DUAL CONVEX COMPOSITION FOR CLIP INTERPRETABILITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept decompositions of CLIP need both inspectable provenance and a clear scale for their activations. We present DCC-CLIP, which constructs each concept as a convex mixture of in-domain sample embeddings and reconstructs each sample as a convex mixture of concepts. A cosine objective fits the directional representation while the two weight systems expose dictionary provenance and sample composition. We establish hull containment, sparse representability, and fixed-weight perturbation bounds, distinguishing these properties from semantic interpretability and optimizer stability. Experiments compare adapted NMF/CRAFT and PCA on image classification and image–text alignment: DCC-CLIP obtains reconstruction cosine similarities of 0.947–0.993 across the reported dataset–backbone pairs. Semantic-consistency, discriminative-retention, and insertion/deletion evaluations examine complementary aspects of the resulting explanations. Code is provided in the supplementary material.

1 INTRODUCTION

A concept explanation of CLIP should connect its embedding coordinates to evidence a reader can inspect. CLIP aligns image and text representations by direction (Radford et al., 2021), but a decomposition of that space need not produce concepts with identifiable sample provenance or comparable coefficients. A large activation may reflect an arbitrary scaling of a factor rather than stronger participation in the reconstruction. A visually plausible exemplar, meanwhile, need not reveal how the concept was constructed.

We study a constrained explanation interface with two explicit weight systems. Each concept is a convex mixture of in-domain sample embeddings, recording its provenance. Each reconstructed sample is a convex mixture of those concepts, recording its composition. The first weights explain where a dictionary element comes from; the second explain how the dictionary reconstructs an image. Both are nonnegative and normalized. We call this factorization *Dual Convex Composition for CLIP Interpretability* (DCC-CLIP).

This interface differs from specifying a named direction with TCAV (Kim et al., 2018) or decomposing nonnegative activations with CRAFT (Fel et al., 2023). CLIP's signed coordinates and cosine comparisons motivate a directional reconstruction objective. Convex coefficients can mix signed vectors without clipping their coordinates, while the common simplex scale removes reciprocal rescaling of a concept and its activation. These structural properties make the decomposition inspectable; whether its concepts are semantically coherent is an empirical question.

We evaluate that question through text-side semantic consistency, reconstruction, discriminative retention, and concept insertion/deletion on classification and image–text alignment tasks. Adapted NMF/CRAFT and PCA use the same CLIP backbones but different input constructions. The evidence covers CUB, an ImageNet subset, Flickr8k, and Conceptual Captions. It evaluates an embedding explanation, rather than downstream improvements in deployed retrieval, recommendation, or moderation.

Our contributions are a dual-simplex decomposition with explicit sample provenance; a geometric analysis separating convex containment, sparse representability, and perturbation sensitivity; and a quantitative comparison of semantic consistency and retained predictive information. This separation

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.

matters: containment alone neither assigns a human name to a concept nor establishes robustness of the learned optimizer.

2 Related Work

Attribution and concept interfaces. Grad-CAM, Integrated Gradients, LIME, and SHAP explain predictions through regions or input features (Selvaraju et al., 2017; Sundararajan et al., 2017; Ribeiro et al., 2016; Lundberg & Lee, 2017). TCAV instead tests sensitivity to a user-specified concept direction learned from positive and negative examples (Kim et al., 2018). Concept bottleneck and concept embedding models place concept-level intermediates inside a predictor (Koh et al., 2020; Espinosa Zarlenga et al., 2022). DCC-CLIP keeps the pretrained encoder fixed and decomposes its representations. Its sample weights provide a provenance interface, not supervised concept labels or a replacement prediction architecture.

Factorizing representations. CRAFT uses nonnegative matrix factorization to discover concepts from network features (Fel et al., 2023; Lee & Seung, 2000); PCA supplies an orthogonal low-rank alternative (Jolliffe, 2002). Their constraints differ from the two simplex constraints here. A nonnegative factorization needs an adaptation for signed CLIP coordinates, whereas a convex combination of signed embeddings remains well defined. Our experiments compare a CLIP-adapted NMF/CRAFT implementation, not every possible adaptation of CRAFT. Concept-based representation comparisons also motivate inspecting differences between learned feature spaces (Kondapaneni et al., 2025; Shah et al., 2023).

Relation to archetypal analysis. The two convex-mixture constraints are the classical structure of archetypal analysis (Cutler & Breiman, 1994), which approximates samples using archetypes that are themselves mixtures of samples. We do not claim that this structure or its hull properties are new. DCC-CLIP uses it as a provenance interface for normalized contrastive embeddings, with a cosine objective and CLIP-specific semantic and reconstruction evaluation. The relevant distinction is the directional objective and explanation setting, rather than a new convex factorization principle.

Explaining CLIP. Work on CLIP explanations diagnoses foreground–background misalignment and noisy activations, and develops regional attribution mechanisms (Li et al., 2025; Zhao et al., 2024; Shalev & Wolf, 2024). These methods address spatial evidence. Our factorization asks how an embedding can be reconstructed through a shared dictionary and how each dictionary vector traces to samples. Convex geometry supplies the containment and representation properties (Matoušek & Matoušek, 2002; Vandenberghe & Boyd, 2004); semantic interpretation still requires inspecting prototypes and evaluating their agreement.

3 Preliminaries

Let $z_i, g_i \in \mathbb{R}^d$ be image and text embeddings from a frozen CLIP model, with

$$\|z_i\|_2 = \|g_i\|_2 = 1, \qquad \cos(z_i, g_i) = z_i^\top g_i. \tag{1}$$

Write $Z \in \mathbb{R}^{n\times d}$ and $G \in \mathbb{R}^{n\times d}$ for the matrices with these embeddings as rows. We factorize Z; the construction applies symmetrically to text. A dictionary $W \in \mathbb{R}^{k\times d}$ and composition matrix $B \in \mathbb{R}^{n\times k}$ produce reconstructions $\hat{Z} = BW$.

4 Method

4.1 Two normalized weight systems

An unconstrained factorization can preserve its reconstruction under reciprocal rescaling of dictionary vectors and coefficients. DCC-CLIP instead constrains the coefficients on both sides of the dictionary:

$$W = C^\top Z, \quad C \in \mathbb{R}^{n\times k}, \quad C_{ij} \geq 0, \quad \sum_{i=1}^{n} C_{ij} = 1. \tag{2}$$

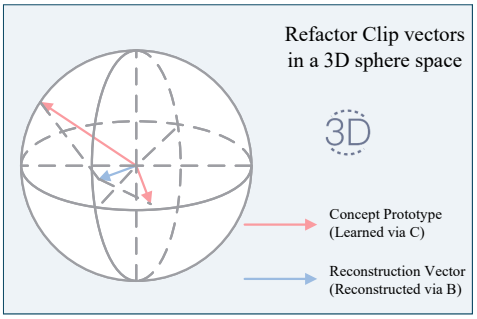


Figure 1: Directional reconstruction and convex composition. Unit-norm sample embeddings define a hull; their convex mixtures generally lie inside the sphere. Cosine compares a nonzero mixture's direction with the sample direction. Hull containment applies before radial normalization.

Column j of C records the samples contributing to concept w_j. A large coefficient identifies a large contribution to that vector, rather than an exemplar selected only after fitting. This is traceability to the training pool; the pool itself determines the available evidence.

The second system represents a sample through the shared concepts:

$$\hat{Z} = BW, \quad B \in \mathbb{R}^{n \times k}, \quad B_{ij} \geq 0, \quad \sum_{j=1}^{k} B_{ij} = 1. \tag{3}$$

Each row of B expresses relative participation on a common $[0, 1]$ scale. It is a composition, not a set of independently calibrated probabilities of semantic presence. Increasing one coefficient necessarily reduces others. Different dictionaries can still assign different meanings to their columns, so cross-model comparisons require inspecting the concepts rather than matching column indices.

Directional objective. Because CLIP compares directions, we fit the factorization by

$$B^*, C^* = \arg\min_{B,C} \frac{1}{n} \sum_{i=1}^{n} -\cos\big(z_i, B_i C^\top Z\big). \tag{4}$$

The cosine is defined for nonzero reconstructions. Softmax maps unconstrained logits onto columns of C and rows of B, enforcing the simplex conditions during optimization. The constraints concern the unnormalized vectors W and $\hat{Z}$. Cosine comparison normalizes their directions but does not require every convex mixture to have unit length. Figure 1 illustrates this distinction; Appendix A shows the pipeline.

4.2 Geometric properties and their scope

Proposition 1: containment and provenance. Every concept lies in $\mathrm{conv}\{z_1, \ldots, z_n\}$. In particular, for any direction a,

$$\min_i a^\top z_i \leq a^\top w_j \leq \max_i a^\top z_i. \tag{5}$$

Moreover, with $S = \mathrm{conv}\{w_1, \ldots, w_k\}$,

$$\hat{z}_i \in S \subseteq \mathrm{conv}\{z_1, \ldots, z_n\}.$$

To prove both claims, substitute Equation 2 into Equation 3: each reconstruction has nonnegative sample coefficients summing to one. The set S is shared across samples and has *at most* k vertices; some dictionary elements may be redundant. This restricts reconstruction geometry but does not ensure membership in a semantic data manifold, immunity to outliers, or unique concepts. Radially normalizing a nonzero mixture can leave this hull.

Sparse representability. Carathéodory's theorem implies that a point in a d-dimensional convex hull has a representation using at most $d + 1$ support points (Matoušek & Matoušek, 2002; Vandenberghe & Boyd, 2004). Thus each concept admits such a representation in samples, and each reconstruction admits one using at most $\min(k, d+1)$ concepts. This is an existence statement, not a sparsity guarantee for learned softmax coefficients, which are generally dense. In particular, when $k \leq d + 1$, it does not improve on using the whole dictionary. The insertion/deletion experiment separately measures whether a small selected subset retains predictive information.

Proposition 2: perturbation sensitivity. For one concept $w = \sum_i c_i z_i$ with $c_i \geq 0$ and $\sum_i c_i = 1$, perturbing the samples while holding weights fixed gives

$$\|\Delta w\| \leq \sum_i c_i \|\Delta z_i\| \leq \max_i \|\Delta z_i\|. \tag{6}$$

The triangle inequality proves this for any norm; replacing only z_j changes w by exactly $c_j \Delta z_j$. If instead only the weights change and $\|z_i\| \leq R$, then

$$\|\Delta w\| \leq R\|\Delta c\|_1. \tag{7}$$

Changing both samples and weights introduces an additional mixed term:

$$\|\Delta w\| \leq \sum_i c_i \|\Delta z_i\| + R\|\Delta c\|_1 + \sum_i |\Delta c_i| \|\Delta z_i\|. \tag{8}$$

Expand $(c_i + \Delta c_i)(z_i + \Delta z_i) - c_i z_i$ and apply the triangle inequality to obtain this bound. These inequalities describe a specified weighted average. They do not bound changes in weights after refitting or directional sensitivity near a zero-norm concept. Consequently, they support local provenance analysis but not an unconditional claim of optimization stability or generalization.

After normalization. A directional explanation also needs to track the norm of its mixture. For nonzero vectors u, v whose norms are at least $\rho > 0$, adding and subtracting $v/\|u\|_2$ gives

$$\left\| \frac{u}{\|u\|_2} - \frac{v}{\|v\|_2} \right\|_2 \leq \frac{2\|u - v\|_2}{\rho}.$$

Together with Equation 6, this controls the normalized change of a fixed-weight concept only when its norm stays away from zero. Convex mixtures of opposing directions can violate that condition. Inspecting mixture norms alongside provenance weights therefore distinguishes a small change in the underlying average from a potentially large change in its normalized direction. This is an analytic condition, not a measured stability result.

A provenance-removal diagnostic. For a fixed concept $w = \sum_i c_i z_i$, removing fitting sample r and renormalizing the remaining coefficients gives, when $c_r < 1$,

$$w_{-r} = \frac{w - c_r z_r}{1 - c_r}, \qquad w_{-r} - w = \frac{c_r}{1 - c_r}(w - z_r), \qquad \|w_{-r} - w\|_2 \leq 2c_r. \tag{9}$$

The bound follows from $w = c_r z_r + (1 - c_r) w_{-r}$ and the fact that both z_r and the remaining convex mixture have norm at most one. It is exact coefficient bookkeeping, not an experiment. A large coefficient permits a larger change but does not force one: if z_r is close to the remaining mixture, its removal changes little. If both concept norms exceed ρ, the normalized-direction change is at most $4c_r/\rho$ by the preceding normalization bound. When $c_r = 1$, there is no remaining mixture to renormalize.

This diagnostic makes the provenance interface actionable without conflating it with saliency. It asks how a specified prototype changes when one supporting sample is excluded with relative weights fixed. Retraining the dictionary can change all coefficients and is outside the bound. Likewise, removing a fitting sample is not the same intervention as deleting a concept from a test reconstruction in E3.

5 EXPERIMENTS

5.1 DATASETS, MODELS AND IMPLEMENTATION SCOPE

The reported classification datasets are CUB-200 Wah et al. (2011) and a 100-class ImageNet subset Deng et al. (2009); the image–text datasets are Flickr8k Hodosh et al. (2013) and Conceptual Captions (CC) Sharma et al. (2018). Features use CLIP Radford et al. (2021) RN50 and ViT-B/16 backbones He et al. (2016); Dosovitskiy (2020). The supplied loader requests train/test caches, but its dataset module is absent, so exact sample identities, subset selection and split construction cannot be verified from this release.

The supplied extraction script fits separate dictionaries for each ground-truth class on CUB and ImageNet, and one joint dictionary on each caption dataset. Test labels select the class-specific dictionary; this is a conditional analysis, not label-free classification of an unseen image. DCC-CLIP fits concepts on training embeddings, freezes them, then optimizes simplex coefficients for each test embedding by cosine reconstruction. The default transform uses 3,000 Adam steps at learning rate 0.01. These implementation facts specify the available artifact, not a certified reconstruction of every historical evaluation.

5.2 CONCEPT DECOMPOSITION AND BASELINES

We compare DCC-CLIP with adapted NMF/CRAFT Lee & Seung (2000) and PCA Jolliffe (2002). All produce a dictionary $W \in \mathbb{R}^{k\times d}$ and activations B, with reconstruction $\hat{Z} = BW$. DCC-CLIP uses $W = C^\top Z$, nonnegative column-stochastic C, and row-stochastic B, minimizing $1-\cos(z_i, \hat{z}_i)$. The supplied baseline NMF uses nonnegative factors and squared reconstruction loss; its PCA routine uses iterative orthogonal low-rank fitting rather than a centered SVD.

Input construction differs: DCC-CLIP fits cached full-image features, while both adapted baselines fit embeddings of overlapping 64-pixel patches (stride 51), resized before encoding. Their transform then encodes full images. Thus comparisons combine factorization and input-construction choices; they do not isolate the effect of the two simplex constraints. A matched-input ablation remains necessary for that attribution. The supplied metric-generation scripts are also incomplete, limiting verification of historical aggregation and insertion/deletion endpoints. We retain the archived numerical results with these boundaries. Runtime and memory trade-offs are in Section 5.9; additional design qualifications appear in Appendix C.

5.3 METRICS

We report three sets of metrics: E1 for within-concept semantic consistency, E2 for reconstruction consistency, and E3 for concept importance via insertion/deletion curves.

E1: Within-Concept Semantic Consistency. For each concept j, we select the Top-K samples based on $B_{i,j}$ with indices $\{i_1^j, \dots, i_K^j\}$. We collect their normalized text embeddings $\mathcal{T}_j = \{t_{i_m^j}\}_{m=1}^K$ and compute pairwise similarities $[M_j]_{m\ell} = \langle t_{i_m^j}, t_{i_\ell^j}\rangle$. The consistency score is the average of the upper-triangular entries: $\text{Consistency}_j = \frac{2}{K(K-1)}\sum_{m<\ell}[M_j]_{m\ell}$, which is aggregated over $j = 1 \dots k$ to report the mean and standard deviation.

E2: Reconstruction Consistency. We measure the fidelity of reconstructions via three metrics: **(a) Discriminative Consistency:** The archived protocol describes test macro-accuracy of a linear head on normalized reconstructions. The supplied downstream script instead trains on cached features, averages batch accuracies and selects checkpoints using the test split. No supplied evaluator links these operations to the DiscCons table. We therefore retain DiscCons as a reported score, not verified macro-accuracy or an independently reproduced held-out result. **(b) Reconstruction Cosine:** $\text{CosSim}(S) = \frac{1}{|S|}\sum_{i\in S}\langle z_i, \tilde{z}_i\rangle$. **(c) Reconstruction MSE:** $\text{MSE}(S) = \frac{1}{|S|\cdot d}\sum_{i\in S}\|z_i - \hat{z}_i\|_2^2$.

E3: Concept Importance via Insertion/Deletion. We rank concept indices $\pi(1), \dots, \pi(k)$ using gradient or saliency Selvaraju et al. (2017); Sundararajan et al. (2017). We define partial reconstructions with row-wise renormalization: $\tilde{z}_i^{\text{ins}}(m) = \text{norm}(\sum_{j\in\Pi_m} B_{i,j}w_j)$ and $\tilde{z}_i^{\text{del}}(m) = \text{norm}(\sum_{j\in\overline{\Pi_m}} B_{i,j}w_j)$, where $\Pi_m = \{\pi(1), \dots, \pi(m)\}$. Letting $f(\cdot)$ denote the linear head's soft-

Table 1: E1 within-concept semantic consistency (higher is better). The metric averages pairwise cosine similarities of text embeddings for the Top-K (K=3) samples per concept, then averages over all concepts; the Δ columns report absolute improvements over baselines.

Dataset	Backbone	DCC-CLIP (Ours)	NMF	PCA	Δ –NMF	Δ –PCA
CC	ViT-B/16	0.4937	0.4722	0.4639	0.0215	0.0298
Flickr8k	ViT-B/16	0.5310	0.5074	0.5108	0.0236	0.0202
CC	RN50	0.5220	0.5014	0.4810	0.0207	0.0410
Flickr8k	RN50	0.4791	0.4714	0.4764	0.0077	0.0027

Protocol. All CLIP text and image embeddings are ℓ_2-normalized prior to similarity computation; consistency scores are averaged across concepts.

Table 2: E2 reconstruction consistency (**DiscCons**, higher is better). Archived discriminative scores are retained; the supplied code does not verify their stated macro-accuracy protocol. The Δ columns give absolute improvements over the baselines.

Dataset	Backbone	DCC-CLIP (Ours)	NMF	PCA	Δ –NMF	Δ –PCA
CUB	RN50	1.0000	0.3864	0.1116	0.6136	0.8884
CUB	ViT-B/16	0.9993	0.5559	0.5637	0.4434	0.4357
ImageNet	RN50	0.9954	0.6818	0.1002	0.3136	0.8952
ImageNet	ViT-B/16	0.9964	0.8110	0.6646	0.1854	0.3318

Protocol. Consistent with the method setup: image/text embeddings are first ℓ_2-normalized; reconstructed vectors are then row-wise normalized (yielding $\tilde{Z}$) for discriminative evaluation on the sphere.

max confidence for the true class (or $f(\cdot) = \langle \cdot, t_i \rangle$ for image-text matching), we compute the partial curves $I(m) = \frac{1}{|S|} \sum_{i \in S} f(\tilde{z}_i^{\text{ins}}(m))$ and $D(m) = \frac{1}{|S|} \sum_{i \in S} f(\tilde{z}_i^{\text{del}}(m))$. We then compute the Area Under the Curve (AUC) via the trapezoidal rule:

$$\text{AUC}_f = \frac{1}{2k} \sum_{m=0}^{k-1} \big[f(m) + f(m+1) \big], \tag{10}$$

where substituting f with I or D yields INS-AUC and DEL-AUC, respectively. Normalization is undefined at a zero partial sum, including the empty insertion and fully deleted endpoints. Their implemented convention cannot be recovered from the supplied release, so the archived curves do not independently establish that endpoint choice.

5.4 Results Analysis

5.5 E1: Within-Concept Semantic Consistency

Table 1 reports higher text-side within-concept consistency for DCC-CLIP in all four dataset–backbone configurations. Absolute improvements over the baselines range from 0.0027 to 0.0410. The Flickr8k RN50 difference from NMF is small, while the largest difference is from PCA on CC RN50. Since factorization uses image embeddings and this metric uses their associated text embeddings, it measures cross-modal coherence of the selected samples. It does not establish that the concepts have unique human-readable meanings or that containment alone caused the improvement.

The distinction between the two weight systems is important here. Top-activation samples selected with B need not be identical to the samples with largest provenance weights in C. Text consistency evaluates the former; inspecting C answers how the vector was constructed. A complete concept audit can use both without treating either as a causal explanation of the other. Component ablations would be needed to isolate the contributions of the cosine objective and each simplex constraint.

5.6 E2: Reconstruction Consistency

Discriminative consistency. Table 2 records DCC-CLIP DiscCons values of 0.995–1.000 on CUB and ImageNet with both backbones, above the adapted NMF and PCA results. The largest displayed

Table 3: E2 reconstruction consistency (**CosSim**, higher is better). For each sample, we compute the cosine similarity between the original embedding and the row-wise normalized reconstruction $\tilde{Z}$, and report the test-set average; the Δ columns give absolute improvements over the baselines.

Dataset	Backbone	DCC-CLIP (Ours)	NMF	PCA	Δ –NMF	Δ –PCA
CUB	RN50	0.9869	0.4203	0.5444	0.5666	0.4425
CUB	ViT-B/16	0.9929	0.7777	0.8247	0.2153	0.1682
ImageNet	RN50	0.9473	0.4038	0.6078	0.5435	0.3396
ImageNet	ViT-B/16	0.9724	0.7845	0.8222	0.1879	0.1502

Protocol. As in the method: image embeddings z_i are ℓ_2-normalized; reconstruction vectors $\hat{z}_i$ are row-wise normalized to obtain $\tilde{z}_i$, and CosSim aggregates $\langle z_i, \tilde{z}_i \rangle$ over the test set.

Table 4: E2 reconstruction consistency (**MSE**, lower is better). Values are scaled by $\times 10^{-4}$ for readability; the Δ columns report *error reductions* (Baseline $-$ Ours; higher is better).

Dataset	Backbone	DCC-CLIP (Ours)	NMF	PCA	Δ –NMF	Δ –PCA
CUB	RN50	0.251	8.023	6.836	7.771	6.585
CUB	ViT-B/16	0.273	7.703	6.214	7.430	5.941
ImageNet	RN50	1.009	8.138	6.120	7.130	5.112
ImageNet	ViT-B/16	1.068	7.490	6.304	6.422	5.236

Protocol. As in the method: z_i are ℓ_2-normalized original embeddings; MSE is computed on the unnormalized reconstructions $\hat{z}_i$ and then averaged over samples and dimensions; reported values are multiplied by 10^{-4}.

difference is on ImageNet RN50: 0.9954 for DCC-CLIP and 0.1002 for PCA. These archived scores should be read separately from reconstruction cosine and with the unresolved evaluation linkage above.

For normalized reconstructions $\tilde{z}_i$ and any fixed linear weight vector w, Cauchy–Schwarz gives

$$|w^\top (z_i - \tilde{z}_i)| \le \|w\|_2 \sqrt{2(1 - \cos(z_i, \tilde{z}_i))}.$$

This controls the change of a specified linear score. Preserving a predicted label additionally requires the original decision margin to exceed the possible changes relative to competing classes. It does not prove that refitting a classifier must recover its original accuracy. The comparison measures discrimination empirically; the bound explains why small angular error can be relevant to a fixed head without supplying a causal account of the baseline differences.

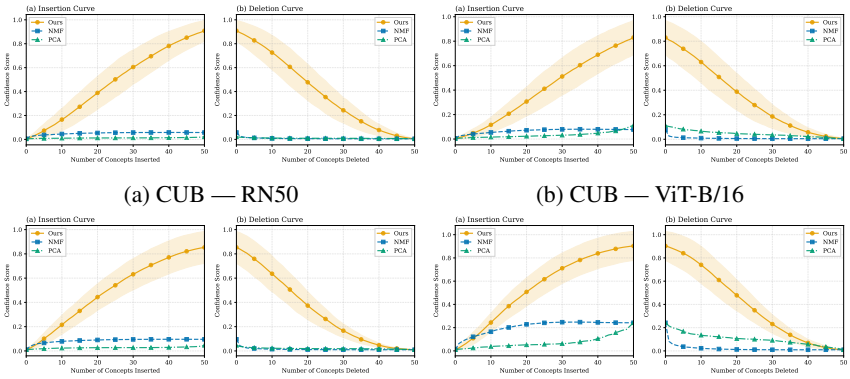
Reconstruction cosine. DCC-CLIP has the highest recorded cosine in every dataset–backbone pair in Table 3, ranging from 0.947 to 0.993. The largest absolute advantage exceeds 0.56. This metric directly matches the directional training objective, so it establishes reconstruction performance in the optimized geometry rather than independent semantic validation. The text-side consistency and discrimination metrics supply the complementary tests.

Hull containment does not itself bound the reconstruction angle: a small-norm convex mixture can point in an unstable direction. Nor does a low concept budget eliminate every concern about memorization. Each fitted dictionary has k elements, but the classification implementation uses one dictionary per ground-truth class and its provenance weights depend on the fitting samples. The comparison therefore needs to be interpreted with the stated data splits and inference procedure; better cosine alone does not establish generalization to a new domain.

Reconstruction MSE. Table 4 reports lower unnormalized reconstruction MSE for DCC-CLIP in all evaluated settings. For CUB RN50, the displayed entries are 0.251 for DCC-CLIP, 8.023 for NMF and 6.836 for PCA; for ImageNet RN50 they are 1.009, 8.138 and 6.120. Keep the table's scaling convention when interpreting these values.

The Euclidean metric also depends on reconstruction length. Writing $r_i = \|\hat{z}_i\|_2$ and θ_i for its angle with z_i gives

$$\|z_i - \hat{z}_i\|_2^2 = (r_i - \cos\theta_i)^2 + \sin^2\theta_i.$$



(a) CUB — RN50 (b) CUB — ViT-B/16

(c) ImageNet — RN50 (d) ImageNet — ViT-B/16

Figure 2: E3 concept importance via insertion/deletion curves. Each subplot shows both insertion and deletion (C=50, Grad-CAM ranking). The desired trend is a faster rise for insertion and a faster drop for deletion.

Although convex containment implies $r_i \leq 1$, it does not make the first term vanish. Perfect directional agreement with $r_i = 1/2$, for example, still gives squared error $1/4$. The lower MSE is therefore a reported empirical finding, not a consequence guaranteed by minimizing cosine loss. Reporting both metrics distinguishes direction preservation from preservation of magnitude.

5.7 E3: Insertion and Deletion Analysis

Figure 2 evaluates partial reconstructions ordered by concept activation. The displayed DCC-CLIP insertion curves rise earlier and deletion curves fall earlier than the adapted NMF and PCA curves. Under this reconstruction intervention, a small leading subset retains substantial predictive information, and removing it reduces the measured score.

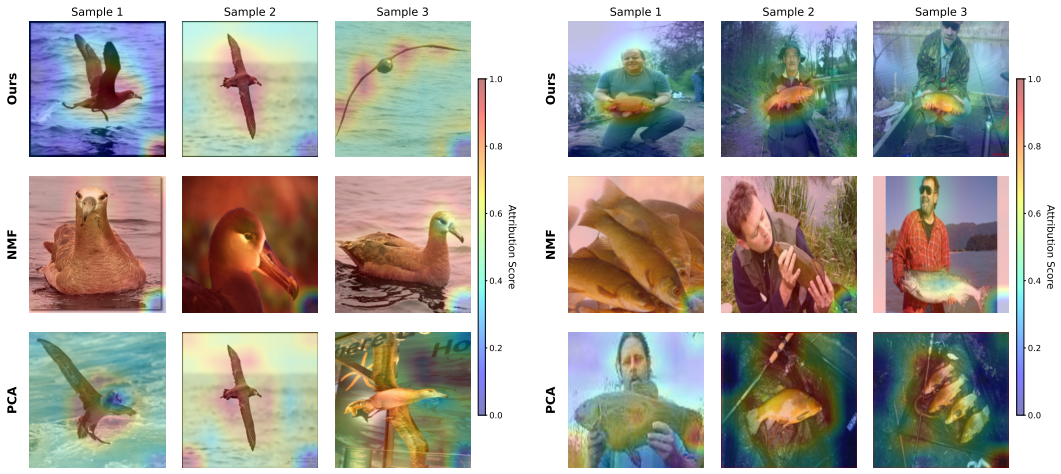
Each partial sum is normalized before evaluation. Consequently, changing one subset changes both its composition and its direction; the curves are not additive contributions on the original feature scale. They establish sufficiency and removal sensitivity for the evaluated score and ordering, not a unique necessary concept set or causal importance of image regions. The sparse-representation existence result does not guarantee these empirical curves. Their value is precisely that they test which information the fitted dictionary retains beyond its formal containment property.

5.8 Qualitative Heatmap Analysis

Figure 3 illustrates concept-associated spatial responses on CUB and ImageNet with RN50. The DCC-CLIP examples emphasize compact object regions, while some baseline responses are more diffuse. These selected images help inspect the learned representation; they do not measure localization accuracy or establish that the dictionaries contain identical semantic factors. The plotted concept indices are internal to each method unless an explicit cross-method alignment is supplied. Provenance is separately available through the sample coefficients in C, while heatmaps visualize a concept's response on the displayed image.

5.9 Runtime and Memory Footprint

Table 5 preserves the reported NVIDIA L40s measurements. DCC-CLIP uses less peak GPU memory in each listed comparison but takes longer than PCA. On CUB RN50, for example, runtime rises from 6.49 to 15.36 seconds while peak memory falls from 1160.77 to 411.33 MB. The time penalty is not uniformly small: CC RN50 takes 16.54 seconds rather than PCA's 3.03 seconds. Against NMF, DCC-CLIP is slower in seven of eight settings and faster on CC ViT-B/16. These are implementation-specific trade-offs, not asymptotic guarantees from convexity. The dense provenance and activation matrices each store nk coefficients, so increasing the sample pool or concept budget still increases



(a) CUB — RN50 (b) ImageNet — RN50

Figure 3: E3 concept importance via insertion/deletion: qualitative heatmaps. Each subplot visualizes selected concepts from the three methods (Ours/NMF/PCA) over multiple samples, together with attribution scores.

Table 5: Runtime (seconds) and peak GPU memory (MB) of concept extraction across datasets. Lower is better.

Model	Method	CUB		ImageNet		Flickr8k		CC	
		Runtime (s)	Peak Mem (MB)	Runtime (s)	Peak Mem (MB)	Runtime (s)	Peak Mem (MB)	Runtime (s)	Peak Mem (MB)
RN50	DCC-CLIP	15.36	411.33	10.58	411.05	9.43	409.04	16.54	409.04
RN50	PCA	6.49	1160.77	7.58	1174.07	5.72	773.12	3.03	773.12
RN50	NMF	13.16	1160.77	9.78	1174.07	8.35	773.12	11.44	773.12
ViT-B/16	DCC-CLIP	16.11	590.91	18.87	590.78	16.66	589.72	9.64	589.72
ViT-B/16	PCA	6.58	1247.02	5.28	1258.13	4.41	914.60	5.50	914.60
ViT-B/16	NMF	14.42	1247.02	15.77	1258.13	7.28	914.60	11.18	914.60

storage. Runtime and memory should therefore accompany semantic quality when selecting a dictionary budget.

6 Limitations

Our current evaluation emphasizes semantic consistency, reconstruction fidelity, and discriminative retention, which provide indirect rather than end-to-end evidence of preserved alignment; dedicated retrieval-oriented system evaluation remains future work. The selected high-error examples in Appendix B include text-heavy images, but do not isolate text as their cause. Finally, the concept budget k reflects a structural trade-off between reconstruction flexibility and human auditability.

7 Conclusion

DCC-CLIP links a shared concept dictionary to the samples that construct it and links each reconstruction to that dictionary through normalized coefficients. Its cosine objective fits CLIP's directional geometry, while the two convex constraints provide explicit provenance. The reported comparisons assess semantic consistency, reconstruction and retained predictive information. Hull containment and bounded changes of specified mixtures explain the interface's geometric scope; human concept interpretation and robustness after refitting remain separate empirical questions.

AI USE STATEMENT

Generative AI tools assisted with drafting and revising parts of the manuscript. The authors reviewed the suggested text and remain responsible for the experimental design, implementation, data, numerical results, methodological claims, and final text.

REFERENCES

Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994. doi: 10.1080/00401706.1994.10485840.

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.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

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 Bethune, 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.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 2016.

Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 593–600. IEEE, 2013.

Ian T. Jolliffe. *Principal Component Analysis*. Springer, 2002.

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.

Neehar Kondapaneni, Oisin Mac Aodha, and Pietro Perona. Representational similarity via interpretable visual concepts. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=ih3BJmIZbC. ICLR 2025; see also arXiv:2503.15699 and project page.

Daniel Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13, 2000.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognition*, 2025.

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

Jiří Matoušek and Jiří Matoušek. *Lectures on discrete geometry*, volume 212. Springer, 2002.

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 (ICML)*, 2021. CLIP.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1135–1144, 2016.

Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *IEEE International Conference on Computer Vision (ICCV)*, 2017.

Harshay Shah, Sung Min Park, Andrew Ilyas, and Aleksander Madry. Modeldiff: A framework for comparing learning algorithms. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202, pp. 30646–30688. PMLR, 2023. URL `https://proceedings.mlr.press/v202/shah23a.html`.

Yoav Shalev and Lior Wolf. Fine-tuning clip via explainability map propagation for boosting image and video retrieval. In *Advances in Information Retrieval: 46th European Conference on Information Retrieval (ECIR 2024)*, volume 14608 of *Lecture Notes in Computer Science*, pp. 355–373, 2024.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2556–2565. ACL, 2018.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

Lieven Vandenberghe and Stephen Boyd. *Convex optimization*, volume 1. Cambridge university press Cambridge, 2004.

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.

Chenyang Zhao, Kun Wang, Xingyu Zeng, Rui Zhao, and Antoni B Chan. Gradient-based visual explanation for transformer-based clip. In *International Conference on Machine Learning*, pp. 61072–61091. PMLR, 2024.

A FLOWCHART DETAILS

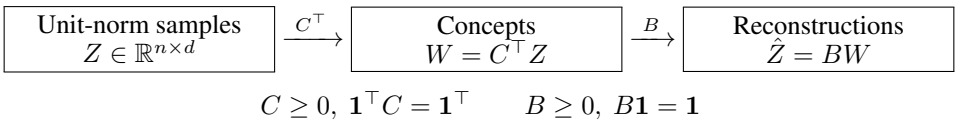


Figure 4: **DCC-CLIP pipeline with consistent row-vector notation.** C records prototype provenance; B records sample composition. The original diagram is retained as an asset but replaced here to avoid its conflicting matrix notation.

Figure 4 visualizes the *DCC-CLIP* pipeline and the roles of Z, W, C, and B.

Notation. Given image–text pairs, the CLIP image encoder produces unit-norm embeddings $Z = [z_i] \in \mathbb{R}^{n \times d}$ with $\|z_i\|_2 = 1$. Concepts are collected in $W = [w_j] \in \mathbb{R}^{k \times d}$. The concept-index matrix $C \in \mathbb{R}^{n \times k}$ is nonnegative and column-stochastic ($\sum_i C_{ij} = 1$); the activation matrix $B \in \mathbb{R}^{n \times k}$ is nonnegative and row-stochastic ($\sum_j B_{ij} = 1$).

Forward flow. (i) Concept synthesis (in-hull):

$$W = C^\top Z \quad \Rightarrow \quad w_j \in \mathrm{conv}(Z).$$

(ii) Sample reconstruction (row-stochastic):

$$\hat{Z} = BW \quad \Rightarrow \quad \hat{z}_i \in S,\ \ S = \mathrm{conv}(W) \subseteq \mathrm{conv}(Z).$$

Row-wise normalization yields $\tilde{Z} = \mathrm{row_norm}(\hat{Z})$ for cosine-based evaluation.

Pipeline Analysis. The two weight systems expose different relationships: C identifies the fitting samples used to construct each prototype, whereas B expresses each reconstruction using the prototypes. Both have normalized nonnegative coefficients. Convex containment holds before row normalization and does not imply containment on the data manifold or preservation of all cross-modal semantics. A low-norm mixture can also have an unstable normalized direction. The diagram describes the algebraic interface, not an additional empirical guarantee.

B FAILURE CASE ANALYSIS

Selected high-error examples. Figure 5 includes images with prominent text or graphic layout. Their stored filenames record cosine-error values between approximately 0.279 and 0.318, and per-coordinate MSE around 0.001. These selected examples identify difficult cases for the recorded dictionary; they do not estimate their prevalence or show that text causes the reconstruction error. Since cosine and per-coordinate MSE have different normalization, these values should not be interpreted as proportional changes of one quantity.

An OCR-aware or domain-specific dictionary is a possible future comparison, not an evaluated mitigation. Establishing the proposed explanation would require controlled text removal or matched text-heavy and text-light examples with the same evaluation protocol.

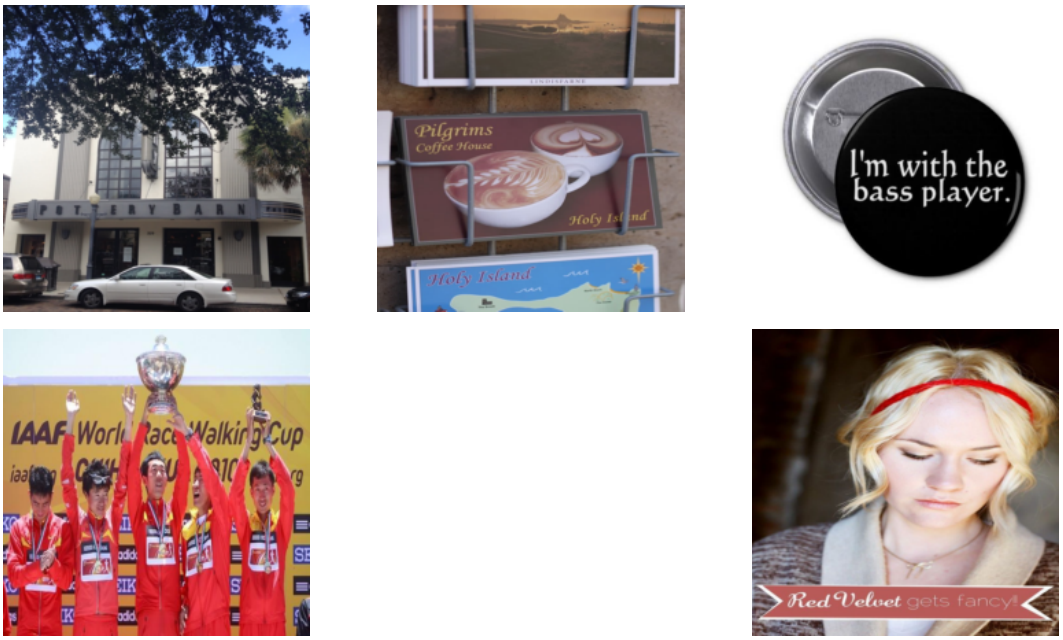


Figure 5: Five retained high-error examples from CC with ViT-B/16. An unavailable-image placeholder in the original panel is omitted; no replacement observation is introduced.

C Questions You May Have

To provide deeper insights into our design choices and address potential questions regarding the positioning of DCC-CLIP, we provide the following clarifications.

Q1: How does DCC-CLIP compare to Sparse Autoencoders (SAEs), and why were SAEs not used as the primary baseline?

A: NMF/CRAFT and PCA provide contrasting additive and low-rank baselines, but do not exhaust concept decomposition methods. SAEs can be adapted to normalized embeddings or cosine objectives and remain a relevant missing comparison. DCC-CLIP's explicit sample-provenance constraint differs from a generic sparse decoder; this distinction does not establish superior explanatory quality without a matched experiment.

Q2: The double convex combination formulation shares mathematical similarities with Archetypal Analysis. What is the distinct contribution here?

A: The idea of representing data via convex hulls (archetypal analysis) is indeed classical. Our novelty lies not in convex-hull modeling itself, but in adapting a dual convex composition to normalized contrastive embeddings with a cosine-preserving objective. By constraining both concept synthesis and sample reconstruction to the simplex while optimizing $\mathbb{E}[1 - \cos(z, \hat{z})]$, we better preserve the directional structure of the shared embedding space. The key contribution of DCC-CLIP is thus to integrate this classical geometry into a cosine-consistent, dual-convex-composition framework for multimodal embeddings, yielding auditable in-hull provenance and comparable concept strengths.

Q3: The experimental evaluation uses a concept budget of $k \leq 50$. Does the method scale to a larger number of concepts?

A: A compact concept budget is a deliberate design choice for auditability and comparability. When extracting macro-level, human-auditable concepts for a specific domain (e.g., CUB bird species), a compact set of concepts is substantially easier to inspect and validate. In our setting, increasing k without bound is not necessarily desirable, because it may reduce interpretability and increase semantic redundancy or superposition, even if reconstruction flexibility increases. The k-vertex representation constrains each reconstruction, but the current results do not measure scaling beyond

this range. Increasing k also increases the dense coefficient storage; larger-budget quality and cost remain empirical questions.

Q4: The paper relies on E1, E2, and E3 metrics. Why not include extensive end-to-end downstream tasks like cross-modal retrieval or human evaluations of concept clarity?

A: Our current evaluation focuses on geometry-preserving concept consistency (E1), reconstruction fidelity, and discriminative retention (E2, E3), rather than end-to-end retrieval benchmarks or subjective human studies. Since CLIP retrieval depends on preserving cosine-based alignment in the shared embedding space, improvements in semantic consistency and reconstruction cosine provide relevant indirect evidence of preserved cross-modal alignment. However, these proxy metrics do not fully replace dedicated retrieval evaluations or human-centered studies of concept clarity, which we explicitly leave to future work as part of broader system-level validation.

Q5: How are the strict simplex constraints ($\sum = 1$ and ≥ 0) practically ensured during the optimization?

A: In our DCC-CLIP implementation, we enforce the simplex constraints through a differentiable softmax reparameterization. Instead of directly optimizing the coefficient matrices, we optimize unconstrained latent logits for the sample-to-concept and concept-to-sample coefficients. During the forward pass, these logits are mapped onto the probability simplex via softmax, ensuring nonnegativity and unit-sum normalization of the corresponding coefficients. This keeps the optimization fully differentiable while preserving the convex-combination structure used for concept synthesis and sample reconstruction.