Proceedings of Machine Learning Research vol 284:1–16, 2026 20th Conference on Neurosymbolic Learning and Reasoning

Weakly supervised concept Bottleneck Learning for Robust Two stage Object centric visual reasoning

Sparsh Tiwari
University of Lübeck, Germany

Gesina Schwalbe
Ulm University, Germany

Bettina Finzel
University of Bamberg, Germany

Editors: Alessandra Mileo, Andrea Passerini and Cogan Shimizu

Abstract

Two-stage neuro-symbolic architectures provide an elegant paradigm for visual problem solving by cleanly separating connectionist perception of predefined symbols from possibly later defined relational reasoning thereon. However, anchoring high-level predicates into visual frames typically necessitates annotations that are expensive to acquire. In this work, we introduce the Dynamic Orthogonal Concept Bottleneck (D-OCB), an object-centric slot-VAE framework designed to extract human-aligned symbolic predicates under extremely weak supervision. D-OCB eliminates the arduous manual tuning of loss-balancing coefficients by dynamically learning optimal hyperparameter allocations during training. To infuse prior knowledge on independence of concept categories, in addition to standard reconstruction self-supervision we penalize correlation across concept subspaces. Crucially, to combat the instability of very low supervision regimes, D-OCB incorporates a dynamic dimensionality allocation mechanism; this adaptive formulation allows well-represented concepts to yield latent dimensions to underperforming concepts that are lagging behind, effectively preventing representation collapse and significantly improving overall concept accuracy. Through an extensive empirical evaluation, we demonstrate that our framework achieves high concept alignment and downstream visual reasoning accuracy using minimal label budgets, matching or outperforming end-to-end paradigms.

1. Introduction

Neuro-symbolic architectures for visual tasks allow to combine the feature extraction capabilities of deep neural networks (DNNs) on raw signals with interpretable and efficient symbolic reasoning (d'Avila Garcez and Lamb, 2023; Kishor, 2022). A central impediment within this paradigm is the classic symbol grounding problem, which asks how abstract symbols acquire semantic meaning beyond circular, text-based definitions and connect to perceptual boundaries in the physical world (Minsky, 1991). Specifically, feature extractors have to aggregate distributed features across visual space into coherent, unified object-centric representations (Brady et al., 2025). Two-stage neuro-symbolic architectures offer a compelling solution when symbol annotations are available, by structurally decoupling perception of symbols from reasoning on them (Mao et al., 2018; Manhaeve et al., 2018). In the first stage, a perception network maps raw inputs onto an explicit concept bottleneck layer; in the second stage, a symbolic engine executes logic programs directly over these

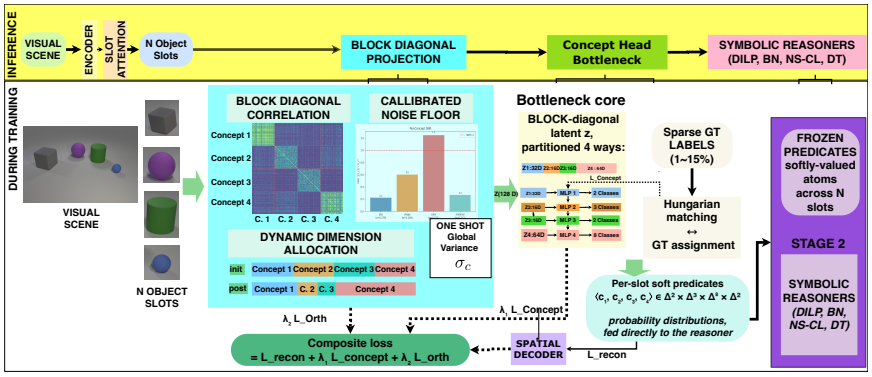

Figure 1: Overview of our D-OCB architecture and training scheme

discovered predicates. This modularity ensures that predictions are fully traceable, reduces susceptibility to shortcut learning, yields faster optimization convergence, and permits the reuse of a static reasoning backend across disparate visual domains (Kikaj et al., 2025).

Despite these advantages, traditional concept bottlenecks suffer from labeling requirements. Existing frameworks rely on full supervision to instantiate intermediate concepts, or turn to reinforcement learning settings with high sample variance (Dai and Wipf, 2019). The use of background knowledge or programmatic constraints to provide weak supervision, where only parts of the training data requires costly labels, remains significantly underexplored (Knab et al., 2026). This particularly is the case for object-centric concept bottleneck architectures and for the use in neuro-symbolic architectures.

To effectively utilize weak supervision for object-centric extraction, three concrete representational hurdles must first be resolved. First, generative bottlenecks trained with sparse labels frequently succumb to **representation collapse** (Jing et al., 2021), where unbounded variance optimization forces latent variables to degrade into uninformative noise. Second, without exhaustive dense annotations to explicitly isolate attributes, feature extractors naturally default to **feature cross-correlation**. This results in the entanglement of distinct conceptual properties (e.g., blending shape constraints into color representations) to exploit spurious visual shortcuts. Finally, traditional bottlenecks assign statically fixed vector sizes across all target concepts. This assumes uniform semantic complexity and inevitably causes **capacity misallocation** (Hou and Behdinan, 2022), where highly nuanced concepts are starved of representational bandwidth while trivial concepts waste feature dimensions.

To address these vulnerabilities, we introduce the Dynamic Orthogonal Concept Bottleneck (D-OCB [1]), illustrated in Figure 1. Rather than serving as a standalone, end-to-end reasoning architecture, D-OCB is explicitly framed as a weakly supervised predicate-grounding module (or perceptual bottleneck) designed specifically for two-stage neuro-symbolic pipelines. By structurally uncoupling perception from deduction, D-OCB anchors human-aligned symbolic predicates from minimal annotation budgets. It prevents representation collapse by anchoring per-concept variances against a calibrated noise floor alongside a variational auto-encoder reconstruction loss (Sutter et al., 2025). To combat feature cross-correlation, D-OCB introduces an explicit penalty that orthogonalizes subspaces between different concept categories. Crucially, to prevent capacity misallocation , D-OCB utilizes a novel dynamic dimensionality allocation mechanism that adaptively redistributes feature space size during training based on concept convergence; this ensures stable representations before matching concepts via the Hungarian algorithm. The result is a specialized percep-

1. The codebase is available at https://github.com/sparshX1993/Dynamic-Orthogonal-Concept-Bottleneck.

tion module capable of providing crisp, human-aligned predicate distributions to any frozen, off-the-shelf symbolic solver. Our key **contributions** amount to:

- We introduce D-OCB (cf. Figure 1), a weakly supervised, two-stage object-centric visual reasoning framework that grounds semantic predicates using as little as 1% of instance-level concept annotations.
- We propose a structural approach for stable concept learning under weak supervision that effectively prevents rank collapse inside each concept block.
- We formalize a dynamic dimensionality allocation protocol, entirely eliminating the need for manual latent-capacity tuning.
- We empirically demonstrate on the CLEVR and confounded CLEVR-Hans3/7 benchmarks that D-OCB predicates drive independent symbolic reasoners to accuracies competitive with end-to-end differentiable paradigms, adding as little as 1%-15% concept supervision to the training data.

2. Related Work

Our architecture is fundamentally motivated by the need to solve two classical cognitive hurdles simultaneously: the symbol grounding problem (Minsky, 1991), which asks how discrete symbols acquire physical semantic meaning from unstructured visual boundaries; and the binding problem (Brady et al., 2025), which concerns how distributed features across visual space are aggregated into unified, coherent entities. To achieve this, the design of our D-OCB intersects three major paradigms in contemporary neuro-symbolic artificial intelligence: concept bottleneck models (CBMs), object-centric representation learning, and differentiable logical rule induction.

CBMs and Generative Realignment. CBMs are interpretable architectures that factor downstream predictions through an intermediate bottleneck layer, where each individual output dimension is explicitly constrained to represent exactly one human-understandable concept(Knab et al., 2026; Kazhdan et al., 2021). Traditional formulations in this domain typically operate over holistic, image-level encodings and rely on dense, exhaustive supervision to map pixels onto categorical concepts (Knab et al., 2026; Losch et al., 2021). To mitigate this data-scarcity issue (Agrawal et al., 2025), contemporary frameworks have introduced semi-supervised (Hu et al., 2025) and completely label-free (Oikarinen et al., 2023) concept learning. Concurrently, efforts have been made to move beyond holistic representations by developing object-centric concept bottlenecks (Steinmann et al., 2025). Alongside these developments, recent frameworks have explored the integration of unsupervised generative constraints into the bottleneck layer (Locatello et al., 2019b). These generative approaches formalize how an image reconstruction loss can actively optimize concept alignment, prevent information leakage, and isolate target attributes (Kim and Mnih, 2018; Locatello et al., 2019a). D-OCB builds directly upon these generative and object-centric foundations but shifts the paradigm entirely to extremely sparse supervision regimes. By introducing a tuning-free dynamic allocation scheme and an explicit cross-correlation constraint, our work explores how balancing self-supervised reconstruction with minimal concept labels can preserve or even enhance grounding accuracy (Chen et al., 2019).

Object-Centric Tokenization and Slot-Based Bottlenecks. To perform relational and compositional reasoning in multi-object scenes, perception models require an object-

centric tokenization (Brady et al., 2025; Seitzer et al., 2023). A recent and superior approach is to use slot attention mechanisms which decompose a raw visual frame into distinct, non-overlapping slot tokens (Didolkar et al., 2025), encouraging spatial specialization and compositionality (Locatello et al., 2020). Despite their success in unsupervised scene decomposition (Mosbach et al., 2025), extending object-centric slots to interpretable concept layers introduces significant training instabilities (Liu et al., 2025). D-OCB resolves these limitations by introducing a strictly block-diagonal projection matrix as additional bias that forces uncorrelated concept subspaces.

Neuro-Symbolic Reasoning and Two-Stage Pipelines. Neuro-symbolic AI seeks to combine the robust perceptual pattern recognition of deep neural networks with the rigorous deductive compositionality of symbolic reasoning backends (Shindo et al., 2021a; Roth et al., 2025). While end-to-end logical frameworks like differentiable inductive logic programming (DILP) (Kikaj et al., 2025; Shindo et al., 2023, 2021b) optimize perception and deduction jointly via gradient descent, they remain highly vulnerable to Likelihood Blow-up (Mattei and Frellsen, 2018) and Posterior Collapse and representation contamination. To make use of available ground truth concept labels for interpretability, and to enable systematic reuse, D-OCB operates as a modular, two-stage neuro-symbolic framework that structurally uncouples perceptual tokenization from downstream symbolic solver engines (Dinu et al., 2024; Wüst et al., 2025).

3. Approach: Dynamic Orthogonal Concept Bottleneck

End-to-end differentiable reasoning frameworks are highly susceptible to feature entanglement (Marconato et al., 2023) and rely heavily on dense supervision. To overcome these inherent limitations, we propose a two-stage neuro-symbolic architecture that cleanly decouples object-centric visual perception from downstream logical reasoning. To ensure the perception module grounds continuous visual data into discrete, disentangled symbols under extreme weak supervision (1%–15%), we introduce D-OCB.

For a formal task definition, let $\mathcal{X} \in \mathbb{R}^{H\times W\times C}$ define the input space of raw visual scenes, and let $\mathcal{S} = (s_1, \ldots, s_K)$ represent an ordered specification of human-interpretable concept predicates. Our goal is to train an object-centric perception module $g : \mathcal{X} \to \mathcal{C}$ that maps an input image x to a discrete valuation tensor $c \in \mathbb{R}^{N\times K}$, representing the presence of K concepts across N distinct object slots, utilizing an extremely limited label budget. These grounded concepts are then passed into a standalone symbolic solver $f : \mathcal{C} \to \mathcal{Y}$ to deduce the final downstream task solution $y \in \mathcal{Y}$ (Muggleton, 1991).

3.1. Overview: Architecture and Optimization Objectives

We first establish the overall architecture for object-centric perceptual grounding, mapping raw pixels to discrete symbolic predicates in a step-by-step manner.An input scene $x \in \mathbb{R}^{h\times w\times c}$ of height h, width w, and c channels is processed by a feature extractor, such as DINOv2 (Oquab et al., 2024), augmented with positional embeddings. The resulting continuous feature maps are routed through an iterative slot attention module (Locatello et al., 2020) to enforce a tokenization prior, outputting N uninterpretable object slots:

$$H = \text{SlotAttention}(\text{Encoder}(x)) \in \mathbb{R}^{N\times D} \tag{1}$$

where D represents the hidden slot dimension (typically $D = 128$).

Following the extraction of these deterministic slots, the model maps the representations to a partitioned latent concept space $Z \in \mathbb{R}^{N \times D_{\text{total}}}$. This concept space has a total feature dimensionality D_{total}, which is explicitly divided into independent subspaces for each target concept $c \in \mathcal{C}$ (e.g., shape, color), such that $D_{\text{total}} = \sum_{c \in \mathcal{C}} D_c$. The specific dimensionality D_c allocated to each concept is not static; it is dynamically adjusted during training (as detailed in Section 3.3) to provide more capacity to struggling concepts.

To derive crisp concept predictions from these latent representations, each isolated concept subspace $z_c \in \mathbb{R}^{D_c}$ per slot is fed into an independent prediction head (a 2-layer MLP). These heads output raw class logits, which are converted into continuous probabilities via a softmax function (for multi-class concepts like color) or a sigmoid function (for binary concepts like material/Size). Finally, an arg max operation (or standard thresholding) discretizes these soft probabilities into the definitive, crisp symbolic predicates (e.g., `color(obj1, red)`) that are ultimately passed to the downstream logical reasoner.

To train this bottleneck effectively without manual hyperparameter engineering, we optimize an object-centric Variational Autoencoder (VAE) objective that balances a self-supervised reconstruction loss $\mathcal{L}_{\text{recon}}$, a concept prediction loss (Zbontar et al., 2021) $\mathcal{L}_{\text{concept}}$, and a structural cross-correlation loss[2] (Bardes et al., 2022) $\mathcal{L}_{\text{orth}}$:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{recon}} + \lambda_1 \mathcal{L}_{\text{concept}} + \lambda_2 \mathcal{L}_{\text{orth}} \tag{2}$$

The scaling coefficients λ_1 and λ_2 are dynamically updated at each epoch based on the moving average of gradient variances. Crucially, the cross-correlation loss $\mathcal{L}_{\text{orth}}$ minimizes the off-diagonal elements of the feature correlation matrix. This explicitly forces the learned concept subspaces to remain orthogonal and causally isolated from one another (Schölkopf et al., 2021; Suter et al., 2019), ensuring that weakly supervised concept signals are not overwhelmed by unconstrained reconstruction errors (Watters et al., 2019) and avoiding the feature interference common in standard dense projections (Goyal and Bengio, 2022; Kirsch et al., 2018).

3.2. Generative Bottleneck: Calibrated Fixed Noise Floors

A critical challenge in this architecture is the representation collapse common to generative bottlenecks (Heek et al., 2026). Generative models typically learn a per-sample variance ($\sigma(x)$). When annotation is scarce, the unsupervised reconstruction loss $\mathcal{L}_{\text{recon}}$ can force the network to map visual noise directly into this learned variance. This unbounded optimization causes a likelihood blow-up (Mattei and Frellsen, 2018), which leads to severe posterior collapse (Lucas et al., 2019a). Consequently, the latent variables become mathematically non-identifiable (Wang et al., 2021), effectively destroying the semantic manifold.

To prevent this, D-OCB discards per-sample learned variance entirely. Instead, we introduce a fixed noise floor σ_c specific to each concept subspace c:

$$z_c = \mu_c(x) + \sigma_c \odot \epsilon \tag{3}$$

Optimized globally during a brief one-shot calibration phase on the labeled subset, σ_c is permanently frozen. This renders the signal-to-noise ratio ($\text{SNR} = ||\mu_c||^2/\sigma_c^2$) mathematically stable and bars the decoder from encoding reconstruction noise into the variance.

2. For more detailed breakdown of our loss function and its mathematical formulations, see supplementary material.

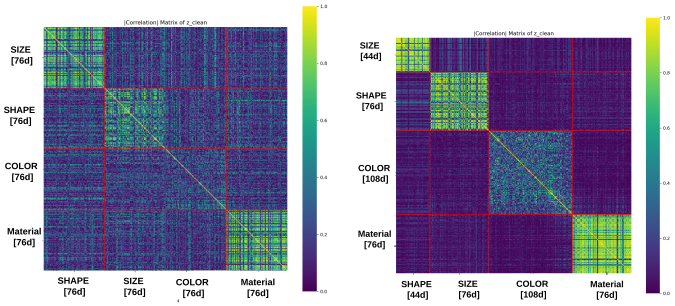

Figure 2: Correleation MATRIX for Clevr HANS7 Showing our DDA assign dimension from one concept to another for epoch 5 (left) to epoch 100 (right)

3.3. Orthogonality: Concept Assignment and Dimensionality Allocation

With the latent space cleanly orthogonalized by the cross-correlation loss and stabilized by the fixed noise floors, the remaining task is to map these dimensions to specific human-interpretable concepts and distribute capacity effectively.

First, to explicitly assign specific latent dimensions to their corresponding ground-truth concepts (e.g., matching a subspace to concept categories "shape" or "color"), alignments are secured via the Hungarian matching algorithm (Kuhn, 1955).

Second, to efficiently allocate latent capacity among these matched concepts without rank collapse (Noci et al., 2022), D-OCB replaces standard Singular Value Decomposition (Vasudevan and Ramakrishna, 2017) with Fisher's Linear Discriminant Analysis (Chen, 2017) (LDA). LDA projects running class centroids by computing a projection matrix $W \coloneqq \arg\max_W \frac{|W^T S_B W|}{|W^T S_W W|}$ that explicitly maximizes between-class scatter S_B relative to within-class scatter S_W. By maximizing class separability, LDA remains invariant to unsupervised reconstruction noise. Combined with Johnson-Lindenstrauss random projections for distance-preserving weight transfers, this dynamic dimensionality allocation ensures that underperforming concepts receive the necessary representational capacity without the need for manual latent-capacity tuning.

4. Experiments and Results

We empirically evaluate the representational fidelity, reasoning capability, and data efficiency of D-OCB, aiming to answer the following questions:
RQ1: Can D-OCB extract highly accurate and well-grounded visual predicates under extreme weak supervision (using only 1%–15% of the training dataset annotations)?
RQ2: Are the extracted predicates sufficiently stable to drive an offline symbolic solver without requiring joint end-to-end logic backpropagation?
RQ3: Does structurally decoupling perception from logical reasoning successfully prevent the exploitation of spurious correlations, thereby enabling D-OCB to achieve robust out-of-distribution (OOD) confounder resistance?

Datasets. We evaluate on the CLEVR object-attribute benchmark (70,000 training images) (Johnson et al., 2017) and its logical scene classification variants CLEVR-Hans3 (3,000 per class, 3 classes) and CLEVR-Hans7 (21,000 training images) (Stammer et al., 2020) and

a real world melanoma detection application using the HAM10000 dataset [3] (Tschandl et al., 2018) (8,012 training images). Hence, the considered sample budgets for 1%/15% weak supervision amount to a mere $700/10.3k$, $90/1.35k$, $210/3.15k$, $80/1.2k$ samples for CLEVR, CLEVR3, CLEVR7, and HAM10000 respectively. In all weakly supervised configurations, the remaining majority of the training pool remains completely unannotated.

Baselines. We benchmark our semantic tokenization pipeline against the following contemporary visual concept extraction architectures: the Concept Embedding Model (CEM) (Zarlenga et al., 2022); the Probabilistic Concept Bottleneck Model (ProbCBM) (Kim et al., 2023); SlotFormer (Biza et al., 2023) for temporal object-centric mechanics; a Standard Slot Attention baseline (Vanilla+SA); SlotVAE-NO-KL (a deterministic ablation); and SlotVAE-KL (Wang et al., 2024), optimized under standard Kullback-Leibler regularizers (Asperti and Trentin, 2020; Wang et al., 2023; Lucas et al., 2019b).To rigorously evaluate out-of-distribution (OOD) reasoning and confounder resistance, we compare our modular D-OCB (coupled with DILP) against end-to-end and dense-supervision paradigms. These include: an end-to-end trained CNN (0% concept supervision); a CNN + DILP baseline (100% supervision); state-of-the-art end-to-end differentiable frameworks (αILP (Shindo et al., 2023) and NEUMANN (Shindo et al., 2024)); DeepProbLog (Manhaeve et al., 2018) (trained jointly with 100% supervision); and a Ground Truth (Oracle) baseline.

Evaluation Metrics. Perceptual quality and grounding alignment are assessed via concept accuracy (acc.) and Macro-F1 scores, computed by aligning predicted object slots to ground-truth annotations via Hungarian matching (Kuhn, 1955). Downstream task execution is measured via rule evaluation acc. and F1-scores (Daniele et al., 2024).

Hyperparameters. The architecture is optimized over 200 epochs with a batch size of 64 and a base learning rate of 4×10^{-4} utilizing a cosine annealing schedule (Oquab et al., 2024). DDA capacity shifts occur every 5 epochs following an initial 30-epoch warmup gate, transferring features in discrete steps of $\Delta = 8$ dimensions to maintain block-diagonal alignment. To prioritize semantic grounding, the concept classification loss is heavily weighted ($\lambda_1 = 8.0$, scaling to 12.0 in later phases) relative to the reconstruction baseline ($\lambda_{recon} = 1.0$). Finally, the object existence head employs a focal loss ($\gamma = 2.0$, positive weight $= 2.5$) to successfully counter slot sparsity.

4.1. RQ1: Visual Concept Extraction and Predicate Accuracy

To evaluate the representational quality of our perception module, we measure its capability to ground continuous visual features into discrete symbolic predicates using the CLEVR and HAM10000 datasets under extremely weak supervision regimes (1% to 15%). Unlike traditional slot models that frequently exhibit performance collapse on complex attributes due to unconstrained reconstruction objectives hijacking the latent space topology, our D-OCB explicitly prevents features from dissolving into unstructured noise.

As shown in Table 2 and Table 1, D-OCB significantly outperforms baselines, amongst others by the combination of VAE reconstruction loss (difference to NoVAE), and by dynamically assigning dimensions to concepts lacking accuracy and discarding per-sample

3. find details on the choice of datasets in the technical supplementary

Table 1: Concept classification acc. on the CLEVR dataset across varying levels of weak supervision. Results are reported as mean ± standard deviation over 5 folds.

		CNN Backbone					DINO Backbone				
Sup.	**Method**	Size	Shape	Color	Material	Mean	Size	Shape	Color	Material	Mean
1%	Vanilla+SA	51.2 ± 0.0	49.3 ± 0.0	10.7 ± 0.0	62.2 ± 0.0	46.4 ± 0.0	72.3 ± 0.0	63.4 ± 0.0	15.8 ± 0.0	70.6 ± 0.0	57.3 ± 0.0
	CEM	51.2 ± 1.3	33.9 ± 0.5	17.2 ± 6.3	50.9 ± 0.5	38.3 ± 2.1	73.4 ± 2.2	41.3 ± 3.9	12.3 ± 0.1	70.8 ± 6.6	49.5 ± 2.6
	ProbCBM	48.2 ± 3.9	34.5 ± 2.1	20.9 ± 8.2	50.7 ± 0.7	38.6 ± 3.3	72.3 ± 4.4	38.4 ± 2.2	12.5 ± 0.3	67.0 ± 7.4	47.5 ± 3.2
	SlotFormer	25.2 ± 3.1	25.0 ± 2.0	17.5 ± 5.1	50.1 ± 1.1	29.4 ± 1.7	48.0 ± 2.8	38.3 ± 2.8	17.7 ± 3.9	55.0 ± 3.1	39.8 ± 2.1
	SlotVAE-NoVAE	52.8 ± 0.0	51.5 ± 0.0	8.9 ± 0.0	60.5 ± 0.0	48.6 ± 0.0	76.2 ± 0.0	62.6 ± 0.0	12.1 ± 0.0	75.6 ± 0.0	56.9 ± 0.0
	SlotVAE-KL	62.0 ± 4.8	64.0 ± 6.6	16.9 ± 3.7	79.2 ± 1.4	55.0 ± 3.2	90.5 ± 6.7	80.7 ± 6.4	**38.3 ± 0.2**	93.6 ± 3.2	75.3 ± 3.3
	D-OCB (Ours)	**75.1 ± 16.1**	**67.5 ± 3.4**	**28.2 ± 5.1**	**84.5 ± 5.0**	**63.8 ± 7.1**	**94.4 ± 1.4**	**82.3 ± 8.6**	36.7 ± 6.3	**94.9 ± 2.3**	**77.1 ± 4.5**
5%	CEM	53.5 ± 1.1	35.9 ± 0.6	24.5 ± 0.5	51.9 ± 1.0	41.5 ± 0.5	80.2 ± 2.9	48.7 ± 1.5	12.4 ± 0.2	82.7 ± 2.9	56.0 ± 1.0
	ProbCBM	48.1 ± 5.2	34.5 ± 1.5	25.0 ± 0.9	50.9 ± 0.7	39.6 ± 1.6	74.6 ± 5.5	46.8 ± 4.9	12.6 ± 0.4	79.8 ± 3.9	53.5 ± 3.1
	SlotFormer	26.7 ± 2.7	26.1 ± 3.2	24.8 ± 0.5	50.3 ± 0.8	32.0 ± 0.9	46.4 ± 3.1	43.9 ± 4.1	14.1 ± 1.8	63.2 ± 3.9	41.9 ± 2.4
	SlotVAE-KL	85.8 ± 5.2	68.5 ± 4.5	43.0 ± 2.3	**91.9 ± 1.6**	72.3 ± 2.6	91.0 ± 7.9	87.3 ± 5.9	62.3 ± 0.1	96.6 ± 5.6	84.1 ± 4.1
	D-OCB (Ours)	**91.8 ± 3.3**	**81.6 ± 1.5**	**58.5 ± 2.3**	91.5 ± 1.8	**80.8 ± 1.0**	**97.6 ± 0.0**	**95.1 ± 1.1**	**75.4 ± 1.5**	**97.1 ± 0.1**	**91.3 ± 0.4**
15%	Vanilla+SA	73.3 ± 0.0	66.5 ± 0.0	59.3 ± 0.0	69.7 ± 0.0	74.1 ± 0.0	77.8 ± 0.0	78.0 ± 0.0	63.9 ± 0.0	78.9 ± 0.0	70.8 ± 0.0
	CEM	61.5 ± 2.7	37.0 ± 1.2	24.7 ± 1.1	53.0 ± 0.9	44.0 ± 1.2	86.2 ± 1.2	54.0 ± 1.5	12.7 ± 0.4	86.7 ± 0.8	60.0 ± 0.4
	ProbCBM	51.4 ± 4.6	34.5 ± 1.1	25.0 ± 0.5	51.4 ± 0.8	40.6 ± 1.3	81.6 ± 3.7	48.6 ± 4.8	13.8 ± 1.6	82.7 ± 2.0	56.7 ± 2.5
	SlotFormer	26.1 ± 3.8	25.7 ± 3.3	24.9 ± 0.2	50.0 ± 1.0	31.7 ± 1.2	47.3 ± 2.7	45.6 ± 4.9	16.7 ± 3.7	65.5 ± 3.6	43.8 ± 2.7
	SlotVAE-NoVAE	75.5 ± 0.0	65.8 ± 0.0	61.2 ± 0.0	74.8 ± 0.0	66.5 ± 0.0	79.8 ± 0.0	78.8 ± 0.0	64.1 ± 0.0	76.2 ± 0.0	77.3 ± 0.0
	SlotVAE-KL	91.6 ± 5.7	74.4 ± 3.7	52.9 ± 1.8	84.2 ± 2.1	75.8 ± 2.4	**97.6 ± 8.2**	92.6 ± 4.3	74.5 ± 0.1	95.9 ± 6.4	89.2 ± 4.0
	D-OCB (Ours)	**94.5 ± 1.6**	**88.6 ± 2.2**	**84.0 ± 0.7**	**93.4 ± 1.3**	**90.1 ± 0.9**	**97.6 ± 0.0**	**95.8 ± 0.3**	**84.6 ± 0.6**	**97.2 ± 0.1**	**93.8 ± 0.1**

Table 2: Concept classification acc. of **D-OCB** on the HAM10000 dataset across varying levels of weak supervision. Results are reported as mean ± standard deviation over five folds.

Backbone	**Sup. (%)**	**Dx**	**Localization**	**Sex**	**Age**	**Mean**
CNN	1	55.3 ± 0.0	84.2 ± 0.0	22.1 ± 0.0	50.0 ± 0.0	52.9 ± 0.0
	5	55.0 ± 0.0	84.4 ± 0.0	60.6 ± 0.0	61.7 ± 0.0	65.4 ± 0.0
	15	59.3 ± 0.0	86.0 ± 0.0	64.3 ± 0.0	64.6 ± 0.0	68.6 ± 0.0
DINO	1	61.8 ± 0.0	82.5 ± 0.0	67.4 ± 0.0	60.2 ± 0.0	68.0 ± 0.0
	5	64.4 ± 0.0	78.5 ± 0.0	69.4 ± 0.0	69.1 ± 0.0	70.4 ± 0.0
	15	64.2 ± 0.0	85.5 ± 0.0	72.8 ± 0.0	70.9 ± 0.0	73.4 ± 0.0

learned variance (difference to SlotVAE-KL). E.g., it maintains a 73.4% mean accuracy on HAM10000 with a DINO backbone at just 15% supervision. D-OCB thus achieves competitive accurate concept alignment without the excessive need for labels and without manual hyperparameter tuning.*Empirically, D-OCB actively prevents representation collapse to extract highly accurate semantic concepts using minimal label budgets, significantly outperforming unconstrained generative baselines giving a positive answer to* ***RQ1****.*

4.2. RQ2: Framework-Agnostic Logical Reasoning Evaluation

We now investigate whether the accurately extracted visual atoms are sufficiently stable to drive downstream symbolic reasoning engines—such as DILP (Shindo et al., 2021b), Decision Trees, Bayesian Networks, and the Neural-Symbolic Concept Learner (NS-CL) (Mao et al., 2018)—without requiring joint end-to-end logic backpropagation. To rigorously test this, we export the frozen valuation tensors generated by D-OCB to evaluate specific compositional and diagnostic rules, as detailed in Table 3.

The empirical results in Table 5 and Table 4 substantiate the universal compatibility and representational stability of D-OCB[4] . Because the softly-valued probabilistic margins generated by D-OCB accurately preserve semantic boundaries even under minimal supervision budgets (1% to 15%), they seamlessly drive diverse, off-the-shelf symbolic reasoning frontends—spanning statistical methods (BN) (Pearl, 1988), rule-based decision trees (DT)(Breiman et al., 1984), and differentiable logic frameworks (DILP, NS-CL(Mao et al., 2019))—to highly accurate downstream task performance. *Our results confirm that the softly-valued probabilistic margins generated by D-OCB maintain strict semantic boundaries, seamlessly driving diverse, frozen symbolic reasoners to accurate downstream performance.*

4. For full table with all rules see supplementary material.

Table 3: Downstream evaluation rules for the CLEVR and HAM10000 datasets.

	ID	Rule Type	Description
CLEVR	**R1**	4-way Conjunction	The scene contains an object that is simultaneously large, red, metal, and a cube.
	R2	3-way Conjunction	The scene contains a small rubber sphere.
	R3	Spatial Relation	A blue cylinder is spatially to the left of a red object (evaluated via 3D x-coordinates).
	R4	Cardinality	The scene contains ≥ 2 metal objects; requires the model to count distinct slots satisfying this constraint.
	R5	Proximity Relation	A large green object is within a continuous Euclidean distance threshold (≤ 2.5) of a yellow object.
HAM	**R1**	Simple	The lesion is malignant.
	R2	Disjunctive	The lesion is melanoma (mel) or basal cell carcinoma (bcc).
	R3	Conjunctive	The lesion is melanoma located on the back (a diagnosis and location conjunction).

Table 4: Downstream reasoning acc. (%) averaged across rules R1-R5 on the CLEVR dataset. Results over multiple folds (complete table in the supplementary)

Backb.	Percept. Model	Decision Tree (DT)					Bayes Net (BN)					NS-CL					DILP				
		1%	15%	25%	50%	75%	1%	15%	25%	50%	75%	1%	15%	25%	50%	75%	1%	15%	25%	50%	75%
CNN	**D-OCB (Ours)**	**63.1**	**70.8**	**72.0**	**71.9**	69.8	**85.0**	**71.6**	72.1	71.8	71.8	**62.2**	**71.0**	**71.5**	71.3	71.2	67.1	**97.0**	**97.1**	**97.4**	**97.1**
	SlotVAE-KL	56.3	33.3	66.6	67.7	**71.6**	52.9	32.7	**83.3**	**84.5**	**85.5**	45.9	31.4	50.1	**74.0**	**74.2**	45.0	47.1	69.5	80.8	82.1
	SlotVAE-NoVAE	55.6	48.3	59.8	68.1	69.4	69.3	53.6	70.0	60.2	61.5	61.4	56.7	66.6	72.2	72.9	**70.9**	74.6	89.0	95.6	96.0
	Vanilla+SA	51.3	51.8	54.1	55.5	58.9	51.4	53.6	53.9	53.8	52.4	**62.2**	60.3	63.8	62.7	66.8	64.2	76.1	85.2	85.2	92.1
DINO	**D-OCB (Ours)**	**67.9**	68.3	**70.3**	**74.0**	69.9	55.6	71.1	71.0	71.0	71.0	70.0	70.8	70.6	70.6	70.8	93.9	**96.9**	**97.3**	97.4	**97.6**
	SlotVAE-KL	65.4	**72.8**	67.5	69.9	**70.5**	**86.6**	**72.2**	**87.7**	**88.1**	**87.5**	64.8	**71.8**	71.4	70.5	70.6	71.7	95.3	81.8	82.0	81.0
	SlotVAE-NoVAE	64.8	68.0	69.2	66.4	66.6	69.3	71.5	72.0	71.4	71.5	**71.4**	71.4	**71.8**	**71.0**	**71.0**	**95.9**	96.8	97.1	**97.6**	**97.6**
	Vanilla+SA	62.2	68.8	68.9	68.5	66.3	68.6	71.3	72.0	70.7	71.5	65.3	71.0	71.3	70.5	70.8	90.6	96.8	96.2	96.5	96.5

4.3. RQ3: Confounder Resistance and Out-of-Distribution Reasoning

A critical vulnerability in end-to-end differentiable logic frameworks is their susceptibility to dataset biases, where the loss from the logical reasoner propagates directly into the perception layer. This phenomenon encourages the perception module to entangle distinct concepts—such as 'shape' and 'color'—to exploit spurious correlations in the training data rather than learning the true underlying causal rules. To evaluate this hypothesis, we utilize the confounded splits of the CLEVR-Hans3 and CLEVR-Hans7 benchmarks (Stammer et al., 2021), where the training sets contain deliberate confounders (e.g., specific shapes artificially co-occurring with specific colors) that are subsequently removed in the out-of-distribution (OOD) test splits. We compare our D-OCB coupled with Differentiable Inductive Logic Programming (DILP) against several end-to-end, two-stage and a Ground Truth (Oracle) baseline. Unlike fully end-to-end neuro-symbolic networks that backpropagate logic-validation errors directly into the perception backbone,

D-OCB structurally decouples these processes by extracting softly-valued visual predicates under weak supervision (1% to 75%) for frozen downstream evaluation. As Table 6 demonstrates, this synergy of object-centric biases and symbolic reasoning effectively suppresses likelihood blowup. Furthermore, our CLEVR-Hans7 analysis confirms that Dynamic Dimensionality Allocation (DDA) autonomously optimizes concept subspace capacities to prevent representation bottlenecks[5]. Even at minimal supervision budgets (1% to 5%), D-OCB paired with DILP (utilizing DINO) achieves robust out-of-distribution (OOD) accuracy with marginal generalization gaps (1.1 ± 0.8 on CLEVR-Hans3; 3.00 ± 0.56 on CLEVR-Hans7), matching densely supervised baselines and resisting confounders without explicit annotations. *In **RQ3**, by isolating the perceptual bottleneck, D-OCB successfully suppresses likelihood blowup and ignores spurious training correlations, achieving robust OOD generalization that matches densely supervised paradigms.*

5. Conclusion

In this work, we proposed the Dynamic Orthogonal Concept Bottleneck (D-OCB) neural network, a visual feature extractor for neuro-symbolic reasoning that can leverage extremely

5. For detailed cross-validation results, hyperparameter configurations, and full DDA ablation see supplementary material.

Table 5: Reasoning acc. (%) of **D-OCB** vs. **Slot-VAE** baseline on HAM10000 (RQ2) for three diagnostic rules across different backbones, supervision levels, and reasoning frameworks. Baseline F1 scores have been converted to % for direct comparison. Mean ± std. dev. over 5 folds.

BB	Sup%	Model	DILP (%)			Decision Tree (%)			Bayes Net (%)			NS-CL (%)		
			R1	R2	R3	R1	R2	R3	R1	R2	R3	R1	R2	R3
CNN	1	**D-OCB (Ours)**	**31.8 ± 0.0**	**34.8 ± 0.0**	**86.6 ± 0.0**	**59.7 ± 0.0**	**69.2 ± 0.0**	**72.6 ± 0.0**	**76.6 ± 0.0**	**82.6 ± 0.0**	**92.5 ± 0.0**	**52.7 ± 0.0**	**60.2 ± 0.0**	**92.5 ± 0.0**
		Slot-VAE	-	-	-	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0	15.6 ± 1.0	18.6 ± 1.0	2.6 ± 1.0	0.0 ± 0.0	0.0 ± 0.0	0.0 ± 0.0
	15	**D-OCB (Ours)**	**69.7 ± 0.0**	**72.6 ± 0.0**	**80.1 ± 0.0**	**63.2 ± 0.0**	**58.2 ± 0.0**	**72.1 ± 0.0**	**84.6 ± 0.0**	**82.6 ± 0.0**	**95.0 ± 0.0**	**69.2 ± 0.0**	**64.2 ± 0.0**	**84.1 ± 0.0**
		Slot-VAE	-	-	-	27.2 ± 3.0	15.4 ± 3.0	0.4 ± 1.0	30.0 ± 1.0	18.5 ± 1.0	0.0 ± 0.0	43.3 ± 0.0	37.9 ± 0.0	8.8 ± 0.0
	75	**D-OCB (Ours)**	**82.6 ± 2.5**	**82.3 ± 2.2**	**92.4 ± 3.3**	**79.9 ± 2.3**	**73.8 ± 3.9**	**84.6 ± 1.6**	**86.4 ± 2.3**	**83.9 ± 0.5**	**93.5 ± 1.1**	**76.3 ± 1.6**	**73.1 ± 3.9**	**89.7 ± 2.9**
		Slot-VAE	-	-	-	40.2 ± 2.0	18.2 ± 1.0	0.0 ± 0.0	49.4 ± 1.0	36.8 ± 1.0	0.0 ± 0.0	50.9 ± 0.0	43.5 ± 0.0	18.9 ± 0.0
DINO	1	**D-OCB (Ours)**	**85.1 ± 0.0**	**78.1 ± 0.0**	**88.6 ± 0.0**	**73.6 ± 0.0**	**64.7 ± 0.0**	**78.1 ± 0.0**	19.9 ± 0.0	**79.6 ± 0.0**	**90.0 ± 0.0**	**84.1 ± 0.0**	**73.1 ± 0.0**	**92.0 ± 0.0**
		Slot-VAE	-	-	-	37.0 ± 3.0	14.4 ± 3.0	0.0 ± 0.0	**37.2 ± 1.0**	23.6 ± 2.0	0.0 ± 0.0	48.9 ± 0.0	41.9 ± 0.0	10.4 ± 0.0
	15	**D-OCB (Ours)**	**87.1 ± 0.0**	**82.1 ± 0.0**	**94.5 ± 0.0**	**68.7 ± 0.0**	**61.7 ± 0.0**	**70.6 ± 0.0**	**87.1 ± 0.0**	**82.1 ± 0.0**	**91.5 ± 0.0**	**79.6 ± 0.0**	**65.2 ± 0.0**	**92.5 ± 0.0**
		Slot-VAE	-	-	-	62.5 ± 2.0	55.1 ± 2.0	0.0 ± 0.0	37.3 ± 2.0	10.7 ± 2.0	0.0 ± 0.0	65.2 ± 0.0	60.4 ± 0.0	25.5 ± 0.0
	75	**D-OCB (Ours)**	**87.6 ± 2.1**	**87.4 ± 2.9**	**93.2 ± 1.3**	**83.1 ± 5.7**	**78.6 ± 6.8**	**90.9 ± 1.2**	**90.7 ± 1.9**	**87.9 ± 2.7**	**93.5 ± 1.1**	**85.9 ± 2.4**	**84.4 ± 4.9**	**92.5 ± 0.7**
		Slot-VAE	-	-	-	70.9 ± 1.0	66.3 ± 2.0	11.8 ± 4.0	47.3 ± 3.0	19.8 ± 2.0	0.0 ± 0.0	70.7 ± 0.0	66.7 ± 0.0	31.0 ± 0.0

Table 6: Confounder Resistance on CLEVR-Hans3 and CLEVR-Hans7. Comparison of D-OCB against end-to-end architectures and dense-supervision baselines. We report train acc. (in-distribution), test acc. (out-of-distribution), and the OOD Gap (↓). Mean ± std. dev. over 5 folds.

CLEVR-Hans3 Benchmark

Method	Sup.	Train Acc ↑	Test Acc ↑	OOD Gap ↓
E2E CNN Basel.	0%	100.0 ± 0.0	70.3 ± 0.0	29.7 ± 0.0
NeSy	100%	98.5 ± 0.0	81.7 ± 0.0	16.8 ± 0.0
NeSy-XIL	100%	100.0 ± 0.0	91.3 ± 0.0	8.7 ± 0.0
αILP	100%	97.5 ± 0.0	97.5 ± 0.0	−0.0 ± 0.0
NEUMANN	100%	96.7 ± 0.0	97.4 ± 0.0	−0.8 ± 0.0
D-OCB + DILP (DINO)	1%	72.8 ± 18.6	71.2 ± 18.2	1.6 ± 0.9
	5%	98.2 ± 0.4	97.0 ± 0.4	1.1 ± 0.8
	15%	**98.7 ± 0.4**	**98.4 ± 0.6**	**0.3 ± 0.3**
D-OCB + DILP (CNN)	1%	36.0 ± 2.3	35.3 ± 0.7	0.8 ± 1.7
	15%	73.9 ± 1.7	72.5 ± 0.9	1.4 ± 2.1
	75%	98.3 ± 0.3	97.3 ± 0.6	1.0 ± 0.8
GT O. + DILP	100%	100.0 ± 0.0	100.0 ± 0.0	0.0 ± 0.0

CLEVR-Hans7 Benchmark

Method	Sup.	Train Acc ↑	Test Acc ↑	OOD Gap ↓
E2E CNN Basel.	0%	69.13 ± 0.00	51.43 ± 0.00	17.70 ± 0.00
CNN Baseline	100%	70.79 ± 0.00	68.46 ± 0.00	2.33 ± 0.00
DeepProbLog	100%	85.35 ± 0.00	67.98 ± 0.00	17.37 ± 0.00
D-OCB + DILP (DINO)	1%	80.86 ± 2.28	78.55 ± 2.57	2.31 ± 0.41
	5%	90.60 ± 0.12	87.60 ± 0.56	3.00 ± 0.56
	15%	92.43 ± 0.16	89.07 ± 0.27	3.36 ± 0.37
D-OCB + DILP (CNN)	1%	22.76 ± 4.16	22.21 ± 3.99	0.55 ± 0.44
	15%	76.91 ± 3.61	69.82 ± 4.49	7.09 ± 0.88
	75%	92.30 ± 0.32	87.26 ± 0.32	5.05 ± 0.43
GT O. + DILP	100%	**94.21 ± 0.00**	**93.56 ± 0.00**	**0.65 ± 0.00**

weak supervision (1% to 15%) for human-alignment of the predicates. Crucially, D-OCB adds self-supervision via a reconstruction and a cross-correlation orthogonal loss, paired with a novel dynamic allocation of extracted features to the labeled human-interpretable concepts. This adaptive mechanism actively redistributes latent capacity, transferring dimensions from well-represented concepts to those lacking accuracy. Our empirical evaluations on CLEVR, CLEVR-Hans, and HAM10000 demonstrate that a trained D-OCB's concept predictions are not only superior in low supervision accuracy, but also seamlessly drive diverse off-the-shelf symbolic reasoners to near-oracle performance, even without joint backpropagation. Furthermore, by the structural decoupling of perception from logical induction D-OCB-based neuro-symbolic pipelines achieve state-of-the-art out-of-distribution generalization against confounding factors. Ultimately, D-OCB establishes a highly stable, efficient, and computationally accessible visual-logical primitive layer for complex downstream reasoning tasks.

As a remaining limitation, the architecture currently relies on a rigid slot attention mechanism, which necessitates the manual, a priori definition of the maximum number of object slots. Additionally, while the latent vector compression effectively isolates structural concepts, it inherently filters out fine-grained details, which can degrade performance on highly textured datasets where critical information is lost during bottleneck projection. Future work can address these constraints by exploring closed-loop abductive learning, wherein the downstream System 2 logical reasoner provides top-down error corrections to iteratively refine the System 1 perceptual module.

Acknowledgments

G.S. and S.T. acknowledge support through the project "chAI" funded by the German Federal Ministry of Research, Technology and Space (BMFTR), grant no. 16IS24058.

References

Susmit Agrawal, Deepika Vemuri, Sri Siddarth Chakaravarthy P, and Vineeth N. Balasubramanian. Walking the web of concept-class relationships in incrementally trained interpretable models, 2025. URL https://arxiv.org/abs/2502.20393.

Andrea Asperti and Matteo Trentin. Balancing reconstruction error and kullback-leibler divergence in variational autoencoders, 2020. URL https://arxiv.org/abs/2002.07514.

Adrien Bardes, Jean Ponce, and Yann LeCun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *International Conference on Learning Representations*, 2022.

Ondrej Biza, Sjoerd van Steenkiste, Mehdi S. M. Sajjadi, Gamaleldin Fathy Elsayed, Aravindh Mahendran, and Thomas Kipf. Invariant slot attention: Object discovery with slot-centric reference frames. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 2507–2527. PMLR, 2023. URL https://proceedings.mlr.press/v202/biza23a.html.

Jack Brady, Julius von Kügelgen, Sébastien Lachapelle, Simon Buchholz, Thomas Kipf, and Wieland Brendel. Interaction asymmetry: A general principle for learning composable abstractions. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=cCl10IU836.

Leo Breiman, Jerome Friedman, Richard A Olshen, and Charles J Stone. *Classification and regression trees*. Routledge, 1984.

Gang Chen. Latent discriminant analysis with representative feature discovery. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31, 02 2017. doi: 10.1609/aaai.v31i1.10879.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in variational autoencoders, 2019. URL https://arxiv.org/abs/1802.04942.

Bin Dai and David Wipf. Diagnosing and enhancing vae models, 2019. URL https://arxiv.org/abs/1903.05789.

Alessandro Daniele, Tommaso Campari, Sagar Malhotra, and Luciano Serafini. Simple and effective transfer learning for neuro-symbolic integration, 2024. URL https://arxiv.org/abs/2402.14047.

Artur d'Avila Garcez and Luís C. Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, November 2023. ISSN 1573-7462. doi: 10.1007/s10462-023-10448-w.

Aniket Didolkar, Andrii Zadaianchuk, Rabiul Awal, Maximilian Seitzer, Efstratios Gavves, and Aishwarya Agrawal. Ctrl-o: Language-controllable object-centric visual representation learning, 2025. URL https://arxiv.org/abs/2503.21747.

Marius-Constantin Dinu, Claudiu Leoveanu-Condrei, Markus Holzleitner, Werner Zellinger, and Sepp Hochreiter. Symbolicai: A framework for logic-based approaches combining generative models and solvers, 2024. URL https://arxiv.org/abs/2402.00854.

Anirudh Goyal and Yoshua Bengio. Inductive biases for deep learning of higher-level cognition, 2022. URL https://arxiv.org/abs/2011.15091.

Jonathan Heek, Emiel Hoogeboom, Thomas Mensink, and Tim Salimans. Unified latents (ul): How to train your latents, 2026. URL https://arxiv.org/abs/2602.17270.

Chun Kit Jeffery Hou and Kamran Behdinan. Dimensionality reduction in surrogate modeling: A review of combined methods. *Data Science and Engineering*, 7(4): 402–427, 2022. doi: 10.1007/s41019-022-00193-5. URL https://doi.org/10.1007/s41019-022-00193-5.

Lijie Hu, Tianhao Huang, Huanyi Xie, Xilin Gong, Chenyang Ren, Zhengyu Hu, Lu Yu, Ping Ma, and Di Wang. Semi-supervised concept bottleneck models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2110–2119, October 2025.

Li Jing, Pascal Vincent, Yann LeCun, and Yuandong Tian. Understanding dimensional collapse in contrastive self-supervised learning. *CoRR*, abs/2110.09348, 2021. URL https://arxiv.org/abs/2110.09348.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross B. Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1988–1997. IEEE Computer Society, 2017. doi: 10.1109/CVPR.2017.215. URL https://doi.org/10.1109/CVPR.2017.215.

Dmitry Kazhdan, Botty Dimanov, Helena Andres Terre, Mateja Jamnik, Pietro Liò, and Adrian Weller. Is disentanglement all you need? Comparing concept-based & disentanglement approaches. *CoRR*, abs/2104.06917, April 2021.

Adem Kikaj, Giuseppe Marra, Floris Geerts, Robin Manhaeve, and Luc De Raedt. DeepGraphLog for Layered Neurosymbolic AI. In *ECAI 2025*, pages 1551–1558. IOS Press, 2025. doi: 10.3233/FAIA250979.

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models, 2023. URL https://arxiv.org/abs/2306.01574.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *Proc. 2018 Int. Conf. Machine Learning*, pages 2649–2658. PMLR, July 2018.

Louis Kirsch, Julius Kunesch, and David Barber. Modular networks: Learning to decompose neural computation. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Rabinandan Kishor. Neuro-symbolic ai: Bringing a new era of machine learning. *International Journal of Research Publication and Reviews*, 2022. URL https://api.semanticscholar.org/CorpusID:260698277.

Patrick Knab, David Steinmann, Christian Bartelt, Kristian Kersting, Bernt Schiele, Thomas Seidl, Udo Schlegel, and Wolfgang Stammer. What's in the bottle? a survey and roadmap of concept bottleneck models. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=IF5vnqxBEW.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955. URL https://api.semanticscholar.org/CorpusID:9426884.

Hongjia Liu, Rongzhen Zhao, Haohan Chen, and Joni Pajarinen. Metaslot: Break through the fixed number of slots in object-centric learning. *CoRR*, abs/2505.20772, 2025. doi: 10.48550/ARXIV.2505.20772. URL https://doi.org/10.48550/arXiv.2505.20772.

Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the Fairness of Disentangled Representations. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019a.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4114–4124. PMLR, May 2019b.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-Centric Learning with Slot Attention. In *Advances in Neural Information Processing Systems*, volume 33, pages 11525–11538. Curran Associates, Inc., 2020.

Max Losch, Mario Fritz, and Bernt Schiele. Semantic Bottlenecks: Quantifying and Improving Inspectability of Deep Representations. *International Journal of Computer Vision*, 129(11):3136–3153, November 2021. ISSN 1573-1405. doi: 10.1007/s11263-021-01498-0.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse. In *Advances in Neural Information Processing Systems*, volume 32, 2019a.

James Lucas, George Tucker, Roger Grosse, and Mohammad Norouzi. Don't blame the elbo! a linear vae perspective on posterior collapse, 2019b. URL https://arxiv.org/abs/1911.02469.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. Deepproblog: Neural probabilistic logic programming, 2018. URL https://arxiv.org/abs/1805.10872.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *Int. Conf. Learning Representations*, September 2018.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *International Conference on Learning Representations*, 2019.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts, 2023. URL https://arxiv.org/abs/2305.19951.

Pierre-Alexandre Mattei and Jes Frellsen. Leveraging the exact likelihood of deep latent variable models. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Marvin Minsky. Logical versus analogical or symbolic versus connectionist or neat versus scruffy. *AI Mag.*, 12:34–51, 1991. URL https://api.semanticscholar.org/CorpusID:28339188.

Malte Mosbach, Jan Niklas Ewertz, Angel Villar-Corrales, and Sven Behnke. Sold: Slot object-centric latent dynamics models for relational manipulation learning from pixels, 2025. URL https://arxiv.org/abs/2410.08822.

Stephen Muggleton. Inductive logic programming. *New Generation Computing*, 8(4):295–318, February 1991. ISSN 1882-7055. doi: 10.1007/BF03037089.

Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse, 2022. URL https://arxiv.org/abs/2206.03126.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Judea Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan kaufmann, 1988.

Stephen Roth, Lennart Baur, Derian Boer, and Stefan Kramer. Enhancing Symbolic Machine Learning by Subsymbolic Representations, June 2025.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954.

Maximilian Seitzer, Max Horn, Andrii Zadaianchuk, Dominik Zietlow, Tianjun Xiao, Carl-Johann Simon-Gabriel, Tong He, Zheng Zhang, Bernhard Schölkopf, Thomas Brox, and Francesco Locatello. Bridging the gap to real-world object-centric learning, 2023. URL https://arxiv.org/abs/2209.14860.

Hikaru Shindo, Devendra Singh Dhami, and Kristian Kersting. Neuro-symbolic forward reasoning, 2021a. URL https://arxiv.org/abs/2110.09383.

Hikaru Shindo, Masaaki Nishino, and Akihiro Yamamoto. Differentiable inductive logic programming for structured examples. *CoRR*, abs/2103.01719, 2021b. URL https://arxiv.org/abs/2103.01719.

Hikaru Shindo, Viktor Pfanschilling, Devendra Singh Dhami, and Kristian Kersting. αILP: thinking visual scenes as differentiable logic programs. *Machine Learning*, 112(4):1465–1497, 2023.

Hikaru Shindo, Viktor Pfanschilling, Devendra Singh Dhami, and Kristian Kersting. Learning differentiable logic programs for abstract visual reasoning. *Machine Learning*, 113, 2024.

Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. *CoRR*, abs/2011.12854, 2020. URL https://arxiv.org/abs/2011.12854.

Wolfgang Stammer, Patrick Schramowski, and Kristian Kersting. Right for the right concept: Revising neuro-symbolic concepts by interacting with their explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3619–3629, 2021.

David Steinmann, Wolfgang Stammer, Antonia Wüst, and Kristian Kersting. Object-centric concept-bottlenecks. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, pages 68899–68924. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/639d992f819c2b40387d4d5170b8ffd7-Paper-Conference.pdf.

Raphael Suter, Dorde Miladinović, Bernhard Schölkopf, and Stefan Bauer. Robustly disentangled causal mechanisms: Validating deep representations for interventional robustness, 2019. URL https://arxiv.org/abs/1811.00007.

Thomas M. Sutter, Yang Meng, Andrea Agostini, Daphné Chopard, Norbert Fortin, Julia E. Vogt, Babak Shahbaba, and Stephan Mandt. Unity by diversity: Improved representation learning in multimodal vaes, 2025. URL https://arxiv.org/abs/2403.05300.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset: A large collection of multi-source dermatoscopic images of common pigmented skin lesions. *CoRR*, abs/1803.10417, 2018. URL http://arxiv.org/abs/1803.10417.

Vinita Vasudevan and M. Ramakrishna. A hierarchical singular value decomposition algorithm for low rank matrices. *CoRR*, abs/1710.02812, 2017. URL http://arxiv.org/abs/1710.02812.

Yanbo Wang, Letao Liu, and Justin Dauwels. Slot-vae: Object-centric scene generation with slot attention, 2024. URL https://arxiv.org/abs/2306.06997.

Yixin Wang, David M Blei, and John P Cunningham. Posterior collapse and latent variable non-identifiability. *Advances in Neural Information Processing Systems*, 34:5443–5455, 2021.

Yixin Wang, David M. Blei, and John P. Cunningham. Posterior collapse and latent variable non-identifiability, 2023. URL https://arxiv.org/abs/2301.00537.

Nicholas Watters, Loïc Matthey, Christopher P. Burgess, and Alexander Lerchner. Spatial broadcast decoder: A simple architecture for learning disentangled representations in VAEs. *CoRR*, abs/1901.07017, 2019. URL http://arxiv.org/abs/1901.07017.

Antonia Wüst, Wolfgang Stammer, Hikaru Shindo, Lukas Helff, Devendra Singh Dhami, and Kristian Kersting. Synthesizing visual concepts as vision-language programs, 2025. URL https://arxiv.org/abs/2511.18964.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off, 2022. URL https://arxiv.org/abs/2209.09056.

Jure Zbontar, Li Jing, Ishan Misra, Yann LeCun, and Stéphane Deny. Barlow twins: Self-supervised learning via redundancy reduction. In *International Conference on Machine Learning*, pages 12310–12320. PMLR, 2021.