# FLACE: EXPLAINING FROZEN CLASSIFIERS THROUGH SIGNED FEATURE-SPACE BOTTLENECKS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Explaining a frozen classifier through intermediate coordinates requires an explicit link between those coordinates and its prediction. FLACE builds this link by learning a signed dictionary from the classifier's own features with classical semi-NMF, inferring nonnegative reconstruction coordinates, and reusing the original linear head. Each coordinate has an exact signed contribution to the reconstructed class logit, equal to the logit change on deleting that coordinate and its zero-baseline integrated gradient. This gives a direct interface for inspecting and editing reconstructed predictions without concept labels or an external text dictionary. On CIFAR-100, CUB, and ImageNet-100 with EfficientNet-B3, the evaluated native prediction paths show lower within-basis correlations and higher insertion AUC for FLACE than for the compared CBMs, with mixed deletion results. The coordinates weight numerical feature directions; optional names aid visual inspection without defining their meaning. Core implementation code is provided in the supplementary material.

## 1 INTRODUCTION

Concept-based explanations are useful when they expose intermediate variables that participate in a model's prediction rather than only highlighting input regions or final logits Kim et al. (2018); Ghorbani et al. (2019). Concept Bottleneck Models (CBMs) formalize this idea by routing prediction through a concept layer, which supports both concept-level inspection and feature-space intervention Koh et al. (2020). Classical CBMs, however, require dense concept annotations, making them difficult to apply when task-relevant concepts are unknown, expensive to label, or too fine-grained for a fixed annotation schema Koh et al. (2020); Yuksekgonul et al. (2023).

Recent label-free CBMs reduce the need for manual concept labels by aligning images with language descriptions, often through CLIP or related vision–language models Radford et al. (2021); Oikarinen et al. (2023); Yang et al. (2023). This strategy is powerful, but it makes the concept space depend on a predefined text pool or an external semantic embedding. If a task-relevant visual factor is missing from that vocabulary, the model may ignore it or force it into an ill-fitting textual category. More importantly, explanations obtained in the external VLM space may not faithfully reflect how the target classifier organizes its own internal features.

FLACE addresses a concrete question: how can a classifier's own features be decomposed into coordinates whose contributions to its predictions can be inspected and edited? Each coordinate is a nonnegative weight on a learned signed feature direction. Its activation specifies how much of that direction enters the reconstructed feature, and its class-specific coefficient determines the associated logit contribution. We use "concept coordinate" in this operational sense throughout. Figure 1 illustrates how optional names can be attached after inspecting highly activating images.



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.

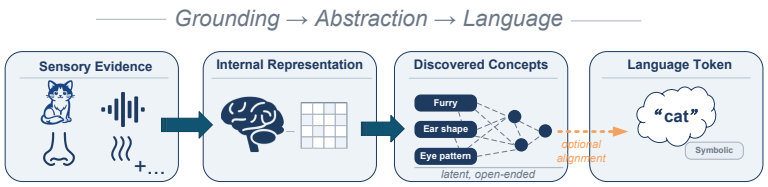



Figure 1: **Features → Coordinates → Optional names.** FLACE first discovers coordinates inside the target model's feature space and then allows optional post-hoc naming. The attribute names in this conceptual diagram illustrate the naming stage; they are not validated outputs of the factorization. Language is therefore treated as an interpretation aid rather than as the source of the concept basis.

Mixed-sign features motivate a signed dictionary with nonnegative coefficients, a classical semi-NMF parameterization Ding et al. (2010). FLACE connects this decomposition to a fixed test-time inference rule and the original linear classifier head. This makes each reconstructed logit an exact sum of coordinate contributions (Sec. 4.1); a negative basis entry alone does not imply opposition to a class. Rescaling fixes a display convention, while naming remains optional.

**Contributions.**

- A post-hoc feature-space bottleneck for a frozen classifier, joining classical semi-NMF, nonnegative test-time inference, and the original prediction head without concept labels or an external text dictionary.
- An explicit account of how coordinates compose each reconstructed logit, with exact deletion and logit-IG identities (valid for any fixed affine decoder/head), and conditions under which rescaling preserves inference and attribution.
- A three-dataset evaluation of coordinate insertion/deletion and within-basis correlation, complemented by sensitivity studies and selected activation grids for visual inspection.

## 2 Related Work

**Concept bottlenecks.** Classical CBMs predict through annotated concepts Koh et al. (2020); CEMs learn richer supervised concept embeddings Espinosa Zarlenga et al. (2022). Post-hoc CBMs project frozen features onto concept axes and learn a concept head Yuksekgonul et al. (2023). Label-free and language-guided CBMs obtain supervision from a text pool and a vision–language model Oikarinen et al. (2023); Yang et al. (2023); residual variants add capacity beyond the concept path Shang et al. (2024). FLACE instead reconstructs the target representation before the original head. It supports numerical coordinate editing, whereas human concept correction additionally requires validated coordinate meanings.

**Decomposition and sparse coding.** NMF is established part-based representation learning Lee & Seung (1999); CRAFT applies nonnegative factorization to neural activations and recursively refines concepts Fel et al. (2023). Semi-NMF permits mixed-sign features and dictionary vectors Ding et al. (2010). Closely related, Parekh et al. (2024) learn a signed dictionary for multimodal-model token representations using nonnegative sparse coefficients, an $\ell_1$ penalty, and dictionary-norm constraints, then ground it through images and vocabulary projections. These constraints affect the learned solution; FLACE's post-fit rescaling does not. SpLiCE Bhalla et al. (2024) sparsely reconstructs CLIP embeddings with a fixed text dictionary, so language defines its coordinates from the outset. Sparse autoencoders (SAEs) learn an encoder and decoder; TopK variants explicitly constrain active width Gao et al. (2025). FLACE uses optimization-based inference with no explicit sparsity penalty or learned encoder; the recovered training fit includes weight decay. Matching dictionary size alone would not match active-coordinate count, reconstruction quality, or computational cost.

FLACE combines these established decomposition tools into an interface for inspecting a frozen classifier through editable reconstruction coordinates. Table 1 summarizes the mechanisms of the closest alternatives; empirical comparisons are presented in Sec. 5.2.

Table 1: Closest decomposition interfaces. All four permit signed dictionary entries. "Encoder" means an additional learned coordinate encoder, beyond the feature extractor. Operations are mechanisms, not claims of human concept correction.

| Method | Input / dictionary | Objective / inference | Output / language |
|---|---|---|---|
| Parekh et al. | LMM tokens; learned | Sparse semi-NMF; nonnegative sparse solve; no encoder | Grounded activations; vocabulary via unembedding |
| SpLiCE | CLIP; fixed text vectors | Sparse recovery; nonnegative sparse solve; no extra encoder | Reconstructed CLIP; language defines basis |
| TopK SAE | Model activations; learned | Reconstruction with TopK; learned encoder | Reconstructed activations; optional later inspection |
| FLACE | Target pooled features; learned | Semi-NMF; NNLS; no encoder | Original linear head; optional later names |

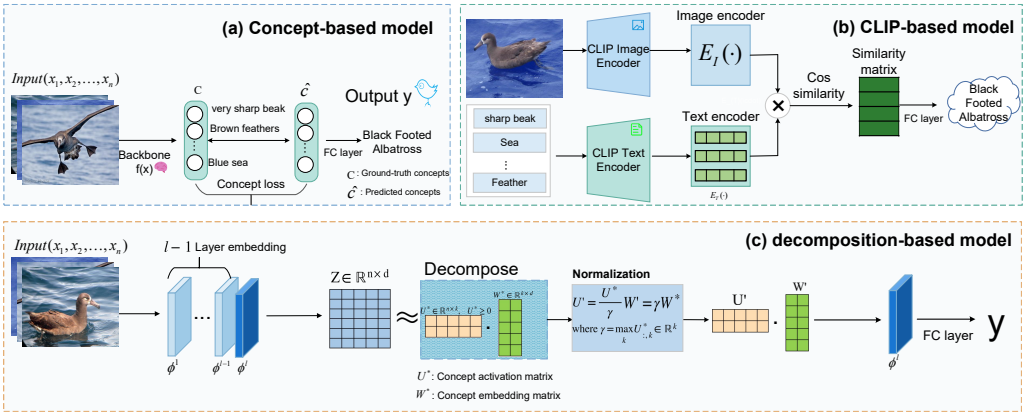



Figure 2: Overview of FLACE compared with traditional CBMs and VLM-aligned label-free CBMs. FLACE learns concept activations and a signed feature-space basis from frozen target-model features, then uses NNLS to infer nonnegative test-time activations against the learned basis; the recovered solver is a finite projected-Adam approximation.

## 3 Problem Definition

Let $f$ be a trained image classifier. We write it as a composition of a feature extractor and a final prediction head, $f(x) = \varphi^l(\varphi^{1\sim(l-1)}(x))$, where $z = \varphi^{1\sim(l-1)}(x) \in \mathbb{R}^d$ denotes the intermediate feature used for concept discovery. Given $n$ samples, their frozen features are stacked into $Z \in \mathbb{R}^{n\times d}$.

Our goal is to learn a concept coordinate system for this fixed feature space. Specifically, we seek nonnegative concept activations $U \in \mathbb{R}_+^{n\times k}$ and a signed concept basis $W \in \mathbb{R}^{k\times d}$ such that $UW$ reconstructs the target features used by the classifier. At test time, a new feature vector should be mapped to a nonnegative concept activation vector and then reconstructed through the same basis before being passed to the original prediction head. This setting makes concept intervention a feature-space operation on the learned coordinates rather than a language-space similarity query.

## 4 Our Method

As shown in Fig. 2, FLACE builds a post-hoc feature-space bottleneck from a trained classifier. We factor its frozen features into nonnegative coordinates and a signed dictionary, then rescale each coordinate relative to its observed training maximum while preserving the reconstructed feature. At test time, coordinates are inferred by nonnegative reconstruction and decoded through the original prediction head. Editing these coordinates changes the reconstructed prediction through the same

fixed head. We define the exact NNLS interface below; Appendix A specifies the available finite projected-Adam implementation.

We split the trained classifier as $\varphi(x) = \varphi^l(\varphi^{1\sim(l-1)}(x))$ and stack frozen intermediate features into $Z \in \mathbb{R}^{n\times d}$. We then perform flexible semi-NMF Ding et al. (2010) with the objective below. The iterative algorithm returns approximate factors; the starred quantities denote global minimizers when attained:

$$(U^*, W^*) \in \arg\min_{U,W} \|Z - UW\|_F^2, \quad \text{s.t. } U \geq 0. \tag{1}$$

The objective in Eq. (1) is unsupervised with respect to concept labels. Unlike conventional NMF Lee & Seung (1999), it keeps the activation matrix nonnegative while allowing the basis matrix $W$ to contain signed entries. The nonnegative rows of $U$ can therefore be read as concept activation strengths, while each row of $W$ represents a direction in the original feature space that may increase some feature entries and decrease others. We use the term *contrastive feature evidence* in a strictly representational sense: a negative basis entry merely shifts the reconstructed feature in the corresponding coordinate, and is not a claim that the negative direction corresponds to a human-semantic "inhibitory" or "opposite" concept. Appendix D discusses the representational motivation for using an unconstrained basis with mixed-sign neural features.

Next, we apply a per-concept rescaling to the concept coefficient matrix and the concept basis matrix:

$$\gamma_r = \max_{1\leq i\leq n} \widehat{U}_{ir} + \varepsilon, \quad U'_{ir} = \frac{\widehat{U}_{ir}}{\gamma_r}, \quad W'_{r:} = \gamma_r \widehat{W}_{r:}, \tag{2}$$

Here $(\widehat{U}, \widehat{W})$ are fitted factors, not assumed global minimizers, and $\varepsilon > 0$. Writing $\Gamma = \mathrm{diag}(\gamma_r)$ gives $U' = \widehat{U}\Gamma^{-1}$, $W' = \Gamma\widehat{W}$ and $U'W' = \widehat{U}\widehat{W}$. The training coefficients have maxima below one (or zero for inactive columns); test coefficients need not. This supplies a display and magnitude-threshold convention, not sparsification or semantic identifiability. For every positive diagonal $\Gamma$, the nonnegative reconstruction cones coincide:

$$\{uW : u \geq 0\} = \{v\Gamma W : v \geq 0\}, \qquad v = u\Gamma^{-1}. \tag{3}$$

Hence exact NNLS has the same optimal reconstruction error after rescaling. Its reconstructed vector is unique, but its coefficients need not be; a finite-iteration solver need not return corresponding solutions. Zero-baseline IG and deletion of corresponding coordinates are already invariant under this reparameterization. Appendix J proves these statements and explains the finite-$\varepsilon$ qualification.

The ideal test-time interface fixes the normalized concept basis $W'$ and defines an exact NNLS solution for each test feature by:

$$u_i^{\text{test}} \in \arg\min_{u\geq 0} \|z_i - uW'\|_2^2, \quad \tilde{z}_i = u_i^{\text{test}} W'. \tag{4}$$

In this interface, we feed the reconstructed feature from Eq. (4) into the original prediction head $\varphi^l(\cdot)$ to obtain $y = \varphi^l(\tilde{z}_i)$. In this setting, $u_i^{\text{test}}$ represents the nonnegative concept activations of sample $i$, and $W'$ represents signed feature-space concept directions. The analogy to Concept Embedding Models (CEMs) Espinosa Zarlenga et al. (2022) is limited to the use of learned concept-related vectors; FLACE is post-hoc, does not require supervised concept embeddings, and performs naming only as an optional visualization step. Appendix E provides qualitative concept visualizations.

### 4.1 Exact composition of a reconstructed prediction

Let the frozen class-$c$ head be $z \mapsto za_c + b_c$, with column vector $a_c \in \mathbb{R}^d$. For a row vector $u \geq 0$, define $\theta_{rc} = W'_{r:}a_c$. Then

$$f_c(u) = b_c + \sum_{r=1}^{k} q_{rc}, \qquad q_{rc} = u_r\theta_{rc}. \tag{5}$$

Each $q_{rc}$ is a signed contribution to the reconstructed logit. Its sign depends on the head as well as the basis, not on a single entry of $W'$. If $u^{(-r)}$ sets only coordinate $r$ to zero, linearity gives

$$f_c(u) - f_c(u^{(-r)}) = q_{rc} = \mathrm{IG}_{r,c}(u; 0). \tag{6}$$

Indeed, the gradient along $tu$, $0 \le t \le 1$, is constant, $\partial f_c/\partial u_r = \theta_{rc}$; integrating and multiplying by $u_r$ proves the IG identity Sundararajan et al. (2017). Thus $\sum_r q_{rc} = f_c(u) - b_c$ exactly. Softmax probabilities are not additive and do not obey this logit identity.

**Coordinate editing.** An analyst (or our evaluation routine) changes selected entries of an already inferred $u$, holds other entries fixed, reconstructs $uW'$, and recomputes the frozen head; NNLS is not re-solved after an edit. For example, setting coordinate 347 to zero changes logit $c$ by $-q_{347,c}$. Turning this into "correct the bird's beak color" requires a stable semantic mapping that this algebra does not supply. The explanation is an additive numerical account, not a sparse logical rule. It explains the reconstructed path; the original logit also includes residual $(z - uW')a_c$, whose magnitude is at most $\|z - uW'\|_2 \|a_c\|_2$.

# 5 Experiments

## 5.1 Datasets and Models

We evaluate on three image classification benchmarks: CIFAR-100 Krizhevsky et al. (2009) (100 classes, 50k/10k), CUB-200-2011 Wah et al. (2011) (200 classes, 5,994/5,794), and ImageNet-100 Deng et al. (2009) (100-class subset, 10k/5k). Unless stated otherwise, we use the official splits with standard augmentation (random crop, horizontal flip) and ImageNet normalization. We train an EfficientNet-B3 Tan & Le (2019) classifier for each dataset and freeze it for evaluation. The decomposition uses the pooled penultimate feature $z \in \mathbb{R}^{1536}$ before the final affine head, which is reused without retraining. EfficientNet's Swish/SiLU activation, $t\sigma(t)$, permits negative values, and global pooling does not impose nonnegativity. This motivates a signed dictionary. Feature extraction is performed in evaluation mode, with dropout inactive. Factorization uses no task labels; labels enter classifier training and evaluation. Implementation versions and dataset-record details are given in Appendices A and I. Raw classifier test accuracies, before decomposition, are 81.31% (CIFAR-100), 81.01% (CUB), and 88.80% (ImageNet-100), with training accuracies of 94.46%, 99.03%, and 99.95%. These are reference accuracies of the original feature path, not reconstructed-path fidelity measurements.

## 5.2 Baselines

For a comprehensive comparison, we evaluate four representative CBM variants that differ in supervision, external dependence, and concept inference mechanism: (1) **Res-CBM** Shang et al. (2024): a residual concept-bottleneck framework that uses explicit concept vectors for class prediction while retaining a residual pathway for incomplete concepts; (2) **Post-hoc CBM** Yuksekgonul et al. (2023): a post-hoc concept discovery and aggregation approach applied to frozen features without retraining the original classifier; (3) **Label-free-CBM** Oikarinen et al. (2023); Yang et al. (2023): a concept-label-free approach that typically uses text-derived concept descriptors or external semantic alignment to construct the concept interface; (4) **CF-CBM** Panousis et al. (2024): a coarse-to-fine concept modeling strategy designed to reduce instability from a single fixed concept granularity. FLACE is evaluated under the same datasets and backbone setting to compare feature-space concept discovery against these representative alternatives. **Protocol matching: shared experimental controls.** All four baselines and FLACE share the same EfficientNet-B3 backbone, the same train/test splits, the same pre-cached features, and the same IG-based ranking procedure (Eq. (8)–(9)) and Insertion/Deletion AUC computation. Each method then performs the intervention along its *own native bottleneck-to-head path*: FLACE uses $\text{NNLS} \to W' \to \varphi^l$ (the original frozen classifier head), Label-free-CBM uses text-pool similarity $\to$ learned linear head, Res-CBM uses its supervised concept-plus-residual head, Post-hoc CBM uses its post-hoc projected linear head, and CF-CBM uses its coarse-to-fine concept-aggregation head. We deliberately keep each method's own head rather than retro-fitting all methods through $\varphi^l$, so that each method is evaluated through its specified prediction path. Consequently this comparison does not isolate the decomposition from head choice.

**Concept counts and correlation comparability.** Table 4 lists each method's reported concept count and native interface. Neither dictionary width nor active-coordinate count is equalized. Correlation measures within-basis association in each method's own representation space, so cross-method differences do not establish a common geometric or semantic quality ranking.

### 5.3 HYPERPARAMETER SETUP

The main-table settings are $(k, T) = (600, 3000)$ for CIFAR-100, $(700, 3000)$ for CUB, and $(900, 3000)$ for ImageNet-100, with seed 42 and $\varepsilon = 10^{-7}$. The reported IG protocol uses 20 steps, a zero baseline and batch size 64; Eq. (6) gives the exact FLACE logit attribution without numerical integration. The optimizer used to produce the main table is not established by the available code record. Appendix A specifies a separate projected-Adam implementation, and Sec. 5.6 reports its separate fidelity measurements, with result sources detailed in Appendix K.

**Finite-solver interpretation.** The documented implementation fits the training factors with 3000 Adam steps and infers test coordinates with 300 steps against the raw basis before rescaling. It uses weight decay in training and a positive coefficient floor in both stages, without a certified NNLS stopping tolerance. These choices affect the returned coordinates, but Eq. (5) remains exact for those coordinates. Full optimizer settings are in Appendix A. Nonunique NNLS solutions can assign different contributions to the same reconstruction, so an explanation must retain its inferred coordinates.

### 5.4 EVALUATION METRICS

*Overview.* We evaluate three complementary properties of the learned concepts that align with our problem setup and reconstruction pathway $\tilde{\mathbf{z}}_i = \mathbf{u}_i W'$ with $p_i(\mathbf{z}) = \mathrm{softmax}(\varphi^l(\mathbf{z}))_{y_i}$. (1) *Concept Importance Ranking (IG)* is used to derive a global ordering of coordinates with the largest positive ground-truth-logit contributions; IG itself is not a performance score but provides the permutation $\pi$ consumed by downstream metrics. The ground-truth class is used here only to keep the evaluation protocol consistent across methods; for unlabeled inspection at deployment time, the same ranking machinery can be applied with respect to the predicted class or any user-specified target class without changing concept discovery, the basis $W'$, or the NNLS test-time inference. (2) *Insertion/Deletion (ID) curves with AUC* quantify the *usability* of this ordering: a larger $\mathrm{AUC}_{\mathrm{ins}}$ records higher average target confidence over the insertion path, while a smaller $\mathrm{AUC}_{\mathrm{del}}$ records lower average target confidence over the deletion path. Neither specifies a sufficient coordinate count. (3) *Concept Basis Correlation* measures basis-level redundancy among the rows of $W'$ via Pearson, Cosine, and Spearman coefficients; lower averages indicate lower average pairwise association among concept directions, rather than direct proof of human-semantic disentanglement or joint linear independence. Therefore, better overall behavior corresponds to **larger** $\mathrm{AUC}_{\mathrm{ins}}$, **smaller** $\mathrm{AUC}_{\mathrm{del}}$, and **smaller** correlation statistics. Deletion probabilities can increase when competing logits fall faster, so neither monotonicity nor an upper bound by initial confidence is guaranteed. Starting confidence and native head calibration affect both AUCs; read them jointly rather than treating either as a semantic quality score.

For *FLACE* specifically, the reconstructed feature $\tilde{\mathbf{z}}_i = \mathbf{u}_i W'$ is passed through the original frozen head $\varphi^l$, and the target-class probability $p_i(\mathbf{z}) = \mathrm{softmax}(\varphi^l(\mathbf{z}))_{y_i}$ is used in the insertion/deletion curves below. For each baseline, the same insertion/deletion ranking protocol is applied along its *own* native bottleneck-to-head path (Sec. 5.2), so that no method is forced through a head it was not trained for.

**Concept Importance Ranking (IG)** For sample $i$ and its ground-truth class $y_i$, define

$$f_{y_i}(\mathbf{u}) \triangleq \big(\varphi^l(\mathbf{u}W')\big)_{y_i}. \tag{7}$$

and take the zero baseline $\mathbf{u}^0 = \mathbf{0}$. Then, using Eq. (7), the Integrated Gradients (IG) importance of the $r$-th concept ($r = 1, \ldots, K$) is

$$\begin{aligned} \alpha_{ir} &= (u_{ir} - u_r^0) \int_0^1 \frac{\partial f_{y_i}\big(\mathbf{u}^0 + t(\mathbf{u}_i - \mathbf{u}^0)\big)}{\partial u_r}\, dt \\ &\approx \frac{u_{ir} - u_r^0}{T} \sum_{t=1}^{T} \left.\frac{\partial f_{y_i}(\mathbf{u})}{\partial u_r}\right|_{\mathbf{u}=\mathbf{u}^0+\frac{t}{T}(\mathbf{u}_i-\mathbf{u}^0)}. \end{aligned} \tag{8}$$

Table 2: Native-path accuracy, within-basis correlation and insertion/deletion AUC. Bold marks the best value within each dataset. Methods use their own heads, spaces and coordinate counts (Table 4); these are descriptive system comparisons.

| Dataset | Method | Native acc. (%) | Basis correlation ↓ | | | AUC | |
|---|---|---|---|---|---|---|---|
| | | | Pearson | Cosine | Spearman | Ins. ↑ | Del. ↓ |
| CIFAR-100 | Res-CBM | 80.75 | 0.726 | 0.726 | 0.187 | 0.507 | 0.259 |
| | Post-hoc CBM | 77.80 | 0.774 | 0.775 | 0.198 | 0.029 | 0.021 |
| | Label-free-CBM | 76.55 | 0.774 | 0.774 | 0.198 | 0.453 | 0.117 |
| | CF-CBM | 20.86 | 0.774 | 0.774 | 0.198 | 0.083 | **0.020** |
| | FLACE | — | **0.050** | **0.049** | **0.044** | **0.562** | 0.335 |
| CUB | Res-CBM | 73.59 | 0.654 | 0.655 | 0.265 | 0.389 | 0.230 |
| | Post-hoc CBM | 42.73 | 0.714 | 0.715 | 0.286 | 0.088 | 0.152 |
| | Label-free-CBM | 69.02 | 0.714 | 0.715 | 0.286 | 0.238 | 0.073 |
| | CF-CBM | 35.95 | 0.714 | 0.715 | 0.286 | 0.215 | **0.035** |
| | FLACE | — | **0.033** | **0.032** | **0.026** | **0.545** | 0.204 |
| ImageNet-100 | Res-CBM | 83.82 | 0.595 | 0.595 | 0.207 | 0.470 | 0.277 |
| | Post-hoc CBM | 78.84 | 0.761 | 0.761 | 0.254 | 0.078 | **0.051** |
| | Label-free-CBM | 78.84 | 0.761 | 0.761 | 0.254 | 0.354 | 0.153 |
| | CF-CBM | 52.22 | 0.761 | 0.761 | 0.254 | 0.370 | 0.061 |
| | FLACE | — | **0.023** | **0.023** | **0.021** | **0.797** | 0.186 |

FLACE's native accuracy is unavailable for these runs (—). Its raw classifier references, 81.31/81.01/88.80%, do not measure reconstruction fidelity. Separate fidelity measurements are reported in Table 3. Correlation and AUC are rounded to three decimals.

Averaging the positive logit contribution in Eq. (8) over samples yields the global ranking score:

$$s_r = \frac{1}{n}\sum_{i=1}^{n} \max(\alpha_{ir}, 0), \tag{9}$$

from which we obtain a descending ranking $\pi$ such that $s_{\pi_1} \geq s_{\pi_2} \geq \cdots \geq s_{\pi_K}$. The score in Eq. (9) ranks concepts by positive contributions to the ground-truth logit; negative logit attributions are excluded for this ordering. Unless explicitly stated otherwise, all Insertion/Deletion AUC values reported in Table 2 and the ablations use this single *global* ranking $\pi$ shared across all test samples; the ranking-estimation split is not documented and is not claimed held-out; the per-sample qualitative diagnostics in Appendix H use sample-specific rankings and are labeled accordingly.

**Usability Evaluation: Insertion / Deletion Curves and AUC Petsiuk et al. (2018).** Using the top-$m$ concept set $S_m = \{\pi_1, \ldots, \pi_m\}$, we construct two probability curves in the reconstructed feature space. **Insertion** gradually adds important concepts from the empty set (here $m = 0, 1, \ldots, K$):

$$\tilde{\mathbf{z}}_i^{\text{ins}}(m) = \sum_{r \in S_m} u_{ir}\, \mathbf{w}'_r, \quad \bar{q}^{\text{ins}}(m) = \frac{1}{n}\sum_{i=1}^{n} p_i\big(\tilde{\mathbf{z}}_i^{\text{ins}}(m)\big). \tag{10}$$

Here $u_{ir}$ is the activation strength of concept $r$ for sample $i$; inclusion is determined by membership in $S_m$.

**Deletion** gradually removes important concepts from the full set:

$$\tilde{\mathbf{z}}_i^{\text{del}}(m) = \sum_{r \notin S_m} u_{ir}\, \mathbf{w}'_r, \quad \bar{q}^{\text{del}}(m) = \frac{1}{n}\sum_{i=1}^{n} p_i\big(\tilde{\mathbf{z}}_i^{\text{del}}(m)\big). \tag{11}$$

The area under each curve (AUC) is computed using equally spaced trapezoidal integration and normalized by the number of steps:

$$\mathrm{AUC}_\star = \frac{1}{K}\sum_{m=1}^{K}\frac{\bar{q}^\star(m-1)+\bar{q}^\star(m)}{2}, \quad \star \in \{\mathrm{ins}, \mathrm{del}\}. \tag{12}$$

Equations (10)–(12) define the insertion/deletion evaluation used in Table 2.

**Redundancy Evaluation: Concept Basis Correlation** We measure redundancy by the mean absolute pairwise correlation among the rows of $W'$, computed over the upper triangle (excluding the diagonal) using three statistics: Pearson (mean-centered), Cosine (no centering), and Spearman (rank-transformed Pearson):

$$\bar{r}^{(\cdot)} = \frac{2}{K(K-1)}\sum_{1\le r<s\le K}\left|r^{(\cdot)}(r,s)\right|, \quad (\cdot)\in\{\mathrm{pear}, \mathrm{cos}, \mathrm{spear}\}. \tag{13}$$

Full formulas are given in Appendix C. The pairwise average requires $K > 1$. Zero/constant rows can make individual coefficients undefined; the historical handling of such rows and Spearman ties is unrecorded, so no replacement convention is silently imposed. A lower value in Eq. (13) indicates lower average pairwise association among concept bases, which we use as a proxy for reduced feature-space redundancy and not as a direct measure of human-semantic disentanglement. Combined with $\mathrm{AUC}_{\mathrm{ins}}$ and $\mathrm{AUC}_{\mathrm{del}}$, it provides a complementary view of whether concepts are useful for feature-space prediction and whether basis directions are well separated.

### 5.5 Experimental Results

Table 2 reports lower within-basis correlations for FLACE (order $10^{-2}$) and the largest insertion AUC among the evaluated CBMs (0.562/0.545/0.797). Deletion results vary: FLACE has higher AUC than Res-CBM on CIFAR-100 (0.335 versus 0.259, lower is better), and lower AUC on CUB and ImageNet-100 (0.204/0.186 versus 0.230/0.277); several other baselines have lower deletion AUC. These are comparisons of native prediction paths, whose coordinate counts, heads and confidence differ. They describe the evaluated systems rather than isolate dictionary quality. Appendix G reports variation across factorization repeats. Sec. 5.6 reports separate fidelity measurements. The PCA/SAE controls in Appendix K have strong AUCs but lack a same-protocol FLACE row with matching reconstruction diagnostics.

### 5.6 Reconstruction Fidelity

To assess how closely the coordinate representation preserves the frozen classifier's predictions, Table 3 compares the original and reconstructed feature paths through the same affine head. Accuracy measures agreement with ground-truth labels; prediction agreement measures whether the two paths select the same class, regardless of correctness. These seed-42 measurements come from the separate archived projected-Adam bundles described in Appendices A and K, rather than the historical runs in Table 2.

Table 3: Reconstruction fidelity in archived seed-42 FLACE bundles. All values are percentages. Original and reconstructed predictions share the frozen affine head.

| Dataset | Original accuracy | Reconstructed accuracy | Prediction agreement |
|---|---|---|---|
| CIFAR-100 | 81.31 | 80.98 | 97.71 |
| CUB | 81.01 | 81.01 | 98.65 |
| ImageNet-100 | 88.82 | 88.78 | 99.50 |

Across these bundles, reconstructed accuracy is within 0.33 percentage points of the original path, with prediction agreement ranging from 97.71% to 99.50%. Thus the reconstruction preserves most top-1 decisions, while the agreement metric also exposes prediction changes that similar aggregate accuracies can conceal.

### 5.7 Ablation Study

We study the sensitivity to the number of NMF components $k$ and the number of NMF iterations $T$. On ImageNet-100, increasing $k$ generally reduces concept-basis correlation and improves insertion-oriented usability, while deletion remains low but can exhibit small high-$k$ trade-offs; performance therefore saturates around a practical plateau rather than a universal optimum. Increasing $T$ yields rapid gains in the early stage and plateaus when $T \gtrsim 2000$. Full ablation curves and additional results on CIFAR-100 and CUB are reported in Appendix F, with additional iteration sweeps for CIFAR-100 and CUB in Appendix F.2.

**Qualitative Analysis.** The activation grids in Appendix E and coordinate-deletion examples in Appendix H illustrate how selected directions participate in reconstructed predictions. The signed-pair example reports the effect of zeroing an inferred coordinate while holding the others fixed.

**Why logit ranking and confidence can disagree.** The frozen head makes coordinate contributions exact, but the reported curves evaluate probabilities. Writing $p_c(u) = \mathrm{softmax}(f(u))_c$ gives

$$\frac{\partial p_c}{\partial u_r} = p_c(u) \left( \theta_{rc} - \sum_j p_j(u) \theta_{rj} \right).$$

Thus even $q_{rc} > 0$ need not make coordinate $r$ increase class-$c$ confidence: competing logits may rise faster. A coordinate-dependent shift common to all logits leaves probabilities unchanged but changes individual logit contributions. The original fixed-head ranking is consequently an explicit evaluation choice. Replacing it by a class-margin or probability ranking would define a different protocol and cannot be presented as the source of the retained AUCs.

## 6 Discussion and Limitations

FLACE provides numerical coordinates for inspecting and editing reconstructed predictions. The exact contribution identity applies to any fixed affine decoder and head. Establishing an advantage over other decompositions therefore requires matched reconstruction and intervention comparisons; the present native-path results do not isolate that advantage. In particular, the signed-basis examples do not establish a unique benefit of signed entries without strict-NMF or sign-clipped controls.

Reconstruction fidelity and semantic meaning are separate evaluation questions. A precise explanation of an inaccurate reconstruction need not describe the original prediction, so inspection should retain both the coordinates and the reconstruction residual. Optional coordinate names support visual browsing; stable attribute meanings and human correction would require attribute-alignment or human-agreement studies. Feature-space edits alone do not establish causal effects in the input world. The experiments use EfficientNet-B3 on three datasets. Broader architectural claims require validating the feature/head split, while hundreds of coordinates and thousands of optimization steps call for matched runtime and memory measurements.

## 7 Conclusion

FLACE exposes reconstructed predictions of a frozen classifier through nonnegative coordinates on signed feature directions. Reusing the original affine head connects each coordinate to an exact logit contribution and makes coordinate deletion directly interpretable as a numerical change in the reconstructed prediction. The three-dataset evaluation characterizes insertion, deletion and basis association under the reported native paths. This interface supports inspecting and editing model features without a predefined text vocabulary, with optional visual descriptions added after learning.

## AI USE STATEMENT

Generative AI tools assisted with manuscript refinement, language clarity and readability, and LaTeX formatting. The authors are responsible for the final manuscript.

## REPRODUCIBILITY STATEMENT

The reconstruction interface and attribution identities are specified in Sec. 4.1 and Appendix J. Appendix A documents the available numerical implementation; Appendices I and K distinguish result sources and identify missing protocol records that limit reproduction of the main comparison. The supplementary code notes its missing dependencies and execution requirements. This revision introduces no new experimental measurements.

## REFERENCES

Usha Bhalla, Alex Oesterling, Suraj Srinivas, Flavio P. Calmon, and Himabindu Lakkaraju. Interpreting CLIP with sparse linear concept embeddings (SpLiCE). *arXiv preprint arXiv:2402.10376*, 2024. URL https://arxiv.org/abs/2402.10376.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. IEEE, 2009.

Chris HQ Ding, Tao Li, and Michael I Jordan. Convex and semi-nonnegative matrix factorizations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(1):45–55, 2010. doi: 10.1109/TPAMI.2008.277.

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.

Thomas Fel, Agustin Picard, Louis Béthune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. CRAFT: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2711–2721, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Fel_CRAFT_Concept_Recursive_Activation_FacTorization_for_Explainability_CVPR_2023_paper.html.

Leo Gao, Tom Dupre la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=tcsZt9ZNKD.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. *Advances in Neural Information Processing Systems*, 32, 2019.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International conference on machine learning*, 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 *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

George A Miller. WordNet: A lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

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

Konstantinos P Panousis, Dino Ienco, and Diego Marcos. Coarse-to-fine concept bottleneck models. *Advances in Neural Information Processing Systems*, 37:105171–105199, 2024.

Jayneel Parekh, Pegah Khayatan, Mustafa Shukor, Alasdair Newson, and Matthieu Cord. A concept-based explainability framework for large multimodal models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 135783–135818, 2024. URL https://proceedings.nips.cc/paper_files/paper/2024/hash/f4fba41b554f9aaa013c4062a1c40518-Abstract-Conference.html.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference*, 2018. URL https://bmvc2018.org/contents/papers/1064.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Chenming Shang, Shiji Zhou, Hengyuan Zhang, Xinzhe Ni, Yujiu Yang, and Yuwang Wang. Incremental residual concept bottleneck models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11030–11040, 2024.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pp. 3319–3328. PMLR, 2017.

Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pp. 6105–6114. PMLR, 2019.

Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. Technical report, California Institute of Technology, 2011.

Yue Yang, Artemis Panagopoulou, Shenghao Zhou, Daniel Jin, Chris Callison-Burch, and Mark Yatskar. Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 19187–19197, 2023.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=nA5AZ8CEyow.

# A FLACE Interface and Numerical Implementation

The objective in Eq. (1) is classical semi-NMF Ding et al. (2010). The available July 25 implementation uses projected Adam as specified below. It is distinct from the alternating least-squares and multiplicative updates described in an earlier manuscript, and its linkage to the main-table runs has not been established. We therefore use it to document the separate fidelity results in Sec. 5.6, without assigning its optimizer to the main table. Features are extracted from torchvision EfficientNet-B3 with PyTorch 2.6.0+cu124 by hooking the dropout output after pooling and before the affine head; dropout is inactive. The software version of the original main-table runs is unrecorded.

**Documented FLACE implementation (seed 42; float32 CUDA)**

1. Extract frozen pooled EfficientNet-B3 features in evaluation mode. Initialize coefficients $U$ and basis $W$ independently with uniform entries in $[0, 1/\sqrt{k})$.
2. For 3000 steps, jointly update $U, W$ with Adam on mean squared reconstruction error, learning rate $0.01$ and weight decay $10^{-6}$. After each step clamp $U$ to at least $10^{-7}$; leave $W$ signed. No class-modulation branch or explicit sparsity penalty is enabled. The stored $\alpha = 0.99$ is unused in this code path.
3. For the entire test feature matrix, fix the raw learned $W$. Initialize test coefficients randomly on CPU in $[0, 1/\sqrt{k})$, clamp to $10^{-7}$ and move to CUDA. Run 300 full-matrix Adam steps on mean squared reconstruction error, learning rate $0.01$, no weight decay; clamp coefficients to at least $10^{-7}$ after each step. Adam uses PyTorch defaults $(\beta_1, \beta_2) = (0.9, 0.999)$ and optimizer epsilon $10^{-8}$; there is no early stopping or certified NNLS tolerance.
4. Set $\gamma_r = \max(\max_i U_{ir}, 10^{-7})$ from training coefficients. Divide training and inferred test coordinates by $\gamma_r$ and multiply basis row $r$ by $\gamma_r$. Reconstruct and reuse the original affine head.

Figure 3: Finite projected-Adam protocol. Unlike the ideal normalized-basis NNLS interface, this implementation infers coefficients against the raw basis and then rescales them. Both parameterizations have the same exact feasible reconstruction cone, but their finite Adam trajectories need not coincide. The positive floor also differs from exact $u \geq 0$.

The archive uses $\max(m_r, \varepsilon)$ rather than $m_r + \varepsilon$ in Eq. (2); both preserve products by matched inverse rescaling, but are different numerical conventions. Likewise, Adam weight decay means this executed fit is not an exact minimizer certificate for the unregularized objective. Theoretical identities in Sec. 4.1 apply to any fixed inferred coordinates and affine head, independently of these optimization differences.

# B Baseline Concept Counts and Native Paths

Table 4 records the reported concept counts, sources and prediction paths. Label-use descriptions refer to method families: Res-CBM uses a supervised residual head; Post-hoc CBM and CF-CBM permit task-label use; Label-free-CBM and FLACE do not use task labels to construct their bases. These descriptions do not establish the label settings of individual runs. Shared controls are specified in Sec. 5.2; missing intervention metadata are collected in Appendix I.

Table 4: Reported coordinate counts and native prediction paths. Counts follow CIFAR-100 / CUB / ImageNet-100 order.

| Method | Coordinate count $k$ | Coordinate source | Native prediction path |
|---|---|---|---|
| Res-CBM | 112 / 312 / 170 | Public supervised concept set | Concept vectors and residual head |
| Post-hoc CBM | 112 / 312 / 170 | Same concept set; frozen backbone | Post-hoc projected linear head |
| Label-free-CBM | 4096 / 4807 / 4523 | GPT-generated text pool and CLIP | Text-pool similarity and learned linear head |
| CF-CBM | 2400 / 4710 / 4500 | Coarse-to-fine concept hierarchy | Coarse-to-fine concept-aggregation head |
| FLACE | 600 / 700 / 900 | Semi-NMF dictionary from target features | Nonnegative reconstruction and original frozen head |

## C Concept-Basis Correlation Formulas

For completeness, we record the three pairwise correlation statistics summarized by Eq. (13) in the main-text Evaluation Metrics subsection. Let $\mathbf{w}_r'^{\top}$ be the $r$-th row of $W'$ and $\bar{w}_r' = \frac{1}{d}\sum_{t=1}^{d} w_{rt}'$ denote the scalar arithmetic mean of that row; $\mathbf{1} \in \mathbb{R}^d$ is the all-ones vector. The Pearson correlation coefficient (with mean-centering) is

$$r^{\mathrm{pear}}(r,s) = \frac{\langle \mathbf{w}_r' - \bar{w}_r'\mathbf{1},\, \mathbf{w}_s' - \bar{w}_s'\mathbf{1}\rangle}{\|\mathbf{w}_r' - \bar{w}_r'\mathbf{1}\|_2 \, \|\mathbf{w}_s' - \bar{w}_s'\mathbf{1}\|_2},$$

the cosine similarity (without mean-centering) is

$$r^{\cos}(r,s) = \frac{\langle \mathbf{w}_r', \mathbf{w}_s'\rangle}{\|\mathbf{w}_r'\|_2 \, \|\mathbf{w}_s'\|_2},$$

and the Spearman coefficient applies a rank transformation followed by Pearson, $r^{\mathrm{spear}}(r,s) = \mathrm{Pearson}(\mathrm{rank}(\mathbf{w}_r'), \mathrm{rank}(\mathbf{w}_s'))$.

## D Representational Notes for Semi-NMF

**Remark 1: mixed-sign bases and contrastive feature evidence.** The following comparison provides intuition for why an unconstrained basis is useful for mixed-sign neural features. Consider the standard NMF reconstruction and the semi-NMF reconstruction that allows negative entries in $W$:

$$\|Z - UW_+\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{d}\left(z_{ij} - \sum_{r=1}^{k} u_{ir} w_{rj}^{+}\right)^2,$$

$$\|Z - U(W_+ + W_-)\|_F^2 = \sum_{i=1}^{n}\sum_{j=1}^{d}\left(z_{ij} - \sum_{r=1}^{k} u_{ir}(w_{rj}^{+} + w_{rj}^{-})\right)^2.$$

By introducing $W_-$, the semi-NMF basis can represent contrastive feature components that are unavailable to a strictly nonnegative basis. This argument is a representational motivation rather than a claim that every learned negative entry has a directly human-semantic interpretation.

**Remark 2: feasible-set relation.** For the same rank $k$ and the same reconstruction objective, semi-NMF with unconstrained $W$ has a feasible set that contains the standard NMF feasible set.

Let

$$\mathcal{F}_{\mathrm{NMF}} = \left\{(U,W) \,\middle|\, U \geq 0,\ W \geq 0\right\}$$

denote the feasible set of standard NMF, and

$$\mathcal{F}_{\text{semi-NMF}} = \left\{(U, W) \,\middle|\, U \geq 0,\ W \in \mathbb{R}^{k \times d}\right\}$$

denote the feasible set of semi-NMF with unconstrained bases. Clearly,

$$\mathcal{F}_{\text{NMF}} \subseteq \mathcal{F}_{\text{semi-NMF}},$$

since any pair $(U, W)$ with $U \geq 0$ and $W \geq 0$ is also feasible for semi-NMF.

Consider the reconstruction objective

$$\mathcal{L}(U, W) = \|Z - UW\|_F^2.$$

Because $\mathcal{F}_{\text{NMF}}$ is a subset of $\mathcal{F}_{\text{semi-NMF}}$, we have

$$\inf_{(U,W) \in \mathcal{F}_{\text{semi-NMF}}} \mathcal{L}(U, W) \quad \leq \quad \inf_{(U,W) \in \mathcal{F}_{\text{NMF}}} \mathcal{L}(U, W).$$

In other words, allowing $W$ to take negative values enlarges the feasible set and therefore cannot increase the minimal reconstruction error under this objective. The inequality can be strict; an infimum need not be attained by finite semi-NMF factors. This is purely a statement about *optimal values* over the two feasible sets; it does not assert that any particular alternating-update or descent algorithm reaches the global optimum, nor does it provide a convergence guarantee for the iterative procedure used in our experiments. The corresponding empirical metric behavior is reported as a sensitivity pattern in Appendix F.2.

### D.1 Toy Example for Per-Concept Normalization

This arithmetic example uses $\varepsilon = 0$ and nonzero column maxima for clarity; it does not illustrate the positive-epsilon implementation convention.

**Given a concept coefficient matrix $U$:**

$$U = \begin{bmatrix} 10 & 1 \\ 20 & 0 \end{bmatrix}$$

**Concept basis matrix $W$:**

$$W = \begin{bmatrix} 0.5 & 2 \\ 1 & 3 \end{bmatrix}$$

If we directly compute $Z = UW$, we obtain:

$$Z = \begin{bmatrix} 10 \times 0.5 + 1 \times 1 & 10 \times 2 + 1 \times 3 \\ 20 \times 0.5 + 0 \times 1 & 20 \times 2 + 0 \times 3 \end{bmatrix} = \begin{bmatrix} 5 + 1 & 20 + 3 \\ 10 + 0 & 40 + 0 \end{bmatrix} = \begin{bmatrix} 6 & 23 \\ 10 & 40 \end{bmatrix}$$

Here, the first concept has larger activation magnitudes than the second concept. FLACE therefore rescales each concept column and the corresponding basis row inversely, so that the reconstructed product is unchanged while activation magnitudes become easier to compare within each concept dimension.

If we perform normalization:

**Per-concept scaling factors:**

$$\gamma_1 = \max(10, 20) = 20, \qquad \gamma_2 = \max(1, 0) = 1.$$

**Normalized $U'$:**

$$U' = \begin{bmatrix} 10/20 & 1/1 \\ 20/20 & 0/1 \end{bmatrix} = \begin{bmatrix} 0.5 & 1 \\ 1 & 0 \end{bmatrix}$$

**Normalized $W'$:**

$$W' = \begin{bmatrix} 20 \cdot 0.5 & 20 \cdot 2 \\ 1 \cdot 1 & 1 \cdot 3 \end{bmatrix} = \begin{bmatrix} 10 & 40 \\ 1 & 3 \end{bmatrix}$$

After normalization, we recompute

$$\begin{aligned} Z' &= U'W' \\ &= \begin{bmatrix} 0.5 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 10 & 40 \\ 1 & 3 \end{bmatrix} \\ &= \begin{bmatrix} 0.5 \cdot 10 + 1 \cdot 1 & 0.5 \cdot 40 + 1 \cdot 3 \\ 1 \cdot 10 + 0 \cdot 1 & 1 \cdot 40 + 0 \cdot 3 \end{bmatrix} \\ &= \begin{bmatrix} 5+1 & 20+3 \\ 10+0 & 40+0 \end{bmatrix} = \begin{bmatrix} 6 & 23 \\ 10 & 40 \end{bmatrix}. \end{aligned}$$

The product is unchanged before and after normalization: $Z' = Z$. The operation changes the scale assigned to each column of $U$ and the corresponding row of $W$, but it preserves reconstruction while making coefficient magnitudes easier to compare within each concept dimension. This sets a coefficient display convention; corresponding IG scores and coordinate deletions are invariant even before choosing that convention.

### D.2 Additional Algebraic Observations for Flexible Semi-NMF

The notes above explain why signed bases are useful for mixed-sign features. This section records a few algebraic observations about feasibility and reconstruction error; these statements are intended to clarify the parameterization, not to introduce a separate theoretical contribution.

**1) Existence and Upper-Bound Constructions**

**Observation 1 (existence via explicit construction).** For any $Z \in \mathbb{R}^{n \times d}$, there always exists an exact factorization $Z = UW$ such that $U \geq 0$. The following constructions show feasibility when the rank is allowed to be sufficiently large.

**Construction A**: Let $k = n$, $U = I_n \geq 0$, and $W = Z$. Then $U \in \mathbb{R}^{n \times n}$, $W \in \mathbb{R}^{n \times d}$, and $\|Z - UW\|_F^2 = 0$.

**Construction B**: Write $Z = Z^+ - Z^-$, where $Z^+ = \max(Z, 0)$ and $Z^- = \max(-Z, 0)$, and take

$$U = [I_n \quad I_n] \in \mathbb{R}_+^{n \times 2n}, \quad W = \begin{bmatrix} Z^+ \\ -Z^- \end{bmatrix} \in \mathbb{R}^{2n \times d},$$

then we still have $Z = UW$.

Therefore, an exact flexible semi-NMF representation always exists when the factorization rank is allowed to be sufficiently large. In practice, FLACE uses smaller ranks and optimizes the reconstruction–interpretability trade-off.

**2) Optimal Subproblem and Error Structure**

**Observation 2 (least-squares $W$ given fixed $U$).** Given a fixed $U \geq 0$, one optimal basis matrix and the corresponding reconstruction error are

$$W^* = U^\dagger Z, \quad \min_W \|Z - UW\|_F^2 = \|(I - P_U)Z\|_F^2,$$

where $U^\dagger$ is the Moore–Penrose pseudoinverse and $P_U = UU^\dagger$ is the projection onto the column space of $U$. Since the error of flexible semi-NMF is the residual of projecting $Z$ outside $\mathrm{span}(U)$, the essence of learning $U$ is to learn a linear subspace with a nonnegative basis.

**3) Relationship with PCA**

**Observation 3 (PCA lower-bound perspective).** Let:

$$\mathcal{E}_k^{\text{PCA}} = \min_{\text{rank}(X) \leq k} \|Z - X\|_F^2$$

be the minimum uncentered rank-$k$ approximation error (truncated SVD; centered PCA only if $Z$ is first centered). Then the optimal value of flexible semi-NMF satisfies

$$\inf_{U \geq 0, W} \|Z - UW\|_F^2 \geq \mathcal{E}_k^{\text{PCA}}.$$

If the semi-NMF infimum is attained, equality requires its product to be an optimal rank-$k$ approximation. Without attainment this characterization fails. For example, let $k = 2$ and

$$Z = \begin{bmatrix} 1 & 0 \\ -1 & 0 \\ 0 & 1 \end{bmatrix}, \quad W_t = \begin{bmatrix} 1 & t \\ -1 & t \end{bmatrix}, \quad U_t = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ (2t)^{-1} & (2t)^{-1} \end{bmatrix}, \quad t > 0.$$

Then $U_t W_t$ has rows $(1, t), (-1, t), (0, 1)$, so its squared error is $2t^2 \to 0$. Yet two nonnegative generating rays cannot contain both $(1, 0)$ and $(-1, 0)$ and also $(0, 1)$: if their cone contains a line, the two rays must be opposite and generate only that line. Thus no exact rank-2 semi-NMF factorization exists, while the unconstrained rank-2 error and semi-NMF infimum both equal zero. This nonattainment does not affect fixed-basis NNLS or fixed-$U$ least squares, whose minima exist.

## E Activation Grids for Selected Coordinates

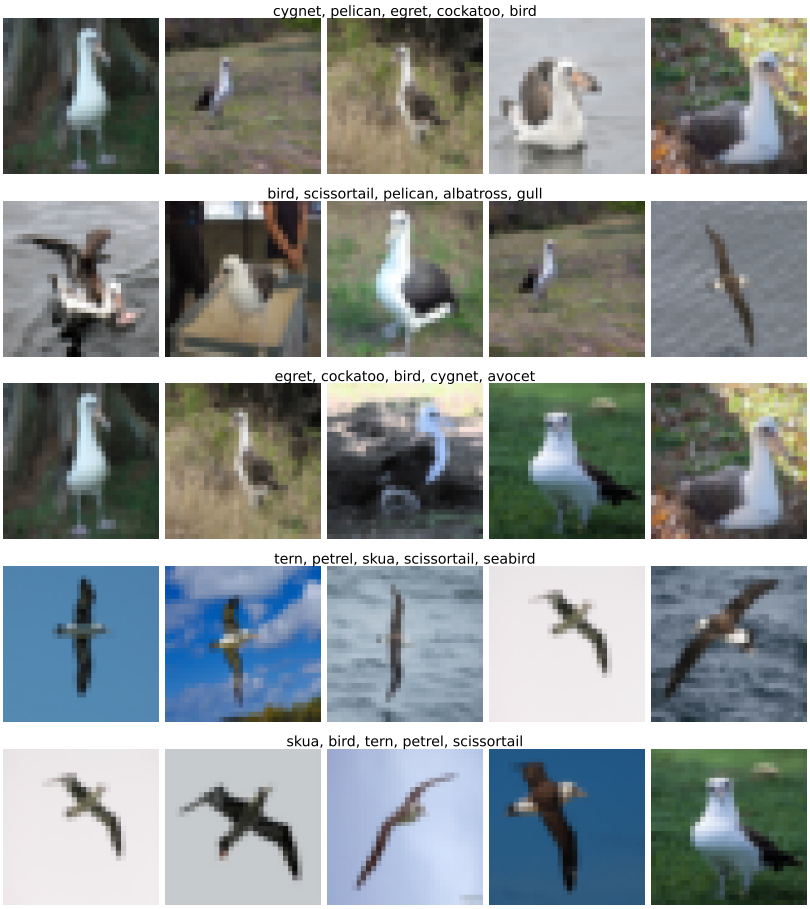



Figure 4: **Selected coordinate activation grids.** Top-activated samples for selected FLACE coordinates. Text labels are post-hoc descriptors used only for visualization; each grid should be read as a qualitative visual cluster, not as a fixed concept vocabulary produced by the model.

Figure 4 shows additional concept visualization examples. For visualization only, we attach post-hoc textual descriptors to selected learned concepts after the FLACE basis has already been learned. Any auxiliary naming resource used in this step, such as WordNet Miller (1995), is applied only after concept discovery and is not used during predictor construction or test-time concept inference. The resulting names should therefore be read as descriptive labels for visual inspection, not as concept vocabulary produced by the model itself.

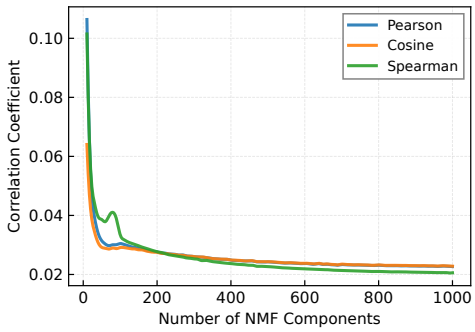



Figure 5: Effect of the number of NMF components $k$ on concept-basis correlation (ImageNet-100). As $k$ increases, the upper-triangular mean absolute Pearson/Cosine/Spearman correlation over the rows of $W'$ decreases rapidly and stabilizes around $k \gtrsim 800$ (lower values indicate reduced redundancy among concept directions).

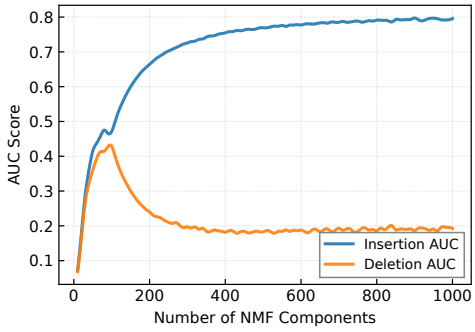



Figure 6: Effect of the number of NMF components $k$ on Insertion/Deletion usability metrics (ImageNet-100). $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$ increases with $k$ and saturates in the high-$k$ regime, indicating higher average confidence along the insertion path. $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$ remains low in the high-$k$ regime but can show small nonmonotonic fluctuations (lower is better).

## F Additional Ablation

**Effect of the Number of NMF Components** The component sweeps below are sensitivity analyses. Plateau ranges indicate where the curves become less sensitive to $k$; they should not be interpreted as new default selections unless explicitly stated in the main experimental setup. Using ImageNet-100 as an example (Fig. 5 and Fig. 6), we fix the number of NMF iterations to $T = 3000$. As the number of concepts $k$ increases from a small value, the correlation of the normalized concept basis $W'$ (measured as the upper-triangular mean absolute Pearson/Cosine/Spearman correlation over its row vectors) drops rapidly, enters a slow-changing region around $k \approx 150$–$200$, and stabilizes at $\sim 2 \times 10^{-2}$ when $k \gtrsim 800$. This trend is compatible with the representational intuition that a larger semi-NMF basis gives the factorization more degrees of freedom to use low-correlated feature directions; it should not be read as isolating a single mechanism for suppressive relationships.

In parallel, the explainability metrics exhibit a "rapid rise then plateau" pattern. The Insertion AUC, $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$, increases with $k$ and saturates around $k \approx 700$–$1000$ ($\sim 0.78$–$0.80$), indicating higher average target confidence along the specified insertion path. The Deletion AUC, $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$, shows a brief increase for very small $k$, then enters a low high-$k$ regime around $\sim 0.18$–$0.20$ with small nonmonotonic fluctuations.

The separate insertion and deletion curves describe this dictionary-width sweep; no combined score or optimal knee point is defined here.

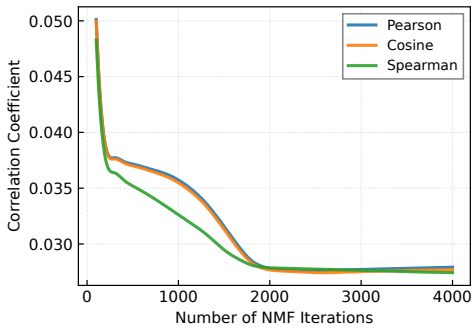



Figure 7: Effect of NMF iterations $T$ on concept-basis correlation on ImageNet-100 with fixed $k = 200$. As $T$ increases, the upper-triangular mean absolute Pearson/Cosine/Spearman correlation over the rows of $W'$ decreases rapidly and plateaus when $T \gtrsim 2000$, indicating lower redundancy among concept directions.

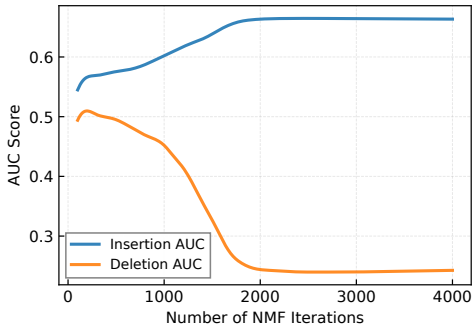



Figure 8: Insertion/Deletion results on ImageNet-100 with fixed $k = 200$. $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$ increases with the number of iterations and plateaus around $T \approx 2000$–$3000$, while $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$ decreases and stabilizes at a low level. Together, the two curves indicate an improved Insertion/Deletion trade-off as the reported curves approach a plateau.

**Effect of NMF Iterations on Performance** On ImageNet-100, we fix the number of concepts to $k = 200$ and vary only the number of NMF iterations $T$. As shown in Fig. 7, the row-wise correlation of the normalized concept basis $W'$ (upper-triangular mean absolute Pearson/Cosine/Spearman correlation) decreases rapidly within the first few hundred steps, enters a slow-decay phase around $T \approx 600$–$800$, and plateaus when $T \gtrsim 2000$. The plateau describes the reported metric curve; it does not certify optimizer convergence or near-orthogonality. The optimizer used for these historical figures remains unlinked to an exact run record.

Regarding explainability (Fig. 8), $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$ increases with $T$ and plateaus around $T \approx 2000$–$3000$, indicating higher average target confidence along the insertion path; neither feature reconstruction nor a small sufficient set follows from this AUC. Meanwhile, $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$ decreases from an initially higher value to around $\sim 0.25$ and then stabilizes, recording lower average target confidence along the deletion path. Overall, the marginal gains from increasing $T$ diminish when $T \gtrsim 2000$, and the computation–performance trade-off typically lies within $T \in [1500, 3000]$. The choice of $T = 3000$ in our main experiments falls in this plateau region and lies within the displayed insertion/deletion plateau. These are empirical sensitivity patterns; they do not identify the mechanism responsible for the changes.

### F.1 Effect of the Number of NMF Components on CIFAR-100 and CUB

**CIFAR-100.** On CIFAR-100 (Fig. 9 and Fig. 10), we again fix the number of NMF iterations to $T = 3000$. As $k$ increases from very small values, the mean absolute Pearson/Cosine/Spearman correlation among the rows of $W'$ exhibits a sharp drop from around $0.10$ to below $0.04$ (for

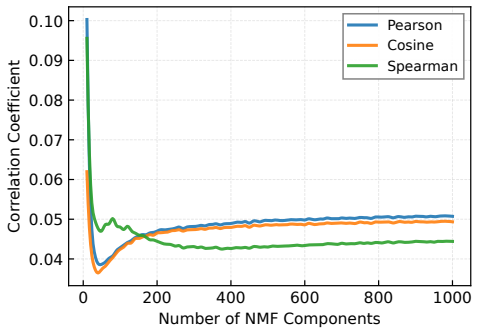



Figure 9: Effect of the number of NMF components $k$ on concept-basis correlation (CIFAR-100). The upper-triangular mean absolute Pearson/Cosine/Spearman correlation drops at small $k$, rebounds, and then approaches a plateau, stabilizing around $k \gtrsim 500$, indicating reduced redundancy among concept directions.

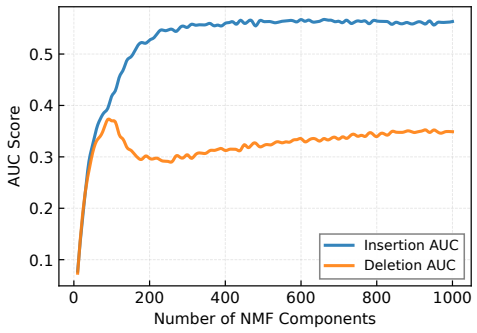



Figure 10: Effect of the number of NMF components $k$ on Insertion/Deletion usability metrics (CIFAR-100). $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$ rises and saturates around $k \approx 600$–$800$, while $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$ has a small high-$k$ increase after its initial decline.

$k \approx 40$–$60$), followed by a mild rebound and then a stable plateau around $0.05$ once $k \gtrsim 500$. Overall, the correlation is substantially reduced compared to the small-$k$ regime, indicating lower collinearity among learned concept directions.

The Insertion/Deletion usability metrics show the familiar "rapid rise then plateau" behaviour. The Insertion AUC, $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$, increases quickly with $k$, reaching $\sim 0.35$ by $k \approx 50$ and saturating around $0.53$–$0.55$ for $k \approx 600$–$800$, which records higher average target confidence along the insertion path. The Deletion AUC, $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$, shows an initial peak at very small $k$, then decreases and stabilizes around $0.34$–$0.36$ once $k \gtrsim 300$. Read jointly, the insertion/deletion curves enter a stable plateau near this range. This sweep should be read as a sensitivity analysis; the default value in the main experiments follows the fixed experimental setup rather than a claim of a universal optimum.

**CUB.** On the fine-grained CUB dataset (Fig. 11 and Fig. 12), the correlation curves exhibit a sharp initial drop as $k$ increases from very small values. Pearson, Cosine, and Spearman correlations decrease in the small-$k$ regime to below $0.05$ by $k \approx 40$. After this sharp decay, the correlations continue to decline more gradually and reach a stable regime around $0.030$–$0.040$ for $k \gtrsim 300$. This indicates that the semi-NMF factorization identifies low-collinearity concept directions in this fine-grained setting.

The Insertion/Deletion metrics follow the same qualitative pattern seen on CIFAR-100 and ImageNet-100. The Insertion AUC rises steeply for small $k$, exceeding $0.40$ by $k \approx 80$ and continuing to improve until it plateaus around $0.50$–$0.52$ in the region $k \approx 300$–$400$. This describes average confidence along the insertion path as dictionary width varies, not a sufficient active-coordinate count. The Deletion AUC also shows a clear early peak at very small $k$, after which it decreases and stabilizes

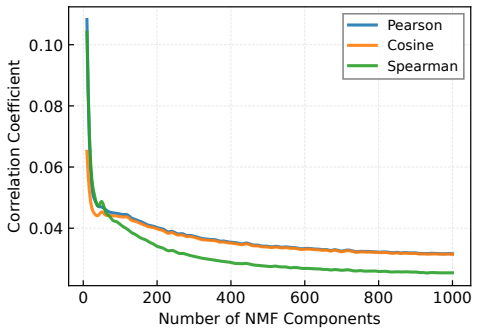



Figure 11: Effect of the number of NMF components $k$ on concept-basis correlation (CUB). Correlation decreases rapidly and stabilizes around $k \gtrsim 300$, indicating lower redundancy in this fine-grained setting.

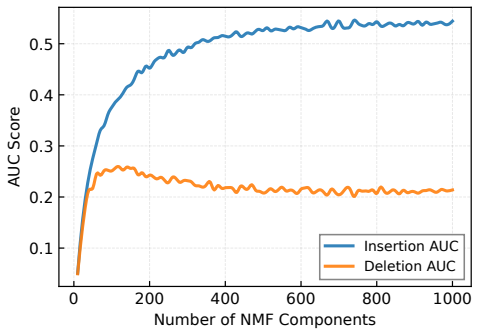



Figure 12: Effect of the number of NMF components $k$ on Insertion/Deletion usability metrics (CUB). The Insertion AUC rises and plateaus near $k \approx 300$–$400$, while the Deletion AUC decreases and stabilizes around $k \gtrsim 250$.

around $0.23$–$0.25$ for $k \gtrsim 250$. These curves describe the reported confidence paths; they do not identify why the metrics change.

The separate insertion and deletion curves describe this dictionary-width sweep; no combined score or optimal knee point is defined here.

### F.2 Effect of the Number of NMF Iterations on CIFAR-100 and CUB

**CIFAR-100.** On CIFAR-100, we fix the number of NMF components to the default used in the main experiments (see Appendix F.1) and vary only the number of iterations $T$. As shown in Fig. 13, the row-wise correlation of the normalized concept basis $W'$ (upper-triangular mean absolute Pearson/Cosine/Spearman correlation) decreases sharply within the first few hundred iterations: starting around $0.08$ at very small $T$, it drops to below $0.05$ by $T \approx 800$–$1000$. After this rapid decay, the curves enter a near-plateau regime with only mild drift, stabilizing around $0.045$–$0.047$ for $T \gtrsim 1500$. This pattern mirrors the ImageNet-100 behavior and indicates that most of the observed redundancy reduction in the concept basis is achieved early in the optimization.

The Insertion/Deletion usability metrics in Fig. 14 follow a complementary trend. The Insertion AUC, $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$, starts around $0.46$–$0.47$ at small $T$ and increases steadily as the factorization refines, reaching about $0.51$ by $T \approx 1500$ and plateauing around $0.52$–$0.53$ for $T \approx 2000$–$3000$. In contrast, the Deletion AUC, $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$, begins near $0.40$, exhibits a slight early peak, and then declines overall to around $0.30$ as $T$ approaches $1500$, remaining essentially flat thereafter. Consequently, the paired insertion/deletion diagnostics improve rapidly in the first $1000$–$1500$ iterations and show diminishing returns once $T \gtrsim 2000$. Choosing $T = 3000$ for CIFAR-100 therefore lies within this plateau region, maintaining a computation–diagnostic trade-off consistent with the conclusions drawn from ImageNet-100.

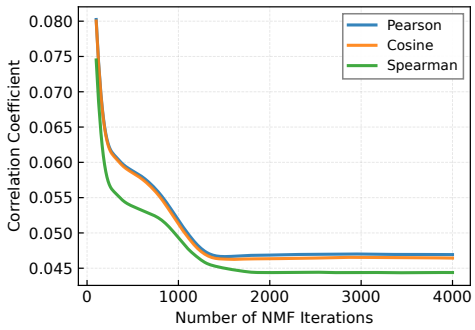



Figure 13: Effect of the number of NMF iterations $T$ on concept-basis correlation (CIFAR-100). The upper-triangular mean absolute Pearson/Cosine/Spearman correlation drops rapidly within the first few hundred iterations and stabilizes around $T \gtrsim 1500$, indicating a plateau in the correlation statistic.

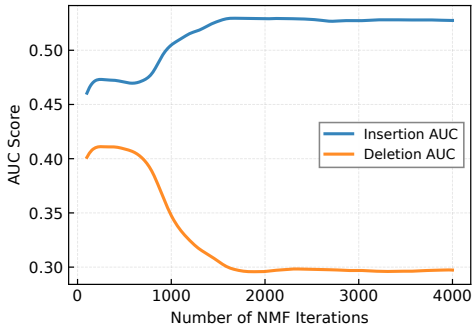



Figure 14: Effect of the number of NMF iterations $T$ on Insertion/Deletion usability metrics (CIFAR-100). $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$ increases with $T$ and plateaus around $T \approx 2000$–$3000$, while $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$ decreases from $\sim 0.40$ to $\sim 0.30$ and then stabilizes.

**CUB.** For the fine-grained CUB dataset, we again fix the number of NMF components and vary only the number of iterations $T$. As shown in Fig. 15, the row-wise correlation of the normalized concept basis $W'$ (upper-triangular mean absolute Pearson/Cosine/Spearman correlation) drops quickly in the early stage of optimization: starting from about $0.055$–$0.058$ at very small $T$, all three curves fall below $0.048$ by $T \approx 300$. After this rapid decay, the correlations continue to decrease more gradually and approach a plateau around $0.040$ (Pearson/Cosine) and $0.033$ (Spearman) when $T \gtrsim 3000$. The reported metric curves approach a lower-correlation plateau, with most of the observed redundancy reduction achieved in the first $\sim 1000$ iterations.

The Insertion/Deletion metrics in Fig. 16 exhibit a complementary pattern. The Insertion AUC, $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$, increases steadily with $T$: it starts around $0.34$–$0.36$ at small $T$, rises to roughly $0.40$ by $T \approx 500$, and further improves to about $0.44$–$0.45$ in the range $T \approx 2500$–$4000$, where it effectively saturates. In contrast, the Deletion AUC, $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$, shows a shallow peak near $0.32$ at $T \approx 200$–$300$, and then declines overall to around $0.24$–$0.25$ as $T$ approaches $3000$–$4000$. The reported insertion AUC increases and deletion AUC decreases overall before reaching their respective plateaus.

## G Additional Robustness Diagnostics

**Independent repeat protocol.** To check that the main conclusions do not depend on a single factorization run, we conducted an additional repeat sweep on a separate GPU machine using the same datasets, EfficientNet-B3 backbone, default component counts, $T = 3000$ NMF iterations, 20 IG steps, and a zero IG baseline. These runs are reported as robustness diagnostics only; they are not used to replace the single-run main comparison in Table 2.

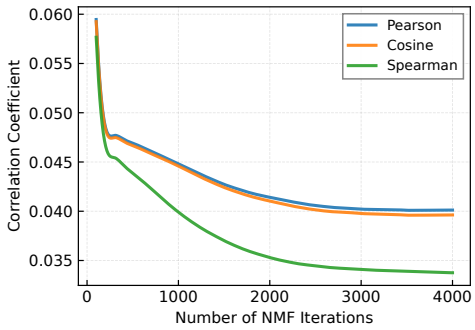



Figure 15: Effect of the number of NMF iterations $T$ on concept-basis correlation (CUB). The upper-triangular mean absolute Pearson/Cosine/Spearman correlation drops rapidly within the first few hundred iterations and then decays slowly, stabilizing around $T \gtrsim 3000$, which indicates a plateau in the correlation statistic.

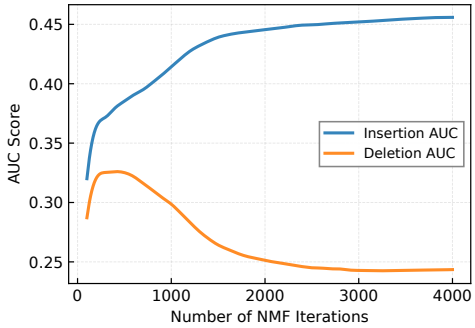



Figure 16: Effect of the number of NMF iterations $T$ on Insertion/Deletion usability metrics (CUB). $\mathrm{AUC}_{\mathrm{ins}}^{\mathrm{interp}}$ increases with $T$ and plateaus around $T \approx 2500$–$4000$, while $\mathrm{AUC}_{\mathrm{del}}^{\mathrm{interp}}$ exhibits an early peak near $T \approx 200$–$300$ and then decreases towards $\sim 0.24$–$0.25$.

**Discussion.** Table 5 describes variation in aggregate historical diagnostics. It does not test coordinate identity or per-sample explanation stability. An unchanged frozen-backbone reference is expected; the ImageNet-100 reference value differs from both the main table and archived audit and remains unreconciled. The reported CUB and ImageNet-100 AUC means are not substituted into the main table.

**Interpretation.** The additional $k$-sensitivity points in Table 6 are consistent with the plateau interpretation in Appendix F. Larger $k$ generally gives the factorization more degrees of freedom to reduce basis redundancy and increase insertion AUC. However, the ImageNet-100 deletion value at $k = 1200$ is slightly worse than at $k = 900$, so the high-$k$ region should be described as a favorable plateau/trade-off rather than as strict monotonic dominance.

## H Additional Qualitative Experimental Results

The following cases illustrate selected activation patterns and coordinate edits in the reconstructed feature space.

### H.1 Selected-Grid Provenance

The two previously displayed top-16 grids are omitted from empirical evidence because the available record does not link their images to dataset/split/sample identifiers and activation ordering. Their image assets are preserved in the source package; no dataset identity or naming validity is inferred from their appearance.

Table 5: Aggregate factorization repeats (mean ± standard deviation). Widths are 600/700/900; seeds are 42,43 for CIFAR-100 and 42,43,44 otherwise. Accuracy is the reported backbone reference, not reconstructed accuracy. Result-source qualifications are given in Appendix K.

| Dataset | Backbone reference | Insertion AUC | Deletion AUC |
|---|---|---|---|
| CIFAR-100 | 0.813100±0.000000 | 0.564424±0.000147 | 0.331951±0.000305 |
| CUB | 0.810148±0.000000 | 0.532683±0.002180 | 0.214241±0.005327 |
| ImageNet-100 | 0.887800±0.000000 | 0.792464±0.000974 | 0.190680±0.001086 |

| Dataset | Pearson | Cosine | Spearman |
|---|---|---|---|
| CIFAR-100 | 0.049950±0.000005 | 0.048743±0.000006 | 0.043471±0.000031 |
| CUB | 0.032861±0.000106 | 0.032670±0.000095 | 0.026399±0.000023 |
| ImageNet-100 | 0.022777±0.000041 | 0.022879±0.000041 | 0.020723±0.000045 |

Table 6: **Additional $k$-sensitivity points from the repeat sweep.** CUB shows a clean improvement from $k = 400$ to $k = 1000$. ImageNet-100 shows lower redundancy and stronger insertion at larger $k$, while deletion has a small high-$k$ trade-off at $k = 1200$.

| Dataset | $k$ | Pearson | Cosine | Spearman | Ins. AUC | Del. AUC |
|---|---|---|---|---|---|---|
| CUB | 400 | 0.035335 | 0.035031 | 0.028762 | 0.511660 | 0.221512 |
| CUB | 700 | 0.032959 | 0.032755 | 0.026383 | 0.530447 | 0.213282 |
| CUB | 1000 | 0.031482 | 0.031322 | 0.025277 | 0.543507 | 0.209643 |
| ImageNet-100 | 600 | 0.023606 | 0.023711 | 0.021865 | 0.778544 | 0.190363 |
| ImageNet-100 | 900 | 0.022809 | 0.022913 | 0.020758 | 0.793408 | 0.191836 |
| ImageNet-100 | 1200 | 0.022651 | 0.022768 | 0.020434 | 0.795894 | 0.195098 |

### H.2 Single-Sample Feature-Space Intervention

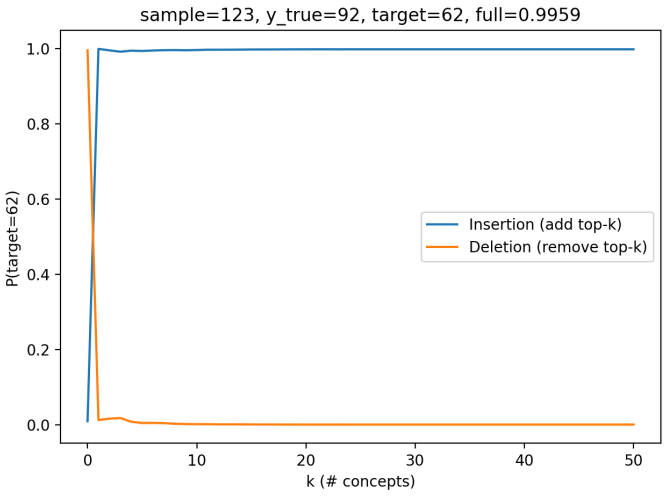



Figure 17: **Feature-Space Intervention.** Insertion and deletion for selected Sample 123: the plot metadata gives true class 92 and target class 62. This is a non-ground-truth-target diagnostic, distinct from the aggregate ground-truth-logit protocol. The steep slopes indicate that the top-ranked FLACE concept is highly influential for the reconstructed feature path of this sample.

To illustrate how discovered concepts affect the target classifier through reconstructed features, we performed single-sample feature-space intervention experiments on a selected sample (Sample

123, true class 92, target class 62 according to the plot metadata), manipulating the activation of its top-ranked concepts. As shown in Fig. 17, setting the top-$k$ concept activations to zero drops model confidence from $\sim 1.0$ to $\sim 0.0$ at $k = 1$ (deletion-oriented sensitivity), while constructing the representation using *only* the top-$k$ concepts recovers prediction confidence to $\sim 1.0$ at $k = 1$ (insertion-oriented sufficiency in this diagnostic example).

### H.3 Signed-Basis Pair Diagnostic

In a CUB diagnostic run, we selected coordinates $r = 20$ and $s = 59$ with negative basis cosine ($\cos(W'_{20}, W'_{59}) = -0.1707$) and weakly negative activation correlation (Pearson $= -0.0930$). Table 7 shows how zeroing each selected coordinate changes the target probability while the remaining coordinates are held fixed.

Table 7: Coordinate deletion in two selected CUB samples. The selected coefficient is set to zero and all others are held fixed. Values are target-class probabilities.

| Sample | Coordinate | Target class | Before deletion | After deletion |
|---|---|---|---|---|
| 3157 | $r = 20$ | 113 | 0.8479 | 0.0084 |
| 1822 | $s = 59$ | 44 | 0.7460 | 0.5581 |

Removing the selected coordinate lowers target confidence in both examples. The table does not establish a change in the predicted class. Negative basis cosine alone determines neither the sign nor the magnitude of this response, which also depends on the head and competing logits. An additional paired-coordinate increase is excluded because its edit magnitude and before/after coefficients were not recorded.

## I Interpretation and Evidence Boundaries

**Logit ranking versus probability.** For $p_c(u) = \mathrm{softmax}(f(u))_c$,

$$\frac{\partial p_c}{\partial u_r} = p_c \left( \theta_{rc} - \sum_j p_j \theta_{rj} \right).$$

Hence a positive target-logit contribution need not increase target probability. Adding the same coordinate-dependent shift to every logit preserves probabilities but can change the positive-logit ranking. The reported fixed-head logit ranking is retained; a margin/probability ranking would be a different, unevaluated protocol.

**Protocol records.** The historical baseline records do not specify whether Res-CBM residual outputs were fixed, zeroed or recomputed during edits, the exact CF-CBM aggregation edit, or all zero/bias endpoints. The labeled samples used to estimate the global ranking are not linked to a calibration/test split. Table 4 summarizes method-family label usage rather than an audited per-run configuration. These comparisons are descriptive native-system diagnostics. The ImageNet-100 class/subsample manifest and training/checkpoint identities are also missing; the stated 10k/5k subset is not uniquely determined by "official splits". Raw accuracy evaluates $\arg\max_c(za_c + b_c)$ against labels; reconstructed accuracy evaluates $\arg\max_c((uW')a_c + b_c)$ against labels; prediction agreement compares these two predictions regardless of correctness. They answer different questions. The main-table caption supplies raw FLACE reference accuracies separately from baseline native accuracies; these cannot quantify the accuracy cost of its bottleneck.

Dense dictionary rows neither prove nor disprove polysemanticity: their feature entries and the visual meanings of their activation sets are different objects. Low pairwise correlation is a geometric statistic; probe completeness tests task-information recovery under a specified probe; class purity tests label concentration among selected activating samples; human alignment tests agreement with interpretable attributes or judgments. None implies the others. Selected activation visualizations do not establish that every coordinate has a consistent meaning. A human correction interface would need stable names, a specified edit operation, and user validation beyond these grids.

The main comparison retains native heads and concept counts. SAE comparisons would additionally need both dictionary width and active TopK reported, alongside reconstruction error and downstream fidelity. A separate archived audit is summarized below; its descriptor inheritance is not human agreement and its matched-width SAE comparison is not sparsity matched. The algebra can be applied to other feature/head splits, but the experiments do not establish performance beyond the stated backbone and datasets.

## J Rescaling, Exact Inference, and Attribution

For $\Gamma = \mathrm{diag}(\gamma_r)$ with every $\gamma_r > 0$, the map $u \mapsto v = u\Gamma^{-1}$ is a bijection on the nonnegative orthant and $v(\Gamma W) = uW$. This proves Eq. (3), equality of optimal NNLS errors, and correspondence of minimizer sets. A finitely generated cone is closed and convex, so its Euclidean projection of $z$ is unique. Coefficients can nevertheless be nonunique when different nonnegative combinations reconstruct the same vector; full row rank of $W$ is sufficient for uniqueness. Statements about corresponding coordinates therefore require choosing corresponding minimizers, not arbitrary representatives.

For any differentiable head $F$, let $g(u) = F(uW)$ and $\widehat{g}(v) = F(v\Gamma W) = g(v\Gamma)$. With $v = u\Gamma^{-1}$ and a zero baseline,

$$\mathrm{IG}_r^{\widehat{g}}(v; 0) = \frac{u_r}{\gamma_r} \int_0^1 \gamma_r \, \partial_r g(tu) \, dt = \mathrm{IG}_r^{g}(u; 0).$$

The same identity holds under the usual almost-everywhere differentiability and path-integrability conditions. Deleting corresponding coordinate $r$ also preserves the reconstructed vector, since $v_r \gamma_r W_{r:} = u_r W_{r:}$. Thus IG invariance is not created by normalization. These facts concern equivalent parameterizations: finite-step optimizers can follow different trajectories under rescaling and need not yield corresponding coefficients or equal approximate errors. An $\ell_1$ penalty or a dictionary-norm constraint changes the optimization problem, unlike this product-preserving transformation.

In Eq. (2), adding fixed $\varepsilon$ ensures strictly positive $\gamma_r$, even for an unused column. With finite $\varepsilon$, independently rescaled inputs need not map to exactly the same normalized representative; training maxima are $m_r/(m_r + \varepsilon)$, not exactly one. Product preservation remains exact algebraically. No accuracy improvement, sparsity, or semantic identifiability follows from this operation.

## K Result Sources and Reconstruction Fidelity

We inspected the completed July 25, 2026 audit archive without rerunning experiments. All 43 result JSON files matched their recorded SHA-256 hashes; the completion marker reports zero failed jobs. The recovered code establishes the distinct numerical protocol in Appendix A. This integrity check verifies stored artifacts, not an independent reproduction. The original main-table optimizer and its linkage to these later bundles remain unresolved.

Table 3 reports the archived bundle fidelity in the main text. The historical manuscript reports 88.80% for ImageNet-100, whereas the archived raw predictions report 88.82%. We retain the historical value with its route label and give the archived value separately; the archive does not establish the origin of the 0.02-point discrepancy. The archived test-feature negative-entry fractions are 60.23%, 51.34%, and 59.56%, supporting the mixed-sign motivation for these bundles.

Table 8: Archived decomposition controls. TopK-SAE widths are 600/700/900 with 32 active coordinates per sample, seed 42, 30 epochs, and the bundle's frozen head. Matching width does not match sparsity or reconstruction error. PCA uses the corresponding rank. Values are retained from stored E6 result files, not recomputed.

| Dataset | Method | Recon. acc. (%) | MSE | Insertion | Deletion |
|---|---|---|---|---|---|
| CIFAR-100 | TopK-SAE | 80.37 | 0.02090 | 0.59605 | 0.31509 |
| | PCA | 81.20 | 0.00615 | 0.74458 | 0.04450 |
| CUB | TopK-SAE | 79.79 | 0.07217 | 0.45128 | 0.31103 |
| | PCA | 81.01 | 0.01543 | 0.71232 | 0.02547 |
| ImageNet-100 | TopK-SAE | 88.26 | 0.06739 | 0.67426 | 0.43210 |
| | PCA | 88.82 | 0.01318 | 0.85676 | 0.04791 |

These separate archived diagnostics do not establish a controlled ranking against historical FLACE: PCA has higher insertion and lower deletion than the historical FLACE rows, and TopK-SAE is stronger on both on CIFAR-100. The comparison does not isolate nonnegativity, active width, or optimizer effects, and the historical/archived provenance boundary prevents treating the numerical gap as a controlled causal ablation. PCA also permits signed coefficients, changing the coordinate interface.

The archived E3 analysis inherits descriptors from existing named spaces: it selects alignments on a calibration split and accepts them only if the correlation sign is stable and the smaller absolute calibration/evaluation correlation is at least 0.3. Coverage is 38/600 (6.33%), 64/700 (9.14%), and 36/900 (4.00%). This supports names for a limited subset under that protocol, not universal naming, a WordNet algorithm, or human agreement. The original figure-caption naming rubric remains unrecorded. No medical, cross-backbone, or human-correction claim is inferred from these results.

## L Existing Assets and Use Terms

Table 9: **Existing assets used in the experiments.** We cite the original sources and do not redistribute raw datasets or third-party model assets in the paper PDF. Any supplementary code release should preserve upstream notices and require users to obtain restricted assets from their original sources.

| Asset | Source / credit | Use terms in this submission |
|---|---|---|
| CIFAR-100 | Krizhevsky and Hinton Krizhevsky et al. (2009) | Public research benchmark; raw data are not redistributed and should be obtained from the original dataset source under its accompanying terms. |
| CUB-200-2011 | Wah et al. Wah et al. (2011) | Public academic benchmark; raw images and annotations are not redistributed and should be obtained from the original dataset source under its accompanying terms. |
| ImageNet-100 subset | ImageNet / ILSVRC Deng et al. (2009) | Derived subset used for evaluation only; ImageNet images are not redistributed and must be accessed under the original ImageNet terms. |
| EfficientNet-B3 | Tan and Le Tan & Le (2019) | Used as the classifier architecture/backbone. Public implementations or pretrained weights, if used in reproduction code, should follow their upstream release terms. |
| Prior CBM baselines | Res-CBM, Post-hoc CBM, Label-free-CBM, and CF-CBM Shang et al. (2024); Yuksekgonul et al. (2023); Oikarinen et al. (2023); Panousis et al. (2024) | Used for method comparison under cited research settings; any third-party baseline code included in supplementary material should retain its original license notices. |
| WordNet | Miller Miller (1995) | Optional resource for post-hoc visualization descriptors, not for FLACE concept discovery, predictor construction, or test-time inference. |