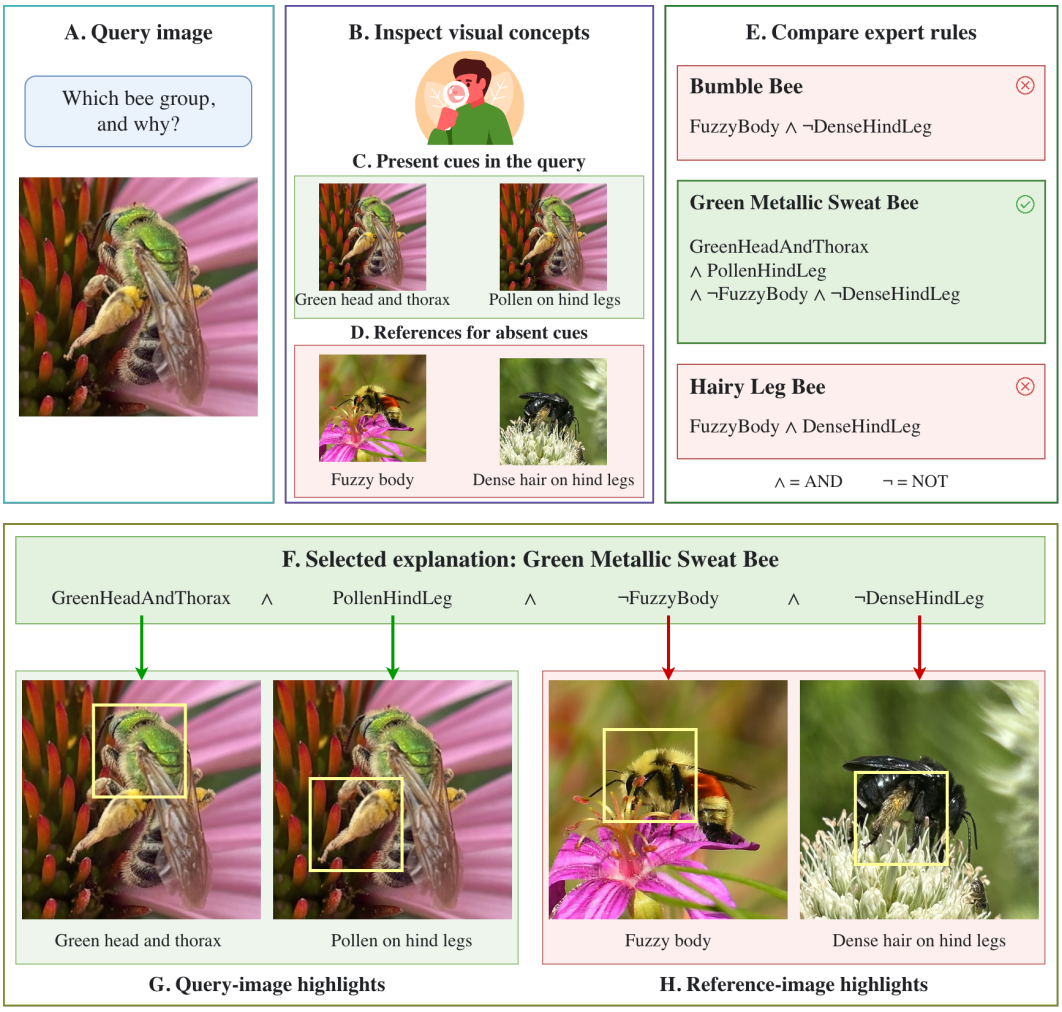

Figure 1: How a human expert would explain a bee image to someone in the domain who may be unfamiliar with its concepts and their appearance. **(A–C)** The expert introduces the concepts and points out green coloration and pollen on the hind legs in the query. **(D)** To explain fuzzy body hair and dense hind-leg hair, which are absent here, the expert shows reference images containing them. **(E–F)** The expert connects these observations through the Green Metallic Sweat Bee rule. Here $\wedge$ means "and" and $\neg$ means "not." **(G–H)** Highlights guide the recipient to the relevant concepts in the query and references. This human explanation process is the running example for FLUENT; Section 3 defines its automated procedure.

While there exists some works that can support some steps of the explanatory workflow described above, no single work can completely replicate it. Prominently, concept-based XAI methods introduce an interpretable vocabulary: TCAV measures concept sensitivity (Kim et al., 2018), and Concept Bottleneck Models (CBMs) predict through explicit concepts (Koh et al., 2020). Uncertainty-aware CBMs already model uncertainty in concept predictions (Kim et al., 2023), and structured prototype methods can reason with evidence for both presence and absence (Nauta et al., 2021; Ayoobi et al., 2025). However, the primary missing piece is a mechanism that can support explanation that leverages complex logical rules over concepts. Such a method should, for a given input, identify the most likely rule that can explain it. Present the rule to the user and then help the user understand the individual concepts used within the rule.

We introduce FLUENT, a post-hoc framework for explaining images in this manner using learned label and concept predictions. FLUENT takes an expert-defined concept set and class-associated propositional rules as inputs. It trains concept predictors on labeled examples and estimates the

error behavior of the label and concept predictors from labeled evaluation data. Independent estimation data are desirable; the bee splits do not fully meet this condition. The rule scorer operates on label and concept predictions and their estimated error rates, without requiring a particular backbone architecture or feature dimension. Our experiments use pretrained backbones with separately trained label and concept heads. Learned backbone parameters were intended to remain fixed, but inspected BatchNorm buffers differ from the pretrained checkpoints; Appendix C.2 distinguishes fixed parameters from fixed inference state.

For a query image, FLUENT combines the classifier's predicted label with predictions about which domain concepts are present or absent to identify the best-supported domain rule. It accounts for the measured error rates of the label and concept predictors, weighing conflicting evidence according to their reliabilities. The selected rule is accompanied by a confidence score expressing its support relative to other candidate rules. The rule provides an interpretation of the predicted evidence using domain criteria; it is not a reconstruction of the classifier's internal decision process. We evaluate the accuracy of the classifier's label and the rule-implied label separately.

Following the expert's demonstration in Figure 1, FLUENT pairs the selected rule with visual examples to show users what its concepts look like. For each positively mentioned concept, it highlights the query patch whose occlusion yields the lowest concept-predictor score. For negative concepts, it retrieves the most similar reference image containing that concept and highlights it there. In the bee example, query highlights show green coloration and pollen, while references show fuzzy body and dense hind-leg hair. These reference images illustrate the concepts named via negative literals.

Succinctly, our contributions are: (I) A post-hoc framework that explains images through domain-specific concepts and class-associated logical rules using features from a pretrained vision backbone. (II) A rule-scoring procedure that incorporates measured label and concept prediction errors and reports normalized support for competing explanations. (III) A visual grounding procedure that links positive concept mentions to query-image highlights and negative mentions to localized reference examples. (IV) An evaluation across 12 backbones and two benchmarks, examining rule recovery, rule-implied classification, concept-supervision budgets, and pretraining choices. We evaluate FLUENT across 12 pretrained vision backbones on a synthetic and a real dataset (Appendix D).

2 RELATED WORK

Saliency-based methods attribute predictions to input features by measuring pixel or region importance. Class saliency and class activation methods (Simonyan et al., 2014; Selvaraju et al., 2017; Wang et al., 2020) highlight discriminative regions through class-specific gradients or forward-pass masking, while LIME (Ribeiro et al., 2016) uses perturbations to fit interpretable local approximations around individual predictions. Propagation and axiomatic methods, including Integrated Gradients (Sundararajan et al., 2017) and LRP (Bach et al., 2015), attribute relevance using reference inputs or conservation rules. Attribution maps can be fragile: small, label-preserving perturbations can substantially alter saliency (Ghorbani et al., 2019a), and sanity checks find some methods insensitive to model parameters or training data (Adebayo et al., 2018). Connecting heatmaps to meaningful concepts requires human interpretation, limiting their usefulness for structured domain criteria and reasoning (Zhang et al., 2021).

Concept-based methods bridge this gap with human-interpretable concepts. TCAV (Kim et al., 2018) measures class-score sensitivity to user-defined concept directions in activation space, while ACE (Ghorbani et al., 2019b) reduces manual effort by discovering concepts through segmentation and clustering. Vision-language approaches associate internal units with natural-language descriptions, as in CLIP-Dissect (Oikarinen & Weng, 2023), or construct language-guided concept bottlenecks, as in LaBo (Yang et al., 2023). These methods provide concept vocabularies or scores; our explanation space instead consists of a supplied set of expert-authored logical rules.

Concept Bottleneck Models (CBMs) (Koh et al., 2020) route predictions through an explicit concept layer, enabling concept-level interventions for human correction at test time. Variants relax supervision or improve concept representations (Yuksekgonul et al., 2023; Oikarinen et al., 2023; Espinosa Zarlenga et al., 2022), while uncertainty-aware ProbCBMs (Kim et al., 2023) propagate concept distributions into downstream decisions. CBMs typically learn concept-to-label mappings

from data rather than encoding expert rules, and often require retraining or architectural constraints that may be undesirable for explaining off-the-shelf backbones.

Prototype networks explain decisions via similarity to prototypical parts or training instances. ProtoPNet (Chen et al., 2019) learns class-specific prototype patches for localized explanations; variants improve compactness through prototype sharing and compositional reasoning (Rymarczyk et al., 2021; Nauta et al., 2021). Recent work further targets human alignment and modern backbones (Nauta et al., 2023; Ma et al., 2024). ProtoArgNet (Ayoobi et al., 2025) reasons over prototypical evidence through structured argumentation, approaching rule-like explanations. ProtoTree and ProtoArgNet already support structured reasoning, including evidence against a prediction. We also rely on visual examples to ground concepts for users, but they aren't a complete explanation.

Contrastive and counterfactual methods ask how input changes would alter a decision (Wachter et al., 2018; Goyal et al., 2019), with diffusion-based generation (Augustin et al., 2022) and semantically constrained steering (Jacob et al., 2022) improving realism. Pertinent negatives (Dhurandhar et al., 2018) highlight features whose absence is decisive, a form of reasoning also relevant to expert rules. Negative evidence therefore has precedents; our method illustrates negatively mentioned domain concepts with reference examples.

We focus on combining uncertainty-aware selection among expert-defined logical rules with visual explanations of their concepts. The selected rule specifies how concept evidence is combined, while query-image highlights and reference examples communicate the rule's positively and negatively mentioned concepts, respectively.

3 FLUENT METHOD

FLUENT combines separately trained label and concept predictors with expert-authored rules. It ranks rules using empirical prediction-error likelihoods, normalizes their support, and illustrates positive and negative concept mentions through query and reference highlights.

3.1 LABELS, CONCEPTS AND EXPERT RULES

We formalize the labels, concepts, and rules that the expert uses to explain the bee image to the recipient in Figure 1. Let $\mathcal{L} = \{l_1, l_2, \ldots, l_n\}$ be the class labels and $\mathcal{X} \subseteq \mathbb{R}^{h_{\text{img}} \times w_{\text{img}} \times 3}$ the modeled domain of images, where h_{img} and w_{img} are the image height and width in pixels and the last dimension contains the three color channels. We train a domain-specific classifier head on features from a pretrained backbone, yielding the label classifier $\phi_{\text{cls}} : \mathcal{X} \to \mathcal{L}$. For an image $x \in \mathcal{X}$, its predicted label is $\hat{\ell}(x) := \phi_{\text{cls}}(x)$.

Domain experts specify the visual concepts used in the rules, such as green coloration of the head and thorax or pollen on the hind legs in Figure 1. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_k\}$ be this concept vocabulary. We treat each concept c_i as a binary proposition about an image. For any image x, let $c_i(x) \in \{0, 1\}$ denote the *concept indicator* that concept c_i is present in x: $c_i(x) := \mathbf{1}\big[\, c_i \text{ is present in } x \,\big] \in \{0, 1\}$. We will use $\mathbf{C}(x)$ to denote the *concept vector* for image x: $\mathbf{C} : \mathcal{X} \to \{0, 1\}^k; \mathbf{C}(x) := \big(c_1(x), \ldots, c_k(x)\big)$. We use lowercase $\mathbf{c} \in \{0, 1\}^k$ for a possible value of this vector.

We use propositional logic to combine present and absent concepts into expert rules. Absence of a concept matters, *e.g.* under the supplied benchmark rule, a Bumble Bee must *not* have dense hair on its hind legs. Rules express concept presence and absence using literals. In propositional logic, a *literal* is either a concept c_i or its negation $\neg c_i$, meaning that the concept is absent. Let $\text{Lit}(\mathcal{C}) = \{c_1, \ldots, c_k, \neg c_1, \ldots, \neg c_k\}$ be the set of literals. FLUENT assumes experts use *propositional logical rules* to express why an image belongs to a class, in terms of which concepts are present or absent, *e.g.* "fuzzy hair on the body *and* dense hair on the hind legs" justifies Hairy Leg Bees. Such explanations combine several concepts, and might also include alternatives for a concept. We express each explanation as a formula over literals. The scoring definition accepts a rule r in conjunctive normal form (CNF), that is, a conjunction of clauses, where each clause is a disjunction of literals: $r = \bigwedge_{j=1}^{m} \big(\bigvee_{p=1}^{s_j} \ell_{j,p} \big), \quad \ell_{j,p} \in \text{Lit}(\mathcal{C})$.

Here $\wedge$ means AND and $\vee$ means OR: every clause must hold, and at least one literal must hold within each clause. For example, $(c_1 \vee c_2) \wedge \neg c_3$ is CNF and requires either c_1 or c_2, together with

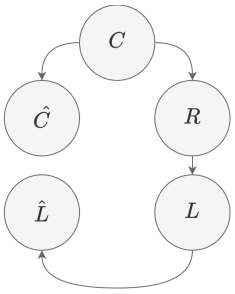

Figure 2: Bayesian model for rule selection. True concepts C determine a rule R and its label L; the observed predictions $\hat{C}$ and $\hat{L}$ are modeled through concept and label prediction likelihoods. Implemented scores use only the concept predictions mentioned by each candidate rule.

the absence of c_3. We write $\mathbf{c} \models r$ when assigning concepts the values in $\mathbf{c}$ makes rule r true. We write $\ell \sqsubseteq r$ when literal ℓ occurs in the written rule; this denotes occurrence, not logical implication.

A class may have several rules describing different qualifying concept configurations. For each class label $l \in \mathcal{L}$, experts provide a non-empty set of rules $\mathcal{R}_l$. Let $\mathcal{R}_{expert} = \bigcup_{l \in \mathcal{L}} \mathcal{R}_l$ be the set of all expert-authored rules. We assume that each supplied rule is assigned to only one class, so the rule sets are disjoint across classes: $\forall\, l_i \neq l_j \in \mathcal{L} :\ \mathcal{R}_{l_i} \cap \mathcal{R}_{l_j} = \varnothing$.

Based on this assumption, each rule belongs to exactly one class. We write $\ell(r)$ for the unique label such that $r \in \mathcal{R}_{\ell(r)}$. We also assume mutual exclusivity: no two distinct rules can be simultaneously satisfied by the same concept vector: $\forall\, r \neq r' \in \mathcal{R}_{expert},\ \forall\, \mathbf{c} \in \{0,1\}^k :\ \mathbf{1}[\mathbf{c} \models r] + \mathbf{1}[\mathbf{c} \models r'] \leq 1$. We also assume *coverage*: every image in the domain satisfies a supplied rule. Together with mutual exclusivity, this means exactly one rule holds for each image:

$$\forall x \in \mathcal{X} : \quad \sum_{r \in \mathcal{R}_{expert}} \mathbf{1}[\mathbf{C}(x) \models r] = 1. \tag{1}$$

The evaluated rule sets use a narrower grammar: positive unit literals or positive disjunctions, together with unit negative literals. The explanation renderer is only supported for this grammar; it does not preserve arbitrary clauses mixing positive and negative literals. The general scoring definition does not imply a validated general-CNF renderer.

3.2 The Explanation Problem

Given an image x, the explanation problem is to recover the unique supplied rule satisfied by its true concept vector $\mathbf{C}(x)$ (Equation (1)). Since the true concepts are unknown at inference time, FLUENT scores candidate rules using predicted labels and concept states.

3.3 Concept Predictors

For each concept c_i, we train a predictor $\phi_{c_i} : \mathcal{X} \to [0,1]$ to infer its presence in a new image. Its score is used for localization, while its binary prediction $\hat{c}_i(x)$ is used for rule inference; we collect these predictions in $\hat{\mathbf{C}}(x) = (\hat{c}_1(x), \ldots, \hat{c}_k(x))$. Writing $\Pr$ for probability, we estimate each predictor's conditional likelihoods $\Pr(\hat{c}_i \mid c_i)$ from its prediction counts on the designated concept-labeled test set (Table 2). For $h, t \in \{0,1\}$, $\Pr(\hat{c}_i = h \mid c_i = t)$ is the probability that concept predictor i outputs h when the true concept state is t. The estimation procedure and conditional-likelihood table are given in Appendix A.1.

3.4 Selecting the Best Rule via Probabilistic Modeling

The label and concept predictors can make errors, so rule selection combines their observed outputs with empirical estimates of their reliability. Figure 2 gives the probabilistic reference model: the true concept vector C determines the satisfied rule R, which determines its class L. The predicted concepts $\hat{C}$ depend on C, and the predicted label $\hat{L}$ depends on L. These dependencies separate the logical rule from the prediction errors used to score it. At inference time, given an image x we observe $\hat{C} = \hat{\mathbf{C}}(x)$ and $\hat{L} = \hat{\ell}(x)$. Below, lower-case values denote assignments to these random variables, and we suppress x when writing the observed predictions.

Let $\Omega = \{\mathbf{c} \in \{0,1\}^k : \exists r \in \mathcal{R}_{expert},\ \mathbf{c} \models r\}$ be the covered concept assignments, and let $A = \{C \in \Omega\}$ be the coverage event. Mutual exclusivity ensures that each assignment in Ω satisfies exactly one rule. We first approximate concept prevalence by an independent product $q(\mathbf{c})$, with $\Pr(c_i = 1) = \pi_i$ and $\Pr(c_i = 0) = 1 - \pi_i$, where π_i is estimated from concept metadata on a sample of the label-training data. The reference model uses this product conditioned on A; conditioning can make the concepts dependent. Its assumptions are:

$$q(\mathbf{c}) = \prod_{i=1}^{k} \Pr(c_i) \qquad \text{(I) independent base prior,} \tag{2}$$

$$\Pr(\hat{\mathbf{C}} \mid \mathbf{c}) = \prod_{i=1}^{k} \Pr(\hat{c}_i \mid c_i) \qquad \text{(II) separate concept channels,} \tag{3}$$

$$\Pr(\hat{\ell} \mid l, r, \mathbf{c}, \hat{\mathbf{C}}) = \Pr(\hat{\ell} \mid l) \qquad \text{(III) label-only error channel,} \tag{4}$$

$$\Pr(r \mid \mathbf{c}, A) = \mathbf{1}[\mathbf{c} \models r], \quad \mathbf{c} \in \Omega \qquad \text{(IV) deterministic rule,} \tag{5}$$

$$\Pr(l \mid r) = \mathbf{1}[l = \ell(r)] \qquad \text{(V) deterministic label.} \tag{6}$$

Assumption (II) means each concept prediction depends only on its own true concept state; (III) means the label prediction depends only on the true class. The concept likelihoods come from concept-test prediction counts (Appendix A.1), and the label likelihoods from the Label Test confusion matrix. These independence assumptions are modeling approximations, not consequences of the predictors sharing a backbone. Appendix A.2 derives the posterior that conditions on all concept predictions under this reference model.

The implemented rule score uses a candidate-specific subset of these predictions. Define the *mentioned concept set* as $\mathcal{I}_\rho(r) := \{i \in \{1, \ldots, k\} : c_i \sqsubseteq r \ \vee\ \neg c_i \sqsubseteq r\}$. Rule satisfaction depends only on the true concepts with indices in $\mathcal{I}_\rho(r)$. For scoring, we additionally choose to use only their predictions $\hat{\mathbf{C}}_r = (\hat{c}_i)_{i \in \mathcal{I}_\rho(r)}$, omitting the prediction likelihoods for other concepts. This choice defines the score used in the experiments; it is not an algebraic simplification of the full-evidence posterior. The unnormalized score is

$$w(r, x) = \Pr(\hat{\ell} \mid \ell(r)) \sum_{\mathbf{c}_{\mathcal{I}_\rho(r)} \in \{0,1\}^{|\mathcal{I}_\rho(r)|}} \mathbf{1}[\mathbf{c}_{\mathcal{I}_\rho(r)} \models r] \prod_{i \in \mathcal{I}_\rho(r)} \Pr(\hat{c}_i \mid c_i) \Pr(c_i). \tag{7}$$

The indicator equals one when the assignment satisfies every clause in r, and zero otherwise. Thus, the sum accounts for all satisfying assignments to the mentioned concepts, weighted by their prevalence and the likelihood of the observed predictions. The label factor accounts for how often a true instance of class $\ell(r)$ receives the observed label $\hat{\ell}$. We normalize these weights over the supplied rules and select the largest:

$$\mathrm{Conf}(r, x) = \frac{w(r, x)}{\sum_{r' \in \mathcal{R}_{expert}} w(r', x)}, \qquad r^*(x) = \operatorname*{arg\,max}_{r \in \mathcal{R}_{expert}} \mathrm{Conf}(r, x). \tag{8}$$

When the denominator is positive, the confidences sum to one and express each rule's share of the total score. A two-concept rule and a four-concept rule use different subsets of the predictions, so these normalized scores are not posterior probabilities based on the same observed evidence. Appendix A.2 gives the exact difference from the full posterior and the implemented fallback when all weights are zero.

3.5 Highlighting Concepts

After selecting $r^*(x)$, FLUENT illustrates its concepts as the expert does for the recipient in Figure 1(G–H), helping people unfamiliar with the concepts connect their names to visible features. Positive mentions receive query-image highlights; negative mentions receive highlights in reference images containing the concept. Define the positive and negative occurrence sets as

$\mathcal{I}_\rho^+(r) := \{i : c_i \sqsubseteq r\}$ and $\mathcal{I}_\rho^-(r) := \{i : \neg c_i \sqsubseteq r\}$. For $i \in \mathcal{I}_\rho^+(r^*(x))$, we highlight the query-image patch whose occlusion produces the lowest concept-predictor score. For $i \in \mathcal{I}_\rho^-(r^*(x))$, we retrieve a similar reference image containing c_i and highlight that concept there. These displays illustrate the literals mentioned in the rule. In a clause such as $c_1 \vee c_2$, only one alternative must hold, so highlighting both concepts does not assert that both are present.

3.5.1 Highlighting Positive Concepts

We localize a positively mentioned concept by testing how masking different image patches changes its predictor's score. For clarity, consider a single $N \times N$ grid of non-overlapping patches in $x \in \mathcal{X} \subseteq \mathbb{R}^{h_{\text{img}} \times w_{\text{img}} \times 3}$. Here h_{img} and w_{img} are the image height and width, distinct from the predictor notation ϕ_{c_i}. Choose N that divides both dimensions, and set the patch height $p_h = h_{\text{img}}/N$ and width $p_w = w_{\text{img}}/N$. Using integer pixel coordinates (u, v) starting at zero, patch (a, b) contains the coordinates: $P_{a,b} = \{(u, v) : (a-1)p_h \le u < ap_h,\ (b-1)p_w \le v < bp_w\}, \quad a, b \in \{1, \ldots, N\}$.

Let $x_{u,v} \in \mathbb{R}^3$ denote the color vector at pixel (u, v). To test patch (a, b), we replace its pixels by black, producing $x_{u,v}^{(a,b)} = x_{u,v}\, \mathbf{1}[(u, v) \notin P_{a,b}]$. We evaluate ϕ_{c_i} on each occluded image and select the patch attaining the lowest score. Keeping that patch and masking the rest defines the highlight $x_{c_i}^*$: $(a^*, b^*) = \arg\min_{a,b \in \{1,\ldots,N\}} \phi_{c_i}(x^{(a,b)}),\ (x_{c_i}^*)_{u,v} = x_{u,v}\, \mathbf{1}[(u, v) \in P_{a^*,b^*}]$.

3.5.2 Highlighting Negative Concepts

A reference image can illustrate a concept that the selected rule mentions negatively, as in Figure 1(H). For each such concept c_i, FLUENT uses a reference dataset containing that concept: $\mathcal{D}_{c_i}^{ref} = \{x_j\}_{j=1}^{M_{c_i}}, \quad c_i(x_j) = 1\ \forall j$. For each $\neg c_i$ with $i \in \mathcal{I}_\rho^-(r^*(x))$, FLUENT retrieves the most similar image $x_{c_i}^\sim \in \mathcal{D}_{c_i}^{ref}$ to image x using cosine similarity in the backbone feature space: $x_{c_i}^\sim = \arg\max_{x_j \in \mathcal{D}_{c_i}^{ref}} \text{similarity}(x, x_j)$. In FLUENT, the similarity function is defined as: $\text{similarity}(x, x_j) = \frac{z \cdot z_j}{\|z\| \|z_j\|}$. Here, $z = f_\lambda(x)$ and $z_j = f_\lambda(x_j)$ where $f_\lambda : \mathcal{X} \to \mathbb{R}^{d_\lambda}$ maps an image to its d_λ-dimensional feature vector at backbone layer λ (typically the penultimate layer). After retrieving $x_{c_i}^\sim$, FLUENT applies the same occlusion procedure to illustrate c_i in that reference image. We denote the resulting highlight in the reference image by $x_{c_i}^{\sim *}$. We collect visual highlights into: $\mathcal{H}^+(x) := \{x_{c_i}^* : i \in \mathcal{I}_\rho^+(r^*(x))\},\ \mathcal{H}^-(x) := \{x_{c_i}^{\sim *} : i \in \mathcal{I}_\rho^-(r^*(x))\}$. These two sets correspond to the query and reference highlights illustrated in Figure 1(G–H). This cosine comparison assumes nonzero feature norms. A faithful implementation must preserve the original query feature vector across all occlusion passes. The qualitative example illustrates the display and does not establish the correctness of the reference ranking (Appendix A.3).

3.6 Explanation Tuple

FLUENT combines the predicted label, selected rule, normalized rule score, and two sets of visual highlights into the explanation tuple: $\mathcal{EXP}(x) = \Big(\hat{\ell}(x), r^*(x), \text{Conf}(r^*(x), x), \mathcal{H}^+(x), \mathcal{H}^-(x)\Big)$. Here $\hat{\ell}(x) = \phi_{\text{cls}}(x)$ is the classifier label, $r^*(x)$ is the selected rule, $\text{Conf}(r^*(x), x)$ is its normalized score, and $\mathcal{H}^+(x)$ and $\mathcal{H}^-(x)$ illustrate its positive and negative concept mentions, respectively. A complete synthetic explanation is shown in Figure 3 in Appendix A.3.

4 Experiments

We report experiments spanning 12 pretrained vision backbones with separately trained label heads and linear concept predictors. The evaluation measures agreement with supplied rules. The experiments use one configured seed and have the split-independence limitations described below.

Benchmarks. The synthetic benchmark has three classes, 16 concepts, and six rules; generated concept metadata determine the unique target rule. The evaluation does not isolate concept evidence from possible background shortcuts. The bee benchmark has three groups, four concepts, and one rule per class. Its Explanation Test concept vectors are class-constant templates, without indepen-

dent per-image verification of visible concepts. Thus bee RuleAcc measures agreement with those templates, not verified recovery of visible concepts. Rules and splits appear in Appendix B.

Protocol. The Explanation Test sets contain 300 synthetic and 600 bee images. The *synthetic reference configuration* uses $T{=}5000$ label-training images per class, $C{=}500$ positive and 500 negative training examples per concept, and prior-estimation sample fraction $P{=}0.1$. Thus C counts examples per polarity per concept, and P controls the fraction of label-training data sampled to estimate concept prevalence. The *bee reference configuration* uses fixed dataset splits and $P{=}0.1$; synthetic T and C budgets do not apply to bees. Appendices D and E give all configurations and separately identified additional evaluations. Likelihood estimates use the Label Test and Concept Test confusion matrices. Appendix C details the experimental settings, inference-state checks, and reproducibility limitations.

Metrics. LabelAcc measures classifier accuracy, RuleAcc measures agreement with the target rule, and RuleLabelAcc measures accuracy of the selected rule's implied label. Selecting the other rule of a correct synthetic class is an error only for RuleAcc; the metrics coincide on bees because each class has one rule. ConceptAcc averages binary accuracy on the Concept Test sets. Avg Conf averages normalized rule support and does not measure calibration. Full definitions are in Appendix C.5.

5 RESULTS

Table 1 reports all five evaluation metrics for the same four representative backbones on both datasets, covering classic and modern convolutional models, plain and hierarchical transformers. Complete results for all 12 backbones are in Appendices D and E. The main findings are that rule recovery depends on concept quality as well as class prediction, and that the effect of rule selection on label accuracy differs across datasets. Concept evidence improves rule-implied labels for the weak synthetic ResNet classifiers, whereas rule-implied labels are less accurate than classifier predictions for every backbone on bees.

Table 1: Four representative backbones in the reference configurations: $T{=}5000$, $C{=}500$, $P{=}0.1$ for the synthetic benchmark; fixed dataset splits and $P{=}0.1$ for bees. Accuracies are percentages; Avg Conf is unitless. ConceptAcc averages accuracy on the concept test sets; the other metrics use the Explanation Test splits (300 synthetic and 600 bee images). Bold marks column maxima among the displayed models within each dataset. Bee values have the overlap limitations in Appendix C.4. Full results appear in Appendix Tables 8 and 24.

Backbone	Label Acc	Concept Acc	Rule Acc	RuleLabel Acc	Avg Conf
Synthetic					
RN50 (TV)	39.0	79.9	69.7	85.3	0.634
CNXv2-L	99.3	93.3	93.7	**99.7**	0.925
ViT-B/16 (DINO)	**99.7**	92.8	**96.7**	**99.7**	0.935
Swin-B (IN-22k)	99.0	**95.2**	94.7	99.0	**0.949**
Bees					
RN50 (TV)	91.2	85.5	87.8	87.8	0.770
CNXv2-L	**99.7**	96.4	85.8	85.8	**0.967**
ViT-B/16 (DINO)	97.7	96.2	84.3	84.3	0.927
Swin-B (IN-22k)	**99.7**	**97.7**	**98.2**	**98.2**	0.947

5.1 SYNTHETIC BENCHMARK

Rule recovery varies across backbones. Under the synthetic reference configuration, ViT-B/16 with DINO achieves the highest rule accuracy, 96.7%. Accurate class predictions alone are insufficient: ConvNeXt V2-L with supervised fine-tuning reaches 93.7% rule accuracy, compared with 54.0% for its FCMAE-only variant, although both exceed 94% label accuracy. Their mean concept accuracies differ substantially (93.3% versus 56.9%). Comparisons of pretraining variants are descriptive, not controlled causal estimates. Complete configuration results are in Appendix D.

Rule-implied labels can correct classifier errors. For ResNet-50 (TV), combining concept and label evidence raises rule-implied label accuracy to 85.3% from the classifier's 39.0%. Across the three ResNet backbones, gains are 26.7–46.3 percentage points under the synthetic reference configuration. These gains concern the label inferred from the selected rule; FLUENT retains the original classifier prediction. For the nine backbones with at least 90% classifier accuracy, the two label accuracies differ by at most 0.7 points under the synthetic reference configuration. Appendix D.3 provides every configuration.

5.2 BEE BENCHMARK

Strong rule recovery does not guarantee label correction. Under the bee reference configuration, rule agreement with the class-derived templates on field photographs ranges from 43.8% to 98.2%, with Swin-B (IN-22k) performing best. Unlike the synthetic setting, rule-implied label accuracy is below the original classifier's accuracy for every backbone, by 1.5–21.8 percentage points. For ConvNeXt V2-L, the gap is 99.7% label accuracy versus 85.8% rule accuracy. Thus, concept evidence can also introduce errors. Concept accuracy measured on curated exemplars need not transfer to field photographs, where cues may be occluded, viewpoint-dependent, or transient. Appendix E gives all backbone results and per-concept accuracies.

Confidence describes relative support, not validated reliability. Mean normalized confidence correlates with rule accuracy across backbones (Spearman ρ=0.92 under the synthetic reference configuration and 0.64 under the bee reference configuration, each across 12 backbones), but these aggregate associations do not establish calibration or individual-error detection. ConvNeXt V2-L has mean confidence 0.967 on bees despite only 85.8% rule accuracy, illustrating that decisive scores can accompany errors. These results do not justify using an unvalidated confidence threshold to present explanations. We report complete results and additional analyses in Appendices D and E.

6 DISCUSSION, LIMITATIONS AND FUTURE WORK

FLUENT combines expert-authored logical rules, empirical label and concept prediction-error likelihoods, and visual examples of positively and negatively mentioned concepts. The selected rule interprets predicted evidence through the supplied vocabulary; it does not reconstruct the label classifier's internal reasoning. The results show that rule-implied labels can be more accurate than weak synthetic label heads: the three ResNets gain 26.7–46.3 percentage points under the synthetic reference configuration. On bees, the direction reverses for every backbone, with gaps of 1.5–21.8 points. These are descriptive comparisons under the evaluated data and estimation procedure. Mean rule confidence is associated with aggregate rule accuracy (Spearman $\rho = 0.92$ for the synthetic reference configuration and 0.64 for the bee reference configuration, each across 12 backbones), but this does not establish calibration or an error-detection threshold.

Experimental limitations. The bee likelihood-estimation pool contains every explanation-evaluation image, concept training and test sources overlap, and observation families span the label splits. Its rule targets are class-derived templates without independent per-image concept verification. Inspected BatchNorm inference buffers differ from pretraining despite unchanged comparable learned parameters. One configured seed, incomplete training details, and uncontrolled possible background cues limit causal and generalization claims. Appendix C.4 details these limitations.

Method and explanation limitations. The scorer assumes mutually exclusive rules covering the modeled domain and does not detect uncovered inputs. Its factorized error channels and prior are approximations. Candidate-specific evidence subsets make the normalized score different from a posterior conditioned on all observations and can favor shorter rules. The evaluated renderer supports positive disjunctions and unit negative literals, not arbitrary signed CNF. Occlusion highlights are sensitive to patch size and do not certify concept presence or anatomical boundaries.

Future work. A new evaluation could run user studies then assess understanding and task performance against other explanation methods like saliency, free-text, LLM-based and no-explanation conditions. Segmentation methods such as SAM 3 (Carion et al., 2026) are candidates for localization, but their effect requires measurement. Overlapping rules and further domains with independently annotated concepts remain extensions.

AI USE STATEMENT

In this work, we used generative AI tools to assist with manuscript editing, readability, formatting, reference-related tasks, interpreting results and implementation of the methods in code. We did not use generative AI tools to generate the expert-authored domain knowledge or rules used in the experiments. We did not use generative AI to directly generate the synthetic dataset, design the method or experiments. We have reviewed all AI-assisted work. LLM-generated code was verified and tested for correctness by authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity Checks for Saliency Maps. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion Visual Counterfactual Explanations. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 364–377. Curran Associates, Inc., 2022. doi: 10.52202/068431-0027. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/025f7165a452e7d0b57f1397fed3b0fd-Paper-Conference.pdf.

Hamed Ayoobi, Nico Potyka, and Francesca Toni. ProtoArgNet: Interpretable Image Classification with Super-Prototypes and Argumentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):1791–1799, Apr 2025. ISSN 2159-5399. doi: 10.1609/aaai.v39i2.32173. URL http://dx.doi.org/10.1609/aaai.v39i2.32173.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10(7):1–46, 07 2015. doi: 10.1371/journal.pone.0130140. URL https://doi.org/10.1371/journal.pone.0130140.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment Anything with Concepts, 2026. URL https://arxiv.org/abs/2511.16719.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9650–9660, October 2021.

Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Cynthia Rudin, and Jonathan K Su. This Looks Like That: Deep Learning for Interpretable Image Recognition. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/adf7ee2dcf142b0e11888e72b43fcb75-Paper.pdf.

Amit Dhurandhar, Pin-Yu Chen, Ronny Luss, Chun-Chen Tu, Paishun Ting, Karthikeyan Shanmugam, and Payel Das. Explanations based on the Missing: Towards Contrastive Explanations with Pertinent Negatives. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/c5ff2543b53f4cc0ad3819a36752467b-Paper.pdf.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lió, and Mateja Jamnik. Concept Embedding Models: Beyond the Accuracy-Explainability Trade-Off. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 21400–21413. Curran Associates, Inc., 2022. doi: 10.52202/068431-1555. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/867c06823281e506e8059f5c13a57f75-Paper-Conference.pdf.

Amirata Ghorbani, Abubakar Abid, and James Zou. Interpretation of Neural Networks Is Fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688, July 2019a. ISSN 2159-5399. doi: 10.1609/aaai.v33i01.33013681. URL http://dx.doi.org/10.1609/aaai.v33i01.33013681.

Amirata Ghorbani, James Wexler, James Zou, and Been Kim. Towards Automatic Concept-based Explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019b. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual Visual Explanations. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2376–2384. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/goyal19a.html.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked Autoencoders Are Scalable Vision Learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, June 2022.

Paul Jacob, Éloi Zablocki, Hédi Ben-Younes, Mickaël Chen, Patrick Pérez, and Matthieu Cord. STEEX: Steering Counterfactual Explanations with Semantics. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner (eds.), *Computer Vision – ECCV 2022*, pp. 387–403, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19775-8.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory sayres. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 10–15 Jul 2018. URL https://proceedings.mlr.press/v80/kim18d.html.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic Concept Bottleneck Models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 16521–16540. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kim23g.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept Bottleneck Models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10012–10022, October 2021.

Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, Furu Wei, and Baining Guo. Swin Transformer V2: Scaling Up Capacity and Resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12009–12019, June 2022.

Chiyu Ma, Jon Donnelly, Wenjun Liu, Soroush Vosoughi, Cynthia Rudin, and Chaofan Chen. Interpretable Image Classification with Adaptive Prototype-based Vision Transformers. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 41447–41493. Curran Associates, Inc., 2024. doi: 10.52202/079017-1312. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/48dfc849640344e2d58df0b5bb78c33b-Paper-Conference.pdf.

Meike Nauta, Ron van Bree, and Christin Seifert. Neural Prototype Trees for Interpretable Fine-Grained Image Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14933–14943, June 2021.

Meike Nauta, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. PIP-Net: Patch-Based Intuitive Prototypes for Interpretable Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2744–2753, June 2023.

Tuomas Oikarinen and Tsui-Wei Weng. CLIP-Dissect: Automatic Description of Neuron Representations in Deep Vision Networks. *International Conference on Learning Representations*, 2023.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free Concept Bottleneck Models. In *International Conference on Learning Representations*, 2023.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pp. 1135–1144. ACM, Aug 2016. doi: 10.1145/2939672.2939778. URL http://dx.doi.org/10.1145/2939672.2939778.

Dawid Rymarczyk, Łukasz Struski, Jacek Tabor, and Bartosz Zieliński. ProtoPShare: Prototypical Parts Sharing for Similarity Discovery in Interpretable Image Classification. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, KDD '21, pp. 1420–1430. ACM, Aug 2021. doi: 10.1145/3447548.3467245. URL http://dx.doi.org/10.1145/3447548.3467245.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Workshop at International Conference on Learning Representations*, 2014.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic Attribution for Deep Networks. In Doina Precup and Yee Whye Teh (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 3319–3328. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/sundararajan17a.html.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. SigLIP 2: Multilingual Vision-Language Encoders with Improved Semantic Understanding, Localization, and Dense Features, 2025. URL https://arxiv.org/abs/2502.14786.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 31(2):841–887, 2018. URL https://jolt.law.harvard.edu/assets/articlePDFs/v31/Counterfactual-Explanations-without-Opening-the-Black-Box-Sandra-Wachter-et-al.pdf.

Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, and Xia Hu. Score-CAM: Score-Weighted Visual Explanations for Convolutional Neural Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2020.

Ross Wightman. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models, 2019.

Ross Wightman, Hugo Touvron, and Hervé Jégou. ResNet strikes back: An improved training procedure in timm, 2021. URL https://arxiv.org/abs/2110.00476.

Sanghyun Woo, Shoubhik Debnath, Ronghang Hu, Xinlei Chen, Zhuang Liu, In So Kweon, and Saining Xie. ConvNeXt V2: Co-Designing and Scaling ConvNets With Masked Autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16133–16142, June 2023.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a Bottle: Language Model Guided Concept Bottlenecks for Interpretable Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19187–19197, June 2023.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc Concept Bottleneck Models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

Yu Zhang, Peter Tino, Ales Leonardis, and Ke Tang. A Survey on Neural Network Interpretability. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 5(5):726–742, Oct 2021. ISSN 2471-285X. doi: 10.1109/tetci.2021.3100641. URL http://dx.doi.org/10.1109/TETCI.2021.3100641.

A Additional Method Details and Explanation Example

This appendix describes concept-likelihood estimation, derives the full-concept reference posterior and its relation to the implemented score, and gives an example explanation tuple.

A.1 Estimating Concept-Prediction Likelihoods

The conditional likelihood $\Pr(\hat{c}_i = h \mid c_i = t)$ describes how often concept predictor i outputs h when the true concept state is t. It differs from the image-specific presence score $\phi_{c_i}(x)$: the former summarizes prediction behavior over labeled examples, whereas the latter is the predictor's score for one image.

For concept c_i, let $\mathcal{D}_{c_i}^{test} = \{(x_j, t_j)\}_{j=1}^{N_{c_i}}$ be the designated concept-labeled evaluation set, with $t_j = c_i(x_j) \in \{0, 1\}$. In the experiments, the binary output is $\hat{c}_i(x_j) = \mathbf{1}[\phi_{c_i}(x_j) \geq 0.5]$. We count true and predicted states using

$$n_i(t, h) = \sum_{j=1}^{N_{c_i}} \mathbf{1}[t_j = t]\, \mathbf{1}[\hat{c}_i(x_j) = h], \qquad t, h \in \{0, 1\}. \tag{9}$$

For a true state represented in the evaluation set, the empirical estimate is

$$\Pr(\hat{c}_i = h \mid c_i = t) \;\approx\; \frac{n_i(t, h)}{n_i(t, 0) + n_i(t, 1)}. \tag{10}$$

The denominator counts all evaluation examples with true state t. The numerator counts those for which the predictor outputs h. Thus, each row of Table 2 is normalized separately. For example, $\Pr(\hat{c}_i = 1 \mid c_i = 1)$ is estimated by $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$, while $\Pr(\hat{c}_i = 1 \mid c_i = 0)$ is estimated by $\mathrm{FP}/(\mathrm{TN} + \mathrm{FP})$.

For a complete numerical specification, an empty true-state row is assigned probability 0.5 for each output, and estimated concept likelihoods are clipped to $[10^{-12}, 1 - 10^{-12}]$ without pseudocounts. These conditional likelihoods supply the concept-prediction factors in Equation (7). The same clipping convention applies to concept-prevalence parameters.

Table 2: Conditional likelihoods for a binary concept predictor, estimated from a designated concept-labeled test set via the confusion matrix.

	$\hat{c}_i = 0$	$\hat{c}_i = 1$
$c_i = 0$	$\Pr(\hat{c}_i = 0 \mid c_i = 0)$	$\Pr(\hat{c}_i = 1 \mid c_i = 0)$
$c_i = 1$	$\Pr(\hat{c}_i = 0 \mid c_i = 1)$	$\Pr(\hat{c}_i = 1 \mid c_i = 1)$

A.2 Full-Concept Posterior and Implemented Rule Score

The coverage event $A = \{C \in \Omega\}$ restricts the reference model to concept assignments that satisfy a supplied rule. The independent product $q(\mathbf{c})$ from Equation (2) is a base approximation before this restriction. The covered prior is

$$\Pr(C = \mathbf{c} \mid A) = \frac{\mathbf{1}[\mathbf{c} \in \Omega]\, q(\mathbf{c})}{Z_A}, \qquad Z_A = \sum_{\mathbf{u} \in \Omega} q(\mathbf{u}) > 0.$$

Its concept marginals need not equal the fitted parameters π_i: conditioning on rule coverage can change their values and introduce dependence. This distinguishes the independent base approximation from the covered domain model.

Using the dependencies in Figure 2, the rule posterior given all predictions is

$$\begin{aligned} \Pr(r \mid \hat{\mathbf{C}}, \hat{\ell}, A) = \alpha_{\text{full}} \sum_{l \in \mathcal{L}} \sum_{\mathbf{c} \in \Omega} \Pr(\hat{\ell} \mid l)\, \Pr(l \mid r) \mathbf{1}[\mathbf{c} \models r] \\ \cdot \Pr(\hat{\mathbf{C}} \mid \mathbf{c}) q(\mathbf{c}). \end{aligned} \tag{11}$$

The normalizer α_{full} absorbs the common coverage constant Z_A and makes the posterior sum to one over rules. Because $\Pr(l \mid r)$ is zero for every $l \neq \ell(r)$, only $l = \ell(r)$ survives the class-label sum. Every assignment satisfying r is already in Ω, so

$$\Pr(r \mid \hat{\mathbf{C}}, \hat{\ell}, A) = \alpha_{\text{full}} \underbrace{\Pr(\hat{\ell} \mid \ell(r)) \sum_{\mathbf{c}:\mathbf{c}\models r} \prod_{i=1}^{k} \Pr(\hat{c}_i \mid c_i) \Pr(c_i)}_{w_{\text{full}}(r,x)}. \tag{12}$$

The corresponding full-evidence selection would be

$$r^*_{\text{full}}(x) = \arg\max_{r \in \mathcal{R}_{expert}} \Pr(r \mid \hat{\mathbf{C}}(x), \hat{\ell}(x), A). \tag{13}$$

This is a reference decision rule; the reported experiments use $r^*(x)$ from Equation (8).

To compare the two scores, define $a_i(t) = \Pr(\hat{c}_i \mid c_i = t) \Pr(c_i = t)$ and $m_i(x) = \sum_{t \in \{0,1\}} a_i(t)$. Since satisfaction of r depends only on concepts in $\mathcal{I}_\rho(r)$, summing over the other true concept states gives

$$w_{\text{full}}(r, x) = w(r, x) \prod_{i \notin \mathcal{I}_\rho(r)} m_i(x). \tag{14}$$

Each omitted factor is the marginal likelihood of an observed concept prediction under the independent base model. The product generally depends on the candidate rule, so it cannot be absorbed into a shared normalization constant. Consequently, the full posterior and the implemented score can rank rules differently.

Conditional independence allows this factorization; it does not make these observed likelihoods disappear. For a concept not mentioned by the rule, if its prediction were unobserved, summing over both its predicted value and its true state would contribute $\sum_t \Pr(c_i = t) \sum_h \Pr(\hat{c}_i = h \mid c_i = t) = 1$. The implemented procedure instead chooses which predictions to use separately for each candidate. Its confidence therefore expresses normalized relative support under that scoring choice. In particular, if every concept prediction is uninformative with likelihood $1/2$, the restricted score contains a factor $2^{-|\mathcal{I}_\rho(r)|}$ multiplying the rule's base-prior satisfaction mass and label likelihood. Rules mentioning different numbers of concepts can therefore receive different evidence-size penalties.

Equation (8) assumes a positive total weight. The numerical convention for zero total weight assigns the uniform value $1/|\mathcal{R}_{expert}|$ and selects the first candidate. Ties at the maximum weight are likewise resolved by candidate order. This fallback is a numerical convention, not additional evidence for the selected rule.

A.3 A Complete Synthetic Explanation

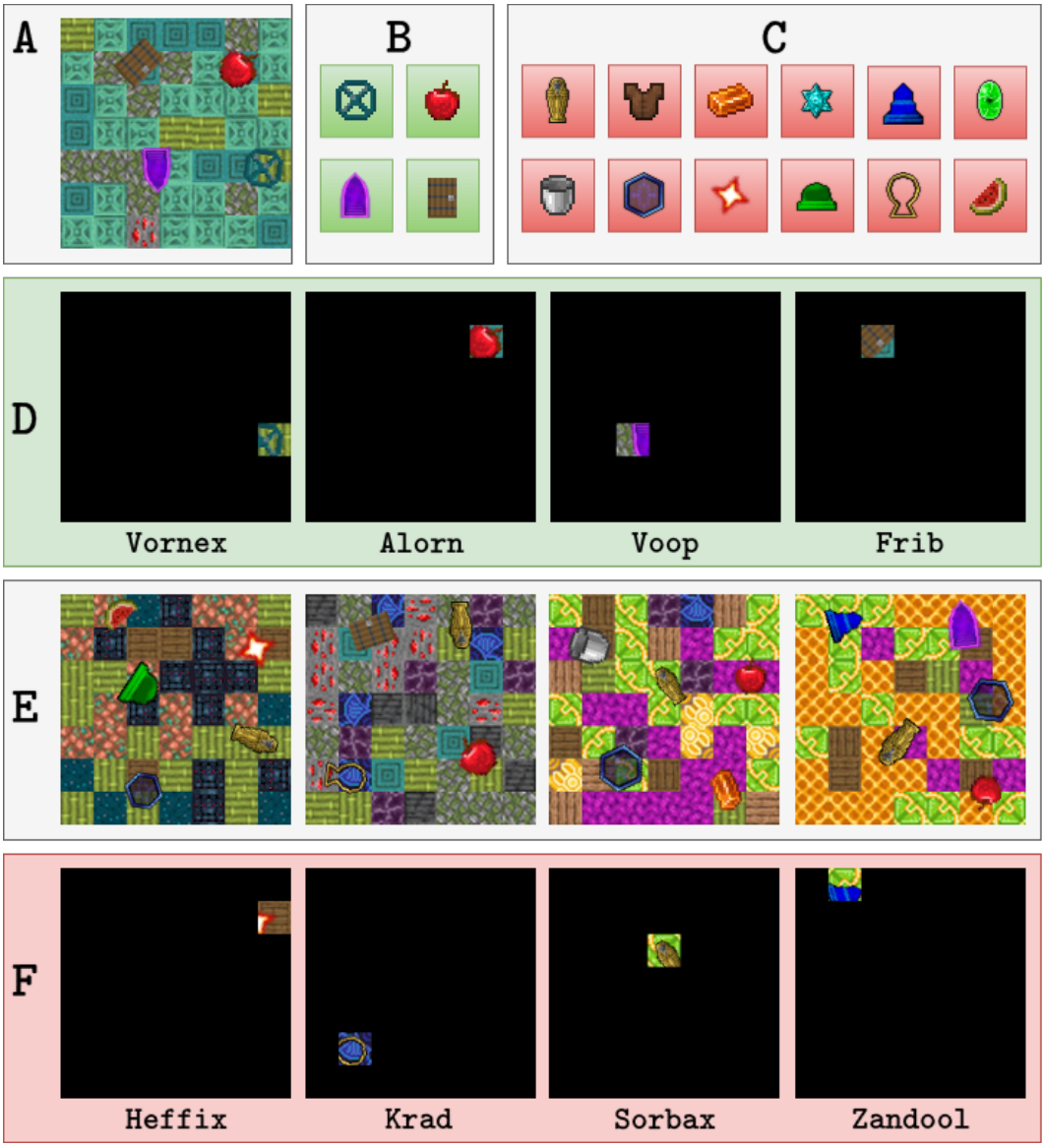

Figure 3: An illustrative FLUENT display for a synthetic *Kallow* image. This figure illustrates the explanation format; it does not establish a numerical confidence or nearest-reference ranking. **(A)** The input image containing multiple concept objects on a tiled background. **(B)** The four concepts predicted as present by the concept predictor: `Alorn`, `Vornex`, `Voop`, and `Frib`. **(C)** The twelve concepts predicted as absent. Rule: $\neg$`Heffix` $\wedge$ $\neg$`Krad` $\wedge$ $\neg$`Sorbax` $\wedge$ $\neg$`Zandool` $\wedge$ (`Alorn` $\vee$ `Vornex`) $\wedge$ (`Frib` $\vee$ `Voop`) is provided as the best explanation. **(D)** Displayed query highlights for the positively mentioned concepts. These panels illustrate the output format; they are not a quantitative localization validation. **(E)** Reference images containing each negatively mentioned concept. **(F)** Occlusion highlights for these concepts in the reference images.

B DATASETS, EXPERT RULES, AND SPLITS

We evaluate FLUENT on a synthetic dataset and a real-world bee dataset. Each dataset is described below; Appendix C presents the experimental protocol and evaluation metrics, and Section 5 summarizes the main results.

B.1 SYNTHETIC DATASET

We evaluate FLUENT on a purpose-built synthetic dataset in which ground-truth labels, concepts, and rules are known. This controlled setting lets us evaluate label predictions, concept predictions, and rule selection against known ground truth, without the annotation uncertainty of natural images.

The synthetic dataset comprises $|\mathcal{L}| = 3$ classes, $|\mathcal{C}| = 16$ abstract concepts, and $|\mathcal{R}_{expert}| = 6$ expert rules (two per class), listed in Table 3. Each concept is rendered as a distinct synthetic visual pattern, and images are composed by placing concept instances at random positions and scales on a tiled 224×224 background. Every ground-truth concept vector satisfies exactly its target rule. Exhaustive Boolean enumeration verifies pairwise mutual exclusivity (Section 3.1); coverage concerns the generated images, rather than every Boolean concept assignment (Equation (1)). Each rule mentions 8 of the 16 concepts. Of the 16 concepts, 12 participate in at least one rule and 4 (`Floom`, `Opple`, `Yintar`, and `Zorf`) appear in images but are irrelevant to classification.

Background choices and the generating software version are unspecified, and the evaluation does not isolate possible background shortcuts. Moreover, changing T changes the evaluation images as well as the training pool: for example, the Vumble rule-1 count is 33, 44, and 62 out of 100 Explanation Test images at $T = 500$, 2000, and 5000, respectively. Cross-T results therefore combine training-size and data-composition differences.

Table 3: CNF rules for the synthetic dataset (two rules per class). Each rule consists of two disjunctive clauses and four negated concepts.

Class	Expert Rule
Kallow	(`Alorn` ∨ `Vornex`) ∧ (`Krad` ∨ `Zandool`) ∧¬`Guffit` ∧ ¬`Smeek` ∧ ¬`Hebda` ∧ ¬`Wexil`
Kallow	(`Alorn` ∨ `Vornex`) ∧ (`Frib` ∨ `Voop`) ∧¬`Krad` ∧ ¬`Zandool` ∧ ¬`Heffix` ∧ ¬`Sorbax`
Vumble	(`Guffit` ∨ `Smeek`) ∧ (`Frib` ∨ `Voop`) ∧¬`Alorn` ∧ ¬`Vornex` ∧ ¬`Heffix` ∧ ¬`Sorbax`
Vumble	(`Guffit` ∨ `Smeek`) ∧ (`Heffix` ∨ `Sorbax`) ∧¬`Frib` ∧ ¬`Voop` ∧ ¬`Krad` ∧ ¬`Zandool`
Dadron	(`Hebda` ∨ `Wexil`) ∧ (`Heffix` ∨ `Sorbax`) ∧¬`Alorn` ∧ ¬`Vornex` ∧ ¬`Guffit` ∧ ¬`Smeek`
Dadron	(`Hebda` ∨ `Wexil`) ∧ (`Krad` ∨ `Zandool`) ∧¬`Guffit` ∧ ¬`Smeek` ∧ ¬`Heffix` ∧ ¬`Sorbax`

Table 4: Data splits used in the synthetic reference configuration (T=5000, C=500, P=0.1). The *Explanation Test* split provides ground-truth concept annotations and is used exclusively for evaluating the FLUENT rule selector (Section 3.4). The *Contrastive* split provides images for retrieval of negated concepts.

Split	Img./Class	Total	Used by
Label Train	5 000	15 000	ϕ_{cls} training
Label Test	1 000	3 000	ϕ_{cls} evaluation
Explanation Test	100	300	FLUENT benchmark
	Img./Concept	**Total**	
Concept Train	500 pos./neg. pairs	16 000	ϕ_{c_i} training
Concept Test	100 pos./neg. pairs	3 200	ϕ_{c_i} evaluation
Contrastive Examples	100	1600	$\mathcal{H}^-$ highlights

B.2 REAL BEE DATASET

The bee benchmark contains field photographs grouped into *Bumble Bee*, *Green Metallic Sweat Bee*, and *Hairy Leg Bee*. The six species are *Bombus griseocollis* and *Bombus huntii* (Bumble Bee), *Agapostemon virescens* and *Augochlorella aurata* (Green Metallic Sweat Bee), and *Anthophora urbana* and *Melissodes bimaculatus* (Hairy Leg Bee). The 600-image Explanation Test split contains 100 images per species and 200 per group.

The benchmark concept vectors are constant within each group: Bumble Bee has only `FuzzyBody`; Green Metallic Sweat Bee has `PollenHindLeg` and `GreenHeadAndThorax`; Hairy Leg Bee has `FuzzyBody` and `DenseHindLeg`. Prior estimation uses the same species-to-concept templates. Independent human verification of which cues are visible in each photograph is not established, and complete source-image attribution and collection details are unavailable. We therefore evaluate agreement with these class-associated templates; these labels do not establish independently annotated image-visible concepts. The concept-classifier test sets are separate curated positive/negative entries and have the overlap limitations described below.

The four concepts used in the experiment are *fuzzy hair on the entire body* (`FuzzyBody`), *dense hair on the hind legs* (`DenseHindLeg`), *pollen carried on the hind legs* (`PollenHindLeg`), and *a green head and thorax* (`GreenHeadAndThorax`). The supplied rule set has one CNF rule per class, shown in Table 5; these are task-specific rules, not universal species-identification criteria.

Table 5: CNF rules for the bee benchmark (one rule per class).

Class	**Expert Rule**
Bumble Bee	`FuzzyBody` $\wedge$ $\neg$`DenseHindLeg`
Green Metallic Sweat Bee	`PollenHindLeg` $\wedge$ $\neg$`FuzzyBody` $\wedge$ `GreenHeadAndThorax` $\wedge$ $\neg$`DenseHindLeg`
Hairy Leg Bee	`FuzzyBody` $\wedge$ `DenseHindLeg`

The three rules are pairwise mutually exclusive and cover 9 of the 16 Boolean assignments; each benchmark template satisfies its group's rule. They are not exhaustive over all Boolean concept assignments, consistent with the restricted coverage assumption in Equation (1). Table 6 gives the dataset counts. Concept entries are counted separately for each concept (positives plus negatives), so their totals need not equal the number of distinct source photographs. The intended augmentation factor is $8\times$, but the augmented training images total 32 928, rather than $4\,119 \times 8 = 32\,952$; three pairs of source files share a filename stem, consistent with the 24-entry shortfall.

Table 6: Counts for the bee experiment. "Test" names identify the dataset partitions; they do not imply independent evaluation (see the overlap analysis below). The outer Label Train count includes 3 643 validation entries. Contrastive counts exclude three non-image metadata files.

Split	**Img./Class**	**Total**	**Used by**
Label Train	24 271–24 301	72 873	ϕ_{cls} training
Label Test	2 698–2 729	8 125	ϕ_{cls} evaluation
Explanation Test	200	600	FLUENT benchmark
	Img./Concept	**Total**	
Concept Train	929–1 160	4 119	ϕ_{c_i} training
Concept Test	118–184	577	ϕ_{c_i} evaluation
Contrastive Examples	186–422	1 236	$\mathcal{H}^-$ highlights

Split overlap and evaluation limitations. All 600 Explanation Test images have byte-identical SHA-256 matches in the 8 125-image Label Test partition used to estimate the label-confusion likelihood. Thus the evaluation examples contribute to a fitted component of the rule scorer. Removing those exact images would leave 7 525 calibration images, but the reported results use the original likelihoods. Additionally, 447 of the 600 benchmark images and 6 191 of the 8 125 Label Test images share a filename-derived observation family with a Label Train image. These family counts indicate sibling-view overlap; they do not quantify its effect on accuracy.

For the concept classifiers, 21 of the 577 Concept Test entries (20 distinct image hashes) match a training source image for the same concept: 3 fuzzy-body, 8 dense-hind-leg, 1 pollen-hind-leg, and 9 green-head/thorax entries. Their test supports are respectively 134, 184, 118, and 141. Training uses augmented sources, so this is source-image overlap even when the final augmented tensors differ. Exact hashing does not detect all crops, re-encodings, or related views. Consequently, the accuracies and likelihood estimates are not independent at the source-image or observation level. Independent evaluation requires grouped, deduplicated splits and new fitting and scoring.

Evaluation scope. The concept configuration is labeled "100 examples"; the actual counts used in the experiment are given in Table 6. This paper reports automatic evaluation only and no completed human study.

C Experimental Protocol and Metrics

C.1 Backbones and Training Settings

We evaluate the 12 `timm` backbones (Wightman, 2019) in Table 7. They include ConvNeXt V2 (Woo et al., 2023), ResNets with torchvision or RSB recipes (He et al., 2016; Wightman et al., 2021), DINO (Caron et al., 2021), MAE (He et al., 2022), SigLIP 2 (Tschannen et al., 2025), BEiT (Bao et al., 2022), Swin (Liu et al., 2021), and SwinV2 (Liu et al., 2022). These references identify architecture or pretraining families; the model identifiers specify the actual variants. SigLIP 2 specifies WebLI.

Table 7: Backbone identifiers and concept-probe layers. CNXv2-L without a qualifier means the fine-tuned variant; RN152 means A1h.

Paper label	Model identifier	Probe layer
CNXv2-L	`convnextv2_large.fcmae_ft_in22k_in1k`	`head.global_pool`
RN152 (A1h)	`resnet152.a1h_in1k`	`global_pool`
ViT-B/16 (DINO)	`vit_base_patch16_224.dino`	`fc_norm`
SwinV2-S	`swinv2_cr_small_ns_224.sw_in1k`	`head.global_pool`
ViT-B/16 (SigLIP 2)	`vit_base_patch16_siglip_gap_224.v2_webli`	`fc_norm`
BEiT-L/16	`beit_large_patch16_224.in22k_ft_in22k_in1k`	`fc_norm`
ViT-B/16 (MAE)	`vit_base_patch16_224.mae`	`fc_norm`
CNXv2-L (FCMAE)	`convnextv2_large.fcmae`	`head.global_pool`
RN50 (TV)	`resnet50.tv_in1k`	`global_pool`
RN50 (A1h)	`resnet50.a1h_in1k`	`global_pool`
Swin-B (IN-1k)	`swin_base_patch4_window7_224.ms_in1k`	`head.global_pool`
Swin-B (IN-22k)	`swin_base_patch4_window7_224.ms_in22k`	`head.global_pool`

The synthetic settings use $T \in \{500, 2000, 5000\}$ label-training images per class, $C \in \{100, 200, 500\}$ positive and equally many negative concept examples, $P \in \{0.1, 0.2\}$ prior sample fractions, and configured seed 42. This gives 216 model/configuration results, or 108 at either prior fraction. Both datasets use linear concept heads, no layer search, augmentation setting 8, and a configured label-head limit of 25 epochs. The synthetic configuration additionally specifies a concept-head limit of 100 epochs. These are epoch limits; individual runs need not reach them. Optimizer and learning-rate settings, complete training histories, exact software revisions, and checkpoint-to-result links are not fully documented, limiting exact reproducibility.

C.2 Backbone Inference State

The stated training design fixes learned backbone parameters and trains task heads, but this does not imply an unchanged feature mapping. Comparing the RN50-TV and RN152-A1h checkpoints with cached pretrained weights finds all comparable non-head learned parameters unchanged, while all inspected BatchNorm running-statistic and counter buffers differ: 159/159 and 465/465, respectively, in each dataset. Such buffers affect inference even without parameter-gradient updates. This comparison covers these two available checkpoint pairs per dataset and does not establish the state of every checkpoint. We consequently avoid attributing cross-backbone differences solely to unchanged pretrained representations.

C.3 Likelihoods and Prior Estimation

The label likelihood $\Pr(\hat{\ell} \mid \ell)$ uses the Label Test confusion matrix, with rows conditioned on the true class. Each binary concept likelihood uses the Concept Test counts as in Appendix A.1. ConceptAcc is an unweighted mean of the concept-specific accuracies, rather than accuracy pooled across differently sized concept sets. Label and rule percentages are computed from their integer correct counts.

Synthetic concept priors use generated concept metadata sampled from Label Train. Bee priors use the species-to-concept mapping described in Appendix B.2. The bee sample sizes are 7 290 at $P = 0.1$ and 14 577 at $P = 0.2$. For the reference-protocol comparison, the two resulting priors differ slightly but all rounded bee metrics agree. The ConvNeXt V2-L label head has 95.4% accuracy on 8 125 Label Test images; the other heads range from 60.1% to 94.7%. This larger pool contains the 600-image Explanation Test used for Explanation Test LabelAcc.

C.4 Evaluation Limitations

Bee data independence. All 600 Explanation Test images have byte-identical copies in the 8 125-image Label Test pool used to estimate the label likelihoods. Removing these copies would leave 7 525 images, but no corrected result is inferred from that subtraction. The Concept Test set includes 21 of 577 entries duplicating a training source for the same concept. Filename-based observation-family grouping also finds training siblings for 447/600 explanation images and 6 191/8 125 Label Test images. These counts distinguish exact-file overlap from related views of the same observation. The scores therefore do not establish generalization to independent observations or independent likelihood-estimation data. The nominal 72 873-image Label Train pool has a 3 643-image internal validation set; its complement contains 69 230 images if that partition is exhaustive.

Scope of comparison. The experiments use one configured seed and do not estimate uncertainty across training runs. Sweep cells share models and data, and are not independent replications. Changing T also changes the synthetic evaluation images and rule proportions; it is not a pure training-size intervention. Background shortcuts are not separately controlled. The evaluation does not compare explanation methods, ablate individual components, assess confidence calibration, quantify localization accuracy, or report user-study outcomes. The qualitative examples illustrate the display format without validating localization or reference retrieval.

C.5 Evaluation Metrics

For each Explanation Test image, let $r^{gt}(x)$ be the target rule implied by the concept vector and $\ell^{gt}(x)$ its label. Synthetic vectors come from generation metadata; bee vectors are the class-associated templates described in Appendix B.2.

LabelAcc is the fraction with $\hat{\ell}(x) = \ell^{gt}(x)$. **ConceptAcc** is the mean binary accuracy over 16 synthetic or four bee Concept Test reports, not accuracy on Explanation Test concepts. **RuleAcc** is the fraction with $r^*(x) = r^{gt}(x)$. **RuleLabelAcc** is the fraction with $\ell(r^*(x)) = \ell^{gt}(x)$. It can exceed LabelAcc without changing the classifier output. RuleAcc cannot exceed RuleLabelAcc, and they are equal for the one-rule-per-class bee benchmark. **Avg** Conf averages the selected rule's normalized score; it measures relative decisiveness, not calibrated correctness.

C.6 Results Collection and Additional Evaluations

The separately provided `experiment_results/` directory contains the complete numerical results, copied source reports, and file hashes. Its `benchmark_results.csv` gives all five aggregate metrics at full precision for every included configuration, with explicit dataset, model, training-budget, prior-fraction, and protocol fields. Per-concept metrics, label-head evaluations, and prior estimates are provided separately. The appendix prints all aggregate configurations, per-concept accuracies, and label-head summaries; the CSV and JSON files retain the underlying counts and additional metric fields.

The reference comparison contains 216 synthetic configurations and 24 bee configurations across 12 backbones. A separate ConvNeXt V2-L bee evaluation uses the `disjoint_label_calibration_v1` protocol. A further ConvNeXt V2-L concept-training evaluation uses different concept data and objectives. These additional evaluations are reported separately in Appendix E; they are not substituted into the reference comparison. The separate bee benchmark does not identify its concept checkpoint, so its ConceptAcc is left unspecified rather than paired with an unrelated concept report.

D Complete Synthetic Results and Sensitivity Analyses

This appendix reports all 216 synthetic configurations: 12 backbones, $T \in \{500, 2000, 5000\}$ label-training images per class, $C \in \{100, 200, 500\}$ positive and the same number of negative training examples per concept, and prior-estimation fractions $P \in \{0.1, 0.2\}$. The *synthetic reference configuration* used in the main paper is T=5000, C=500, P=0.1. Each configuration is evaluated once; repeated-seed uncertainty estimates are unavailable. Appendix C defines the metrics and training protocol.

Benchmark accuracy percentages are rounded half-up to one decimal, and mean confidence to three decimals. LabelAcc, RuleAcc, RuleLabelAcc, and Avg Conf use the 300-image Explanation Test split; ConceptAcc is the unweighted mean of 16 binary concept-test accuracies. Each concept test has 100 positive and 100 negative examples. The separately provided `experiment_results` directory contains full-precision metrics, correct-prediction counts, concept confusion matrices, classification reports, and the source reports used to generate these tables. Tables in this appendix present all five benchmark metrics for every configuration and the individual concept and label-head accuracies. The complete sweep tables also give Switch%, the fraction of images for which the selected rule implies a different class from the classifier prediction; a switch is not necessarily an error.

D.1 Synthetic reference configuration

Table 8: All 12 backbones in the synthetic reference configuration (T=5000, C=500, P=0.1). Accuracies are percentages and confidence is unitless. Bold marks the maximum in each column.

Backbone	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf
CNXv2-L	99.3%	93.3%	93.7%	**99.7%**	0.925
CNXv2-L (FCMAE only)	94.7%	56.9%	54.0%	94.0%	0.687
RN152	40.3%	73.6%	47.7%	67.3%	0.540
RN50 (TV)	39.0%	79.9%	69.7%	85.3%	0.634
RN50 (A1h)	51.0%	87.0%	67.7%	77.7%	0.698
ViT-B/16 (DINO)	**99.7%**	92.8%	**96.7%**	**99.7%**	0.935
ViT-B/16 (MAE)	97.7%	79.4%	75.0%	97.7%	0.804
ViT-B/16 (SigLIP 2)	99.0%	91.6%	90.0%	99.0%	0.908
BEiT-L/16	99.3%	89.4%	91.3%	**99.7%**	0.938
Swin-B (IN-1k)	99.0%	90.0%	90.7%	99.0%	0.918
Swin-B (IN-22k)	99.0%	**95.2%**	94.7%	99.0%	**0.949**
SwinV2-S	99.3%	91.9%	94.0%	99.3%	0.918

D.2 Detailed synthetic findings

Rule-implied labels can correct classifier errors. The three ResNet backbones illustrate this property dramatically: their transfer-trained heads reach only 39.0–51.0% label accuracy on the Rule Selector Test split (RN50 (TV) 39.0%, RN152 40.3%, RN50 (A1h) 51.0%), yet the rule-implied label accuracy (RuleLabelAcc) reaches 85.3%, 67.3%, and 77.7%, respectively — improvements of +46.3, +27.0, and +26.7 points obtained by incorporating concept evidence into the joint rule score. This occurs because the score in Equation (7) can rank a rule of another class above rules associated with the classifier prediction $\hat{\ell}(x)$. The concept classifiers, while individually imperfect (73.6–87.0% mean accuracy for the three ResNets), collectively provide enough evidence for the rule selector to infer a different label from an incorrect classifier prediction. The correction is positive in every (T, C) cell of the sweep at P=0.1 for all three ResNets (RN50 (TV) +6.7 to +54.3 points, RN50 (A1h) +15.3 to +29.3, RN152 +13.7 to +27.0; Appendix D.3). For the nine backbones whose heads are at least 90% accurate in the synthetic reference configuration, rule-implied label accuracy remains close to classifier accuracy: RuleLabelAcc differs from LabelAcc by at most 0.7 points in the synthetic reference configuration (*e.g.* CNXv2-L 99.3% → 99.7%, CNXv2-L (FCMAE only) 94.7% → 94.0%) and by between −1.7 and +0.7 points over the sweep at P=0.1. More generally, each synthetic class has two mutually exclusive rules, and each image satisfies exactly one of them. RuleLabelAcc can exceed RuleAcc when the selector chooses the other rule of the correct class: the rule is incorrect for that image, but its implied label is correct.

FLUENT *rule accuracy tracks concept quality, not just label accuracy.* Across the 12 backbones in the synthetic reference configuration, rule accuracy is more strongly rank-correlated with mean concept accuracy (Spearman $\rho = 0.90$) than with label accuracy ($\rho = 0.87$); over their 108 (backbone, T, C) cells at P=0.1 the correlations are $\rho = 0.91$ and 0.79, respectively. These are descriptive associations among related configurations, not independent observations or significance tests. Backbones with near-perfect heads but weak concept probes are the clearest cases: ViT-B/16 (MAE) reaches 97.7% label accuracy but only 79.4% mean concept accuracy and 75.0% rule accuracy, and CNXv2-L (FCMAE only) reaches 94.7% label accuracy but 56.9% concept and 54.0% rule accuracy; conversely, RN50 (TV) attains 69.7% rule accuracy from a 39.0% head because its concept probes average 79.9%. The same effect is visible within a single backbone as the concept-annotation budget grows: at T=5000, P=0.1, raising C from 100 to 200 to 500 increases mean concept accuracy monotonically for all 12 backbones, and rule accuracy rises from C=100 to C=500 for all 12 of them (strictly at every step for 10; ViT-B/16 (SigLIP 2) is flat at 85.0% between C=100 and C=200, and BEiT-L/16 dips at C=200, $81.7\% \rightarrow 80.0\% \rightarrow 91.3\%$), *e.g.* CNXv2-L $90.0\% \rightarrow 92.3\% \rightarrow 93.7\%$, SwinV2-S $87.0\% \rightarrow 93.3\% \rightarrow 94.0\%$, RN50 (TV) $35.3\% \rightarrow 41.7\% \rightarrow 69.7\%$, while label accuracy is unaffected because the head does not depend on C. Because each rule mentions 8 concepts, uncertainty across their predictions compounds in the rule score, and rule selection is sensitive to per-concept errors.

Pretraining variants differ in rule recovery. For related architecture families, the variants differ by tens of points in rule accuracy: ConvNeXt V2-L with supervised fine-tuning reaches 93.7% rule accuracy against 54.0% for the same architecture with FCMAE pretraining only (mean concept accuracy 93.3% vs. 56.9%); ViT-B/16 reaches 96.7% with DINO, 90.0% with SigLIP 2, and 75.0% with MAE weights; and Swin-B improves from 90.7% to 94.7% when pretrained on IN-22k instead of IN-1k. Label accuracy remains 94.7–99.7% across these seven variants. These comparisons are consistent with an association between concept-prediction quality and rule recovery, but they do not isolate the effects of pretraining data, objectives, and implementation differences.

Average rule confidence correlates with rule accuracy. Across the 12 backbones in the synthetic reference configuration, mean normalized rule confidence correlates with rule accuracy with Spearman $\rho = 0.92$ (Pearson $r = 0.96$): every backbone with Avg $\mathrm{Conf} \geq 0.90$ attains at least 90.0% rule accuracy (seven backbones, 90.0–96.7%), and every backbone below 0.90 attains at most 75.0% (five backbones, 47.7–75.0%). Over their 108 sweep cells at P=0.1 the rank correlation is $\rho = 0.95$. Within a single backbone the nine (T, C) cells at P=0.1 span a narrower range and the relation is weaker (ρ between 0.20 and 0.93 depending on the backbone), so the observed association is stronger across the evaluated configurations than within individual backbones. These aggregate associations do not establish that confidence reliably identifies individual explanation errors.

D.3 COMPLETE SYNTHETIC SWEEP

Tables 10–15 give every benchmark metric at both prior-estimation fractions. Label and concept predictors are shared between the two P values, and label predictors are shared across C at fixed T. Different T settings use separately generated training and evaluation sets; consequently, changes across T combine training-size and data-composition effects. For example, Vumble rule 1 appears in 163/500, 954/2000, and 2749/5000 training images and 33/100, 44/100, and 62/100 Explanation Test images, respectively. These comparisons therefore do not isolate sample efficiency.

Table 9 summarizes rule accuracy at $P{=}0.1$ across all nine (T, C) pairs.

Table 9: Rule accuracy (%) over the complete synthetic sweep at $P{=}0.1$: training images per class $T \in \{500, 2000, 5000\}$, positive and negative examples per concept $C \in \{100, 200, 500\}$, all 12 backbones. Every metric of every cell is listed in Appendix D.3.

	$T{=}500$			$T{=}2000$			$T{=}5000$		
Backbone	$C{=}100$	$C{=}200$	$C{=}500$	$C{=}100$	$C{=}200$	$C{=}500$	$C{=}100$	$C{=}200$	$C{=}500$
CNXv2-L	85.7	89.0	89.0	87.0	89.7	89.0	90.0	92.3	93.7
CNXv2-L (FCMAE only)	55.7	46.3	51.7	46.0	48.7	47.0	49.7	53.0	54.0
RN152	31.7	33.0	35.0	41.7	45.3	46.0	41.7	43.3	47.7
RN50 (TV)	45.7	59.0	62.3	42.3	52.3	73.0	35.3	41.7	69.7
RN50 (A1h)	53.7	57.7	63.7	55.3	61.7	69.7	53.7	59.0	67.7
ViT-B/16 (DINO)	84.0	83.7	92.7	93.3	97.0	97.3	92.7	95.3	96.7
ViT-B/16 (MAE)	63.7	61.0	67.3	62.0	68.0	73.0	61.7	66.7	75.0
ViT-B/16 (SigLIP 2)	79.3	82.7	89.7	75.7	85.0	87.7	85.0	85.0	90.0
BEiT-L/16	73.7	82.3	86.7	78.3	84.0	92.7	81.7	80.0	91.3
Swin-B (IN-1k)	78.0	83.0	84.3	84.7	90.0	93.0	87.3	90.0	90.7
Swin-B (IN-22k)	87.7	89.0	91.0	87.3	92.3	94.7	91.0	94.0	94.7
SwinV2-S	75.3	79.7	83.3	81.7	83.7	92.7	87.0	93.3	94.0

Table 10: Complete synthetic results at T=500 and P=0.1: all 12 backbones and all three concept-training budgets. Accuracies and label-switch rates are percentages; confidence is unitless.

Backbone	C	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf	Switch%
CNXv2-L	100	99.3%	89.8%	85.7%	98.7%	0.903	0.7
	200	99.3%	91.7%	89.0%	99.7%	0.895	0.3
	500	99.3%	93.9%	89.0%	99.3%	0.916	0.0
CNXv2-L (FCMAE only)	100	95.7%	55.2%	55.7%	95.3%	0.664	1.0
	200	95.7%	58.5%	46.3%	94.3%	0.681	1.3
	500	95.7%	57.8%	51.7%	95.7%	0.668	0.3
RN152	100	35.3%	67.1%	31.7%	49.0%	0.414	51.0
	200	35.3%	67.6%	33.0%	51.7%	0.438	59.7
	500	35.3%	71.9%	35.0%	52.7%	0.516	59.3
RN50 (TV)	100	40.7%	71.6%	45.7%	66.0%	0.567	58.3
	200	40.7%	76.3%	59.0%	83.3%	0.616	55.3
	500	40.7%	80.1%	62.3%	84.0%	0.696	59.3
RN50 (A1h)	100	48.7%	78.5%	53.7%	69.7%	0.630	39.0
	200	48.7%	79.4%	57.7%	77.3%	0.616	40.3
	500	48.7%	84.4%	63.7%	78.0%	0.694	48.0
ViT-B/16 (DINO)	100	100.0%	89.2%	84.0%	99.3%	0.912	0.7
	200	100.0%	91.8%	83.7%	99.7%	0.915	0.3
	500	100.0%	93.9%	92.7%	100.0%	0.935	0.0
ViT-B/16 (MAE)	100	95.7%	71.3%	63.7%	95.7%	0.772	0.7
	200	95.7%	75.3%	61.0%	95.3%	0.783	2.0
	500	95.7%	79.1%	67.3%	96.3%	0.819	1.3
ViT-B/16 (SigLIP 2)	100	100.0%	83.9%	79.3%	100.0%	0.863	0.0
	200	100.0%	90.3%	82.7%	100.0%	0.881	0.0
	500	100.0%	91.2%	89.7%	100.0%	0.913	0.0
BEiT-L/16	100	99.0%	83.6%	73.7%	98.7%	0.857	0.3
	200	99.0%	87.7%	82.3%	99.0%	0.886	0.0
	500	99.0%	89.3%	86.7%	99.0%	0.924	0.0
Swin-B (IN-1k)	100	99.0%	85.6%	78.0%	99.0%	0.868	0.0
	200	99.0%	89.0%	83.0%	98.7%	0.884	0.3
	500	99.0%	90.8%	84.3%	98.7%	0.901	0.3
Swin-B (IN-22k)	100	99.7%	89.8%	87.7%	99.7%	0.927	0.0
	200	99.7%	93.3%	89.0%	99.7%	0.929	0.0
	500	99.7%	95.3%	91.0%	99.0%	0.914	0.7
SwinV2-S	100	98.3%	85.3%	75.3%	97.7%	0.860	0.7
	200	98.3%	88.3%	79.7%	98.0%	0.875	0.3
	500	98.3%	91.6%	83.3%	98.0%	0.887	1.0

Table 11: Complete synthetic results at T=500 and P=0.2: all 12 backbones and all three concept-training budgets. Accuracies and label-switch rates are percentages; confidence is unitless.

Backbone	C	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf	Switch%
CNXv2-L	100	99.3%	89.8%	85.7%	98.7%	0.902	0.7
	200	99.3%	91.7%	89.0%	99.7%	0.895	0.3
	500	99.3%	93.9%	89.0%	99.3%	0.912	0.0
CNXv2-L (FCMAE only)	100	95.7%	55.2%	56.3%	95.3%	0.672	1.0
	200	95.7%	58.5%	48.0%	94.3%	0.686	1.3
	500	95.7%	57.8%	52.7%	95.7%	0.672	0.3
RN152	100	35.3%	67.1%	31.3%	49.3%	0.410	51.0
	200	35.3%	67.6%	32.0%	50.3%	0.438	59.3
	500	35.3%	71.9%	34.0%	52.3%	0.515	58.7
RN50 (TV)	100	40.7%	71.6%	46.0%	67.7%	0.566	58.3
	200	40.7%	76.3%	58.7%	83.3%	0.617	55.0
	500	40.7%	80.1%	63.3%	84.3%	0.700	59.0
RN50 (A1h)	100	48.7%	78.5%	53.7%	70.0%	0.634	38.7
	200	48.7%	79.4%	57.0%	76.0%	0.620	39.7
	500	48.7%	84.4%	64.0%	78.3%	0.701	48.0
ViT-B/16 (DINO)	100	100.0%	89.2%	83.7%	99.3%	0.912	0.7
	200	100.0%	91.8%	83.7%	99.7%	0.915	0.3
	500	100.0%	93.9%	92.7%	100.0%	0.933	0.0
ViT-B/16 (MAE)	100	95.7%	71.3%	64.7%	95.7%	0.775	0.7
	200	95.7%	75.3%	61.7%	95.7%	0.789	2.3
	500	95.7%	79.1%	67.0%	96.0%	0.825	1.0
ViT-B/16 (SigLIP 2)	100	100.0%	83.9%	79.3%	100.0%	0.863	0.0
	200	100.0%	90.3%	79.0%	100.0%	0.887	0.0
	500	100.0%	91.2%	89.3%	100.0%	0.913	0.0
BEiT-L/16	100	99.0%	83.6%	73.7%	98.7%	0.859	0.3
	200	99.0%	87.7%	83.3%	99.0%	0.889	0.0
	500	99.0%	89.3%	86.3%	99.0%	0.924	0.0
Swin-B (IN-1k)	100	99.0%	85.6%	77.3%	99.0%	0.868	0.0
	200	99.0%	89.0%	85.0%	98.7%	0.884	0.3
	500	99.0%	90.8%	84.3%	98.7%	0.900	0.3
Swin-B (IN-22k)	100	99.7%	89.8%	87.7%	99.7%	0.926	0.0
	200	99.7%	93.3%	89.3%	99.7%	0.930	0.0
	500	99.7%	95.3%	91.0%	99.0%	0.919	0.7
SwinV2-S	100	98.3%	85.3%	75.0%	97.7%	0.860	0.7
	200	98.3%	88.3%	80.0%	98.0%	0.876	0.3
	500	98.3%	91.6%	83.3%	98.0%	0.886	1.0

Table 12: Complete synthetic results at T=2000 and P=0.1: all 12 backbones and all three concept-training budgets. Accuracies and label-switch rates are percentages; confidence is unitless.

Backbone	C	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf	Switch%
CNXv2-L	100	99.7%	90.3%	87.0%	99.7%	0.901	0.0
	200	99.7%	91.1%	89.7%	99.7%	0.902	0.0
	500	99.7%	94.1%	89.0%	99.3%	0.917	0.3
CNXv2-L (FCMAE only)	100	95.0%	55.1%	46.0%	93.3%	0.613	1.7
	200	95.0%	57.1%	48.7%	95.0%	0.723	0.0
	500	95.0%	57.3%	47.0%	94.3%	0.683	0.7
RN152	100	44.7%	67.0%	41.7%	61.3%	0.461	42.3
	200	44.7%	69.8%	45.3%	61.7%	0.450	42.0
	500	44.7%	71.0%	46.0%	64.7%	0.503	48.0
RN50 (TV)	100	32.3%	73.6%	42.3%	60.3%	0.528	63.0
	200	32.3%	77.1%	52.3%	67.7%	0.580	58.3
	500	32.3%	81.4%	73.0%	86.7%	0.695	69.3
RN50 (A1h)	100	53.0%	78.8%	55.3%	69.7%	0.600	34.3
	200	53.0%	79.3%	61.7%	73.0%	0.603	40.7
	500	53.0%	84.3%	69.7%	80.7%	0.697	43.3
ViT-B/16 (DINO)	100	99.3%	89.1%	93.3%	99.0%	0.918	1.0
	200	99.3%	91.6%	97.0%	99.0%	0.945	0.3
	500	99.3%	94.4%	97.3%	99.7%	0.954	1.0
ViT-B/16 (MAE)	100	98.7%	71.5%	62.0%	98.7%	0.718	0.0
	200	98.7%	76.5%	68.0%	98.7%	0.758	0.0
	500	98.7%	79.8%	73.0%	98.7%	0.810	0.0
ViT-B/16 (SigLIP 2)	100	99.3%	84.2%	75.7%	99.3%	0.801	0.0
	200	99.3%	89.2%	85.0%	99.3%	0.892	0.0
	500	99.3%	91.9%	87.7%	99.3%	0.926	0.0
BEiT-L/16	100	99.7%	83.5%	78.3%	99.7%	0.830	0.0
	200	99.7%	87.3%	84.0%	99.7%	0.886	0.0
	500	99.7%	88.6%	92.7%	99.3%	0.936	0.3
Swin-B (IN-1k)	100	99.0%	86.6%	84.7%	99.0%	0.899	0.0
	200	99.0%	88.2%	90.0%	99.0%	0.883	0.0
	500	99.0%	91.2%	93.0%	99.0%	0.944	0.0
Swin-B (IN-22k)	100	98.7%	90.0%	87.3%	98.3%	0.916	0.3
	200	98.7%	93.4%	92.3%	98.7%	0.946	0.0
	500	98.7%	94.8%	94.7%	98.7%	0.950	0.0
SwinV2-S	100	97.7%	86.4%	81.7%	97.7%	0.872	0.0
	200	97.7%	88.3%	83.7%	98.0%	0.858	0.3
	500	97.7%	91.7%	92.7%	98.0%	0.901	0.3

Table 13: Complete synthetic results at T=2000 and P=0.2: all 12 backbones and all three concept-training budgets. Accuracies and label-switch rates are percentages; confidence is unitless.

Backbone	C	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf	Switch%
CNXv2-L	100	99.7%	90.3%	87.0%	99.7%	0.898	0.0
	200	99.7%	91.1%	90.0%	99.7%	0.899	0.0
	500	99.7%	94.1%	89.0%	99.3%	0.915	0.3
CNXv2-L (FCMAE only)	100	95.0%	55.1%	45.7%	93.3%	0.618	1.7
	200	95.0%	57.1%	49.0%	95.0%	0.731	0.0
	500	95.0%	57.3%	46.3%	94.3%	0.690	0.7
RN152	100	44.7%	67.0%	41.7%	61.3%	0.463	42.3
	200	44.7%	69.8%	46.3%	63.3%	0.451	44.0
	500	44.7%	71.0%	45.7%	64.0%	0.503	47.0
RN50 (TV)	100	32.3%	73.6%	42.7%	60.3%	0.532	63.0
	200	32.3%	77.1%	52.0%	67.3%	0.584	58.7
	500	32.3%	81.4%	73.3%	87.0%	0.698	69.7
RN50 (A1h)	100	53.0%	78.8%	55.3%	69.7%	0.601	35.0
	200	53.0%	79.3%	61.3%	73.3%	0.604	41.0
	500	53.0%	84.3%	69.7%	81.0%	0.696	43.7
ViT-B/16 (DINO)	100	99.3%	89.1%	93.3%	99.0%	0.915	1.0
	200	99.3%	91.6%	97.3%	99.0%	0.942	0.3
	500	99.3%	94.4%	97.3%	99.7%	0.953	1.0
ViT-B/16 (MAE)	100	98.7%	71.5%	61.3%	98.7%	0.722	0.0
	200	98.7%	76.5%	68.0%	98.7%	0.760	0.0
	500	98.7%	79.8%	73.0%	98.7%	0.811	0.0
ViT-B/16 (SigLIP 2)	100	99.3%	84.2%	75.7%	99.3%	0.803	0.0
	200	99.3%	89.2%	85.0%	99.3%	0.892	0.0
	500	99.3%	91.9%	87.7%	99.3%	0.926	0.0
BEiT-L/16	100	99.7%	83.5%	78.0%	99.7%	0.828	0.0
	200	99.7%	87.3%	84.3%	99.7%	0.886	0.0
	500	99.7%	88.6%	92.7%	99.3%	0.935	0.3
Swin-B (IN-1k)	100	99.0%	86.6%	84.7%	99.0%	0.897	0.0
	200	99.0%	88.2%	90.0%	99.0%	0.881	0.0
	500	99.0%	91.2%	93.0%	99.0%	0.942	0.0
Swin-B (IN-22k)	100	98.7%	90.0%	87.3%	98.3%	0.916	0.3
	200	98.7%	93.4%	92.3%	98.7%	0.945	0.0
	500	98.7%	94.8%	94.7%	98.7%	0.950	0.0
SwinV2-S	100	97.7%	86.4%	81.7%	97.7%	0.873	0.0
	200	97.7%	88.3%	83.7%	98.0%	0.857	0.3
	500	97.7%	91.7%	92.7%	98.0%	0.901	0.3

Table 14: Complete synthetic results at T=5000 and P=0.1: all 12 backbones and all three concept-training budgets. Accuracies and label-switch rates are percentages; confidence is unitless.

Backbone	C	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf	Switch%
CNXv2-L	100	99.3%	90.0%	90.0%	99.3%	0.908	0.0
	200	99.3%	91.6%	92.3%	99.7%	0.904	0.3
	500	99.3%	93.3%	93.7%	99.7%	0.925	0.3
CNXv2-L (FCMAE only)	100	94.7%	54.9%	49.7%	94.0%	0.666	1.3
	200	94.7%	56.8%	53.0%	94.0%	0.752	0.7
	500	94.7%	56.9%	54.0%	94.0%	0.687	0.7
RN152	100	40.3%	69.1%	41.7%	61.0%	0.459	58.3
	200	40.3%	70.6%	43.3%	63.7%	0.474	52.7
	500	40.3%	73.6%	47.7%	67.3%	0.540	58.0
RN50 (TV)	100	39.0%	71.3%	35.3%	45.7%	0.493	65.7
	200	39.0%	74.5%	41.7%	55.0%	0.550	73.7
	500	39.0%	79.9%	69.7%	85.3%	0.634	66.7
RN50 (A1h)	100	51.0%	78.9%	53.7%	66.3%	0.587	41.3
	200	51.0%	80.9%	59.0%	68.3%	0.644	39.0
	500	51.0%	87.0%	67.7%	77.7%	0.698	43.0
ViT-B/16 (DINO)	100	99.7%	89.4%	92.7%	99.7%	0.935	0.0
	200	99.7%	92.0%	95.3%	99.7%	0.930	0.0
	500	99.7%	92.8%	96.7%	99.7%	0.935	0.0
ViT-B/16 (MAE)	100	97.7%	71.9%	61.7%	97.7%	0.720	0.0
	200	97.7%	75.9%	66.7%	97.7%	0.764	0.0
	500	97.7%	79.4%	75.0%	97.7%	0.804	0.0
ViT-B/16 (SigLIP 2)	100	99.0%	84.2%	85.0%	99.0%	0.817	0.0
	200	99.0%	90.4%	85.0%	99.0%	0.909	0.0
	500	99.0%	91.6%	90.0%	99.0%	0.908	0.0
BEiT-L/16	100	99.3%	82.7%	81.7%	99.3%	0.844	0.0
	200	99.3%	87.3%	80.0%	99.3%	0.895	0.0
	500	99.3%	89.4%	91.3%	99.7%	0.938	0.3
Swin-B (IN-1k)	100	99.0%	86.3%	87.3%	99.0%	0.892	0.0
	200	99.0%	88.9%	90.0%	99.0%	0.910	0.0
	500	99.0%	90.0%	90.7%	99.0%	0.918	0.0
Swin-B (IN-22k)	100	99.0%	90.1%	91.0%	99.0%	0.925	0.0
	200	99.0%	92.7%	94.0%	99.0%	0.937	0.0
	500	99.0%	95.2%	94.7%	99.0%	0.949	0.7
SwinV2-S	100	99.3%	86.1%	87.0%	99.3%	0.871	0.0
	200	99.3%	88.6%	93.3%	99.3%	0.909	0.0
	500	99.3%	91.9%	94.0%	99.3%	0.918	0.0

Table 15: Complete synthetic results at T=5000 and P=0.2: all 12 backbones and all three concept-training budgets. Accuracies and label-switch rates are percentages; confidence is unitless.

Backbone	C	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf	Switch%
CNXv2-L	100	99.3%	90.0%	90.0%	99.3%	0.907	0.0
	200	99.3%	91.6%	92.3%	99.7%	0.902	0.3
	500	99.3%	93.3%	93.7%	99.7%	0.922	0.3
CNXv2-L (FCMAE only)	100	94.7%	54.9%	48.3%	94.0%	0.657	1.3
	200	94.7%	56.8%	53.0%	94.0%	0.754	0.7
	500	94.7%	56.9%	53.0%	94.0%	0.684	0.7
RN152	100	40.3%	69.1%	43.0%	63.0%	0.456	57.7
	200	40.3%	70.6%	43.3%	63.7%	0.470	52.0
	500	40.3%	73.6%	47.3%	67.0%	0.537	57.3
RN50 (TV)	100	39.0%	71.3%	36.0%	47.0%	0.495	63.3
	200	39.0%	74.5%	42.3%	55.0%	0.552	74.0
	500	39.0%	79.9%	62.7%	77.7%	0.636	59.0
RN50 (A1h)	100	51.0%	78.9%	53.3%	66.3%	0.587	41.3
	200	51.0%	80.9%	59.0%	68.3%	0.644	39.3
	500	51.0%	87.0%	68.0%	78.0%	0.697	43.3
ViT-B/16 (DINO)	100	99.7%	89.4%	92.7%	99.7%	0.933	0.0
	200	99.7%	92.0%	95.3%	99.7%	0.927	0.0
	500	99.7%	92.8%	96.7%	99.7%	0.933	0.0
ViT-B/16 (MAE)	100	97.7%	71.9%	62.3%	97.7%	0.718	0.0
	200	97.7%	75.9%	67.0%	97.7%	0.770	0.0
	500	97.7%	79.4%	75.3%	97.7%	0.801	0.0
ViT-B/16 (SigLIP 2)	100	99.0%	84.2%	79.3%	99.0%	0.816	0.0
	200	99.0%	90.4%	85.0%	99.0%	0.910	0.0
	500	99.0%	91.6%	90.0%	99.0%	0.906	0.0
BEiT-L/16	100	99.3%	82.7%	81.7%	99.3%	0.845	0.0
	200	99.3%	87.3%	80.0%	99.3%	0.897	0.0
	500	99.3%	89.4%	91.3%	99.7%	0.938	0.3
Swin-B (IN-1k)	100	99.0%	86.3%	87.3%	99.0%	0.890	0.0
	200	99.0%	88.9%	90.3%	99.3%	0.908	0.3
	500	99.0%	90.0%	90.7%	99.0%	0.915	0.0
Swin-B (IN-22k)	100	99.0%	90.1%	91.0%	99.0%	0.925	0.0
	200	99.0%	92.7%	94.0%	99.0%	0.934	0.0
	500	99.0%	95.2%	93.7%	99.0%	0.947	0.7
SwinV2-S	100	99.3%	86.1%	87.3%	99.3%	0.868	0.0
	200	99.3%	88.6%	93.3%	99.3%	0.907	0.0
	500	99.3%	91.9%	93.7%	99.3%	0.915	0.0

D.4 PRIOR-ESTIMATION FRACTION

Table 16 compares the two prior-estimation fractions while holding the label and concept predictors fixed. It gives RuleAcc at $T{=}5000, C{=}500$ and the maximum absolute change across all nine (T, C) pairs. Tables 10–15 report both fractions in full; no configuration is represented only by this sensitivity summary. The comparison measures sensitivity to the two estimated priors, rather than to arbitrary changes in concept prevalence.

The synthetic sweep compares $P{=}0.1$ with $P{=}0.2$ using the same label and concept predictors. Their accuracies are unchanged. The maximum change in rule accuracy across the nine (T, C) settings is at most 2.0 percentage points for 10 of the 12 backbones (0.3 points for CNXv2-L). The exceptions are ViT-B/16 (SigLIP 2) and RN50 (TV), with maxima of 5.7 and 7.0 points. Mean confidence changes by less than 0.02 for every backbone. These comparisons support limited sensitivity for most evaluated configurations; they do not establish a general sample-complexity guarantee for prior estimation.

Table 16 reports the rule accuracies at $T{=}5000, C{=}500$ and maximum absolute changes. At $T{=}5000, C{=}500$ the two samples' estimated concept priors differ by at most 0.012. The bee results have identical aggregate metrics at the displayed precision for the two fractions; Appendix E describes their prior-fraction comparison.

Table 16: Prior-fraction sensitivity. The two RuleAcc columns use $T{=}5000, C{=}500$; maximum absolute accuracy changes across the nine (T, C) pairs are in percentage points. LabelAcc and ConceptAcc are unchanged.

	RuleAcc, $T{=}5000, C{=}500$		Maximum absolute change		
Backbone	$P{=}0.1$	$P{=}0.2$	RuleAcc (pp)	RuleLabelAcc (pp)	Avg Conf
CNXv2-L	93.7%	93.7%	0.3	0.0	0.003
CNXv2-L (FCMAE only)	54.0%	53.0%	1.7	0.0	0.009
RN152	47.7%	47.3%	1.3	2.0	0.004
RN50 (TV)	69.7%	62.7%	7.0	7.7	0.005
RN50 (A1h)	67.7%	68.0%	0.7	1.3	0.007
ViT-B/16 (DINO)	96.7%	96.7%	0.3	0.0	0.003
ViT-B/16 (MAE)	75.0%	75.3%	1.0	0.3	0.007
ViT-B/16 (SigLIP 2)	90.0%	90.0%	5.7	0.0	0.006
BEiT-L/16	91.3%	91.3%	1.0	0.0	0.003
Swin-B (IN-1k)	90.7%	90.7%	2.0	0.3	0.003
Swin-B (IN-22k)	94.7%	93.7%	1.0	0.0	0.004
SwinV2-S	94.0%	93.7%	0.3	0.0	0.003

D.5 PER-CONCEPT SYNTHETIC TEST ACCURACIES

The following tables report the 16 concept accuracies for each of the 108 (backbone, T, C) combinations. The two prior-estimation fractions share these predictors, so each accuracy is reported once. Every concept is tested on 200 balanced examples. The arithmetic mean across both eight-concept panels for a row gives ConceptAcc in the benchmark tables.

Table 17: Synthetic concept-test accuracy (%) at T=500, panel 1 of 2. C is the number of positive and negative training examples per concept. Values are shared by P=0.1 and P=0.2.

Backbone	C	Alorn	Floom	Frib	Guffit	Hebda	Heffix	Krad	Opple
CNXv2-L	100	86.0	91.5	76.0	94.0	98.5	95.5	89.5	88.5
	200	90.0	93.0	81.0	92.0	98.0	94.0	92.0	90.5
	500	93.5	95.5	85.0	94.0	99.0	96.5	92.5	89.5
CNXv2-L (FCMAE only)	100	58.5	56.0	50.0	62.0	61.5	50.5	58.0	55.0
	200	72.0	54.0	57.0	72.0	62.0	51.5	64.0	52.0
	500	61.5	48.5	51.5	80.5	65.5	62.0	52.0	55.0
RN152	100	70.0	60.0	70.0	76.5	66.5	64.0	65.5	65.0
	200	72.0	62.5	68.5	79.5	70.0	69.0	71.5	60.0
	500	81.0	74.5	75.5	78.0	77.0	70.5	65.5	72.5
RN50 (TV)	100	69.0	72.5	64.5	81.0	72.5	76.5	76.5	72.0
	200	77.0	80.0	80.5	89.0	81.0	86.0	60.5	64.5
	500	82.5	81.5	89.0	83.0	83.5	90.0	81.5	78.5
RN50 (A1h)	100	73.5	73.5	85.5	86.0	76.5	85.5	79.0	79.0
	200	76.5	84.5	81.5	86.5	80.5	84.5	82.5	73.5
	500	86.5	89.5	90.5	87.5	85.0	93.5	89.0	87.5
ViT-B/16 (DINO)	100	79.5	93.0	97.0	91.5	94.5	93.0	91.0	77.5
	200	83.0	96.0	99.5	94.5	95.5	92.0	91.0	82.5
	500	94.0	96.0	100.0	97.0	96.5	92.0	94.5	93.5
ViT-B/16 (MAE)	100	72.0	66.0	72.5	84.0	75.0	75.0	76.0	58.0
	200	81.5	75.0	80.0	90.0	77.5	73.5	85.0	65.5
	500	84.5	72.5	88.5	77.5	82.0	81.0	88.5	72.0
ViT-B/16 (SigLIP 2)	100	85.5	78.5	84.0	94.5	85.0	88.5	89.0	82.0
	200	94.5	89.5	91.5	95.5	89.5	87.5	91.5	88.0
	500	90.0	87.0	87.5	96.0	91.0	92.5	94.5	90.5
BEiT-L/16	100	81.5	84.0	83.0	90.0	73.0	75.0	89.0	77.0
	200	83.0	91.0	89.0	91.0	85.0	88.0	90.0	75.5
	500	87.5	91.5	93.5	94.0	82.5	90.5	92.0	81.5
Swin-B (IN-1k)	100	83.5	88.0	84.5	90.0	92.5	88.5	86.0	77.0
	200	87.0	91.5	87.5	89.5	97.5	93.5	92.5	85.5
	500	89.5	91.0	91.0	90.5	96.0	92.5	91.5	90.0
Swin-B (IN-22k)	100	83.0	92.0	83.5	96.0	95.5	92.0	96.0	85.0
	200	90.5	93.5	90.0	94.0	97.5	97.5	95.5	92.0
	500	91.5	98.5	93.0	95.0	98.5	98.0	96.0	95.0
SwinV2-S	100	84.5	75.5	80.0	89.0	88.0	92.5	89.5	86.0
	200	88.0	89.5	85.0	86.0	92.0	89.0	94.0	84.0
	500	91.0	91.0	88.0	91.5	96.5	96.5	94.5	92.0

Table 18: Synthetic concept-test accuracy (%) at T=500, panel 2 of 2. C is the number of positive and negative training examples per concept. Values are shared by P=0.1 and P=0.2.

Backbone	C	Smeek	Sorbax	Voop	Vornex	Wexil	Yintar	Zandool	Zorf
CNXv2-L	100	83.5	83.0	94.5	97.0	90.0	92.0	95.0	82.5
	200	88.0	86.0	95.0	91.5	94.0	96.5	96.5	88.5
	500	87.5	88.0	98.0	97.5	98.5	98.0	98.0	92.0
CNXv2-L (FCMAE only)	100	61.5	57.5	51.0	52.0	53.5	52.5	54.0	49.0
	200	64.5	57.5	48.5	54.0	64.0	50.5	62.0	50.5
	500	75.5	55.0	54.0	51.5	51.0	53.0	55.0	53.5
RN152	100	70.5	70.5	62.0	69.5	68.5	63.5	72.5	59.0
	200	72.0	67.0	62.5	72.5	65.5	56.5	75.5	56.5
	500	72.0	70.5	65.0	78.0	72.0	53.5	75.0	69.5
RN50 (TV)	100	81.0	78.0	66.0	74.0	71.5	65.0	58.5	67.5
	200	73.5	80.5	66.0	87.0	77.0	70.0	79.0	69.0
	500	86.0	80.0	75.5	87.5	82.0	61.0	73.5	67.0
RN50 (A1h)	100	74.5	76.5	71.5	81.0	79.0	78.0	85.5	71.5
	200	75.5	77.5	74.5	85.5	80.0	73.0	83.0	71.0
	500	78.0	77.0	72.0	89.5	77.5	78.0	90.5	79.0
ViT-B/16 (DINO)	100	89.5	85.0	93.0	90.5	92.0	83.0	96.5	80.0
	200	86.0	92.0	88.0	95.0	94.0	90.5	98.0	90.5
	500	94.5	91.5	97.0	96.5	93.5	80.0	96.5	89.5
ViT-B/16 (MAE)	100	80.5	68.5	62.0	91.0	79.0	61.5	58.0	61.0
	200	78.5	74.5	66.5	86.5	70.0	58.5	67.0	74.5
	500	79.0	78.5	66.5	93.5	85.0	67.0	75.0	74.0
ViT-B/16 (SigLIP 2)	100	89.5	80.5	80.0	96.5	81.5	89.5	70.0	68.0
	200	92.5	86.5	86.0	93.0	90.5	91.5	92.0	84.5
	500	94.0	89.5	89.5	96.5	95.0	87.5	93.5	85.0
BEiT-L/16	100	74.5	81.0	91.0	92.5	87.0	82.0	88.5	89.0
	200	87.0	81.0	95.0	92.5	91.5	87.5	91.5	85.0
	500	86.0	86.0	95.5	94.0	88.5	77.5	97.0	91.5
Swin-B (IN-1k)	100	82.5	79.5	86.0	88.0	93.0	80.0	91.0	79.0
	200	82.5	86.0	90.5	84.5	92.5	88.5	89.0	85.5
	500	85.5	83.0	96.0	91.0	95.5	89.0	95.0	86.5
Swin-B (IN-22k)	100	84.0	87.5	96.0	95.5	93.5	80.5	89.0	88.0
	200	89.0	89.0	97.5	94.5	95.5	90.5	95.5	91.5
	500	92.5	91.0	98.0	95.0	94.0	94.0	98.0	96.0
SwinV2-S	100	83.0	77.5	87.0	92.0	93.5	83.5	86.0	77.5
	200	88.5	84.5	91.5	88.5	92.0	88.0	91.0	81.5
	500	91.5	84.5	92.5	93.0	96.5	91.5	92.5	82.5

Table 19: Synthetic concept-test accuracy (%) at T=2000, panel 1 of 2. C is the number of positive and negative training examples per concept. Values are shared by P=0.1 and P=0.2.

Backbone	C	Alorn	Floom	Frib	Guffit	Hebda	Heffix	Krad	Opple
CNXv2-L	100	85.0	90.5	79.5	95.5	98.5	96.0	89.5	85.5
	200	88.0	85.0	79.5	92.0	99.0	96.5	89.5	89.5
	500	90.5	94.5	85.0	95.0	98.5	95.5	95.5	93.0
CNXv2-L (FCMAE only)	100	58.5	52.0	52.0	61.0	59.5	52.5	53.5	58.0
	200	68.0	52.0	56.0	72.5	63.0	54.0	59.0	52.0
	500	58.5	49.0	53.0	73.0	63.5	62.0	52.0	55.0
RN152	100	74.0	61.5	71.5	72.5	64.5	66.0	67.5	66.0
	200	75.5	70.5	75.5	82.0	71.5	69.0	70.5	63.5
	500	76.0	70.5	74.5	69.5	73.5	73.0	72.5	68.0
RN50 (TV)	100	75.0	74.0	63.5	81.5	70.5	85.5	79.0	76.5
	200	78.5	78.5	83.5	83.5	75.5	85.5	84.5	64.5
	500	80.0	75.0	92.0	88.0	85.0	91.5	83.0	82.5
RN50 (A1h)	100	71.5	78.5	83.0	87.0	83.5	80.0	85.5	76.5
	200	79.5	83.0	84.0	84.0	82.0	81.0	82.0	74.5
	500	81.5	86.5	83.5	86.0	89.0	87.5	91.0	86.5
ViT-B/16 (DINO)	100	77.0	93.0	97.0	92.5	95.0	93.5	93.5	76.5
	200	82.5	95.5	99.5	93.5	97.5	89.5	94.5	85.0
	500	93.5	96.0	100.0	98.0	97.0	97.0	94.0	93.0
ViT-B/16 (MAE)	100	77.5	65.0	72.5	85.5	76.5	72.5	77.5	59.5
	200	81.5	73.5	77.0	91.5	82.5	77.0	84.0	65.5
	500	86.0	76.0	86.0	86.5	82.0	84.5	85.0	72.0
ViT-B/16 (SigLIP 2)	100	85.5	78.5	84.0	91.5	85.5	87.0	90.0	81.5
	200	92.0	88.0	86.0	94.5	89.0	90.5	90.5	82.0
	500	94.5	85.5	93.0	94.0	91.5	89.0	94.5	93.0
BEiT-L/16	100	77.5	84.0	85.5	89.0	75.5	74.5	88.0	74.5
	200	77.5	93.5	91.0	85.5	84.0	89.0	90.5	79.0
	500	83.0	89.0	94.5	95.0	81.0	93.0	91.5	81.0
Swin-B (IN-1k)	100	82.5	89.5	88.0	89.5	96.0	88.5	91.5	85.0
	200	87.0	90.0	87.0	91.5	95.0	91.0	91.5	84.0
	500	89.5	91.0	92.5	89.0	96.5	97.0	95.0	89.5
Swin-B (IN-22k)	100	84.5	94.0	81.5	95.0	93.0	92.5	95.5	86.0
	200	83.0	94.0	92.0	94.0	95.5	98.5	97.0	93.5
	500	82.0	96.0	91.5	94.5	99.0	98.0	98.5	96.0
SwinV2-S	100	87.0	81.5	83.5	90.0	90.0	94.0	89.0	82.0
	200	87.0	89.5	87.0	86.0	91.0	89.0	91.5	86.0
	500	90.5	92.5	86.0	93.0	97.0	94.5	95.0	93.0

Table 20: Synthetic concept-test accuracy (%) at T=2000, panel 2 of 2. C is the number of positive and negative training examples per concept. Values are shared by P=0.1 and P=0.2.

Backbone	C	Smeek	Sorbax	Voop	Vornex	Wexil	Yintar	Zandool	Zorf
CNXv2-L	100	85.0	83.5	94.5	97.5	91.5	92.5	96.0	84.0
	200	87.5	82.5	94.5	94.0	94.5	97.5	95.5	92.0
	500	89.5	90.0	99.0	95.5	97.5	96.5	98.0	92.5
CNXv2-L (FCMAE only)	100	67.0	57.0	51.5	51.0	54.0	51.0	53.5	49.0
	200	59.0	56.5	48.5	57.5	57.0	51.5	52.5	55.0
	500	66.0	58.0	55.0	50.5	55.0	52.0	61.5	53.5
RN152	100	65.5	61.5	61.0	80.5	69.0	55.5	77.0	58.0
	200	67.0	70.0	60.0	76.0	72.0	59.0	74.5	60.0
	500	73.0	75.5	63.5	80.5	80.0	52.5	78.5	55.5
RN50 (TV)	100	82.5	80.0	64.5	80.5	78.0	63.0	60.5	62.5
	200	78.0	75.0	65.0	84.5	81.5	63.5	80.0	72.0
	500	87.5	84.5	71.5	91.0	82.0	55.5	78.5	75.0
RN50 (A1h)	100	72.0	76.0	78.5	79.5	79.5	74.5	83.0	71.5
	200	72.5	83.5	70.5	83.5	82.0	77.0	77.0	72.5
	500	76.5	78.0	78.0	90.0	83.5	81.5	89.5	80.5
ViT-B/16 (DINO)	100	90.0	82.0	91.0	90.5	92.0	85.0	96.0	81.5
	200	81.0	94.0	90.0	95.5	94.0	87.5	98.5	88.0
	500	91.0	91.5	98.5	97.0	92.5	81.0	97.5	93.5
ViT-B/16 (MAE)	100	82.0	67.0	60.5	92.5	79.0	63.0	57.5	56.5
	200	83.0	75.5	67.5	86.5	74.5	63.0	72.5	68.5
	500	80.0	76.0	72.0	90.5	85.0	64.0	75.0	75.5
ViT-B/16 (SigLIP 2)	100	90.0	82.0	83.0	97.5	82.5	92.0	68.0	68.0
	200	91.5	88.0	88.5	96.0	88.0	90.5	87.5	84.0
	500	93.5	92.5	90.5	96.0	94.5	86.0	94.0	88.5
BEiT-L/16	100	72.5	82.5	90.0	93.5	86.5	84.0	89.5	89.5
	200	90.0	86.0	93.5	91.0	92.0	83.0	91.5	80.5
	500	86.0	82.5	92.0	94.5	89.0	76.0	97.5	92.5
Swin-B (IN-1k)	100	79.5	77.5	85.5	86.5	92.5	81.0	93.5	79.5
	200	82.0	86.0	91.5	83.0	92.0	87.5	88.5	84.0
	500	81.5	86.5	93.5	92.5	95.5	89.5	94.5	85.0
Swin-B (IN-22k)	100	87.5	84.5	92.0	96.5	93.0	81.5	94.5	89.0
	200	90.5	91.5	95.0	95.0	95.5	90.5	96.5	92.5
	500	92.5	91.5	97.0	96.5	97.5	92.0	97.5	97.5
SwinV2-S	100	84.0	78.0	87.5	93.5	95.0	84.5	86.5	77.0
	200	88.5	86.5	94.5	87.0	91.0	87.5	89.0	82.0
	500	92.0	85.0	94.5	93.0	97.0	89.5	92.0	82.5

Table 21: Synthetic concept-test accuracy (%) at T=5000, panel 1 of 2. C is the number of positive and negative training examples per concept. Values are shared by P=0.1 and P=0.2.

Backbone	C	Alorn	Floom	Frib	Guffit	Hebda	Heffix	Krad	Opple
CNXv2-L	100	84.5	89.0	81.5	95.5	95.5	96.5	90.0	86.5
	200	91.5	92.0	82.0	92.5	99.0	93.0	91.0	93.0
	500	93.5	89.0	83.0	95.5	98.5	94.5	95.5	88.0
CNXv2-L (FCMAE only)	100	57.5	54.0	51.0	63.0	57.0	50.0	52.0	59.0
	200	64.5	58.0	55.5	72.5	61.0	50.5	63.5	51.5
	500	62.5	50.5	53.5	77.5	61.0	62.0	48.0	56.0
RN152	100	75.0	68.0	73.0	74.0	65.0	67.5	59.5	65.0
	200	75.5	70.0	77.0	79.0	75.5	67.5	68.5	60.0
	500	78.5	71.5	80.5	82.0	77.5	79.0	72.0	66.0
RN50 (TV)	100	70.5	70.5	73.5	77.5	71.0	81.5	78.5	71.0
	200	75.5	74.0	86.0	78.5	69.0	82.0	82.5	57.5
	500	78.0	84.0	88.0	85.0	80.5	87.5	81.0	81.0
RN50 (A1h)	100	73.5	76.5	85.0	84.5	75.5	82.5	80.0	81.0
	200	81.0	86.0	89.5	82.0	78.5	85.0	85.5	71.5
	500	84.0	90.5	92.5	88.5	87.0	93.5	90.0	81.5
ViT-B/16 (DINO)	100	82.5	94.0	97.5	90.5	93.5	93.5	93.5	76.5
	200	80.5	95.5	99.5	94.0	94.5	91.5	93.5	85.5
	500	87.5	95.5	99.5	96.5	97.0	95.5	93.5	87.0
ViT-B/16 (MAE)	100	78.0	63.5	72.5	85.0	76.0	73.5	76.5	60.0
	200	82.5	73.5	76.5	84.0	85.5	74.5	82.5	67.5
	500	85.0	73.5	87.0	87.0	80.0	84.0	81.0	73.0
ViT-B/16 (SigLIP 2)	100	85.5	80.0	85.0	94.5	85.0	87.5	90.5	81.0
	200	92.0	88.0	90.5	95.0	90.0	90.5	91.0	87.0
	500	94.0	91.5	92.5	95.0	91.0	87.0	94.0	91.5
BEiT-L/16	100	81.5	85.0	86.5	87.0	69.5	74.5	87.5	73.5
	200	81.0	92.0	88.0	91.0	84.5	88.0	90.5	76.0
	500	88.0	92.0	94.0	90.5	83.0	91.5	94.5	82.5
Swin-B (IN-1k)	100	84.0	89.0	89.0	88.5	94.0	89.0	94.0	80.0
	200	87.0	90.0	86.0	89.5	93.0	92.5	93.0	83.5
	500	89.0	90.5	91.5	89.0	96.0	95.5	92.0	89.0
Swin-B (IN-22k)	100	83.0	92.0	83.0	95.5	97.0	93.5	96.0	84.5
	200	86.0	94.5	90.5	93.0	96.5	97.0	95.0	88.5
	500	89.0	97.5	93.5	96.0	98.0	96.5	96.0	93.0
SwinV2-S	100	87.5	82.5	81.5	90.5	89.5	92.0	88.0	83.0
	200	84.5	88.0	86.0	87.0	91.0	91.0	93.0	85.0
	500	91.5	91.5	88.5	92.0	97.0	96.0	94.0	89.0

Table 22: Synthetic concept-test accuracy (%) at T=5000, panel 2 of 2. C is the number of positive and negative training examples per concept. Values are shared by P=0.1 and P=0.2.

Backbone	C	Smeek	Sorbax	Voop	Vornex	Wexil	Yintar	Zandool	Zorf
CNXv2-L	100	84.0	85.0	94.5	96.5	93.0	91.0	93.5	83.5
	200	87.0	85.0	93.5	91.5	91.0	97.0	94.5	91.5
	500	89.5	85.0	97.5	96.5	98.0	97.0	97.5	94.5
CNXv2-L (FCMAE only)	100	63.0	57.0	54.0	50.5	54.5	51.5	53.5	50.5
	200	58.5	56.0	49.0	56.0	60.0	51.0	53.0	48.0
	500	63.5	56.0	56.0	52.0	51.0	53.0	56.0	52.0
RN152	100	71.5	65.0	65.5	80.5	73.5	57.0	79.0	67.0
	200	69.5	66.5	65.5	79.0	74.5	59.0	82.0	60.5
	500	70.0	68.5	69.0	87.0	77.5	56.0	80.0	63.0
RN50 (TV)	100	80.0	74.5	62.5	76.0	69.0	61.5	56.5	66.0
	200	75.5	75.0	64.5	88.0	75.0	61.0	80.5	67.5
	500	84.5	80.0	72.0	85.0	82.5	59.0	75.5	75.5
RN50 (A1h)	100	70.0	77.0	73.0	83.5	80.0	82.0	82.0	76.0
	200	79.0	76.5	75.0	86.5	82.5	79.0	83.5	73.0
	500	80.5	83.5	87.0	90.5	82.5	90.0	88.0	83.0
ViT-B/16 (DINO)	100	90.5	84.0	91.5	87.0	93.5	84.5	95.0	82.5
	200	86.5	92.0	89.5	95.5	95.5	90.0	97.0	92.0
	500	89.5	93.5	94.5	94.5	91.5	77.5	97.0	94.0
ViT-B/16 (MAE)	100	80.0	68.0	61.0	92.5	79.0	66.0	57.5	61.0
	200	80.0	77.0	66.5	84.5	76.5	65.0	72.5	65.5
	500	77.5	79.0	71.0	91.5	80.5	67.5	77.0	76.0
ViT-B/16 (SigLIP 2)	100	91.0	82.0	81.5	96.5	81.5	91.5	67.0	67.0
	200	93.0	89.5	87.5	94.5	92.5	90.5	91.0	84.5
	500	93.5	93.0	88.5	97.0	94.5	84.5	91.0	87.5
BEiT-L/16	100	70.0	78.5	90.0	93.0	87.0	83.0	88.5	87.5
	200	87.0	83.5	93.5	92.5	90.5	87.0	87.5	83.5
	500	88.0	88.5	90.5	94.5	90.0	72.0	97.5	93.0
Swin-B (IN-1k)	100	81.0	78.0	86.0	87.0	92.5	78.5	92.5	78.5
	200	82.5	88.0	90.5	85.0	91.5	89.0	95.0	86.0
	500	81.5	88.5	94.5	91.0	95.5	89.0	85.0	82.5
Swin-B (IN-22k)	100	83.5	87.0	93.0	97.0	93.5	81.0	93.0	89.5
	200	89.5	91.0	95.0	95.0	94.5	90.5	94.5	92.5
	500	92.5	93.5	98.0	97.0	98.5	94.0	97.5	92.5
SwinV2-S	100	80.0	80.0	87.0	91.5	93.5	84.5	88.0	79.0
	200	86.5	87.0	94.0	90.0	94.0	87.5	90.5	83.0
	500	91.5	85.0	94.0	92.0	97.5	93.5	95.5	82.5

D.6 SYNTHETIC LABEL-HEAD ACCURACIES

Table 23 reports the validation and test accuracies of every synthetic label head. These splits are distinct from the 300-image Explanation Test split used for LabelAcc in the benchmark tables. Validation accuracy is available as a fraction rounded to two decimals in the classification reports and is therefore shown to the nearest percentage point; test accuracy is computed from exact confusion counts. At fixed T, the same label head is used for all three concept-training budgets and both prior-estimation fractions.

Table 23: Synthetic label-head accuracy (%) on the validation and test splits. The test split contains 3 000 images; validation has 75, 300, and 750 images at T=500, 2000, 5000, respectively. All 36 trained heads are shown.

	T=500		T=2000		T=5000	
Backbone	Validation	Test	Validation	Test	Validation	Test
CNXv2-L	100%	99.0%	100%	99.4%	100%	99.5%
CNXv2-L (FCMAE only)	89%	93.0%	95%	95.4%	97%	96.6%
RN152	49%	43.9%	42%	43.5%	40%	41.5%
RN50 (TV)	43%	37.7%	33%	34.0%	39%	37.0%
RN50 (A1h)	60%	49.3%	52%	49.6%	53%	49.0%
ViT-B/16 (DINO)	100%	99.6%	99%	99.2%	100%	99.8%
ViT-B/16 (MAE)	96%	93.9%	99%	98.2%	99%	98.9%
ViT-B/16 (SigLIP 2)	99%	99.3%	99%	99.5%	100%	99.7%
BEiT-L/16	99%	99.0%	99%	99.1%	100%	99.6%
Swin-B (IN-1k)	100%	98.6%	99%	99.1%	99%	99.5%
Swin-B (IN-22k)	100%	99.0%	100%	99.1%	100%	99.6%
SwinV2-S	100%	97.5%	99%	98.2%	99%	99.5%

E COMPLETE BEE RESULTS AND PREDICTOR ACCURACIES

The *bee reference configuration* uses the fixed dataset splits described in Appendix B.2, prior-estimation fraction $P{=}0.1$, and seed 42. This appendix reports all 24 backbone/prior-fraction combinations: 12 backbones at each of $P \in \{0.1, 0.2\}$. It also reports the individual concept and label-head accuracies and two additional ConvNeXt V2-L experiments separately. All results are single-run measurements. Accuracies are rounded half-up to one decimal and confidence to three decimals; full-precision reports and counts are supplied separately in `experiment_results`.

LabelAcc, RuleAcc, RuleLabelAcc, and Avg Conf use the 600-image Explanation Test split. ConceptAcc is the unweighted mean of the four concept-test accuracies. There is one rule per class, so RuleAcc and RuleLabelAcc coincide. The overlap and class-template limitations in Appendix B.2 apply: these results do not constitute an independent evaluation on new observations or independently annotated visible concepts.

E.1 COMPLETE BEE BENCHMARK RESULTS

Table 24: All 12 backbones in the bee reference configuration ($P{=}0.1$, seed 42). Accuracies are percentages; confidence is unitless. Bold marks column maxima. The split-overlap limitations in Appendix B.2 apply.

Backbone	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf
CNXv2-L	**99.7%**	96.4%	85.8%	85.8%	**0.967**
CNXv2-L (FCMAE only)	63.3%	68.3%	43.8%	43.8%	0.633
RN152	84.7%	87.0%	62.8%	62.8%	0.803
RN50 (TV)	91.2%	85.5%	87.8%	87.8%	0.770
RN50 (A1h)	87.0%	87.0%	66.7%	66.7%	0.792
ViT-B/16 (DINO)	97.7%	96.2%	84.3%	84.3%	0.927
ViT-B/16 (MAE)	81.5%	85.5%	66.2%	66.2%	0.793
ViT-B/16 (SigLIP 2)	78.2%	84.4%	58.5%	58.5%	0.763
BEiT-L/16	80.8%	84.1%	63.2%	63.2%	0.798
Swin-B (IN-1k)	96.0%	95.3%	82.2%	82.2%	0.946
Swin-B (IN-22k)	**99.7%**	**97.7%**	**98.2%**	**98.2%**	0.947
SwinV2-S	97.3%	95.2%	83.5%	83.5%	0.943

Table 25: Complete bee results at $P{=}0.2$, seed 42. All five metrics are shown for every backbone. At the displayed precision they equal the corresponding $P{=}0.1$ values.

Backbone	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf
CNXv2-L	**99.7%**	96.4%	85.8%	85.8%	**0.967**
CNXv2-L (FCMAE only)	63.3%	68.3%	43.8%	43.8%	0.633
RN152	84.7%	87.0%	62.8%	62.8%	0.803
RN50 (TV)	91.2%	85.5%	87.8%	87.8%	0.770
RN50 (A1h)	87.0%	87.0%	66.7%	66.7%	0.792
ViT-B/16 (DINO)	97.7%	96.2%	84.3%	84.3%	0.927
ViT-B/16 (MAE)	81.5%	85.5%	66.2%	66.2%	0.793
ViT-B/16 (SigLIP 2)	78.2%	84.4%	58.5%	58.5%	0.763
BEiT-L/16	80.8%	84.1%	63.2%	63.2%	0.798
Swin-B (IN-1k)	96.0%	95.3%	82.2%	82.2%	0.946
Swin-B (IN-22k)	**99.7%**	**97.7%**	**98.2%**	**98.2%**	0.947
SwinV2-S	97.3%	95.2%	83.5%	83.5%	0.943

E.2 DETAILED BEE FINDINGS

Agreement with the bee rules varies across backbones. On the 600-image Explanation Test partition, rule accuracy ranges from 43.8% (CNXv2-L, FCMAE only; 263 correct) to 98.2% (Swin-B, IN-22k; 589 correct). The five backbones with at least 96% label accuracy on this partition (Swin-B (IN-22k), CNXv2-L, ViT-B/16 (DINO), SwinV2-S, Swin-B (IN-1k)) agree with the class-associated rule for 82.2–98.2% of images, with mean normalized rule confidence 0.927–0.967. CNXv2-L has 99.7% label accuracy, 85.8% rule accuracy, and mean confidence 0.967. For these five backbones, individual concept-classifier accuracies are 90.2–99.3% (means 95.2–97.7%; CNXv2-L 93.2–98.6%, mean 96.4%). The concept sets contain same-source training/test overlap, and the benchmark vectors are class templates, so these measurements do not establish independent recognition of each cue on the benchmark photographs.

Rule selection can reduce label accuracy on field photographs. Because each class has exactly one rule, every rule error is also a rule-label error, and the gap between label accuracy and rule accuracy measures the net effect of rule selection relative to the classifier; this gap is present for every backbone, from 1.5 points (Swin-B (IN-22k), 99.7% vs. 98.2%) to 21.8 points (RN152, 84.7% vs. 62.8%), and rule accuracy never exceeds label accuracy on this dataset. For CNXv2-L (99.7% vs. 85.8%), on 13.8% of images (83 of 600) the scoring procedure in Equation (7) selects a rule of a different class from the (almost always correct) backbone prediction. Concept accuracy is measured on the curated concept exemplars, whereas Explanation Test images are unconstrained field photographs in which cues can be occluded, viewpoint-dependent, or transient (*e.g.* visible pollen loads vary across photographs); concept accuracy on these images may therefore differ from the curated estimate, and errors can compound across the two to four literals per rule. The high mean confidence (Avg $\mathrm{Conf} = 0.967$) despite imperfect rule accuracy shows that decisive rule scores can accompany errors; these results do not validate a confidence threshold for presenting explanations.

Backbone quality and rule confidence vary on real data. Some backbones have lower accuracy in both evidence channels. For RN152, label accuracy on the Explanation Test partition is 84.7% (its fine-tuned head reaches 73.1% on the Label Test split) and mean concept accuracy is 87.0% (81.4–92.9% per concept), yielding a rule accuracy of 62.8%; CNXv2-L with FCMAE pretraining only is the weakest backbone on both channels (63.3% label, 68.3% concept) and reaches 43.8% rule accuracy. Rule-implied label accuracy is lower than classifier accuracy for every backbone on this dataset. Because there is one rule per class, rule accuracy and rule-implied label accuracy coincide; individual concept errors need not change the selected rule. Across backbones, rule accuracy follows the label channel more closely (Spearman $\rho = 0.90$ with label accuracy vs. 0.76 with mean concept accuracy across the 12 backbones), the reverse of the synthetic ordering. The aggregate confidence pattern is: the five backbones with Avg $\mathrm{Conf} \geq 0.92$ all exceed 82% rule accuracy, and six of the seven below 0.81 fall below 67%; the exception is RN50 (TV), whose mean rule confidence is comparatively low (0.770) despite 87.8% rule accuracy. The rank correlation between confidence and rule accuracy across backbones is $\rho = 0.64$ (Pearson $r = 0.84$), weaker than on the synthetic data. This association does not establish reliable confidence-based ranking for new deployments. The checkpoint comparison gives Swin-B 82.2% and 98.2% rule accuracy for IN-1k and IN-22k checkpoints, respectively; ViT-B/16 reaches 84.3% with DINO against 66.2% (MAE) and 58.5% (SigLIP 2), and the FCMAE-only and supervised-fine-tuned CNXv2-L checkpoints obtain 43.8% and 85.8%, respectively. These comparisons use a single run per configuration, with no repeated-seed uncertainty estimates.

Table 26: Bee benchmark counts out of 600 images. Every count is identical at $P{=}0.1$ and $P{=}0.2$. A label switch means that the selected rule implies a class different from the classifier prediction; it is not necessarily an additional error.

Backbone	Correct labels	Correct rules	Correct rule labels	Switches	Switch%
CNXv2-L	598	515	515	83	13.8
CNXv2-L (FCMAE only)	380	263	263	219	36.5
RN152	508	377	377	163	27.2
RN50 (TV)	547	527	527	43	7.2
RN50 (A1h)	522	400	400	151	25.2
ViT-B/16 (DINO)	586	506	506	80	13.3
ViT-B/16 (MAE)	489	397	397	134	22.3
ViT-B/16 (SigLIP 2)	469	351	351	213	35.5
BEiT-L/16	485	379	379	162	27.0
Swin-B (IN-1k)	576	493	493	91	15.2
Swin-B (IN-22k)	598	589	589	9	1.5
SwinV2-S	584	501	501	87	14.5

E.3 BEE PER-CONCEPT AND LABEL-HEAD ACCURACIES

Table 27 gives every concept-test accuracy and the corresponding macro-average. The four concept tests contain 134 fuzzy-body, 184 dense-hind-leg, 118 pollen-on-hind-leg, and 141 green-head/thorax entries, respectively (577 entries total). The same concept predictors are used at both prior-estimation fractions. Label-head validation and test accuracies use 3 643 and 8 125 images, respectively, and therefore differ from LabelAcc on the 600-image Explanation Test split. Validation accuracy is recorded as a fraction rounded to two decimals and is shown to the nearest percentage point; test accuracy is computed from exact confusion counts. The Label Test split supplies the label-confusion likelihood and includes the Explanation Test images, as detailed in Appendix B.2.

Table 27: Bee concept and label-head accuracies (%). The four concept columns correspond to fuzzy body, dense hair on hind legs, pollen on hind legs, and green head/thorax. Mean is the unweighted mean over concepts. Validation and test refer to the label-head splits, containing 3 643 and 8 125 images, respectively. Validation accuracy is available only to the nearest percentage point.

	Concept accuracy					Label-head accuracy	
Backbone	Fuzzy	Dense	Pollen	Green	Mean	Validation	Test
CNXv2-L	97.8	96.2	93.2	98.6	96.4	95	95.4
CNXv2-L (FCMAE only)	80.6	73.4	69.5	49.6	68.3	59	60.1
RN152	88.8	84.8	81.4	92.9	87.0	75	73.1
RN50 (TV)	87.3	78.3	84.7	91.5	85.5	80	79.5
RN50 (A1h)	93.3	80.4	81.4	92.9	87.0	79	77.7
ViT-B/16 (DINO)	97.0	95.7	95.8	96.5	96.2	91	90.3
ViT-B/16 (MAE)	94.0	72.8	86.4	88.7	85.5	73	72.9
ViT-B/16 (SigLIP 2)	91.0	64.7	88.1	93.6	84.4	72	71.7
BEiT-L/16	90.3	75.5	79.7	90.8	84.1	74	73.5
Swin-B (IN-1k)	98.5	91.8	92.4	98.6	95.3	90	91.1
Swin-B (IN-22k)	99.3	95.7	97.5	98.6	97.7	95	94.7
SwinV2-S	96.3	90.2	95.8	98.6	95.2	91	90.4

E.4 PRIOR-FRACTION SENSITIVITY

For the bee experiment, the two prior samples contain 7 290 and 14 577 Label Train images, respectively. Their class-associated templates yield nearly identical marginal priors: the dense-hair, pollen, and green-head/thorax prior changes from 0.3334705075 to 0.3334705358, and fuzzy-body prevalence is the complement. All correctness and switch counts are unchanged, as are all metrics at the displayed precision. The largest absolute change in mean confidence is approximately 2.05×10^{-8}. An exhaustive comparison of the $3 \times 2^4 = 48$ possible predicted-label/concept-vector inputs for each backbone also finds no selected-rule change under the two priors and identical likelihood tables. This establishes insensitivity to these two almost identical prior estimates; it does not establish robustness to substantial prevalence shift.

Per-image confidence calibration and selective prediction are not evaluated. The aggregate results do not determine the accuracy that would follow from removing the overlaps; quantifying that effect requires experiments with independently partitioned data.

Tables 24 and 25 report all benchmark metrics for the two fractions, and Table 26 gives their correctness and switch counts.

E.5 ADDITIONAL CONVNEXT V2-L EXPERIMENTS

Two additional experiments use CNXv2-L with the `fcmae_ft_in22k_in1k` checkpoint. They are reported separately because their protocol and predictor identities cannot be combined into a single matched experiment. Neither replaces a row of the 12-backbone reference comparison.

Disjoint label-calibration experiment. The benchmark report identifies the label-calibration protocol as `disjoint_label_calibration_v1`. Table 28 gives all its benchmark metrics; ConceptAcc is unreported because the report does not identify a matching concept-test result. Of 600 evaluation images, 598 labels, 581 rules, and 581 rule-implied labels are correct, and 17 selected rules imply a different class from the classifier prediction. The rule-accuracy difference from the reference CNXv2-L experiment cannot be attributed solely to label calibration without establishing the identity of all other experiment components. A disjoint label-calibration split also does not by itself resolve the other dataset-overlap or concept-template limitations.

Table 28: Additional CNXv2-L benchmark using the disjoint label-calibration protocol. Accuracies are percentages; confidence is unitless. A dash denotes an unreported matching concept accuracy, not zero.

Backbone	LabelAcc	ConceptAcc	RuleAcc	RuleLabelAcc	Avg Conf
CNXv2-L	99.7%	–	96.8%	96.8%	0.977

Separate concept-training experiment. Table 29 reports the four concept classifiers from the separate CNXv2-L concept-training run. Each concept is tested on 200 examples, for 800 test entries in total. Its mean accuracy is 86.4% (full precision 86.375%). The available experiment records do not establish that these concept predictors produced the disjoint-calibration benchmark above, so that mean is not inserted into Table 28.

Table 29: Separate CNXv2-L concept-training experiment. Per-concept correctness counts and accuracies are reported independently of the disjoint label-calibration benchmark.

Concept	Test examples	Correct	Accuracy
Fuzzy body	200	176	88.0%
Dense hair on hind legs	200	171	85.5%
Pollen on hind legs	200	147	73.5%
Green head and thorax	200	197	98.5%