WHERE COMPOSITIONALITY HIDES: SHARED-CONTENT INTERFERENCE IN CONTRASTIVE VISION-LANGUAGE RETRIEVAL

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrastive vision–language models struggle with compositional retrieval, but a low retrieval score does not by itself establish that the relevant information is absent from their representations. We identify an exact, format-dependent bottleneck in the scoring geometry. Removing the mean of two normalized candidate captions from a frozen image embedding preserves every 1×2 caption decision, yet forces the 2×2 Group Score to equal the original Text Score. On CLIP ViT-B/32 this closes the full 22-point Winoground Group–Text gap, increasing Group Score from 9.0% to 31.0% and Image Score from 11.25% to 64.5%, with replication across five models and ColorSwap. An equivalent score decomposition shows that 213 of 355 raw Image failures have a positive joint-matching margin but fail the cross-image offset threshold. A complementary probe-free analysis reports that the highest-magnitude 20% of coordinates carry 57.3% of absolute similarity mass but only 20.2% of the reported discriminative share on object swaps. We distinguish this basis-dependent observation from the exact projection result. Tested global corrections do not recover the paired diagnostic's gain, while a top-K approximation recovers part of it in the evaluated setting. Together, the results establish retrieval geometry as a concrete bottleneck in paired compositional evaluation: some useful discrimination is already present in frozen pooled features, although gap closure alone neither proves complete compositional representation nor establishes open-gallery retrieval improvement.

1 INTRODUCTION

Contrastive Language-Image Pre-training (Radford et al., 2021) and its successors (Zhai et al., 2023; Cherti et al., 2023) dominate vision-language representation learning, yet fail systematically at compositional reasoning (Yuksekgonul et al., 2023; Hsieh et al., 2023; Ma et al., 2023). Shown "a mug on a book" versus "a book on a mug," zero-shot CLIP performs near chance (Thrush et al., 2022).

One explanation is loss of compositional information: the contrastive objective rewards object-level bag-of-words features at the expense of spatial and relational structure, and the cross-modal projection is where that structure is lost. This diagnosis has motivated a large literature that either retrains the encoder with hard negatives (Yuksekgonul et al., 2023; Doveh et al., 2023; Patel et al., 2024; Peleg et al., 2025) or injects structure at inference time (Miranda et al., 2026; Bhattacharya, 2026; Kang et al., 2025). These approaches need not assume total information loss; binding failures (Lewis et al., 2024) can reflect both representation and readout limitations.

We isolate a readout limitation without changing the pooled features. Our claims are deliberately scoped, and we state them as three separate hypotheses so that each can be evaluated against its own evidence:

- **(H1) Readout bottleneck.** Useful discrimination is present in frozen pooled features on a subset of examples, yet the original cross-image comparisons fail.
- **(H2) Magnitude allocation.** In the evaluated basis, high-magnitude dimensions carry disproportionate absolute similarity mass relative to measured discriminability.

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.

(H3) Candidate dependence. The paired diagnostic is content-adaptive, and the global corrections tested here do not reproduce its gain.

Our central instrument is a closed-form *mean-erasure* projection. Given two candidate captions C_0, C_1, we project the image embedding orthogonally away from their mean $C_{\text{mean}} = \frac{1}{2}(C_0 + C_1)$. This is a diagnostic, not a deployable retrieval method, and we treat it as such throughout: it carries no parameters, no training, and no classifier, which is precisely what makes it a controlled test of the readout. The transformation uses both candidate captions but no correct-assignment label. It changes cross-image rankings while preserving every original within-image caption decision.

Contributions.

1. **A structural account of 1×2 versus 2×2 difficulty** (Section 4). We prove mean-erasure is a no-op for 1×2 rankings, that it can alter 2×2 rankings, and that under erasure the Group Score is algebraically *forced* to equal the Text Score (Corollary 2). This gives the gap an exact score-level account without assuming that the removed direction is exclusively semantic.
2. **Empirical closure of that gap** (Section 5, evidence for H1). Mean-erasure lifts CLIP ViT-B/32's Winoground Group Score from 9.0% to 31.0% and Image Score from 11.25% to 64.5%, replicates on five models and on ColorSwap (Burapacheep et al., 2024), with localized controls and a separate score-level audit (Appendix M).
3. **A probe-free measurement of what cosine weights** (Section 6, evidence for H2). Element-wise interaction features sum exactly to the similarity score, making each dimension's contribution directly measurable without any probe. The top-20% highest-magnitude dimensions carry 57.3% of cosine mass but 20.2% of discriminative signal for object swaps, and are over-represented by 2.35–$3.07\times$ in all seven SugarCrepe categories.
4. **Candidate-dependent correction and its alternatives** (Sections 7–8, evidence for H3). Tested global corrections do not recover the paired diagnostic's full gain. A top-K approximation increases Group Score by 5.5 points in the reported evaluation. An LLaVA comparison is treated separately because it uses richer patch features.

Scope. The exact format result covers both object and attribute binding; ColorSwap is specifically a color–object binding test. Recoverability beyond these evaluated tasks is not established by gap closure. The spectral and mean-erasure observations are distinct, and the cross-model comparison does not isolate architecture, training data, or loss function.

2 Related Work

Representation collapse and the modality gap. Liang et al. (2022) characterized the modality gap, showing image and text embeddings occupy distinct cone-like manifolds. Lewis et al. (2024) find CLIP composes concepts in single-object settings but degrades sharply when binding is required, and Hendricks & Nematzadeh (2021) probe verb and relation understanding in vision-language transformers. Our account differs in locating the failure downstream of the representation, in the scoring function.

Benchmarks and their artifacts. Compositional failure is documented across Winoground (Thrush et al., 2022), ARO (Yuksekgonul et al., 2023), VL-CheckList (Zhao et al., 2022), CREPE (Ma et al., 2023), SugarCrepe (Hsieh et al., 2023), SugarCrepe++ (Dumpala et al., 2024), VALSE (Parcalabescu et al., 2022), ColorSwap (Burapacheep et al., 2024), and BiVLC (Miranda et al., 2024), which converts 1×2 SugarCrepe into a 2×2 benchmark by generating a synthetic hard-negative image — precisely the transformation our Proposition 1 predicts should make the benchmark structurally harder. Hsieh et al. (2023) showed ARO's hard negatives contain distributional artifacts, and Udandarao et al. (2025) showed blind heuristics rival CLIP on most benchmarks in this family. We take both findings seriously: in Section 6.4 we report a length-only heuristic that beats our own probe on two SugarCrepe categories, and we consequently decline to rest any claim on probe accuracy. Our Proposition 1 adds a structural rather than statistical explanation for why 1×2 benchmarks are easier: the shared component cancels by construction.

Recovery methods. Training-time methods fine-tune with hard negatives (Yuksekgonul et al., 2023; Doveh et al., 2023), synthetic negative image–caption pairs (Patel et al., 2024), or multi-image objectives (Peleg et al., 2025). Our analysis shows which offset cancels in 1×2 evaluation; it does not imply that hard-negative training cannot improve 2×2 scores. FSC-CLIP (Oh et al., 2024), CE-CLIP (Zhang et al., 2024), and READ-CLIP (Kwon et al., 2025) change representations through complementary objectives and require direct evaluation before drawing such a conclusion.

Concurrent geometric accounts. Several recent works converge with ours on the conclusion that CLIP's compositional failure is better explained by how matching happens than by what is encoded; we differentiate precisely. Kang et al. (2025) prove an impossibility result for embeddings built by additive superposition of object-level parts and propose dense per-patch similarity maps — a conditional result about an idealized encoding scheme, where ours is an empirical claim about trained models operating on the single pooled embedding. Koishigarina et al. (2025) show CLIP's text encoder linearly encodes attribute–object binding that cross-modal matching fails to use, and propose a fixed globally-trained linear correction; we supply the mechanism they do not (cosine implicitly weights by magnitude, Section 6) and show in Section 7 that a correction in that structural class does not transfer across compositional axes here. Miranda et al. (2026) is closest in thesis, but their recovery requires a trained transformer over frozen patch and token embeddings; ours needs no patch-level access and carries a proof (Proposition 1) of exactly when it can and cannot succeed, which none of these three provide. Alshehri et al. (2026) independently describe a "bag-of-concepts" effect in which similarity approximates mean-pooling of concept evidence, consistent with the mechanism we measure directly, and Bhattacharya (2026) reach a high Winoground Group Score using scene-graph parsing with a multimodal LLM — complementary, since it abandons the dual-encoder cosine setting whose limits we characterize.

Joint matching and evaluation format. Zhu et al. (2025) introduce GroupMatch and test-time matching. For two images and captions their matching criterion is $s_{00} + s_{11} > s_{01} + s_{10}$, exactly the Image criterion after unnormalized row centering. Our contribution connects mean-erasure to that existing criterion and explains the Group–Text identity and its normalization boundary; we do not introduce a new matching metric. VisMin (Awal et al., 2024) and ConMe (Huang et al., 2024) offer complementary structural evaluations, but are not additional completed experiments here.

Retrieval geometry and steering. Hubness correction is the standard post-hoc fix for cross-modal retrieval pathologies (Smith et al., 2017); we evaluate it directly in Section 7. Activation steering (Turner et al., 2023), the linear representation hypothesis (Park et al., 2023), and text-based decomposition of CLIP image representations (Gandelsman et al., 2024) inform our content-agnostic baselines. Unlike cross-attention architectures such as BLIP (Li et al., 2022), dual-encoder models score entirely through global cosine similarity, which is what makes them susceptible to the interference we describe.

3 Setup and Notation

Let $I \in \mathbb{R}^d$ and $C \in \mathbb{R}^d$ denote L_2-normalized *post-projection, post-pooling* image and text embeddings — the vectors a dual encoder actually scores. All operations below are vector operations in this final embedding space; none are performed on tokens, strings, or patch-level features. No model weights are modified anywhere in this paper, and no encoder is fine-tuned.

For a candidate caption pair (C_0, C_1) define

$$C_{\text{mean}} = \tfrac{1}{2}(C_0 + C_1), \qquad \Delta = \tfrac{1}{2}(C_0 - C_1), \tag{1}$$

so $C_0 = C_{\text{mean}} + \Delta$ and $C_1 = C_{\text{mean}} - \Delta$. The mean is the component shared by the two vectors and Δ their contrast. Assigning them exclusively to object identity and syntax is a semantic interpretation, not a consequence of this decomposition. *Mean-erasure* projects an image embedding orthogonally away from C_{mean}:

$$I' = I - (I \cdot \hat{C}_{\text{mean}})\, \hat{C}_{\text{mean}}, \qquad \hat{C}_{\text{mean}} = C_{\text{mean}} / \|C_{\text{mean}}\|. \tag{2}$$

Evaluation formats. A 1×2 benchmark presents one image and two captions and asks the model to rank the correct caption higher (SugarCrepe, ARO, VALSE, SugarCrepe++). A 2×2 benchmark

presents two images *and* two captions, all four sharing content (Winoground, ColorSwap), scored by **Text Score** (each image ranks its caption first), **Image Score** (each caption ranks its image first), and **Group Score** (both hold). That distinction is not cosmetic, as we now show.

4 WHY 2×2 EVALUATION IS STRUCTURALLY HARDER

Proposition 1 (1×2 invariance)**.** *Let C_0, C_1 be L_2-normalized, so $\|C_0\| = \|C_1\| = 1$. For any image I with a nonzero projection residual and $C_{mean} \neq 0$, mean-erasure (Eq. 2) leaves the ranking of $\mathrm{sim}(I, C_0)$ against $\mathrm{sim}(I, C_1)$ unchanged.*

Proof. The ranking depends on the sign of $I \cdot C_0 - I \cdot C_1$. Substituting Eq. 1, $I \cdot C_0 - I \cdot C_1 = 2\,(I \cdot \Delta)$. The L_2-normalization assumption gives $C_{\text{mean}} \cdot \Delta = \frac{1}{4}(\|C_0\|^2 - \|C_1\|^2) = 0$, so $C_{\text{mean}} \perp \Delta$. Erasure removes a component lying entirely along $\hat{C}_{\text{mean}}$, which contributes zero to $I \cdot \Delta$. The sign, and hence the ranking, is unchanged. Normalizing I' rescales both sides by the same positive factor $1/\|I'\|$ and cannot flip the sign. □

The L_2-normalization assumption is not a technicality we can drop: without it $C_{\text{mean}} \cdot \Delta = \frac{1}{4}(\|C_0\|^2 - \|C_1\|^2) \neq 0$ in general and the cancellation fails. It holds by construction in every model we evaluate, since cosine-similarity inference normalizes before scoring.

Corollary 1 (2×2 non-invariance)**.** *Winoground's Image Score compares $\mathrm{sim}(I_0, C_0)$ against $\mathrm{sim}(I_1, C_0)$. The unnormalized difference contains the term $(I_0 - I_1) \cdot C_{mean}$, which does **not** cancel, because I_0 and I_1 have different projections onto C_{mean}. When both images depict the same objects, both projections are large and similar, so this term contributes a large common-mode quantity with a small, noise-dominated difference. Erasure sets it to zero exactly, leaving the numerator proportional to $I' \cdot \Delta$. The denominators $\|I'_0\|, \|I'_1\|$ differ after projection and rescale the two sides unequally, but are strictly positive and reintroduce no C_{mean} component.*

Corollary 2 (Group Score is forced to equal Text Score)**.** *Under mean-erasure, if a pair passes Text Score then $I'_0 \cdot \Delta > 0$ and $I'_1 \cdot \Delta < 0$. Since $I' \cdot C_{mean} = 0$, the Image Score condition reduces to $\frac{I'_0 \cdot \Delta}{\|I'_0\|} > \frac{I'_1 \cdot \Delta}{\|I'_1\|}$, a strictly positive quantity compared against a strictly negative one, which holds unconditionally. Passing Text Score therefore implies passing Image Score, and Group Score is forced to equal Text Score exactly. The second Image inequality is $-I_1'^{op}\Delta/\|I'_1\| > -I_0'^{op}\Delta/\|I'_0\|$, the identical condition. Zero means and zero residuals are handled as specified in Appendix P.*

What Corollary 2 does and does not buy us. We state the consequence plainly, because it bounds our own result. Mean-erasure cannot raise Group Score above the model's *baseline* Text Score; the 31.0% we report in Section 5 is exactly CLIP ViT-B/32's unmodified Text Score, and Text Score itself does not improve at all. So mean-erasure is not a method for making a model more compositional. What it is, is an exact decomposition. For CLIP ViT-B/32 the 22.0-point gap is removed exactly by this candidate-dependent operation. Algebraically, unnormalized erasure subtracts the mean of each score row. This explains the identity without requiring a semantic label for the removed direction. Text Score is the conjunction of two within-image decisions, not the accuracy of a single 1×2 example. Sections 7 and 8 then ask what it takes to remove that interference without a foil.

5 MEAN-ERASURE CLOSES THE 2×2 GAP

5.1 WINOGROUND

Table 1 reports mean-erasure and localized controls on CLIP ViT-B/32 over all 400 Winoground pairs. Group Score rises from 9.0% to 31.0% ($+22.0$pp, bootstrap 95% CI $[18.25, 26.25]$, $p < 10^{-4}$, $N = 10^4$), driven entirely by Image Score rising from 11.25% to 64.5%, with Text Score exactly invariant — zero difference on every one of the $10{,}000$ bootstrap resamples, as Proposition 1 requires.

The controls probe candidate dependence. Single-caption erasure is excluded: it makes both image scores against that caption exactly zero, so strict Image and Group Scores must be zero; floating-point residuals can otherwise create spurious passes. For the noun control, NLTK-extracted nouns

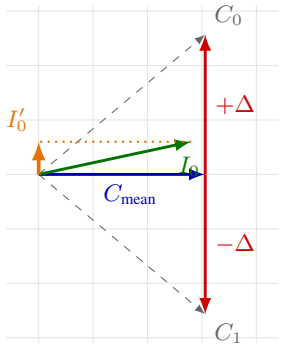


Figure 1: **Mean-erasure geometry.** C_0 and C_1 are L_2-normalized and therefore equidistant from the origin, which forces $C_{\text{mean}} \perp \Delta$ (Proposition 1). The diagram illustrates removal of the candidate-mean component from I_0. The schematic does not establish that the mean is a pure object-identity direction. Within the caption span, the residual lies along Δ; in $\mathbb{R}^d$ it can also retain a component orthogonal to both captions.

Table 1: Mean-erasure and localized controls on Winoground (CLIP ViT-B/32, $n = 400$). Matched mean-erasure changes cross-image comparisons while preserving within-image ranks. Controls have the interpretive limits described in the text.

Erasure target	**Text**	**Image**	**Group**
None (baseline)	31.0%	11.25%	9.0%
C_{mean} (matched pair)	31.0%	**64.5%**	**31.0%**
C_{mean} (random unrelated pair)	26.25%	8.5%	7.0%
Isolated nouns only	25.5%	10.25%	7.25%

from caption 0 are deduplicated, joined with "and," inserted into "A photo of a {nouns}," and encoded by CLIP's full text tower (Appendix P). This is a contextualized noun prompt, not a token average. The random-pair direction control breaks the candidate-mean identity; its failure does not by itself identify shared object semantics. A sweep over partial strength $\alpha \in [0, 1]$ shows Image and Group rise monotonically while Text stays exactly invariant at every α (Appendix B), an empirical pattern rather than a uniqueness argument. A symmetric image-side variant, projecting text embeddings away from I_{mean}, doubles Text Score with Image Score exactly invariant, as the symmetric form of Proposition 1 predicts (Appendix C).

Gains hold across all three Winoground swap tags, with object swaps gaining most in absolute terms, without implying uniform recovery of all binding types (Appendix A).

5.2 A SECOND 2×2 BENCHMARK

Winoground tests spatial arrangement. To test whether the effect is an artifact of one benchmark's construction, we ran the identical protocol — same scoring, same full control suite — on ColorSwap (Burapacheep et al., 2024), which is independently constructed and tests *color–attribute binding* rather than arrangement (300 test groups).

All five models improve with $p < 10^{-4}$ (Table 2); on Winoground the same five improve 3.0–4.2$\times$ (Appendix A). ColorSwap extends the format result to attribute binding; it does not establish universal attribute recoverability. The baseline Text Scores implied by the theorem are $35.7, 32.7, 56.3, 61.3, 67.3\%$ in table order; these inferred values are not an independent empirical check. We note that no 1×2 benchmark could serve as evidence here: Proposition 1 makes mean-erasure a provable no-op on SugarCrepe, ARO, VALSE and SugarCrepe++, so a null result there would be a theorem, not a finding.

Table 2: Mean-erasure on ColorSwap (Burapacheep et al., 2024), $n = 300$ groups, all five models. One-sided bootstrap p-values. ColorSwap tests color–object binding. These aggregate results are carried over from the submitted experiment; per-pair replication code accompanies this revision.

Model	Baseline GS	Mean-erasure GS	Rel. gain
CLIP ViT-B/32 (OpenAI)	12.0%	**35.7%**	2.97×
CLIP ViT-L/14 (OpenAI)	7.3%	**32.7%**	4.45×
CLIP ViT-B/32 (LAION)	23.3%	**56.3%**	2.41×
SigLIP-base	30.3%	**61.3%**	2.02×
SigLIP-SO400M	37.0%	**67.3%**	1.82×

6 What Cosine Similarity Actually Weights

Section 5 identifies a score-geometry bottleneck using fixed pooled features. The Group–Text identity preserves existing caption decisions; it is not independent evidence that all compositional information survives. We complement this result with a coordinate-level test of H2 using a measurement requiring no probe, no training, no seeds, and no feature standardization. An earlier version of this work instead estimated importance from MLP probe weights; Section 6.4 documents why we abandoned that, and the measurement below supersedes it under strictly weaker assumptions.

6.1 A probe-free decomposition

For L_2-normalized I, C, the element-wise interaction features $f_i = I_i C_i$ sum exactly to the score: $\sum_i f_i = \cos(I, C)$. Each dimension's contribution to the discriminative margin between a true and a false caption is therefore directly computable, with no estimation step anywhere. We pair this with each dimension's *discriminability*: the AUC of f_i used alone as a univariate match/mismatch classifier. Comparing the two answers H2 directly — does cosine spend its mass on the dimensions that actually discriminate?

On CLIP ViT-B/32, SugarCrepe swap-object ($n = 245$), it does not. The top-5% most discriminative dimensions receive only $0.484\times$ their proportional share of cosine weight (95% CI $[0.448, 0.530]$, 10^4-sample bootstrap; $P(\text{ratio} \geq 1) = 0$ across every resample). Read from the other direction: the top-20% highest-magnitude dimensions carry 57.3% of total cosine mass but only 20.2% of the discriminative signal, a $2.84\times$ concentration ratio between the two reported shares. This is an association in the evaluated basis, not an invariant assignment of coordinates to semantic roles. A second, independently written implementation of the same quantity returns 0.47, inside the first's confidence interval.

Generality across categories. Extending the top-20% measurement to all seven SugarCrepe categories, the over-representation holds in every one, ranging 2.35–$3.07\times$ ($n = 245$–1000 per category). The magnitude mechanism is therefore not specific to object swaps, even though our evidence for its behavioral consequences is strongest there. A finer per-dimension version of the measurement does flip sign for the two *addition* categories (`add_obj`, `add_att`), where the false caption is a strict superset of the true one ("broccoli and noodles" vs. "broccoli, noodles, and carrots"). That is a lexical-length asymmetry rather than a compositional swap, so cosine can solve it without resolving arrangement at all — independently locating, in our own measurements, the benchmark artifacts reported by Hsieh et al. (2023) and Udandarao et al. (2025).

6.2 Object identity, not arrangement, tracks magnitude

The decisive control holds everything fixed except the negative caption. Over identical embeddings and identical dimensions we compute the correlation between per-dimension contribution and magnitude for two tasks: *object identity* (negative = a caption from a different image) and *compositional* (negative = the same words rearranged).

Across the reported 50-condition analysis, object-identity contribution correlates with magnitude at mean $\rho = +0.78$ (range $+0.29$ to $+0.96$) while compositional contribution correlates at $\rho = -0.14$ (range -0.51 to $+0.04$). The gap is negative in 50 of 50: object-identity information is strongly

magnitude-aligned in every model, dataset, split and basis tested, and compositional information is not. This is a basis-dependent association, not a causal identification of semantic axes. The two-task design also rules out the obvious artifact: if a negative correlation arose mechanically from measuring any weak signal it would appear in both tasks, whereas under otherwise identical conditions the two point in opposite directions. A permuted-label run, where no signal exists, returns $+0.007$.

6.3 Two mechanisms, not one

It is tempting to assume the spectral pattern above *causes* the shared-content confound of Section 5. We tested this and it does not. Decomposing C_{mean} and Δ into the SVD basis for 400 Winoground pairs, spectral placement does not predict per-pair mean-erasure gain ($\rho \approx 0.05$, $p = 0.31$), even though Δ does concentrate in low-magnitude dimensions (34.8% of its energy in the bottom quintile against a 20% uniform baseline). We therefore distinguish these two geometric observations; a nonsignificant correlation does not establish statistical independence. Each rests on its own evidence; neither requires the other.

6.4 Why we no longer report probe-based evidence

We report this negative result because it changed our claims. Run across all seven SugarCrepe categories, our held-out probe is at chance for the three categories closest to genuine rearrangement (`swap_obj` 50.0%, `replace_att` 50.0%, `swap_att` 51.5%) and high only for the two addition categories (80.5%, 70.5%) — where a zero-information heuristic predicting the *shorter* caption reaches 98.8% and 99.1%, beating it outright, exactly the artifact Udandarao et al. (2025) report across 17 benchmarks. Separately, the ℓ_1-norm of first-layer MLP weights is not well-defined: a hidden unit rescales by λ in and $1/\lambda$ out with the function unchanged, and our unstandardized features scale with σ_i, biasing weights toward low-magnitude dimensions regardless of content. We therefore withdraw the probe-based claim entirely and base the revised diagnosis on the closed-form projection of Section 5 and the probe-free measurement above, neither of which fits a parameter. Appendix E gives the full per-category table, the length-heuristic control, and the seven importance measures we compared before abandoning the approach.

7 Limits of the Tested Global Corrections

We compare candidate-dependent erasure with global alternatives along two structurally different axes: fixed transformations of the *representation*, and a fixed replacement for the *scoring function*.

Fixed transformations of the representation. We evaluate SVD steering (Gandelsman et al., 2024), contrastive activation vectors (Turner et al., 2023), Mahalanobis reweighting, and two matched null controls, all on Winoground with paired bootstrap p-values (Appendix F). Every method matches baseline or degrades it; all have $p \geq 0.6$ and three have $p \geq 0.91$ (Table 3). Mahalanobis reweighting, which reweights low-singular-value directions, collapses Image Score to 2.5%: compositional signal and noise are entangled in the low-magnitude regime and equalizing amplifies both. Critically, the random null controls match the targeted methods, so these tests do not establish an advantage for the chosen directions. We add the standard published correction for this failure class — gallery-side hubness normalization via the inverted softmax (Smith et al., 2017), each image penalized against a $4{,}126$-caption bank. It is content-agnostic, precomputable, foil-free, and moves Group Score $+0.75$pp.

A fixed replacement for cosine similarity. A sharper test of H3 replaces the scoring function rather than the representation. Holding the frozen embeddings fixed, we substitute cosine with a learned bilinear scorer $s(I, C) = I^\top M C$, with M initialized at identity, trained on ColorSwap and evaluated zero-shot on Winoground. Across a weight-decay sweep the best setting beats cosine by only $+0.25$ to $+0.75$pp, and relaxing the regularizer to let M move further from identity makes it *worse* — the signature of fitting one training distribution rather than learning a transferable rule (Appendix G).

Table 3: Winoground under content-agnostic corrections (CLIP ViT-B/32). Mean-erasure is shown as the reference paired diagnostic. †evaluated on the 300-pair CAV split. ‡requires both candidate captions at evaluation time.

Intervention	Text	Image	Group	Bootstrap p
Baseline (400 pairs)	31.0%	11.3%	9.0%	—
Mean-erasure (reference)‡	**31.0%**	**64.5%**	**31.0%**	$< \mathbf{10^{-4}}$
Inverted softmax (Smith et al., 2017), $\beta = 20$	31.0%	12.5%	9.75%	0.31
SVD bottom 20% ($\alpha = 2$)	28.5%	10.3%	7.5%	0.912
Random subspace ($\alpha = 2$)	22.3%	7.8%	5.8%	0.981
Mahalanobis reweighting	20.8%	2.5%	1.8%	1.000
Baseline† (300 pairs)	32.0%	12.3%	9.3%	—
CAV ($\alpha = 5$)†	33.3%	11.3%	9.0%	0.691
Random pre-projection vector†	31.0%	12.3%	9.3%	0.604

The bilinear result measures cross-task transfer: training on color binding and evaluating on Winoground introduces domain shift. It is not an impossibility result for a fixed scorer. The supplementary code adds an in-domain split and validation-selected regularization to separate these explanations. The supported H3 conclusion is that the tested global alternatives do not reproduce the candidate-pair diagnostic; seven negative controls cannot establish universal necessity.

8 Beyond the Diagnostic: Foil-Free Recovery

Mean-erasure needs both candidate captions, which makes it a diagnostic rather than a retrieval method. But "content-adaptive" and "foil-dependent" are separable requirements, and the evaluated approximation is a first test of that distinction.

Top-K approximation: no explicit foil label. Given a query image, score it against the full caption pool with ordinary cosine, take the top-K captions, use their mean as $\hat{C}_{\text{mean}}$, project, and re-rank. Captions sharing an image's objects rank highly by cosine even from a large pool, so retrieval itself surfaces the shared content. $K = 2$ yields $9.0\% \rightarrow 14.5\%$ Group Score ($+5.5$pp; reported raw bootstrap tail 0.0012, Bonferroni-adjusted 0.006 over five K values). The effect decays monotonically toward baseline as K grows ($+1.5$pp at K=4, -0.75pp at K=32), tracing the curve Section 7 predicts: large K approaches a corpus-wide mean, the content-agnostic intervention we already showed fails. One honest trade-off: because each image's direction is computed independently rather than shared, Text Score is no longer invariant and falls 2.5–8pp at every K — a trade, not a free lunch (Appendix H).

Architectural comparison: a richer readout. LLaVA-1.5-7B (Liu et al., 2024) has a reported Winoground Group Score of 30.0% (Text 46.75%, Image 42.50%) under the yes/no likelihood protocol in Appendix I. This is a system-level comparison, not evidence about identical pooled vectors: LLaVA uses unpooled CLIP patch features, a learned connector, and an LLM. Its additional spatial information and training prevent attributing the difference to cosine alone. The fixed-pooled-feature diagnostic remains the controlled evidence for our readout claim.

The top-K pool must also be distinguished from open retrieval. If it consists of paired benchmark captions, top-2 can recover the true caption and its foil implicitly. The supplementary replication evaluates this condition alongside external distractors and foil removal, and records pair co-retrieval. The existing gain supports only the evaluated setting, not established open-gallery generalization.

9 Discussion and Limitations

The exact result establishes a format-dependent readout bottleneck (H1); the coordinate analysis is basis-dependent (H2), and candidate dependence is demonstrated for this diagnostic rather than proved necessary for every correction (H3). We now bound that conclusion.

Limitations. **(i)** Mean-erasure is a diagnostic; Corollary 2 caps its Group Score at the model's baseline Text Score, and Text Score itself never improves. The top-K variant recovers 25% of the gap to that ceiling and costs Text Score to do it. **(ii)** The geometric identity is not restricted to object-level binding; the interpretation of H2 is basis- and task-specific. For attribute and relation swaps, we do not infer absence of information from an unsuccessful probe or diagnostic, and Section 6.4 documents in full why we no longer treat probe accuracy on this benchmark family as evidence either way. **(iii)** 266 of 400 Winoground pairs (66.5%) fail under erasure in *both* modality directions (Appendix D); shared-content interference is a major but not exclusive source of compositional failure, and Appendix L categorizes three residual failure modes. **(iv)** Our five models vary simultaneously in architecture, scale, and training data, so we make no causal attribution for cross-model differences: an earlier version of this work attributed a magnitude-alignment reversal to SigLIP-SO400M's scale, and we withdraw that as unidentifiable from five uncontrolled points. We isolate the matching-function axis (Section 7) but not the objective-versus-architecture axis; both need matched-architecture training runs we have not done.

10 Conclusion

A substantial component of paired compositional retrieval failure is exposed by changing the scoring geometry while keeping the pooled representations fixed. Mean-erasure is provably inert on 1×2 benchmarks, provably forces 2×2 Group Score to equal the baseline Text Score, and closes the observed Group–Text gaps across five models and two benchmarks. This identifies a concrete readout limitation without equating gap closure with complete compositional recovery. The probe-free measurements and global-control experiments motivate further study of scoring and representations together.

ETHICS STATEMENT

This work analyzes publicly available frozen models (CLIP, SigLIP, LLaVA-1.5) on publicly available benchmarks (Winoground, ColorSwap, SugarCrepe, ARO) under licenses permitting academic research use. No human subjects, private data, or personally identifying information are involved, and no new model or dataset is released. Our analysis is diagnostic and introduces no deployable capability with foreseeable potential for misuse. We note one responsible-reporting consideration: the benchmark artifacts we document in Section 6.4 imply that reported compositional accuracies in this literature — including, previously, our own — may partly reflect lexical shortcuts rather than compositional competence, and we have revised our claims accordingly rather than retaining the stronger published numbers.

REPRODUCIBILITY STATEMENT

All models and datasets are public and used at their canonical splits. Section 3 and Eq. 2 fully specify mean-erasure, which is a closed-form projection with no fitted parameters and no stochasticity; Winoground and ColorSwap scores are therefore deterministic given a frozen model. The formal results (Proposition 1, Corollaries 1 and 2) state their assumptions — L_2-normalized caption embeddings — explicitly in Section 4, with complete proofs. Hyperparameters for all seven content-agnostic corrections are in Appendix F, the bilinear scorer's training protocol in Appendix G, the LLaVA scoring procedure in Appendix I, and the probe protocol we report but no longer rely on in Appendix E. The supplementary package provides the original scripts and reconstructed replication modules for the score audit, benchmark extraction, spectral estimands, top-K retrieval, bilinear scoring, and LLaVA likelihood evaluation. A manifest distinguishes retained experimental outputs from reconstructed protocols. It records dependencies, seeds, split rules, and commands; the provided algebra and aggregate-result checks run independently of model downloads. The reconstructed modules make the missing protocols executable and do not constitute new experimental results until run.

AI USE STATEMENT

Generative AI assisted writing, literature checks, mathematical exposition, and implementation and checking of replication code. This included identifying limits in earlier interpretations. The authors are responsible for verifying the final claims, experimental outputs, references, and supplementary artifacts.

REFERENCES

Sultan Alshehri, Zhantao Yang, Han Zhang, and Marios Savvides. Similarity is not logic: Factored inference for dual-encoder vision-language models. *arXiv preprint arXiv:2607.23052*, 2026.

Rabiul Awal, Saba Ahmadi, Le Zhang, and Aishwarya Agrawal. VisMin: Visual minimal-change understanding. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://arxiv.org/abs/2407.16772.

Amartya Bhattacharya. Inference-time structural reasoning for compositional vision-language understanding. *arXiv preprint arXiv:2603.27349*, 2026.

Jirayu Burapacheep, Ishan Gaur, Agam Bhatia, and Tristan Thrush. ColorSwap: A color and word order dataset for multimodal evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1716–1726, 2024.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2818–2829, 2023.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Eli Schwartz, Roei Herzig, Raja Giryes, Rogerio Feris, Rameswar Panda, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision & language concepts to vision & language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2657–2668, 2023.

Sri Harsha Dumpala, Aman Jaiswal, Chandramouli Sastry, Evangelos Milios, Sageev Oore, and Hassan Sajjad. SUGARCREPE++ dataset: Vision-language model sensitivity to semantic and lexical alterations. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, 2024.

Yossi Gandelsman, Alexei A. Efros, and Jacob Steinhardt. Interpreting CLIP's image representation via text-based decomposition. In *International Conference on Learning Representations (ICLR)*, 2024.

Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3635–3644, 2021.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. SugarCrepe: Fixing hackable benchmarks for vision-language compositionality. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023.

Irene Huang, Wei Lin, M. Jehanzeb Mirza, Jacob A. Hansen, Sivan Doveh, Victor Ion Butoi, Roei Herzig, Assaf Arbelle, Hilde Kuehne, Trevor Darrell, Chuang Gan, Aude Oliva, Rogerio Feris, and Leonid Karlinsky. ConMe: Rethinking evaluation of compositional reasoning for modern VLMs. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL `https://arxiv.org/abs/2406.08164`.

Raphi Kang, Yue Song, Georgia Gkioxari, and Pietro Perona. Is CLIP ideal? no. can we fix it? yes! *arXiv preprint arXiv:2503.08723*, 2025.

Darina Koishigarina, Arnas Uselis, and Seong Joon Oh. CLIP behaves like a bag-of-words model cross-modally but not uni-modally. *arXiv preprint arXiv:2502.03566*, 2025.

Jihoon Kwon, Kyle Min, and Jy-yong Sohn. Enhancing compositional reasoning in CLIP via reconstruction and alignment of text descriptions. *arXiv preprint arXiv:2510.16540*, 2025. URL `https://arxiv.org/abs/2510.16540`.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Jack Merullo, Qinan Yu, Stephen H. Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pp. 1487–1500, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, pp. 12888–12900, 2022.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26296–26306, 2024.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10910–10921, 2023.

Imanol Miranda, Ander Salaberria, Eneko Agirre, and Gorka Azkune. BiVLC: Extending vision-language compositionality evaluation with text-to-image retrieval. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 37, 2024.

Imanol Miranda et al. Revisiting compositionality in dual-encoder vision-language models: The role of inference. *arXiv preprint arXiv:2604.11496*, 2026.

Youngtaek Oh, Jae Won Cho, Dong-Jin Kim, In So Kweon, and Junmo Kim. Preserving multi-modal capabilities of pre-trained VLMs for improving vision-linguistic compositionality. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 19060–19076, 2024. URL https://aclanthology.org/2024.emnlp-main.1062/.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 8253–8280, 2022.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. *arXiv preprint arXiv:2311.03658*, 2023.

Maitreya Patel, Abhiram Kusumba, Sheng Cheng, Changhoon Kim, Tejas Gokhale, Chitta Baral, and Yezhou Yang. TripletCLIP: Improving compositional reasoning of CLIP via synthetic vision-language negatives. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024.

Amit Peleg, Naman Deep Singh, and Matthias Hein. Advancing compositional awareness in CLIP with efficient fine-tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, 2025.

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

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations (ICLR)*, 2017.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5238–5248, 2022.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*, 2023.

Vishaal Udandarao, Mehdi Cherti, Shyamgopal Karthik, Jenia Jitsev, Samuel Albanie, and Matthias Bethge. A good CREPE needs more than just sugar: Investigating biases in compositional vision-language benchmarks. *arXiv preprint arXiv:2506.08227*, 2025.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations (ICLR)*, 2023.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11975–11986, 2023.

Le Zhang, Rabiul Awal, and Aishwarya Agrawal. Contrasting intra-modal and ranking cross-modal hard negatives to enhance visio-linguistic compositional understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://arxiv.org/abs/2306.08832.

Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. VL-CheckList: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022.

Yinglun Zhu, Jiancheng Zhang, and Fuzhi Tang. Test-time matching: Unlocking compositional reasoning in multimodal models. *arXiv preprint arXiv:2510.07632*, 2025. URL https://arxiv.org/abs/2510.07632.

A Cross-Model Replication on Winoground

Table 5 extends the Winoground result of Section 5.1 to all five contrastive models. All improve by 3.0–4.2×, showing the effect across these checkpoints without isolating scale, data, or training loss. SigLIP-SO400M's higher baseline (12.25% against CLIP ViT-B/32's 9.0%) does not translate into a smaller relative gain, indicating that whatever advantage the larger model holds does not resolve the interference.

Table 4 breaks the CLIP ViT-B/32 result down by Winoground swap tag. Improvement is consistent across all three; object swaps gain most in absolute terms, consistent with the object-level scoping of H1. The both-swap cell has only 26 pairs and should be read with care.

Table 4: Per-category Winoground Group Score before and after mean-erasure (`clip-vit-base-patch32`).

Swap tag	n	Baseline	Mean-erasure
Relation	233	6.0%	**22.7%**
Object	141	7.8%	**35.5%**
Both	26	42.3%	**80.8%**
All	400	9.0%	**31.0%**

Table 5: Mean-erasure across five contrastive models on Winoground ($n = 400$). Per Corollary 2, each post-erasure Group Score equals that model's own baseline Text Score exactly.

Model	Baseline GS	After erasure	Rel. gain
CLIP ViT-B/32 (OpenAI) (Radford et al., 2021)	9.00%	**31.00%**	3.4×
CLIP ViT-B/32 (LAION) (Cherti et al., 2023)	8.25%	**35.00%**	4.2×
CLIP ViT-L/14 (OpenAI) (Radford et al., 2021)	7.25%	**28.25%**	3.9×
SigLIP-base (Zhai et al., 2023)	10.25%	**33.00%**	3.2×
SigLIP-SO400M (Zhai et al., 2023)	12.25%	**37.25%**	3.0×

B Sensitivity to Erasure Strength

We sweep a partial-erasure strength $\alpha \in [0, 1]$ with $I'_\alpha = I - \alpha(I \cdot \hat{C}_{\text{mean}})\hat{C}_{\text{mean}}$. Image and Group Scores rise monotonically while Text Score is exactly invariant at every α — a joint pattern stochastic perturbation cannot produce, and a confirmation that Proposition 1 holds at every erasure strength rather than only at $\alpha = 1$.

Table 6: Winoground scores versus partial-erasure strength α (`clip-vit-base-patch32`). $\alpha = 0$ is baseline; $\alpha = 1$ is full mean-erasure.

α	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Text	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0	31.0
Image	11.3	11.8	12.8	13.8	15.8	18.0	21.3	26.0	30.8	42.0	**64.5**
Group	9.0	9.5	10.3	11.3	12.8	13.8	16.8	18.5	21.0	27.8	**31.0**

C Symmetric Erasure: Bidirectional Evidence

Define $I_{\text{mean}} = \frac{1}{2}(I_0 + I_1)$ and project each text embedding away from it:

$$T' = T - (T \cdot \hat{I}_{\text{mean}})\,\hat{I}_{\text{mean}}, \qquad \hat{I}_{\text{mean}} = I_{\text{mean}}/\|I_{\text{mean}}\|. \tag{3}$$

The symmetry is a structured algebraic prediction, not a coincidence. Under text-side erasure, image embeddings are modified while caption embeddings are not, so the caption ranking for a fixed image

Table 7: Symmetric erasure across five models on Winoground. Text-side erasure roughly quintuples Image Score; image-side erasure roughly doubles Text Score. In each case the other score is exactly at baseline.

	Baseline		Text-side erasure		Image-side erasure	
Model	**Text**	**Image**	**Text**	**Image**	**Text**	**Image**
CLIP ViT-B/32 (OpenAI)	31.0%	11.3%	31.0%	**64.5%**	**64.5%**	11.3%
CLIP ViT-B/32 (LAION)	35.0%	11.5%	35.0%	**66.8%**	**66.8%**	11.5%
CLIP ViT-L/14 (OpenAI)	28.5%	11.3%	28.5%	**65.5%**	**65.5%**	11.3%
SigLIP-base	32.8%	12.8%	32.8%	**67.0%**	**67.0%**	12.8%
SigLIP-SO400M	37.3%	16.5%	37.3%	**70.5%**	**70.5%**	16.5%

(Text Score) cannot change. Under image-side erasure the roles swap and the image ranking for a fixed caption (Image Score) cannot change by identical logic. Only the cross-modality comparisons — those involving two differently modified embeddings — move, as Corollary 1 describes. The prediction is confirmed exactly on five independent models.

D Per-Pair Overlap of Text-Side and Image-Side Erasure

Table 8: Per-pair pass-set overlap between the two erasure directions (`clip-vit-base-patch32`, $n = 400$).

Set	**n**	**%**
Baseline passes	36	9.0%
Text-side erasure passes	125	31.3%
Image-side erasure passes	45	11.2%
Both sides pass	36	9.0%
Only text-side passes	89	22.3%
Only image-side passes	9	2.2%
Neither passes (residual)	266	66.5%

Three observations. First, the "both sides pass" set coincides exactly with the baseline-pass set: no pair is recovered jointly by both directions that was not already passing. Second, the asymmetry is large — text-side erasure adds 89 unique recoveries against image-side's 9 — and is locked in by Corollary 2: text-side erasure forces Group Score to equal Text Score (31.3%) while image-side forces it to equal the baseline Image Score (11.2%), and CLIP's baseline Image Score is strictly lower. Third, the 266-pair joint residual (66.5%) is the genuine representation gap: no instance-specific shared-content removal on either modality recovers these pairs. This is the quantitative form of limitation (iii) in Section 9.

E Probe Protocol and the Length-Bias Confound

We report this appendix in full because it documents a claim we withdrew. An earlier version of this work reported 80.3% held-out probe accuracy on SugarCrepe as evidence for H1. That figure was produced by a script that sampled from an unfiltered SugarCrepe pool rather than isolating the `swap_obj` split; because the unfiltered pool is dominated by `add_obj` and `add_att`, the aggregate was inflated. Table 9 gives the corrected per-category breakdown.

Protocol. Pair-wise 80/20 held-out split (both members of a pair assigned to the same side, so no pair straddles the split), random seed 42, `MLPClassifier(hidden_layer_sizes=(512,256), max_iter=1000, early_stopping=True)` with Adam defaults, on element-wise interaction features. Accuracy is reported on held-out test pairs only. The same configuration is used for every category.

Table 9: Held-out probe accuracy across all seven SugarCrepe categories (CLIP ViT-B/32), against a zero-information heuristic that predicts the *shorter* caption is the true one — no image, no model, no embeddings. "Ties" counts pairs where both captions have equal token length, on which the heuristic cannot fire.

Category	n	Length heuristic	Held-out probe	Ties
add_obj	1000	**98.8%**	80.5%	20/1000
add_att	692	**99.1%**	70.5%	8/692
replace_rel	1000	54.2%	61.0%	505/1000
replace_obj	1000	55.0%	58.2%	775/1000
swap_att	666	51.1%	51.5%	569/666
replace_att	788	51.0%	50.0%	660/788
swap_obj	245	52.4%	50.0%	221/245

Two conclusions follow, both of which narrow our claims. **(i)** On the two categories where the probe appears strong, a zero-information heuristic beats it outright. Probe accuracy there reflects a lexical-length asymmetry in benchmark construction, exactly as Udandarao et al. (2025) report across 17 benchmarks, and is not evidence of compositional survival. **(ii)** On the three categories closest to genuine compositional rearrangement, the probe is at chance. These probes do not establish general recovery or absence of attribute and relational information; ColorSwap supplies a separate color-binding result.

Why the weight-norm importance measure was also unsound. Independently of the data bug, the ℓ_1-norm of first-layer MLP weights is not a well-defined quantity. Scaling a hidden unit's incoming weights by λ and its outgoing weights by $1/\lambda$ leaves the learned function exactly unchanged while moving the reported value, and we verified the reported statistic can be driven across both signs under such rescalings. It was at least stable under random rescalings (0 sign flips in 600 trials), so the published value was not an initialization artifact — but stability is not well-definedness. Compounding this, the interaction features have scale proportional to σ_i and were not standardized, so the network assigns larger weights to low-magnitude dimensions irrespective of content; retraining the original pipeline on randomly permuted labels, where no signal exists, returns values comparable to those originally reported. For completeness we record the full battery of importance measures we ran on SigLIP-base/ARO before abandoning the approach: ℓ_1 weight norm $\rho = -0.320$; ℓ_2 weight norm -0.322; linear probe on the same features $+0.131$; feature ablation -0.011; permutation importance -0.011; gradient$\times$input -0.032; direct cosine contribution -0.139. Ablation and permutation importance are near zero by construction here: with several hundred correlated dimensions, removing or shuffling any single one leaves accuracy essentially unchanged, so neither is informative at single-dimension granularity. The disagreement among these measures is itself the reason we moved to the probe-free decomposition of Section 6, which has no free parameters to disagree about.

F Content-Agnostic Intervention Details

SVD steering (Gandelsman et al., 2024). Project the post-projection image embedding into the SVD basis of the visual projection matrix, $\mathbf{v}' = \mathbf{v}U$, multiplicatively boost the bottom 20% of singular-value dimensions (102 of 512 for ViT-B/32) by $1 + \alpha$ with $\alpha = 2$, and project back via $U^\top$. The same amplification is applied to every image regardless of content.

Contrastive activation vectors (Turner et al., 2023). Compute $\mathbf{d}_{\text{CAV}} = \frac{1}{100}\sum_{k=1}^{100}(I_0^{(k)} - I_1^{(k)})$ over a held-out set of 100 Winoground pairs (the remaining 300 are used for evaluation, which is why CAV rows in Table 3 carry their own baseline), unit-normalize, and add $\alpha = 5$ times this direction to every test image's *pre*-projection embedding before applying W.

The 0/1 pair orientations are not aligned to a common semantic axis, so the CAV may cancel; its null result is not evidence against semantically supervised directions.

Mahalanobis reweighting. In the SVD basis, set $v'_i \leftarrow v'_i/(\sigma_i + \epsilon)$ with $\epsilon = 10^{-3}$, reweighting dimensions according to singular value; contributions still depend on the actual embeddings.

Inverted softmax (Smith et al., 2017). Gallery-side hubness normalization. Each image I receives a penalty computed against a query bank of 4,126 captions drawn from the Winoground caption pool and the COCO validation captions, and scores become $s_\beta(I, C) = \exp(\beta \cos(I, C))/\sum_{C' \in \text{bank}} \exp(\beta \cos(I, C'))$. We sweep $\beta \in \{1, 5, 10, 20, 50\}$ and report the best. The penalty depends only on the gallery, is precomputable, requires no foil, and is fully content-agnostic — which is exactly why we include it as the strongest published member of this class.

Random subspace control (post-projection). Sample a random orthogonal subspace of the same dimensionality as the bottom-20% SVD band (102 dimensions) and boost it additively by the *same* $\alpha = 2$ used for SVD steering, so the two are compared on equal footing: $\mathbf{v}_{\text{steered}} = \mathbf{v} + \alpha P_{\text{rand}}\mathbf{v}$. Averaged over 10 random draws.

Random pre-projection vector control. Sample $\mathbf{r} \sim \mathcal{N}(0, I)$, unit-normalize, and add $\alpha = 5$ times it to every pre-projection embedding — the matched null to CAV at identical magnitude. Averaged over 10 random draws.

G Learned Bilinear Scorer

To test whether a fixed replacement for the *scoring function* can substitute for content-adaptive erasure, we hold the frozen embeddings fixed and replace cosine with $s(I, C) = I^\top MC$, where $M \in \mathbb{R}^{d \times d}$ is initialized at the identity (so training begins exactly at cosine similarity). We train on ColorSwap with a pairwise ranking loss and evaluate zero-shot on Winoground, sweeping weight decay over $\{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ to control how far M may travel from identity.

The best setting yields $+0.25$ to $+0.75$pp Group Score over cosine. Critically, the relationship with weight decay is monotone in the wrong direction: weaker regularization, which permits M to move further from identity and fit the training distribution better, produces *worse* zero-shot Winoground performance. This is consistent with domain shift and does not prove that an in-domain fixed scorer must fail. The reconstructed replication includes an in-domain validation protocol.

H Foil-Free Top-K Approximation

Procedure. (1) Score the query image against the full caption pool with standard cosine. (2) Take the top-K captions. (3) Set C_{mean} to their mean. (4) Project the image away from it and re-rank. No explicit foil label is used. A paired-caption gallery can nevertheless include and retrieve the foil implicitly; the replication records pool composition and pair co-retrieval.

Table 10: Foil-free top-K erasure on Winoground (`clip-vit-base-patch32`). p-values are one-sided bootstrap, $N = 10^4$.

K	**Group Score**	**Δ vs. baseline**	**p**
baseline	9.00%	—	—
2	**14.50%**	+5.50pp	0.0012
4	10.50%	+1.50pp	0.182
8	9.25%	+0.25pp	0.469
16	8.75%	−0.25pp	0.611
32	8.25%	−0.75pp	0.749

$K = 2$ survives Bonferroni correction across the five tested values ($0.0012 \times 5 = 0.006 < 0.05$). The reported K sweep is exploratory; future selection must use validation pairs, including under gallery perturbation. We report the full sweep specifically to show the predicted decay — as K

grows the estimate approaches a corpus-wide mean, which is precisely the content-agnostic intervention Section 7 shows fails. The curve traced here and the failure there are the same phenomenon measured two ways.

In absolute terms 14.5% closes 25% of the distance to the diagnostic's 31.0% ceiling ($+5.5$pp of $+22.0$pp available). One trade-off we state plainly: because each image's erasure direction is computed independently rather than shared across the pair, Text Score is no longer invariant and decreases by 2.5–8pp at every K. Group Score improves at small K because it requires both conditions to pass on the same pair and the Image Score gain outweighs the Text Score loss, but this is a genuine trade, not a free improvement.

I LLaVA Evaluation Protocol

LLaVA-1.5-7B (Liu et al., 2024) pairs a frozen CLIP ViT-L/14 visual encoder — a CLIP-family backbone with a different feature extraction path from the pooled dual-encoder setting — with Vicuna-7B through a two-layer MLP connector. No cosine similarity appears anywhere in its scoring path.

For each of the four (image, caption) combinations in a Winoground group, we present the image with the prompt `"Does this image match the following description? {caption} Answer yes or no."` and record the model's token-level log-probability of the `"Yes"` token, normalized against `"No"`. We use log-probabilities rather than greedy generation so that scores are continuous and comparable, and so that the comparison logic — which pair of scores must exceed which — is bit-for-bit the same function used for the cosine baseline throughout the paper. The model sees one image and one caption at a time and is never given the foil.

This is a comparison of systems, not a controlled comparison of readouts from identical pooled features. LLaVA has substantially more parameters and more training than a dual encoder, so it should not be read as a controlled comparison: LLaVA uses dense patch features rather than the pooled projected embedding used by the cosine system. Its result therefore does not establish H1 for pooled vectors.

J Winoground Example Pairs

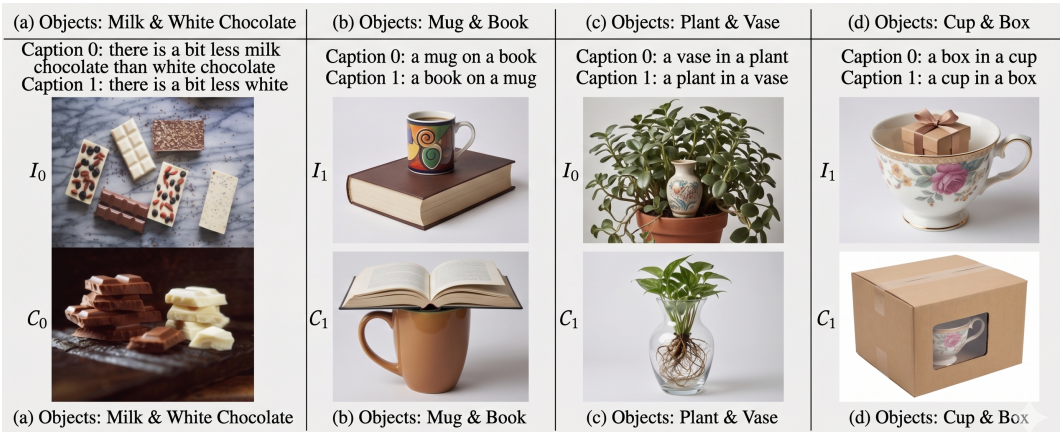


Figure 2: Representative Winoground pairs. Each group (a)–(d) shows two images containing the same objects in different spatial arrangements, paired with two captions using identical words in different syntactic order. The caption-pair mean is the shared vector component used in the diagnostic.

K Qualitative Examples: Successful Recovery

Figures 3–5 illustrate successful recoveries: the caption decisions already pass and mean-erasure also makes the image comparisons pass. These are illustrative panels from the original run.

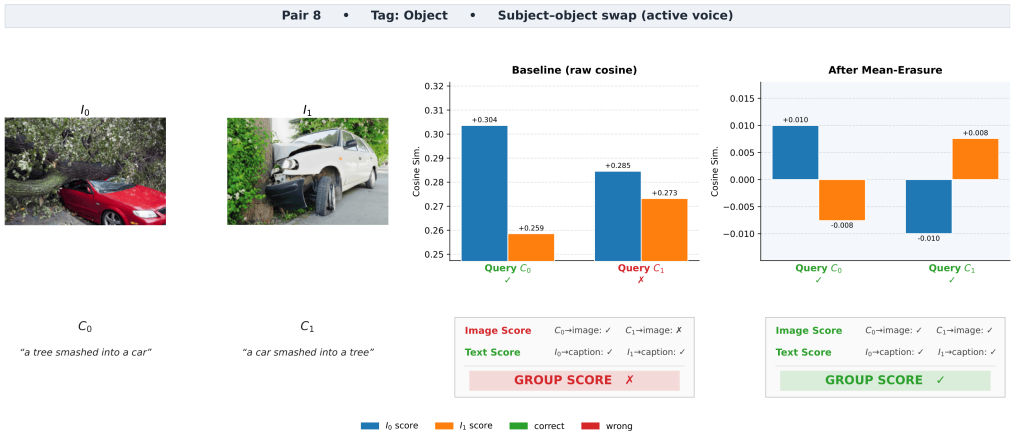


Figure 3: **Object swap (Pair 8).** "A tree smashed into a car" vs. "a car smashed into a tree" share the bag-of-words *{tree, car, smashed}*, which dominates both image embeddings under raw cosine similarity (left). After mean-erasure (right) the residual aligns with Δ and both Image and Text Score pass.

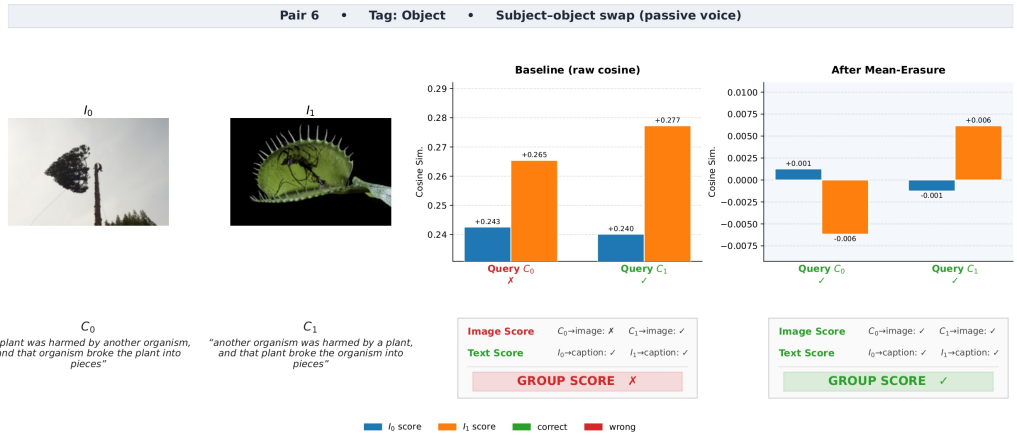


Figure 4: **Passive-voice Object swap (Pair 6).** "A plant was harmed by another organism" vs. "another organism was harmed by a plant." Baseline cosine confuses the two queries because both images share the same bag-of-words content. After mean-erasure, both Image and Text Score pass.

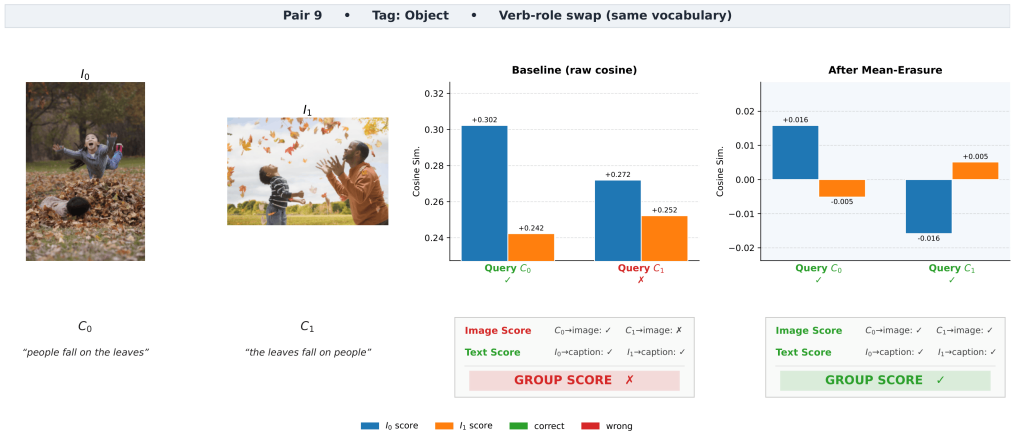


Figure 5: **Verb-role swap (Pair 9).** "People fall on the leaves" vs. "the leaves fall on people" reuse identical vocabulary while swapping the agent of the verb.

L QUALITATIVE EXAMPLES: RESIDUAL FAILURE MODES

The 266 pairs unsolved after erasure in both directions (Table 8) fall into distinct modes. Figures 6–8 show one representative pair per mode. Scores alone do not localize the residual failure to one encoder.

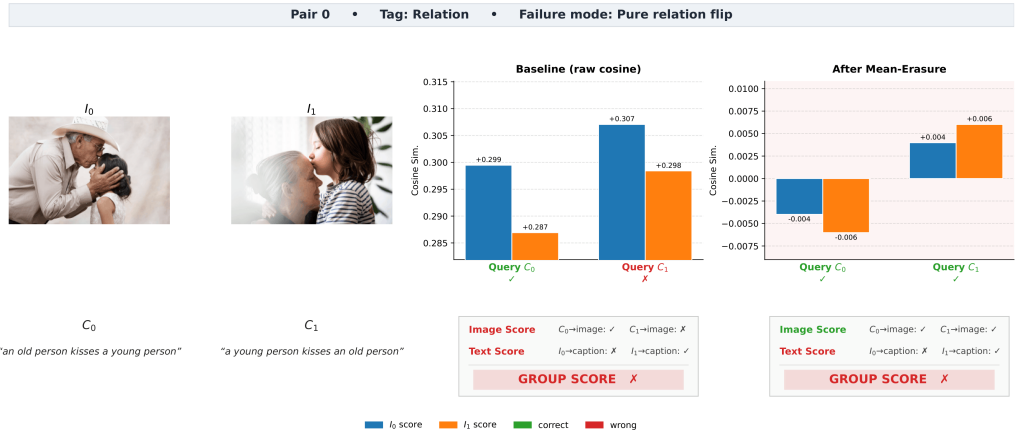


Figure 6: **Pure relation flip (Pair 0, Tag: Relation).** "An old person kisses a young person" vs. "a young person kisses an old person" differ only in the order of two noun phrases. Mean-erasure removes the shared component correctly, but the visual residual does not preserve the role assignment with enough magnitude to recover Group Score.

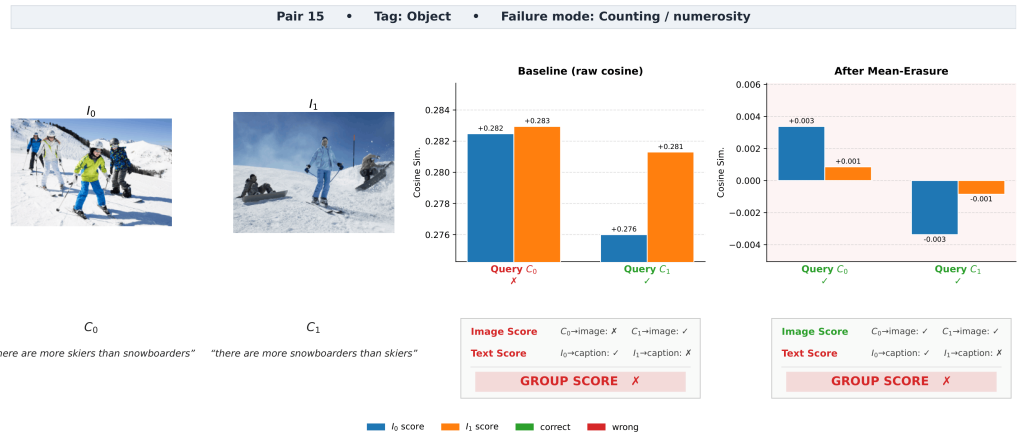


Figure 7: **Counting / numerosity (Pair 15, Tag: Object).** "There are more skiers than snowboarders" vs. the reverse. Mean-erasure recovers Image Score — the visual encoder does discriminate the two scenes — but Text Score still fails: CLIP's text encoder produces no stable "more vs. fewer" axis in the joint space.

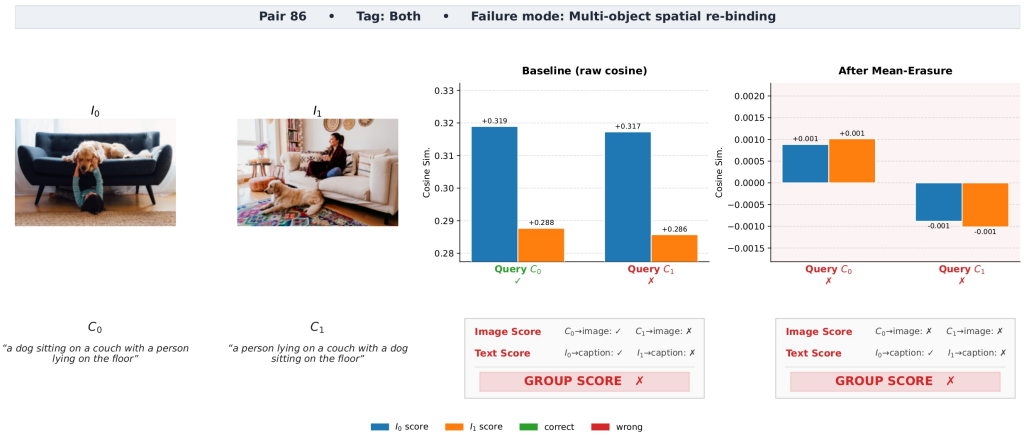


Figure 8: **Multi-object spatial re-binding (Pair 86, Tag: Both).** "A dog sitting on a couch with a person lying on the floor" vs. the swapped assignment. After erasure both image residuals project onto the same side of Δ with similar magnitude. This residual Group failure is not explained by the candidate-mean correction alone.

M SCORE-LEVEL AUDIT OF THE MAIN RESULT

For $s_{ij} = I_i^\top C_j$, write $a_i = (s_{i0} + s_{i1})/2$, $b_i = (s_{i0} - s_{i1})/2$, $m = b_0 - b_1$, and $o = |a_0 - a_1|$. Text passes iff $b_0 > 0 > b_1$; Image passes iff $m > o$. Mean-erasure before normalization gives $s'_{ij} = s_{ij} - a_i$ exactly. After normalization its two Image inequalities both reduce to $b_0/r_0 > b_1/r_1$, where r_i is the residual norm. This supplies both directions of the main proof. Group equals baseline Text in either variant, although Image outcomes may differ.

Table 11: Retained Winoground audit, $n = 400$. T, V, G are baseline percentages; J is row-centered Image (two-way GroupMatch); V_{ME} is normalized mean-erased Image. F is the percentage of baseline Image failures with $0 < m \leq o$.

Model	T	V	G	J	V_{ME}	F
CLIP ViT-B/32 (OpenAI)	31.00	11.25	9.00	64.50	64.50	60.0
CLIP ViT-B/32 (LAION)	35.00	11.25	8.25	66.25	66.75	62.0
CLIP ViT-L/14 (OpenAI)	28.25	10.50	7.25	65.50	65.50	61.5
SigLIP-base	33.00	13.00	10.25	67.25	67.00	62.4
SigLIP-SO400M	37.25	16.50	12.25	70.00	70.25	64.1

For CLIP ViT-B/32, $213/355 = 60.0\%$ of Image failures have $0 < m \leq o$. Its $258/400$ joint matches give a 95% Wilson interval $[59.69, 69.03]\%$ and an exact one-sided binomial probability 3.53×10^{-9} under a $1/2$ null. The $88/400$ additional Group passes have Wilson interval $[18.22, 26.32]$ percentage points. These are newly computed statistics of the retained run, not new encoder experiments. Joint matching has 50% chance performance versus 25% for independent Image retrieval, so the increase is not a comparable multiple of compositional ability.

N PROOF DETAILS AND REFERENCE EXAMPLES

Chance rates are tied to a null model. For i.i.d. continuous scores, the two within-row comparisons involve disjoint pairs of random variables, so Text passes with probability $1/4$. The same argument applies to raw Image Score. Group passes exactly when the two diagonal scores are the largest two of the four, giving probability $1/\binom{4}{2} = 1/6$. Centering preserves Text, makes Group equal Text, and makes Image equivalent to a comparison of two exchangeable sums. Thus centered (T, V, G) chance rates are $(1/4, 1/2, 1/4)$ when assignment-sum ties have probability zero. These are not universal null rates for arbitrarily correlated image–caption scores. In particular, shuffling both assignment labels of a fixed pair does not make its four scores independent.

Text and joint matching are different. For $S = \left(\begin{smallmatrix} 0.4 & 0.2 \\ 0.3 & 0.2 \end{smallmatrix}\right)$, both images prefer caption 0, so Text fails. Yet the diagonal sum 0.6 exceeds the off-diagonal sum 0.5, so joint matching passes. This example prevents the mistaken interpretation that every positive joint margin corresponds to two correct 1×2 decisions.

Renormalization can change the Image outcome. Let $C_0 = (\sqrt{3}/2, 1/2, 0)$, $C_1 = (\sqrt{3}/2, -1/2, 0)$, $I_0 = (0, 0.4, \sqrt{0.84})$ and $I_1 = (\sqrt{0.96}, 0.2, 0)$. All four vectors have unit length. The caption mean lies along the first axis, so $b_0 = 0.2$, $b_1 = 0.1$, $r_0 = 1$, and $r_1 = 0.2$. Row-centered Image passes ($0.2 > 0.1$), but normalized mean-erased Image fails ($0.2 < 0.5$). Text and Group fail in both versions, as the theorem requires. Thus the distinction is geometric and not merely a floating-point artifact.

Partial erasure. With $I_i^{(\alpha)} = I_i - \alpha(I_i^\top \hat{C})\hat{C}$, the unnormalized scores are $(1 - \alpha)a_i \pm b_i$. For $0 \leq \alpha \leq 1$, Image passes iff $m > (1 - \alpha)o$. Hence unnormalized Image and Group pass sets grow monotonically, while Text is invariant. Renormalization changes row scales and removes that simple monotonicity guarantee for Image Score. A monotone empirical curve under cosine normalization should be reported as an observation rather than as a consequence of this unnormalized result.

A symmetric score decomposition. Writing $u_i = (-1)^i$ and $v_j = (-1)^j$, every score matrix has the unique form

$$s_{ij} = \mu + u_i r + v_j c + u_i v_j h,$$

where $\mu = (s_{00}+s_{01}+s_{10}+s_{11})/4$, $r = (s_{00}+s_{01}-s_{10}-s_{11})/4$, $c = (s_{00}-s_{01}+s_{10}-s_{11})/4$, and $h = (s_{00}-s_{01}-s_{10}+s_{11})/4$. Then $T \iff h > |c|$, $V \iff h > |r|$, $G \iff h > \max(|r|, |c|)$, and $J \iff h > 0$. Row centering removes μ, r; column centering removes μ, c; double centering retains only the interaction term $u_i v_j h$. Consequently, column-centered Group equals raw Image, and double-centered Group equals joint matching. These are algebraic restatements of different evaluation criteria, not three independent demonstrations of representational recovery.

O Beyond Two Candidates

Subtracting one constant per row preserves within-row rankings for any matrix size. The identity between centered Group and raw Text is specific to two columns and two rows. For example,

$$S = \begin{pmatrix} 2 & 1 & 0 \\ 99 & 100 & -100 \\ 0 & 0 & 1 \end{pmatrix}$$

has each diagonal entry as its row maximum. After row centering, the first column is $(1, 66, -1/3)^\top$, so the first caption retrieves the second image. Group therefore fails despite Text passing.

For arbitrary numbers of candidates, joint assignment sums are invariant to additive row and column offsets: every bijection contains each row and column once, so all assignments receive the same added constant. This property underlies the distinction between matching and independent retrieval, but does not imply that general galleries admit one-to-one ground truth or that centering improves their retrieval accuracy.

P Evaluation Protocol and Evidence Coverage

The five identifiers used by the retained Winoground script are:

Model	Checkpoint identifier
CLIP ViT-B/32 (OpenAI)	`openai/clip-vit-base-patch32`
CLIP ViT-B/32 (LAION)	`laion/CLIP-ViT-B-32-laion2B-s34B-b79K`
CLIP ViT-L/14 (OpenAI)	`openai/clip-vit-large-patch14`
SigLIP-base	`google/siglip-base-patch16-224`
SigLIP-SO400M	`google/siglip-so400m-patch14-384`

Images are decoded as RGB and passed directly to the checkpoint processor. The saved run uses the `haideraltahan/wds_winoground` mirror, float32 feature extraction, normalized image/text vectors, and seed 0 for caption-pair permutation. SigLIP uses fixed-length text padding. This run is associated with the supplied calibration script; original model/dataset revision hashes and per-pair scores are absent from its aggregate JSON. A rerun should record both and compare against the official dataset distribution. The portable audit supplied with this revision records such metadata and saves scores and embeddings for further checks.

What can be checked without new encoder runs? The retained aggregate output supports Table 11, the normalization differences, the reported medians, and the failure-partition counts. The analysis script reconstructs integer counts at $n = 400$, generates the table, recomputes Wilson intervals and exact binomial probabilities, and verifies the projection identities on randomly generated unit embeddings. Those numerical checks support implementation consistency and do not substitute for the proofs or a fresh dataset run.

The original supplementary archive contains code for probes, SVD analyses, fixed interventions, noun erasure, erasure-strength sweeps, and symmetric erasure. ColorSwap, top-K, and LLaVA were added in later manuscript versions. Their reported values should be paired with corresponding run artifacts before making stronger reproduction claims. The manuscript retains the reported experiments; the replication manifest identifies which outputs still need to be regenerated. The code

package also provides a score-matrix interface so independent model or benchmark evaluations can use precisely the same comparisons.

Ties and numerical stability. All reported historical scores use strict $>$. For new audits, compute score arithmetic in float64 from the extracted vectors and additionally report a tolerance sweep, declaring a comparison positive only when its difference exceeds τ. Exact erasure-induced zero columns should be assigned zero analytically. Record minimum projection norms and the count of degenerate residuals. A nonzero tolerance can change measured equality near the boundary, so the exact-arithmetic theorem and a numerical tolerance audit must be reported separately.

Q Interpretation of Earlier Controls

Noun-only control. The supplied procedure tokenizes caption 0 with NLTK and retains tags beginning with `NN`. It deduplicates with `set`, joins using " and ", substitutes "objects" if no nouns are found, and forms `A photo of a {noun_string}`. It runs that entire string through CLIP's text tower and projection, normalizes the pooled vector, then projects and renormalizes both images. This procedure does not create an uncontextualized token average. Because hash-dependent set ordering can change the prompt, a reproducible rerun should save the generated strings or deterministically preserve first occurrence. Changing the ordering policy requires rerunning the control; the old value cannot be assumed unchanged.

CAV direction and sample denominators. The original CAV uses $d \propto \sum_k (I_0^{(k)} - I_1^{(k)})$ over 100 pairs without a documented semantic alignment of the 0/1 assignments. Opposite orientations can cancel. Its null result characterizes that construction, not supervised concept directions generally. Further, percentages on a 300-pair evaluation must lie on multiples of $100/300$ unless averaged over runs. Several earlier quarter-point entries do not satisfy that constraint; we do not retain that table without count-level verification.

Bilinear scoring and domain shift. A fixed score $s(I, C) = I^\top MC$ trained on ColorSwap and transferred to Winoground changes both the scoring function and the train/test distribution. Weak transfer cannot establish impossibility of a global correction. A suitable control fits M on training pairs, selects regularization on validation pairs, and evaluates unseen pairs within the same task, with a separately reported cross-task transfer test. No labels from the evaluation pairs may enter training or hyperparameter choice.

Corpus normalization. For a fixed reference bank, an image-dependent normalizer may remove a component correlated with a_i, but it need not equal the mean against the current candidate pair. Different terms need not be equally effective. Conversely, the dependence of a_i on candidates does not prohibit a learned fixed scorer from improving the margins or reducing relevant offsets on the data distribution. Finite negative controls do not justify a universal necessity claim.

R Feature-Importance Analysis: Explicit Estimands

To make a future audit unambiguous, let x_n, c_n^+, c_n^- be normalized embeddings of an image, its true caption and its foil. Fix and report a common orthonormal basis Q for both modalities. Define

$$
\begin{aligned}
f_{nk}^{\pm} &= (Q^\top x_n)_k (Q^\top c_n^{\pm})_k, \\
M_k &= \frac{1}{2N} \sum_n (|f_{nk}^+| + |f_{nk}^-|), \\
A_k &= \mathrm{AUC}(\{f_{nk}^+\}_n, \{f_{nk}^-\}_n), \\
D_k &= |A_k - 1/2|.
\end{aligned}
$$

For a coordinate subset H, an absolute-mass share is $\sum_{k\in H} M_k / \sum_k M_k$, while an AUC-based discriminability share is $\sum_{k\in H} D_k / \sum_k D_k$. These are different estimands from the signed margin share based on $\sum_n (f_{nk}^+ - f_{nk}^-)$. None is an invariant fraction of semantic information. If the denominator vanishes, the share is undefined, not zero.

These definitions specify the reconstructed audit; they do not establish the exact estimator used for the earlier reported 57.3% mass and 20.2% signal shares. In particular, raw cosine mass cannot be substituted for absolute mass, and AUC excess cannot be silently substituted for a margin-based numerator. The earlier "50 conditions" description also does not identify the two magnitude definitions: $5 \times 2 \times 4 \times 2 = 80$, whereas $5 \times (1 + 4) \times 2 = 50$. Recovering the condition manifest is necessary to choose between those designs.

Selection must be separated from evaluation. Select high-mass or discriminative coordinates on training pairs, evaluate shares on held-out pairs, and rerun the entire selection under within-pair label permutations. Report identity, singular-vector, and seeded random-rotation bases, paired bootstrap uncertainty, and sensitivity to magnitude definitions. A small SVD correlation or an isolated sign reversal across uncontrolled models cannot by itself explain mean-erasure. The latter is fully determined by four scores and is invariant to a common orthogonal rotation.