MITIGATING REASONING SHORTCUTS IN ABDUCTIVE LEARNING WITH SEMANTIC-LAYER INITIALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Abductive Learning (ABL) is a promising framework for integrating machine learning with logical reasoning. However, its generalization can be undermined by reasoning shortcuts, where seemingly strong reasoning performance masks unintended meanings of intermediate concepts. We trace this vulnerability to the learning–reasoning interface in ABL: associating input patterns with reasoning concepts is entangled with representation learning, leaving semantic alignment to random initialization and training dynamics. Meanwhile, the prevailing weak-binding practice treats inputs independently during reasoning, weakening cross-example constraints on what a predicted label means. To address these problems, we reformulate ABL with an explicit semantic layer that decouples semantic alignment from representation learning while enabling strong binding across related inputs. Under idealized perfect strong binding, we formalize shortcuts as concept transformations preserving reasoning outcomes and distinguish benign permutations from malignant non-injective collapses, yielding an identifiability limit up to benign equivalence. Guided by this analysis, we propose BASIL (Binding-Alignment for Semantic Initialization via Logic-consistency), a two-stage initialization. Stage I clusters inputs for strong binding; Stage II performs training-free semantic alignment by maximizing logic consistency. Experiments on visual and symbolic tasks show that BASIL substantially improves ABL's robustness to reasoning shortcuts.

1 INTRODUCTION

The principled integration of data-driven machine learning with knowledge-driven logical reasoning is a central challenge in artificial intelligence (Marra et al., 2024). Abductive Learning (ABL) (Zhou, 2019; Zhou & Huang, 2022) offers a promising framework for this integration. In ABL, a machine learning model predicts discrete pseudo-labels from sub-symbolic inputs, and a knowledge base (KB) reasons over the corresponding pseudo-groundings to produce the final reasoning result. When this result disagrees with supervision, abductive reasoning revises the pseudo-groundings and maps them back to pseudo-label supervision. This mechanism enables ABL to learn intermediate concepts without concept labels. However, reasoning-result supervision alone cannot, in general, uniquely identify their semantics. This ambiguity allows ABL to learn concept predictions that deviate from the underlying semantics while retaining spuriously high reasoning accuracy, giving rise to *reasoning shortcuts* (RSs) that undermine generalization (Marconato et al., 2023b; Yang et al., 2024).

To mitigate RSs, existing approaches intervene at two points in the ABL pipeline. Some strengthen the representations used to initialize ABL through self-supervised pretraining (Dai et al., 2019; Cai et al., 2021). Others refine how abductive solutions guide learning, using collective consistency (Huang et al., 2021), ambiguity-aware objectives (He et al., 2024), or KB curricula (Hu et al., 2025). In the broader Neuro-Symbolic literature, RS mitigation also includes partial concept supervision (Andolfi & Giunchiglia, 2025), entropy-based regularization (Marconato et al., 2023b), and reconstruction-based disentanglement (Misino et al., 2022). Even with these interventions, RSs can persist when initial representations are weak (Jia et al., 2025) or when reasoning results provide highly ambiguous supervision for the intermediate concepts (Tao et al., 2024). See Appendix A for further discussion.

We trace the persistence of RSs in standard ABL to two implicit modeling choices at the learning–reasoning interface. First, ABL relies on predictor training to align input patterns with logical concepts. Pseudo-labels are tied to concept predicates by a predefined mapping, so semantic alignment proceeds through the same updates that shape representations. When abductive solutions are ambiguous, this

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.

coupling can sustain incorrect concept meanings even when representations separate the concepts. Second, the prevailing *weak-binding* practice assigns each input occurrence a separate logical constant. Constraints from different examples can thus be satisfied independently without determining what a shared pseudo-label means, leaving the semantic signal highly ambiguous.

Motivated by this diagnosis, we reformulate ABL with an explicit *semantic layer* between learning and reasoning. The layer consists of an *abstraction*, which maps inputs to logical constants, and an *interpretation*, which assigns concept predicates to pseudo-labels. The abstraction enables *strong binding* by assigning a shared constant to related inputs, so evidence from different examples jointly constrains their meaning. To distinguish harmful shortcuts from superficial ones, we analyze RSs under an idealized *perfect strong binding* setting that decouples representation errors from logical ambiguity. In this setting, RSs can be formalized as transformations over intermediate concepts that preserve reasoning outcomes. Building on this view, we introduce a taxonomy that distinguishes *benign* shortcuts (permutations of concept names) from *malignant* shortcuts (non-injective collapses that merge distinct concepts). Furthermore, we establish an identifiability limit: under reasoning-result supervision, the intermediate semantics can be determined only up to benign permutations.

Guided by this analysis, we propose BASIL (**B**inding-**A**lignment for **S**emantic **I**nitialization via **L**ogic-consistency), a two-stage initialization method for ABL. Stage I learns a cluster predictor and assigns one shared logical constant to each cluster, allowing constraints from different examples to jointly constrain the same cluster meaning. Stage II keeps the predictor fixed and performs training-free semantic alignment by maximizing logic consistency with the KB. The aligned predictor then initializes standard ABL training, where abductive feedback guides further updates.

Our contributions are summarized as follows:

- **A principled formulation of ABL with an explicit semantic layer.** We expose abstraction and interpretation as distinct choices in converting pseudo-labels into pseudo-groundings, enabling their separate analysis and initialization (Section 2).
- **A formal taxonomy of reasoning shortcuts.** We distinguish benign permutations from malignant collapses and establish an identifiability limit for reasoning-result supervision (Section 3).
- **A practical two-stage initialization for ABL.** We develop BASIL to establish cluster-level strong binding and perform training-free semantic alignment before standard ABL training (Section 4).

Section 5 evaluates BASIL across four visual tasks and a scalable symbolic task, demonstrating superior performance in mitigating shortcut-induced training collapse.

2 Revisiting Abductive Learning

We revisit Abductive Learning (ABL) through an explicit semantic layer between machine learning and logical reasoning. This view makes clear how neural predictions acquire logical meaning and where semantic ambiguity enters the learning process.

Pipeline. Let $\mathcal{X}$ be the input space, $\mathcal{P}$ a set of k unary concept predicates, $\mathcal{C}$ a set of logical constants, and $\mathcal{G} \coloneqq \{p(c) \mid p \in \mathcal{P},\, c \in \mathcal{C}\}$ the set of *pseudo-groundings*. Given an input sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$, a predictor $f_\theta : \mathcal{X} \to \mathbb{R}^k$ assigns each input x_t a *pseudo-label*

$$z_\theta(x_t) \coloneqq \arg\max_{z \in [k]} f_\theta(x_t)_z, \qquad [k] \coloneqq \{1, \ldots, k\}.$$

These pseudo-labels are converted into a sequence of pseudo-groundings, which the knowledge base KB uses to deduce a reasoning result. We denote this reasoning process by $\mathrm{R}_{\mathrm{KB}} : \mathcal{G}^{<\infty} \to \mathcal{Y}$.

Semantic Layer. The semantic layer specifies how pseudo-labels are converted into pseudo-groundings. It consists of two mappings. The *interpretation* $\iota : [k] \to \mathcal{P}$ assigns a concept predicate to each pseudo-label, while the *abstraction* $\alpha : \mathcal{X} \to \mathcal{C}$ assigns a logical constant to each input. Together they map x_t to the pseudo-grounding

$$g^{\iota,\alpha}(x_t) \coloneqq \iota(z_\theta(x_t))(\alpha(x_t)). \tag{1}$$

For an input $\boldsymbol{x} = (x_1, \ldots, x_T)$, the semantic layer induces $\boldsymbol{g}^{\iota,\alpha}(\boldsymbol{x}) \coloneqq \big(g^{\iota,\alpha}(x_t)\big)_{t=1}^{T}$. We denote this layer as (α, ι). The abstraction groups inputs by logical constant, whereas the interpretation determines what each pseudo-label means. We call the selection of ι *semantic alignment*.

Learning. ABL is trained on a dataset $D = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ with supervision only on the reasoning results $y^{(i)}$. For an example $(\boldsymbol{x}^{(i)}, y^{(i)})$, let $\boldsymbol{g} = \boldsymbol{g}^{\iota,\alpha}(\boldsymbol{x}^{(i)})$ be the current pseudo-grounding sequence. If it does not produce the supervised result, ABL revises it by solving

$$\boldsymbol{g}' = \arg\min_{\widetilde{\boldsymbol{g}}} \mathcal{L}(\widetilde{\boldsymbol{g}}, \boldsymbol{g}) \quad \text{s.t.} \quad \mathrm{R}_{\mathrm{KB}}(\widetilde{\boldsymbol{g}}) = y^{(i)}, \tag{2}$$

where $\mathcal{L}$ measures the inconsistency between the current pseudo-groundings and a candidate revision. The revised pseudo-groundings $\boldsymbol{g}'$ are then mapped back to pseudo-label targets for updating f_θ. When the interpretation ι is bijective, this step reverses the predicate assignment in Equation (1): a corrected grounding $p(c)$ yields the pseudo-label target $\iota^{-1}(p)$ for the corresponding input.

3 Taxonomy of Reasoning Shortcuts

We analyze reasoning shortcuts under an idealized setting that removes perceptual errors. This lets us isolate the ambiguity introduced by the knowledge base and reasoning supervision.

Definition 1 (Strong Binding)**.** Given an abstraction $\alpha : \mathcal{X} \to \mathcal{C}$, a set $\mathcal{X}_S \subseteq \mathcal{X}$ is *strongly bound* if there exists $c \in \mathcal{C}$ such that $\alpha(x) = c$ for every $x \in \mathcal{X}_S$.

Perfect strong binding strengthens Definition 1 by requiring the groups induced by the abstraction to coincide exactly with the ground-truth concepts.

Assumption 1 (Perfect Strong Binding)**.** Let $s^\star : \mathcal{X} \to \mathcal{P}$ be the ground-truth concept mapping. There exists an abstraction $\alpha^\star : \mathcal{X} \to \mathcal{C}$ such that, for all $x, x' \in \mathcal{X}$,

$$\alpha^\star(x) = \alpha^\star(x') \quad \Longleftrightarrow \quad s^\star(x) = s^\star(x').$$

Perfect strong binding removes perceptual grouping errors and leaves only the ambiguity induced by the KB and reasoning supervision. Under this assumption, an input sequence $\boldsymbol{x} = (x_1, \ldots, x_T)$ induces the ground-truth pseudo-grounding sequence $\boldsymbol{g}^\star(\boldsymbol{x}) \coloneqq \big(s^\star(x_t)(\alpha^\star(x_t))\big)_{t=1}^{T}$.

Definition 2 (Reasoning Shortcut)**.** Let $\mathcal{D}_\mathcal{X}$ denote the distribution of input sequences, and let $\pi : \mathcal{P} \to \mathcal{P}$ be a mapping on concept predicates. It acts on a pseudo-grounding as $\pi(p(c)) \coloneqq (\pi(p))(c)$ and element-wise on pseudo-grounding sequences. We call π a *reasoning shortcut* if it preserves the reasoning result on the support of $\mathcal{D}_\mathcal{X}$:

$$\mathrm{R}_{\mathrm{KB}}(\boldsymbol{g}^\star(\boldsymbol{x})) = \mathrm{R}_{\mathrm{KB}}(\pi(\boldsymbol{g}^\star(\boldsymbol{x}))), \qquad \forall \boldsymbol{x} \in \mathrm{supp}(\mathcal{D}_\mathcal{X}).$$

Definition 3 (Benign and Malignant Shortcuts)**.** A reasoning shortcut is *benign* if π is bijective and *malignant* otherwise. In particular, the identity mapping that preserves all concept names is benign.

A benign shortcut preserves reasoning results under the fixed KB by relabeling concept names without losing their distinctions. A malignant shortcut preserves those results while merging at least two concepts, since every non-bijective map on the finite set $\mathcal{P}$ is non-injective. Given an interpretation $\iota : [k] \to \mathcal{P}$ and a benign shortcut π, we call ι and $\pi \circ \iota$ *benign-equivalent*. The following theorem shows that, under mild symmetry conditions such as permutation invariance of the inconsistency measure $\mathcal{L}$ in Equation (2), benign-equivalent interpretations remain indistinguishable throughout ABL training. Appendix B formalizes the full conditions and provides a detailed proof.

Theorem 1 (Benign-equivalence invariance of ABL)**.** *Fix an abstraction α. Let $\iota : [k] \to \mathcal{P}$ be bijective, and let π be a benign shortcut. Consider two ABL runs with this abstraction, one under ι and the other under $\pi \circ \iota$, initialized from the same predictor and optimizer state. Under the mild symmetry conditions stated in Appendix B, the two runs induce the same distribution over parameter trajectories. Consequently, ABL using only reasoning-result supervision cannot distinguish the two interpretations or select a unique representative of their benign-equivalence class.*

Under the symmetry conditions of Theorem 1, relabeling the interpretation also relabels the corresponding abductive corrections. When these corrections are mapped back using the inverse of the relabeled interpretation, the relabeling cancels, giving the same pseudo-label targets and hence the same learning signal. Benign ambiguity is therefore a limitation of reasoning-result supervision rather than a learning failure. Recovering the intended concept names requires additional information that distinguishes these interpretations. We accordingly treat recovery of the semantics of intermediate concepts up to a benign shortcut as valid, while malignant shortcuts remain failures.

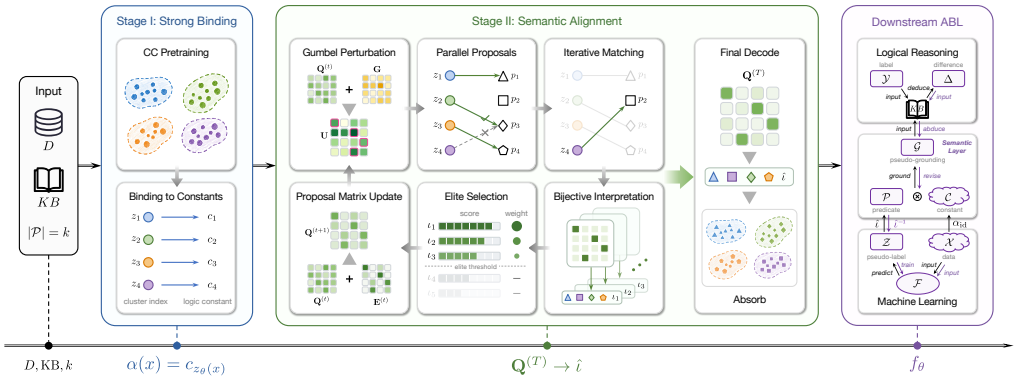

Figure 1: BASIL→ABL training pipeline. **Stage I** trains the cluster predictor f_θ and constructs the abstraction α. **Stage II** keeps both fixed, selects an interpretation $\hat{\iota}$ by maximizing logic-consistency, and applies the decoded interpretation to the predictor's final linear classification layer. Downstream ABL training then proceeds from this semantically aligned predictor.

4 Our Approach

We propose BASIL (Binding-Alignment for Semantic Initialization via Logic-consistency) to initialize the semantic layer before standard ABL training (Figure 1). In the ABL setting of Section 2, our goal is to learn (i) a k-way pseudo-label predictor f_θ with well-structured representations and (ii) a bijective interpretation $\iota : [k] \to \mathcal{P}$ such that, aside from prediction errors, interpreting the predicted pseudo-labels through ι recovers the ground-truth concepts up to a benign shortcut.

BASIL performs this initialization in two stages:

- **Stage I (Strong Binding in Practice)** pretrains a k-way cluster predictor f_θ and constructs α by assigning one logical constant to each cluster.
- **Stage II (Training-Free Semantic Alignment)** fixes (f_θ, α) and searches for a high-scoring bijection $\hat{\iota}$ under logic-consistency with KB on the training data D.

Finally, BASIL applies $\hat{\iota}$ by permuting the predictor's final linear classification layer, after which the aligned predictor resumes training in the standard ABL loop.

4.1 Stage I: Strong Binding in Practice

Construction. We use Contrastive Clustering (CC) (Li et al., 2021) to pretrain a k-way cluster predictor f_θ, with predicted cluster $z_\theta(x) = \arg\max_{z \in [k]} f_\theta(x)_z$. More details about CC are provided in Appendix C.1. After pretraining, we freeze f_θ, assign a distinct logical constant c_z to each cluster z, and set $\alpha(x) = c_{z_\theta(x)}$. This realizes strong binding at the cluster level, so all occurrences assigned to one cluster contribute to the same predicate assignment in Stage II.

Why strong binding. Under the standard abstraction $\alpha(x) = x$, each input occurrence receives a separate logical constant. Abduction can therefore revise the predicate assigned to one occurrence without imposing the same revision on another, even when both initially share a pseudo-label. Strong binding instead assigns one constant to all inputs in a cluster, so constraints from different examples act on the same concept assignment. For example, suppose f_θ assigns pseudo-labels z_0, z_1, z_2 to images of digits $0, 1, 2$ in the equations $0 + 1 = 1$, $0 + 2 = 2$, and $1 + 2 = 3$, while their predicate meanings are unknown. With separate constants, each equation can be satisfied by independently revising its input pseudo-labels. With strong binding, let v_j denote the digit assigned to constant c_{z_j}. The shared assignments must satisfy $v_0 + v_1 = 1$, $v_0 + v_2 = 2$, and $v_1 + v_2 = 3$, uniquely giving $(v_0, v_1, v_2) = (0, 1, 2)$. Thus, the three equations jointly identify the cluster meanings.

4.2 Stage II: Training-free semantic alignment

Stage II aligns the Stage I clusters with concept predicates using the iterative procedure summarized in Figure 1. Starting from a proposal matrix over cluster-to-predicate assignments, BASIL repeatedly samples candidate bijections, scores them by logic-consistency, retains a top-scoring subset, and updates the proposal matrix using the score-weighted frequency of each cluster-to-predicate pair in that subset. Repeating this cycle reinforces assignment pairs that consistently appear in high-scoring candidates, after which a final matching step returns one bijective interpretation.

Objective. Stage II searches over the bijective interpretations $\mathcal{I}_{\text{bij}} \coloneqq \{\iota : [k] \to \mathcal{P} \mid \iota \text{ is bijective}\}$. For each candidate ι, the fixed Stage I abstraction α induces the pseudo-grounding sequence defined in Equation (1), and we evaluate ι using the following empirical *logic-consistency score*:

$$\widehat{s}_D(\iota) \coloneqq \frac{\sum_{i=1}^{N} T_i \, \mathbb{I}\big[\mathrm{R}_{\mathrm{KB}}\big(\boldsymbol{g}^{\iota,\alpha}(\boldsymbol{x}^{(i)})\big) = y^{(i)}\big]}{\sum_{i=1}^{N} T_i}, \tag{3}$$

where T_i is the length of input $\boldsymbol{x}^{(i)}$. The score is the length-weighted fraction of training examples whose reasoning result matches the supervision, and Stage II seeks a bijection that maximizes it.

Interpretation sampling. Stage II maintains a row-stochastic proposal matrix $\mathbf{Q}^{(t)} \in [0,1]^{k\times k}$, where $\mathbf{Q}^{(t)}_{z,p}$ is the current proposal weight for assigning cluster z to predicate p. At each round, BASIL combines these weights with independent Gumbel noise (Jang et al., 2017):

$$U_{z,p} = \frac{\log \mathbf{Q}^{(t)}_{z,p}}{\tau_g} + G_{z,p}, \qquad G_{z,p} \sim \mathrm{Gumbel}(0,1), \tag{4}$$

where τ_g controls the balance between the proposal matrix and the Gumbel perturbations. To construct a bijection, the procedure begins with all clusters and predicates unmatched. At each step, every unmatched cluster proposes to the remaining predicate with the highest utility, and each predicate accepts the highest-utility proposal it receives. The accepted cluster–predicate pairs are then removed from further consideration, and the process repeats until a complete matching is formed.

Elite update. After sampling m bijections, BASIL scores them using Equation (3) and retains the top $\lceil \rho m \rceil$. It then aggregates the cluster-to-predicate pairs across these elite interpretations, giving higher-scoring elites more weight. Specifically, $\mathbf{E}^{(t)}_{z,p}$ is the normalized total weight of elites that map cluster z to predicate p, allowing the same pair to be reinforced by multiple interpretations. Appendix C.2 gives the exact weighting scheme. BASIL uses $\mathbf{E}^{(t)}$ to update the proposal matrix with a cross-entropy-method (Rubinstein, 1997) step, followed by uniform mixing to preserve exploration:

$$\bar{\mathbf{Q}}^{(t+1)} = (1-\lambda)\mathbf{Q}^{(t)} + \lambda \mathbf{E}^{(t)}, \qquad \mathbf{Q}^{(t+1)} = (1-\eta)\bar{\mathbf{Q}}^{(t+1)} + \eta \mathbf{U}^{\text{unif}}, \qquad \mathbf{U}^{\text{unif}}_{z,p} = 1/k. \tag{5}$$

Here, λ controls the update rate and η maintains exploration. Repeating this cycle concentrates later proposals on pairs that consistently appear in high-scoring interpretations.

Final decode. After T rounds, $\mathbf{Q}^{(T)}$ summarizes the accumulated evidence for cluster-to-predicate pairs, from which BASIL decodes one bijection by maximum-weight matching. For $\iota \in \mathcal{I}_{\text{bij}}$, let $A_\iota \in \{0,1\}^{k\times k}$ be its permutation matrix, with $(A_\iota)_{z,p} = \mathbb{I}[\iota(z) = p]$. We select one entry from each row and column to maximize the total assignment weight:

$$\hat{\iota} = \mathrm{Dec}(\mathbf{Q}^{(T)}) \coloneqq \arg\max_{\iota \in \mathcal{I}_{\text{bij}}} \sum_{z=1}^{k} \mathbf{Q}^{(T)}_{z,\iota(z)} = \arg\max_{\iota \in \mathcal{I}_{\text{bij}}} \langle \mathbf{Q}^{(T)}, A_\iota \rangle. \tag{6}$$

BASIL absorbs $\hat{\iota}$ into the predictor by permuting its final linear classification layer, after which standard ABL starts from the aligned predictor. The complete algorithm is provided in Appendix C.2.

4.3 Theory: Logic consistency and final decoding

Score separation under cluster errors. Let $\iota^\star : [k] \to \mathcal{P}$ be a bijective cluster-to-concept alignment minimizing ϵ_b, the length-weighted probability that a sequence contains at least one incorrectly aligned input. Define the *benign interpretation set* $\mathcal{A} \coloneqq \{\pi \circ \iota^\star \mid \pi \text{ is a benign shortcut}\}$. For $\iota \in \mathcal{I}_{\text{bij}}$, let $s(\iota)$ and $\widetilde{s}(\iota)$ denote the length-weighted population logic-consistency scores under perfect strong binding and the Stage I clusters, respectively. When $\mathcal{A} \subsetneq \mathcal{I}_{\text{bij}}$, define the ideal gap

$$\Delta_{\mathcal{A}} \coloneqq \min_{\iota^\circ \in \mathcal{A},\ \iota \in \mathcal{I}_{\text{bij}} \setminus \mathcal{A}} \big(s(\iota^\circ) - s(\iota)\big).$$

Theorem 2 (Score separation under cluster errors)**.** *Assume $\mathcal{A} \subsetneq \mathcal{I}_{\text{bij}}$, noise-free reasoning-result supervision, bounded sequence lengths, distinct cluster constants, and KB reasoning invariant to renaming logical constants. For every $\iota^\circ \in \mathcal{A}$ and $\iota \in \mathcal{I}_{\text{bij}} \setminus \mathcal{A}$,*

$$\widetilde{s}(\iota^\circ) - \widetilde{s}(\iota) \geq \Delta_{\mathcal{A}} - 2\epsilon_b.$$

Thus, if $\Delta_{\mathcal{A}} > 2\epsilon_b$, every maximizer of $\widetilde{s}$ belongs to $\mathcal{A}$.

The condition connects the task's semantic information to the cluster quality needed for alignment. Appendix D.1 provides the proof, a finite-sample counterpart, and a fixed-dataset certificate.

Elite evidence and final decoding. Theorem 2 concerns logic-consistency maximizers. We now connect the elite updates to the final matching in Equation (6). Assume $\mathcal{A} \subsetneq \mathcal{I}_{\text{bij}}$ and fix a benign interpretation $\iota^\dagger \in \mathcal{A}$. Its *elite margin* at round t is

$$\gamma_t(\iota^\dagger) := \langle \mathbf{E}^{(t)}, A_{\iota^\dagger} \rangle - \max_{\iota \in \mathcal{I}_{\text{bij}} \setminus \mathcal{A}} \langle \mathbf{E}^{(t)}, A_\iota \rangle,$$

the difference between its matching score and the best score outside $\mathcal{A}$.

Theorem 3 (Benign final decoding)**.** *Let* $\mathbf{Q}^{(0)} = \mathbf{U}^{\text{unif}}$, $\lambda \in (0, 1]$, $\eta \in [0, 1)$, *and* $a = (1-\eta)(1-\lambda)$. *Under the update in Equation* (5)*,*

$$\sum_{t=0}^{T-1} a^{T-1-t} \gamma_t(\iota^\dagger) > 0 \quad \Longrightarrow \quad \text{Dec}(\mathbf{Q}^{(T)}) \in \mathcal{A}.$$

Theorem 3 makes the effect of the elite updates explicit. In Equation (5), each elite matrix $\mathbf{E}^{(t)}$ contributes to $\mathbf{Q}^{(T)}$, with its contribution discounted by a for each subsequent round. The final decode belongs to $\mathcal{A}$ whenever the discounted margins $\gamma_t(\iota^\dagger)$ sum to a positive value for one fixed benign interpretation $\iota^\dagger$, even if some rounds have negative margins. Appendix D.2 provides the proof, a quantitative margin bound, and post-hoc certificate evaluations.

5 Experiments

We evaluate BASIL in visual and symbolic ABL settings through end-to-end comparisons, controlled ablations, and scaling analyses.[1] The experiments are organized around three questions:

Q1. *Does BASIL reliably mitigate reasoning shortcuts in ABL?*

Q2. *Does BASIL require each component for reliable semantic alignment?*

Q3. *Does BASIL remain reliable as concept count increases and cluster purity decreases?*

Metrics. Our primary metric tests whether every concept has the correct majority prediction up to a benign shortcut. Let $R \in \mathbb{N}^{k \times k}$ be the concept confusion matrix on the test set, where $R_{p,q}$ counts inputs with ground-truth concept p and predicted concept q. Let Π_{ben} denote the benign shortcuts for the task, with $\Pi_{\text{ben}} = \{\text{id}\}$ when the target semantics is unique. For $\pi \in \Pi_{\text{ben}}$, let $R^{(\pi)}$ be the confusion matrix after relabeling the predicted concepts by π. We define

$$\text{Success}(R) = \mathbf{1}\left[\exists\, \pi \in \Pi_{\text{ben}} \text{ s.t. } R^{(\pi)}_{p,p} > \max_{q \neq p} R^{(\pi)}_{p,q},\ \forall p \in \mathcal{P}\right]. \tag{7}$$

We also report reasoning accuracy (RA), concept accuracy (CA), and shortcut severity (SSV). RA is the fraction of test examples for which the reasoning result obtained by the KB from the predicted concepts agrees with the observed outcome. CA is the corresponding concept accuracy. On tasks with non-identity benign shortcuts, it is computed after the best benign relabeling. We define SSV as the strongest off-diagonal confusion after minimizing over benign relabelings:

$$\text{SSV}(R) = \min_{\pi \in \Pi_{\text{ben}}} \max_{\substack{p,q \in \mathcal{P} \\ p \neq q}} \frac{R^{(\pi)}_{p,q}}{\sum_{r \in \mathcal{P}} R^{(\pi)}_{p,r}}. \tag{8}$$

Thus, SSV measures the largest fraction of any true concept assigned to a single wrong concept.

Competitors. We compare BASIL with standard ABL (Zhou, 2019), A^{3}BL (He et al., 2024), and DeepStochLog (DSL) (Winters et al., 2022). We include L+PT as a strong representation-level RS-mitigation baseline (Marconato et al., 2023b). It initializes learner L with the Stage I pretrained model and then follows the standard training procedure of L. L+BASIL-init starts from the same Stage I checkpoint and additionally applies Stage II before the standard training of L. BASIL denotes ABL+BASIL-init, while the A^{3}BL and DSL variants test whether BASIL-init transfers to other learners. L+SA applies Stage II without Stage I and is included only in the stage ablation.

[1] Code is available at https://anonymous.4open.science/r/BASIL.

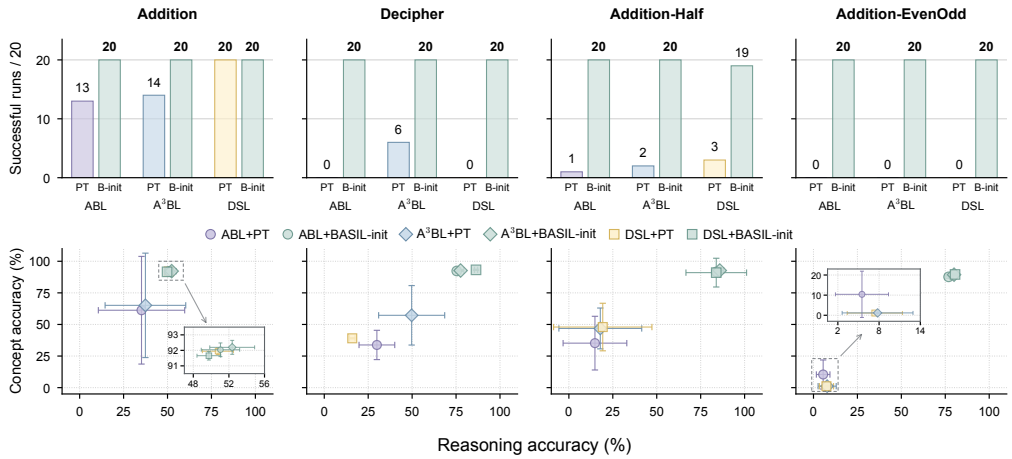

Figure 2: Reasoning-shortcut mitigation for ABL and transfer of BASIL-init to A^{3}BL and DSL. Top: successful runs out of 20. Bottom: mean RA versus CA with standard-deviation bars. Dashed boxes enlarge overlapping points. B-init denotes BASIL-init.

Tasks. We evaluate four visual tasks rendered with CIFAR-10 images and one symbolic task, with full protocols in Appendix E. *Addition* provides only the sum of two n-digit image sequences as supervision (Manhaeve et al., 2018). We report the most challenging setting, $n = 4$, here and $n \in \{1, 2, 3\}$ in Appendix F. *Decimal Equation Decipher* (Decipher) presents complete addition equations, with supervision indicating only whether each equation satisfies the addition rules (Huang et al., 2021). *Addition-Half* and *Addition-EvenOdd* are shortcut tasks from RSBench (Bortolotti et al., 2024), evaluated on out-of-distribution (OOD) operand pairs absent from training. Addition-EvenOdd's OOD RA is computed after the best post-hoc benign relabeling. We introduce *Large-K Relational Decipher* (Large-K), a scalable symbolic task with $k \in \{20, 30, 50\}$ concepts and k benign shortcuts. To control cluster purity, each relation endpoint remains in its correct cluster with probability $p = 0.7$ and is otherwise assigned to a uniformly sampled incorrect cluster.

5.1 BASIL Reliably Mitigates Reasoning Shortcuts in ABL

Overall Results. Figure 2 answers the effectiveness question of Q1. BASIL succeeds in all 80 ABL runs across the four visual tasks. More importantly, the +PT baselines, initialized from the same Stage I checkpoints, can retain substantial RA while their CA remains low. This gap shows why high reasoning accuracy alone cannot establish that the model has learned the intended concept semantics. BASIL improves successful-run counts and CA together. The gain therefore comes from aligning the predictor outputs with the logical predicates, rather than from stronger pretraining or from fitting the final outcomes more closely. BASIL-init also improves A^{3}BL and DSL over their +PT variants without changing their training procedures. These transfer results indicate that BASIL-init acts at the semantic layer before downstream learning, rather than relying on a learner-specific update rule.

Task Diversity. The tasks differ in supervision strength and shortcut structure. Addition is relatively simple and provides denser outcome information, yet ABL+PT and A^{3}BL+PT still fail on some seeds. Decipher provides the opposite stress test: its validity label is much weaker, and all +PT baselines struggle, whereas BASIL-init makes all three learners consistently successful. Under perfect strong binding, Decipher and Addition-Half each have a single benign interpretation, while Addition-EvenOdd has two. Addition-Half and Addition-EvenOdd also admit malignant shortcuts. These results demonstrate BASIL's robustness to reasoning shortcuts across diverse tasks.

5.2 Both Stages Are Necessary and Stage II Design Matters

Stage Ablation. To answer Q2, Table 1 tests whether both stages are necessary for reliable semantic alignment. Stage I alone can help on the comparatively simple Addition task, but it leaves cluster semantics unresolved and remains unreliable on the other three tasks, where supervision information is much weaker. Stage II alone faces the complementary problem: it lacks stable cluster distinctions to align. The full method succeeds consistently across all four tasks, supporting the two-stage design in which semantic alignment builds on the cluster structure learned during pretraining.

Table 1: Stage ablation study. PT and SA denote Stage I pretraining and Stage II semantic alignment. **Succ.** reports successful runs out of 20. CA and SSV are mean ± standard deviation over 20 seeds.

Initialization		Addition			Decipher			Addition-Half			Addition-EvenOdd		
PT	SA	Succ. ↑	CA ↑	SSV ↓	Succ. ↑	CA ↑	SSV ↓	Succ. ↑	CA ↑	SSV ↓	Succ. ↑	CA ↑	SSV ↓
		0	12.8±4.5	63.4±10.0	0	28.4±8.6	78.4±12.5	0	40.3±12.2	72.3±17.5	0	11.0±4.8	90.4±7.0
✓		13	61.2±42.6	38.6±38.0	0	33.8±11.6	95.6±2.7	1	35.2±21.2	92.2±19.5	0	10.4±11.5	98.5±1.3
	✓	5	32.3±26.3	49.2±19.3	0	10.4±6.7	96.0±12.2	0	20.0±12.7	84.6±13.3	0	5.5±3.1	96.2±3.7
✓	✓	**20**	**92.0±0.4**	**10.8±1.5**	**20**	**92.3±0.2**	**11.1±0.9**	**20**	**91.6±0.8**	**8.6±1.6**	**20**	**87.7±1.2**	**15.5±5.6**

Table 2: Comparison of Stage II alignment search methods. Orbit is the normalized Hamming distance (%) from the Stage II decoded interpretation to the nearest benign relabeling of the Stage I majority interpretation. Orbit and RA are reported as mean ± standard deviation over 20 seeds.

Method	Addition			Decipher			Addition-Half			Addition-EvenOdd		
	Succ. ↑	Orbit ↓	RA ↑	Succ. ↑	Orbit ↓	RA ↑	Succ. ↑	Orbit ↓	RA ↑	Succ. ↑	Orbit ↓	RA ↑
Best-of-M	**20**	31.0±9.4	51.4±2.2	14	37.5±13.7	64.5±16.9	**20**	**3.0±7.1**	**84.0±1.3**	4	35.5±17.7	42.7±24.2
Hungarian	**20**	1.0±4.4	**52.0±1.8**	19	12.5±14.4	73.4±9.1	6	33.0±23.0	32.0±34.3	2	52.5±15.5	29.0±20.6
BASIL (w/o bij.)	**20**	**0.0±0.0**	51.8±2.0	1	85.5±19.6	51.2±5.4	0	57.0±7.1	10.1±0.7	0	87.5±21.7	10.5±6.7
BASIL	**20**	**0.0±0.0**	51.8±2.0	**20**	**0.0±0.0**	**75.5±0.8**	**20**	**3.0±7.1**	**84.0±1.3**	**20**	**0.0±0.0**	**77.4±1.4**

Stage II Variants. Table 2 compares BASIL's Stage II alignment search method with three variants, using the same Stage I checkpoints and a total budget of $M = 50{,}000$ sampled candidates. Best-of-M selects the highest-scoring interpretation from M uniformly sampled bijections. Hungarian aggregates evidence from uniform candidates into a weighted assignment matrix and decodes a single interpretation via maximum-weight matching. BASIL (w/o bij.) retains BASIL's iterative update strategy but assigns predicates independently to clusters, allowing predicate reuse.

Adaptivity. Best-of-M relies on a fixed candidate set, so it can select a benign interpretation only if one is sampled. Uniform sampling works well on Addition-Half, whose five concepts admit just $5! = 120$ bijections. As the number of concepts grows, the factorial expansion of the search space makes direct sampling impractical. Hungarian can construct an unsampled interpretation, but when several benign interpretations receive similar scores, its weighted matrix can mix their cluster-to-predicate assignments and decode outside the benign set. BASIL instead updates its proposal matrix over multiple rounds. Assignment pairs favored by high-scoring elites receive greater weights and are sampled more frequently in later rounds. This feedback can progressively favor one benign interpretation without requiring a complete benign candidate at the outset. We call the resulting concentration on one benign interpretation *mode locking*. Addition-EvenOdd tests this behavior with two benign interpretations. The next section extends this test to k benign interpretations on Large-K.

Bijectivity. Table 2 shows that bijectivity is useful when alignment is vulnerable to concept collapse. Addition admits no malignant shortcut, and BASIL (w/o bij.) performs similarly to BASIL. The variant is unreliable on Addition-Half and Addition-EvenOdd, which admit malignant shortcuts, and on Decipher, where cluster errors can make collapsed assignments score well (Appendix E.3). BASIL's bijective search rules out these collapses, preserving the distinctions learned in Stage I.

5.3 BASIL AT SCALE AND THE ROLE OF CLUSTER PURITY

Scalability. Large-K has k concepts, k benign shortcuts, and no malignant shortcuts (Appendix E.3). Figure 3 shows that BASIL maintains 20/20 successes as k increases from 20 to 50, whereas neither one-shot control succeeds. BASIL's CA remains near 70%, matching the prescribed endpoint purity ($p = 0.7$), which suggests that errors remain dominated by cluster corruption rather than semantic misalignment. The non-bijective variant also performs well, with only one failure at $k = 50$. This is consistent with the earlier findings on bijectivity, as Large-K admits no malignant shortcuts. Due to space constraints, we provide a detailed report of computational costs in Appendix F.3.

Cluster Purity. We examine the role of cluster purity in Figure 4 by varying Stage I pretraining epochs over 20, 50, 100, and 300 on Decipher, the most challenging visual task. The resulting mean cluster purities range from 51.8% to 67.7%, yet BASIL succeeds in all 80 runs (see Appendix F.4 for full results), showing that reliable recovery does not require highly pure clusters. Achieving higher RA and CA and lower SSV, however, still depends on purer clusters. This dependence points to Stage I cluster purity as the main observed limit on final performance: Stage II can reassign predicates to clusters, but it cannot separate concepts that Stage I has already grouped into the same cluster.

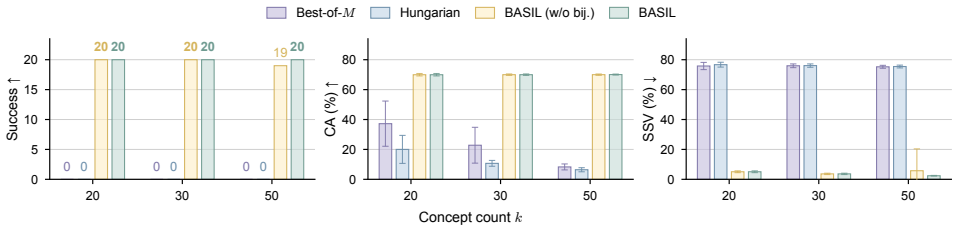

Figure 3: Scalability analysis on Large-K after downstream ABL training. Success reports successful runs out of 20, while CA and SSV show means with standard-deviation bars.

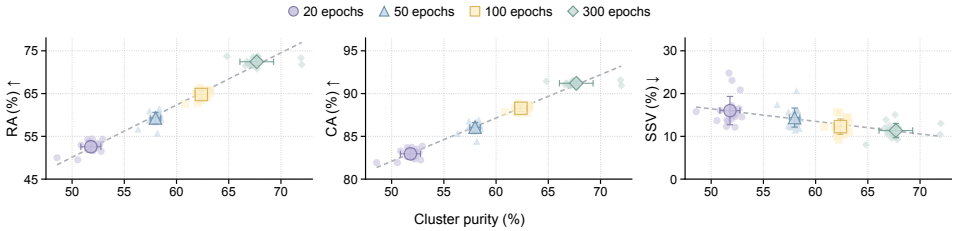

Figure 4: Analysis of Stage I cluster purity on Decipher. RA, CA, and SSV versus purity after 20, 50, 100, and 300 pretraining epochs (means with standard-deviation error bars over 20 runs).

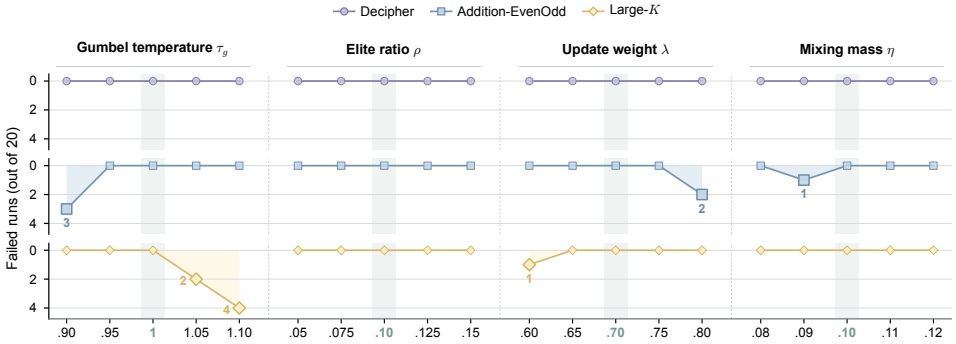

Figure 5: Hyperparameter sensitivity analysis. Each panel varies one parameter around its default and reports failed runs out of 20 after downstream ABL training. Shaded bands mark the default values.

Hyperparameter Sensitivity. Figure 5 shows that BASIL is robust to the hyperparameter variations, with 20/20 successes in most settings. The lowest success rate is 16/20 on Large-K ($k = 50$) at $\tau_g = 1.1$. Raising the temperature makes sampling more random, so assignments favored by earlier rounds are less likely to be chosen again. Increasing τ_g by just 0.1 from its default of 1.0 cuts the upper bound on the probability ratio between candidate assignments in a cluster's first proposal by about 43% (Appendix F.5). More of the fixed sampling budget is then spent trying alternative assignments, leaving fewer opportunities to reinforce promising ones and settle on one benign interpretation.

6 Conclusion

We introduced BASIL, a two-stage initialization that mitigates reasoning shortcuts in ABL without concept supervision. BASIL combines cluster-level strong binding with training-free bijective alignment, separating semantic alignment from representation learning. Our analysis distinguishes benign relabelings from malignant collapses and establishes an identifiability limit under reasoning-result supervision. Across four visual tasks, BASIL succeeds in all 80 ABL runs and transfers to A^{3}BL and DeepStochLog, while controlled symbolic experiments demonstrate BASIL's scalability. Matched pretraining comparisons establish the benefit of explicit alignment beyond representation quality alone. BASIL thus provides a reliable, transferable initialization that strengthens existing learners without changing their downstream training procedures. A limitation of BASIL is that unsupervised pretraining may group instances in ways that do not reflect the intended semantic structure. BASIL also assumes inputs are already separated into individual objects, limiting its direct use in multi-object settings. In future work, we will explore more flexible correspondences between learned clusters and logical concepts to reduce dependence on the initial clustering, and investigate integration with object detection or segmentation to support reasoning over multiple objects.

AI USE STATEMENT

We used generative AI tools to assist with proof writing, provide feedback on research methodology and experiments, and support qualitative and thematic analysis of experimental behavior, including discussion of the results. We did not use these tools to generate synthetic datasets, develop theoretical models or conceptual frameworks, formulate mathematical claims, supply the key arguments underlying the proofs, propose or refine research hypotheses, implement BASIL or the baseline learners, or clean or reformat the experimental datasets. Translation was not applicable to this work. We also used AI tools for manuscript polishing and the preparation of figures and tables. We reviewed AI-assisted material, checked proof arguments against their stated assumptions, and verified numerical results against experimental records. The authors take responsibility for the final content of this work, including AI-assisted text, claims, and figures.

REPRODUCIBILITY STATEMENT

Appendices B and D give the assumptions and full proofs of the theoretical results. Appendix C details both stages of BASIL, including pretraining hyperparameters, the alignment algorithm, and initialization of downstream learners. Appendix E specifies task construction, data splits, evaluation protocols, model architectures, training schedules, and hardware. Appendix F provides additional numerical results, run statistics, and settings for the controlled analyses.

REFERENCES

Luca Andolfi and Eleonora Giunchiglia. Right for the right reasons: Avoiding reasoning shortcuts via prototypical neurosymbolic AI. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 157309–157343, 2025.

Samuele Bortolotti, Emanuele Marconato, Tommaso Carraro, Paolo Morettin, Emile van Krieken, Antonio Vergari, Stefano Teso, and Andrea Passerini. A neuro-symbolic benchmark suite for concept quality and reasoning shortcuts. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, pp. 115861–115905, 2024.

Samuele Bortolotti, Emanuele Marconato, Paolo Morettin, Andrea Passerini, and Stefano Teso. Shortcuts and identifiability in concept-based models from a neuro-symbolic lens. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 173204–173243, 2025.

Le-Wen Cai, Wang-Zhou Dai, Yu-Xuan Huang, Yu-Feng Li, Stephen Muggleton, and Yuan Jiang. Abductive learning with ground knowledge base. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1815–1821, 2021.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 1597–1607, 2020.

Wang-Zhou Dai, Qiuling Xu, Yang Yu, and Zhi-Hua Zhou. Bridging machine learning and logical reasoning by abductive learning. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, pp. 2815–2826, 2019.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1735–1742, 2006.

Hao-Yuan He, Hui Sun, Zheng Xie, and Ming Li. Ambiguity-aware abductive learning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 18019–18042, 2024.

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)*, pp. 770–778, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Wen-Chao Hu, Qi-Jie Li, Lin-Han Jia, Cunjing Ge, Yu-Feng Li, Yuan Jiang, and Zhi-Hua Zhou. Curriculum abductive learning. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, pp. 174455–174483, 2025.

Yu-Xuan Huang, Wang-Zhou Dai, Le-Wen Cai, Stephen H. Muggleton, and Yuan Jiang. Fast abductive learning by similarity-based consistency optimization. In *Advances in Neural Information Processing Systems 34 (NeurIPS)*, pp. 26574–26584, 2021.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

Lin-Han Jia, Wen-Chao Hu, Jie-Jing Shao, Lan-Zhe Guo, and Yu-Feng Li. Verification learning: Make unsupervised neuro-symbolic system feasible. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, pp. 27292–27306, 2025.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, pp. 5338–5348, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Yunfan Li, Peng Hu, Zitao Liu, Dezhong Peng, Joey Tianyi Zhou, and Xi Peng. Contrastive clustering. In *Proceedings of the 35th AAAI Conference on Artificial Intelligence*, pp. 8547–8555, 2021.

Zenan Li, Zehua Liu, Yuan Yao, Jingwei Xu, Taolue Chen, Xiaoxing Ma, and Jian Lü. Learning with logical constraints but without shortcut satisfaction. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023.

Robin Manhaeve, Sebastijan Dumančić, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, pp. 3749–3759, 2018.

Emanuele Marconato, Stefano Teso, and Andrea Passerini. Neuro-Symbolic reasoning shortcuts: Mitigation strategies and their limitations. In *Proceedings of the 17th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy)*, pp. 162–166, 2023a.

Emanuele Marconato, Stefano Teso, Antonio Vergari, and Andrea Passerini. Not all neuro-symbolic concepts are created equal: Analysis and mitigation of reasoning shortcuts. In *Advances in Neural Information Processing Systems 36 (NeurIPS)*, pp. 72507–72539, 2023b.

Emanuele Marconato, Samuele Bortolotti, Emile van Krieken, Antonio Vergari, Andrea Passerini, and Stefano Teso. BEARS make neuro-symbolic models aware of their reasoning shortcuts. In *Proceedings of the 40th Conference on Uncertainty in Artificial Intelligence (UAI)*, pp. 2399–2433, 2024.

Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, 328:104062, 2024.

Eleonora Misino, Giuseppe Marra, and Emanuele Sansone. VAEL: Bridging variational autoencoders and probabilistic logic programming. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, pp. 4667–4679, 2022.

Reuven Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

Luciano Serafini and Artur d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. In *Proceedings of the 11th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy)*, 2016.

Matthew Stephens. Dealing with label switching in mixture models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 62(4):795–809, 2000.

Akihiro Takemura and Katsumi Inoue. Differentiable logic programming to mitigate reasoning shortcuts in neurosymbolic systems. In *Proceedings of the 42nd International Conference on Logic Programming (ICLP)*, pp. 29–51, 2026.

Akihiro Takemura, Katsumi Inoue, and Masaaki Nishino. Constraint-based analysis of reasoning shortcuts in neurosymbolic learning. In *Proceedings of the 23rd International Conference on Principles of Knowledge Representation and Reasoning (KR)*, pp. 1132–1142, 2026.

Lue Tao, Yu-Xuan Huang, Wang-Zhou Dai, and Yuan Jiang. Deciphering raw data in neuro-symbolic learning with provable guarantees. In *Proceedings of the 38th AAAI Conference on Artificial Intelligence*, pp. 15310–15318, 2024.

Thomas Winters, Giuseppe Marra, Robin Manhaeve, and Luc De Raedt. DeepStochLog: Neural stochastic logic programming. In *Proceedings of the 36th AAAI Conference on Artificial Intelligence*, pp. 10090–10100, 2022.

Xiao-Wen Yang, Wen-Da Wei, Jie-Jing Shao, Yu-Feng Li, and Zhi-Hua Zhou. Analysis for abductive learning and neural-symbolic reasoning shortcuts. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, pp. 56524–56541, 2024.

Zhun Yang, Adam Ishay, and Joohyung Lee. NeurASP: Embracing neural networks into answer set programming. In *Proceedings of the 29th International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1755–1762, 2020.

Zhi-Hua Zhou. Abductive learning: towards bridging machine learning and logical reasoning. *Science China Information Sciences*, 62(7):076101, 2019.

Zhi-Hua Zhou and Yu-Xuan Huang. Abductive learning. In Pascal Hitzler and Md. Kamruzzaman Sarker (eds.), *Neuro-Symbolic Artificial Intelligence: The State of the Art*, pp. 353–369. IOS Press, 2022.

A Related Work

Reasoning Shortcuts. Reasoning shortcuts (RS) arise when models satisfy logical constraints through unintended intermediate semantics. They remain a central obstacle in data–knowledge dual-driven learning. Marconato et al. (2023b) characterize RS as unintended optima of the learning objective, while Yang et al. (2024) analyze shortcut risk and generalization in ABL. Takemura et al. (2026) formalize shortcut detection as uniqueness of a CSP over concept mappings and analyze value-permutation symmetries through its automorphism group, focusing on pre-deployment shortcut verification. In differentiable logic programming, Takemura & Inoue (2026) show that one-to-one grounding of neural outputs to logical atoms substantially reduces reasoning shortcuts on MNIST-scale tasks. BASIL turns this structural constraint into a learner-agnostic initialization, adds symmetry-aware analysis, and evaluates concept counts up to $k = 50$. Bortolotti et al. (2025) derive identifiability conditions for concepts and the inference layer in concept-based models from the same shortcut perspective, and the benign-renaming ambiguity we formalize is the neuro-symbolic analogue of classical label switching in mixture models (Stephens, 2000).

Existing mitigation strategies differ mainly in where they intervene. At the learning-objective level, Marconato et al. (2023b) analyze overconfidence-oriented heuristics, Misino et al. (2022) use reconstruction-based disentanglement, He et al. (2024) weight ambiguous abductive candidates to provide soft supervision, and Li et al. (2023) introduce a logic-derived distributional loss with dual variables. Controlled studies examine reconstruction and concept supervision, while RSBench benchmarks reasoning models alongside partially concept-supervised and pretrained-representation baselines (Marconato et al., 2023a; Bortolotti et al., 2024). Other methods constrain ambiguity with sparse concept anchors (Andolfi & Giunchiglia, 2025) or a curriculum over the existing KB (Hu et al., 2025), while Marconato et al. (2024) make predictors shortcut-aware through calibrated concept uncertainty. BASIL makes the semantic layer explicit and performs a one-time, gradient-free Stage II alignment from final reasoning outcomes, preserving the outcome-only supervision setting and producing an initialization consumable by multiple ABL learners. Semantic deciphering and label-free verification provide complementary end-to-end formulations: Tao et al. (2024) decipher raw neuro-symbolic data with guarantees tied to knowledge-base efficacy, while Jia et al. (2025) replace label-based reasoning with label-free verification. This decomposition exposes two questions that standard ABL leaves implicit: whether the representation preserves concept distinctions and whether the KB interprets the induced classes correctly.

Representation-level strategies are closest to our approach. Concept bottleneck models (Koh et al., 2020) make intermediate semantics explicit under concept supervision, whereas BASIL aligns the bottleneck from outcomes alone. Dai et al. (2019) retrain perception on abduction-revised pseudo-labels, while Huang et al. (2021) optimize embedding-space similarity consistency. Marconato et al. (2023b) further show that concept-extractor disentanglement sharply reduces shortcut density. BASIL makes the remaining alignment step explicit: Stage I learns a cluster structure, and Stage II assigns those clusters bijectively to predicates. Every learner grid in Section 5 therefore includes +PT, the strongest matched outcome-only intervention in our experiments. Methods that use concept anchors or change the KB curriculum operate through different information interfaces.

Logic-Guided Neural Learning. Other neuro-symbolic frameworks integrate logical structure directly into neural learning. DeepProbLog (Manhaeve et al., 2018) combines neural predicates with probabilistic logic programming and supports end-to-end training from examples. Logic Tensor Networks (Serafini & d'Avila Garcez, 2016) learn numerical groundings under many-valued logical constraints, while NeurASP (Yang et al., 2020) uses answer set programs to constrain neural predictions and guide training. BASIL intervenes before downstream learning by establishing the correspondence between learned clusters and logical predicates. The initialized learner then continues its own logic-guided updates, as illustrated by our A^{3}BL and DeepStochLog transfer experiments. Semantic initialization thus complements logical feedback during downstream training.

Contrastive Learning. Contrastive learning learns invariant representations by pulling similar pairs together while separating dissimilar pairs in embedding space (Hadsell et al., 2006). SimCLR (Chen et al., 2020) operationalizes this principle via strong stochastic augmentations and large-batch instance discrimination, whereas MoCo (He et al., 2020) uses a momentum encoder and a queue to provide a large and consistent set of negatives. BASIL uses Contrastive Clustering (Li et al., 2021), coupling instance-level contrast with a cluster-level objective over soft assignments.

B BENIGN-EQUIVALENCE INVARIANCE OF ABL

We state the precise relabeling conditions used in Theorem 1 and prove equality in distribution of the resulting parameter trajectories. Let $D = \{(\boldsymbol{x}^{(i)}, y^{(i)})\}_{i=1}^{N}$ be the training set and fix an abstraction α. For $\iota : [k] \to \mathcal{P}$, let $\boldsymbol{g}^{\iota,\alpha}(\boldsymbol{x}^{(i)})$ denote the pseudo-groundings in Equation (1).

We represent an ABL run by a training state $H = (\theta, \sigma, t)$, where σ contains the optimizer and scheduler state. At each iteration, ABL: (i) constructs pseudo-groundings with the current predictor; (ii) samples corrected pseudo-groundings $\hat{\boldsymbol{G}} = (\hat{\boldsymbol{g}}^{(i)})_{i=1}^{N}$ from a conditional distribution $M_D(\cdot \mid (\boldsymbol{g}^{(i)})_i)$ supported on $F_y := \{\widetilde{\boldsymbol{g}} : \mathrm{R}_{\mathrm{KB}}(\widetilde{\boldsymbol{g}}) = y\}$; and (iii) maps the corrections back to pseudo-label targets $S = (\iota^{-1}(\hat{\boldsymbol{g}}^{(i)}))_i$ and updates the state as $H^+ = U(H, S; \xi)$, where ξ collects the randomness of the training update. When ABL uses the correction objective in Equation (2), M_D is supported on its minimizers. The parameter trajectory is the θ component of the state sequence $(H_t)_{t\geq 0}$.

Full statement of Theorem 1. Fix an abstraction α and a bijective interpretation $\iota \in \mathcal{I}_{\mathrm{bij}}$. Let π be a benign reasoning shortcut acting on pseudo-groundings as $\pi(P(c)) = (\pi(P))(c)$ and element-wise on sequences. Consider two ABL runs with this abstraction, one under ι and the other under $\pi \circ \iota$, initialized from the same state $H^{(0)} = (\theta^{(0)}, \sigma^{(0)}, 0)$. Assume:

1. (*Relabeling-invariant correction loss*) $\mathcal{L}(\pi(\widetilde{\boldsymbol{g}}), \pi(\boldsymbol{g})) = \mathcal{L}(\widetilde{\boldsymbol{g}}, \boldsymbol{g})$ for all $\boldsymbol{g}, \widetilde{\boldsymbol{g}} \in \mathcal{G}^{<\infty}$.
2. (*Feasibility is preserved*) for every $i \in [N]$ and candidate $\widetilde{\boldsymbol{g}}$,
$$\mathrm{R}_{\mathrm{KB}}(\widetilde{\boldsymbol{g}}) = y^{(i)} \quad \Longleftrightarrow \quad \mathrm{R}_{\mathrm{KB}}(\pi(\widetilde{\boldsymbol{g}})) = y^{(i)}. \tag{9}$$
3. (*Abduction respects relabeling*) $M_D(\cdot \mid (\pi\boldsymbol{g}^{(i)})_i) = (\pi^{\otimes})_* M_D(\cdot \mid (\boldsymbol{g}^{(i)})_i)$, where $(\pi^{\otimes})_*$ is the distribution obtained by applying π to every corrected pseudo-grounding in the batch.
4. (*Training does not depend on predicate names*) conditional on H, the law of ξ is the same under both interpretations and independent of the abduction draw, and both runs share U.

Let $\Lambda(\iota) = \{\theta_\iota^{(t)}\}_{t\geq 0}$ be the parameter trajectory under ι and fixed α. Then
$$\Lambda(\iota) \overset{d}{=} \Lambda(\pi \circ \iota). \tag{10}$$
ABL can neither distinguish benign-equivalent interpretations nor uniquely select one.

Proof. Write $\iota' := \pi \circ \iota$ and couple the two runs on a common probability space. They share the training randomness ξ_t. Whenever the ι-run samples $\hat{\boldsymbol{G}}$, define the correction in the ι'-run as $\hat{\boldsymbol{G}}' := \pi^{\otimes}(\hat{\boldsymbol{G}})$. The third condition gives this correction the required conditional distribution.

Base case. Both runs start from $H_\iota^{(0)} = H_{\iota'}^{(0)} = H^{(0)}$.

Inductive step. Suppose $H_\iota^{(t)} = H_{\iota'}^{(t)} =: H$ almost surely. The runs use the same predictor and therefore produce the same pseudo-labels. Because α is fixed, $\boldsymbol{g}^{\iota',\alpha}(\boldsymbol{x}^{(i)}) = \pi(\boldsymbol{g}^{\iota,\alpha}(\boldsymbol{x}^{(i)}))$ for every training example. The first two conditions map the correction problem and its minimizers one-to-one under π. The coupled correction therefore satisfies $\hat{\boldsymbol{g}}'^{(i)} = \pi(\hat{\boldsymbol{g}}^{(i)})$ almost surely.

Since $\iota'^{-1} = \iota^{-1} \circ \pi^{-1}$ element-wise, mapping the corrections back to pseudo-label targets gives
$$S'^{(i)} = \iota'^{-1}(\hat{\boldsymbol{g}}'^{(i)}) = \iota^{-1}(\pi^{-1}\pi(\hat{\boldsymbol{g}}^{(i)})) = \iota^{-1}(\hat{\boldsymbol{g}}^{(i)}) = S^{(i)} \quad \text{a.s.} \tag{11}$$
Thus $S' = S$. With the same update map and shared ξ_t, $H_{\iota'}^{(t+1)} = U(H, S'; \xi_t) = U(H, S; \xi_t) = H_\iota^{(t+1)}$ almost surely. Induction proves equality of the coupled state and parameter trajectories.

For independently randomized runs, the third condition gives $S' \overset{d}{=} S$ conditional on any common state H. The fourth condition then gives the two state processes the same initial state and transition distribution. Their parameter trajectories therefore have the same distribution. □

Scope of the result. The correction problem and all random choices must respect benign relabeling. The assumptions cover randomized abduction, random tie-breaking, mini-batch sampling, data shuffling, dropout, and optimizer randomness, provided none of these mechanisms depends on predicate names. An external signal that favors one naming can break the invariance.

C BASIL: ALGORITHM AND IMPLEMENTATION DETAILS

This appendix details Stage I pretraining and Stage II search, decoding, and model handoff.

C.1 STAGE I: CLUSTER PRETRAINING AND STRONG BINDING

For our visual experiments, Contrastive Clustering (CC) (Li et al., 2021) jointly learns image representations and cluster assignments from unlabeled data. For each image x, it samples two augmented views $x^{(1)}$ and $x^{(2)}$ and passes them through a shared encoder. An instance-level head produces embeddings, while a cluster-level head produces soft assignments over k clusters.

Objectives. The instance-level contrastive loss treats the two views of the same image as a positive pair and other images in the batch as negatives. The cluster-level loss applies the same idea to the columns of the soft-assignment matrices, aligning each cluster across views while separating different clusters. CC optimizes the sum of these two losses end to end.

After pretraining, the cluster-head argmax gives $z_\theta(x)$, and one logical constant per cluster defines α. The instance embedding has dimension $d = 256$. With this clustering fixed, Stage II can rename a cluster but cannot split inputs within it. This restriction applies to initialization, while subsequent ABL training can change the predictor. Tables 3 and 4 list the training settings and augmentations.

The cluster head has $k = |\mathcal{P}|$ outputs, matching the task's logical concept vocabulary. This sets the number of available clusters without assuming that their assignments are perfectly pure; Theorem 2 explicitly allows cluster errors. An overclustered representation with $k' > k$ would require a more flexible correspondence to the k predicates. Mapping several subclusters of the same concept to one predicate is distinct from a malignant shortcut that merges different concepts. The present Stage II bijection does not implement this more general correspondence.

Table 3: Stage I pretraining hyperparameters.

Hyperparameter	Value
epochs	1000
batch size	1024
instance embedding dimension d	256
instance temperature	0.5
optimizer	Adam, lr $= 3 \times 10^{-3}$, betas $(0.9, 0.999)$

Two stochastic views are generated with the pipelines in Table 4.

Table 4: Image augmentations used for Stage I pretraining.

Dataset	Augmentation pipeline
MNIST	Random resized crop to 28×28 (scale 0.8–1.0), rotation up to 15°, affine transformation (translation up to 10% per axis, scale 0.9–1.1, shear up to 10°), and random erasing ($p = 0.5$, area fraction 0.02–0.1)
CIFAR-10	Random resized crop to 32×32 (scale 0.2–1.0, aspect ratio 0.75–1.33), horizontal flip ($p = 0.5$), color jitter (brightness, contrast, saturation 0.4 and hue 0.1, $p = 0.8$), and grayscale conversion ($p = 0.2$)

C.2 STAGE II: SEMANTIC ALIGNMENT AND MODEL HANDOFF

Algorithm 1 searches over bijective interpretations while keeping the Stage I predictor and abstraction fixed. The final matching is applied directly to the classifier outputs.

Algorithm 1 Stage II semantic alignment by stabilized Gumbel-Parallel CEM

Input: Dataset D, knowledge base KB, fixed z_θ and α, total budget M, rounds T, elite ratio ρ, update weight λ, uniform-mixing coefficient η, Gumbel temperature τ_g
Output: Decoded bijection $\hat{\iota} \in \mathcal{I}_{\text{bij}}$

Set $m \leftarrow M/T$; initialize $\mathbf{Q}^{(0)} \leftarrow \mathbf{U}^{\text{unif}}$.
for $t = 0, \ldots, T-1$ **do**
  for each of m candidates **do**
    Draw one utility matrix U using Equation (4); set $R \leftarrow [k]$, $C \leftarrow \mathcal{P}$.
    while $R \neq \varnothing$ **do**
      Each $z \in R$ proposes to $p_z = \arg\max_{p \in C} U_{z,p}$.
      Each predicate receiving proposals accepts the proposing cluster with largest $U_{z,p}$.
      Fix all accepted pairs permanently and remove their clusters and predicates from R and C.
  Score each proposal by the empirical logic-consistency score in Equation (3).
  Let $\mathcal{J}_t$ be the top $\lceil \rho m \rceil$ proposals.
  Compute score-softmax elite weights and the elite-frequency matrix $\mathbf{E}^{(t)}$.
  Update $\mathbf{Q}^{(t+1)}$ by Equation (5).
return $\hat{\iota} = \text{Dec}(\mathbf{Q}^{(T)})$.

Consistency score and elite weights. Equation (3) weights each example by its sequence length T_i. For the score-softmax calculation, we use the equivalent unnormalized count

$$S_D(\iota) = \sum_{i=1}^{N} T_i \, \mathbb{I}\Big[\mathrm{R}_{\mathrm{KB}}\Big(\boldsymbol{g}^{\iota,\alpha}(\boldsymbol{x}^{(i)})\Big) = y^{(i)}\Big] .$$

Normalization by $\sum_i T_i$ leaves candidate rankings unchanged. The nonlinear weighting below uses counts $S_D \in [0, \sum_i T_i]$ rather than normalized scores $\widehat{s}_D \in [0, 1]$.

All main experiments score $M = 50{,}000$ bijections over $T = 50$ rounds, giving $m = 1{,}000$ candidates per round. Each round retains the top 10% candidates and aggregates them with score-softmax weights over shifted log consistency scores. For elite count $S_j = S_D(\iota_j)$, we use

$$a_j = \log\left(1 + S_j - \min_{\ell \in \mathcal{J}_t} S_\ell\right), \tag{12}$$

$$\omega_j = \frac{\exp((a_j - \max_{\ell \in \mathcal{J}_t} a_\ell)/\tau_s)}{\sum_{\ell \in \mathcal{J}_t} \exp((a_\ell - \max_{u \in \mathcal{J}_t} a_u)/\tau_s)}, \quad \tau_s = \max(\mathrm{std}\{a_\ell\}, 10^{-6}), \tag{13}$$

and

$$\mathbf{E}^{(t)}_{z,p} = \sum_{j \in \mathcal{J}_t} \omega_j \, \mathbf{1}[\iota_j(z) = p]. \tag{14}$$

Subtracting the smallest elite count makes every offset nonnegative, and the logarithm compresses differences between counts while preserving their order. The temperature τ_s adapts the softmax scale to the spread of the transformed counts in the current elite set; equal counts yield uniform weights. Thus the logarithm changes relative count differences, while τ_s controls their overall scale. The decoding analysis depends on the resulting elite-frequency matrices and does not require this particular score transform. The proposal update uses $\lambda = 0.7$, uniform-mixing coefficient $\eta = 0.10$, and Gumbel temperature $\tau_g = 1.0$. The returned interpretation is the maximum-weight matching decoded from the final proposal matrix. To apply it to the predictor, let the final linear classification layer have weights W and bias b. We replace them with $A_\iota^\top W$ and $A_\iota^\top b$, respectively, which permutes the k output classes from cluster order into predicate order without gradient updates.

One-shot controls. Best-of-M scores $M = 50{,}000$ uniformly sampled bijections and returns the best one. For the visual tasks, the Hungarian control draws the same number of uniform candidates and retains the top 20% by logic consistency. It uses the shifted-log, automatic-temperature softmax in Equations (12)–(13), but accumulates assignment-specific evidence from correctly reasoned examples: for candidate j, each occurrence of cluster z in a consistent example adds one to entry $(z, \iota_j(z))$ of a matrix V_j. A single maximum-weight matching is then decoded from $\sum_j \omega_j V_j$. Unlike the CEM elite-frequency matrix, these entries weight how often an assignment is supported by the data,

Table 5: Stage II alignment hyperparameters.

Hyperparameter	Value
total scored bijections M	50,000
CEM rounds T	50
elite ratio ρ	0.10
elite aggregation	score-softmax over shifted log scores
update weight λ	0.70
uniform-mixing coefficient η	0.10
realized per-entry floor	$0.10/k$
Gumbel temperature τ_g	1.0
primary output	maximum-weight matching

and the proposal distribution is never updated. For Large-K, the top 10% of uniform candidates form a rank-weighted assignment-frequency matrix. For h candidates ranked by descending score, rank $j \in \{0, \ldots, h-1\}$ has normalized weight proportional to $\exp[-j/\max(h/5, 1)]$. It uses maximum-weight matching without proposal updates or uniform mixing.

Ablation variants. The SA-only ablation initializes the predictor randomly before Stage II. BASIL (w/o bij.) treats clusters independently during proposal generation and assigns each cluster the predicate with the largest final proposal weight, so predicates may be duplicated or left unused. For the visual-task constraint comparisons, both the non-bijective variant and its paired bijective control apply the decoded mapping through five epochs of neural training on the targets $\hat{\iota}(z_\theta(x))$, using Adam with learning rate 10^{-3} and batch size 256. Clusters mapped to the same predicate supply the same target label, while unused predicates receive no assigned examples. All k classifier outputs remain trainable during this handoff and subsequent ABL training. For Large-K, both variants set each $(z, \hat{\iota}(z))$ entry of the $k \times k$ logit table to 6 and all others to zero, without gradient updates.

Cost. Stage II scores M interpretations on N training instances, requiring $\mathcal{O}(MN)$ KB evaluations, followed by an $\mathcal{O}(k^3)$ final matching problem. If one KB evaluation costs C_{KB}, the scoring cost is $\mathcal{O}(MNC_{\mathrm{KB}})$, in addition to proposal construction and matrix updates. Our implementation reuses the fixed Stage I predictions and evaluates candidate assignments in batches; the arithmetic tasks also admit vectorized consistency checks. These operations avoid repeated neural inference during search, but the cost of reasoning still depends on the knowledge base. For a KB requiring expensive inference, its evaluation cost must therefore be included when choosing the sampling budget.

D Analysis of Stage II

We first establish score separation and final decoding, then examine decoding under benign symmetry and conditional continuation of the actual sampler. Each empirical certificate accompanies the result it evaluates. Ground-truth concept labels are used only for post-hoc analysis, never by BASIL.

D.1 Score Separation under Cluster Errors

Standing assumptions. Write $\mathcal{D} := \mathcal{D}_{\mathcal{X}}$ for the input-sequence distribution. We assume throughout that (i) sequence lengths satisfy $1 \le T_{\min} \le T(\boldsymbol{x}) \le T_{\max}$, (ii) supervision is correct, so $y = \mathrm{R}_{\mathrm{KB}}(\boldsymbol{g}^\star(\boldsymbol{x}))$, and (iii) injectively renaming logical constants does not change R_{KB}. Thus KB reasoning depends on predicate labels and shared constants, irrespective of the constants' names.

Benign interpretations and expected scores. Because Stage I assigns one distinct constant per cluster, write $\alpha(x) = c(z_\theta(x))$ for an injective map $c : [k] \to \mathcal{C}$. Choose a bijection $\iota^\star : [k] \to \mathcal{P}$ that minimizes the length-weighted Stage I cluster error defined below. For any $\iota \in \mathcal{I}_{\mathrm{bij}}$, the induced concept relabeling is $\pi_\iota := \iota \circ (\iota^\star)^{-1}$. The benign interpretation set is therefore

$$\mathcal{A} := \{\iota \in \mathcal{I}_{\mathrm{bij}} : \pi_\iota \text{ is a benign shortcut}\} = \{\pi \circ \iota^\star : \pi \text{ is a benign shortcut}\}.$$

The implementation weights an example by its sequence length. Set $w(\boldsymbol{x}) := T(\boldsymbol{x})/T_{\max} \in [1/\kappa, 1]$, where $\kappa = T_{\max}/T_{\min}$, and let $\mu_w := \mathbb{E}_{\mathcal{D}}[w(\boldsymbol{x})]$. Let I_ι be the consistency indicator under perfect strong binding and $\widetilde{I}_\iota$ the indicator under the actual Stage I clusters. Then

$$s(\iota) = \frac{\mathbb{E}_{\mathcal{D}}[wI_\iota]}{\mu_w}, \qquad \widetilde{s}(\iota) = \frac{\mathbb{E}_{\mathcal{D}}[w\widetilde{I}_\iota]}{\mu_w},$$

and $\widehat{s}_D$ in Equation (3) is the empirical counterpart of $\widetilde{s}$.

Define the sequence-level cluster-error event

$$\mathcal{E}_b(\boldsymbol{x}) \coloneqq \{\exists t : \iota^\star(z_\theta(x_t)) \neq s^\star(x_t)\}$$

and its length-weighted probability

$$\epsilon_b \coloneqq \frac{\mathbb{E}_{\mathcal{D}}[w\,\mathbb{I}[\mathcal{E}_b(\boldsymbol{x})]]}{\mu_w} = \frac{\mathbb{E}_{\mathcal{D}}[T(\boldsymbol{x})\,\mathbb{I}[\mathcal{E}_b(\boldsymbol{x})]]}{\mathbb{E}_{\mathcal{D}}[T(\boldsymbol{x})]}. \tag{15}$$

Perfect strong binding implies $\epsilon_b = 0$; conversely, $\epsilon_b = 0$ means that the Stage I clusters match the ground-truth concepts $\mathcal{D}$-almost surely after the best bijective alignment.

When $\mathcal{A} \subsetneq \mathcal{I}_{\mathrm{bij}}$, define the ideal consistency gap

$$\Delta_{\mathcal{A}} \coloneqq \min_{\iota^\circ \in \mathcal{A},\ \iota \notin \mathcal{A}} \big(s(\iota^\circ) - s(\iota)\big). \tag{16}$$

Benign shortcuts preserve every reasoning result on the support of $\mathcal{D}$, so each member of $\mathcal{A}$ has ideal score one. An interpretation outside $\mathcal{A}$ has expected score one only if it disagrees on a zero-probability subset of that support; then $\Delta_{\mathcal{A}} = 0$, and the theorem makes no identification claim.

Proof of Theorem 2. **Step 1: ideal scores.** Perfect strong binding implies $z_\theta(x_t) = (\iota^\star)^{-1}(s^\star(x_t))$ at every position. The Stage I constants therefore reproduce the same co-reference pattern as the ground-truth concepts, up to an injective renaming of constants. By invariance to logical-constant renaming, $\mathrm{R}_{\mathrm{KB}}(\boldsymbol{g}^{\iota,\alpha}(\boldsymbol{x})) = \mathrm{R}_{\mathrm{KB}}(\pi_\iota(\boldsymbol{g}^\star(\boldsymbol{x})))$. With correct supervision, this gives

$$s(\iota) = \frac{\mathbb{E}_{\mathcal{D}}[w\,\mathbb{I}[\mathrm{R}_{\mathrm{KB}}(\pi_\iota(\boldsymbol{g}^\star)) = \mathrm{R}_{\mathrm{KB}}(\boldsymbol{g}^\star)]]}{\mu_w}.$$

For $\iota^\circ \in \mathcal{A}$, the induced relabeling is a benign shortcut, so the indicator equals one throughout the support and $s(\iota^\circ) = 1$. The remaining inequality follows from Equation (16); if the gap is positive, every interpretation outside $\mathcal{A}$ has strictly smaller score.

Step 2: cluster perturbations. Outside $\mathcal{E}_b(\boldsymbol{x})$, every Stage I cluster agrees with the aligned ground-truth concept. The actual pseudo-grounding and the pseudo-grounding under perfect strong binding then differ only by an injective renaming of constants, so their consistency indicators are identical for every ι. The expected scores can therefore differ only on sequences in $\mathcal{E}_b$, whose length-weighted probability is ϵ_b. Thus $\sup_\iota |\widetilde{s}(\iota) - s(\iota)| \leq \epsilon_b$. Applying this bound to $\iota^\circ \in \mathcal{A}$ and $\iota \notin \mathcal{A}$ gives

$$\widetilde{s}(\iota^\circ) - \widetilde{s}(\iota) \geq \Delta_{\mathcal{A}} - 2\epsilon_b.$$

The condition $\Delta_{\mathcal{A}} > 2\epsilon_b$ therefore places every score maximizer in $\mathcal{A}$. □

Corollary D.1 (Finite-sample identification)**.** *Under the assumptions of Theorem 2, write $\kappa = T_{\max}/T_{\min}$ and fix $(z_\theta, \alpha, \iota^\star)$ before drawing $D \sim \mathcal{D}^N$ i.i.d. If $\Delta_{\mathcal{A}} > 2\epsilon_b$, then with probability at least $1-\delta$, every maximizer of the empirical score $\widehat{s}_D$ lies in $\mathcal{A}$ whenever*

$$N \geq 2\kappa^2(\Delta_{\mathcal{A}} - 2\epsilon_b)^{-2} \log \frac{2k!}{\delta}.$$

With perfect strong binding and $\Delta_{\mathcal{A}} > 0$, the empirical argmax set equals $\mathcal{A}$ on this event.

Proof. Set $\gamma = \Delta_{\mathcal{A}} - 2\epsilon_b > 0$ and $Z_i = w_i(\widetilde{I}_{\iota,i} - \widetilde{s}(\iota))$. These i.i.d. centered variables have range at most one, and $N^{-1}\sum_i w_i \geq 1/\kappa$. Hoeffding's inequality and a union bound over $k!$ bijections give

$$\Pr\left[\sup_\iota |\widehat{s}_D(\iota) - \widetilde{s}(\iota)| \geq \gamma/2\right] \leq 2k! \exp\left(-\frac{N\gamma^2}{2\kappa^2}\right) \leq \delta.$$

On the complementary event, estimation cannot close the population gap. Under perfect strong binding, every benign interpretation has empirical score one. □

Corollary D.2 (Fixed-dataset certificate)**.** *Fix any finite scoring dataset D with correct reasoning-result labels and Stage I quantities $(z_\theta, \alpha, \iota^\star)$ defined above. Let $\widehat{s}^\star_D$ be the length-weighted empirical score under perfect strong binding, and let $\widehat{\mathcal{M}}^\star_D \coloneqq \arg\max_\iota \widehat{s}^\star_D(\iota) \subsetneq \mathcal{I}_{\mathrm{bij}}$. Define*

$$\widehat{\Delta}_D \coloneqq 1 - \max_{\iota \notin \widehat{\mathcal{M}}^\star_D} \widehat{s}^\star_D(\iota), \qquad \widehat{\epsilon}_{b,D} \coloneqq \frac{\sum_i T_i \mathbb{I}[\mathcal{E}_b(\boldsymbol{x}^{(i)})]}{\sum_i T_i}.$$

If $\widehat{\Delta}_D > 2\widehat{\epsilon}_{b,D}$, every maximizer of the empirical score under the actual Stage I clusters lies in $\widehat{\mathcal{M}}^\star_D$. Under perfect strong binding, the argmax sets coincide. Stage I need not be independent of D.

Proof. For every interpretation, the two empirical consistency indicators can differ only on sequences in $\mathcal{E}_b$, so $|\widehat{s}_D(\iota) - \widehat{s}^\star_D(\iota)| \le \widehat{\epsilon}_{b,D}$ simultaneously. Every member of $\widehat{\mathcal{M}}^\star_D$ has ideal empirical score one, whereas every interpretation outside it has score at most $1 - \widehat{\Delta}_D$. Perturbing both sides by at most $\widehat{\epsilon}_{b,D}$ preserves a strict gap when $\widehat{\Delta}_D > 2\widehat{\epsilon}_{b,D}$. □

Uniform-search budget. Under the concentration event of Corollary D.1, Best-of-M succeeds whenever its M independent uniform draws include a benign interpretation. Consequently,

$$\Pr[\hat{\iota}_M \notin \mathcal{A}] \le \delta + (1 - |\mathcal{A}|/k!)^M \le \delta + \exp(-M|\mathcal{A}|/k!).$$

The chance of even proposing a benign interpretation is at most $M|\mathcal{A}|/k!$, explaining why a fixed budget can cover the 120 permutations in Addition-Half yet miss the benign set at large k.

A sharper bound for the Large-K noise channel. The symmetric endpoint corruption in Large-K permits an exact score decomposition, beyond the worst-case perturbation bound above.

Proposition D.1 (Unbiased endpoint corruption rescales the score landscape)**.** *Let $k \ge 2$ and $p \in [0,1]$. Consider a circulant relation with difference colors $c : \mathbb{Z}_k \setminus \{0\} \to \{0, \dots, q-1\}$, observations over all ordered off-diagonal pairs, and the Large-K channel: each endpoint is kept with probability p and otherwise replaced by a uniformly random wrong cluster, independently, while the supervised relation label is computed from the true pair; equal-endpoint predictions map to a sentinel that never matches a relation color. Then for every $\iota \in \mathcal{I}_{\mathrm{bij}}$,*

$$\tilde{s}_p(\iota) = \beta\, s(\iota) + c_0, \qquad \beta = \left(p - \tfrac{1-p}{k-1}\right)^2,$$

where s is the expected score under perfect strong binding and c_0 is independent of ι. For $\beta > 0$, the expected-score maximizers are unchanged and the consistency gap is $\beta\,\Delta_{\mathcal{A}}$.

Proof. Fix a pair (i, j) with relation label $y = c(j - i)$ and let $m_y := |\{d \ne 0 : c(d) = y\}|$. Write π for the composite concept relabeling induced by ι and $\chi := \mathbb{I}[c(\pi(j) - \pi(i)) = y]$ for the clean agreement indicator. A corrupted endpoint is uniform over the $k - 1$ wrong clusters, and a bijection maps these onto the $k - 1$ concepts other than the true image, while an equal-endpoint prediction hits the diagonal sentinel and never agrees. If only the source is corrupted, its image a is uniform over the $k - 1$ values in $\mathcal{P} \setminus \{\pi(i)\}$. Since $a \mapsto \pi(j) - a$ sweeps every nonzero difference exactly once as a ranges over $\mathcal{P} \setminus \{\pi(j)\}$, exactly m_y values of $a \ne \pi(j)$ agree; restricting to the actual domain removes $a = \pi(i)$ (which agrees iff $\chi = 1$) and adds $a = \pi(j)$ (the sentinel, which never agrees), giving agreement probability $(m_y - \chi)/(k-1)$; the target-only case is symmetric. If both endpoints are corrupted, inclusion–exclusion gives $((k-2)m_y + \chi)/(k-1)^2$. Combining the four cases yields

$$p^2\chi + 2p(1-p)\tfrac{m_y - \chi}{k-1} + (1-p)^2\tfrac{(k-2)m_y + \chi}{(k-1)^2} = \beta\chi + \text{terms independent of } \iota,$$

with $\beta = p^2 - \frac{2p(1-p)}{k-1} + \frac{(1-p)^2}{(k-1)^2} = (p - \frac{1-p}{k-1})^2$. Averaging over pairs proves the claim, since the offset depends only on the relation's color counts $m_{c(j-i)}$ and not on ι. □

Large-K. For the Large-K generator used in the experiments ($q = 8$ relation colors and seeds 1–20 for each k), exact backtracking verifies that the only concept permutations preserving every relation color are the k cyclic rotations. Thus $|\mathcal{A}| = k$ on all 60 instances. An exact branch-and-bound search then finds the best non-benign permutation and computes the corresponding consistency gap $\Delta_{\mathcal{A}}$. Rotational symmetry lets the search fix one vertex, while disagreements already determined and a minimum-cost matching over the unassigned vertices provide a valid pruning bound. The globally best non-benign competitor is a transposition in every instance.

The endpoint corruption used in the experiments rescales the expected score gap by the factor β in Proposition D.1 without changing the score ordering. A second exact branch-and-bound certificate operates directly on each sampled noisy scoring set and proves that every interpretation outside the benign orbit has strictly lower score. Table 6 summarizes both certificates.

The mean clean gap is 14.1% at $k = 20$ and 6.2% at $k = 50$. Despite this smaller gap, exact certification confirms strict optimality of the benign orbit on all 60 noisy scoring sets.

Table 6: Exact Large-K score certificates over relation seeds 1–20. Expected-score gaps are fractions of the $k(k-1)$ clean ordered pairs; β is the exact rescaling under the $p = 0.7$ endpoint channel. The final column certifies strict optimality of the benign orbit on each sampled scoring set.

k	$\lvert\mathcal{A}\rvert$	$\Delta_{\mathcal{A}}$ range (mean)	β	$\beta\Delta_{\mathcal{A}}$ range	Strict empirical optima
20	20	10.5–16.3% (14.1%)	0.468	4.9–7.6%	20/20
30	30	9.2–10.8% (9.9%)	0.476	4.4–5.1%	20/20
50	50	5.7–6.4% (6.2%)	0.481	2.8–3.1%	20/20

Decipher. For each Decipher rendering, we evaluate 20 Stage I predictors on the same training set and exhaustively score all $10! = 3{,}628{,}800$ bijections. Under perfect strong binding, the intended interpretation is the unique empirical optimum, with exact length-weighted gaps of 40.6% on MNIST and 40.7% on CIFAR-10. Under the actual Stage I clusters, exhaustive evaluation returns the intended majority interpretation as the global maximizer in all 40 checkpoint-level analyses. On MNIST, the ideal empirical gap exceeds twice the length-weighted fraction of equations containing a Stage I cluster error for all 20 checkpoints, and the intended maximizers have a mean gap of 38.9%. For CIFAR-10, the generic perturbation bound is more conservative, but exact enumeration still certifies the intended maximizer in all 20 checkpoints, with a mean gap of 4.6%.

D.2 Final-Decoding Guarantee

Proof of Theorem 3. Let $\varnothing \neq \mathcal{A} \subsetneq \mathcal{I}_{\mathrm{bij}}$ and fix $\iota^\dagger \in \mathcal{A}$. Define its final decode margin as

$$m_T(\iota^\dagger) \coloneqq \langle \mathbf{Q}^{(T)}, A_{\iota^\dagger} \rangle - \max_{\iota \in \mathcal{I}_{\mathrm{bij}} \setminus \mathcal{A}} \langle \mathbf{Q}^{(T)}, A_\iota \rangle,$$

which compares its matching score with the best interpretation outside $\mathcal{A}$. For $\lambda, \eta \in [0,1]$ and $\mathbf{Q}^{(0)} = \mathbf{U}^{\mathrm{unif}}$, we establish the following stronger bound on the final decode margin:

$$m_T(\iota^\dagger) \geq (1-\eta)\lambda \sum_{t=0}^{T-1} a^{T-1-t} \gamma_t(\iota^\dagger), \tag{17}$$

where $a = (1-\eta)(1-\lambda)$ and $\gamma_t(\iota^\dagger)$ is the elite margin defined in Section 4.3.

Fix a competitor $\iota \notin \mathcal{A}$ and write its matching-score difference from $\iota^\dagger$ as

$$d_t(\iota) \coloneqq \langle \mathbf{Q}^{(t)}, A_{\iota^\dagger} \rangle - \langle \mathbf{Q}^{(t)}, A_\iota \rangle.$$

At uniform initialization, every bijection selects k entries, so $d_0(\iota) = 0$. Equation (5) gives

$$\mathbf{Q}^{(t+1)} = (1-\eta)(1-\lambda)\mathbf{Q}^{(t)} + (1-\eta)\lambda \mathbf{E}^{(t)} + \eta \mathbf{U}^{\mathrm{unif}}.$$

The uniform term contributes $\langle \mathbf{U}^{\mathrm{unif}}, A_\iota \rangle = 1$ for every assignment matrix, so it cancels in the score difference. With $a = (1-\eta)(1-\lambda)$, the remaining terms give

$$d_{t+1}(\iota) = a\, d_t(\iota) + (1-\eta)\lambda \left(\langle \mathbf{E}^{(t)}, A_{\iota^\dagger} \rangle - \langle \mathbf{E}^{(t)}, A_\iota \rangle \right).$$

The parenthesized term is at least $\gamma_t(\iota^\dagger)$ for each $\iota \notin \mathcal{A}$. Unrolling from $d_0(\iota) = 0$ yields

$$d_T(\iota) \geq (1-\eta)\lambda \sum_{t=0}^{T-1} a^{T-1-t} \gamma_t(\iota^\dagger).$$

Taking the minimum over $\iota \notin \mathcal{A}$ gives $m_T(\iota^\dagger)$ and establishes Equation (17). Under the assumptions of Theorem 3, $(1-\eta)\lambda > 0$, so a positive weighted sum makes $m_T(\iota^\dagger) > 0$. No interpretation outside $\mathcal{A}$ can then match the score of $\iota^\dagger$ in Equation (6), so the final decode belongs to $\mathcal{A}$. □

Post-hoc certificate evaluation. The right-hand side of Equation (17) is computed from the elite matrices along a CEM trajectory. The proof applies to any nonempty proper target set, so these diagnostics use the empirical reference set $\mathcal{A}_r$ defined in Appendix E.1. A positive value certifies that the final matching belongs to this set. A trajectory is *Stage-I-supported* if its majority cluster-to-concept assignment is bijective, providing a benign reference interpretation for the certificate.

All 57 supported trajectories end with a positive certificate, and the proposal mass concentrates on the same benign mode. For Decipher and Addition-Half, the final mode masses and margins match the closed-form locked-state values from Appendix D.3 to four decimals. For Large-K, the reference shift selected at round 25 equals the final decode in all 20 runs; the median final fixed-mode mass is 0.833, and the final certificates range from 0.859 to 1.577.

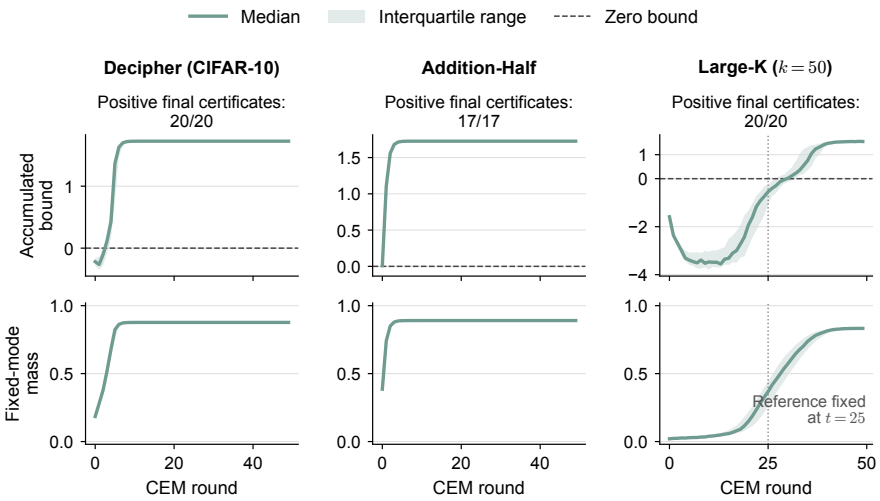

Figure 6: Empirical final-decode certificates across CEM rounds. Top: the accumulated margin bound on the right-hand side of Equation (17), including the factor $(1-\eta)\lambda$. Positive values certify a benign final decode. Bottom: proposal mass on the same fixed benign interpretation. Curves show medians and interquartile ranges over 20 Decipher, 17 Addition-Half, and 20 Large-K Stage I-supported trajectories. The three excluded Addition-Half runs have non-bijective Stage I majority assignments. The annotations give the number of positive final certificates; dotted vertical lines mark the selection of the Large-K reference at round 25, after which it is held fixed.

D.3 Decoding under Benign Symmetry

A benign interpretation need not absorb all proposal mass to be decoded correctly. For a benign set with disjoint assignment supports, the elite updates reduce to one scalar per mode, and a strict leader suffices. Cyclic shifts in Large-K have this structure, since distinct shifts differ in every row.

Proposition D.2 (Elite dynamics and benign-set decoding)**.** *Let $\lambda \in (0,1]$, $\eta \in [0,1)$, and let $\mathcal{A} = \{\iota_1, \dots, \iota_L\}$, $L \geq 1$, have pairwise-disjoint assignment-matrix supports, let $\mathbf{Q}^{(0)} = \mathbf{U}^{\mathrm{unif}}$, and suppose in every round the elite multiset consists of orbit members with weight fractions $(w_1^{(t)}, \dots, w_L^{(t)})$ in the probability simplex. Then every entry of $\mathbf{Q}^{(t)}$ in the support of ι_g equals a scalar $q_g^{(t)}$, every entry outside the orbit supports equals $q_0^{(t)}$, and with $a = (1-\eta)(1-\lambda)$,*

$$q_g^{(t+1)} = a\, q_g^{(t)} + (1-\eta)\lambda\, w_g^{(t)} + \eta/k, \qquad q_0^{(t+1)} = a\, q_0^{(t)} + \eta/k.$$

Assume $\mathcal{A} \subsetneq \mathcal{I}_{\mathrm{bij}}$ and let $q_{(1)}^{(T)}$ be the largest benign-mode mass. Define $q_{\mathrm{alt}}^{(T)} = q_0^{(T)}$ when $L = 1$, and $q_{\mathrm{alt}}^{(T)} = \max(q_{(2)}^{(T)}, q_0^{(T)})$ when $L \geq 2$, where $q_{(2)}^{(T)}$ is the second-largest benign-mode mass.

$$m_T(\iota_{(1)}) \;\geq\; 2\big(q_{(1)}^{(T)} - q_{\mathrm{alt}}^{(T)}\big),$$

so a strict lead over $q_{\mathrm{alt}}^{(T)}$ suffices for a benign final decode without complete lock-in. If all elites belong to one member ($w = 1$) from round t_0 on, the leader mass converges geometrically to $q^\star = ((1-\eta)\lambda + \eta/k)/(1-a)$ and every other class to $(\eta/k)/(1-a)$. The final decoding margin converges to $2(1-\eta)\lambda/(1-a)$ regardless of the concept count k.

Proof. $\mathbf{E}^{(t)} = \sum_g w_g^{(t)} A_{\iota_g}$ adds $w_g^{(t)}$ to every entry of the (disjoint) support of ι_g and nothing elsewhere; the uniform term adds η/k to every entry. Within-class equality therefore propagates by induction from the uniform start. Subtracting the outside-support recurrence from a benign-mode recurrence gives $q_g^{(t+1)} - q_0^{(t+1)} = a(q_g^{(t)} - q_0^{(t)}) + (1-\eta)\lambda w_g^{(t)} \geq 0$ by induction from the uniform start. For $L \geq 2$ this also gives $q_{(2)}^{(T)} \geq q_0^{(T)}$; for $L = 1$, the alternative mass is $q_0^{(T)}$ by definition. A strict leader over $q_{\mathrm{alt}}^{(T)}$ therefore makes the displayed lower bound positive. Any $\iota \notin \mathcal{A}$ agrees with the leader $\iota_{(1)}$ in at most $k-2$ rows, so at most $k-2$ of its entries lie in the leading support and the remaining at least two have mass at most $q_{\mathrm{alt}}^{(T)}$; hence $\langle \mathbf{Q}^{(T)}, A_\iota \rangle \leq (k-2) q_{(1)}^{(T)} + 2 q_{\mathrm{alt}}^{(T)}$, while

$\langle \mathbf{Q}^{(T)}, A_{\iota_{(1)}} \rangle = k\, q_{(1)}^{(T)}$. The fixed points follow by iterating the displayed affine recursions. At lock, every nonleading entry converges to the same value; swapping the predicates assigned to two rows of the leader therefore yields an off-orbit bijection that differs in exactly those rows, so the limiting lower bound is attained and the margin converges to the stated value. □

D.4 Prospective Continuation

Theorem 3 certifies an observed elite trajectory. The following complementary result starts from a current proposal matrix and lower-bounds the probability that future random proposals keep reinforcing a unique score maximizer until the final decode.

Theorem D.1 (Prospective finite-horizon lock for Gumbel-Parallel CEM)**.** *Let $k \geq 2$ and let $\mathcal{F}_0$ determine a positive row-stochastic $\mathbf{Q}^{(0)}$ and a score over $\mathcal{I}_{\mathrm{bij}}$ with strict maximizer $\iota^\dagger$. Consider $L \geq 1$ future rounds indexed by $j = 0, \ldots, L-1$, with integers $m_j \geq 1$, $r_j = \max\{1, \lceil \rho m_j \rceil\}$, and fresh independent standard Gumbels in Equation (4), independent of $\mathcal{F}_0$. Assume $\tau_g > 0$, $\lambda \in (0, 1]$, and $\eta \in (0, 1)$. Let $a = (1-\eta)(1-\lambda)$, $b = (1-\eta)\lambda$, and define*

$$\bar{\mathbf{Q}}^{(0)} = \mathbf{Q}^{(0)}, \qquad \bar{\mathbf{Q}}^{(j+1)} = a\bar{\mathbf{Q}}^{(j)} + bA_{\iota^\dagger} + \eta \mathbf{U}^{\mathrm{unif}},$$

$$q_j = \prod_{z=1}^{k} \frac{(\bar{\mathbf{Q}}^{(j)}_{z,\iota^\dagger(z)})^{1/\tau_g}}{\sum_{p=1}^{k} (\bar{\mathbf{Q}}^{(j)}_{z,p})^{1/\tau_g}}, \qquad p_j = \Pr[\mathrm{Bin}(m_j, q_j) \geq r_j],$$

and

$$m_{\iota^\dagger}(\mathbf{Q}) = \min_{\iota \neq \iota^\dagger} \langle \mathbf{Q}, A_{\iota^\dagger} - A_\iota \rangle.$$

If

$$a^L m_{\iota^\dagger}(\mathbf{Q}^{(0)}) + \frac{2b(1-a^L)}{1-a} > 0,$$

then, almost surely,

$$\Pr\Big[\mathrm{Dec}(\mathbf{Q}^{(L)}) = \iota^\dagger \mid \mathcal{F}_0\Big] \geq \prod_{j=0}^{L-1} p_j.$$

Proof. For one proposal from a fixed matrix $\mathbf{Q}$, consider the event that every row z ranks $\iota^\dagger(z)$ first under Equation (4). The Gumbel-max identity and rowwise independence give this event probability $q_{\iota^\dagger}(\mathbf{Q}) = \prod_z \big(\mathbf{Q}^{1/\tau_g}_{z,\iota^\dagger(z)} / \sum_p \mathbf{Q}^{1/\tau_g}_{z,p}\big)$. Because $\iota^\dagger$ is bijective, all first-sweep proposals are distinct and the proposal–accept sampler returns $\iota^\dagger$ on this event. Across m_j proposals its occurrence count is therefore $\mathrm{Bin}(m_j, q_j)$. Whenever this count is at least r_j, strict score optimality makes every selected elite a copy of $\iota^\dagger$; score-softmax aggregation then gives $\mathbf{E}^{(j)} = A_{\iota^\dagger}$.

Conditional on mode-pure elites in the preceding rounds, the actual matrix is $\bar{\mathbf{Q}}^{(j)}$. Applying the preceding binomial bound at each step and the conditional chain rule yields $\Pr[\mathbf{E}^{(j)} = A_{\iota^\dagger}\ \forall j < L \mid \mathcal{F}_0] \geq \prod_{j<L} p_j$. Along this event, for every competitor $\iota \neq \iota^\dagger$,

$$\langle \bar{\mathbf{Q}}^{(j+1)}, A_{\iota^\dagger} - A_\iota \rangle = a\langle \bar{\mathbf{Q}}^{(j)}, A_{\iota^\dagger} - A_\iota \rangle + b\, d_H(\iota^\dagger, \iota) \geq a\langle \bar{\mathbf{Q}}^{(j)}, A_{\iota^\dagger} - A_\iota \rangle + 2b,$$

since distinct permutations differ in at least two rows and the uniform term cancels. Iteration gives the displayed margin bound; its positivity makes $\iota^\dagger$ the unique maximum-weight decode. □

The per-entry floor is $e = \eta/k$. For the Decipher certificate below, $e = 0.01$, so numerical clipping is inactive and the theorem applies to the implemented proposal–accept sampler's real-valued law. This bound concerns continuation from the current state, not convergence from a uniform start.

Prospective continuation. The complementary bound in Theorem D.1 asks whether a current proposal state is already concentrated enough to preserve a unique score maximizer through the remaining rounds. For all 20 CIFAR-10 Decipher trajectories, the certificate activates by the start of round 8. From that point, each recorded state lower-bounds the probability of retaining the same interpretation through round 50 by 0.99999, with a certified final decode margin of 1.7260.

E Experimental Setup and Reproducibility

E.1 Evaluation Protocol

We use Success, reasoning accuracy (RA), concept accuracy (CA), and shortcut severity (SSV) as defined in Section 5. For tasks with multiple benign shortcuts, CA and SSV use the best benign alignment before aggregation. On Addition-EvenOdd, we apply identity and $\pi_5(p) = (p+5) \bmod 10$ separately to all predicted concepts in each run. OOD RA is the larger of the two accuracies against the held-out sums, and CA is the larger accuracy against the held-out concept labels. SSV takes the smaller severity, and Success accepts either relabeling under Equation (7). Each candidate relabeling is global across the run's evaluation examples, but the best relabeling is selected separately for each metric before aggregation. The same rule applies to every method, using labels only for post-hoc evaluation. Because shift-by-five preserves the training support rather than all OOD sums, these results measure recovery up to benign equivalence and OOD reasoning after semantic relabeling.

For a decoded Stage II interpretation $\hat{\iota} : [k] \to \mathcal{P}$, let $r : [k] \to \mathcal{P}$ be the evaluation reference assignment and define $\mathcal{A}_r = \{\pi \circ r : \pi \in \Pi_{\text{ben}}\}$. We report

$$\text{Orbit}(\hat{\iota}) = \frac{100}{k} \min_{\iota^\circ \in \mathcal{A}_r} \sum_{z=1}^{k} \mathbf{1}[\hat{\iota}(z) \neq \iota^\circ(z)]. \tag{18}$$

For visual tasks, $r(z)$ is the most frequent ground-truth concept among inputs assigned to cluster z on the scoring data. For Large-K, r is the known hidden cluster permutation. Thus r is an empirical evaluation reference, not the bijection $\iota^\star$ minimizing sequence-level error in Theorem 2. Visual majority assignments can be non-bijective, in which case $\mathcal{A}_r$ contains non-bijective reference maps and Orbit remains their normalized Hamming distance. Such runs remain in the aggregate Orbit statistics, but are excluded from certificates that require a benign bijective reference. Stage II success tests whether Orbit is zero, whereas final success evaluates the downstream predictor after training. Concept labels are used to compute these evaluation references, not for Stage II alignment. For Large-K, downstream RA is relation accuracy and final success accepts every cyclic benign shortcut.

For visual cluster purity, let $C_{z,p}$ count training inputs in cluster z with ground-truth concept p, and let Z_+ contain the nonempty clusters. Purity is the unweighted mean of their majority proportions:

$$\text{Purity} = \frac{100}{|Z_+|} \sum_{z \in Z_+} \frac{\max_{p \in \mathcal{P}} C_{z,p}}{\sum_{q \in \mathcal{P}} C_{z,q}}.$$

Purity is measured before Stage II on training inputs, with equal weight for each nonempty cluster.

Unless stated otherwise, every aggregate uses 20 runs. Success is reported as a count out of 20, and continuous metrics are reported as mean $\pm$ standard deviation. The mechanism comparisons and main effectiveness experiments use independent sets of runs. Within each paired comparison, methods share the Stage I checkpoint, data seed, candidate budget, and downstream training schedule. A run is Stage I-supported when the per-cluster majority labels form a bijection over the task concepts. Diagnostics that require this condition report the eligible trajectory count in their captions.

E.2 Tasks and Data Splits

E.2.1 Visual Tasks

All visual tasks render concepts as MNIST or CIFAR-10 images and use held-out images at test time. Table 7 summarizes the task sizes and supervision.

Addition. Here k denotes the concept vocabulary size, and n denotes the number of digits in each operand; Addition uses the ten digit concepts 0–9 for every n. For digit length $n \in \{1, 2, 3, 4\}$, each instance contains two n-digit image sequences and only their sum is observed. We shuffle the 50,000-image CIFAR-10 training split and consume images sequentially, giving $N_{\text{train}}(n) = \lfloor 50{,}000/(2n) \rfloor$ instances. Each test set contains 500 independently sampled instances from the CIFAR-10 test split.

Decimal Equation Decipher. Each example renders both operands and the result, and the observed label states only whether the decoded equation satisfies base-10 addition. For ABL and BASIL, each equation contains two single-digit operands and their one- or two-digit sum, with a nonzero leading digit for a two-digit result. Each digit is rendered by an independently sampled image of its class, and operator symbols are omitted from the predictor inputs. Each split contains equal numbers of three- and four-digit sequences: 20,000 training and 2,000 test equations on CIFAR-10, and 16,384 in each split on MNIST. All observed validity labels are positive (valid:invalid ratio 1:0).

Table 7: Visual-task construction for ABL and BASIL. Addition lists each digit length n separately; ID and OOD denote the shortcut-stress test splits. External-learner budgets are given below.

Task	Rendering	n	Train	Test	Supervision
Addition ($k = 10$)	CIFAR-10	1	25,000	500	sum of two n-digit operands
	CIFAR-10	2	12,500	500	
	CIFAR-10	3	8,333	500	
	CIFAR-10	4	6,250	500	
Decipher ($k = 10$)	CIFAR-10	—	20,000	2,000	validity of a complete addition equation
	MNIST	—	16,384	16,384	
Addition-Half ($k = 5$)	CIFAR-10	—	5,000	ID: 1,000 OOD: 1,000	sum of two single-digit operands
Addition-EvenOdd ($k = 10$)	CIFAR-10	—	6,720	ID: 960 OOD: 5,040	sum of two single-digit operands

Addition-Half and Addition-EvenOdd. These tasks follow the RSBench shortcut designs (Bortolotti et al., 2024). Addition-Half uses digits $\{0, 1, 2, 3, 4\}$ and training pairs

$$\{(0, 0), (0, 1), (1, 0), (2, 3), (3, 2), (2, 4), (4, 2)\}.$$

The remaining 18 ordered pairs form the OOD split. Addition-EvenOdd uses digits 0–9 with pairs

$$\{(0, 6), (2, 8), (4, 6), (4, 8), (1, 5), (3, 7), (1, 9), (3, 9)\}$$

together with their reverses, giving 16 ordered training pairs and 84 OOD pairs. Its identity and shift-by-five interpretations are both benign on the training support. Each task uses a fixed assignment of CIFAR-10 classes to digit concepts. Within each split, instances are distributed evenly across the specified operand pairs, with held-out images used for ID and OOD testing.

E.2.2 Large-K Relational Decipher

Large-K isolates semantic alignment from representation learning. It generates a symbolic relation whose names are identifiable up to cyclic shifts, corrupts the observed cluster endpoints, and continues with a symbolic table-ABL learner. We evaluate $k \in \{20, 30, 50\}$ at endpoint purity $p = 0.7$.

For each of 20 relation seeds, the generator performs four steps:

1. Draw colors $c(d) \sim \mathrm{Unif}\{0, \ldots, 7\}$ for cyclic differences $d \in \{1, \ldots, k-1\}$ using NumPy PCG64, rejecting colorings with an additional affine automorphism $x \mapsto ax + b$.
2. Set $\mathrm{relation}(i, j) = c((j - i) \bmod k)$ for every ordered off-diagonal pair; diagonal entries use a sentinel.
3. Emit every ordered off-diagonal pair four times, giving $N = 4k(k-1)$ observations.
4. Independently keep each endpoint in its correct cluster with probability p; with probability $1 - p$, assign it to a uniformly sampled incorrect cluster.

The train and test sets share the generated relation but use independent endpoint-noise realizations. The benign set used for evaluation contains exactly the k cyclic shifts.

E.3 Reasoning-Shortcut Structure

We identify the concept transformations π preserving all reasoning results on the training support under perfect strong binding (Definitions 2–3). These properties describe task constraints before cluster errors. For a fixed reference alignment, each benign shortcut induces one benign interpretation.

Addition. Preserving the observed sum for $(0, 0)$ forces $\pi(0) = 0$, and preserving it for $(d, 0)$ then forces $\pi(d) = d$ for every digit d. The same argument applies to multi-digit inputs by setting all other positions to zero, which the input construction permits. Thus identity is the only benign shortcut, and no malignant shortcut exists on the full training support.

Decipher. The support contains every pair of single-digit operands with its valid one- or two-digit result. Preserving $0 + 0 = 0$ gives $\pi(0) = 0$, while the non-carry equations impose $\pi(d + 1) = \pi(d) + \pi(1)$ for $d = 0, \ldots, 8$. Thus $\pi(d) = d\pi(1)$, and the digit range restricts $\pi(1)$ to 0 or 1. The zero choice collapses all digits to zero, but transforms $1 + 9 = 10$ into $0 + 0 = 00$, which the KB rejects because a two-digit result cannot start with zero. Hence identity is the only reasoning shortcut under perfect strong binding. With imperfect clusters, the all-zero assignment continues to satisfy every three-digit equation regardless of the cluster errors. Such partial consistency can favor collapsed assignments during empirical alignment, as seen in the non-bijective control (Table 2).

Addition-Half. The seven training pairs force $\pi(0) = 0$, $\pi(1) = 1$, $\pi(2) + \pi(3) = 5$, and $\pi(2) + \pi(4) = 6$. Writing $a = \pi(2)$ gives $(\pi(2), \pi(3), \pi(4)) = (a, 5 - a, 6 - a)$ with $a \in \{2, 3, 4\}$. Only $a = 2$ is bijective, so identity is the unique benign shortcut. With $a = 3$, the malignant map $(\pi(0), \ldots, \pi(4)) = (0, 1, 3, 2, 3)$ merges digits 2 and 4 but preserves every training sum.

Addition-EvenOdd. The sixteen training pairs constrain π to the form

$$\pi(d) = \begin{cases} d + a, & d \in \{0, 2, 4\}, \\ d - a, & d \in \{6, 8\}, \\ d + b, & d \in \{1, 3\}, \\ d - b, & d \in \{5, 7, 9\}, \end{cases} \qquad a, b \in \{0, \ldots, 5\}.$$

A bijection must satisfy $\sum_{d=0}^{9} \pi(d) = 45$. Summing the four cases gives $(6 + 3a) + (14 - 2a) + (4 + 2b) + (21 - 3b) = 45 + a - b$, which forces $a = b$. Its image must also contain zero, leaving only $a = b = 0$ and $a = b = 5$. Both are bijections, yielding exactly identity and $\pi_5(d) = (d+5) \bmod 10$. In contrast, $a = b = 1$ gives the malignant map $(1, 2, 3, 4, 5, 4, 5, 6, 7, 8)$. These transformations preserve the training pairs' sums, but need not preserve sums on the OOD support.

Large-K. Every cyclic shift $\pi_b(i) = (i + b) \bmod k$ preserves the difference $(j - i) \bmod k$ and hence every relation label. The exact certificates in Appendix D.1 establish that these are the only benign shortcuts for the evaluated relations. Moreover, a non-injective map merges some distinct concepts i and j. Their ordered pair occurs in the training support, but its image lies on the diagonal, whose sentinel differs from every observed relation label. Thus no malignant shortcut can preserve the Large-K training support under perfect strong binding.

E.4 Models and Downstream Learners

CIFAR-10 uses ResNet-18 (He et al., 2016) with a 3×3, stride-1 input convolution and no initial max pooling. MNIST uses a LeNet-5-style CNN (LeCun et al., 1998). Both output k logits.

We evaluate ABL, A^{3}BL, and DSL with their standard learner-specific training procedures. For learner L, L+PT starts from the Stage I pretrained predictor without semantic alignment, and L+SA applies Stage II to a randomly initialized predictor. L+BASIL-init uses Stage I followed by Stage II before the standard training of L; BASIL denotes the ABL instance of this initialization. Paired +PT and +BASIL-init variants share the same Stage I checkpoint. For transfer, the aligned predictor initializes A^{3}BL or DSL before its standard training procedure.

E.5 Downstream Training, Compute, and Reproducibility

The predictor handoff and the paired non-bijective control are specified in Appendix C.2. Visual ABL uses Adam with learning rate 3×10^{-4}, betas $(0.9, 0.999)$, and mini-batch size 256. It runs 20 downstream loops with segment size 0.1 and one neural epoch per loop, using cross-entropy loss with label smoothing 0.5. A^{3}BL uses Adam with learning rate 10^{-3} and betas $(0.9, 0.99)$, one neural epoch per update with mini-batch size 256, and reasoning batches of 1024. Its top-K candidate count is 32 and its aggregation temperature is 0.2. We train A^{3}BL for 40 epochs on Addition and 20 on Decipher, Addition-Half, and Addition-EvenOdd. For the two shortcut-stress tasks, A^{3}BL uses 10,000 training and 500 test instances, with evaluation after the final epoch. DSL runs for 10 epochs on Addition and CIFAR-10 Decipher and 20 on MNIST Decipher and the two shortcut-stress tasks. DSL uses Adam with learning rate 10^{-3} and batch size 128. For both shortcut-stress tasks, DSL uses 10,000 training, 500 validation, and 500 test instances.

For A^{3}BL and DSL, Addition and Decipher use sequentially drawn operand images. Decipher additionally samples result images from the matching digit classes. The number of training examples is limited by the available operand images, and DSL reserves 500 examples for validation. On

CIFAR-10 Decipher, A^{3}BL uses 25,000 training and 500 test instances, while DSL uses 24,500 training, 500 validation, and 500 test instances. All Decipher equations are valid. Within each learner, the +PT and +BASIL-init comparison keeps these budgets and schedules fixed. Large-K instead uses 20 hard-abduction loops over ten fixed segments with independent train and test endpoint noise.

The continuation timing runs use 16 logical CPUs on an Intel Xeon Gold 6430 host and four PyTorch/BLAS threads, with PyTorch 2.4.1; visual runs additionally use one NVIDIA RTX A6000, while Large-K runs on the CPU. Paired ABL+PT and BASIL continuations reuse the same 1,000-epoch Stage I checkpoint, with the pretraining cost counted for both methods. Stage I timings span the first epoch report through the final checkpoint in the original pretraining runs. Stage II includes cluster inference and alignment search. Handoff + ABL includes the five-epoch pseudo-target fitting when used, downstream training, and evaluation; Total additionally includes data, model, and KB setup. All continuations retain the schedules above; Appendix F.3.2 reports their runtimes.

F Additional Experimental Results

This appendix follows the experimental analyses in Section 5. Appendix F.1 gives additional effectiveness and transfer results for Q1, and Appendix F.2 analyzes Stage II search and downstream handoff for Q2. Appendix F.3 presents scaling and computational cost results, Appendix F.4 examines Stage I cluster purity, and Appendix F.5 details hyperparameter sensitivity. Unless stated otherwise, Success is a count out of 20 and continuous metrics are mean ± standard deviation over 20 runs.

F.1 Additional Effectiveness and Transfer Results

This subsection extends Figure 2 across renderings, sequence lengths, learners, and shortcut structures to test whether BASIL-init remains useful across these settings.

F.1.1 Addition Across Digit Lengths

Figure 7 compares Addition digit lengths $n \in \{1, 2, 3, 4\}$; Table 8 gives the complete metrics.

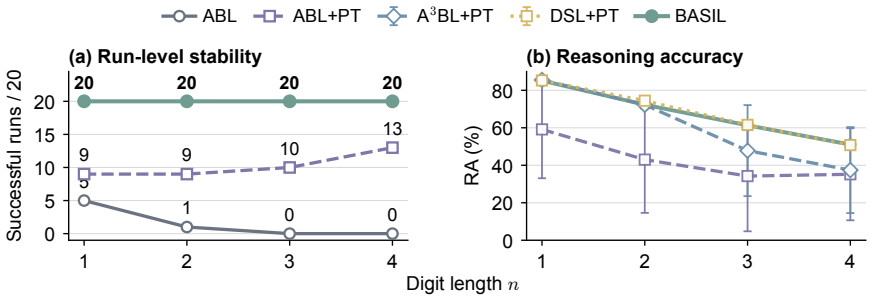

Figure 7: Analysis of CIFAR-10 Addition across digit lengths after downstream training: (a) success counts; (b) mean RA with standard-deviation bars. Each setting uses 20 runs.

BASIL succeeds in all 20 runs at every digit length, with concept accuracy remaining near 92%. Reasoning accuracy decreases as the sequences grow longer, even though semantic recovery remains reliable. Longer instances require predictions for more constituent inputs, so residual concept errors can affect more reasoning outcomes even when concept accuracy remains stable. This distinction separates errors in predicting a full sequence from failures of semantic alignment.

Table 8: Complete Addition results across digit lengths on CIFAR-10, including the pretrained learners in Figure 7. Continuous metrics are percentages (mean ± standard deviation over 20 runs).

n	Method	Success ↑	RA ↑	CA ↑	SSV ↓
1	ABL	5/20	26.2 ± 17.7	33.9 ± 26.7	53.2 ± 22.5
	ABL+PT	9/20	59.1 ± 26.1	43.5 ± 44.9	58.1 ± 43.3
	BASIL	20/20	85.2 ± 1.6	92.3 ± 0.9	11.5 ± 3.1
	A^3BL+PT	20/20	85.5 ± 1.1	92.4 ± 0.6	2.0 ± 0.3
	DSL+PT	20/20	85.2 ± 1.2	92.0 ± 0.6	12.9 ± 3.4
2	ABL	1/20	1.6 ± 2.2	20.7 ± 10.8	59.3 ± 13.1
	ABL+PT	9/20	43.0 ± 28.4	43.3 ± 44.9	57.0 ± 41.8
	BASIL	20/20	72.3 ± 1.6	92.2 ± 0.4	10.9 ± 1.8
	A^3BL+PT	20/20	72.3 ± 2.1	92.2 ± 0.6	2.0 ± 0.3
	DSL+PT	20/20	74.5 ± 1.3	92.7 ± 0.4	10.2 ± 1.3
3	ABL	0/20	0.2 ± 0.4	16.9 ± 8.5	63.7 ± 8.1
	ABL+PT	10/20	34.2 ± 29.5	47.9 ± 44.8	50.6 ± 40.6
	BASIL	20/20	61.3 ± 2.2	92.2 ± 0.5	11.0 ± 1.7
	A^3BL+PT	15/20	47.8 ± 24.3	70.4 ± 38.1	23.6 ± 36.3
	DSL+PT	20/20	61.6 ± 1.2	92.3 ± 0.2	10.9 ± 0.9
4	ABL	0/20	0.0 ± 0.0	12.8 ± 4.5	63.4 ± 10.0
	ABL+PT	13/20	35.2 ± 24.6	61.2 ± 42.6	38.6 ± 38.0
	BASIL	20/20	51.1 ± 2.2	92.0 ± 0.4	10.8 ± 1.5
	A^3BL+PT	14/20	37.4 ± 22.9	65.1 ± 41.3	28.7 ± 38.1
	DSL+PT	20/20	50.8 ± 1.7	91.9 ± 0.2	11.2 ± 1.0

F.1.2 Decipher Across Renderings and Learners

Panel A of Table 9 compares ABL, Stage I pretraining, and BASIL across MNIST and CIFAR-10 renderings. Panel B reports the paired +PT and +BASIL-init comparison for three downstream learners on CIFAR-10. The two panels vary rendering and downstream learner separately.

Table 9: Decipher results across renderings and downstream learners. Panel A compares ABL variants on MNIST and CIFAR-10; Panel B compares paired initializations on CIFAR-10.

Panel A: Rendering comparison

Rendering	Method	Success ↑	RA ↑	CA ↑	SSV ↓
MNIST	ABL	0/20	37.5 ± 8.4	39.8 ± 12.4	99.1 ± 0.5
MNIST	ABL+PT	1/20	48.2 ± 13.5	48.9 ± 17.3	94.4 ± 21.5
MNIST	**BASIL**	20/20	97.4 ± 0.2	99.2 ± 0.1	0.8 ± 0.2
CIFAR-10	ABL	0/20	18.0 ± 13.6	28.4 ± 8.6	78.4 ± 12.5
CIFAR-10	ABL+PT	0/20	29.9 ± 10.2	33.8 ± 11.6	95.6 ± 2.7
CIFAR-10	**BASIL**	20/20	75.4 ± 0.6	92.3 ± 0.2	11.1 ± 0.9

Panel B: Cross-learner transfer on CIFAR-10

Learner	Initialization	Success ↑	RA ↑	CA ↑	SSV ↓
ABL	+PT	0/20	29.9 ± 10.2	33.8 ± 11.6	95.6 ± 2.7
ABL	+BASIL-init	20/20	75.4 ± 0.6	92.3 ± 0.2	11.1 ± 0.9
A^3BL	+PT	6/20	49.7 ± 18.8	57.2 ± 23.5	63.5 ± 39.9
A^3BL	+BASIL-init	20/20	77.6 ± 1.8	92.7 ± 0.7	2.3 ± 0.1
DSL	+PT	0/20	15.9 ± 1.6	39.0 ± 0.4	93.1 ± 2.0
DSL	+BASIL-init	20/20	86.4 ± 0.9	93.2 ± 0.5	10.5 ± 2.2

F.1.3 Shortcut-Stress Tasks

Addition-Half tests a malignant many-to-one shortcut, while Addition-EvenOdd tests two benign interpretations. Tables 10 and 11 report the OOD results behind Figure 2.

Table 10: Addition-Half results on the OOD operand pairs.

Learner	Initialization	Success ↑	RA ↑	CA ↑	SSV ↓
ABL	—	0/20	17.9 ± 6.5	40.3 ± 12.2	72.3 ± 17.5
ABL	+PT	1/20	14.8 ± 18.1	35.2 ± 21.2	92.2 ± 19.5
ABL	+SA	0/20	9.3 ± 5.9	20.0 ± 12.7	84.6 ± 13.3
ABL	+BASIL-init	20/20	83.9 ± 1.5	91.6 ± 0.8	8.6 ± 1.6
A^3BL	—	0/20	12.4 ± 2.2	40.9 ± 2.5	82.7 ± 5.8
A^3BL	+PT	2/20	17.9 ± 23.6	46.8 ± 16.2	85.9 ± 26.8
A^3BL	+BASIL-init	20/20	85.9 ± 2.2	92.7 ± 1.2	7.2 ± 1.9
DSL	—	0/20	12.0 ± 1.8	41.2 ± 2.0	84.1 ± 5.2
DSL	+PT	3/20	19.3 ± 28.0	48.0 ± 18.8	82.3 ± 31.5
DSL	+BASIL-init	19/20	83.9 ± 17.2	91.0 ± 11.4	10.8 ± 19.9

Table 11: Addition-EvenOdd OOD results after alignment to identity or shift-by-five.

Learner	Initialization	Success ↑	RA ↑	CA ↑	SSV ↓
ABL	—	0/20	2.4 ± 1.2	11.0 ± 4.8	90.4 ± 7.0
ABL	+PT	0/20	5.5 ± 3.9	10.4 ± 11.5	98.5 ± 1.3
ABL	+SA	0/20	1.3 ± 0.7	5.5 ± 3.1	96.2 ± 3.7
ABL	+BASIL-init	20/20	76.9 ± 2.1	87.7 ± 1.2	15.5 ± 5.6
A^3BL	—	0/20	4.1 ± 1.6	9.6 ± 6.6	90.9 ± 5.4
A^3BL	+PT	0/20	7.8 ± 5.2	1.2 ± 0.4	99.1 ± 1.1
A^3BL	+BASIL-init	20/20	80.0 ± 1.6	89.4 ± 0.8	14.1 ± 3.9
DSL	—	0/20	3.0 ± 1.4	11.8 ± 6.7	92.9 ± 4.0
DSL	+PT	0/20	7.4 ± 4.0	1.1 ± 0.3	99.2 ± 0.9
DSL	+BASIL-init	20/20	80.6 ± 2.0	89.6 ± 1.2	13.7 ± 2.6

On Addition-EvenOdd, BASIL and both transferred BASIL-init variants succeed in 10/20 runs under identity-only evaluation. The remaining runs select the other benign interpretation.

F.2 Additional Mechanism Analysis

This subsection examines optimizer handoff and the role of bijectivity during search.

F.2.1 Stage II Output and Downstream Handoff

Final training outcomes can obscure how much semantic recovery the initialization itself provides. Figure 8 evaluates both stages on an independent paired CIFAR-10 Decipher cohort.

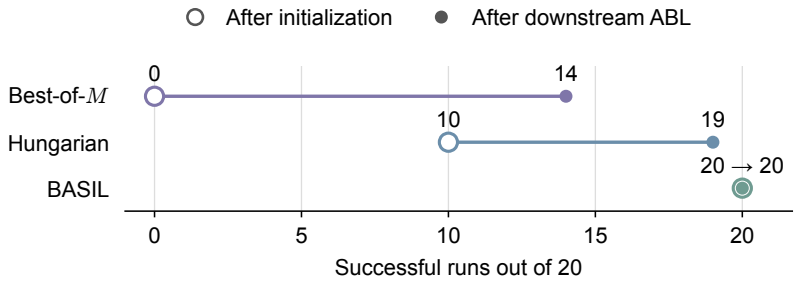

Figure 8: Semantic recovery on 20 paired CIFAR-10 Decipher runs. Hollow and filled markers report aggregate success counts after initialization and downstream training, respectively; concentric markers indicate coincident counts. Methods share Stage I checkpoints, scoring data, a 50,000-candidate budget, and the downstream schedule. Stage II and initialization counts coincide.

Best-of-M and Hungarian reach 14/20 and 19/20 final successes, despite providing correct initial interpretations in 0/20 and 10/20 runs. BASIL achieves 20/20 at both evaluations, aligning semantics before downstream training. The final counts thus understate BASIL's advantage at initialization.

F.2.2 Where Bijectivity Matters

The following controls examine why a high training alignment score can conceal incorrect semantics and whether a final bijective projection can replace bijectivity during search.

Addition-Half output space. Figure 9 compares Stage II consistency and OOD reasoning in 20 paired CIFAR-10 Addition-Half runs; Table 12 gives the full metrics for both stages.

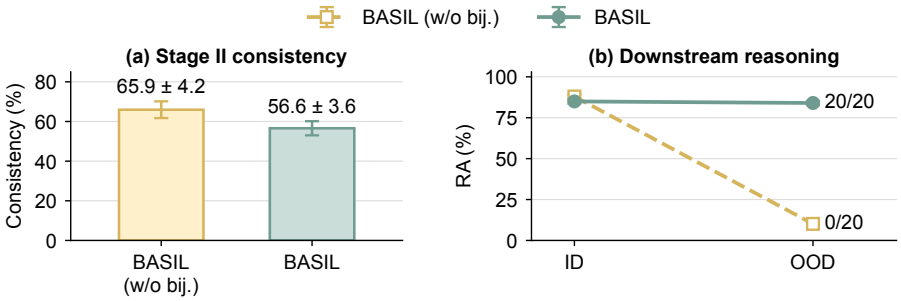

Figure 9: Higher training consistency need not imply correct semantic alignment. On 20 paired CIFAR-10 Addition-Half runs, (a) shows Stage II consistency and (b) downstream RA on ID and OOD operand pairs, with OOD success counts. Means are plotted with standard-deviation bars.

BASIL (w/o bij.) uses four of the five predicates on average and attains higher training consistency, as well as slightly higher downstream ID accuracy. Its OOD accuracy falls to 10.1%, whereas BASIL retains 84.0% and succeeds in all 20 runs. Concept merging can therefore fit the observed operand pairs while losing the semantic distinctions needed on unseen combinations.

Table 12: Complete Addition-Half output-space control results for Figure 9. Hamming distance is measured from the Stage I majority interpretation. Continuous metrics are mean ± standard deviation over 20 paired runs; predicate usage is averaged, and output and success counts are out of 20.

Evaluation stage	Metric	BASIL (w/o bij.)	**BASIL**
Stage II	Training consistency (%)	65.9 ± 4.2	56.6 ± 3.6
	Hamming distance	2.9 ± 0.4	0.1 ± 0.4
	Mean predicates used / unused	4.0 / 1.0	5.0 / 0.0
	Bijective outputs	0/20	20/20
Downstream ABL (ID)	Successful runs	0/20	20/20
	RA (%)	88.0 ± 1.3	85.0 ± 1.3
	CA (%)	44.7 ± 0.5	92.0 ± 0.6
	SSV (%)	97.8 ± 0.8	8.3 ± 1.8
Downstream ABL (OOD)	Successful runs	0/20	20/20
	RA (%)	10.1 ± 0.7	84.0 ± 1.3
	CA (%)	42.4 ± 0.4	91.7 ± 0.7
	SSV (%)	97.7 ± 0.8	8.5 ± 1.7

Search and final projection. Figure 10 separates sampling from final decoding.

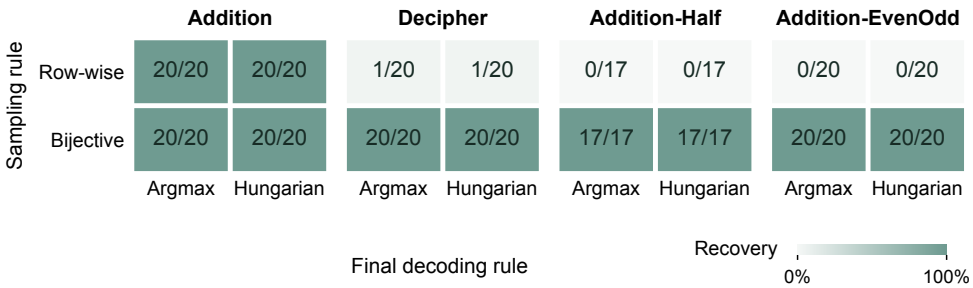

Figure 10: Stage II benign-interpretation recovery before ABL. Cells give s/N, with color indicating the recovery proportion. Decoders share the final $\mathbf{Q}$ within each sampling rule; argmax is row-wise. Addition-Half includes 17 runs with bijective Stage I majority assignments; other tasks use 20.

Bijective sampling recovers a benign interpretation on every eligible trajectory across the four tasks. For row-wise sampling, Hungarian decoding gives the same recovery counts as row-wise argmax. In these runs, a bijective final decode does not recover distinctions lost during search.

F.3 Scaling and Computational Cost

We supplement Figure 3 with full scaling results and computational costs for BASIL and its controls.

F.3.1 Large-K Scaling

Table 13 gives Stage II recovery and downstream ABL metrics at $p = 0.7$ for Figure 3.

Table 13: Large-K recovery at endpoint purity $p = 0.7$. Stage II recovery is evaluated before training; Success, RA, CA, and SSV are measured after downstream symbolic ABL. Continuous metrics are percentages (mean $\pm$ standard deviation over 20 runs).

k	Method	Stage II recovery	After downstream ABL			
			Success ↑	RA ↑	CA ↑	SSV ↓
20	Best-of-M	0/20	0/20	30.9 ± 9.4	37.2 ± 15.1	75.8 ± 2.4
	Hungarian	0/20	0/20	20.1 ± 4.6	20.0 ± 9.4	76.7 ± 1.6
	BASIL (w/o bij.)	20/20	20/20	55.6 ± 1.6	70.0 ± 0.8	5.1 ± 0.6
	BASIL	20/20	20/20	55.6 ± 1.6	70.0 ± 0.8	5.1 ± 0.6
30	Best-of-M	0/20	0/20	21.4 ± 6.1	22.8 ± 12.0	75.9 ± 1.2
	Hungarian	0/20	0/20	15.7 ± 1.7	10.7 ± 2.0	76.1 ± 1.1
	BASIL (w/o bij.)	20/20	20/20	55.6 ± 1.1	70.0 ± 0.5	3.6 ± 0.4
	BASIL	20/20	20/20	55.6 ± 1.1	70.0 ± 0.5	3.6 ± 0.4
50	Best-of-M	0/20	0/20	15.2 ± 0.8	8.3 ± 2.0	75.3 ± 1.0
	Hungarian	0/20	0/20	13.9 ± 1.0	6.4 ± 1.4	75.4 ± 0.9
	BASIL (w/o bij.)	18/20	19/20	55.5 ± 0.9	70.0 ± 0.4	5.8 ± 14.6
	BASIL	20/20	20/20	55.5 ± 0.8	70.1 ± 0.3	2.4 ± 0.2

Appendix E.3 explains why Large-K has k benign shortcuts and no malignant shortcuts under perfect strong binding. The difficulty also grows through weaker score separation: Table 6 shows that the mean ideal consistency gap narrows from 14.1% at $k = 20$ to 6.2% at $k = 50$, while the benign orbit remains strictly optimal on every sampled scoring set. BASIL's stable recovery therefore reflects effective search over a larger space with finer score distinctions.

F.3.2 Computational Cost

Table 14 compares the computational costs of ABL, ABL+PT, and BASIL on CIFAR-10 Addition-Half and Decipher, and reports the complete symbolic Large-K pipeline at $k = 50$. Hardware, training schedules, and timing definitions are given in Appendix E.5.

Table 14: Representative runtime costs in minutes. Stage I reports medians over 20 complete pretraining runs per task; other stages report medians over three timing runs per method. Totals combine the pretraining median with the median continuation time, including setup.

Task	Method	Stage I	Stage II	Handoff + ABL	Total
Addition-Half	ABL	—	—	0.9	1.3
	ABL+PT	94.2	—	1.0	95.4
	BASIL	94.2	2.4	1.1	98.2
Decipher	ABL	—	—	6.4	6.7
	ABL+PT	169.4	—	6.4	176.1
	BASIL	169.4	4.4	7.5	181.9
Large-K	BASIL	—	2.2	0.03	2.2

For the visual tasks, Stage I pretraining dominates the total runtime and is also required by ABL+PT. Stage II takes 2.4 minutes on Addition-Half and 4.4 minutes on Decipher. The complete symbolic Large-K pipeline, including downstream training, takes 2.2 minutes at $k = 50$.

F.4 Stage I Cluster Purity

Table 15 supplements Figure 4 with Addition and Decipher results. BASIL succeeds in all 80 Decipher runs across the four pretraining settings. Purity describes the Stage I partition on training inputs, whereas CA evaluates test predictions after downstream training; initial purity therefore imposes no fixed ceiling on final accuracy. We vary Stage I cluster quality through a 20–300 epoch sweep, separate from the default 1000-epoch pretraining schedule.

Table 15: Stage I cluster purity and downstream metrics across pretraining epochs on CIFAR-10. Success is a count out of 20; other entries are percentages (mean ± standard deviation).

PT epochs	Purity	Success ↑	RA ↑	CA ↑	SSV ↓
Addition ($n = 4$)					
20	53.0 ± 1.2	20/20	20.2 ± 1.9	81.6 ± 0.8	17.6 ± 4.2
50	59.1 ± 0.6	20/20	28.5 ± 2.1	85.4 ± 0.7	15.0 ± 3.6
100	63.2 ± 0.5	20/20	34.6 ± 1.4	87.7 ± 0.5	14.0 ± 3.0
300	67.9 ± 1.6	20/20	47.5 ± 2.0	91.1 ± 0.3	11.5 ± 1.3
Decipher					
20	51.8 ± 0.9	20/20	52.6 ± 1.2	83.0 ± 0.6	16.0 ± 3.2
50	58.0 ± 0.5	20/20	59.3 ± 1.3	86.1 ± 0.5	14.4 ± 2.2
100	62.4 ± 0.6	20/20	64.8 ± 1.0	88.3 ± 0.4	12.3 ± 1.7
300	67.7 ± 1.6	20/20	72.4 ± 0.8	91.2 ± 0.2	11.4 ± 1.6

F.5 Hyperparameter Sensitivity

Figure 5 shows one-parameter scans on CIFAR-10 Decipher, Addition-EvenOdd, and Large-K ($k = 50$, $p = 0.7$), using $M = 50{,}000$ candidates over $T = 50$ rounds. The defaults are $(\tau_g, \rho, \lambda, \eta) = (1.0, 0.10, 0.70, 0.10)$, with score-softmax elite aggregation and matching-based final decoding. Each setting uses 20 paired runs sharing Stage I checkpoints or symbolic cluster realizations, scoring data, and downstream schedules. Default runs are reused across the four sweeps. Success after downstream ABL accepts either benign mode on Addition-EvenOdd's OOD set and any cyclic shift on Large-K.

Within the tested neighborhood of the defaults, 54 of the 60 task–parameter settings have no failed runs. Decipher succeeds throughout, while Large-K has 16/20 successes at $\tau_g = 1.1$.

Effect of Gumbel temperature. For a fixed proposal matrix $\mathbf{Q}$, let $P_{\tau_g}(p \mid z, \mathbf{Q})$ denote the probability that cluster z first proposes to predicate p, before acceptance resolves collisions. For $\eta > 0$, uniform initialization and the mixing step in Equation (5) bound each entry of $\mathbf{Q}$; Equation (4) then bounds the first-proposal odds between any two predicates within a cluster:

$$\frac{\eta}{k} \le \mathbf{Q}_{z,p} \le 1 - \eta + \frac{\eta}{k},$$

$$\frac{P_{\tau_g}(p \mid z, \mathbf{Q})}{P_{\tau_g}(p' \mid z, \mathbf{Q})} = \left(\frac{\mathbf{Q}_{z,p}}{\mathbf{Q}_{z,p'}}\right)^{1/\tau_g} \le C_{\mathrm{mix}}(\tau_g) := \left(\frac{1 - \eta + \eta/k}{\eta/k}\right)^{1/\tau_g}.$$

For Large-K ($k = 50$, $\eta = 0.1$), increasing τ_g from 1.0 to 1.1 lowers this ceiling from 451 to approximately 259, a 42.6% reduction (Figure 11).

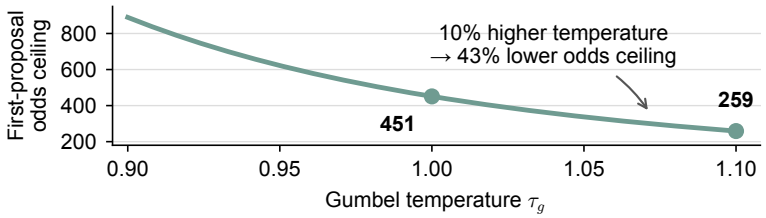

Figure 11: Analytic first-proposal odds ceiling ($k = 50$, $\eta = 0.1$). This bounds a probability ratio, not measured concentration or complete-matching probability.

In the single-mode limit of Proposition D.2, the additional setting $\lambda = 0.7$ gives a weight ratio $((1-\eta)\lambda + \eta/k)/(\eta/k) = 316$. Its corresponding first-proposal odds fall from 316 to approximately 187.26, a 40.7% reduction for the same temperature change.