CONCEPT ENVELOPE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Pretrained embeddings can be highly predictive while encoding task-relevant concepts in a form that is not readable as a single direction. Yet concept-bottleneck and dictionary-score methods often expose each named concept through a scalar score, which can obscure semantic structure spread across multiple embedding directions. Motivated by the Linear Representation Hypothesis and semantic structure in embedding space, we introduce Concept Envelope, which connects predictive sufficiency with concept-based interpretability and offers a way to retain task-related semantic variation beyond prediction alone. Concept Envelope is a supervised neural method that jointly learns a concept subspace and a nonlinear predictor with soft covariance separation, followed by post-hoc semantic grounding. Across seven visual benchmarks, it maintains competitive predictive accuracy. Two quantitative interpretability metrics show retention of independently defined semantic directions, including those missed by accurate prediction-only subspaces, and transfer of estimated edits to held-out intervention pairs. Controlled language tasks further show that dictionary scores can follow surface wording rather than latent state.

1 INTRODUCTION

Pretrained model embeddings are widely used as features for downstream prediction. A fixed embedding vector X can contain rich semantic information for predicting a response Y, as illustrated by transfer performance and linear-probe evaluations (Radford et al., 2021; Alain & Bengio, 2016). Within such an embedding, only a subset of the encoded semantic factors may be relevant for a given downstream response. However, the coordinates of X are not named concepts. A predictor trained on X may perform well without making clear which semantic factors in the embedding are used for prediction. Our goal is to learn a low-dimensional, task-relevant subspace of the embedding that retains the semantic variations and can be interpreted at the concept level.

Concept-based interpretability provides a natural interface for this goal. Rather than explaining a prediction through raw embedding coordinates, concept methods describe behavior using human-level variables such as attributes, topics, or visual properties (Kim et al., 2018; Koh et al., 2020). The Linear Representation Hypothesis suggests that semantic changes are encoded by approximately linear directions or subspaces (Park et al., 2024; 2025; Engels et al., 2025). When a concept spans several directions, a single scalar score can obscure its internal geometry. Empirical studies further indicate that distinct semantic factors are often approximately orthogonal in the native Euclidean geometry of pretrained representations, including language-model concept directions and visual attribute representations (Park et al., 2024; Stein et al., 2024).

Given this special geometry, **our insight is that a statistical envelope provides a natural framework for this joint goal of prediction and concept interpretability.** Envelope methods were originally developed to improve parameter estimation efficiency in linear models (Cook et al., 2010). An envelope contains the predictive information while keeping variation within it uncorrelated with variation in its orthogonal complement (Cook et al., 2013; Cook & Zhang, 2015). In pretrained representations, linear semantic structure gives this separation a concept-level meaning. When a predictive score mixes several concepts (Figure 1), the retained and residual components can remain correlated because both contain variation from those concepts. Covariance separation favors retaining this associated variation together, allowing the subspace to preserve concept structure beyond what prediction alone requires.

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

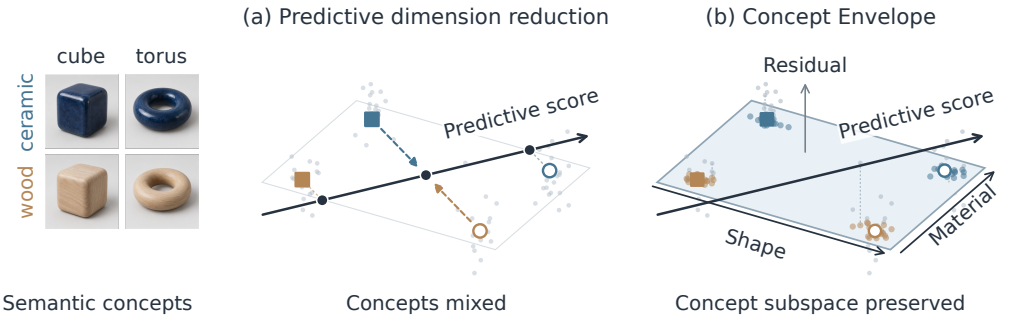


Figure 1: An illustrative shape–material example. Shape and material vary independently. (a) Two distinct combinations project to the same predictive score. (b) The envelope retains the concept subspace, preserving both concepts and the predictive direction while separating independent residual variation. Marker shape and color encode shape and material; gray dots show ambient variation.

We introduce *Concept Envelope*, a response-supervised neural method that implements this perspective in fixed pretrained embeddings. It learns a linear subspace jointly with a nonlinear predictor, using a soft covariance-separation penalty to accommodate the approximate semantic geometry of the embedding. The training objective uses the task response Y rather than concept labels. Concept evidence is used after fitting to select a candidate subspace and ground it in a human-readable coordinate system.

Our contributions. First, we connect statistical envelopes with concept-interpretable learning in pretrained representations. Under a linear semantic decomposition, we explain when covariance separation expands a prediction-sufficient subspace to retain the associated concept variation. This gives envelopes a new interpretation beyond their traditional role in parametric estimation efficiency, with a target distinct from predictive compression and predefined concept scores. Second, based on this insight, we develop a response-supervised neural estimator that jointly learns the subspace and a nonlinear predictor through soft covariance separation. Third, across seven visual benchmarks, we show competitive prediction and evaluate concept preservation through retention of independently defined semantic directions and transfer to held-out intervention pairs. Constructed day and month tasks further show how one-dimensional dictionary scores can follow surface wording rather than latent state. Fourth, we derive an oracle rate for an idealized concept-dictionary estimator under semantic sparsity, formalize the prediction–preservation gap, and establish conditions under which covariance separation selects the full active concept subspace.

Relation to prior work. Concept-based methods such as concept activation vector probes, concept bottleneck models, and post-hoc concept bottleneck models provide concept-level probes or predictors, typically using concept examples, concept labels, or an explicit concept layer (Kim et al., 2018; Koh et al., 2020; Yuksekgonul et al., 2023). In Concept Envelope, the task response is used to learn the subspace, while concept evidence enters only afterward, for subspace selection and interpretation. Sparse autoencoders and sparse concept decompositions instead seek global features or dictionaries in dense embeddings; our object is a compact subspace tied to a particular downstream response (Huben et al., 2024; Bhalla et al., 2024).

On the statistical side, sufficient dimension reduction estimates response-relevant subspaces, including neural methods that learn the reduction jointly with a predictor (Kapla et al., 2022; Huang et al., 2024; Tang & Li, 2026). Envelope methods additionally impose covariance separation between the retained subspace and its complement, with most formulations developed for linear models (Cook et al., 2010; 2013; Cook & Zhang, 2015). The closest nonlinear envelope formulation is the enveloped central space of Ma et al. (2024), defined through conditional sufficiency and independence. They estimate the subspace using the generalized method of moments together with smoothing-kernel regression or density estimation before downstream prediction. As a result, their approach is infeasible for a high-dimensional pretrained embedding, say $p = 4096$. By contrast, we learn the subspace jointly with a neural predictor from a pretrained embedding, with covariance separation regularization incorporated into the training objective to facilitate differentiable learning.

2 CONCEPT ENVELOPE

Let $X \in \mathbb{R}^d$ be an embedding from a fixed pretrained model and Y the downstream response.

Geometric motivation. To explain the connection between envelopes and concept interpretation, we consider an idealized semantic decomposition. Let $C \in \mathbb{R}^s$ collect task-related semantic variables, and let $A \in \mathbb{R}^{d\times s}$ have orthonormal columns spanning their embedding space (Stein et al., 2024). A named concept may occupy several coordinates of C. Consider

$$X = AC + R, \qquad Y = h(B^\top C) + \varepsilon, \qquad B \in \mathbb{R}^{s\times q}, \qquad q < s,$$

where R is independent of C and lies orthogonal to $\mathrm{span}(A)$, and ε is response noise.

For prediction, $\mathcal{M} = \mathrm{span}(AB)$ is sufficient, since its projection determines $B^\top C$. However, its projection removes directions $A\delta$ for which $B^\top \delta = 0$. Changes in these directions leave the predictive summary unchanged, although they may change the constituent concepts. Preserving the concept space therefore asks more than retaining a sufficient score. Proposition 1 in Section 4 formalizes this distinction through the prediction–preservation gap.

An envelope connects these two targets through covariance structure. For $\Sigma_X = \mathrm{Cov}(X)$, a subspace with orthogonal projector P is "reducing" when $P\Sigma_X(I_d - P) = 0$. This means that the retained and residual projections are uncorrelated. The Σ_X-envelope of $\mathcal{M}$ is the smallest such subspace containing $\mathcal{M}$. Under the decomposition above, $\mathcal{S}_C$ is reducing and contains $\mathcal{M}$, so the envelope lies within $\mathcal{S}_C$. It can retain additional semantic directions around the predictive core without adding the independent residual R. More rigorously, the envelope equals $\mathcal{S}_C$ when no proper reducing subspace within $\mathcal{S}_C$ contains $\mathcal{M}$. Appendix B.4 gives a condition in terms of the covariance of C and the predictive directions B. Under the assumptions of Proposition 2, fixing the retained dimension at s then allows covariance separation to select the full concept space among equally predictive subspaces.

Prediction identifies response-relevant variation, while covariance separation favors retaining the semantic variation associated with it. The target can therefore be larger than the most compressed sufficient representation. It also differs from a predefined concept-score bottleneck: the subspace is learned before its coordinates are grounded. Subspace preservation alone does not name or disentangle concepts. We use a soft covariance penalty to accommodate the approximate linear and orthogonal geometry of pretrained embeddings.

Subspace parameterization. We seek a u-dimensional task subspace $\mathcal{S} \subset \mathbb{R}^d$, with $u \ll d$. Let $\Gamma \in \mathbb{R}^{d\times u}$ be a semi-orthogonal basis for $\mathcal{S}$, so that $\Gamma^\top\Gamma = I_u$, and define

$$P_\Gamma = \Gamma\Gamma^\top, \qquad P_\Gamma^\perp = I_d - P_\Gamma, \qquad Z_\Gamma = \Gamma^\top X.$$

Here $Z_\Gamma \in \mathbb{R}^u$ is the retained coordinate representation used by the predictor.

The estimand is the subspace, not a particular basis. For any orthogonal $Q \in \mathbb{R}^{u\times u}$, the bases Γ and ΓQ define the same projection P_Γ. Accordingly, we interpret the learned object through $\mathrm{span}(\Gamma)$ or P_Γ; different bases within the learned subspace are chosen only for post-hoc interpretation.

Population objective. Let ℓ denote the task loss, and let $\mathcal{F}_u$ denote a predictor class on $\mathbb{R}^u$; in experiments we use a neural-network class. For $f \in \mathcal{F}_u$, write the population risk as

$$R(\Gamma, f) = \mathbb{E}\{\ell(Y, f(\Gamma^\top X))\}.$$

We penalize covariance between the retained and residual projections to encourage the separation described above. The penalty is the squared Frobenius norm of the off-diagonal covariance block induced by the decomposition $\mathbb{R}^d = \mathrm{span}(\Gamma) \oplus \mathrm{span}(\Gamma)^\perp$:

$$\mathcal{C}(\Gamma) = \left\|\mathrm{Cov}(P_\Gamma X, P_\Gamma^\perp X)\right\|_F^2 = \left\|P_\Gamma \Sigma_X P_\Gamma^\perp\right\|_F^2, \qquad \Sigma_X = \mathrm{Cov}(X). \tag{1}$$

The block captures cross-covariance left when predictive structure is split across the retained subspace and its complement, so penalizing it discourages such partial retention. It is a second-order separation criterion, weaker than independence, and does not by itself guarantee semantic disentanglement or concept recovery.

The population objective trades off prediction risk and covariance separation through $\lambda \geq 0$:

$$\mathcal{L}(\Gamma, f) = R(\Gamma, f) + \lambda\, \mathcal{C}(\Gamma). \tag{2}$$

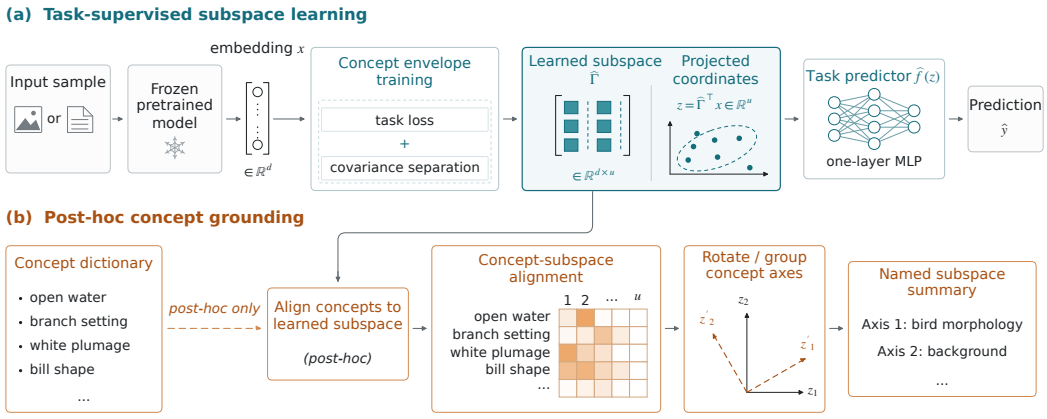


Figure 2: Concept Envelope separates subspace estimation from post-hoc grounding. The estimator learns $\widehat{\Gamma}$ from task loss and a covariance-separation penalty. After estimation, post-hoc grounding uses concept evidence to choose a reporting rotation and summarize the retained subspace.

A minimizer over semi-orthogonal Γ and $f \in \mathcal{F}_u$ defines a population concept envelope:

$$(\Gamma^\star, f^\star) \in \arg\min_{\Gamma^\top\Gamma = I_u,\ f\in\mathcal{F}_u} \mathcal{L}(\Gamma, f). \tag{3}$$

When $\lambda = 0$ the objective reduces to prediction-only supervised subspace learning.

Neural estimator and post-hoc interpretation. Given data $\{(X_i, Y_i)\}_{i=1}^n$, let $\bar{X} = n^{-1}\sum_{i=1}^n X_i$ and $\widehat{\Sigma}_X = n^{-1}\sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top$. We estimate the concept envelope by

$$(\widehat{\Gamma}, \widehat{f}) \in \arg\min_{\Gamma^\top\Gamma = I_u,\ f\in\mathcal{F}_u} \left[\frac{1}{n}\sum_{i=1}^n \ell\{Y_i, f(\Gamma^\top X_i)\} + \lambda\widehat{\mathcal{C}}(\Gamma)\right], \qquad \widehat{\mathcal{C}}(\Gamma) = \left\|P_\Gamma \widehat{\Sigma}_X P_\Gamma^\perp\right\|_F^2, \tag{4}$$

where $\widehat{\mathcal{C}}(\Gamma)$ is the empirical covariance-separation penalty. The tuning parameters are the retained dimension u and the covariance-separation weight λ.

Figure 2 shows the two stages. Once the subspace has been estimated, the grounding stage interprets the retained coordinates using concept evidence, choosing an orthogonal reporting rotation for display.

3 Experiments

The experiments evaluate whether Concept Envelope keeps accuracy competitive while learning an interpretable low-dimensional subspace of the frozen Qwen3-VL-Embedding-8B embedding (Li et al., 2026). Inputs and concept-dictionary atoms come from the same model, so every method is compared in one embedding space. We first compare test accuracy on seven visual benchmarks, then measure CAV Retention on three of them and Edit Fit on the two with source-paired interventions, and finally use constructed day and month tasks to test whether dictionary scores preserve a latent state when its surface description changes.

3.1 Visual Benchmarks

Table 1 reports test accuracy on Waterbirds, CelebA, COCO-BAM, CIFAR-10, CIFAR-100, CUB, and FunnyBirds (Sagawa et al., 2020; Liu et al., 2015; Yang & Kim, 2019; Krizhevsky, 2009; Wah et al., 2011; Hesse et al., 2023). The comparison covers unsupervised representations, classical and neural sufficient dimension reduction, envelope-related methods, and concept bottlenecks. Appendix A.2 describes each family and its relation to Concept Envelope. The full-embedding MLP serves as an accuracy reference.

A compact subspace remains competitive with the full embedding. Across the seven benchmarks, Concept Envelope at $\lambda > 0$ remains competitive with the full-embedding MLP, and adding covariance

Table 1: Predictive accuracy across seven benchmarks.

		Waterbirds		CelebA		COCO-BAM		CIFAR		CUB	Funny Birds
Family	Method	Avg	WG	Avg	WG	Avg	WG	10	100	Avg	Avg
Unsupervised representations	PCA + MLP	68.96 (0.42)	10.84 (0.12)	94.83 (0.06)	32.89 (4.01)	79.75 (0.16)	57.80 (0.40)	90.55 (0.08)	65.06 (0.29)	79.12 (0.25)	39.20 (1.23)
	SAE + MLP	92.16 (0.21)	77.13 (1.34)	94.50 (0.18)	36.22 (2.39)	76.30 (1.26)	49.00 (4.69)	92.84 (0.15)	75.27 (0.18)	82.96 (0.18)	36.72 (0.69)
Classical SDR	SIR + MLP	77.84 (0.00)	0.00 (0.00)	95.24 (0.02)	29.22 (2.60)	91.35 (0.20)	78.60 (2.58)	93.94 (0.07)	57.34 (0.17)	48.32 (1.71)	81.04 (0.77)
	SAVE + MLP	77.84 (0.00)	0.00 (0.00)	92.39 (2.68)	9.78 (5.40)	50.45 (0.95)	13.80 (16.17)	76.28 (1.15)	17.96 (0.75)	1.26 (0.17)	25.72 (3.32)
Envelope-related methods	PLS + MLP	76.59 (0.58)	49.44 (0.53)	95.21 (0.04)	34.89 (2.71)	89.85 (0.37)	74.00 (1.10)	92.86 (0.05)	69.11 (0.21)	80.99 (0.25)	59.32 (1.14)
Neural SDR (nonlinear)	Deep SDR	92.97 (0.39)	80.93 (1.35)	95.25 (0.06)	39.56 (0.65)	90.90 (0.98)	78.00 (3.41)	93.90 (0.06)	73.30 (0.39)	81.70 (0.39)	73.64 (0.37)
Concept bottlenecks	PCBM	80.12 (0.45)	49.94 (0.94)	94.61 (0.09)	35.89 (4.85)	89.60 (0.73)	76.40 (2.94)	91.00 (0.12)	71.00 (0.34)	69.61 (0.26)	72.96 (1.12)
	PCBM-h	92.60 (1.59)	81.16 (1.11)	95.29 (0.03)	40.78 (3.89)	91.15 (0.37)	79.20 (1.17)	93.90 (0.20)	77.94 (0.16)	82.29 (0.12)	79.32 (1.16)
Concept Envelope	CE, $\lambda = 0$	92.78 (0.88)	74.86 (2.13)	95.29 (0.08)	46.11 (3.14)	89.95 (0.62)	75.80 (2.23)	93.93 (0.13)	76.09 (0.39)	81.44 (0.51)	76.12 (1.10)
	CE, $\lambda > 0$	92.45 (1.23)	71.15 (0.75)	95.29 (0.04)	46.22 (3.47)	91.30 (1.94)	79.60 (5.82)	93.74 (0.13)	75.87 (0.19)	81.93 (0.57)	77.28 (1.75)
Full-embedding reference	MLP	93.60 (0.17)	84.52 (0.25)	93.98 (0.15)	41.33 (4.67)	90.50 (0.16)	77.00 (0.63)	93.87 (0.05)	76.18 (0.15)	82.50 (0.13)	79.88 (1.29)

Accuracy (%): mean over five seeds, with population SD below. Avg / WG: average / worst-group accuracy; CIFAR columns report average accuracy. CE: Concept Envelope ($\lambda = 0$: prediction-only linear reduction).

separation does not systematically reduce average accuracy relative to $\lambda = 0$. Other compact reductions are also competitive. Concept Envelope has the highest worst-group accuracy on CelebA, while on Waterbirds it remains below PCBM-h, Deep SDR, SAE, and the full-embedding MLP. It is more accurate than the pure concept-score predictor PCBM on every benchmark. PCBM-h also reads the full embedding through a residual predictor, so its accuracy does not establish that the concept scores alone are sufficient. By contrast, Concept Envelope makes all predictions from the learned low-dimensional subspace.

3.2 Concept Structure in the Learned Subspace

After a post-hoc rotation of the two learned coordinates, held-out Waterbirds examples organize by bird morphology along one axis and by background along the other (Figure 3). A single score cannot represent this structure: the two factors stay in the predictive subspace as separate directions instead of collapsing into one.

Two metrics measure how much of that structure each subspace keeps. *CAV Retention* measures how far an independently defined semantic direction aligns with the retained subspace: the squared cosine between that subspace and a label-defined concept activation vector v_c, fitted once in the frozen embedding (Kim et al., 2018). *Edit Fit* measures whether an edit estimated from one set of paired interventions transfers to held-out pairs of the same type. It aggregates the held-out fit of each directed transition, weighted by the fraction of that transition's ambient edit energy retained by the subspace; cells below 1% retained edit energy are not reported (Appendix A.3).

CAV Retention is evaluated on Waterbirds, COCO-BAM, and FunnyBirds. Edit Fit uses source-paired object or scene replacements on COCO-BAM and single-part interventions on FunnyBirds (Appendix A.3). Figure 4 compares methods that return a fixed linear subspace of the same embedding, with dimensions matched to Concept Envelope where the method's rank permits. The training objective uses only task labels. The reported u and λ are fixed from dataset-specific pilot studies before the formal five-seed evaluation; Appendix A.4 describes the pilot-selection protocol.

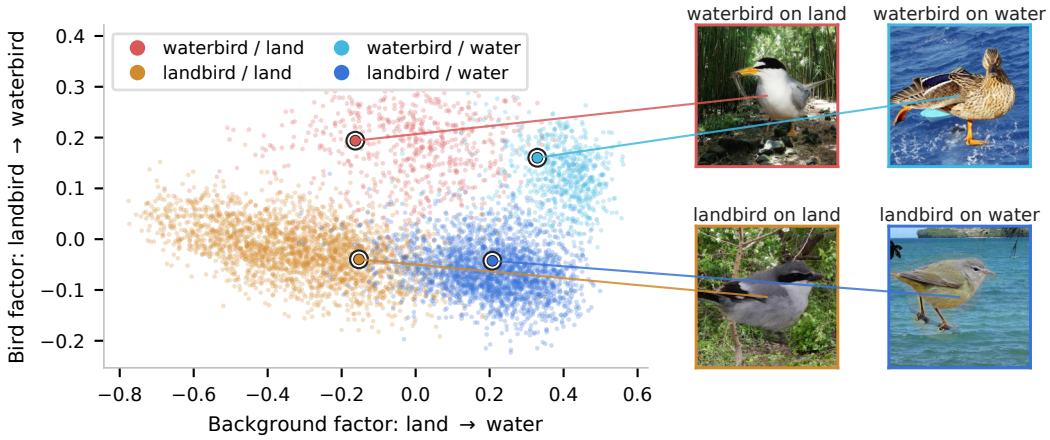


Figure 3: Learned Waterbirds concept plane after a post-hoc rotation of the two envelope coordinates. Marked held-out samples illustrate the four bird–background groups.

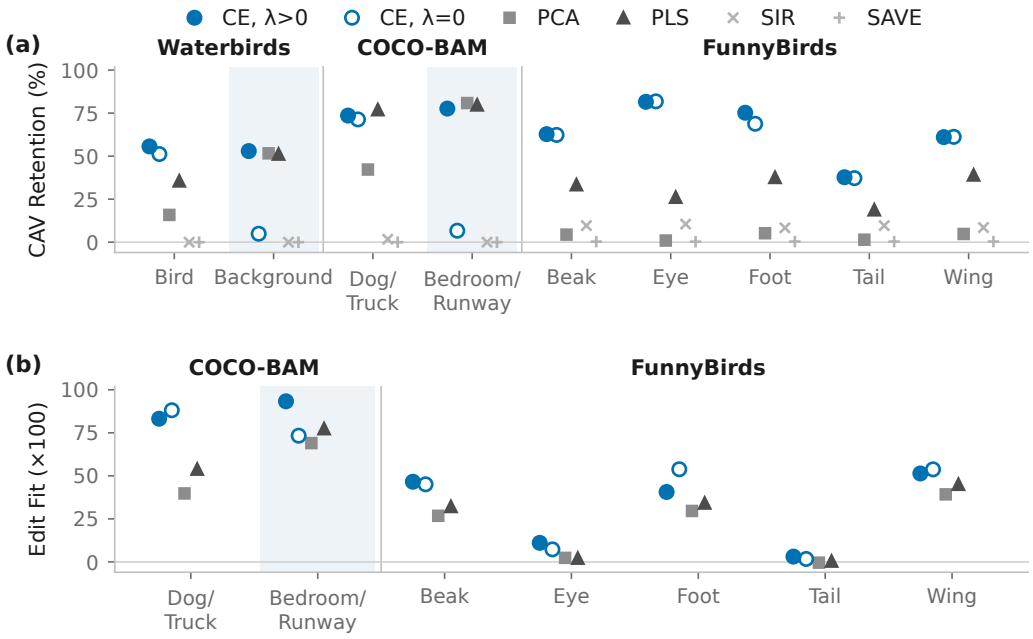


Figure 4: CAV Retention (a) on Waterbirds, COCO-BAM, and FunnyBirds, and Edit Fit (b) on source-paired interventions in COCO-BAM and FunnyBirds, for linear subspaces of the frozen embedding. Shaded columns are factors other than the prediction target; a missing Edit Fit marker means the subspace captures less than 1% of the intervention change. Exact values are in Appendix A.3.

Covariance separation keeps the second factor. At the same retained dimension, $\lambda = 0$ gives competitive average accuracy (Table 1) and retains the Waterbirds bird and COCO-BAM Dog/Truck directions about as well as $\lambda > 0$. The difference is in the second factor: Waterbirds background retention rises from 5% to 53%, and COCO-BAM scene retention from 7% to 78% (Figure 4). COCO-BAM scene Edit Fit also rises from 73 to 93. In the Waterbirds sweep, background retention stays near 53% at $\lambda \in \{20, 40, 80\}$, while average accuracy varies by less than 0.4 percentage points across the sweep (Appendix Table 8). On FunnyBirds, where the class is jointly determined by the five parts, the two settings are mixed rather than systematically different. The penalty's main gain is thus retaining a second factor that prediction alone leaves out.

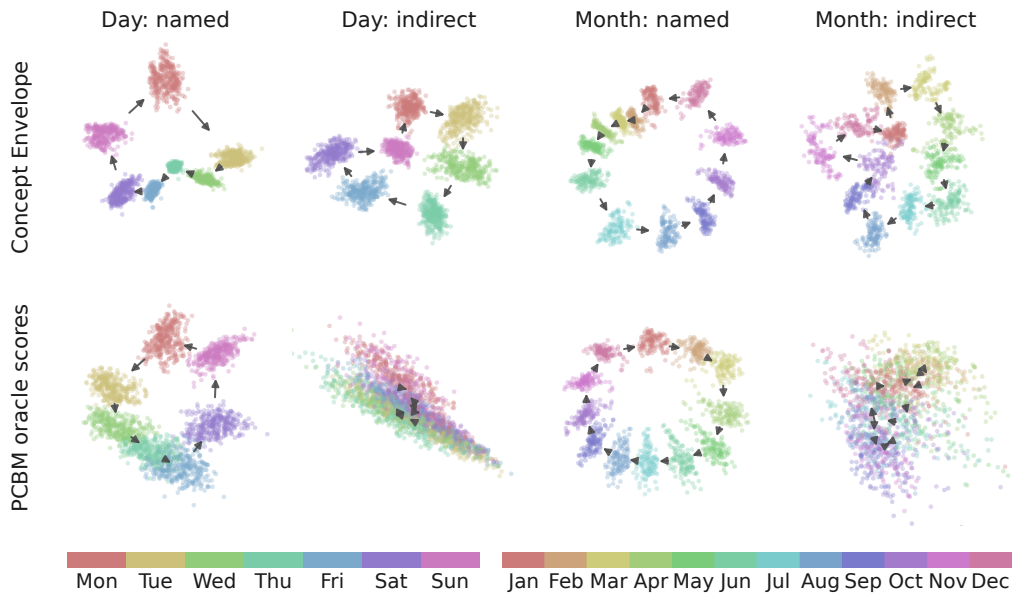


Figure 5: State geometry on the day and month tasks. Top: the two learned Concept Envelope coordinates. Bottom: the first two principal components of PCBM scores from the state-name dictionary. Colors mark the state and arrows the next state; each panel is rotated so that the first state points upward and scaled to a common radius.

The learned subspace retains more of the evaluated concept geometry. PCA and PLS retain the Waterbirds background and COCO-BAM scene about as well as Concept Envelope. They retain less of other directions: Concept Envelope retains 56% of the Waterbirds bird direction, compared with 36% for PLS and 16% for PCA, and exceeds both baselines on all five FunnyBirds part directions. Its Edit Fit also exceeds PCA and PLS in every reported column. SIR gives the clearest contrast with predictive accuracy: it leads average accuracy on COCO-BAM and FunnyBirds, yet retains at most 11% of any evaluated concept direction and never reaches the support threshold for Edit Fit.

The qualitative structure is stable across dimensions and encoders. Increasing u leaves the dominant Waterbirds geometry largely unchanged, and replacing Qwen3 with a vision-only MoonViT encoder preserves the same evaluated factors (Appendix A.4).

3.3 Dictionary Scores Can Follow Wording Instead of State

We construct two cyclic next-state tasks: predicting the next day of the week and the next month of the year. A named input states the current state explicitly ("The meeting is scheduled for Monday. What day is tomorrow?"), whereas an indirect input identifies it without using any day or month name ("The workweek just started, so I am prepping for tomorrow. What day is tomorrow?"); both are labeled Tuesday. Targets wrap from Sunday to Monday and from December to January. PCBM represents each input by one score per dictionary atom. We give it the true state names as its dictionary (Monday–Sunday; January–December), so the comparison tests whether the score representation recognizes indirect descriptions of the same state rather than whether the dictionary covers those states.

PCBM is accurate on named inputs but drops to 58% on days and 68% on months when the same state is described indirectly. The cyclic organization is also absent from the first two principal components of the score representation (Figure 5, bottom). A same-size generated dictionary improves indirect accuracy without closing the gap, showing that the result also depends on the choice of dictionary (Appendix Table 13). PCBM-h remains accurate because its residual predictor has access to information outside the score bottleneck. Concept Envelope learns its subspace from the task and uses a dictionary only afterward to name its coordinates. Its indirect accuracy reaches

99.4% on days and 94.4% on months, while the coordinates preserve the cyclic state order known to occur in language-model representations (Figure 5, top; Engels et al., 2025).

4 Theoretical Analysis

We first derive an oracle rate under *semantic sparsity*, where a small number of latent concepts generate the task-relevant subspace. We then formalize the *prediction–preservation gap* and establish conditions under which covariance separation selects the full active concept subspace.

4.1 Statistical Complexity under Semantic Sparsity

Let $A = (a_1, \dots, a_K) \in \mathbb{R}^{d\times K}$ be a fixed dictionary of normalized concept directions. Although its atoms may be dense in $\mathbb{R}^d$, a task-relevant concept subspace is generated by an active set $S_\star \subset [K]$ of size s; sparsity is therefore over concepts rather than embedding coordinates. For any support $S \subset [K]$ of size s with nonsingular $A_S^\top A_S$, let $P_S = A_S(A_S^\top A_S)^{-1}A_S^\top$ be the projector onto $\mathrm{span}(A_S)$, and fix an orthonormal basis Q_S for its range. We analyze a dictionary-constrained oracle surrogate whose retained dimension is s: the projection ranges over these dictionary-generated subspaces, while the prediction loss and covariance-separation penalty remain those of Section 2. Define

$$\mathcal{C}(P_S) = \|P_S\Sigma_X(I - P_S)\|_F^2, \qquad \widehat{\mathcal{C}}(P_S) = \|P_S\widehat{\Sigma}_X(I - P_S)\|_F^2.$$

For a population predictor g, write $\mathcal{L}_S(g) = \mathbb{E}\{\ell(Y, g(Q_S^\top X))\} + \lambda\mathcal{C}(P_S)$, and let

$$M(S) = \inf_g \mathcal{L}_S(g),$$

where the infimum is over measurable predictors with finite risk. Let $\mathcal{A}_s$ denote the collection of supports with $|S| = s$ and $A_S^\top A_S$ nonsingular. For i.i.d. observations $(X_i, Y_i)_{i=1}^n$, define the empirical criterion

$$\widehat{\mathcal{L}}_S(f) = \frac{1}{n}\sum_{i=1}^n \ell\{Y_i, f(Q_S^\top X_i)\} + \lambda\widehat{\mathcal{C}}(P_S).$$

The dictionary-constrained empirical concept envelope is

$$(\widehat{S}, \widehat{f}) \in \arg\min_{S\in\mathcal{A}_s,\ f\in\mathcal{F}_{n,s}} \widehat{\mathcal{L}}_S(f). \tag{5}$$

Here $\mathcal{F}_{n,s}$ is a predictor class on $\mathbb{R}^s$; α is the Hölder smoothness of the oracle predictor, W_n is the network size, and δ_n is the operator-norm covariance-estimation error. Appendix B.1 states the precise regularity conditions.

Theorem 1 (Concept-dictionary oracle rate)**.** *Assume the structural, prediction-class, and covariance-estimation conditions stated in Appendix B.1, and choose $W_n \asymp n^{s/(4\alpha+2s)}$. Then, up to logarithmic factors,*

$$\mathcal{L}_{\widehat{S}}(\widehat{f}) - M(S_\star) = O_p\Big(n^{-\frac{\alpha}{2\alpha+s}} + \sqrt{\frac{s\log K}{n}} + \lambda\{s\delta_n + s\delta_n^2\}\Big).$$

If, in addition, K grows polynomially in n and $\lambda\delta_n \lesssim \sqrt{\log n/(sn)}$, then

$$\mathcal{L}_{\widehat{S}}(\widehat{f}) - M(S_\star) = O_p\Big(n^{-\frac{\alpha}{2\alpha+s}} + \sqrt{\frac{s\log n}{n}}\Big).$$

The three terms control prediction, support selection, and covariance plug-in error. Because the candidate subspaces are restricted to dictionary-generated supports, this is an oracle rate for the surrogate, not a finite-sample guarantee for the estimator of Section 2, which optimizes over arbitrary semi-orthogonal bases. Under the simplified conditions, the bound consists of a slow nonparametric prediction term and a nearly parametric remainder. The proof is in Appendix B.

4.2 Gap Between Minimal Sufficiency and Interpretability

The oracle result targets the active concept subspace. We now ask why this target can be larger than a minimal predictive representation: prediction sufficiency retains enough information for Y, whereas interpretation also preserves the active concept subspace.

For an orthogonal projector P, let $R(P) = \inf_g \mathbb{E}\{\ell(Y, g(PX))\}$ and $R^\star = \inf_P R(P)$, where the infimum in $R(P)$ is over measurable predictors and optimizations over P are over orthogonal projectors. Let $P_\star$ be the projector onto the active concept subspace, with rank s. Define the concept omission score

$$D(P; P_\star) = \frac{1}{s}\|(I - P)P_\star\|_F^2 = 1 - \frac{1}{s}\operatorname{tr}(PP_\star).$$

Thus $D(P; P_\star) = 0$ exactly when P contains the active concept subspace. Assume $R^\star$ is attained by a projector of finite rank, and define

$$m_{\text{pred}} = \min\{\operatorname{rank}(P) : R(P) = R^\star\}, \quad m_{\text{pres}} = \min\{\operatorname{rank}(P) : R(P) = R^\star,\ D(P; P_\star) = 0\}.$$

The prediction–preservation gap is $\Delta_{\text{pres}} = m_{\text{pres}} - m_{\text{pred}}$. A positive gap means that the most compressed predictive representation is too small to preserve the active concept subspace.

Proposition 1 (Linear concept example)**.** *Suppose*

$$X = A_{S_\star} C_\star + R,$$

where the active atoms are orthonormal, so that $A_{S_\star} \in \mathbb{R}^{d\times s}$ has orthonormal columns, $C_\star \in \mathbb{R}^s$, and R is independent of $(C_\star, Y)$ and takes values orthogonal to $\operatorname{span}(A_{S_\star})$. Let $R_C^\star = \inf_g \mathbb{E}\{\ell(Y, g(C_\star))\}$. Suppose the minimal rank of a linear summary of $C_\star$ attaining $R_C^\star$ is $q < s$, attained by $B^\top C_\star$ for some $B \in \mathbb{R}^{s\times q}$ with rank q. Then $m_{\text{pred}} = q$, $m_{\text{pres}} = s$, so the gap reduces to $\Delta_{\text{pres}} = s - q$.

Suppose further that every risk-optimal projector contains $\mathcal{M} = \operatorname{span}(A_{S_\star} B)$, and that the smallest Σ_X-invariant subspace containing $\mathcal{M}$ is $\operatorname{range}(P_\star)$. At $u = s$, every projector containing $\mathcal{M}$ minimizes prediction risk. The penalty $\mathcal{C}(P)$ vanishes exactly when $\operatorname{range}(P)$ is Σ_X-invariant, so $P_\star$ attains both minimum risk and zero penalty. Thus every $\lambda > 0$ uniquely selects $P_\star$ as the minimizer of $R(P) + \lambda\mathcal{C}(P)$ (Appendix B.4).

5 Discussion

Statistical envelopes offer a representation for prediction and concept interpretation together, beyond their original role in efficient parameter estimation. In embeddings with approximately linear and orthogonal semantic structure, covariance separation favors keeping related variation together within the predictive subspace. Our results show why this matters: subspaces with similar predictive accuracy can retain very different concept geometry. This geometry may contain several distinct factors or multiple directions for one concept, consistent with the multi-dimensional structure of categorical concepts (Park et al., 2025).

Concept dictionaries remain useful, but their role should be separated from the representation used for prediction. In Concept Envelope, the estimator learns a low-dimensional subspace inside the fixed embedding from the task objective; dictionaries, representative examples, and grounding diagnostics are then used to provide human-readable evidence for that subspace. This changes the interpretation target from an isolated concept axis to the geometry of the retained subspace: what matters is not only whether a dictionary atom has a high score, but how semantic factors are organized within the predictive subspace.

Concept Envelope is task- and distribution-conditioned. The learned subspace is optimized for predictive contrasts in the observed data, and the post-hoc grounding describes concept evidence associated with it. Consequently, if a nuisance or spurious factor is predictive under the training distribution, the learned subspace itself may align more strongly with that factor than with a causally relevant one. Concept Envelope does not identify concepts that causally affect Y from predictive data alone: it recovers predictive concept evidence, but cannot in general distinguish causal factors from observationally similar non-causal alternatives. Such distinctions require stronger structural assumptions, counterfactual or interventional data, or explicit task design.

AI USE STATEMENT

We used generative AI tools to assist with literature retrieval, software implementation, experimental execution, preparation of figures, and drafting and editing the manuscript. We also used LLM-based agents in the post-hoc grounding workflow described in Section 2 and Appendix A.1: they proposed candidate concept descriptions from sampled training inputs and summarized quantitative evidence packets after subspace estimation. The agents did not enter the estimation objective or generate labels. The authors reviewed the AI-assisted text, citations, code, figures, candidate concepts, and summaries, verified the reported results, and take responsibility for the final content of this work.

REFERENCES

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016. URL https://arxiv.org/abs/1610.01644.

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). In *Advances in Neural Information Processing Systems*, volume 37, 2024.

R. Dennis Cook. SAVE: A method for dimension reduction and graphics in regression. *Communications in Statistics—Theory and Methods*, 29(9-10):2109–2121, 2000. doi: 10.1080/03610920008832598.

R. Dennis Cook and Xin Zhang. Foundations for envelope models and methods. *Journal of the American Statistical Association*, 110(510):599–611, 2015. doi: 10.1080/01621459.2014.983235.

R. Dennis Cook, Bing Li, and Francesca Chiaromonte. Envelope models for parsimonious and efficient multivariate linear regression. *Statistica Sinica*, 20(3):927–960, 2010.

R. Dennis Cook, Inge S. Helland, and Zhihua Su. Envelopes and partial least squares regression. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 75(5):851–877, 2013. doi: 10.1111/rssb.12018.

Joshua Engels, Eric J. Michaud, Isaac Liao, Wes Gurnee, and Max Tegmark. Not all language model features are one-dimensionally linear. In *International Conference on Learning Representations*, 2025.

Robin Hesse, Simone Schaub-Meyer, and Stefan Roth. FunnyBirds: A synthetic vision dataset for a part-based analysis of explainable AI methods. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3981–3991, 2023.

Jian Huang, Yuling Jiao, Xu Liao, Jin Liu, and Zhou Yu. Deep dimension reduction for supervised representation learning. *IEEE Transactions on Information Theory*, 70(5):3583–3598, 2024. doi: 10.1109/TIT.2023.3340658.

Robert Huben, Hoagy Cunningham, Logan Riggs Smith, Aidan Ewart, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models. In *International Conference on Learning Representations*, 2024.

Daniel Kapla, Lukas Fertl, and Efstathia Bura. Fusing sufficient dimension reduction with neural networks. *Computational Statistics & Data Analysis*, 168:107390, 2022. doi: 10.1016/j.csda.2021.107390.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2668–2677. PMLR, 2018.

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*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 2020.

Michael Kohler and Sophie Langer. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021. doi: 10.1214/20-AOS2034.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. URL https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

Ker-Chau Li. Sliced inverse regression for dimension reduction. *Journal of the American Statistical Association*, 86(414):316–327, 1991. doi: 10.1080/01621459.1991.10475035.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-vl-embedding and qwen3-vl-reranker: A unified framework for state-of-the-art multimodal retrieval and ranking. *arXiv preprint arXiv:2601.04720*, 2026. URL https://arxiv.org/abs/2601.04720.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 17612–17625, 2022.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.

Linquan Ma, Jixin Wang, Han Chen, and Lan Liu. Semiparametrically efficient method for enveloped central space. *Journal of the American Statistical Association*, 119(547):2166–2177, 2024. doi: 10.1080/01621459.2023.2252134.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 39643–39666. PMLR, 2024.

Kiho Park, Yo Joong Choe, Yibo Jiang, and Victor Veitch. The geometry of categorical and hierarchical concepts in large language models. In *International Conference on Learning Representations*, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

Adam Stein, Aaditya Naik, Yinjun Wu, Mayur Naik, and Eric Wong. Towards compositionality in concept learning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 46530–46555. PMLR, 2024.

Yin Tang and Bing Li. Belted and ensembled neural network for linear and nonlinear sufficient dimension reduction. *Journal of the American Statistical Association*, 121(554):1244–1255, 2026. doi: 10.1080/01621459.2025.2590775.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011. URL https://authors.library.caltech.edu/records/cvm3y-5hh21.

Mengjiao Yang and Been Kim. Benchmarking attribution methods with relative feature importance. *arXiv preprint arXiv:1907.09701*, 2019. URL https://arxiv.org/abs/1907.09701.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023.

A Additional Experiments and Grounding Details

This appendix gives the experimental details needed to reproduce and interpret the claims in Section 3. It records the grounding procedure, benchmark setup, and supporting diagnostics for the main figures.

A.1 Post-Hoc Grounding Pipeline

Post-hoc grounding begins from a fitted Concept Envelope model. At this stage, the learned subspace and predictor are fixed; grounding collects evidence used to interpret the learned coordinates.

For evaluation samples $X_1, \ldots, X_n$, define

$$z_i = \widehat{\Gamma}^\top X_i, \qquad \mathbf{Z} = (z_1, \ldots, z_n)^\top.$$

Let $\mathcal{D} = \{c_1, \ldots, c_m\}$ be the concept dictionary used for semantic evaluation, and let $s_j \in \mathbb{R}^n$ be the activation vector of atom c_j on the same evaluation samples. Sample inputs and dictionary atoms are embedded with the same frozen Qwen3-VL-Embedding-8B checkpoint. Before activations are computed, both embedding sets receive the same centering and normalization preprocessing, so dictionary activations are evaluated in a common embedding geometry (Li et al., 2026; Liang et al., 2022; Bhalla et al., 2024).

Figure 6 separates the agent-assisted language stages from the quantitative evaluation. For agent-generated dictionaries, the first language stage proposes candidate atoms from sampled training inputs without target labels, class names, or group attributes. The quantitative step evaluates dictionary activations against the learned coordinates and forms an evidence packet. A second language stage may summarize that packet from accepted atoms, correlations, metrics, and representative samples. The language stages are advisory; estimation, acceptance, metrics, and predictions remain with the quantitative step.

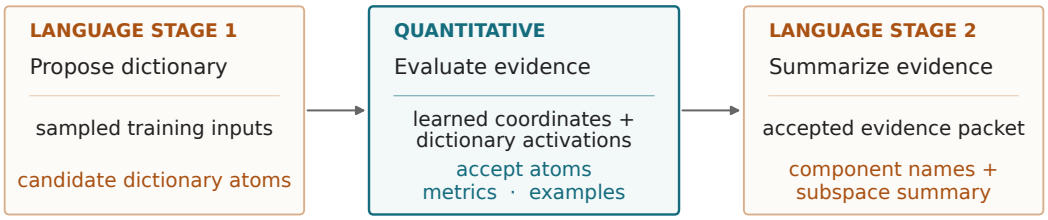


Figure 6: Agent-assisted post-hoc grounding. Language stages propose candidate atoms and may summarize the final evidence packet. The quantitative step evaluates dictionary activations against learned coordinates, accepts atoms, and computes the diagnostics returned by Algorithm 1.

Table 2 gives coarse grounding summaries for the selected benchmarks and diagnostics.

Table 2: Coarse grounding summaries. For higher-dimensional visual subspaces, the reported factor groups are coarsened from the corresponding concept-space reports rather than assigned one component at a time.

Task	u	Factor families	Observed organization
Waterbirds	2	Perching songbird versus aquatic waterbird morphology	One dominant bird-morphology axis opposes perch-oriented songbird traits to aquatic and wading waterbird structure.
CelebA	16	Light scalp hair; facial hair; dark hair	The primary split contrasts light scalp hair with darker hair and facial-hair traits; later axes add weaker facial-hair subtype and dark-hair refinements.
CIFAR-10	16	Animal morphology family; road vehicle family; marine and aircraft scene family	The main split separates animal traits from transport or scene concepts, with weaker structure separating road vehicles from marine or aircraft scenes.
CIFAR-100	24	Animal versus built-world content; building and urban structure; aquatic/reptilian morphology; animal surface patterning and plumage	The main split separates animal anatomy and texture from streets, buildings, furniture, and appliances; weaker axes isolate urban structure, aquatic or reptilian traits, and fur or plumage patterning.
CUB	32	Perched woody vegetation versus aquatic shoreline habitat; foliage enclosure and occlusion; composition scale and perch geometry	The main split separates perched woody-vegetation scenes from open aquatic or shoreline settings, with weaker axes reflecting foliage occlusion and perch geometry.
Day, named	2	Weekend versus midweek; early-week versus late-week position	The named day task is organized by a weekend/midweek contrast together with a second factor for early- versus late-week position.
Day, indirect	2	Workweek progression; weekend versus midweek	The indirect day task keeps the weekend/midweek factor but shifts the dominant organization toward early-to-late workweek progression.
Month, named	2	Summer versus winter/year-boundary; spring/early-summer versus autumn	The named month task follows an annual position geometry with one factor contrasting summer against winter or year-boundary cues and a second factor separating spring or early summer from autumn.
Month, indirect	2	Summer-to-winter seasonality; spring-to-autumn transition	The indirect month task preserves a seasonal-cycle geometry organized by a summer/winter contrast and a secondary spring/autumn transition.

Algorithm 1 Post-hoc grounding of a learned Concept Envelope subspace

Require: learned coordinates $\mathbf{Z}$, concept dictionary $\mathcal{D}$, activations $\{s_j\}_{j=1}^m$, threshold τ, fit/holdout split
Ensure: accepted atoms $\mathcal{A}$, joint fit J, effective rank r, evidence packet $\mathcal{E}$

for $c_j \in \mathcal{D}$ **do**
  $\rho_j \leftarrow \big(\mathrm{corr}((s_j)_{\mathrm{fit}}, (\mathbf{Z}_{\mathrm{fit}})_{:,1}), \ldots, \mathrm{corr}((s_j)_{\mathrm{fit}}, (\mathbf{Z}_{\mathrm{fit}})_{:,u})\big)$
  $a_j \leftarrow u^{-1} \sum_{k=1}^{u} |\rho_{jk}|$
end for
$\mathcal{A} \leftarrow \{j : a_j \geq \tau\}$
if $\mathcal{A} = \varnothing$ **then**
  return $\mathcal{A}$, $J = 0$, $r = 0$, empty evidence packet
end if
form $\mathbf{S}_{\mathcal{A}}$ from the accepted activation vectors
$\widehat{B}_{\mathcal{A}} \leftarrow \arg\min_B \|\mathbf{Z}_{\mathrm{fit}} - \mathbf{S}_{\mathcal{A},\mathrm{fit}} B\|_F^2$
$\widehat{\mathbf{Z}}_{\mathcal{A},\mathrm{hold}} \leftarrow \mathbf{S}_{\mathcal{A},\mathrm{hold}} \widehat{B}_{\mathcal{A}}$
$J \leftarrow R^2(\mathbf{Z}_{\mathrm{hold}}, \widehat{\mathbf{Z}}_{\mathcal{A},\mathrm{hold}})$
$\widehat{\mathbf{Z}}_{\mathcal{A}} \leftarrow \mathbf{S}_{\mathcal{A}} \widehat{B}_{\mathcal{A}}$
$r \leftarrow$ entropy effective rank of $\widehat{\mathbf{Z}}_{\mathcal{A}}$
choose post-hoc orthogonal Q and set $\mathbf{Z}_{\mathrm{rot}} \leftarrow \mathbf{Z}Q$ for interpretation
$\mathcal{E} \leftarrow \{\mathcal{A}, \{\rho_j\}_{j\in\mathcal{A}}, J, r, \mathbf{Z}_{\mathrm{rot}}, \text{representative samples}\}$
return $\mathcal{A}$, J, r, $\mathcal{E}$

The acceptance statistic a_j screens for dictionary atoms whose activations correlate with the learned coordinates. The joint fit J is the held-out R^2 from accepted dictionary activations to learned coordinates. The effective-rank diagnostic r is computed from the normalized squared singular values of $\widehat{\mathbf{Z}}_{\mathcal{A}}$. The rotation Q is only for interpretation: because the basis inside $\mathrm{span}(\widehat{\Gamma})$ is not identifiable, rotating coordinates for reporting does not change the learned subspace or predictions.

The evidence packet $\mathcal{E}$ contains the accepted atoms, correlations, diagnostics, rotated coordinates, and representative samples; only this packet is passed to the optional summary stage.

A.2 Experimental Setup and Baselines

Table 3 lists the dataset tasks, grounding dictionaries, and prediction metrics used in the experiments. The visual benchmarks are Waterbirds, CelebA, COCO-BAM, CIFAR-10, CIFAR-100, CUB, and FunnyBirds (Sagawa et al., 2020; Liu et al., 2015; Yang & Kim, 2019; Krizhevsky, 2009; Wah et al., 2011; Hesse et al., 2023). We use dataset-provided train, validation, and test splits when available. For CIFAR-10, CIFAR-100, and CUB, we retain the standard test split and reserve a fixed seed-42 10% subset of the training split for validation. COCO-BAM uses our source-connected 800/100/100 quartet split, and FunnyBirds uses the registered class-balanced 45,000/5,000/500 split. The day and month tasks are constructed diagnostics for named versus indirect descriptions of the same latent state.

Table 3: Benchmark setup and prediction metrics.

Dataset / task	Target	Group or nuisance	Grounding dictionary	Metrics
Waterbirds	bird type	background	agent-generated	Avg, WG
CelebA	blond hair	male attribute	agent-generated	Avg, WG
COCO-BAM	dog versus truck	bedroom versus runway	agent-generated	Avg, WG
CIFAR-10	object class	–	agent-generated	Avg
CIFAR-100	object class	–	agent-generated	Avg
CUB	bird species	–	agent-generated	Avg
FunnyBirds	50-way class	–	agent-generated	Avg
Day of week	next state	description form	true state names	Avg
Month of year	next state	description form	true state names	Avg

Table 4 lists what each compared family learns: unsupervised representations (PCA, SAE; Huben et al., 2024), classical SDR (SIR, SAVE; Li, 1991; Cook, 2000), envelope methods (PLS; Cook et al., 2013), neural SDR, linear and nonlinear (Kapla et al., 2022; Tang & Li, 2026; Huang et al., 2024), concept bottlenecks (PCBM, PCBM-h; Koh et al., 2020; Yuksekgonul et al., 2023), and the full-embedding MLP as an accuracy reference. Envelope methods also include the predictor envelope (Cook et al., 2013) and the enveloped central space (Ma et al., 2024); PLS is the envelope method evaluated in Table 1.

Table 4: What each compared method family learns.

Family	Methods	What is learned
Unsupervised representations	PCA + MLP; SAE + MLP	a high-variance linear subspace (PCA) or sparse autoencoder features (SAE), fitted without the response
Classical SDR	SIR + MLP; SAVE + MLP	directions of the response-sufficient linear central subspace, estimated from inverse conditional moments
Envelope methods	PLS + MLP	response-relevant linear structure given by PLS components
Neural SDR, linear reduction	CE, $\lambda = 0$	a linear reduction learned jointly with a nonlinear predictor
Neural SDR, nonlinear reduction	Deep SDR	a nonlinear sufficient representation learned jointly with a predictor
Concept bottlenecks	PCBM; PCBM-h	scalar concept scores (PCBM), plus a residual predictor (PCBM-h)
Concept Envelope	CE, $\lambda > 0$	a linear subspace and nonlinear predictor trained end to end with covariance separation
Full-embedding reference	MLP	the full frozen embedding

Relation to Concept Envelope. The families differ in what they offer an interpretation. The unsupervised reductions see no response. Classical SDR targets response sufficiency, which need not produce factor-aligned semantic geometry; a K-class response yields at most $K-1$ SIR directions (Li, 1991). Envelope methods supply the covariance structure that our penalty comes from, but fit their reduction separately from a nonlinear predictor. A linear reduction trained jointly with a nonlinear predictor is the case closest to ours, and is represented by Concept Envelope at $\lambda = 0$; its nonlinear counterpart, Deep SDR, is a strong predictive comparator without directions in the original embedding space. Concept-bottleneck models predict from scalar concept scores, and PCBM-h adds a residual predictor. Concept Envelope at $\lambda > 0$ is the proposed method: a linear subspace and a nonlinear predictor trained end to end with covariance separation, interpreted after estimation.

On Waterbirds, CelebA, CIFAR-10, CIFAR-100, and CUB, PCBM and PCBM-h use the same generated concept-dictionary protocol as the grounding stage. Concept scores are computed as embedding-space similarities between frozen example embeddings and generated concept embeddings. This is a PCBM-style scalar concept-score interface, but it differs from the task-specific CAV concept banks used in some original PCBM experiments, where concept vectors can be learned from positive and negative concept examples. On these benchmarks, the PCBM rows should therefore be interpreted as PCBM-style concept-score predictors under our dictionary construction protocol rather than as a reproduction of a task-specific CAV setup. The effect is most visible on fine-grained datasets such as CUB, where semantically meaningful dictionary atoms may capture bird appearance, pose, or background while still being insufficiently discriminative for 200-way species classification. On COCO-BAM and FunnyBirds, whose concept annotations are available, PCBM and PCBM-h use the CAV variant of the concept bank, fitted from the object and context labels (COCO-BAM) and the 26 official part values (FunnyBirds).

A.3 Concept Metrics

CAV Retention. For a concept c, let v_c be its concept activation vector (Kim et al., 2018), fitted in the frozen embedding by class-balanced ℓ_2-regularized logistic regression on the concept labels and held fixed across methods. For a retained subspace $\mathcal{S}$ with orthogonal projector P,

$$\mathrm{CAVRet}(\mathcal{S}, c) = \frac{\|Pv_c\|_2^2}{\|v_c\|_2^2} = \cos^2 \angle(v_c, \mathcal{S}).$$

Concept labels are used only to fit these evaluation vectors; they do not enter any projector's training objective. For FunnyBirds, each part family averages the one-versus-rest scores of its part values.

Edit Fit. For a directed intervention t, let Δ_i be the change in the frozen embedding from a source image to its intervened pair. From the fit pairs we estimate the mean projected edit $\hat{d}_t = P\overline{\Delta}_{\mathrm{fit},t}$. On disjoint held-out pairs, define the transition fit f_t and its retained edit-energy share s_t as

$$f_t = 1 - \frac{\sum_{i \in \mathrm{test}(t)} \|P\Delta_i - \hat{d}_t\|_2^2}{\sum_{i \in \mathrm{test}(t)} \|P\Delta_i\|_2^2}, \qquad s_t = \frac{\sum_{i \in \mathrm{test}(t)} \|P\Delta_i\|_2^2}{\sum_{i \in \mathrm{test}(t)} \|\Delta_i\|_2^2}.$$

For a factor c with directed-transition set $\mathcal{T}_c$, we report the retained-energy-weighted fit

$$\mathrm{EditFit}(\mathcal{S}, c) = \frac{\sum_{t \in \mathcal{T}_c} s_t f_t}{\sum_{t \in \mathcal{T}_c} s_t}.$$

This is equivalent to normalizing the projected residual and projected change of each transition by its ambient edit energy before aggregation. We report $100 \times \mathrm{EditFit}$ and keep negative values.

Support threshold. A factor's support is the average of s_t over its directed transitions. A cell is reported only if this support is at least 1% in every seed, which excludes projections that contain almost none of the edit. Every threshold from 0.1% to 1.79% gives the same table: the lowest retained change among reported cells is 1.797%, and the highest among excluded cells is 0.097%.

Matched dimensions. Subspace dimensions match Concept Envelope on each dataset ($u = 2$ on Waterbirds, $u = 8$ on COCO-BAM, and $u = 21$ on FunnyBirds). Only SIR is class-count limited: a K-class response yields at most $K-1$ directions; SAVE uses the requested dimension. For reference, an isotropic u-dimensional subspace of the 4096-dimensional embedding retains $u/4096$ of a random direction in expectation (0.049% at $u = 2$), well below every reported value.

Table 5: CAV Retention (%). Mean over five seeds; population SD in parentheses for Concept Envelope. The other methods are deterministic given the split.

	Waterbirds		COCO-BAM		FunnyBirds				
Method	Bird	Background	Dog/Truck	Bedroom/Runway	Beak	Eye	Foot	Tail	Wing
CE, $\lambda > 0$	55.64 (1.28)	52.94 (0.64)	73.57 (1.36)	77.63 (0.97)	62.77 (1.53)	81.56 (2.35)	75.20 (2.85)	37.74 (1.59)	61.05 (2.52)
CE, $\lambda = 0$	51.21 (7.65)	4.91 (3.13)	71.32 (2.76)	6.67 (1.10)	62.32 (2.38)	81.83 (2.95)	68.80 (2.10)	37.21 (1.32)	61.20 (2.01)
PCA	15.86	51.63	42.22	80.83	4.39	0.95	5.20	1.45	4.79
PLS	35.61	51.19	77.00	79.76	33.40	26.13	37.63	18.83	39.06
SIR	<0.01	<0.01	1.72	0.02	9.63	10.53	8.34	9.62	8.49
SAVE	<0.01	<0.01	0.01	<0.01	0.30	0.29	0.36	0.34	0.36

FunnyBirds columns average the one-versus-rest scores of the part values in each family.

Table 6: Edit Fit ($\times 100$). Mean over five seeds; population SD in parentheses for Concept Envelope. The other methods are deterministic given the split.

	COCO-BAM		FunnyBirds				
Method	Dog/Truck	Bedroom/Runway	Beak	Eye	Foot	Tail	Wing
CE, $\lambda > 0$	83.19 (0.31)	93.31 (0.12)	46.58 (1.88)	11.07 (2.62)	40.64 (1.18)	3.09 (0.66)	51.39 (1.82)
CE, $\lambda = 0$	88.11 (0.23)	73.34 (2.65)	45.10 (1.88)	7.30 (4.85)	53.85 (2.49)	1.73 (1.20)	53.78 (3.12)
PCA	39.79	69.04	26.80	2.37	29.62	−0.41	39.26
PLS	53.92	77.44	32.22	2.20	34.22	0.50	45.11
SIR	—	—	—	—	—	—	—
SAVE	—	—	—	—	—	—	—

A dash marks a cell where the subspace keeps less than 1% of the ambient change in some seed. The gate is on the retained share of the ambient change, while Edit Fit itself is normalized within the projected change.

Paired interventions. Edit Fit needs image pairs that differ in one factor. From BAM assets (Yang & Kim, 2019), we construct source-connected dog/truck $\times$ bedroom/runway replacements as complete 2×2 quartets and split them 800/100/100. Using the FunnyBirds rendering factors (Hesse et al., 2023), we form exact single-part replacement pairs while preserving all other rendering parameters. Tables 5 and 6 report the values plotted in Figure 4.

A.4 Ablation Studies

This subsection reports two Waterbirds ablations in the original PCA-2 view of the learned coordinates. We do not treat u as the number of true concepts; it controls the capacity of a compact retained subspace. The top row of Figure 7 varies the retained dimension at fixed λ, and the bottom row varies the covariance-separation weight at fixed u.

Retained dimension. At fixed $\lambda = 20$, increasing u adds capacity without materially changing the dominant Waterbirds geometry, as the top row of Figure 7 shows. Table 7 reports the corresponding prediction and grounding values.

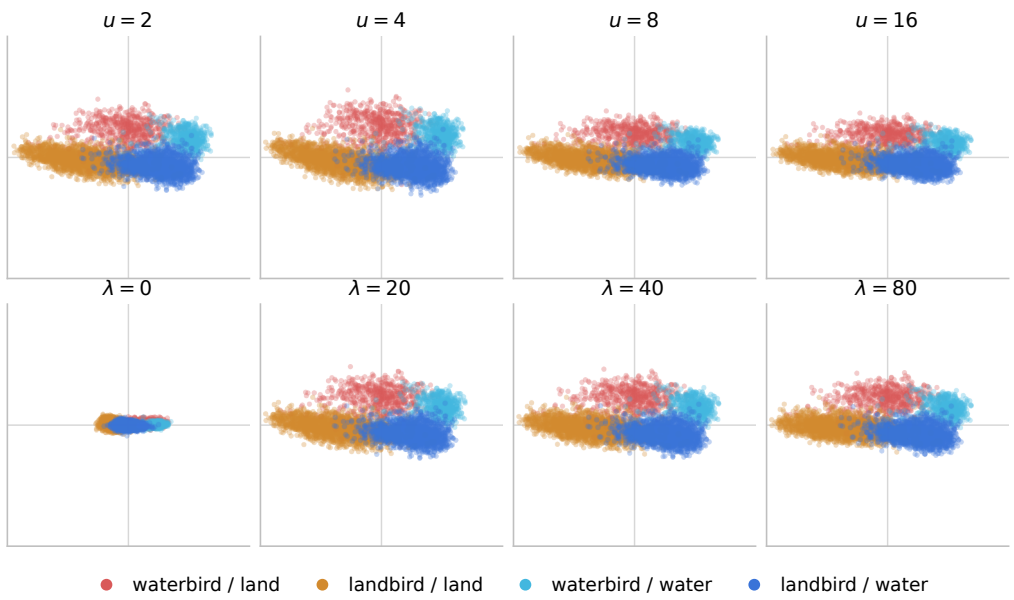


Figure 7: Waterbirds ablations in PCA-2 coordinates for the fixed representative seed 42. Top row: retained dimension at fixed $\lambda = 20$. Bottom row: covariance-separation weight at fixed $u = 2$. Positive-λ runs use the global covariance estimator, and each row uses a shared axis scale.

Table 7: Waterbirds retained-dimension ablation at fixed $\lambda = 20$. Values are means over seeds 42–46, with population SD in parentheses. Accuracy is reported as a percentage; the final column is the percentage of variance in the accepted-concept reconstruction explained by its first two principal components. Joint fit and effective rank are defined in Algorithm 1.

u	Avg. acc. (%)	WG acc. (%)	Joint fit	Eff. rank	First two comp. (%)
2	92.45 (1.23)	71.15 (0.75)	0.886 (0.008)	1.149 (0.019)	100.00 (0.00)
4	92.82 (0.71)	72.87 (3.11)	0.886 (0.009)	1.155 (0.055)	99.70 (0.13)
8	93.29 (0.88)	71.50 (9.84)	0.891 (0.005)	1.149 (0.060)	99.35 (0.08)
16	94.28 (0.50)	75.83 (3.35)	0.886 (0.006)	1.122 (0.040)	98.63 (0.10)

Covariance-separation weight. At fixed $u = 2$, the bottom row of Figure 7 shows a nearly one-dimensional predictive mixture at $\lambda = 0$ and a visible two-factor geometry at positive λ. Table 8 quantifies the difference using the same fixed bird and background CAVs as Figure 4. Background retention rises from 4.91% at $\lambda = 0$ to approximately 53% at each of $\lambda \in \{20, 40, 80\}$, while bird retention stays between 51% and 56%. Joint fit also rises from 0.781 to approximately 0.887. Average accuracy varies by less than 0.4 percentage points across the sweep, while worst-group accuracy falls from 74.86% to 70.25–71.15%. At this fixed dimension, the penalty therefore preserves more background information with little change in average accuracy, but a cost in worst-group accuracy.

Table 8: Waterbirds covariance-separation ablation at fixed $u = 2$. Values are means over seeds 42–46, with population SD in parentheses. Accuracy and CAV Retention are percentages; CAVs are fixed across all weights and seeds. Joint fit and effective rank summarize the post-hoc grounding (Algorithm 1).

					CAV Retention (%)	
λ	Avg. acc. (%)	WG acc. (%)	Joint fit	Eff. rank	Bird	Background
0	92.78 (0.88)	74.86 (2.13)	0.781 (0.052)	1.051 (0.033)	51.21 (7.65)	4.91 (3.13)
20	92.45 (1.23)	71.15 (0.75)	0.886 (0.008)	1.149 (0.019)	55.64 (1.28)	52.94 (0.64)
40	92.63 (1.30)	70.97 (2.47)	0.887 (0.006)	1.122 (0.033)	53.33 (4.51)	52.96 (0.87)
80	92.46 (1.23)	70.25 (2.62)	0.888 (0.004)	1.099 (0.027)	51.81 (4.61)	52.92 (0.81)

Selecting u and λ. We tune in two steps over the fixed dataset-specific grids recorded in the released configurations. For the retained dimension, we fit a moderately large pilot subspace, use the effective rank of its grounded evidence to identify the relevant range, and choose the smallest u on the validation-accuracy plateau. With u fixed, a dataset-specific pilot balances validation accuracy against the geometry of the grounded evidence, summarized by joint fit and effective rank; semantic coverage is an additional diagnostic when task-specific concept labels are available. The resulting λ is frozen before the formal five-seed summary rather than reselected from it.

Encoder sensitivity. Tables 9 and 10 repeat the semantic evaluation under a vision-only MoonViT-SO-400M encoder. The evaluated factor families and interventions remain represented under both encoders, but their relative strengths change: MoonViT increases Waterbirds and several FunnyBirds retention values while reducing both COCO-BAM retention values and the COCO-BAM dog/truck Edit Fit. The larger Waterbirds seed spread under MoonViT also shows that the detailed geometry is encoder-dependent.

Table 9: CAV Retention (%) of Concept Envelope ($\lambda > 0$) under two encoders. Mean over five seeds; population SD in parentheses.

	Waterbirds		COCO-BAM		FunnyBirds				
Encoder	Bird	Background	Dog/Truck	Bedroom/Runway	Beak	Eye	Foot	Tail	Wing
Qwen3	55.64 (1.28)	52.94 (0.64)	73.57 (1.36)	77.63 (0.97)	62.77 (1.53)	81.56 (2.35)	75.20 (2.85)	37.74 (1.59)	61.05 (2.52)
MoonViT	77.89 (15.40)	63.68 (22.75)	72.22 (1.96)	49.77 (0.45)	62.06 (2.23)	84.53 (2.05)	76.24 (2.09)	46.59 (1.63)	72.82 (2.43)

Table 10: Edit Fit ($\times 100$) of Concept Envelope ($\lambda > 0$) under two encoders. Mean over five seeds; population SD in parentheses.

	COCO-BAM		FunnyBirds				
Encoder	Dog/Truck	Bedroom/Runway	Beak	Eye	Foot	Tail	Wing
Qwen3	83.19 (0.31)	93.31 (0.12)	46.58 (1.88)	11.07 (2.62)	40.64 (1.18)	3.09 (0.66)	51.39 (1.82)
MoonViT	45.78 (0.40)	81.59 (0.24)	57.22 (2.40)	40.52 (1.81)	56.31 (3.19)	10.61 (2.58)	57.65 (2.32)

Dictionary size and alignment threshold. The dictionary size and the acceptance threshold τ decide which concepts are named. We vary each on Waterbirds, COCO-BAM, and FunnyBirds, at the settings used for the concept-metric benchmarks ($K = 24$, $\tau = 0.15$), with one agent-generated dictionary per size held fixed across seeds. Concepts are selected and the regression is fitted on the training split; joint fit is evaluated on the test split, and effective rank is computed from the resulting reconstruction (Algorithm 1).

Table 11 varies $K \in \{12, 24, 48\}$ at $\tau = 0.15$, and Table 12 varies $\tau \in \{0.10, \ldots, 0.30\}$ at $K = 24$. A larger dictionary fits the learned coordinates more closely but reconstructs them with a less concentrated component, so joint fit and effective rank grow together. $K = 24$ fits all three datasets well; on FunnyBirds the fit still improves at $K = 48$, at the cost of a higher effective rank. The alignment threshold is nearly flat on Waterbirds and COCO-BAM. On FunnyBirds, both diagnostics fall as τ increases, with the sharpest decline above $\tau = 0.20$; the reported setting $\tau = 0.15$ lies before this decline.

Table 11: Grounding diagnostics against dictionary size, at fixed $\tau = 0.15$. Mean over five seeds; population SD in parentheses. One dictionary per size, held fixed across seeds.

Dataset	Metric	$K = 12$	$K = 24$	$K = 48$
Waterbirds	Joint fit	0.830 (0.003)	0.908 (0.001)	0.970 (0.002)
	Effective rank	1.074 (0.012)	1.102 (0.018)	1.170 (0.023)
FunnyBirds	Joint fit	0.487 (0.024)	0.660 (0.040)	0.779 (0.041)
	Effective rank	1.777 (0.078)	2.207 (0.206)	2.405 (0.235)
COCO-BAM	Joint fit	0.949 (0.001)	0.969 (0.001)	0.981 (0.001)
	Effective rank	1.292 (0.020)	1.331 (0.023)	1.356 (0.024)

Table 12: Grounding diagnostics against the alignment threshold τ, at fixed $K = 24$. Mean over five seeds; population SD in parentheses.

Dataset	Metric	$\tau = 0.10$	$\tau = 0.15$	$\tau = 0.20$	$\tau = 0.25$	$\tau = 0.30$
Waterbirds	Joint fit	0.916 (0.002)	0.908 (0.001)	0.896 (0.007)	0.892 (0.009)	0.884 (0.007)
	Effective rank	1.113 (0.019)	1.102 (0.018)	1.096 (0.017)	1.094 (0.016)	1.088 (0.015)
FunnyBirds	Joint fit	0.708 (0.012)	0.660 (0.040)	0.528 (0.066)	0.215 (0.119)	0.042 (0.058)
	Effective rank	2.227 (0.200)	2.207 (0.206)	1.994 (0.146)	1.646 (0.340)	0.404 (0.495)
COCO-BAM	Joint fit	0.969 (0.001)	0.969 (0.001)	0.968 (0.001)	0.968 (0.001)	0.968 (0.001)
	Effective rank	1.331 (0.023)	1.331 (0.023)	1.330 (0.023)	1.330 (0.022)	1.330 (0.023)

A.5 Additional Results and Implementation Details

Optimization. In the experiments, $\widehat{\Sigma}_X$ is computed once from the training embeddings and held fixed during optimization. We optimize the semi-orthogonal basis Γ on the Stiefel manifold, using gradient updates followed by QR retraction to enforce $\Gamma^\top \Gamma = I_u$; the predictor parameters are updated jointly with Γ under the empirical objective in (4). Unless otherwise noted, the downstream predictor is an MLP head with one hidden layer of width 64.

Compute resources. The released five-seed benchmark, ablation, and semantic-evaluation runs were executed as single-GPU jobs on the Nova cluster using NVIDIA A100 GPUs. Development and local smoke checks used a workstation with an Intel Core i5-13400 CPU, an NVIDIA GeForce RTX 4070 SUPER GPU with 12GB VRAM, and 32GB system RAM. The experiments do not require distributed training; exact commands and dataset-specific configurations are included in the accompanying code release.

Assets and licenses. The experiments use the Qwen3-VL-Embedding-8B encoder together with standard benchmark datasets, including Waterbirds, CelebA, COCO-BAM, CIFAR-10, CIFAR-100, CUB, and FunnyBirds, and the paper cites the original sources. The accompanying code release records the source links and the relevant license or terms-of-use information for these

assets, including Apache-2.0 for Qwen3-VL-Embedding-8B, dataset-card or release-agreement information for Waterbirds and CelebA, citation and host-terms guidance for CIFAR-10, CIFAR-100, COCO-BAM, and FunnyBirds, and the non-commercial research and educational-use restriction for CUB-200-2011.

Code release. An anonymized code release accompanies the core experiments. It includes the main training and evaluation entry points, configuration files for the benchmark settings, and instructions for reproducing the principal tables and figures. The release also documents required external downloads and the license for the released code.

Constructed day and month diagnostics. Each example describes a current state and is labeled by its cyclic successor, with Sunday mapping to Monday and December to January. The named weekday corpus instantiates 320 prompt templates for each of the seven days (2,240 examples); the indirect corpus combines 40 schedule scenarios with eight question wrappers for the same class-balanced total and contains no literal weekday names. The named and indirect month corpora each instantiate 120 prompt templates for each of the 12 months (1,440 examples); the indirect prompts use seasonal, holiday, and routine cues and contain no literal month names. These are separately generated corpora with the same cyclic labeling rule. Each corpus is split independently and stratified by successor label into 70%/15%/15% train, validation, and test sets: 1,568/336/336 examples for days and 1,008/216/216 for months.

Concept Envelope predicts from the learned subspace rather than from dictionary scores. The main PCBM and PCBM-h rows use dictionaries containing the true state names. In Figure 5, the Concept Envelope panels show both learned coordinates ($u = 2$) directly, the principal components of the PCBM scores are fitted on the training split, and each panel is centered, rotated so that the first state points upward, and scaled to a common radius.

Table 13: Accuracy (%) on constructed day and month diagnostics. Oracle rows are the formal comparison; agent rows are auxiliary diagnostics with size-matched generated dictionaries.

	Day of week		Month of year	
Method	Named	Indirect	Named	Indirect
Concept Envelope	99.11	99.40	97.22	94.44
PCBM (oracle)	99.70	58.33	100.00	67.59
PCBM (agent)	100.00	76.49	100.00	82.87
PCBM-h (oracle)	100.00	100.00	100.00	100.00
PCBM-h (agent)	100.00	100.00	100.00	100.00

The oracle rows show the main effect: PCBM drops on indirect descriptions even when the dictionary contains the true state names, and the agent-generated dictionaries, size-matched at seven atoms for days and twelve for months, raise indirect accuracy by 15–18 points without closing the gap. PCBM-h remains accurate because its residual predictor can carry information outside the score bottleneck; its accuracy therefore does not imply that the concept-score representation alone is sufficient.

B PROOFS

B.1 CONDITIONS FOR THEOREM 1

The theorem for the dictionary-constrained oracle surrogate uses the following conditions. They separate the semantic-sparsity assumption from standard prediction-class and covariance-estimation controls.

Condition 1 (Latent concept target)**.** *There is an active set $S_\star \in \mathcal{A}_s$ such that $P_\star = P_{S_\star}$ is well defined and the nonzero eigenvalues of $A_{S_\star}^\top A_{S_\star}$ are bounded away from zero and infinity. The active set $S_\star$ minimizes $M(S)$ over all $S \in \mathcal{A}_s$.*

Condition 2 (Loss regularity and tails)**.** *The sieve is uniformly clipped,* $\sup_{f\in\mathcal{F}_{n,s}}\|f\|_\infty \le C_f$. *The loss is* L_ℓ*-Lipschitz on* $[-C_f, C_f]$. *The anchored loss* $\ell(Y,0)$ *is sub-exponential, and* $Q_S^\top X$ *is uniformly sub-Gaussian over* $S \in \mathcal{A}_s$.

Condition 3 (Smooth oracle predictor)**.** *A population minimizer* $g_\star \in \arg\min_g \mathcal{L}_{S_\star}(g)$ *belongs to a bounded Hölder ball* $\mathcal{H}^\alpha(C_g)$ *on* $\mathbb{R}^s$ *and satisfies* $\|g_\star\|_\infty \le C_f$.

Condition 4 (Neural-network class)**.** *The prediction class* $\mathcal{F}_{n,s}$ *is a clipped fixed-width/deep fully connected ReLU sieve on* $\mathbb{R}^s$*, with output range contained in* $[-C_f, C_f]$ *and parameter size* W_n.

Condition 5 (Covariance estimation)**.** *For a deterministic sequence* $\delta_n \to 0$,

$$\|\widehat{\Sigma}_X - \Sigma_X\|_{\mathrm{op}} = O_p(\delta_n), \qquad \|\Sigma_X\|_{\mathrm{op}} \le C_\Sigma.$$

The conditions have three roles. Condition 1 specifies the oracle target: the optimal dictionary-induced subspace is generated by s active atoms, and the active Gram matrix is uniformly nonsingular. Conditions 2–4 control the prediction part of the empirical criterion, through uniform concentration over the $\binom{K}{s}$ candidate supports and approximation of the oracle regression function on the active s-dimensional coordinates. Condition 5 controls the plug-in error from replacing Σ_X by $\widehat{\Sigma}_X$ in the covariance-separation penalty. These roles correspond to the three terms in the bound: nonparametric prediction error on the active coordinates, dictionary-support selection error, and covariance-estimation error. The additional assumptions $K \le n^c$ and $\lambda\delta_n \lesssim \sqrt{\log n/(sn)}$ are used only for the simplified display, where the latter two terms are absorbed into the $\sqrt{s\log n/n}$ remainder.

B.2 Proof of Theorem 1

Proof. Let $\mathbb{P}_n$ and $\mathbb{P}$ denote the empirical and population measures. Define

$$A_{\mathrm{pred}} = \sup_{S\in\mathcal{A}_s,\ f\in\mathcal{F}_{n,s}} \left|(\mathbb{P}_n - \mathbb{P})\ell\{Y, f(Q_S^\top X)\}\right|,$$

and

$$A_{\mathrm{cov}} = \sup_{S\in\mathcal{A}_s} |\widehat{\mathcal{C}}(P_S) - \mathcal{C}(P_S)|, \qquad A_{\mathrm{app}} = \inf_{f\in\mathcal{F}_{n,s}} \{\mathcal{L}_{S_\star}(f) - M(S_\star)\}.$$

For any $f_0 \in \mathcal{F}_{n,s}$, empirical optimality gives

$$\widehat{\mathcal{L}}_{\widehat{S}}(\widehat{f}) \le \widehat{\mathcal{L}}_{S_\star}(f_0).$$

Adding and subtracting the population criteria yields

$$\begin{aligned}\mathcal{L}_{\widehat{S}}(\widehat{f}) - M(S_\star) &\le \{\mathcal{L}_{\widehat{S}}(\widehat{f}) - \widehat{\mathcal{L}}_{\widehat{S}}(\widehat{f})\} + \{\widehat{\mathcal{L}}_{S_\star}(f_0) - \mathcal{L}_{S_\star}(f_0)\} \\ &\quad + \{\mathcal{L}_{S_\star}(f_0) - M(S_\star)\} \\ &\le 2A_{\mathrm{pred}} + 2\lambda A_{\mathrm{cov}} + \{\mathcal{L}_{S_\star}(f_0) - M(S_\star)\}.\end{aligned}$$

Taking the infimum over $f_0 \in \mathcal{F}_{n,s}$ gives

$$\mathcal{L}_{\widehat{S}}(\widehat{f}) - M(S_\star) \le 2A_{\mathrm{pred}} + 2\lambda A_{\mathrm{cov}} + A_{\mathrm{app}}. \tag{6}$$

The three terms are controlled separately. By Lemma 1,

$$A_{\mathrm{pred}} = O_p\left(\frac{W_n}{\sqrt{n}} + \sqrt{\frac{s\log(eK/s)}{n}}\right).$$

By Lemma 2,

$$A_{\mathrm{app}} = O(W_n^{-2\alpha/s}).$$

By Lemma 3,

$$\lambda A_{\mathrm{cov}} = O_p\{\lambda(s\delta_n + s\delta_n^2)\}.$$

Substitution into (6) gives

$$\mathcal{L}_{\widehat{S}}(\widehat{f}) - M(S_\star) = O_p\left(W_n^{-2\alpha/s} + \frac{W_n}{\sqrt{n}} + \sqrt{\frac{s\log(eK/s)}{n}} + \lambda\{s\delta_n + s\delta_n^2\}\right).$$

Choosing $W_n \asymp n^{s/(4\alpha+2s)}$ balances $W_n^{-2\alpha/s}$ and $W_n/\sqrt{n}$. Since $\log(eK/s) \lesssim \log K$ for $1 \le s \le K$ and $K \ge 2$, this proves the theorem. □

Lemma 1 (Prediction deviation over dictionary supports)**.** *Under Conditions 2 and 4,*

$$A_{\text{pred}} = O_p\left(\frac{W_n}{\sqrt{n}} + \sqrt{\frac{s\log(eK/s)}{n}}\right).$$

Here and in Lemma 2, logarithmic factors are suppressed.

Proof. For any $S \in \mathcal{A}_s$ and $f \in \mathcal{F}_{n,s}$, write

$$\ell\{Y, f(Q_S^\top X)\} = \ell(Y,0) + \Big[\ell\{Y, f(Q_S^\top X)\} - \ell(Y,0)\Big].$$

The first term satisfies

$$|(\mathbb{P}_n - \mathbb{P})\ell(Y,0)| = O_p(n^{-1/2})$$

by Condition 2. For the centered classes, define

$$\mathcal{D}_S = \left\{(x,y) \mapsto \ell\{y, f(Q_S^\top x)\} - \ell(y,0) : f \in \mathcal{F}_{n,s}\right\},$$
$$\mathcal{D} = \bigcup_{S\in\mathcal{A}_s} \mathcal{D}_S.$$

Each $h \in \mathcal{D}$ is bounded by $L_\ell C_f$. The covering bound for fixed-width/deep fully connected ReLU sieves gives, uniformly over empirical measures,

$$\log N\{\varepsilon, \mathcal{D}, L_1(P_n)\} \le CW_n^2 \log^c(n/\varepsilon) + s\log(eK/s),$$

where $|\mathcal{A}_s| \le \binom{K}{s} \le (eK/s)^s$ (Kohler & Langer, 2021, Supplement B, Lemma 19). Symmetrization and a bounded empirical-process maximal inequality therefore give

$$\sup_{h\in\mathcal{D}} |(\mathbb{P}_n - \mathbb{P})h| = O_p\left(\frac{W_n}{\sqrt{n}} + \sqrt{\frac{s\log(eK/s)}{n}}\right).$$

Combining the centered bound with the anchor term proves the lemma. □

Lemma 2 (Approximation on active concept coordinates)**.** *Under Conditions 2–4,*

$$A_{\text{app}} = O(W_n^{-2\alpha/s}).$$

Proof. Let $Z = Q_{S_\star}^\top X$, choose $a_n \asymp \sqrt{\log n}$, and set $\Omega_{a_n} = [-a_n, a_n]^s$. By Kohler–Langer's ReLU approximation theorem (Kohler & Langer, 2021, Corollary 1), after output clipping there is $f_n \in \mathcal{F}_{n,s}$ such that

$$\sup_{z\in\Omega_{a_n}} |f_n(z) - g_\star(z)| \lesssim a_n^\alpha W_n^{-2\alpha/s} =: \varepsilon_n.$$

The factor a_n^α is logarithmic and is absorbed in the suppressed logarithmic factors. Since f_n and $g_\star$ are bounded and Z is sub-Gaussian,

$$\mathbb{E}|f_n(Z) - g_\star(Z)| \le \varepsilon_n + 2C_f\mathbb{P}(Z \notin \Omega_{a_n}) \lesssim \varepsilon_n.$$

The covariance penalty is fixed on $S_\star$, and $M(S_\star) = \mathcal{L}_{S_\star}(g_\star)$. Therefore Lipschitzness of the loss gives

$$\mathcal{L}_{S_\star}(f_n) - M(S_\star) \lesssim \mathbb{E}|f_n(Z) - g_\star(Z)| = O(W_n^{-2\alpha/s}).$$

Taking the infimum over $f \in \mathcal{F}_{n,s}$ proves the claim. □

Lemma 3 (Covariance penalty perturbation)**.** *Under Condition 5,*

$$\lambda A_{\text{cov}} = O_p\{\lambda(s\delta_n + s\delta_n^2)\}.$$

Proof. Write $\Delta_n = \widehat{\Sigma}_X - \Sigma_X$. For any rank-$s$ projector P,

$$\widehat{\mathcal{C}}(P) - \mathcal{C}(P) = \|P\Sigma_X(I-P) + P\Delta_n(I-P)\|_F^2 - \|P\Sigma_X(I-P)\|_F^2.$$

Hence

$$|\widehat{\mathcal{C}}(P) - \mathcal{C}(P)| \le 2\|P\Sigma_X(I-P)\|_F\|P\Delta_n(I-P)\|_F + \|P\Delta_n(I-P)\|_F^2.$$

Since P has rank s, Condition 5 gives

$$\|P\Sigma_X(I-P)\|_F \le \sqrt{s}\,\|\Sigma_X\|_{\mathrm{op}} \le \sqrt{s}\,C_\Sigma,$$

while

$$\|P\Delta_n(I-P)\|_F \le \sqrt{s}\,\|\Delta_n\|_{\mathrm{op}}.$$

Therefore, uniformly over $S \in \mathcal{A}_s$,

$$|\widehat{\mathcal{C}}(P_S) - \mathcal{C}(P_S)| \le 2sC_\Sigma\|\Delta_n\|_{\mathrm{op}} + s\|\Delta_n\|_{\mathrm{op}}^2.$$

Using $\|\Delta_n\|_{\mathrm{op}} = O_p(\delta_n)$ and multiplying by λ yields the stated bound. □

B.3 Proof of Proposition 1

Proof. Let P_B be the projector onto $\mathrm{span}(A_{S_\star}B)$. Since $A_{S_\star}$ has orthonormal columns and R is orthogonal to $\mathrm{span}(A_{S_\star})$, the projected vector P_BX determines $B^\top C_\star$. Also, $P_\star X = A_{S_\star}C_\star$ determines $C_\star$, while the residual vector R is independent of $(C_\star, Y)$. Thus the best risk attainable from projections of X is $R_C^\star$. By assumption, $B^\top C_\star$ attains this risk, so

$$R(P_B) = R_C^\star = R^\star, \qquad \mathrm{rank}(P_B) = q,$$

so $m_{\mathrm{pred}} \le q$. Conversely, let P be a rank-r projector. Then

$$PX = PA_{S_\star}C_\star + PR.$$

The signal part is a rank-at-most-r linear image of $C_\star$, and PR is independent of $(C_\star, Y)$. Thus observing PX cannot improve on the oracle risk attainable from a rank-at-most-r linear summary of $C_\star$. If $r < q$, the definition of q implies that this risk is larger than $R_C^\star = R^\star$. Therefore no rank smaller than q can attain $R^\star$, and

$$m_{\mathrm{pred}} = q.$$

For the preservation dimension, note that

$$D(P; P_\star) = 0 \iff (I-P)P_\star = 0 \iff \mathrm{range}(P_\star) \subseteq \mathrm{range}(P).$$

Thus any projector with zero concept omission has rank at least s. The projector $P_\star$ itself has rank s and attains $R^\star$, since it retains all coordinates $C_\star$ in the active concept subspace. Hence

$$m_{\mathrm{pres}} = s.$$

Combining the two identities gives

$$\Delta_{\mathrm{pres}} = m_{\mathrm{pres}} - m_{\mathrm{pred}} = s - q.$$

□

B.4 Covariance Separation and Concept Preservation

For a symmetric covariance matrix Σ_X, a subspace is reducing if it and its orthogonal complement are invariant under Σ_X. The Σ_X-envelope of a subspace $\mathcal{M}$, denoted by $\mathcal{E}_{\Sigma_X}(\mathcal{M})$, is the smallest reducing subspace containing $\mathcal{M}$ (Cook et al., 2013). The following result connects this construction to the population objective of Concept Envelope.

Proposition 2 (Covariance separation and concept preservation)**.** *Under the model of Proposition 1, suppose X has finite second moments and $R^\star$ is finite. Write $\mathcal{M} = \mathrm{span}(A_{S_\star}B)$, and assume that every orthogonal projector attaining $R^\star$ contains $\mathcal{M}$:*

$$R(P) = R^\star \implies \mathcal{M} \subseteq \mathrm{range}(P).$$

Let $\mathcal{E} = \mathcal{E}_{\Sigma_X}(\mathcal{M})$, $e = \dim(\mathcal{E})$, and fix a retained dimension $e \le u \le d$. Among rank-u orthogonal projectors, the minimizers of

$$R(P) + \lambda\mathcal{C}(P), \qquad \mathcal{C}(P) = \|P\Sigma_X(I-P)\|_F^2,$$

are precisely those containing $\mathcal{M}$ when $\lambda = 0$, and precisely those onto reducing subspaces containing $\mathcal{E}$ when $\lambda > 0$. In particular, at $u = e$, every $\lambda > 0$ has the unique minimizer given by the orthogonal projector onto $\mathcal{E}$.

Writing $\Sigma_C = \mathrm{Cov}(C_\star)$, the envelope is

$$\mathcal{E} = A_{S_\star} \operatorname{span}\big(B, \Sigma_C B, \ldots, \Sigma_C^{s-1} B\big).$$

Consequently, if

$$\operatorname{rank}\big[B, \Sigma_C B, \ldots, \Sigma_C^{s-1} B\big] = s,$$

then at $u = s$ every $\lambda > 0$ uniquely selects $P_\star$. At the same dimension, prediction alone is minimized by every projector containing the q-dimensional $\mathcal{M}$.

Proof. Since Σ_X is symmetric, $\mathcal{C}(P) = 0$ if and only if $P\Sigma_X = \Sigma_X P$, or equivalently $\operatorname{range}(P)$ reduces Σ_X. Any projector containing $\mathcal{M}$ attains $R^\star$, because its projection determines $B^\top C_\star$, as in the proof of Proposition 1. Together with the assumed necessity of $\mathcal{M}$, this proves the assertion for $\lambda = 0$.

The envelope $\mathcal{E}$ is reducing. Its orthogonal complement has an orthonormal eigenbasis for the restriction of Σ_X. Adjoining any $u - e$ of these eigenvectors to $\mathcal{E}$ therefore gives a rank-u subspace with risk $R^\star$ and zero covariance penalty. Thus the minimum objective equals $R^\star$. For $\lambda > 0$, any minimizer must separately attain $R(P) = R^\star$ and $\mathcal{C}(P) = 0$. Its range consequently contains $\mathcal{M}$ and reduces Σ_X, so it contains $\mathcal{E}$ by the definition of the envelope. Conversely, every reducing subspace containing $\mathcal{E}$ attains both equalities. When $u = e$, containment forces equality of the subspaces and hence uniqueness of the orthogonal projector.

Independence and orthogonality of the residual in Proposition 1 give

$$\Sigma_X A_{S_\star} = A_{S_\star} \Sigma_C.$$

It follows that every Σ_X-invariant subspace containing $\mathcal{M}$ also contains the columns of $A_{S_\star} \Sigma_C^j B$ for all $j \geq 0$. By the Cayley–Hamilton theorem, the span for $0 \leq j < s$ already contains all these columns and is itself invariant. Symmetry makes it reducing, proving the displayed formula for $\mathcal{E}$. Under the rank condition, $\mathcal{E} = \operatorname{range}(P_\star)$ and $e = s$, so the uniqueness result applies at $u = s$. □

For example, the rank condition holds when Σ_C has distinct eigenvalues and every eigenvector has a nonzero projection onto $\operatorname{span}(B)$. Covariance separation then retains all s concept directions even though only $q < s$ directions are needed for prediction.