Exploring the Rashomon Set for Concept-Based Models

Anonymous authors
Paper under double-blind review

Abstract

In many machine learning problems, there may exist multiple models that achieve nearly identical predictive performance while relying on fundamentally different internal logic. However, standard training procedures produce a single model, offering no practical way to explore alternatives that may better suit downstream needs. The set of these equally accurate models is known as the Rashomon set. Exploring the Rashomon set is particularly challenging in large and complex hypothesis spaces, such as Concept Bottleneck Models (CBMs), which are widely used in computer vision to make predictions through intermediate, human-understandable concepts. In this paper, we provide a method for efficiently exploring the Rashomon set of CBMs. Our framework introduces a specialized parallel adapter-based construction, combined with a checkpointing scheme and a concept diversity objective, to generate multiple equally accurate CBMs from a single training process. Empirical results show that our method finds models with better diversity than baselines while using much less memory. We further demonstrate that access to these diverse yet accurate CBMs enables trustworthy model selection, resolution of inter-class confusion, and reliable abstention in decision-making.

1 Introduction

Interpretability is important for deploying machine learning models in high-stakes domains, as it can help practitioners to troubleshoot errors, audit for biases, and align model behavior with domain expertise. For image classification, concept bottleneck models (CBMs) are designed to be interpretable by first predicting human-interpretable concepts and then using those concepts for the final classification. However, current CBM frameworks operate under a "unique truth" paradigm: they are optimized to produce a single model. This paradigm is fundamentally limited; if the resulting model relies on a flawed rationale or a spurious correlation, practitioners have no immediate alternative "good" models to turn to.

Indeed, in many machine learning problems, there could be different models that achieve similar predictive accuracy yet differ substantially in reasoning patterns, known as *the Rashomon set* (Semenova et al., 2022; Rudin et al., 2024; Xin et al., 2022). For example, a bird species may be identified with equal accuracy by its wing pattern or its beak shape. This diversity is inherent in many real-world datasets and in human decision-making, and the Rashomon set makes it explicit (but the unique truth paradigm ignores it). Essentially, the Rashomon set provides a "menu" of good options, from which experts can select models, compare them, or audit.

However, finding the Rashomon set in neural networks is notoriously challenging. While Rashomon sets have been successfully constructed for simpler hypothesis spaces like sparse decision trees (Xin et al., 2022) or generalized additive models (Zhong et al., 2023), these techniques do not scale to neural networks. The continuous, high-dimensional parameter spaces of deep models often admit infinitely many near-optimal solutions that are numerically distinct but *semantically redundant*, meaning they differ in weights but rely on very similar underlying reasoning. Methods such as varying random seeds, applying dropout at inference (Hsu et al., 2024), or perturbing the final linear layer (Donnelly et al., 2025) usually induce diversity concentrated in the output layer or local variation rather than genuine functional divergence.

In this paper, we propose a novel method to explore the Rashomon set of the CBMs by constructing and analyzing a Rashomon slice, which provides a representative subset of models with diverse conceptual

rationales. We achieve this through a specialized architecture and a customized objective function. Specifically, we combine small, parallel adapter modules with a shared and frozen backbone, enabling the simultaneous learning of multiple models within a single training pipeline. We then apply a customized objective function with a diversity-promoting regularization on the concept representations to force models toward distinct rationales. To make joint training scalable, we introduce a model-axis gradient checkpointing scheme that serializes the backward pass across models, reducing memory usage to nearly that of a single model. Together, these components allow us to efficiently find a meaningful set of diverse yet accurate models with minimal additional training cost.

We evaluate the Rashomon slice quantitatively and qualitatively on multiple datasets (AwA2, CUB, CIFAR-10, CelebA, HAM10000). Our results show that the slice has rich conceptual diversity without sacrificing accuracy. We demonstrate the practical utility of our method in multiple settings:

1. Expert Alignment: Selecting models that prioritize domain-relevant concepts or satisfy trustworthiness criteria (such as fairness) without retraining.
2. Resolving Inter-Class Confusion: By forcing models toward discriminative rationales, we reduce the accuracy gap in visually similar species (e.g., Indigo Bunting vs. Blue Grosbeak).
3. Reliable Abstention: We use the disagreement across models as a safety signal to identify ambiguous samples where a single model might otherwise fail.

Contributions: (1) we introduce the first method for exploring the Rashomon set of CBMs; (2) on multiple datasets, our method discovers accurate yet different models, outperforming baselines in efficiency and diversity; (3) we demonstrate three practical use cases, showing that diversity across equally accurate models can be directly leveraged to improve downstream tasks.

2 Related Work

Rashomon Sets. The Rashomon Effect describes that many different models from a function class can explain a dataset almost equally well (Breiman, 2001), and the Rashomon set is the collection of these good models (Semenova et al., 2022; Fisher et al., 2019; Rudin et al., 2024). Existing work in this area can be broadly categorized into efforts to compute and characterize the Rashomon Effect and the Rashomon set (Xin et al., 2022; Zhong et al., 2023; Hsu & Calmon, 2022; Dong & Rudin, 2020) and efforts that study the implications of the Rashomon set for different trustworthy applications (Semenova et al., 2023; Hsu et al.; Boner et al., 2024; Ganesh et al., 2025; Black et al., 2022; Marx et al., 2020; Watson-Daniels et al., 2023; Turbal et al., 2025; Hsu et al.; 2026).

For methods exploring Rashomon sets of neural networks, some vary random seeds during training (D'Amour et al., 2020), use affine transformations to intermediate network activations (Eerlings et al., 2026), apply dropout at inference (Hsu et al., 2024), or perturb the final layer (Donnelly et al., 2025). Although these approaches produce models with different weights, injected randomness often yields solutions that rely on similar underlying representations, and restricting diversity to the linear layer probes only a small part of the model space. As a result, these methods offer limited insight into alternative reasoning pathways supported by the data, which the Rashomon slice aims to fix.

Concept Bottleneck Models (CBMs). CBMs (Koh et al., 2020) have a special architecture which first predicts concepts from the input and then uses these concepts to predict the label. This two-stage architecture enables users to inspect how concept information is represented and used, and supports human intervention to modify model behavior. Beyond the standard formulation, recent work explores concept representations by learning richer concept embeddings (Espinosa Zarlenga et al., 2022), modeling concept uncertainty (Kim et al., 2023), capturing dependencies between concepts (Vandenhirtz et al., 2024), and disentangling concepts in latent space (Chen et al., 2020). Other extensions formulate inference using energy-based objectives to support concept correction (Xu et al., 2024; Kim et al., 2024) or utilize internal ensemble-based gating to refine concept accuracy (Ahn et al., 2026). However, these methods focus on producing a single CBM without considering the Rashomon set.

Deep Ensembles. Deep ensembles train multiple networks and average their predictions to improve accuracy, calibration, and out-of-distribution behavior (Lakshminarayanan et al., 2017; Stickland & Murray, 2020; Fort et al., 2019; Yang et al., 2020; Zhou et al., 2018; Yang et al., 2021). Among many studies in this field, more closely related to our setting are methods that encourage diversity among ensemble members (Shui et al., 2018; Zhang et al., 2020; Dereka et al., 2024). However, the goal of deep ensembles is **fundamentally different** from ours. Deep ensembles aggregate multiple models to produce a single improved predictor, and the individual members are not required to be strong or meaningfully different on their own. In contrast, our goal is to find a *diverse set of individually accurate models*, where each model achieves high performance but relies on distinct internal reasoning. Each model stands on its own, rather than being combined with others.

3 Methods

3.1 Preliminaries and Definitions

Concept Bottleneck Models (CBMs). Let $\mathcal{D} = \{\mathbf{x}_i, y_i, \mathbf{c}_i\}_{i=1}^n$ be a dataset, where $\mathbf{x}_i \in \mathbb{R}^{s_1 \times s_2 \times s_3}$ is an input image, $y_i \in \{1, \ldots, K\}$ is the class label with K possible classes, and $\mathbf{c}_i \in \mathbb{R}^p$ is a vector of p human-interpretable concepts that describe input $\mathbf{x}_i$. A typical CBM consists of a concept encoder $g(\cdot) : \mathbb{R}^{s_1 \times s_2 \times s_3} \to \mathbb{R}^p$ that maps input data $\mathbf{x}$ to a latent concept representation $\hat{\mathbf{c}} = g(\mathbf{x})$, and a classifier $f(\cdot) : \mathbb{R}^p \to \mathbb{R}$ that maps these concepts to output predictions $\hat{y} = f(\hat{\mathbf{c}})$. The empirical risk of a CBM over a dataset $\mathcal{D}$ is defined as the weighted sum of task and concept prediction losses:

$$\mathcal{L}_{\text{total}}(f, g; \mathcal{D}) = \frac{1}{n} \sum_{(\mathbf{x}, y, \mathbf{c}) \in \mathcal{D}} \mathcal{L}_Y(f(g(\mathbf{x})), y) + \lambda \mathcal{L}_C(g(\mathbf{x}), \mathbf{c}),$$

where $\mathcal{L}_Y$ and $\mathcal{L}_C$ denote the task and concept prediction losses, respectively (such as cross-entropy loss or misclassification error), and λ is a hyperparameter that controls the trade-off between the two losses.

Rashomon Set. Following Semenova et al. (2022), we define the Rashomon set of CBMs as follows:

Definition 1 (ε-Rashomon set of CBMs)**.** Given a validation dataset $\mathcal{D}_{val}$, hypothesis spaces $\mathcal{F}$ and $\mathcal{G}$, and a reference model $(f_{\text{ref}}, g_{\text{ref}}) \in \mathcal{F} \times \mathcal{G}$, the ε-Rashomon set $R_{\text{set}}(\varepsilon, \mathcal{D}_{\text{val}})$ is the set of all models (f, g) whose total loss is within a tolerance ε of the reference loss:

$$R_{\text{set}}(\varepsilon, \mathcal{D}_{val}) = \{(f, g) \in \mathcal{F} \times \mathcal{G} : \mathcal{L}_{\text{total}}(f, g; \mathcal{D}_{val}) \leq \mathcal{L}_{\text{total}}(f_{\text{ref}}, g_{\text{ref}}; \mathcal{D}_{val}) + \varepsilon\}. \tag{1}$$

Typically, the reference model is an empirical risk minimizer $(f_{\text{ref}}, g_{\text{ref}}) \in \arg\min_{f \in \mathcal{F}, g \in \mathcal{G}} \mathcal{L}_{\text{total}}(f, g, \mathcal{D})$ (Xin et al., 2022; Semenova et al., 2022). However, obtaining the global minimizer can be challenging for neural networks, and practically, any well-trained CBM can serve as the reference model (Fisher et al., 2019). For example, one approach is to choose as a reference model the one that performs well on the validation data. Similarly, we evaluate the Rashomon set on validation data to mitigate overfitting in neural networks and ensure that the resulting Rashomon set contains models with comparable generalization performance.

Exploring Models in the Rashomon Set. $R_{\text{set}}(\varepsilon, \mathcal{D}_{\text{val}})$ formally contains all models that satisfy the performance bound, but it is typically an uncountable and computationally intractable set to fully compute, particularly when the hypothesis spaces $\mathcal{F}$ and $\mathcal{G}$ are parameterized by neural networks. Moreover, even if we could compute many near-optimal neural-network solutions, the resulting collection may still be uninformative: neural networks often admit numerous parameterizations that induce very similar functions, resulting in decision boundaries that are nearly indistinguishable on the data distribution.

Therefore, a more practical alternative is to focus on a finite *diverse* subset of near-optimal models. Unlike previous works that primarily aim to obtain arbitrary members of the Rashomon set (Hsu et al., 2024), our goal is to identify a set of M models that are both *accurate and meaningfully distinct* in their decision-making logic. Concretely, given a user-defined similarity metric, we seek a set of near-optimal models that are well-separated under this metric. To do so, we compute a set of M accurate models that minimize total pairwise similarity among its elements. We refer to this set as a *Rashomon slice* and discuss next how we compute it. Since the slice explicitly encourages pairwise separation, it can be viewed as an *approximate* M-model packing (Semenova et al., 2022) of the Rashomon set under the similarity metric.

3.2 Finding the Rashomon Slice

To compute the Rashomon slice, we jointly optimize a collection of M models using a single objective that balances predictive performance with an explicit separation regularizer, encouraging the learned models to remain near-optimal while promoting diversity across the models. The objective is optimized on the training data, while Rashomon set membership (i.e., the performance bound) is evaluated on the validation set $\mathcal{D}_{\text{val}}$ to better reflect generalization.

Diversity-Aware Minimax Optimization. We discover the Rashomon slice by training M models simultaneously. To encourage that every individual model in the slice satisfies the Rashomon performance bound, (i.e., $\mathcal{L}_{\text{total}}(f_{\text{ref}}, g_{\text{ref}}; \mathcal{D}_{val}) + \varepsilon$), we minimize the maximum task and concept prediction losses across the models. As for similarity constraint, the primary source of diversity in CBMs lies in their internal concept representations. For two CBMs to represent distinct decision-making strategies, they should produce different concept activation patterns for the same input. Therefore, we maximize diversity by penalizing the average similarity between the predicted concept vectors across models. The resulting joint objective is:

$$\min_{f_m, g_m \forall m} \left[\max_m \mathcal{L}_Y(f_m(g_m(\mathbf{x})), y) + \lambda \cdot \max_m \mathcal{L}_C(g_m(\mathbf{x}), \mathbf{c}) - \frac{\alpha}{M} \sum_{m=1}^{M} \mathcal{L}_{\text{div}}(g_m) \right], \tag{2}$$

where $\alpha > 0$ controls the strength of the diversity regularization and the diversity loss for model m is defined as $\mathcal{L}_{\text{div}}(g_m) = 1 - \frac{1}{M-1} \sum_{m' \neq m} \text{sim}(g_m(\mathbf{x}), g_{m'}(\mathbf{x}))$. This minimax surrogate (first two terms in Eq. 2) encourages each model's total loss to remain small, serving as a proxy to satisfy the validation-based Rashomon bound in Definition 1. $\mathcal{L}_{\text{div}}$ leverages training set cosine similarity as a differentiable surrogate to practically encourage this separation during optimization. Overall, this formulation encourages the models to remain near-optimal, while the diversity term encourages the models to occupy distinct regions of the decision space.

Parallel Architecture. Solving the optimization problem in Eq. 2 by training M neural networks simultaneously is computationally intensive. To address this, we propose a specialized architecture where all models share a pretrained backbone and utilize model-specific adapters that operate in parallel.

In our framework, each model $m \in \{1, \ldots, M\}$ consists of a concept encoder $g_m(\cdot)$ and a classifier $f_m(\cdot)$, i.e., $\hat{y}^{(m)} = f_m(g_m(\mathbf{x}))$. We construct M parallel models by inserting adapter modules into specific layers of a shared backbone network B (e.g., ViT). Each model m maintains its own set of trainable adapters, denoted as $A_m = \{A_m^{(1)}, A_m^{(2)}, \ldots, A_m^{(L)}\}$, where the superscript indicates the layer index at which an adapter is inserted. During a forward pass, only the adapters associated with model m are activated, resulting in an effective backbone $B_m(\mathbf{x}) = B(\mathbf{x}, A_m)$. The backbone parameters are shared and frozen across all models, while each adapter set A_m is trainable and model-specific (and thus unique). We denote $B^{(l)}$ as shared parameters without adapters of layer l.

Given the adapted features $B_m(\mathbf{x})$, each model m uses a set of per-concept multilayer perceptrons (MLPs) $\{h_{m,j}\}_{j=1}^{p}$ to predict the activation of concept j. The resulting concept prediction vector is then

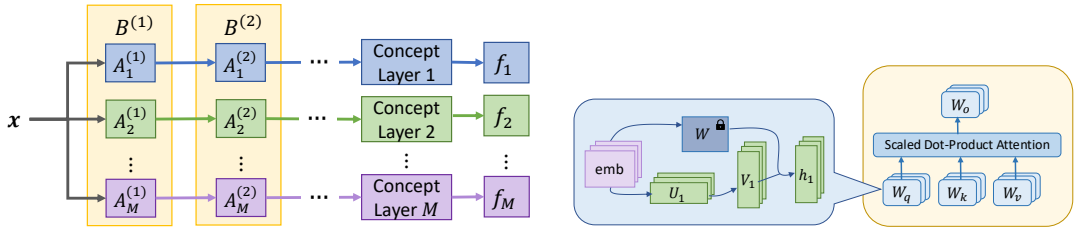
$$\hat{\mathbf{c}}^{(m)} = g_m(\mathbf{x}) = [h_{m,1}(B_m(\mathbf{x})), ..., h_{m,p}(B_m(\mathbf{x}))].$$

This design follows Espinosa Zarlenga et al. (2022) and Kim et al. (2023) to improve the flexibility of learning different concept representations.

Each classifier $f_m(\cdot)$ then maps the predicted concept vector $\hat{\mathbf{c}}^{(m)}$ to task label, $\hat{y}^{(m)} = f_m(\hat{\mathbf{c}}^{(m)})$. Together, the collection $\{g_m, f_m\}_{m=1}^{M}$ forms a data-driven Rashomon slice. Figure 1a summarizes the proposed architecture.

Implementation. Though the proposed architecture is backbone- and adapter-agnostic, we implement it using a Vision Transformer (ViT) backbone with Low-Rank Adaptation (LoRA) (Hu et al., 2022) modules as an example.

Low-Rank Parameterization. Each model m in the Rashomon slice is equipped with a unique set of LoRA modules inserted into the query, key, value, and projection matrices of every attention block. For a



(a) Overall structure (b) Zoom-in on a backbone block

Figure 1: Proposed architecture for finding the Rashomon slice. (a) Overall structure: an input image passes through a frozen backbone with attached adapters, followed by multiple parallel concept layers, each linked to a classifier. (b) Zoom-in on a backbone block: at each Attention layer, multiple adapter modules are inserted in parallel into the Q, K, V, and projection mappings, operating concurrently and independently. All adapters are trained jointly, while at inference, each model activates only its own adapters while sharing the same backbone weights.

pretrained weight matrix $W \in \mathbb{R}^{d_{out} \times d_{in}}$, the weight update is constrained to a low-rank decomposition:

$$W + \Delta W_m = W + U_m V_m$$

where $U_m \in \mathbb{R}^{d_{out} \times r}$ and $V_m \in \mathbb{R}^{r \times d_{in}}$ are trainable parameters with rank $r \ll \min(d_{in}, d_{out})$. The collection of these low-rank parameters across all L blocks, denoted as $\{U_m^{(l)}, V_m^{(l)}\}_{l=1}^{L}$, constitutes the adapter set A_m that differentiates model m from other members of the Rashomon slice (see Figure 1b).

Surrogate Losses and Similarity. In the definition of the Rashomon set, misclassification loss is usually used for task and concept accuracy, as this directly reflects the predictive performance we aim to guarantee. However, since the misclassification loss is non-differentiable and cannot be optimized via gradient descent, we utilize cross-entropy loss for $\mathcal{L}_Y$ and $\mathcal{L}_C$ during the training process. For the similarity metric $\mathrm{sim}(\cdot, \cdot)$ used in $\mathcal{L}_{\text{div}}$, we compute the cosine similarity between the predicted concept vectors $g_m(\mathbf{x})$ and $g_{m'}(\mathbf{x})$, thereby encouraging models to learn distinct concept representations.

Model-axis Checkpointing Scheme for Efficient Memory Usage. Jointly training M CBMs to explore the Rashomon slice is memory-intensive: activation memory scales nearly linearly with M, even when conventional layer-wise checkpointing is applied to each model. To avoid this blow-up, we introduce a *model-axis checkpointing scheme*: we wrap each model's full forward (and backward) invocation inside a checkpoint boundary, so that during backpropagation only one model is active and holds its activations at any given time. Empirically, this reduces peak memory to a level comparable to training a single model, rather than growing linearly with the size of the slice, which is useful when M is large but GPU memory is limited.

4 Experiments

Our evaluation addresses the following questions: (1) how the Rashomon slice compares to baseline methods in discovering diverse yet well-performing models and what qualitative differences arise in how individual models rely on concepts (Section 4.1); (2) how the diversity across members can be leveraged for practical downstream tasks (Section 4.2); (3) whether our method remains memory- and parameter-efficient (Section 4.3); and (4) whether the framework generalizes beyond the default ViT+LoRA configuration (Section 4.4).

Datasets. We conduct experiments on five datasets: AwA2 (Xian et al., 2019), CIFAR-10 (Krizhevsky, 2009), CUB-200-2011 (Wah et al., 2011), HAM10000 (Tschandl et al., 2018), and CelebA (Liu et al., 2015). AwA2, CUB and CelebA have human-annotated concepts. Therefore, we directly use them for training. CIFAR-10 and HAM10000 lack explicit concept annotations, and we follow the automatic concept discovery pipeline proposed in Oikarinen et al. (2023) and Chowdhury et al. (2024), respectively, to get concepts.

Baselines. Since, as discussed in the related work section, we know of no previous method that constructs the Rashomon set for CBMs, we adapt and extend three related approaches to serve as baselines, covering brute-force retraining, inference-time approximation, and diversity-regularized ensembles:

- *Random initialization* represents a brute-force method to find a collection of models. We train M CBMs using different random seeds for parameter initialization. Each CBM is fully tunable and updated during training.
- *Dropout CBMs* approximates a Rashomon set through inference-time sampling rather than full retraining. Following Hsu et al. (2024), we insert dropout layers throughout the backbone and train a single CBM with dropout enabled. By keeping dropout active during inference and varying the random seed for each pass, we create a set of distinct models, each defined by a unique dropout mask applied to the same trained architecture.
- *Diverse Ensemble (DivEnsemble)* is adapted from the diversified deep ensemble method (Zhang et al., 2020). It uses a shared, fully tunable backbone B and a shared concept predictor g, while each model maintains its own task classifier f_m. All models are trained jointly, and the diversity regularization term is applied at the task prediction level.

Diversity metrics. We evaluate all methods using both concept and task prediction accuracy. Beyond accuracy, we report a set of diversity metrics that provide a more comprehensive analysis of how models differ. We consider the following diversity metrics:

- *Centered Kernel Alignment (CKA)* compares the concept representations of two models (Kornblith et al., 2019). Using a linear kernel $k(\mathbf{x}, \mathbf{y}) = \mathbf{x}^\top \mathbf{y}$, we form Gram matrices $K^{(m_1)}$ and $K^{(m_2)}$ from the two models' concept representations evaluated on the same inputs, center them as $\tilde{K}^{(m)} = HK^{(m)}H$ with $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$. CKA is then computed as the cosine similarity between these centered Gram matrices, normalized by their Frobenius norms.
- *SHAP similarity* (Lundberg & Lee, 2017) measures how similarly different models use learned concepts for prediction. We obtain SHAP importance scores over concepts for each model, forming a SHAP importance vector. Diversity is measured by the 1-cosine similarity between the SHAP vectors of two models, where lower similarity indicates more distinct usage of concepts for decision-making.
- *Concept coverage* measures the size of the union of the most important concepts across models. For each model, we rank concepts by their SHAP importance scores on the full dataset and select the top 10. We then compute the size of the union of these top-10 concept sets across models. This quantity reflects how much the models' most influential concepts overlap: a larger union indicates more diverse concept usage.

For both CKA and SHAP similarity, we compute all pairwise scores across models to form an $M \times M$ similarity matrix S on a fixed test set. We summarize similarity using the off-diagonal mean $\bar{S}_{\text{off}} = \frac{2}{M(M-1)} \sum_{m_1 < m_2} S_{m_1 m_2}$, where a higher value indicates stronger agreement among models.

Experimental Setup. We set $M = 10$ for each method and use a pretrained ViT-S/16 as the backbone, except for the HAM10000 dataset where we use medical ViT-B/16 (Eslami et al., 2023). The concept predictor $\{h_{m,j}\}_{j=1}^{p}$ and label classifier f_m are implemented as linear layers. To ensure a balance between concept and task accuracy, we set $\lambda = 1$. In our method, we insert LoRA modules into the Q, K, V, and projection matrices of every attention block, with a 0.1 dropout ratio, a scaling multiplier $s = 2$, and a rank of $r = 8$ for AwA2, CIFAR-10, and CelebA, $r = 16$ for HAM10000, and $r = 32$ for CUB. All models are optimized using AdamW, with learning rate and weight decay tuned on a single L40 GPU.

Once all methods are trained, we first identify the models belonging to the Rashomon set. We select the model from the random initialization baseline that achieves the lowest validation misclassification error-based total loss (see Eq. 1) to serve as a reference model. We set $\varepsilon = 0.03$, corresponding to an absolute performance tolerance of approximately 1.5% in both task and concept accuracy – a threshold small enough to ensure

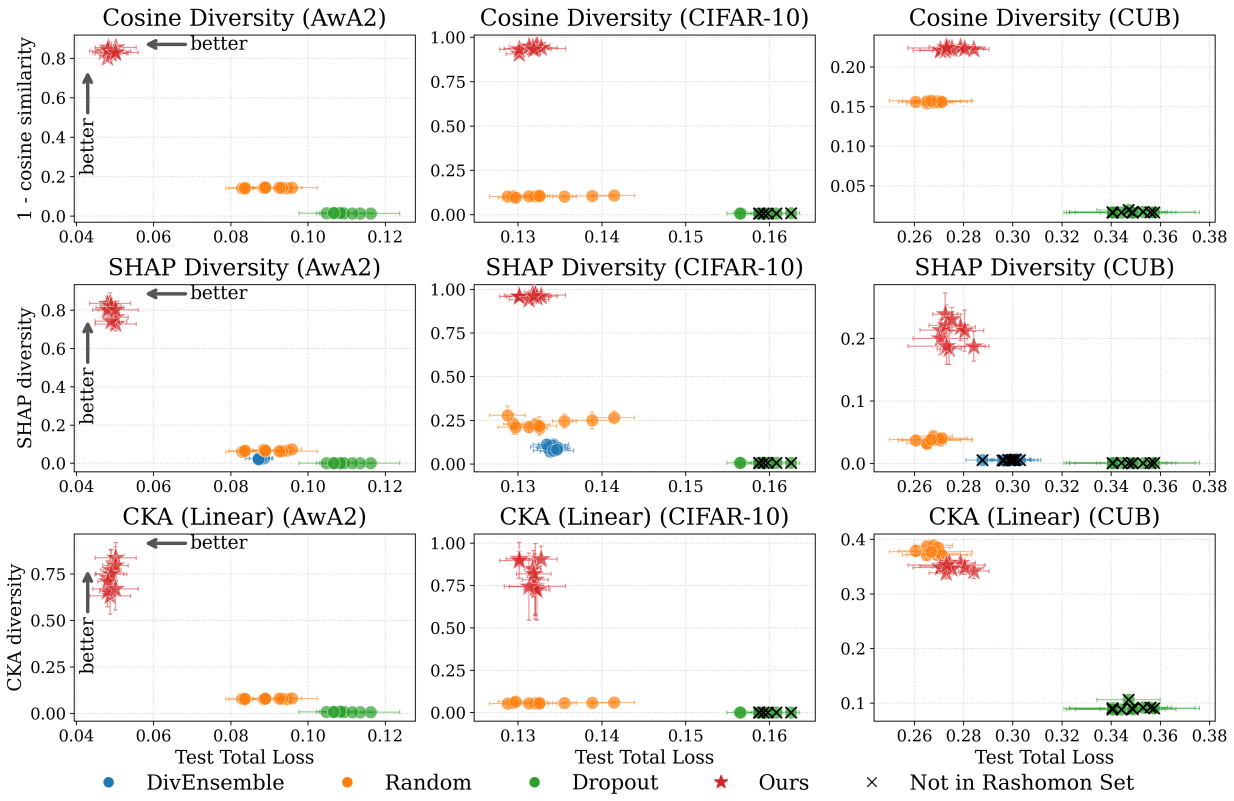


Figure 2: Comparison of model diversity versus total loss(misclassification loss) on test set across different methods on the AwA2, CIFAR-10, and CUB datasets. The top row displays cosine diversity, the middle row displays SHAP diversity, and the bottom row shows CKA diversity. Each point represents an individual model. Black crosses indicate models that do not satisfy the Rashomon set belonging criterion on the validation data. Our method achieves higher diversity while remaining within the Rashomon set.

all models remain competitive with the reference. We identify all models whose validation loss falls within an ε-margin of the reference model's validation misclassification error-based total loss as members of the Rashomon set. We then evaluate the diversity of these members using the following metrics: cosine similarity and centered kernel alignment (CKA) (Kornblith et al., 2019) of concept predictions, cosine similarity of SHAP importance for the concepts (Lundberg & Lee, 2017), and concept coverage (the size of the union of the top-10 most important concepts on the whole dataset evaluated by SHAP). More details about datasets and configurations are in Appendix A.

4.1 The Rashomon Slice Contains Accurate and Diverse Models

Quantitative Results. In this section, we show that our method can find more diverse models yet stay within the Rashomon set (See Figure 2). It is important to note that each point in Figure 2 represents an individual model. We evaluate the standalone performance of each model obtained from different methods.

As we can see, our method successfully finds models (red stars) that consistently achieve low test total loss with no crosses on them, indicating that all of these models fall within the $\varepsilon-$threshold of the reference model. Meanwhile, models found by our method also stay in the top-left region across the AwA2, CIFAR-10, and CUB datasets. This indicates that our method is able to find models that achieve both low test total loss and higher diversity measured by cosine similarity, SHAP, and CKA compared to baselines. Beyond pairwise diversity, we report the concept coverage – the size of the union of the top 10 most important concepts based on SHAP across all models from a given method that successfully fall within the Rashomon set. Table 1

Table 1: Comparison of concept coverage across five datasets for models **in the Rashomon set**. N/A indicates that no models produced by that method fell within the Rashomon set.

Method	AwA2	CIFAR10	CUB	HAM10000	CelebA
DivEnsemble	33	19	N/A	N/A	N/A
Random	43	36	51	36	**6**
Dropout	19	14	N/A	N/A	**6**
Ours	**59**	**61**	**63**	**57**	**6**

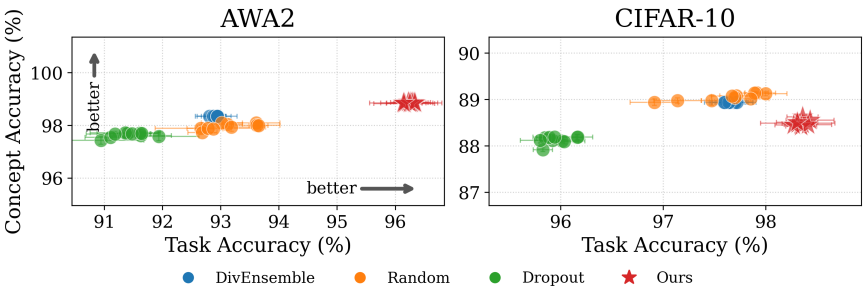


Figure 3: Per-model task and concept accuracy on AwA2 and CIFAR-10. Each point represents an individual model. Our method achieves task accuracy comparable to or better than baselines, with stable variance across members and only a marginal decrease in concept accuracy on the CIFAR-10 dataset.

shows that our method achieves the highest coverage, indicating that it finds models that rely on a broader set of concepts for prediction. It is worth noting that both our method and some baselines achieve a concept coverage of six on the CelebA dataset. This is because our setup of CelebA provides only six available concepts, which limits the space of concept-based explanations. More results are shown in Appendix B.

Figure 3 shows that on both AwA2 and CIFAR-10 datasets, the models found by our method attain task accuracy that is comparable to or better than the baselines with low variance across Rashomon members. Concept accuracy decreases marginally compared to the strongest baselines. Combining with Figure 2 and Figure 9, the results show that our method achieves high concept accuracy and high cosine diversity simultaneously, so there is enough inherent data diversity for the cosine penalty to not fight the concept loss for reasonable diversity regularization. Instead, it selects among the multiple high-accuracy concept encoders that the data already admits.

Qualitative Results. In this section, we provide qualitative results for the diversity of models discovered by our method. We visualize five arbitrary models from the Rashomon slice for the prediction of **beaver**, **deer**, and **tiger** classes, highlighting the top five concepts that each model relies on. As shown in Figure 4, different members of the slice justify the same prediction using different rationales. We categorize these concepts into four decision rationale groups: positive evidence (having it supports the prediction, green), negative evidence (having it has a negative effect on the predictions, red), excluding evidence (does not having it supports the predictions, blue), and spurious correlation (the concept is problematic, purple).

Taking **beaver** class as an example, the five Rashomon members justify the same prediction through different conceptual strategies. Model 0 and Model 4 rely almost exclusively on exclusion reasoning, identifying absent traits from different perspectives such as “fly”, “meat”, “stripes”, and “desert” for Model 0 and “orange”, “tree” and “longneck” for Model 4 to rule out non-beaver classes, while using a small set of physical characteristics as direct evidence. Model 1, on the other hand, combines positive evidence from physical traits (“chewteeth”, “small”) with environmental exclusion (“desert”, “bush”) and treats “strong” as negative evidence. Model 2 also mixes positive physical traits (“fast”, “paws”) with excluding evidence (“tusk”, “orange”), while using “quadrupedal” as negative evidence. Model 3 is dominated by a single large exclusion signal “hands” and is complemented by two behavioral traits “timid” and “slow” as positive and excluding evidence respectively. Interestingly, Model 3 also includes a spurious concept “jungle”, which means that the model has a wrong

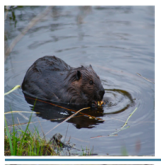
Class		Model 0	Model 1	Model 2	Model 3	Model 4
Beaver		- fly (+4.56)	- desert (+3.71)	- tusk (+2.68)	- hands (+9.21)	- orange (+3.66)
		- meat (+4.51)	+ chewteeth (+1.07)	+ fast (+2.19)	- jungle (-2.69)	- tree (+1.28)
		- stripes (+1.42)	+ small (+1.02)	+ quadrupedal (-1.71)	- slow (+0.70)	+ fish (+1.08)
		- desert (+1.23)	+ strong (-1.00)	+ paws (+1.12)	- mountains (+0.69)	- longneck (+0.79)
		+ buckteeth (+1.20)	- bush (+0.99)	- orange (+1.12)	+ timid (+0.64)	+ buckteeth (+0.75)
Deer		- meatteeth (+1.93)	- orange (+8.68)	- jungle (+14.78)	- meat (+17.74)	- insect (+5.02)
		+ strong (-1.44)	- plankton (-1.52)	- hands (-8.00)	- coastal (+1.12)	+ furry (+1.35)
		+ grazer (+1.04)	- muscle (+0.98)	+ plains (+1.24)	- hairless (+1.02)	+ muscle (+1.30)
		+ hooves (+1.03)	- meatteeth (+0.97)	+ group (+1.04)	- pads (-0.80)	+ longleg (+1.21)
		- water (+0.99)	+ horns (0.92)	- weak (+1.02)	- stalker (+0.75)	- hunter (+1.17)
Tiger		+ strong (+5.76)	- swims (+3.66)	- tusk (-5.63)	+ fierce (+1.39)	- hooves (+16.77)
		+ meatteeth (+2.12)	- longneck (+1.46)	+ quadrupedal (+3.22)	- grazer (+1.32)	+ claws (+1.38)
		- small (+1.81)	- gray (+1.38)	+ fast (+1.95)	- skimmer (+1.02)	- newworld (+1.00)
		- desert (-1.10)	+ orange (+1.24)	- tunnels (-1.53)	- longneck (+0.92)	+ ground (-0.96)
		- domestic (+1.01)	+ straightteeth (+1.13)	- inactive (+1.48)	- vegetation (+0.74)	- plankton (+0.79)

Figure 4: Top-5 concepts based on contribution scores $W_{p,k}\hat{c}_p$ of each concept p to the prediction of class k (value in parentheses) for beaver (top), deer (middle), and tiger (bottom) classes across five Rashomon slice members. The sign stands for whether the concept is predicted to be present ($+$, $\hat{c}_p > 0$) or absent in the corresponding class ($-$, $\hat{c}_p \leq 0$). All the predictions for concepts shown are correct and correspond to ground-truth concept labelings for the classes. Colors indicates the decision rationale in which the concept contributes to the prediction: green denotes positive evidence ($W_{p,k} > 0, \hat{c}_p > 0$); red represents negative evidence ($W_{p,k} \leq 0, \hat{c}_p > 0$); blue means excluding evidence ($W_{p,k} > 0, \hat{c}_p \leq 0$); purple stands for spurious evidence ($W_{p,k} \leq 0, \hat{c}_p \leq 0$). The diverse patterns highlight that different members rely on different rationales.

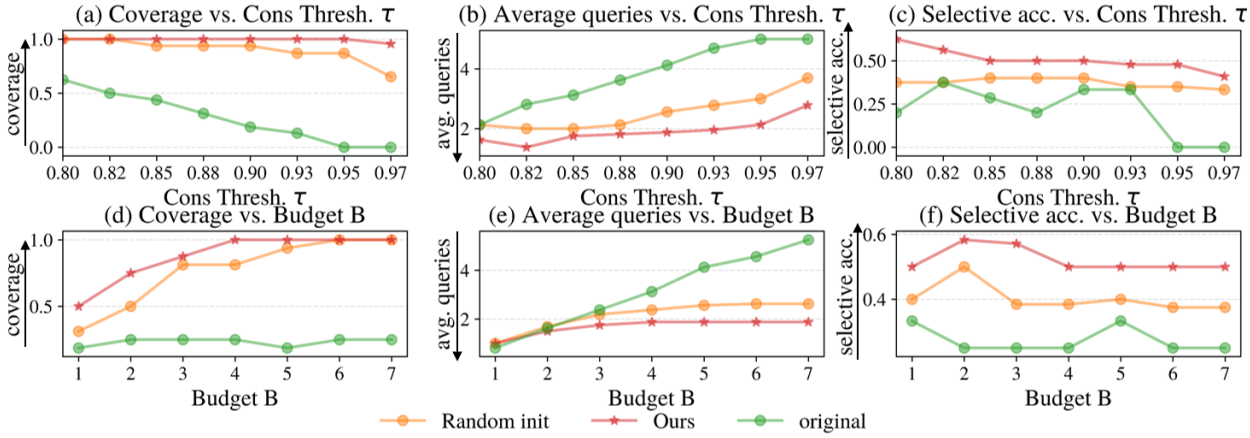


Figure 5: Performance of abstention on Dermatofibroma class from the HAM10000 dataset across varying the confidence thresholds τ (top row) and query budgets B (bottom row) on the coverage (fraction of non-abstained predictions), average number of concept queries, and selective accuracy (accuracy on non-abstained samples). Arrows indicate the preferred direction. Compared to the original single-CBM safeguard (*original*) and the random initialization baseline, the Rashomon-member safeguard achieves higher coverage with fewer queries while improving selective accuracy across settings.

perception of the surroundings. Overall, the five members span different combinations of decision-making rationales and use different profiles of traits ranging from physical, behavioral to contextual, illustrating that equivalent predictions within a Rashomon set can arise from fundamentally different concept-level decision mechanisms.

Beyond inspection, the Rashomon slice enables personalized intervention by allowing a practitioner to select a member whose concept reliance best matches domain preferences. For example, if a user finds a highly weighted concept, such as *jungle* in Model 2 for the **deer** class, undesirable or potentially spurious, they can instead deploy a different Rashomon slice member that achieves similar accuracy while placing less emphasis on that concept (e.g., Model 4). This provides a lightweight alternative to retraining: practitioners can choose among models in the slice based on the concepts they want to encourage or avoid. We discuss more about model choice and other use-cases in Section 4.2.

Diversity Analysis. A diverse Rashomon slice is useful in practice if its members genuinely reason differently. Just like DivEnsemble, if members just re-weight the same underlying concept representations through different output heads, changing between models can reduce a concept's influence on the final prediction, but cannot offer a genuinely different way of perceiving the input. In other words, since the underlying concept detectors remain identical, any spurious or flawed concept detection will happen across all members. Since our architecture gives each member its own LoRA adapters inside the shared backbone, we expect diversity to emerge at the level of concept representations themselves, not just at the output head. To show this, we performed intervention experiments on the tiger class on the AwA2 dataset, and also traced the origin of concept-representation diversity through layer-wise eigenvector visualizations with details in Appendix F. Overall, we observed that diversity is concentrated in the deeper LoRA adapters rather than the shallow layers and confirmed that representational diversity in the Rashomon slice models is architecturally grounded.

4.2 Use Cases: The Rashomon Slice Can Help with Abstention, Mitigation, and Trustworthy Model Selection

In this section, we provide three distinct use cases once we have the Rashomon slice for CBMs.

1. Efficient and Reliable Abstention. In cases when CBM is not confident, prior work introduced conceptual safeguards (Joren et al., 2024) that abstain from predicting on high-uncertainty samples and instead seek human intervention to confirm key concepts. Traditional safeguards rely on Monte Carlo sampling from a single model to quantify uncertainty. However, this often reflects local variation around a single rationale. A Rashomon set provides a more natural framework for this task, as it contains multiple, near-optimal, but conceptually distinct rationales for the same input. Therefore, we use individual Rashomon members to estimate uncertainty. If the members disagree on a task prediction beyond a threshold τ, the system abstains from the sample and queries an expert for a specific concept label (within a budget B). We then reweight the members based on their consistency with the expert's feedback, repeating this process until the weighted consensus exceeds τ or the budget B is exhausted.

We compare our Rashomon-member safeguard against the single-CBM safeguard and the random initialization baseline across various budgets B and thresholds τ on both Beaver class from the AwA2 dataset (see Appendix E) and Dermatofibroma class from the HAM10000 dataset. As shown in Figure 5, our method returns a prediction on a larger fraction of test examples without needing to abstain, even at high confidence thresholds (left column). Our models require fewer human concept queries to reach a confident consensus (middle column), and for the samples where a prediction is provided, the accuracy is higher than the baselines (right column). These results indicate that diverse Rashomon members can provide a more precise indicator of uncertainty than sampling from a single model.

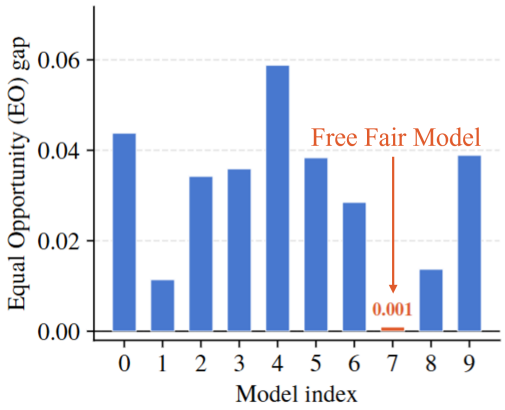


Figure 6: The Rashomon slice finds a fair model for free (Young → Wavy Hair). All ten models achieve comparable accuracy, but Equal Opportunity (EO) gaps span $[0.001, 0.059]$. Model 7 achieves near-zero EO gap at no accuracy cost.

2. Trustworthy Model Selection. Beyond diversity, the Rashomon slice enables a broader form of trustworthy model selection: since all members achieve near-identical accuracy by construction, any variation in deployment-relevant properties across members is a free gain with no accuracy trade-off required. Whenever a practitioner has a secondary trustworthiness objective in mind, they can simply select the Rashomon member that best satisfies it, rather than retraining under complex constraints (Black et al., 2022; Hsu et al.). We demonstrate this principle with fairness using the constructed Rashomon slice.

On CelebA, despite similar accuracy, Equal Opportunity gaps span $[0.001, 0.059]$ across Rashomon slice members (Figure 6), indicating that fairness disparity is an artifact of a particular decision rationale rather than an inherent property of the task and a fairer model can be obtained for free, a gain which is unattainable for single-model pipelines. The same principle might extend naturally to other trustworthiness properties such as robustness or privacy, since each Rashomon slice member emphasizes a distinct set of concepts and may therefore interact differently with distributional shifts or membership inference (Hsu et al., 2026).

Table 2: Conceptual collapse and class-wise accuracy gaps for confounded class pairs. Random initialization yields highly aligned concept rationales and large inter-class gaps, whereas Rashomon slice reduce cross-model concept similarity and shrink the accuracy gap.

	Random Init	Dropout	DivEnsemble	Ours
Indigo Bunting vs. Blue Grosbeak				
Avg Cross-Model Concept Cos Sim ↓	0.8045	0.9770	N/A	**0.6085**
Avg Acc (IB/BG) ↑	88.67% / 72.67%	82.67% / 58.67%	86.67% / 68.00%	**92.00% / 81.33%**
Avg Acc Gap (IB-BG) ↓	16.00%	24.00%	18.67%	**10.67%**
Bobcat vs. Leopard				
Avg Cross-Model Concept Cos Sim ↓	0.8549	0.9883	N/A	**0.1629**
Avg Acc (BC/LP) ↑	94.21% / 99.58%	93.65% / 98.96%	93.10% / 99.93%	**98.17% / 99.93%**
Avg Acc Gap (LP-BC) ↓	5.37%	5.31%	6.83%	**1.76%**

3. Confounding Classes. Fine-grained datasets often contain confounding class pairs with substantial visual overlap (e.g., CUB: Indigo Bunting vs. Blue Grosbeak; AwA2: Bobcat vs. Leopard), where a model may lock onto evidence that is predictive for one class but insufficiently discriminative for its confounder, resulting in a large per-class accuracy gap. We address this by using the Rashomon slice as a diversified ensemble: rather than relying on a single decision rationale, we aggregate across members whose diverse concept representations collectively cover more discriminative evidence. This mitigates conceptual collapse, where independently trained models converge on the same shared, potentially non-discriminative concepts. As shown in Table 2, the random initialization baseline exhibits strong conceptual collapse (cross-model cosine similarity ≥ 0.8) and a large accuracy gap between the two confounded classes. In contrast, Rashomon slice reduces cross-model concept similarity substantially, and as a result both per-class accuracies improve and the gap between them shrinks.

4.3 Memory and Parameter Efficiency

We also report the memory cost and the number of trainable parameters for each method on AwA2 dataset in Table 3 (left). Our method is very memory efficient, using only 7.6% of the memory and 1.2% of the trainable parameters required by the method based on random initialization. This substantial reduction stems from our adapter structure and model-axis checkpointing schema. Further, ablations in Table 3 (right) show that these design choices reduce different components of the memory (parameters vs. activations) and their effects accumulate. More results are available in Appendix C.

4.4 Generalizability beyond ViT+LoRA Architecture

We further evaluate whether our method can be generalized to other backbone and adapter choices other than ViT+LoRA. We add two additional implementation settings: (1) ViT-S/16 with DoRA, which changes the adapter structure and (2) ResNet-18 with ConvAdapters, which completely changes the backbone and the

Table 3: **Left:** Exact and relative memory and trainable-parameter costs on AwA2. Random initialization serves as the baseline for ratios. **Right:** Peak GPU memory under different combinations of LoRA and checkpointing on AwA2. In our method, checkpointing and LoRA are designed to reduce different components of the memory footprint, and combining them gives the largest memory reduction.

Metrics	Rand.	DivE.	Ours
Mem (GiB)	37.09	4.66	**2.81**
Mem (%)	100.0	12.5	**7.6**
# Params (M)	217.0	21.7	**2.7**
# Params (%)	100.0	10.0	**1.2**

Variant	Mem (GiB)	Reduction (%)
no-LoRA / no-ckpt	37.09	0%
no-LoRA / ckpt	4.66	87.4%
LoRA / no-ckpt	25.54	31.1%
LoRA / ckpt	**2.81**	**92.4%**

corresponding adapters. In each setting, all baselines (Random, Dropout, DivEnsemble) are also implemented with the same backbone.

As shown in Figure 7, in both settings our method identifies Rashomon slices with better or comparable task loss and higher SHAP diversity than baselines using the same backbone. These results indicate that the proposed framework is not tied to certain backbones or adapters, but can be applied to different parameter-efficient adaptation mechanisms and backbone architectures.

Figure 7: **Generalizability beyond ViT+LoRA.** We evaluate our method under two additional settings: ViT-S/16 with DoRA and ResNet-18 with ConvAdapters. Each point corresponds to one model in the Rashomon slice. Our method achieves higher SHAP diversity at comparable test total loss, suggesting that the framework generalizes beyond the ViT+LoRA configuration.

5 Conclusions and Future Work

This work introduced the first lightweight framework for constructing a diverse set of equally accurate yet conceptually distinct CBMs from a single training process. We demonstrated that access to this Rashomon slice unlocks practical capabilities unavailable to single-model approaches, including reliable abstention by treating member disagreement as an uncertainty signal, trustworthy model selection by identifying members that satisfy secondary objectives such as fairness at no accuracy cost, and resolution of inter-class confusion by leveraging diverse discriminative rationales across members. Taken together, these use cases show that model multiplicity in CBMs is a practical resource that practitioners can actively exploit for downstream tasks. Beyond the current framework, several directions remain open for future work. The number of models M is currently user-specified. Developing data-driven criteria to determine it automatically remains an important extension. Our method requires more training epochs to converge than single-model baselines, a common trade-off with parameter-efficient adaptation, which we view as an acceptable cost given the substantial memory savings that allow many more slice members to be trained under fixed GPU budgets. Nonetheless, developing training schedules that narrow this gap remains a future direction. Introducing sparsity regularization could encourage each member to rely on a smaller, more interpretable subset of concepts. Finally, interactive visualization tools that allow domain experts to browse and compare Rashomon members would make the diversity of reasoning strategies accessible beyond machine learning practitioners.

References

Dasom Ahn, Sangwon Kim, and Byoung Chul Ko. Energy-ensemble concept bottleneck models for enhancing interpretability and accuracy in concept-based learning. *Knowledge-Based Systems*, 341:115821, 2026. ISSN 0950-7051. doi: 10.1016/j.knosys.2026.115821. 2

Dan Biderman, Jacob Portes, Jose Javier Gonzalez Ortiz, Mansheej Paul, Philip Greengard, Connor Jennings, Daniel King, Sam Havens, Vitaliy Chiley, Jonathan Frankle, Cody Blakeney, and John Patrick Cunningham. LoRA learns less and forgets less. *Transactions on Machine Learning Research*, 2024. ISSN 2835-8856. Featured Certification. 21

Emily Black, Manish Raghavan, and Solon Barocas. Model multiplicity: Opportunities, concerns, and solutions. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 850–863, 2022. 2, 11

Zachery Boner, Harry Chen, Lesia Semenova, Ronald Parr, and Cynthia Rudin. Using noise to infer aspects of simplicity without learning. *Advances in Neural Information Processing Systems*, 37:131824–131858, 2024. 2

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, 16(3):199–231, 2001. 2

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020. 2

Townim F Chowdhury, Vu Minh Hieu Phan, Kewen Liao, Minh-Son To, Yutong Xie, Anton van den Hengel, Johan W Verjans, and Zhibin Liao. Adacbm: An adaptive concept bottleneck model for explainable and accurate diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 35–45. Springer, 2024. 5, 17

Stanislav Dereka, Ivan Karpukhin, Maksim Zhdanov, and Sergey Kolesnikov. Diversifying deep ensembles: A saliency map approach for enhanced ood detection, calibration, and accuracy. In *2024 IEEE International Conference on Image Processing (ICIP)*, pp. 437–443, 2024. doi: 10.1109/ICIP51287.2024.10647334. 3

Jiayun Dong and Cynthia Rudin. Exploring the cloud of variable importance for the set of all good models. *Nature Machine Intelligence*, 2(12):810–824, 2020. 2

Jon Donnelly, Zhicheng Guo, Alina Jade Barnett, Hayden McTavish, Chaofan Chen, and Cynthia Rudin. Rashomon sets for prototypical-part networks: Editing interpretable models in real-time. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4528–4538, 2025. 1, 2

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. *Journal of Machine Learning Research*, 2020. 2

Gilles Eerlings, Brent Zoomers, Jori Liesenborgs, Gustavo Rovelo Ruiz, and Kris Luyten. DIVERSE: Disagreement-inducing vector evolution for rashomon set exploration. In *The Fourteenth International Conference on Learning Representations*, 2026. 2

Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In Andreas Vlachos and Isabelle Augenstein (eds.), *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 1181–1193, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-eacl.88. 6, 17

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, et al. Concept embedding models: Beyond the accuracy-explainability trade-off. *Advances in neural information processing systems*, 35:21400–21413, 2022. 2, 4, 17

Aaron Fisher, Cynthia Rudin, and Francesca Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019. 2, 3

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019. 3

Prakhar Ganesh, Afaf Taik, and Golnoosh Farnadi. Systemizing multiplicity: The curious case of arbitrariness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 1032–1048, 2025. 2

Ethan Hsu, Tony Cao, Lesia Semenova, and Chudi Zhong. The rashomon set has it all: Analyzing trustworthiness of trees under multiplicity. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*. 2, 11

Ethan Hsu, Harry Chen, Chudi Zhong, and Lesia Semenova. The double-edged nature of the rashomon set for trustworthy machine learning. In *Forty-third International Conference on Machine Learning*, 2026. 2, 11

Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. In *Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 28988–29000, 2022. 2

Hsiang Hsu, Guihong Li, Shaohan Hu, and Chun-Fu Chen. Dropout-based rashomon set exploration for efficient predictive multiplicity estimation. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Representation Learning*, volume 2024, pp. 33104–33144, 2024. 1, 2, 3, 6

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 4

Hailey Joren, Charles Thomas Marx, and Berk Ustun. Classification with conceptual safeguards. In *The Twelfth International Conference on Learning Representations*, 2024. 10, 21

Eunji Kim, Dahuin Jung, Sangha Park, Siwon Kim, and Sungroh Yoon. Probabilistic concept bottleneck models. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, ICML'23. JMLR.org, 2023. 2, 4

Sangwon Kim, Dasom Ahn, Byoung Chul Ko, In-su Jang, and Kwang-Ju Kim. Eq-cbm: A probabilistic concept bottleneck with energy-based models and quantized vectors. In *Proceedings of the Asian Conference on Computer Vision*, pp. 3432–3448, 2024. 2

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020. 2

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International conference on machine learning*, pp. 3519–3529. PMlR, 2019. 6, 7

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical Report TR-2009, University of Toronto, 2009. 5, 17

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017. 3

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 5, 17

Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017. 6, 7

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In *Proceedings of the International Conference on Machine Learning (ICML)*, pp. 6765–6774, 2020. 2

Dominik J. Mühlematter, Michelle Halbheer, Alexander Becker, Dominik Narnhofer, Helge Aasen, Konrad Schindler, and Mehmet Ozgur Turkoglu. LoRA-ensemble: Efficient uncertainty modelling for self-attention networks. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. 23

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. 5, 17

Cynthia Rudin, Chudi Zhong, Lesia Semenova, Margo Seltzer, Ronald Parr, Jiachang Liu, Srikar Katta, Jon Donnelly, Harry Chen, and Zachery Boner. Amazing things come from having many good models. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024. 1, 2

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In *ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT)*, 2022. 1, 2, 3

Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. *Advances in neural information processing systems*, 36:3362–3401, 2023. 2

Changjian Shui, Azadeh Sadat Mozafari, Jonathan Marek, Ihsen Hedhli, and Christian Gagné. Diversity regularization in deep ensembles. *CoRR*, abs/1802.07881, 2018. 3

Asa Cooper Stickland and Iain Murray. Diverse ensembles improve calibration. volume abs/2007.04206, 2020. ICML 2020 Workshop on Uncertainty and Robustness in Deep Learning, ICML UDL 2020; Conference date: 17-07-2020 Through 17-07-2020. 3

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):180161, 2018. 5, 17

Bohdan Turbal, Iryna Voitsitska, and Lesia Semenova. ElliCE: Efficient and provably robust algorithmic recourse via the rashomon sets. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 2

Moritz Vandenhirtz, Sonia Laguna, Ričards Marcinkevičs, and Julia Vogt. Stochastic concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:51787–51810, 2024. 2

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. Caltech-ucsd-birds-200-2011. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. 5, 17

Jamelle Watson-Daniels, David C Parkes, and Berk Ustun. Predictive multiplicity in probabilistic classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 10306–10314, 2023. 2

Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning—a comprehensive evaluation of the good, the bad and the ugly. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9):2251–2265, 2019. doi: 10.1109/TPAMI.2018.2857768. 5, 17

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. *Advances in neural information processing systems*, 35:14071–14084, 2022. 1, 2, 3

Xinyue Xu, Yi Qin, Lu Mi, Hao Wang, and Xiaomeng Li. Energy-based concept bottleneck models: Unifying prediction, concept intervention, and probabilistic interpretations. In *The Twelfth International Conference on Learning Representations*, 2024. 2

Huanrui Yang, Jingyang Zhang, Hongliang Dong, Nathan Inkawhich, Andrew Gardner, Andrew Touchet, Wesley Wilkes, Heath Berry, and Hai Li. Dverge: diversifying vulnerabilities for enhanced robust generation of ensembles. *Advances in Neural Information Processing Systems*, 33:5505–5515, 2020. 3

Zhuolin Yang, Linyi Li, Xiaojun Xu, Shiliang Zuo, Qian Chen, Pan Zhou, Benjamin Rubinstein, Ce Zhang, and Bo Li. Trs: Transferability reduced ensemble via promoting gradient diversity and model smoothness. *Advances in Neural Information Processing Systems*, 34:17642–17655, 2021. 3

Shaofeng Zhang, Meng Liu, and Junchi Yan. The diversified ensemble neural network. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546. 3, 6

Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models. In *Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

Tianyi Zhou, Shengjie Wang, and Jeff A Bilmes. Diverse ensemble evolution: Curriculum data-model marriage. *Advances in Neural Information Processing Systems*, 31, 2018. 3

A Experimental Setup

A.1 Datasets

In our experiments, we evaluate the proposed method across five datasets that span both fine-grained and coarse-grained visual domains. This allows for a comprehensive assessment of our model's performance under different concept supervision settings, including both human-annotated concepts and automatically discovered concepts.

Animals with Attributes 2 (AwA2) (Xian et al., 2019). The AwA2 dataset is a widely used benchmark for zero-shot learning and concept-based models. It consists of 37,322 images spanning 50 animal classes. The dataset provides 85 human-annotated binary attributes (e.g., "stripes", "water", "domestic"). We directly utilize these expert-defined attributes as our ground-truth concept space for training, following the standard train-test splits provided in the original benchmark.

Caltech-UCSD Birds-200-2011 (CUB) (Wah et al., 2011). The CUB dataset is a fine-grained image classification dataset containing 11,788 images across 200 bird species. It includes dense annotations with 312 binary attributes representing specific bird parts and characteristics (e.g., "wing color", "beak shape"). Similar to AwA2, these human-annotated attributes serve directly as the concept-level supervision in our experiments, providing a high-dimensional and highly granular concept space.

CIFAR-10 (Krizhevsky, 2009). CIFAR-10 contains 60,000 images across 10 coarse-grained object classes. Because CIFAR-10 lacks explicit concept annotations, we construct the concept space using the automatic, label-free concept discovery pipeline proposed by Oikarinen et al. (2023). This involves using a Large Language Model to generate a comprehensive candidate set of descriptive concepts for each class, which are then filtered and aligned using a pretrained CLIP model to ensure the concepts accurately reflect the visual semantics of the images.

HAM10000 (Tschandl et al., 2018). HAM10000 is a medical dataset that consists of 10,015 dermatoscopic images of pigmented skin lesions across 7 diagnostic categories. Since explicit concept labels are unavailable, we construct concepts following the method in Chowdhury et al. (2024). Specifically, we start with the class-specific concept pool Chowdhury et al. constructed using GPT-4. To identify a compact and highly discriminative subset of concepts, we apply a selection procedure: for each of the 7 task classes, candidate concepts are ranked using a t-test on their CLIP similarity margins computed on the training set. We select the top 20 concepts per class and take the union across all classes, leading to 139 unique binary concepts.

CelebFaces Attributes (CelebA) (Liu et al., 2015). CelebA is a large-scale face attributes dataset containing over 200,000 celebrity images. Following the experimental setup in prior concept-based learning research (Espinosa Zarlenga et al., 2022), we filter the dataset to focus on the eight most balanced facial attributes and use them to construct both concept and task supervision. Specifically, the first six attributes are used as the binary concepts. To ensure a structured target space with sufficient label diversity, we use all eight attributes to define $2^8 = 256$ composite classes.

A.2 Configurations

We use a pretrained ViT-S/16 as the backbone, except for the HAM10000 dataset where we use a medical ViT-B/16 (Eslami et al., 2023). For all methods, the concept predictor $\{h_{m,j}\}_{j=1}^{p}$ and label classifier f_m are implemented as linear layers. We set $\lambda = 1$ to achieve a balance between concept and task accuracy and train the model using the AdamW optimizer with learning rate in $\{0.0001, 0.0003, 0.00075\}$ and weight decay in $\{0, 0.0004\}$ on a single L40 GPU.

For our method, we insert LoRA modules into the Q, K, V, and projection matrices of every attention block, with scaling multiplier $s = 2$, dropout ratio=0.1, and a rank of $r = 8$ for AwA2, CIFAR-10 and CelebA dataset. We set $r = 16$ and $r = 32$ for HAM10000 and CUB dataset respectively to increase the capacity of

adapters. All backbones are frozen, and only the adapter parameters and concept heads are updated during training.

Another hyperparameter we have is α (see Eq. 2). We dynamically update α during training based on the gradient magnitude. Specifically, let $g^{(t)} = \mathbb{E}_{w \in \mathcal{P}}\left[|\nabla_w \mathcal{L}_{\text{total}}^{(t)}|\right]$ denote the average absolute gradient across all parameters $\mathcal{P}$ in the concept heads. To ensure the regularization strength remains within a valid range $[\alpha_{\min}, \alpha_{\max}]$, we map the gradient using a shifted and scaled sigmoid function σ:

$$\alpha^{(t+1)} = \alpha_{\min} + 2(\alpha_{\max} - \alpha_{\min})\left(\sigma(g^{(t)}) - 0.5\right).$$

$\alpha_{\max}$ and $\alpha_{\min}$ are tunable hyperparameters to further control the strength of diversity regularization. Intuitively, in early stages of training, the actual α will be close to $\alpha_{\max}$ and gradually decay to $\alpha_{\min}$ as the model converges. In our implementation, we tune $\alpha_{\max}$ in $\{0.75,1\}$ and $\alpha_{\min}$ in $\{0.25,0.5,0.75\}$.

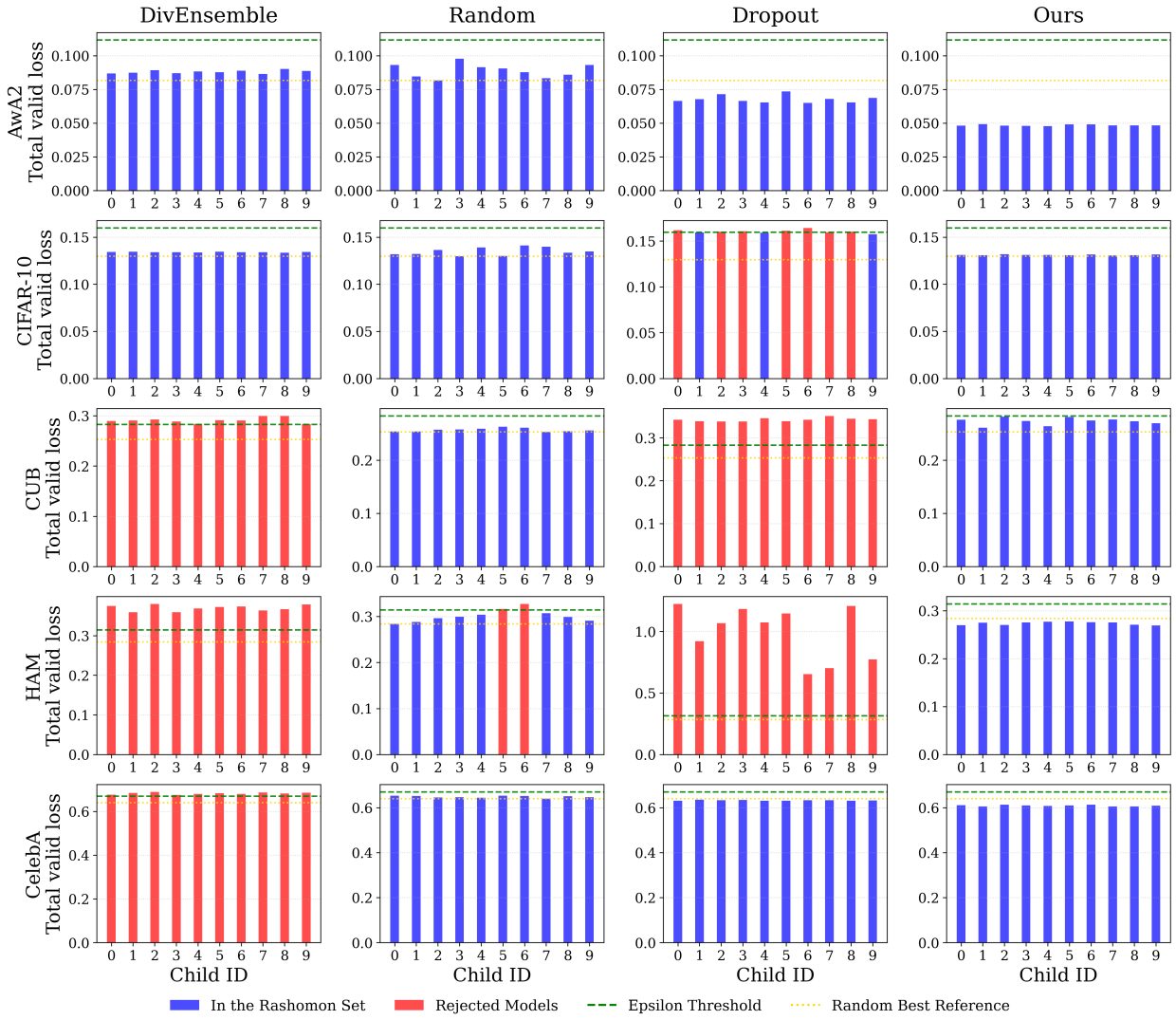


Figure 8: Identification of Rashomon Set Membership across Five Datasets. We evaluate whether models trained with different methods fall within the Rashomon set (blue bars) or are rejected (red bars) based on the ε-threshold (green dashed line) on the validation set. The reference performance (yellow dotted line) is determined by the best model on the validation set from the random initialization baseline.

B Additional Quantitative Results

B.1 Rashomon Set Membership of Rashomon slice and Baselines

In this section, we evaluate whether the models discovered via different methods successfully stay within the ε-Rashomon set as defined in Definition 1. We select the model from the random initialization baseline that achieves the lowest validation misclassification error-based total loss (see Eq. 1) to serve as a reference model. We set $\varepsilon = 0.03$, corresponding to an absolute performance tolerance of approximately 1.5% in both task and concept accuracy – a threshold small enough to ensure all models remain competitive with the reference. We identify all models whose validation loss falls within an ε-margin of the reference model's validation total loss as members of the Rashomon set.

Figure 8 shows the membership of the Rashomon set across different methods. The green dashed line denotes the Rashomon bound. Blue bars represent models whose validation loss falls within this ε-margin, and red bars indicate models that exceed it. As shown in Figure 8, models found by our method are usually stay within the Rashomon set across all five datasets.

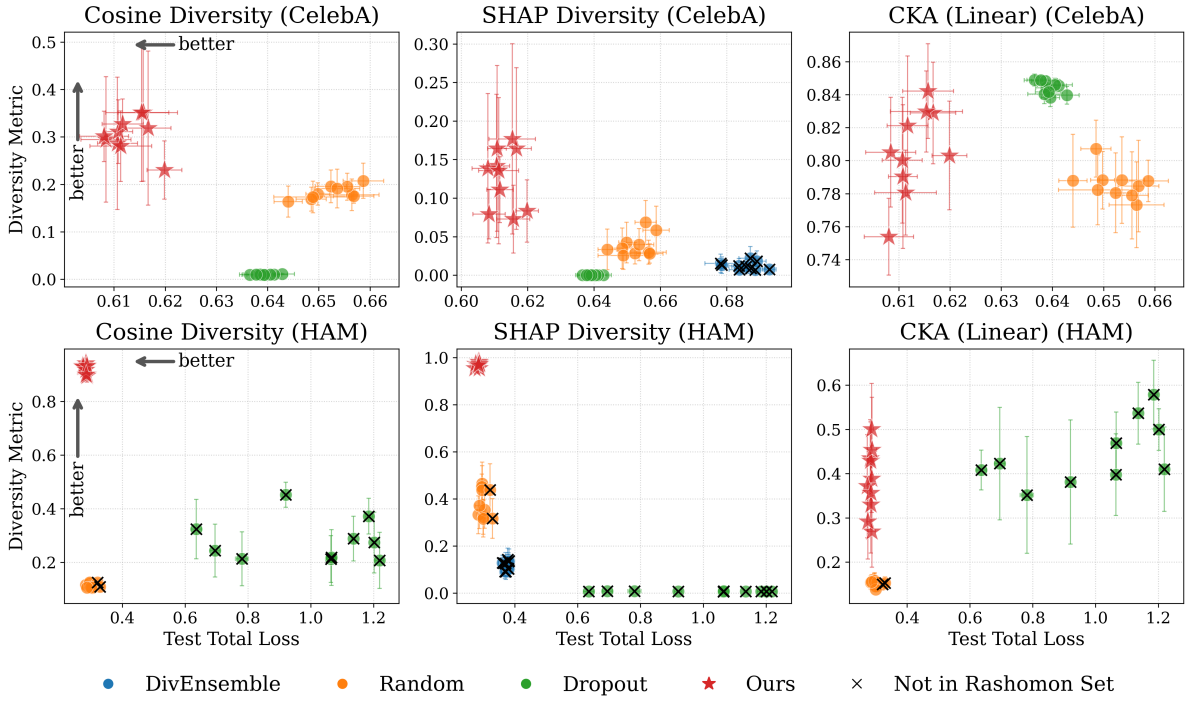


Figure 9: Comparison of model diversity versus total loss on test set across different methods on the CelebA and HAM10000 datasets. The first column displays cosine diversity, the second column displays SHAP diversity, and the last column shows CKA diversity. Each point represents an individual model. Black crosses indicate models that do not satisfy the Rashomon set membership criterion on the validation data. Our method achieves higher diversity while remaining within the Rashomon set. Note that Diverse Ensemble baseline is excluded from the plots for the cosine and CKA diversity metrics. For this baseline, a single concept encoder is shared across all m models, resulting in identical concept representations where diversity metrics are trivially equal to 1.

B.2 Diversity Results on More Datasets

Figure 9 illustrates the performance of our proposed method compared to baselines on the HAM10000 and CelebA datasets. Similar to what we observed in Figure 2, models found by our method stay in the top-left region for both CelebA and HAM10000, indicating that our method is able to find models that achieve both

low test total loss and higher diversity measured by cosine similarity, SHAP, and CKA compared to baselines. Note that because DivEnsemble baseline has a shared concept encoder for all m models, its cosine similarity and CKA similarity are trivially equal to 1. Therefore, we omit these metrics for DivEnsemble.

C Efficiency Analysis

Table 4: Exact memory and parameter costs of different methods and their ratios relative to the random initialization across four datasets. Our method achieves strong diversity with substantially lower memory and parameter cost than baselines. Lower is better for all metrics.

	Method	Mem (GiB) ↓	Mem ratio (%) ↓	Params (M) ↓	Param ratio (%) ↓
CUB	random init	36.01	-	217.3	-
	dropout	4.15	11.5%	21.8	10.0%
	Diverse Ensemble	3.75	10.4%	21.9	10.1%
	Ours	**2.81**	**7.8%**	**2.9**	**1.3%**
CIFAR-10	random init	37.09	-	217.2	-
	dropout	4.16	11.2%	21.8	10.0%
	Diverse Ensemble	3.75	10.1%	21.7	10.0%
	Ours	**2.81**	**7.6%**	**2.9**	**1.3%**
AwA2	random init	37.09	-	217.0	-
	dropout	4.16	11.2%	21.7	10.0%
	Diverse Ensemble	4.66	12.5%	21.7	10.0%
	Ours	**2.81**	**7.6%**	**2.7**	**1.2%**
CelebA	random init	37.09	-	216.7	-
	dropout	4.16	11.2%	21.7	10.0%
	Diverse Ensemble	3.75	10.1%	21.7	10.0%
	Ours	**2.80**	**7.6%**	**2.3**	**1.1%**
HAM	random init	27.19	-	875.6	-
	dropout	2.80	10.3%	87.56	10.0%
	Diverse Ensemble	2.54	9.4%	87.57	10.0%
	Ours	**1.60**	**5.8%**	**12.27**	**1.4%**

C.1 Comparison of Memory Use and Trainable Parameters

Table 4 shows the memory usage and number of trainable parameters for each method across all datasets. Our method requires fewer than 2% of the trainable parameters compared to the random initialization baseline. Our method consistently achieves the lowest memory among all methods and remains substantially below that of random initialization. Together with the strong accuracy and diversity results reported earlier, these findings show that our method achieves accuracy, diversity, and efficiency in both memory usage and trainable parameters.

C.2 Memory Ablations

In Section 4.3, we discussed the memory reduction from LoRA and model-axis checkpointing on AwA2 dataset. Our results generalize beyond it to CUB dataset, as we show in Table 5. Checkpointing alone reduces peak memory by 82%, LoRA alone reduces it by 35%, and combining both reduces it by 93%. These results show that LoRA and checkpointing reduce different components of memory (parameters and optimizer states versus

activations) and their effects accumulate. This reduction is architecture-level rather than dataset-specific, and we expect the same trend to hold for other datasets.

Table 5: Peak GPU memory under different combinations of LoRA and checkpointing on CUB. In our method, checkpointing and LoRA are designed to reduce different components of the memory footprint, and combining them gives the largest memory reduction.

Variant	no-LoRA/no-ckpt	no-LoRA/ckpt	LoRA/no-ckpt	**LoRA/ckpt**
Mem (GiB)	37.75	6.84	24.75	**2.81**
% reduced	0%	82%	35%	**93%**

C.3 Wall-Clock running time

We further evaluate whether the memory savings in Section 4.3 come at the cost of additional per-epoch runtime, comparing wall-clock time and convergence epochs against random initialization. Table 6 shows that the per-epoch time of our method is comparable to random initialization e.g., on AwA2, 833s vs. 843s), with moderate overhead on CUB and CIFAR-10 (ratios of 1.35 and 1.36). Our method requires more epochs to converge, which is consistent with the observation that LoRA-based adaptation trains slower than full fine-tuning (Biderman et al., 2024). However, memory is the fundamental bottleneck for exploring the Rashomon slice of CBMs: training 10 randomly initialized models requires 37 GiB and fits at most 12 models on a single L40 GPU, whereas our method can fit at least 50 models.

D Additional Qualitative Analysis

In this section, we provide additional qualitative analysis on the HAM10000 dataset. Figure 10 visualizes three arbitrary models from the Rashomon slice for the prediction of **Basel Cell Carcinoma (BCC)**, **Benign Keratosis-like Lesion (BKL)**, and **Dermatofibroma (DF)** classes. The results are similar to those from the AwA2 dataset (see Figure 4): different models in the Rashomon slice use distinct combinations of concepts while arriving at the same accurate diagnosis.

Taking dermatofibroma (DF, bottom) as an example, Model 0 is primarily driven by negative evidence. It identifies texture trait "granular/pebbly texture" and size "approximately 20mm", but assigns them negative weights, treating these present concepts as counter-evidence of DF class. Only "pinkish tone" serves as direct positive support. In contrast, Model 1 relies primarily on positive shape concepts such as "linear/streak-like shape", "round/oval shapes", and "small lesion clusters", with a single exclusion cue (absence of "elevated growth") and negative evidence ("no shiny white structures"). Model 2 relies on exclusion reasoning. The absence of pathological features such as "yellow necrosis", "nodular form", and "form asymmetry" serves as the core reasoning, complemented by positive influence of "light brown" and "pinkish color". Overall, the three members span negative-evidence-dominated, positive-shape-driven, and exclusion-based rationales for the DF diagnosis, illustrating that equivalent predictions within a Rashomon slice can emerge from fundamentally different concept-level mechanisms.

E More Results on Abstention Use Case

In cases when CBM is not confident, prior work introduced conceptual safeguards (Joren et al., 2024) that abstain from predicting on high-uncertainty samples and instead seek human intervention to confirm key concepts. Traditional safeguards rely on Monte Carlo sampling from a single model to quantify uncertainty. However, this often reflects local variation around a single rationale. A Rashomon set provides a more natural framework for this task, as it contains multiple, near-optimal, but conceptually distinct rationales.

Therefore, we replace Monte Carlo samples with the Rashomon slice members, treating their disagreement on task predictions as a direct uncertainty signal. When the disagreement rate exceeds the threshold τ, the system abstains and queries an expert for up to B concept values. We then reweight members by their

Table 6: Per-epoch running time and the number of epochs to converge of random initialization vs. our method. The ratio is computed as the per-epoch time of our method divided by that of random initialization. Our method takes comparable per-epoch running time, but requires more epochs to converge.

Metric	**CUB**	**AwA2**	**CIFAR-10**
Per-epoch Time: Random	106s	843s	702s
Per-epoch Time: Ours	143s	833s	954s
Ratio (Ours/Random)	1.35	0.99	1.36
# Epochs: Random	46	34	61
# Epochs: Ours	251	176	290

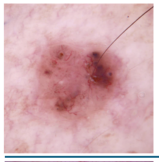
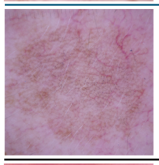
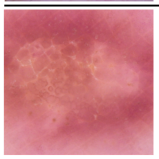
Class		**Model 0**	**Model 1**	**Model 2**
BCC		+ no ulceration or bleeding (-7.69) - irregular pattern/texture (+1.60) - irregular, rough surface (+1.22) + pinkish color (+0.99) + small lesion cluster (+0.96)	- gray region (+1.12) + pinkish tone (+0.91) + no atypical vessel/streak (+0.86) + small lesion cluster (+0.77) + is pink (+0.73)	- yellow necrosis (+2.77) - button/coin-like shape (+1.55) + pink growths (+0.95) - irregular pattern/texture - nodular form (+0.76)
BKL		- yellow tones (+7.46) + no shiny white structures (-0.70) - grey regions (-0.64) - irregular pattern/texture (-0.58) + linear/streak-like shape (-0.57)	- yellow necrosis (+7.39) - nodular form (+0.99) - irregular surface pattern (+0.67) - blue shades (+0.60) - red base (+0.60)	- flat macular spots (+3.12) + lesion ~20mm (+1.01) + flat smooth surface (+0.54) + rough, uneven texture (+0.33) + purple tones (+0.41)
DF		+ granular/pebbly texture (-9.21) - irregular pattern/texture (-1.00) + lesion ~20mm (-0.70) + pinkish tone (+0.64) + is pink (-0.55)	+ linear/streak-like shape (+3.67) + round/oval shapes (+0.74) + small lesion clusters (+0.58) - elevated growth(+0.55) + no shiny white structures (-0.54)	- yellow necrosis (+5.15) - nodular form (+2.87) + light brown (+1.06) + pinkish color (+0.65) - form asymmetry (+0.64)

Figure 10: Top-5 concepts based on contribution scores $W_{p,k}\hat{c}_p$ of each concept p to the prediction of class k (value in parentheses) for Basal Cell Carcinoma (BCC, top), Benign Keratosis-like Lesions (BKL, middle), and Dermatofibroma (DF, bottom) classes from the HAM10000 dataset across three Rashomon slice members. The sign stands for whether the concept is predicted to be present ($+$, $\hat{c}_p > 0$) or absent in the corresponding class ($-$, $\hat{c}_p \leq 0$). All the predictions for concepts shown are correct and correspond to ground truth concept labelings for the classes. Color indicates the decision rationale in which the concept contributes to the prediction: green denotes positive evidence ($W_{p,k} > 0, \hat{c}_p > 0$); red represents negative evidence ($W_{p,k} \leq 0, \hat{c}_p > 0$); blue means excluding evidence ($W_{p,k} > 0, \hat{c}_p \leq 0$); purple stands for spurious evidence ($W_{p,k} \leq 0, \hat{c}_p \leq 0$). The diverse patterns highlight that different members rely on different rationales.

consistency with the expert's confirmed concepts:

$$w_m \propto \exp\left(-\mathcal{L}_C^{(S)}(g_m(\mathbf{x}), \mathbf{c})\right),$$

where S is the set of queried concepts and $\mathbf{c}$ is the expert-confirmed concept vector. We accept a prediction when the weighted consensus exceeds the threshold:

$$\max_y \sum_m w_m \, \mathbb{1}_{[f_m(g_m(\mathbf{x}))=y]} \geq \tau,$$

and otherwise query further concepts or abstain once budget B is exhausted.

In addition to an example shown in Figure 5, we compare our Rashomon-based safeguard against both a single-CBM safeguard and the random initialization baseline on Beaver class from the Awa2 dataset in Figure 11 across various budgets B and thresholds τ. Our method returns a prediction on a larger fraction of test examples without needing to abstain, even at high confidence thresholds (left column). Our models require

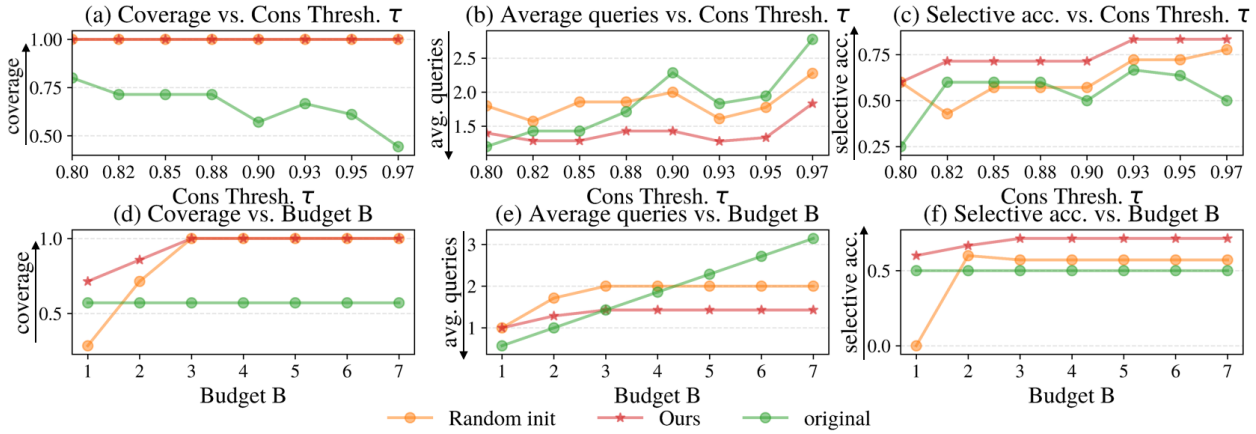


Figure 11: Performance of abstention on beaver class from the AwA2 dataset across varying the confidence thresholds τ (top row) and query budgets B (bottom row) on the coverage (fraction of non-abstained predictions), average number of concept queries, and selective accuracy (accuracy on non-abstained samples). Arrows indicate the preferred direction. The Rashomon-member safeguard achieves higher coverage with fewer queries while improving selective accuracy across settings compared to baselines.

fewer human concept queries to reach a confident consensus (middle column), and for the samples where a prediction is provided, the accuracy is higher than the baselines (right column). Note that Diverse Ensemble can not be applied in this use case because all child models in Diverse Ensemble share the same encoder and output the same concept predictions. Our results indicate that diverse Rashomon members can provide a more precise indicator of uncertainty than sampling from a single model.

F Diversity Within the Rashomon Slice Members is Meaningful

As discussed in the main text, a practically useful Rashomon slice requires models that exhibit diverse reasoning pathways. To further verify the diversity of our models, we first use concept interventions to reveal the large variance in the models' final decision logic across different Rashomon slice members and then use layer-wise eigenvector visualizations to trace the origin of this diversity to the LoRA adapters.

We first demonstrate that members of our Rashomon slice have distinct dependencies on visual features. We perform interventions on the tiger class in AwA2 using eight concepts, specifically: "orange", "stripes", "furry", "big", "paws", "tail", "meatteeth", "claws". For each model m, we intervene on subsets of five concepts at a time by changing the signs of the predicted concept logits and record the resulting accuracy drop over all testing data points belonging to the tiger class. We chose to intervene on five concepts at a time to ensure perturbations are strong enough to produce measurable accuracy differences across models, while leaving enough concepts intact so that models relying on different subsets can respond distinctly.

For a fixed set of concepts we intervene on, the change in accuracy varies wildly across different models m: some models display little or no accuracy change (0% change in accuracy), while others suffer near complete degradation (losing up to 99.5% accuracy). From these results we find that different models in the Rashomon slice can respond differently to the same concept intervention indicating different reasoning pathways exist.

To localize the source of diversity in concept representations, we follow Mühlematter et al. (2026) and quantify weight-space variation via Singular Value Decomposition (SVD). For each transformer block, we compute the cosine similarity between the top-16 right singular vectors of the LoRA-augmented projection matrices across Rashomon slice members. We perform this analysis for the QKV and output projection matrices.

As shown in Figure 12, the leading directions in early layers remain similar across models, while similarity gradually declines with depth and diversity concentrates in the adapters of the deeper blocks. This means

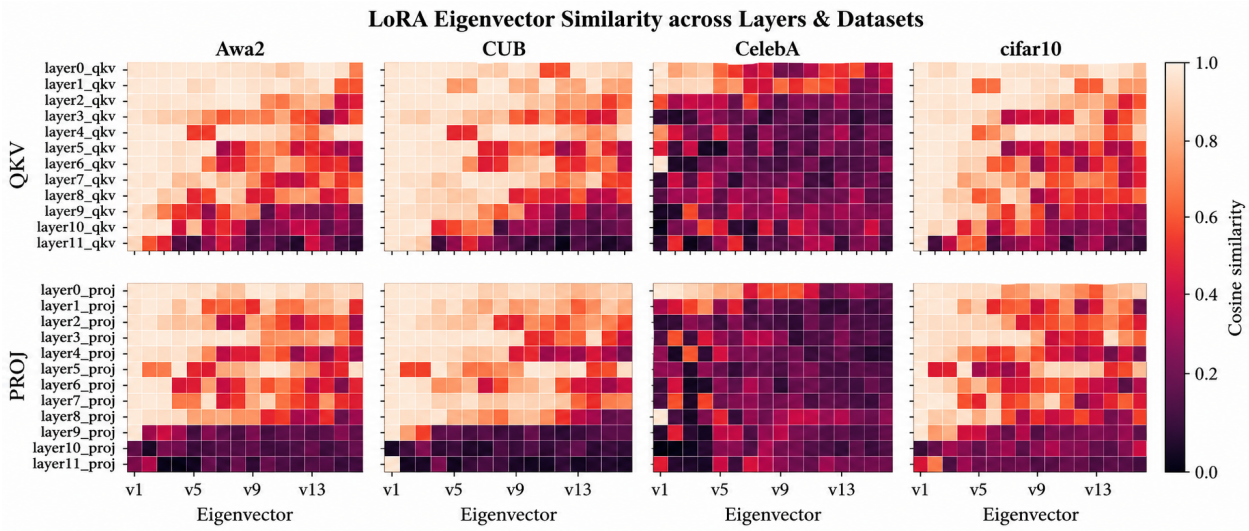


Figure 12: Layerwise eigenvector similarity calculated on the QKV matrices (top) and projection matrices (bottom). Similarity between Rashomon slice members decreases with depth across all four datasets, indicating that diversity concentrates in deeper layers.

that models can learn low-level concepts in the same way, but diverge when learning higher-level concepts, which explains why our framework can achieve a good balance between accuracy and diversity.

Across both datasets and both QKV and projection matrices, for Rashomon slice we observe a clear pattern: similarity decreases noticeably with depth, and within each layer the later eigenvectors show lower similarity than the leading ones. This indicates that Rashomon models share similar low-level representations but diverge increasingly in deeper layers. This structured reduction in similarity reflects the effect of our method in encouraging meaningful representational diversity by using the adapter modules.

G Sensitivity Analysis of the Rashomon Slice Size

In this section, we study the influence of the Rashomon slice size on the model diversity and accuracy. We vary $M \in \{10, 15, 20, 25, 30\}$ for the CIFAR-10 dataset and report diversity metrics and test total loss in Figure 13.

As we can see, the test total loss tends to increase for models obtained at larger values of M. The test total loss for models obtained by $M = 30$ (purple triangles) is shifted toward the right compared to smaller slices. This suggests that as we increase the required number of models, the search process must "reach further" into the loss landscape, incorporating candidates that are further from the global optimum.

The impact of increasing M from 10 to 30 varies significantly across the different diversity metrics. While cosine diversity (left) remains relatively stable, indicating that the fundamental parameter orientations are robust to scaling, SHAP diversity and CKA (middle and right) show more sensitivity. Specifically, for $M = 30$, there is a visible decreasing trend in diversity scores. This indicates that as the size of the Rashomon slice grows, it becomes increasingly difficult to maintain high pairwise diversity, as additional models begin to converge toward a more redundant latent subspace.

Collectively, these results reveal a trade-off between the size of the Rashomon slice and individual model performance. While a smaller size (e.g., $M = 10, 15$) yields a tighter cluster with lower test loss and higher diversity, increasing M beyond 25 could lead to "representation saturation". The divergence between weight-based (Cosine) and representation-based (CKA/SHAP) metrics shows that while the models remain distinct in parameter space, their functional behavior becomes increasingly coupled at higher M. In practice,

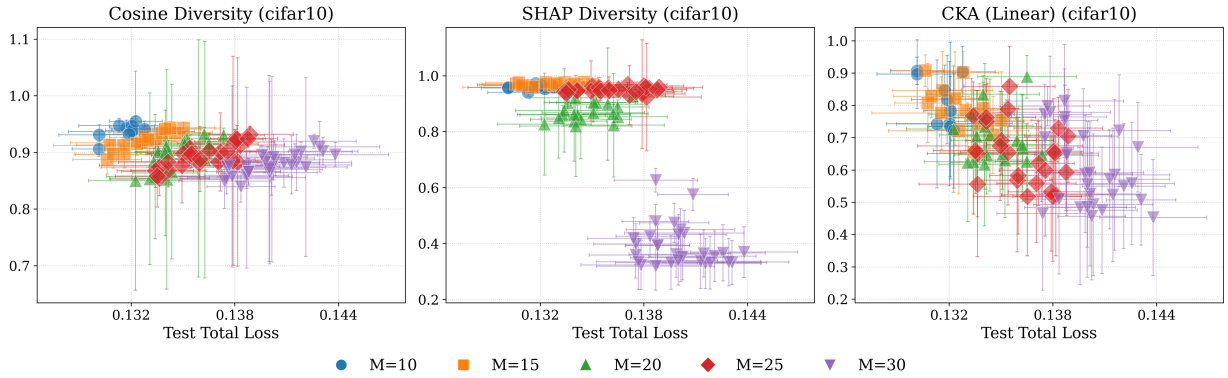


Figure 13: Impact of Rashomon slice size $M \in \{10, 15, 20, 25, 30\}$ on CIFAR-10. As M increases, the test loss for the Rashomon slice members shifts slightly higher. While cosine diversity remains stable, both SHAP and CKA diversity have a visible decrease as M approaches 30. All the models shown are verified members in the Rashomon set.

an optimal configuration must balance the quantity of models discovered with the meaningful diversity and accuracy of each member.

H Sensitivity Analysis of α

To characterize the relationship between diversity and concept accuracy across values of the regularization strength α in Eq. 2, we conduct a grid search over α on the AwA2 and CUB datasets using $M = 5$ models, with all other hyperparameters and training process the same as in Section 4. We evaluate the effect of varying α on concept accuracy and three diversity metrics: cosine similarity, SHAP diversity, and CKA. Task accuracy remains stable across models and values of α due to the minimax objective in Eq. 2, so the analysis focuses on the interplay between concept accuracy and diversity.

The results are presented in Figure 14. Increasing α generally improves diversity at some cost to concept accuracy, though the sharpness of this trade-off varies by dataset and metric. On CUB, the frontier is sharper under cosine distance, with high α values risking Rashomon set membership, while on AwA2 the trade-off is softer and higher diversity can often be achieved with little cost to concept accuracy. Overall, in our experiments, values of $\alpha \approx 0.5$ generally provided a good balance between diversity and concept accuracy while keeping models within the Rashomon set. This motivates our dynamic α schedule, which is initialized around this range and adapts during training without requiring expensive per-dataset grid searches.

I Optimization Stability

In this section, we report the training curves of all Rashomon slice members during optimization. Figure 15 shows that the training losses of all members converge smoothly with no oscillation or collapse, confirming that our optimization remains stable in practice.

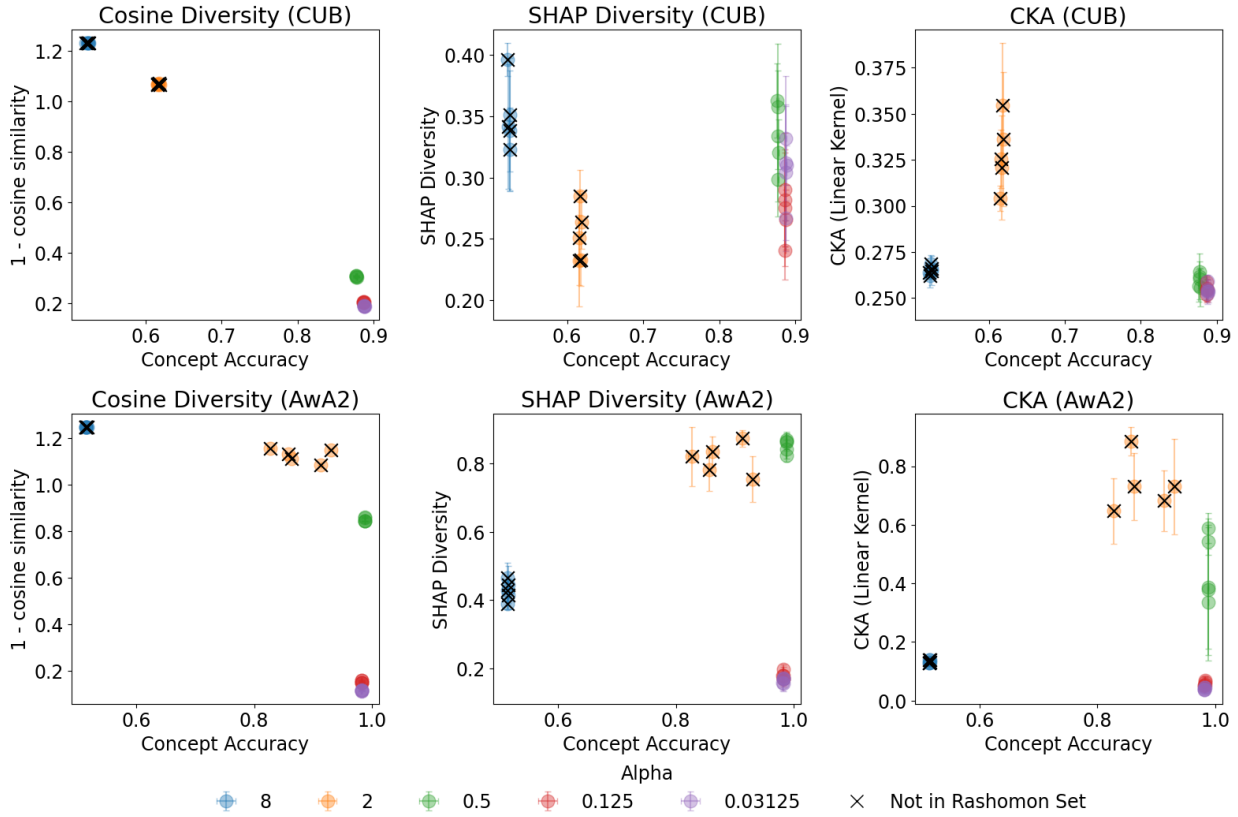


Figure 14: Diversity measures versus concept accuracy for different values of α on CUB (top) and AwA2 (bottom). Crosses indicate models outside the Rashomon set. The relationship between diversity and concept accuracy is metric- and dataset-dependent, with CUB showing a sharper frontier under cosine distance than AwA2.

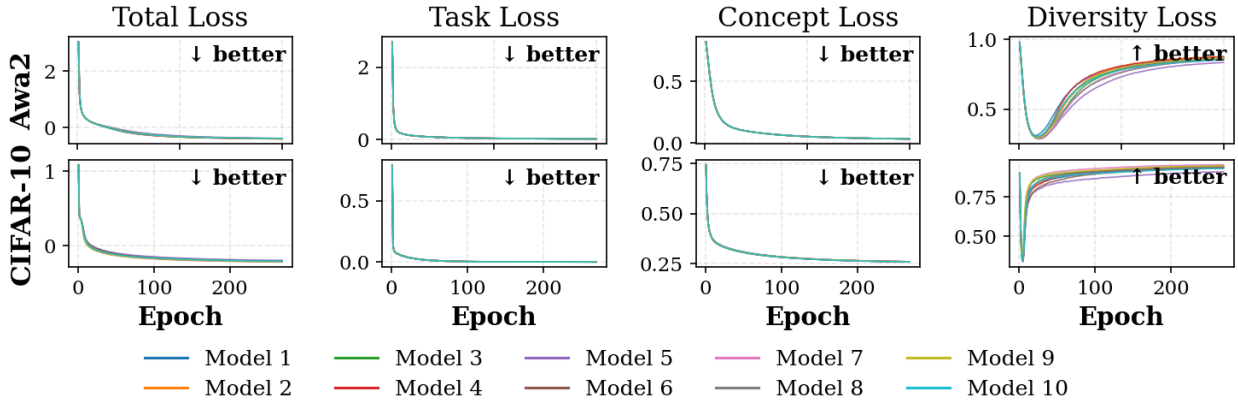


Figure 15: Training dynamics of all Rashomon slice members during optimization on CIFAR-10 and AwA2. Total, task, concept, and diversity losses all converge smoothly without oscillation or collapse, indicating stable training in practice.