BEYOND RECONSTRUCTION: PRESERVING AND DECOMPOSING COMPOSITIONAL BEHAVIOR WITH MULTIMODAL SPARSE AUTOENCODERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse autoencoders (SAEs) decompose VLM representations into interpretable features, but recovering objects, attributes, and relations does not establish that an SAE preserves how the model combines them. Reconstruction, sparsity, retrieval, and feature purity therefore provide limited evidence of compositional faithfulness. We identify two requirements for valid multimodal analysis: independently inferred image and text codes to prevent pair-identity leakage, and determinate scores for every selected feature. Without these safeguards, random dictionaries can yield near-perfect retrieval, while fixed product-TopK can select coordinates arbitrarily. We then train a multimodal SAE using Group-Hoyer sparsity, cross-modal decoding, and a variable-size positive-only paired auxiliary. Across matched CLIP, SigLIP, and SigLIP2 experiments, it improves reconstructed SugarCrepe accuracy over MGSAE by 3.7-5.1 percentage points while preserving retrieval geometry and automatically measured shared-concept coverage with 64-256 active features. Exact compositional-margin decompositions identify compact, cross-seed-recoverable circuits whose ablation and activation swapping affect held-out decisions more than energy-matched random controls. Ordered relation binding, however, remains distributed rather than localized to individual symbolic features. These results show that multimodal SAEs can preserve and causally expose compositional computations.

1 INTRODUCTION

Vision-language models (VLMs) must represent not only which concepts appear in an image or text, but how those concepts are bound together: which object carries a given attribute and which entity acts upon another. A representation that detects *person*, *horse*, and *riding* without preserving this binding cannot support the distinction. Compositionality is therefore a question about the structured use of concepts in a model's decisions, not about their availability. Contrastively trained VLMs such as CLIP, SigLIP, and SigLIP2 learn shared image-text embedding spaces with strong retrieval and transfer performance (Radford et al., 2021; Zhai et al., 2023; Tschannen et al., 2025), yet controlled benchmarks reveal persistent sensitivity to object and word presence alongside weak sensitivity to attributes, order, and relations (Thrush et al., 2022; Yuksekgonul et al., 2023; Hsieh et al., 2023). When a model confuses a caption with its relation-swapped foil, the relevant relation may be absent, weakly accessible, distributed across many directions, or attenuated during global pooling. Separating these possibilities requires a representation-level tool that exposes candidate variables and tests how they affect behavior. Sparse autoencoders (SAEs) offer one such tool. Given a dense representation $x \in \mathbb{R}^d$, an SAE learns an encoder f, an overcomplete dictionary $D \in \mathbb{R}^{d \times p}$ with $p > d$, and a decoder bias b, producing a sparse code $z = f(x) \in \mathbb{R}^p$ and reconstruction $\hat{x} = Dz + b$. The learned coordinates can be inspected through highly activating examples and manipulated through ablation, activation exchange, or steering. Multimodal SAEs extend this idea by seeking features that are meaningful across visual and linguistic representations, and recent methods recover recognizable concepts, cross-modal correspondences, and grounded image-text features (Papadimitriou et al., 2025; Kaushik et al., 2026; Nasiri-Sarvi et al., 2026; Gu et al., 2026). However, *concept recovery is not compositional decomposition*. Features associated with *person*, *horse*, and *riding* do not show whether the SAE preserves their roles in a caption or whether those features affect the backbone's decision. Reconstruction error, retrieval, and semantic purity can miss this distinc-

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

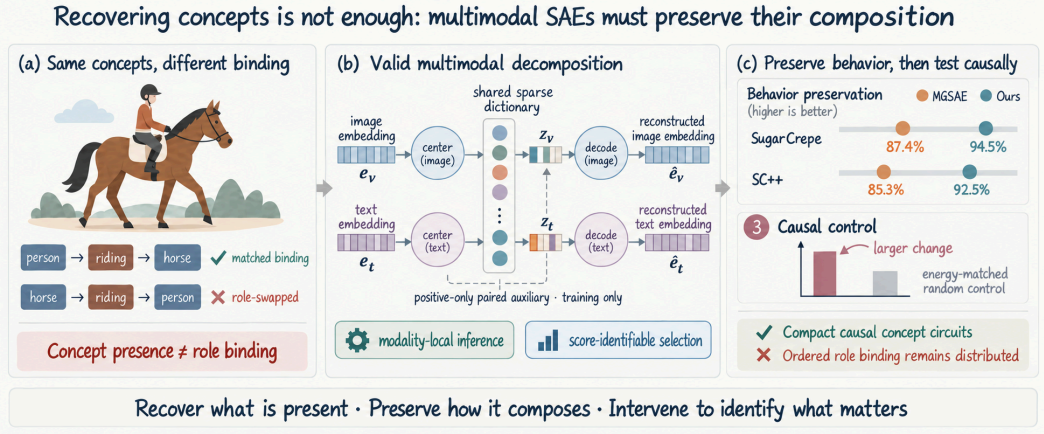

Figure 1: **Concept recovery alone does not guarantee compositional faithfulness.** (a) The same concepts can support different role bindings. (b) Our multimodal SAE uses a shared dictionary with modality-local image and text inference, the positive-only paired auxiliary is used during training. (c) We assess preservation of compositional decisions and test selected circuits through held-out interventions against matched random controls.

tion: reconstruction may reverse a small contrastive margin, while a recognizable feature need not causally support the decision. Multimodal analysis also requires care in selecting active features. Using a paired caption to select an image's support can encode pair identity, yielding near-perfect retrieval even with random dictionaries. Fixed product-TopK selection is ambiguous when $K > N_+$ positive bilateral products: the remaining coordinates come from tied zeros rather than cross-modal evidence. We therefore require *modality-local* inference and *score-identifiable* selection.

We ask: *can a multimodal SAE preserve a pretrained VLM's compositional behavior, and do its features causally mediate that behavior?* We test three distinct levels: *concept presence*, whether sparse codes recover object-, attribute-, and relation-sensitive information; *compositional faithfulness*, whether reconstruction preserves decisions between minimally different structured alternatives; and *causal compositional control*, whether interventions on independently discovered features change those decisions. Recovering a concept alone establishes neither decision preservation nor causal mediation. Our SAE independently infers image and text codes to prevent pair leakage. Modality-specific centering and a shared encoder-decoder define a common feature space; Group-Hoyer regularization, shared latent dropout, and cross-decoding discourage modality-private shortcuts. A variable-size, positive-only paired auxiliary encourages shared features without forcing zero-score coordinates into a fixed support (Sec. 4). At matched exact sparsity, the models retain shared-concept coverage with 64-256 active features. Across CLIP, SigLIP, and SigLIP2, they improve reconstructed SugarCrepe accuracy over MGSAE (Kaushik et al., 2026) while maintaining comparable retrieval. Closed-form margin decomposition identifies compact circuits whose ablation and activation exchange affect held-out decisions more than activation-energy-matched random controls, functional matching recovers related circuits across seeds. Prior multimodal SAEs emphasize shared-feature discovery through reconstruction, retrieval, and semanticity. We additionally test whether sparse features preserve and causally influence the frozen VLM's compositional decisions. While object, attribute, action, and scene information is accessible, and relation-sensitive features contribute to compositional margins, ordered role binding does not consistently localize to individual features. The resulting decomposition supports causal analysis without yielding a fully symbolic account of composition. Multimodal SAEs therefore provide a causal coordinate system for studying compositional decisions, rather than automatically converting dense VLM representations into symbolic programs. Our main contributions are:

- **Compositional evaluation beyond reconstruction.** We distinguish concept presence, compositional faithfulness, and causal compositional control, and evaluate each through behavioral, geometric, semantic, and intervention-based measures.
- **Validity conditions for multimodal SAE analysis.** We formalize modality-local inference and selector-score identifiability, demonstrate identity leakage through pair-constructed supports, and characterize the indeterminate regime of fixed product-TopK selection.

- **A behavior-preserving multimodal SAE.** We combine independent inference, modality-specific centering, a shared dictionary, structured sparsity, cross-decoding, and a positive-only paired auxiliary, improving compositional preservation over matched MGSAE across three VLM backbones.
- **Causal circuits and limits of symbolic interpretation.** Margin decomposition, cross-seed matching, ablation, and activation exchange reveal reproducible causal circuits, while showing that ordered relation binding remains distributed.

Our central claim is that multimodal SAE evaluation requires behavioral and causal validation in addition to reconstruction and semanticity, and that such validation is meaningful only under modality-local, score-identifiable inference.

2 Related Work

Sparse autoencoders and dictionary learning. Sparse autoencoders (SAEs) reconstruct neural representations using sparse codes over overcomplete dictionaries (Cunningham et al., 2023; Gao et al., 2024). Reconstruction, sparsity, and semantic labels do not establish that the recovered coordinates are independent factors used by the model: feature splitting, absorption, and polysemanticity can distribute one factor across several coordinates or combine correlated concepts within one latent (Chanin et al., 2025).

Multimodal sparse autoencoders. Recent multimodal SAEs recover cross-modal concepts, modality-specialized features, latent correspondences, and grounded image-text structure (Pach et al., 2025; Papadimitriou et al., 2025; Kaushik et al., 2026; Nasiri-Sarvi et al., 2026; Gu et al., 2026; Shen et al., 2025). Their evaluations emphasize reconstruction, retrieval, semanticity, support overlap, and localization. Our closest baseline, MGSAE (Kaushik et al., 2026), combines paired group sparsity with cross-modal masking to encourage shared concepts in pooled, aligned embeddings. We additionally require modality-local inference when evaluating all compared methods and test whether their coordinates preserve and causally mediate compositional decisions. Modality-specific geometry can confound semantic correspondence (Liang et al., 2022; Dhimoïla et al., 2026); pair-conditioned or score-indeterminate selection can create apparent agreement without independent recognition. Direct index overlap is also unreliable across independently trained dictionaries whose features may permute, split, or merge (Paulo & Belrose, 2025). We address these concerns through modality-local inference, score-identifiable selection, and functional cross-seed matching.

Compositionality and causal interpretation in VLMs. Controlled benchmarks such as ARO, Winoground, SugarCrepe, and SugarCrepe++ reveal VLM failures involving attributes, order, relations, and semantic roles (Yuksekgonul et al., 2023; Thrush et al., 2022; Hsieh et al., 2023; Dumpala et al., 2024). Likewise, decodability does not establish that a model uses the decoded property in its decisions (Elazar et al., 2021). Mechanistic interpretability therefore supplements correlational evidence with interventions such as activation and path patching or causal tracing (Meng et al., 2023; Goldowsky-Dill et al., 2023; Conmy et al., 2023). We apply this principle to multimodal sparse dictionaries using held-out feature interventions, energy-matched controls, and cross-seed matching.

3 Beyond Semanticity: Compositional Faithfulness in Multimodal SAEs

3.1 Problem Setup and Compositional Faithfulness

Let g_v and g_t be the frozen image and text encoders of a contrastive VLM, producing $x_v = g_v(I), x_t = g_t(T) \in \mathbb{R}^d$. Writing $\mathcal{N}(x) = x/\|x\|_2$, the model scores an image-text pair by

$$s(x_v, x_t) = \mathcal{N}(x_v)^\top \mathcal{N}(x_t). \tag{1}$$

A multimodal SAE maps each representation to a nonnegative overcomplete code and reconstructs it as

$$z_m = F_m(x_m) \in \mathbb{R}^p_{\geq 0}, \qquad \hat{x}_m = H_m(z_m), \qquad m \in \{v, t\}, \;\; p > d. \tag{2}$$

The backbone remains frozen. Sec. 4 specifies our shared-dictionary parameterization. The conventional criterion of accurate reconstruction under sparse codes is insufficient here, because a VLM is used through comparisons, a reconstruction can be close to its input in average vector-space error while changing the decision that representation induces.

We therefore study compositional examples $e = (I, T^+, T^-)$, where T^+ describes the image and T^- is a minimally different foil differing in an attribute assignment, object identity, action, word order, relation, or subject-object binding. The dense and reconstructed decision margins are

$$\Delta(e) = s(x_v, x_t^+) - s(x_v, x_t^-), \qquad \hat{\Delta}(e) = s(\hat{x}_v, \hat{x}_t^+) - s(\hat{x}_v, \hat{x}_t^-), \tag{3}$$

and the VLM prefers the positive caption when $\Delta(e) > 0$. Our aim is not to minimize $\|x - \hat{x}\|$, but to learn a sparse coordinate system that retains the geometry relevant to $\Delta(e)$ and supports intervention on the features contributing to it. This form is image-anchored and covers SugarCrepe and SugarCrepe++, bidirectional benchmarks such as Winoground require both anchoring directions to be preserved and are treated in App. B.1.

We distinguish three levels of evidence. *Concept presence* asks whether objects, attributes, actions, or relations are recoverable from the code. *Compositional faithfulness* asks whether reconstruction preserves how the backbone distinguishes structured alternatives. *Causal compositional control* further requires that independently identified features change those decisions under intervention. These levels are not equivalent: decodability does not imply behavioral use, and decision preservation does not identify the features responsible for the decision. Assuming nonzero margins, an SAE preserves example e when $\Delta(e)\hat{\Delta}(e) > 0$. Let $\mathcal{E}$ be a benchmark and let $\mathcal{E}_0 = \{e \in \mathcal{E} : \Delta(e) \neq 0,\ \hat{\Delta}(e) \neq 0\}$ be the subset on which Proposition 1 applies. Dataset-level decision preservation is

$$\mathrm{DP}(\mathcal{E}) = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbf{1}\left[\Delta(e)\hat{\Delta}(e) > 0\right]. \tag{4}$$

This measures fidelity to the frozen backbone rather than correctness. We therefore report it together with dense and reconstructed accuracy and with the rates of *corrections* ($C(\mathcal{E})$) $\Delta < 0 < \hat{\Delta}$ and *degradations* ($D(\mathcal{E})$) $\hat{\Delta} < 0 < \Delta$.

Proposition 1 (Exact behavioral accounting)**.** *On* $\mathcal{E}_0$,

$$\mathrm{DP} = 1 - C(\mathcal{E}) - D(\mathcal{E}), \qquad A_{\mathrm{SAE}} - A_{\mathrm{dense}} = C(\mathcal{E}) - D(\mathcal{E}). \tag{5}$$

Thus, unchanged aggregate accuracy can conceal extensive behavioral changes.

Finally, compositional preservation depends on the original margin and the direction of reconstruction error, not only its average magnitude. Define

$$\delta_v = \|\mathcal{N}(x_v) - \mathcal{N}(\hat{x}_v)\|_2, \quad \delta_\pm = \left\|\mathcal{N}(x_t^\pm) - \mathcal{N}(\hat{x}_t^\pm)\right\|_2, \quad \hat{\rho} = \left\|\mathcal{N}(\hat{x}_t^+) - \mathcal{N}(\hat{x}_t^-)\right\|_2. \tag{6}$$

Proposition 2 (Compositional margin stability)**.** *Let* e *be an example for which* $x_v, \hat{x}_v, x_t^\pm, \hat{x}_t^\pm$ *are all nonzero. Then*

$$\left|\hat{\Delta}(e) - \Delta(e)\right| \leq \hat{\rho}\,\delta_v + \delta_+ + \delta_- \leq 2\delta_v + \delta_+ + \delta_-, \tag{7}$$

and the dense decision is preserved whenever

$$|\Delta(e)| > \hat{\rho}\,\delta_v + \delta_+ + \delta_-. \tag{8}$$

The bound explains why reconstruction error alone cannot certify compositional faithfulness: small-margin examples may flip under minor angular distortion, while large-margin decisions tolerate greater error. Moreover, only error along the caption contrast affects the decision. The proof, tie handling, behavioral-accounting derivation, and bidirectional generalization are provided in App. B.

3.2 Validity of Multimodal Sparse Inference

A shared coordinate is meaningful only if its activation is independently supported by both modalities. We therefore distinguish paired supervision during training from pair-conditioned inference.

Definition 1 (Modality-local inference)**.** *Inference is modality-local if*

$$z_v = F_v(x_v), \qquad z_t = F_t(x_t), \tag{9}$$

so that each primary code is invariant to the input paired with it from the other modality.

This condition permits paired training objectives but prohibits the known pair from constructing the codes used for retrieval, interpretation, or intervention. Without it, pair identity alone can produce apparent alignment.

Proposition 3 (Pair-conditioned supports can encode identity)**.** *For* N *image-text pairs, latent width* p*, and support size* K *satisfying* $\binom{p}{K} \geq N$*, there exists a pair-conditioned coding rule that achieves perfect rank-one latent retrieval without using any semantic content.*

The construction assigns each pair a distinct size-K support and gives its image and caption the same binary code. Matching similarity is then 1, while every nonmatching similarity is strictly smaller. Thus, retrieval and support overlap do not identify semantic alignment when pair identity enters inference (App. B.4). Even with independently inferred codes, a cross-modal selector can be ambiguous.

Definition 2 (Selector-score identifiability)**.** *A selected support is score-identifiable if it is uniquely determined by the intended score vector q. For size-K TopK selection, this requires the strict boundary*

$$q_{(K)} > q_{(K+1)}, \tag{10}$$

where $q_{(1)} \geq \cdots \geq q_{(p)}$.

For nonnegative image and text codes, consider the bilateral product score

$$q_j = z_{v,j} z_{t,j}, \qquad N_+ = \#\{j : q_j > 0\}. \tag{11}$$

Proposition 4 (Indeterminacy of fixed product-TopK)**.** *For $1 \leq K < p$, if $K > N_+$, product-TopK selects exactly $K - N_+$ zero-product coordinates and*

$$q_{(K)} = q_{(K+1)} = 0. \tag{12}$$

Its support is therefore not uniquely determined by bilateral evidence.

After the N_+ positive products are selected, the remaining slots must be filled from a tied zero-score set. A deterministic implementation can return the same indices repeatedly, but this reflects its tie-breaking convention rather than semantic evidence. We measure the affected support fraction by

$$\rho_{\text{zero}} = \frac{(K - N_+)_+}{K}. \tag{13}$$

The full proof and diagnostic details are provided in App. B.5.

Together, Definitions 1 and 2 impose two requirements:

i. **locality:** each primary code is inferred using only its own modality, and
ii. **identifiability:** each selected coordinate is distinguished by the evidence the selector is intended to measure.

The SAE introduced next satisfies the first requirement through independent primary inference. It addresses the second by using paired information only as a training signal and by restricting its paired auxiliary to the variable-size set of coordinates with strictly positive bilateral evidence, rather than padding that set to a fixed support with zero-score features.

4 Learning Shared Concepts with Multimodal SAEs

Paired supervision can encourage a shared image-text concept space, but pair-conditioned inference can construct artificial agreement (Sec. 3.2). We therefore separate independently inferred *primary codes* from a paired auxiliary used only during training.

4.1 Modality-Specific Centering

Image and text embeddings exhibit a systematic modality offset. To prevent the shared dictionary from spending capacity on this first-order difference, we initialize separate modality-specific centers from the training embeddings:

$$\mu_m^{(0)} = \frac{1}{|\mathcal{D}_{\text{train}}|} \sum_{i \in \mathcal{D}_{\text{train}}} x_m^{(i)}, \qquad \tilde{x}_m = x_m - \mu_m, \qquad m \in \{v, t\}. \tag{14}$$

The centers μ_v and μ_t are initialized to $\mu_v^{(0)}$ and $\mu_t^{(0)}$, respectively, and optimized jointly with the SAE. During decoding, the corresponding learned center is added back. This accounts for the modality offset without assuming identical higher-order geometry across image and text embeddings.

4.2 Shared Dictionary and Independent Primary Codes

Both modalities use the same encoder and decoder dictionary:

$$z_m = \text{ReLU}\left(W_{\text{enc}} \tilde{x}_m + b_{\text{enc}}\right), \qquad m \in \{v, t\}, \tag{15}$$

where $W_{\text{enc}} \in \mathbb{R}^{p \times d}$ and the shared decoder $D \in \mathbb{R}^{d \times p}$ has unit-norm columns. Sharing parameters gives coordinate j a common image-text parameterization, while z_v and z_t remain independently inferred and need not have identical supports. Thus, the dictionary is shared architecturally, but support overlap must emerge from the data. The primary reconstruction objective is

$$\mathcal{L}_{\text{self}} = \mathbb{E}_{(x_v,x_t)\sim\mathcal{D}_{\text{train}}} \left[\ell(x_v, \hat{x}_v) + \ell(x_t, \hat{x}_t)\right], \tag{16}$$

where ℓ is mean-squared reconstruction error.

4.3 Cross-Decoding

A shared dictionary can still divide into modality-specialized subsets. Cross-decoding therefore requires each independently inferred code to retain information useful to the paired modality:

$$\begin{aligned} \hat{x}_{t\leftarrow v} &= Dz_v + \mu_t, & \hat{x}_{v\leftarrow t} &= Dz_t + \mu_v, \\ \mathcal{L}_{\text{cross}} &= \mathbb{E}_{(x_v,x_t)} \left[\ell(x_t, \hat{x}_{t\leftarrow v}) + \ell(x_v, \hat{x}_{v\leftarrow t})\right]. \end{aligned} \tag{17}$$

The paired representation appears only as a reconstruction target, it does not participate in producing the primary code.

4.4 Group-Hoyer Concentration

We encourage paired activation energy to concentrate over fewer coordinate groups. For $g_j = (z_{v,j}^2 + z_{t,j}^2)^{1/2}$, we use

$$H(g) = \frac{\sqrt{p} - \|g\|_1/(\|g\|_2 + \epsilon)}{\sqrt{p} - 1}, \qquad \mathcal{L}_{\text{group}} = \mathbb{E}_{(x_v,x_t)}[1 - H(g)]. \tag{18}$$

This promotes a small effective support without enforcing an exact number of nonzero activations. The model may therefore emit many positive ReLU coordinates while concentrating most of its energy in substantially fewer features.

4.5 Strictly Positive Paired Auxiliary

Cross-decoding and Group-Hoyer concentration encourage transferable structure but do not require selected coordinates to have positive evidence in both modalities. After independently obtaining z_v and z_t, we define

$$S_+(z_v, z_t) = \{j : z_{v,j} z_{t,j} > 0\}, \qquad z_{m,j}^{\text{pair}} = z_{m,j} \mathbf{1}[j \in S_+], \qquad m \in \{v, t\}. \tag{19}$$

Its support size $|S_+| = N_+$ varies across pairs. Unlike fixed product-TopK, the auxiliary never introduces zero-product coordinates merely to reach a prescribed support size.

We train the paired codes to retain information sufficient for reconstructing their respective modalities:

$$\mathcal{L}_{\text{pair}} = \mathbb{E}_{(x_v,x_t)} \left[\ell(x_v, Dz_v^{\text{pair}} + \mu_v) + \ell(x_t, Dz_t^{\text{pair}} + \mu_t)\right]. \tag{20}$$

These codes are used only to provide an auxiliary training signal. All reconstruction, retrieval, semantic analysis, and causal interventions use the independently inferred primary codes.

Training objective. The core objective is

$$\mathcal{L}_{\text{core}} = \mathcal{L}_{\text{self}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{group}}\mathcal{L}_{\text{group}} + \lambda_{\text{pair}}\mathcal{L}_{\text{pair}}. \tag{21}$$

The VLM remains frozen and only the SAE parameters are optimized. Shared latent masking and remaining training details are in Appendix C.

4.6 Decomposing Compositional Decisions

Semantic labels reveal what correlates with an SAE coordinate, but not how it supports a VLM decision. We therefore decompose the reconstructed margin exactly. Let

$$u_v = \frac{\hat{x}_v}{\|\hat{x}_v\|_2}, \qquad r_\pm = \|\hat{x}_{t,\pm}\|_2,$$

where $\hat{x}_{t,\pm} = Dz_{t,\pm} + \mu_t$ are the reconstructed positive and negative captions. The contribution of coordinate j is

$$c_j(e) = u_v^\top D_j \left(\frac{z_{t,+,j}}{r_+} - \frac{z_{t,-,j}}{r_-} \right), \tag{22}$$

and the reconstructed margin satisfies the exact identity

$$\hat{\Delta}(e) = \sum_{j=1}^{p} c_j(e) + c_\mu(e), \qquad c_\mu(e) = u_v^\top \mu_t \left(\frac{1}{r_+} - \frac{1}{r_-} \right). \tag{23}$$

Thus, $c_j > 0$ supports the correct caption and $c_j < 0$ supports the foil. Contribution depends on both the activation contrast and the alignment of decoder direction D_j with the decision direction, activation magnitude alone is insufficient. We use a one-sided decomposition that fixes the common image representation, avoiding the non-unique allocation of bilinear image-text interactions. The derivation and image-side analogue are given in App. D.1 and D.2.

4.7 Causal Circuits and Functional Stability

Per-feature contributions decompose the reconstructed margin but do not establish causal effects. We therefore select circuits on a discovery set and test interventions on a disjoint evaluation set, these interventions affect SAE-reconstructed decisions, not the frozen backbone. For each coordinate, we compute its mean absolute contribution on $\mathcal{E}_{\text{disc}}$ and select the top L:

$$a_j = \frac{1}{|\mathcal{E}_{\text{disc}}|} \sum_{e \in \mathcal{E}_{\text{disc}}} |c_j(e)|, \qquad C_L = \text{TopL}_j(a_j). \tag{24}$$

On $\mathcal{E}_{\text{eval}}$, we ablate C_L to test necessity and exchange its activations between positive and negative captions to test contrast-specific control. We compare margin and decision changes with random sets matched for size and aggregate activation energy. To assess cross-seed stability despite dictionary permutations, we match features one-to-one by Hungarian assignment on decoder cosine similarity and test whether they recur in independently discovered contribution rankings. Intervention, matching, and uncertainty details appear in App. D.

5 Experiments

Our experiments address four questions: **(Q1)** Do standard multimodal SAE evaluations admit non-semantic or score-indeterminate solutions? **(Q2)** Does sparse reconstruction preserve the retrieval geometry and compositional decisions of the frozen VLM? **(Q3)** What semantic and grounded structure is exposed by the learned coordinates? **(Q4)** Do compact feature circuits causally and reproducibly control compositional decisions?

5.1 Experimental Setup

Data, backbones, and baselines. We train SAEs on paired MS COCO embeddings from frozen CLIP, SigLIP, and SigLIP2 backbones, primarily using pooled representations; patches and tokens are used only for grounding and relation analyses. Baselines are the dense VLM, a reconstruction-only SAE, an independently inferred TopK SAE, and MGSAE (Kaushik et al., 2026). Where applicable, we match data, dictionary width, sparsity, and optimization budget. All primary evaluations use modality-local inference, avoiding the pair-conditioned support leakage shown in Sec. 5.2.

Operating points and evaluation. We evaluate native ReLU codes and modality-local exact-sparse TopK codes, primarily at $K \in \{64, 128, 256\}$. Evaluation covers reconstruction and geometry, compositional accuracy and decision preservation, feature semanticity and grounding, and held-out causal interventions with energy-matched and cross-seed controls.

Statistical protocol. Primary results average three independently trained SAEs (seeds 17-19); hyperparameters are selected without benchmark test labels. Circuit discovery, causal evaluation, and cross-seed matching use separate splits. Full datasets, model variants, metrics, optimization settings, sparsity sweeps, and uncertainty procedures appear in Appendices E, F, and H.3.

5.2 Q1: Are Multimodal SAE Evaluations Identifiable?

Pair-conditioned inference can manufacture apparent cross-modal alignment by assigning each image-caption pair a jointly constructed support that encodes pair identity rather than indepen-

Table 1: **Latent retrieval under pair-conditioned and modality-local inference.** Ranges aggregate the audited models and retrieval directions.

Codes/supports	Inference	R@1
Shared random supports	Pair-conditioned	1.000
Frozen random dictionaries	Pair-conditioned	0.973-0.998
Learned dictionaries	Pair-conditioned	0.955-1.000
Learned dictionaries	Modality-local	0.246-0.353

dently recognized semantics (Sec. 3). As shown in Table 1, shared random supports achieve perfect retrieval, while frozen random dictionaries reach R@1 of 0.973-0.998. Learned pair-conditioned codes are similarly near-perfect, but retrieval drops substantially under modality-local inference. A three-seed SPARC reproduction confirms this pattern: at $K = 768$, pair-conditioned R@1 is 0.9997 ± 0.0001 for image-to-text and 0.9996 ± 0.0000 for text-to-image, falling to 0.0795 ± 0.0030 and 0.1220 ± 0.0076 under its modality-local safeguard. Thus, near-perfect latent retrieval does not establish semantic alignment when the known pair participates in support construction.

Fixed product-TopK introduces a separate ambiguity. Let $N_+ = \#\{j : z_{v,j} z_{t,j} > 0\}$ and $\rho_{\text{zero}} = (K - N_+)_+ / K$. Among 29 compatible historical checkpoints, 26 satisfy $K > N_+$, forcing the remaining support to be selected among tied zero scores. Deterministic tie-breaking concentrates these unsupported selections on particular coordinate ranges, even while reconstruction, retrieval, and geometry metrics improve. These audits motivate two requirements: primary codes must be inferred *modality-locally*, and every selected coordinate must be *score-identifiable*. All remaining experiments satisfy both conditions. Complete random controls, per-checkpoint results, index histograms, and training trajectories are provided in App. G.

5.3 Q2: Do Sparse Representations Preserve VLM Behavior?

Compositional faithfulness. Table 2 shows that reconstruction fidelity alone does not predict behavioral preservation. Across CLIP, SigLIP, and SigLIP2, our TopK-256 model achieves higher reconstructed SugarCrepe accuracy than MGSAE-768 while using one third as many active coordinates. On SigLIP2, for example, MGSAE reconstructs with less than half the FVU of our model (0.045 vs. 0.112), yet our model achieves higher SugarCrepe accuracy (0.840 vs. 0.815) and decision preservation (0.892 vs. 0.874). Behavioral accounting using Eq. 5 shows that these gains arise primarily from fewer degradations and greater preservation of backbone decisions, rather than indiscriminate prediction changes. The reconstructed margins likewise remain closer to those of the dense backbone, particularly for vulnerable small-margin examples.

Fidelity-sparsity trade-off. We additionally evaluate independently inferred exact-K codes for $K \in \{64, 128, 256\}$. Our model maintains meaningful shared-concept coverage and stronger compositional preservation than matched MGSAE across these operating points. Larger K generally improves reconstruction and geometry preservation but yields less concentrated representations, while reconstruction-only TopK can achieve lower FVU. The results therefore define a Pareto trade-off among vector fidelity, sparsity, semantic coverage, and behavioral faithfulness. Complete sweeps are reported in Appendix H.3.

Component analysis. One-at-a-time sensitivity (App. C.9) shows that stronger cross-decoding raises automatically measured shared-concept coverage at a cost in FVU. Architecture controls (Appendix H.7) show that reconstruction-only and ordinary TopK variants preserve compositional decisions more closely, while our objective trades part of that fidelity for a shared cross-modal dictionary.

5.4 Q3: What Semantic and Compositional Structure Emerges?

Semantic coverage and grounding. Using top-activating image and text examples, we identify shared object, attribute, action, and scene features with substantial purity and lift over their dataset priors. Exact-sparse variants retain meaningful shared-concept coverage at $K = 64$-256. Figure 13 illustrates matched image-text coordinates for three COCO pairs. Figure 14 illustrates how a selected compositional circuit affects dense patch and token representations. The circuit produces more held-out decision flips than matched random controls, while patch and token localization scores remain comparable to those controls (Table 36). We therefore interpret these visualizations as circuit-level diagnostics rather than evidence about the spatial grounding of individual pooled features. Nevertheless, automatic semanticity does not establish a reusable variable, since concepts may split across coordinates or appear as conjunctive features. We therefore treat semantic coverage, grounding, and

Table 2: **Reconstruction and behavioral results across backbones.** FVU, GPI, and retrieval use the 5,000 COCO validation pairs, SugarCrepe and SC++ ITT report reconstructed accuracy, and DP is decision preservation (Eq. 5). All SAE inference is modality-local. Means over seeds 17-19, bold marks the best SAE per backbone. The SigLIP2 matched-support sweep is in Tables 16 and 24.

Backbone	Method	FVU↓	GPI↑	I2T R@1↑	T2I R@1↑	SugarCrepe↑	SC DP↑	SC++ ITT↑	SC++ DP↑
CLIP	Dense VLM	-	1.0000	0.3348	0.2850	0.7835	1.0000	0.6504	1.0000
	Ours TopK-256	0.1009	0.9298	0.2713	0.2468	0.7448	0.8290	0.6118	0.8276
	MGSAE-768	0.0440	0.9738	0.2629	0.2227	0.7325	0.8286	0.6140	0.8315
	Ours dense	**0.0232**	**0.9798**	**0.3192**	**0.2877**	**0.7903**	**0.9220**	**0.6460**	**0.9189**
SigLIP	Dense VLM	-	1.0000	0.4472	0.4414	0.8369	1.0000	0.7030	1.0000
	Ours TopK-256	0.1230	0.9496	0.3696	0.3684	0.7935	0.8645	0.6654	0.8416
	MGSAE-768	0.0504	0.9816	0.3707	0.3746	0.7783	0.8601	0.6567	0.8485
	Ours dense	**0.0257**	**0.9891**	**0.4307**	**0.4317**	**0.8265**	**0.9430**	**0.6890**	**0.9234**
SigLIP2	Dense VLM	-	1.0000	0.4922	0.5034	0.8667	1.0000	0.7538	1.0000
	Ours TopK-256	0.1121	0.9521	0.4228	0.4310	0.8397	0.8922	0.7149	0.8636
	MGSAE-768	0.0453	**0.9824**	0.4197	0.4423	0.8153	0.8739	0.6989	0.8526
	Ours dense	**0.0396**	0.9790	**0.4679**	**0.4793**	**0.8592**	**0.9365**	**0.7332**	**0.9118**

behavioral faithfulness as separate evaluation axes. MGSAE, for example, can produce more concentrated codes and stronger automatic shared-concept scores at its native operating point, while our model better preserves compositional behavior.

Relations are present but not factorized. Relation categories remain decodable from pooled and token-level representations, and relation-sensitive features contribute to compositional margins. However, these features often mix predicates with common arguments or scene context and do not consistently distinguish reversed subject-object roles. Relation decodability can therefore remain high even when ARO relation and Winoground performance is weak. The learned dictionaries expose relation-sensitive information, but not stable symbolic subject-relation-object variables. Complete labeling, grounding, and relation analyses are provided in Appendix I.

5.5 Q4: Do Sparse Circuits Causally and Reproducibly Control Decisions?

Held-out causal control. The exact margin decomposition in Section 4.6 shows that a relatively small feature circuit captures a disproportionate share of the reconstructed compositional margin. This concentration is strongest for object and attribute decisions, whereas ordered relations depend on broader circuits. As shown in Figure 17, ablating circuits discovered on separate data changes held-out margins and decisions substantially more than ablating size- and activation-energy-matched random features. Activation exchange has an even stronger contrast-specific effect, shifting margins toward the opposite composition when positive- and negative-caption activations are exchanged.

Functional stability and scope. Literal feature indices differ across independently trained dictionaries. Matching features by decoder geometry, 31% of the seed-17 top-32 SugarCrepe circuit recurs in each target seed's top-32 ranking (chance: 0.3%), although matched decoder cosine is modest (0.22), indicating rotation or splitting across seeds. We report this as rank recovery; causal transfer of mapped circuits is left to future work. These results show that selected SAE coordinates mediate part of the reconstructed model's compositional behavior, but not that composition is implemented by a small symbolic program. Ordered relation binding remains distributed; complete analyses appear in Appendix J.

6 Conclusion

We introduced a framework for evaluating multimodal SAEs through concept presence, compositional faithfulness, and causal compositional control, while requiring modality-local inference and selector-score identifiability. Across CLIP, SigLIP, and SigLIP2, our model preserves compositional decisions better than matched MGSAE while retaining retrieval geometry and meaningful shared-concept coverage under exact sparsity. Exact margin decomposition and held-out interventions further reveal reproducible feature circuits that causally influence reconstructed decisions. However, ordered relation binding remains distributed rather than localized to clean symbolic neurons. Multimodal SAEs therefore provide a useful causal coordinate system for studying composition, provided that their apparent cross-modal structure is learned rather than introduced by inference. Broader implications and limitations are discussed in Appendix K.

AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: (i) providing feedback on research methodology and experimental design, in the form of critique of draft analyses and suggestions for additional controls and evaluations, and (ii) checking the correctness of mathematical proofs. All propositions and proofs were formulated and written by the authors, AI tools were used only to verify the authors' derivations and flag potential errors, and any flagged issue was resolved by the authors. We have not used generative AI tools to generate synthetic data, formulate mathematical claims, provide critical ingredients for or write proofs, propose hypotheses, implement methods, clean or reformat datasets, or interpret results, translation and qualitative data analysis are not applicable to this work.

Additionally, we used generative AI tools for tasks with recommended disclosure: brainstorming, creating and modifying charts and scientific figures, and checking the manuscript for errors, internal consistency, and readability. The manuscript text was written by the authors.

We have reviewed all AI-assisted work. All figures were checked against the underlying experimental outputs, all experimental results were produced by the authors' code, and every AI-suggested correction or edit was verified by an author before inclusion. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

We will release anonymized code as supplementary material, including SAE training, modality-local inference, all evaluation pipelines, the exact margin decomposition, and the causal-intervention and random-control procedures, together with configuration files for every reported checkpoint. All VLM backbones (CLIP, SigLIP, SigLIP2) and datasets (MS COCO, SugarCrepe, SugarCrepe++, ARO, Winoground) are publicly available. Data and split construction, including the disjoint discovery and evaluation splits used for causal analysis, are described in Appendices E.1 and D.4 (Table 13). The complete training objective and all loss coefficients are given in Appendix C (Table 4 and Eq. 76), and optimization hyperparameters and per-backbone configurations are given in Appendix C.8 (Tables 7 and 8). Compared methods and model selection are specified in Appendices E.2-E.3, and the executed causal protocol is summarized in Appendix D (Table 10). All evaluation metrics are defined in Appendix F. Complete proofs and assumptions for all propositions appear in Appendix B. Main results are averaged over three independently trained seeds (17-19) and causal results over five seeds (17-21), the statistical protocol and uncertainty estimates are described in Appendices E.4 and D.10 (Table 15), and per-seed results are reported in Appendices H.10 and J.8.

REFERENCES

Matteo Attimonelli, Alessandro De Bellis, Franco Maria Nardini, Claudio Pomo, Cosimo Rulli, Rossano Venturini, and Tommaso Di Noia. Puma: Post-hoc sparsification of universal multimodal embeddings for efficient retrieval, 2026. URL https://arxiv.org/abs/2608.25780.

David Chanin, James Wilken-Smith, Tomáš Dulka, Hardik Bhatnagar, Satvik Golechha, and Joseph Bloom. A is for absorption: Studying feature splitting and absorption in sparse autoencoders, 2025. URL https://arxiv.org/abs/2409.14507.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Stefan Heimersheim, and Adrià Garriga-Alonso. Towards automated circuit discovery for mechanistic interpretability, 2023. URL https://arxiv.org/abs/2304.14997.

Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL https://arxiv.org/abs/2309.08600.

Grégoire Dhimoïla, Thomas Fel, Victor Boutin, and Agustin Picard. Cross-modal redundancy and the geometry of vision-language embeddings, 2026. URL https://arxiv.org/abs/2602.06218.

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, 2024. URL https://arxiv.org/abs/2406.11171.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals, 2021. URL https://arxiv.org/abs/2006.00995.

Leo Gao, Tom Dupré la Tour, Henk Tillman, Gabriel Goh, Rajan Troll, Alec Radford, Ilya Sutskever, Jan Leike, and Jeffrey Wu. Scaling and evaluating sparse autoencoders, 2024. URL https://arxiv.org/abs/2406.04093.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. Localizing model behavior with path patching, 2023. URL https://arxiv.org/abs/2304.05969.

Difei Gu, Yunhe Gao, Gerasimos Chatzoudis, Zihan Dong, Guoning Zhang, Bangwei Guo, Yang Zhou, Mu Zhou, and Dimitris Metaxas. Lucid-sae: Learning unified vision-language sparse codes for interpretable concept discovery, 2026. URL https://arxiv.org/abs/2602.07311.

Patrik O. Hoyer. Non-negative matrix factorization with sparseness constraints, 2004. URL https://arxiv.org/abs/cs/0408058.

Cheng-Yu Hsieh, Jieyu Zhang, Zixian Ma, Aniruddha Kembhavi, and Ranjay Krishna. Sugarcrepe: Fixing hackable benchmarks for vision-language compositionality, 2023. URL https://arxiv.org/abs/2306.14610.

Chiraag Kaushik, Davis Barch, and Andrea Fanelli. Decomposing multimodal embedding spaces with group-sparse autoencoders, 2026. URL https://arxiv.org/abs/2601.20028.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022. URL https://arxiv.org/abs/2203.02053.

Weiduo Liao, Yunqiao Yang, and Ying Wei. When structured sparse autoencoders learn consistent concepts across modalities, 2026. URL https://arxiv.org/abs/2607.08605.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt, 2023. URL https://arxiv.org/abs/2202.05262.

Ali Nasiri-Sarvi, Hassan Rivaz, and Mahdi S. Hosseini. Sparc: Concept-aligned sparse autoencoders for cross-model and cross-modal interpretability, 2026. URL https://arxiv.org/abs/2507.06265.

Mateusz Pach, Shyamgopal Karthik, Quentin Bouniot, Serge Belongie, and Zeynep Akata. Sparse autoencoders learn monosemantic features in vision-language models, 2025. URL https://arxiv.org/abs/2504.02821.

Isabel Papadimitriou, Huangyuan Su, Thomas Fel, Sham Kakade, and Stephanie Gil. Interpreting the linear structure of vision-language model embedding spaces, 2025. URL https://arxiv.org/abs/2504.11695.

Gonçalo Paulo and Nora Belrose. Sparse autoencoders trained on the same data learn different features, 2025. URL https://arxiv.org/abs/2501.16615.

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, 2021. URL https://arxiv.org/abs/2103.00020.

Shufan Shen, Junshu Sun, Qingming Huang, and Shuhui Wang. Vl-sae: Interpreting and enhancing vision-language alignment with a unified concept set, 2025. URL https://arxiv.org/abs/2510.21323.

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, 2022. URL https://arxiv.org/abs/2204.03162.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, Olivier Hénaff, Jeremiah Harmsen, Andreas Steiner, and Xiaohua Zhai. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features, 2025. URL https://arxiv.org/abs/2502.14786.

Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006. doi: https://doi.org/10.1111/j.1467-9868.2005.00532.x. URL https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9868.2005.00532.x.

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?, 2023. URL https://arxiv.org/abs/2210.01936.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023. URL https://arxiv.org/abs/2303.15343.

APPENDIX ROADMAP

A Extended Related Work and Positioning **16**

A.1 Sparse Autoencoders as Representation Decompositions 16

A.2 Hoyer Sparsity and Our Multimodal Group-Hoyer Extension 16

A.3 Sparse Decompositions of Vision-Language Representations 17

A.4 Validity of Cross-Modal Feature Alignment 18

A.5 Compositional Evaluation of Vision-Language Models 19

A.6 From Semantic Correlation to Causal Control 19

A.7 Summary of Positioning 20

B Formal Foundations and Proofs **20**

B.1 Compositional Margins and Bidirectional Benchmarks 20

B.2 Proof of Compositional Margin Stability 21

B.3 Proof of Exact Behavioral Accounting 22

B.4 Proof of Pair-Identity Leakage 23

B.5 Proof of Fixed Product-TopK Indeterminacy 24

B.6 Implications for the Experimental Analysis 24

C Training Details **24**

C.1 Self-Reconstruction and Cross-Decoding 25

C.2 Strictly-Positive Paired Auxiliary 25

C.3 Group-Hoyer Concentration 25

C.4 Inactive Balance and Mismatch Terms 26

C.5 Shared Latent Masking 26

C.6 Complete Objective 26

C.7 Auxiliary Support Statistics 26

C.8 Optimization and Hyperparameters 27

C.9 Training-Weight Sensitivity 27

D Exact Margin Decomposition and Causal Analysis Protocol **27**

D.1 Caption-Side Margin Decomposition 28

D.2 Image-Side Margin Decomposition 30

D.3 Why We Use a One-Sided Decomposition 31

D.4 Discovery and Evaluation Split 32

D.5 Circuit Discovery and Held-Out Concentration 32

D.6 Circuit Ablation and Activation Exchange 32

D.7 Size- and Activation-Matched Controls 33

D.8 Permutation Non-Identifiability and Executed Seed Matching 33

D.9 Cross-Seed Transfer: Defined but Not Executed 34

D.10 Effect Aggregation and Uncertainty 34

E Additional Experimental Details **34**

E.1 Data and Split Construction . . . 34
E.2 Compared Methods . . . 35
E.3 Optimization and Model Selection . . . 37
E.4 Statistical Protocol . . . 37

F Evaluation Metrics and Protocols **38**
F.1 Representational Fidelity . . . 38
F.2 Compositional Behavior . . . 38
F.3 Semantic Structure and Grounding . . . 38
F.4 Causal Structure . . . 39

G Q1: Detailed Validity Audit **39**
G.1 Matched Exact-Sparsity Audit . . . 39
G.2 Pair-Conditioned Support Leakage . . . 40
G.3 Score Indeterminacy Under Fixed Product-TopK . . . 41
G.4 Selector Validity During Training . . . 43
G.5 Implications for the Remaining Experiments . . . 44

H Q2: Detailed Behavioral-Preservation Analysis **44**
H.1 Behavioral Accounting . . . 45
H.2 Compositional-Margin Preservation . . . 47
H.3 Exact Sparsity and the Fidelity-Sparsity Frontier . . . 47
H.4 Matched-Support and Nearest-Fidelity Comparisons . . . 49
H.5 Benchmark- and Perturbation-Level Results . . . 49
H.6 Representation Geometry and Retrieval . . . 50
H.7 Dense-Family Architecture and Objective Controls . . . 51
H.8 Objective-Ladder and AuxK Controls . . . 51
H.9 CLIP ViT-L/14 Scale Replication . . . 52
H.10 Per-Seed Results and Uncertainty . . . 53

I Q3: Detailed Semantic and Compositional Structure Analysis **53**
I.1 Feature-Labeling Protocol . . . 53
I.2 Semantic Structure at Matched Exact Support . . . 54
I.3 Qualitative Image-Text Coordinate Matches . . . 55
I.4 What Has and Has Not Been Tested Out of Sample . . . 55
I.5 Patch/Token Causal Diagnostic . . . 56
I.6 Relation Presence, Decodability, and Ordered Binding . . . 57
I.7 Interpretation Boundary . . . 58

J Q4: Detailed Causal-Circuit Results **58**
J.1 Concentration of Margin Contributions . . . 58
J.2 Held-Out Circuit Ablation . . . 59
J.3 Activation Exchange . . . 60

J.4 Comparison with Local-Activity-Matched Controls 60
J.5 Functional Stability Across Seeds 61
J.6 Cross-Seed Circuit Transfer 62
J.7 Ranking-Criterion Steering Extension 62
J.8 Per-Seed Results and Uncertainty 62
J.9 Scope of the Causal Claim 63

K Discussion and Limitations **63**

A Extended Related Work and Positioning

This appendix expands the positioning in Section 2. We distinguish four questions that are often conflated in evaluations of sparse multimodal representations: whether a sparse code reconstructs the original representation, whether its coordinates admit semantic descriptions, whether it preserves the behavior of the underlying model, and whether the identified coordinates causally mediate that behavior. Our work primarily addresses the latter two questions in the setting of pooled dual-encoder vision-language embeddings.

A.1 Sparse Autoencoders as Representation Decompositions

Sparse autoencoders (SAEs) approximate a dense activation or embedding $x \in \mathbb{R}^d$ as

$$z = f(x), \qquad \hat{x} = Dz + b, \tag{25}$$

where $D \in \mathbb{R}^{d \times p}$ is typically overcomplete, with $p > d$, and $z \in \mathbb{R}^p$ is encouraged or constrained to be sparse. The motivating hypothesis is that neural representations contain more functionally relevant features than can be assigned one-to-one to individual neurons, and that an overcomplete sparse dictionary can separate directions otherwise represented in superposition (Cunningham et al., 2023). Subsequent work has scaled SAEs to substantially larger dictionaries and introduced controlled sparsity mechanisms, including TopK activation, together with metrics for reconstruction, dead latents, feature explainability, and downstream effects (Gao et al., 2024).

These developments make SAEs useful candidate tools for mechanistic analysis, but neither sparsity nor reconstruction uniquely determines a factorized representation. Increasing dictionary width can split a broad semantic feature into several narrower features. Conversely, feature absorption can cause a specific latent to capture a subset of examples that would otherwise activate a broader feature, preventing the broad latent from firing consistently even when its apparent concept is present (Chanin et al., 2025). Correlated factors may also be represented through conjunctive or polysemantic coordinates. A latent associated with *red car*, for example, does not imply the existence of reusable and independently manipulable *red* and *car* variables. Conversely, a single semantic factor may be distributed across several coordinates.

SAE solutions are additionally non-identifiable across random initializations. Dictionaries trained on the same model activations and data can recover substantially different literal coordinates, particularly under exact TopK sparsity (Paulo & Belrose, 2025). Consequently, a latent index should not be treated as a canonical semantic identifier. Comparisons across runs must instead account for permutation, sign or scale conventions where applicable, and feature splitting or merging. We therefore evaluate cross-seed stability through functional matching based on activation behavior and semantic response rather than direct equality of neuron indices.

These observations motivate a hierarchy of evidence. Reconstruction shows that the dictionary retains information, semantic labeling shows that some coordinates correlate with recognizable concepts, cross-context evaluation tests whether those interpretations generalize, and intervention tests whether the coordinates influence the model behavior under study. Our evaluation explicitly separates these levels rather than treating semanticity as sufficient evidence of mechanistic faithfulness.

A.2 Hoyer Sparsity and Our Multimodal Group-Hoyer Extension

Hoyer sparsity characterizes the concentration of a vector through the relationship between its ℓ_1 and ℓ_2 norms (Hoyer, 2004). Unlike an unnormalized ℓ_1 penalty, the Hoyer measure separates sparsity from the overall scale of the activations and provides a direct measure of how strongly activation mass is concentrated on a small number of coordinates. Group-sparsity methods separately encourage predefined sets of variables to become active or inactive together (Yuan & Lin, 2006).

Existing Hoyer regularization, however, is formulated for a single activation vector and does not directly encode the correspondence between independently inferred image and text codes. We introduce *Group-Hoyer regularization* as a multimodal extension of Hoyer sparsity. The regularizer organizes corresponding image and text activations into modality-aligned groups and applies Hoyer-style concentration at the group level. It thereby encourages sparse use of a shared concept dictionary while permitting a coordinate to receive different activation magnitudes from the two modalities.

Importantly, Group-Hoyer does not construct one modality's support using its paired observation from the other modality. Image and text codes remain independently inferred, and the multimodal grouping enters only as a training regularizer. This distinguishes Group-Hoyer from pair-conditioned selection rules whose inferred support depends directly on the paired example.

Our novelty claim is therefore not the introduction of Hoyer sparsity itself. Rather, to our knowledge, we introduce its grouped multimodal formulation for SAEs and combine it with modality-specific centering, independently inferred primary codes, cross-decoding, shared latent dropout, and a variable-size positive-only paired auxiliary. Group-Hoyer encourages structured bilateral feature use, while the remaining components preserve modality-local inference and discourage modality-private reconstruction shortcuts.

A.3 Sparse Decompositions of Vision-Language Representations

Recent work has extended sparse dictionary learning from language-model activations to vision and multimodal representations. Papadimitriou et al. (Papadimitriou et al., 2025) train SAEs on the embedding spaces of CLIP, SigLIP, SigLIP2, and AIMv2. They analyze reconstruction, sparsity, semantic concepts, modality specialization, and stability across training runs. Their Bridge Score combines co-activation across paired image-text inputs with geometric alignment between dictionary directions, showing that even predominantly unimodal features can participate in cross-modal integration. Their study primarily characterizes the sparse linear organization of VLM embedding spaces, it does not ask whether SAE reconstruction preserves fine-grained compositional decisions or whether the recovered features causally mediate those decisions.

Pach et al. (2025) train SAEs on the vision encoders of VLMs such as CLIP and evaluate neuron-level monosemanticity with a score validated against a large-scale user study. They find that sparsity and latent width most strongly determine monosemanticity, and that interventions on SAE features in CLIP's vision encoder steer the outputs of a downstream multimodal LLM. Their analysis is unimodal and targets the semantic purity of individual features, it does not ask whether reconstruction preserves contrastive image-text decisions between minimally different captions.

VL-SAE (Shen et al., 2025) maps vision and language representations into a single concept set, encouraging semantically similar image and text representations to produce consistent neuron activations during self-supervised training. The resulting concepts are used both to interpret vision-language alignment and to strengthen it at the concept level. VL-SAE is therefore close to our goal of a shared cross-modal coordinate system, but it evaluates concept-level alignment rather than preservation of compositional decisions or causal mediation by held-out feature circuits.

MGSAE (Kaushik et al., 2026) addresses the tendency of standard SAEs to learn split dictionaries in which most coordinates are active for only one modality. It combines paired group sparsity with cross-modal random masking, encouraging corresponding image and text embeddings to use common dictionary elements. Experiments on CLIP and CLAP show improvements in multimodal feature alignment, semanticity, and dictionary utilization. MGSAE is our closest baseline because it operates on aligned pooled embeddings and directly targets shared feature learning.

Our method differs from MGSAE in both its objective and evaluation. We introduce multimodal Group-Hoyer regularization and combine it with modality-specific centering, independently inferred primary codes, cross-decoding, shared latent dropout, and a variable-size positive-only auxiliary. Unlike paired support construction, our primary image and text codes remain modality-local at inference. Beyond learning shared concepts, we evaluate whether sparse reconstruction preserves compositional decisions and whether the identified features causally control those decisions.

SPARC (Nasiri-Sarvi et al., 2026) learns a common concept space across heterogeneous models and modalities. Its Global TopK mechanism encourages input streams to activate common latent dimensions, while cross-reconstruction promotes semantic consistency between their representations. This enables direct concept comparison, cross-model retrieval, and text-guided localization. SPARC is closely related in its goal of concept alignment, but its central question is whether heterogeneous models can share an interpretable coordinate system. Our focus is instead whether independently inferred sparse codes preserve the structured decisions of a fixed dual-encoder VLM. Strong support overlap alone does not establish preservation of attribute binding, relation direction, or contrastive margins.

LUCID-SAE (Gu et al., 2026) learns unified sparse codes for image-patch and text-token representations while retaining private capacity for modality-specific information. It aligns the shared components through an optimal-transport-based objective and evaluates concept discovery and visual grounding. This shared/private decomposition is complementary to our use of modality-specific centering and shared latent structure. However, LUCID-SAE studies token- and patch-level grounding, whereas we operate on global dual-encoder embeddings and ask whether their sparse reconstruction preserves compositional behavior.

The structured sparse autoencoder of Liao et al. (Liao et al., 2026) incorporates semantic and spatial grouping of image patches. Its exclusive and group-sparsity terms encourage inter-group concept separation and intra-group spatial consistency, improving semantic alignment and monosemanticity in multimodal model features. This approach uses visual spatial structure as supervision for feature organization. Our method instead targets pooled image and text embeddings, for which explicit patch-level grouping is unavailable. Moreover, its structured regularization should not be conflated with our Group-Hoyer objective: the former organizes spatial image-patch groups, whereas Group-Hoyer extends Hoyer sparsity to modality-aligned latent groups.

Cross-modal redundancy provides another perspective on multimodal dictionary structure. Dhimoïla et al. (Dhimoïla et al., 2026) introduce an aligned SAE based on an iso-energy assumption and distinguish bimodal atoms carrying alignment information from unimodal atoms associated with modality-specific structure. Their results reinforce the importance of separating semantic alignment from the modality gap. Our modality-specific centering serves a related geometric purpose, but our primary objective is not to identify all sources of the modality gap. Instead, we ask whether removing systematic modality offsets facilitates a common sparse coordinate system without discarding behaviorally relevant compositional information.

PUMA (Attimonelli et al., 2026) also sparsifies global multimodal embeddings, but for a different objective. It preserves dense dot-product geometry and subsequently fine-tunes sparse codes for efficient retrieval, reducing storage and scoring costs. Its target is retrieval fidelity and efficiency rather than semantic factorization or causal explanation. We therefore treat PUMA as related work in post-hoc multimodal sparsification rather than as a direct mechanistic-interpretability baseline.

A.4 Validity of Cross-Modal Feature Alignment

A shared embedding space need not have identical geometry across modalities. Image and text embeddings can occupy systematically separated regions even when paired samples have high relative similarity (Liang et al., 2022). Differences in means, covariance, and concentration can therefore be absorbed by dictionary atoms and mistaken for semantic specialization. Modality-specific centering removes the dominant mean offset before a shared sparse decomposition is learned, while the reconstruction objective retains the residual within-modality variation.

A separate issue concerns how sparse supports are inferred. Let x_v and x_t denote paired image and text embeddings. If an image explanation is claimed to be independently recognized from the image, its primary code must have the form

$$z_v = F_v(x_v), \qquad z_t = F_t(x_t), \tag{26}$$

rather than

$$z_v = F_v(x_v, x_t) \quad \text{or} \quad z_t = F_t(x_t, x_v). \tag{27}$$

In the latter case, the paired example can influence which coordinates are active. The resulting support may encode pair identity even when the dictionary directions themselves carry little shared semantic information. Retrieval based on such supports then measures the effectiveness of pair-conditioned code construction rather than independent cross-modal recognition. Our random-dictionary control demonstrates that this mechanism can produce near-perfect retrieval without a meaningful learned decomposition.

Fixed product-TopK selection introduces a related but distinct ambiguity. For nonnegative modality-specific activations $a^{(v)}$ and $a^{(t)}$, suppose bilateral scores are defined as

$$q_j = a_j^{(v)} a_j^{(t)}, \tag{28}$$

and exactly K coordinates are selected according to q_j. Let

$$N_+ = |\{j : q_j > 0\}| \tag{29}$$

denote the number of coordinates with strictly positive bilateral evidence. When $K > N_+$, at least $K - N_+$ selected coordinates must come from the tied set $\{j : q_j = 0\}$. Their identities are then determined by implementation-dependent tie-breaking rather than by cross-modal evidence. Although the resulting tensor is computationally well defined, the selected zero-score coordinates are not semantically identifiable from the stated scoring rule.

We therefore distinguish two validity requirements. *Modality-local inference* requires that each primary code be inferred from its own modality. *Selector-score identifiability* requires every selected coordinate to be supported by a defined selection score rather than an unresolved tie. Our variable-size positive-only auxiliary couples modalities only during training and retains only coordinates with positive bilateral evidence, avoiding the need to populate a fixed support with tied zeros.

These requirements do not prohibit paired supervision during training. They separate the training signal from the representation claimed at inference: paired data may encourage a common dictionary, but the image and text codes used for evaluation must remain independently computable.

A.5 Compositional Evaluation of Vision-Language Models

Standard image-text retrieval can often be solved through object and word presence without requiring detailed sensitivity to structure. ARO evaluates attribute assignment, relations, and word order, and shows that strong VLMs can behave similarly to bags of words (Yuksekgonul et al., 2023). Winoground uses carefully controlled image-caption sets in which the same constituent words participate in different compositions, exposing failures involving relations and semantic roles (Thrush et al., 2022).

SugarCrepe was introduced in response to linguistic shortcuts in earlier compositional benchmarks. It uses fluent hard negatives and adversarial refinement to reduce the ability of text-only systems to exploit benchmark artifacts (Hsieh et al., 2023). SugarCrepe++ extends this setting by pairing lexically different but semantically equivalent positive captions with a hard negative, separating sensitivity to surface form from sensitivity to semantic alteration (Dumpala et al., 2024). Together, these benchmarks show that successful retrieval over broad datasets does not imply sensitivity to attribute binding, relation direction, or word order.

Our use of these benchmarks differs from standard backbone evaluation. We compare decisions made from the original embeddings with those made after SAE reconstruction. For a visual embedding v, a correct caption embedding t^+, and a compositional foil t^-, the relevant quantity is the contrastive margin

$$m(v, t^+, t^-) = s(v, t^+) - s(v, t^-), \tag{30}$$

where s is the backbone similarity function. The SAE is behaviorally faithful when the reconstructed margin

$$\hat{m} = s(\hat{v}, \hat{t}^+) - s(\hat{v}, \hat{t}^-) \tag{31}$$

preserves both the sign and structure of the original decision. A small average reconstruction error does not ensure this property because the original margin may itself be small. We therefore report decision preservation, corrections, degradations, accuracy, and margin distortion rather than treating vector reconstruction as a sufficient proxy.

A.6 From Semantic Correlation to Causal Control

A feature can encode information that the model does not use. This distinction is well established in the probing literature: successful decoding demonstrates that a property is accessible to a probe, but not that it causally contributes to the model's prediction (Elazar et al., 2021). The same limitation applies to SAE feature labeling. A coordinate may activate on examples containing a relation without mediating the model's preference for the correctly bound caption.

Mechanistic interpretability therefore supplements correlational evidence with interventions. Causal tracing and activation patching replace or suppress internal states to measure their effect on a target behavior (Meng et al., 2023). Path patching refines this idea by isolating contributions transmitted through particular computational paths (Goldowsky-Dill et al., 2023), while automated circuit-discovery methods search for sparse subgraphs sufficient for a behavior (Conmy et al., 2023).

We adapt this principle from transformer components to sparse multimodal features. Because the reconstructed embedding is linear in the sparse code, its contribution to a contrastive decision can be decomposed into coordinate-level and interaction terms. We use this decomposition to identify compact candidate circuits on a discovery split and test them on held-out examples. We then intervene through feature ablation and activation exchange and compare the resulting change in the decision margin with activation-energy-matched random controls. This controls for the possibility that any sufficiently energetic set of coordinates would perturb the output.

Cross-seed functional matching provides an additional robustness test. We do not require independently trained SAEs to assign the same semantic function to the same literal index. Instead, we ask whether functionally corresponding features or circuits can be recovered through their activation profiles and behavioral effects. Reproducibility at this functional level is stronger evidence than index overlap, while remaining compatible with permutation, splitting, and merging of learned dictionary elements.

A.7 Summary of Positioning

Table 3 summarizes the closest multimodal sparse-representation methods and the principal distinction of our work.

Overall, prior work establishes that SAEs can recover interpretable structure, encourage cross-modal alignment, support grounding, and preserve retrieval-relevant geometry. Our work addresses a different unresolved question: whether a multimodal SAE preserves how the original VLM combines concepts in its decisions and whether the recovered sparse coordinates causally mediate those decisions. Methodologically, we introduce Group-Hoyer regularization as a multimodal extension of Hoyer sparsity and integrate it into a validity-aware SAE with independently inferred primary codes. Evaluatively, we distinguish concept presence from compositional faithfulness and causal control.

B Formal Foundations and Proofs

This section extends the formalism in Section 3.1 and provides proofs of the theoretical claims used in the main paper. Throughout, $\mathcal{N}(x) = x/\|x\|_2$, and all normalized vectors are assumed to be nonzero.

B.1 Compositional Margins and Bidirectional Benchmarks

For an image I, a correct caption T^+, and a minimally different foil T^-, the main paper considers the image-anchored dense and reconstructed margins

$$\Delta(e) = s(x_v, x_t^+) - s(x_v, x_t^-), \qquad \hat{\Delta}(e) = s(\hat{x}_v, \hat{x}_t^+) - s(\hat{x}_v, \hat{x}_t^-), \tag{32}$$

where

$$s(x_v, x_t) = \mathcal{N}(x_v)^\top \mathcal{N}(x_t). \tag{33}$$

This form directly covers caption-selection benchmarks such as SugarCrepe and SugarCrepe++.

Benchmarks such as Winoground evaluate a quadruple (I^+, I^-, T^+, T^-) through multiple comparisons. The two image-anchored margins are

$$\Delta_{I^+} = s(x_v^+, x_t^+) - s(x_v^+, x_t^-), \tag{34}$$

$$\Delta_{I^-} = s(x_v^-, x_t^-) - s(x_v^-, x_t^+), \tag{35}$$

and the two text-anchored margins are

$$\Delta_{T^+} = s(x_v^+, x_t^+) - s(x_v^-, x_t^+), \tag{36}$$

$$\Delta_{T^-} = s(x_v^-, x_t^-) - s(x_v^+, x_t^-). \tag{37}$$

Their reconstructed counterparts are obtained by replacing every dense embedding with its SAE reconstruction.

We apply the definitions of preservation, correction, degradation, and margin distortion separately to each required comparison. An image score is preserved when the signs of both image-anchored

Table 3: **Positioning relative to closely related multimodal sparse-representation methods.** The methods are complementary but differ in their representation level, alignment mechanism, evaluation target, and the evidence required to support an interpretability claim.

Method	Representation	Primary objective	Relation to our work
Papadimitriou et al.	VLM embedding spaces	Characterize sparse concepts, modality structure, bridges, and stability	Provides evidence for sparse linear structure in VLM embeddings; does not evaluate preservation or causal mediation of compositional decisions.
MGSAE	Pooled CLIP and CLAP embeddings	Reduce modality-split dictionaries using group sparsity and cross-modal masking	Closest matched baseline for learning shared concepts; our method introduces Group-Hoyer and adds validity-aware independent inference, compositional preservation, and causal feature circuits.
SPARC	Representations from heterogeneous models and modalities	Learn a unified concept space through Global TopK and cross-reconstruction	Targets concept alignment and transfer across models; our work targets behavior preservation within fixed dual-encoder VLMs.
LUCID-SAE	Image-patch and text-token representations	Learn shared sparse codes with modality-private capacity and grounding	Studies local grounding and shared/private structure; our analysis concerns global embeddings and contrastive compositional decisions.
Structured SAE	Multimodal model features with image-patch structure	Improve semantic and spatial consistency through structured sparsity	Uses patch grouping as a visual structural prior; our Group-Hoyer objective instead extends Hoyer sparsity to modality-aligned latent groups in pooled dual-encoder embeddings.
PUMA	Universal multimodal embeddings	Efficient sparse retrieval with preserved dot-product geometry	Shares post-hoc sparsification of global embeddings, but optimizes retrieval efficiency rather than mechanistic or causal interpretation.
Ours	Global CLIP, SigLIP, and SigLIP2 embeddings	Preserve and causally expose compositional decisions under valid independent inference	Introduces multimodal Group-Hoyer regularization and distinguishes concept presence, compositional faithfulness, and causal control through validity tests, margin decomposition, interventions, and cross-seed functional matching.

margins are preserved, a text score is preserved when the signs of both text-anchored margins are preserved, and a group decision is preserved when all four required margin signs are preserved. This extension preserves the distinction between behavioral fidelity to the dense backbone and benchmark correctness.

Examples with an exactly zero dense or reconstructed margin are excluded from sign-based preservation metrics and reported separately as ties. In practice, the same fixed numerical tolerance is applied to all compared methods.

B.2 Proof of Compositional Margin Stability

Proof of Proposition 2. For an example $e = (I, T^+, T^-)$, write

$$a = \mathcal{N}(x_v), \qquad \hat{a} = \mathcal{N}(\hat{x}_v), \qquad b^{\pm} = \mathcal{N}(x_t^{\pm}), \qquad \hat{b}^{\pm} = \mathcal{N}(\hat{x}_t^{\pm}), \tag{38}$$

and define

$$u = b^+ - b^-, \qquad \hat{u} = \hat{b}^+ - \hat{b}^-. \tag{39}$$

The dense and reconstructed margins can then be written as

$$\Delta(e) = a^\top u, \qquad \hat{\Delta}(e) = \hat{a}^\top \hat{u}. \tag{40}$$

Adding and subtracting $a^\top \hat{u}$ gives the exact identity

$$\hat{\Delta}(e) - \Delta(e) = (\hat{a} - a)^\top \hat{u} + a^\top (\hat{u} - u). \tag{41}$$

Applying the triangle inequality and Cauchy-Schwarz to Equation 41, and using $\|a\|_2 = 1$, yields

$$\begin{aligned} \left|\hat{\Delta}(e) - \Delta(e)\right| &\le \|\hat{a} - a\|_2 \|\hat{u}\|_2 + \|\hat{u} - u\|_2 \\ &= \hat{\rho}\, \delta_v + \|\hat{u} - u\|_2, \end{aligned} \tag{42}$$

where

$$\hat{\rho} = \|\hat{u}\|_2 = \left\|\mathcal{N}(\hat{x}_t^+) - \mathcal{N}(\hat{x}_t^-)\right\|_2. \tag{43}$$

Because

$$\hat{u} - u = (\hat{b}^+ - b^+) - (\hat{b}^- - b^-), \tag{44}$$

another application of the triangle inequality gives

$$\|\hat{u} - u\|_2 \le \delta_+ + \delta_-. \tag{45}$$

Therefore,

$$\left|\hat{\Delta}(e) - \Delta(e)\right| \le \hat{\rho}\, \delta_v + \delta_+ + \delta_-. \tag{46}$$

Since $\hat{b}^+$ and $\hat{b}^-$ are unit vectors,

$$\hat{\rho} = \|\hat{b}^+ - \hat{b}^-\|_2 \le \|\hat{b}^+\|_2 + \|\hat{b}^-\|_2 = 2, \tag{47}$$

which establishes

$$\left|\hat{\Delta}(e) - \Delta(e)\right| \le \hat{\rho}\, \delta_v + \delta_+ + \delta_- \le 2\delta_v + \delta_+ + \delta_-. \tag{48}$$

Finally, suppose

$$|\Delta(e)| > \hat{\rho}\, \delta_v + \delta_+ + \delta_-. \tag{49}$$

The preceding bound implies

$$|\hat{\Delta}(e) - \Delta(e)| < |\Delta(e)|. \tag{50}$$

Consequently, the reconstruction error is too small to move the margin across zero, and $\operatorname{sign}(\hat{\Delta}(e)) = \operatorname{sign}(\Delta(e))$. □

Remark 1. *The factor $\hat{\rho}$ is the separation between the two reconstructed captions and is directly measurable at evaluation time. Since $\hat{\rho} \le 2$, the first bound is never weaker than treating the two similarity terms independently. It can be substantially tighter for minimally edited foils, for which $\hat{\rho} \ll 2$.*

The result applies to every margin defined in Appendix B.1. Consequently, preservation of a grouped benchmark decision can be certified by checking the corresponding condition for each constituent comparison.

B.3 Proof of Exact Behavioral Accounting

Proof of Proposition 1. Let

$$\mathcal{E}_0 = \left\{ e \in \mathcal{E} : \Delta(e) \neq 0,\ \hat{\Delta}(e) \neq 0 \right\}. \tag{51}$$

For every $e \in \mathcal{E}_0$, exactly one of the following mutually exclusive cases holds:

1. $\Delta(e)$ and $\hat{\Delta}(e)$ have the same sign,
2. $\Delta(e) < 0 < \hat{\Delta}(e)$, corresponding to a correction, or
3. $\hat{\Delta}(e) < 0 < \Delta(e)$, corresponding to a degradation.

These cases are exhaustive. Averaging their indicators over $\mathcal{E}_0$ therefore gives

$$\mathrm{DP} = 1 - \mathrm{Corr} - \mathrm{Deg}\,. \tag{52}$$

For the accuracy difference, observe that

$$\mathbf{1}[\hat{\Delta}(e) > 0] - \mathbf{1}[\Delta(e) > 0] = \begin{cases} +1, & \Delta(e) < 0 < \hat{\Delta}(e), \\ -1, & \hat{\Delta}(e) < 0 < \Delta(e), \\ 0, & \text{otherwise.} \end{cases} \tag{53}$$

Averaging over $\mathcal{E}_0$ yields

$$A_{\mathrm{SAE}} - A_{\mathrm{dense}} = \mathrm{Corr} - \mathrm{Deg}\,. \tag{54}$$

□

This identity explains why reconstructed benchmark accuracy is not a complete measure of behavioral faithfulness. Accuracy reports only the net balance between corrections and degradations, whereas decision preservation reports the total amount of behavior retained from the dense backbone.

B.4 Proof of Pair-Identity Leakage

Proof of Proposition 3. Let

$$\mathcal{D} = \left\{ (x_v^{(i)}, x_t^{(i)}) \right\}_{i=1}^{M}, \tag{55}$$

and suppose that the latent width p and support size K satisfy

$$1 \le K < p, \qquad \binom{p}{K} \ge M. \tag{56}$$

There are therefore at least M distinct K-element subsets of $\{1, \dots, p\}$. Assign each pair i a distinct support S_i and define

$$z_v^{(i)} = z_t^{(i)} = \mathbf{1}_{S_i}. \tag{57}$$

For a matched image-text pair,

$$\cos\left(z_v^{(i)}, z_t^{(i)}\right) = 1. \tag{58}$$

For $i \neq j$, the supports are distinct and have the same cardinality, so $|S_i \cap S_j| < K$. Hence,

$$\cos\left(z_v^{(i)}, z_t^{(j)}\right) = \frac{|S_i \cap S_j|}{K} < 1. \tag{59}$$

Every matched similarity strictly exceeds every mismatched similarity involving the same query. The construction therefore achieves perfect image-to-text and text-to-image rank-one retrieval. It also gives matched support overlap equal to one and strictly smaller overlap for every mismatched pair.

The assignment $i \mapsto S_i$ depends only on pair identity and does not use the semantic content of either representation. Thus, perfect retrieval and support overlap do not establish semantic alignment when the known pair is allowed to construct the evaluated codes. □

Remark 2. *The combinatorial condition is weak. When $K = 1$, it requires only $p \ge M$, and all mismatched codes are orthogonal. The construction is a worst-case identifiability result: it shows that retrieval cannot certify semantic alignment under pair-conditioned inference, not that every pair-conditioned method necessarily learns an explicit identity code.*

B.5 Proof of Fixed Product-TopK Indeterminacy

Proof of Proposition 4. Let

$$q_j = z_{v,j} z_{t,j} \geq 0, \qquad N_+ = \#\{j : q_j > 0\}, \tag{60}$$

and define the zero-score set

$$Z_0 = \{j : q_j = 0\}. \tag{61}$$

Then $|Z_0| = p - N_+$. Because all scores are nonnegative, a size-K TopK support must select all N_+ strictly positive coordinates before selecting any coordinate from Z_0.

Suppose

$$N_+ < K < p. \tag{62}$$

The selector must then include exactly $K - N_+$ coordinates from Z_0. Moreover,

$$K - N_+ < p - N_+ = |Z_0|, \tag{63}$$

so at least one zero-score coordinate is selected and at least one is not. Consequently,

$$q_{(K)} = q_{(K+1)} = 0, \tag{64}$$

where $q_{(1)} \geq \cdots \geq q_{(p)}$ are the ordered scores.

Let $j \in Z_0$ be selected and $k \in Z_0$ be unselected. Exchanging j and k produces a different size-K support with exactly the same multiset of selected scores. Both supports are therefore equally compatible with the bilateral product criterion. The selected support is not uniquely identified by its stated score and depends on an additional tie-breaking convention. □

The affected fraction of the selected support is

$$\rho_{\text{zero}} = \frac{(K - N_+)_+}{K}. \tag{65}$$

It is zero whenever $K \leq N_+$ and positive whenever fixed product-TopK must fill part of its support from tied zero scores. A deterministic implementation may repeatedly return the same indices, but this reproducibility reflects its tie-breaking rule rather than positive cross-modal evidence.

B.6 Implications for the Experimental Analysis

The results above establish three principles used throughout the experiments. First, reconstruction fidelity cannot certify preservation of a compositional decision without considering the original margin. Second, reconstructed accuracy must be decomposed into preserved, corrected, and degraded decisions. Third, retrieval and support overlap are semantically interpretable only when the evaluated codes are inferred modality-locally and their selected coordinates are score-identifiable.

The detailed empirical audits of these validity conditions are provided in Appendix G. Exact margin attribution and the causal intervention protocol are developed separately in Appendix D.

C Training Details

This appendix gives the complete objective and optimization protocol summarized in Section 4. Expectations are over matched pairs $(x_v, x_t) \sim \mathcal{D}_{\text{train}}$. We use the per-example reconstruction loss

$$\ell(x, \hat{x}) = \frac{1}{d}\|x - \hat{x}\|_2^2. \tag{66}$$

The image and text centering vectors μ_v and μ_t are initialized to their respective training-set means and remain learnable. The primary codes z_v and z_t are inferred independently with the shared affine encoder and ReLU. No paired input or paired support is used by the inference map. Table 4 summarizes the coefficients and activation status of every objective term in the selected checkpoint.

Table 4: **Objective terms in the canonical checkpoint.** The total optimized loss is $\mathcal{L}_{self} + 0.1\mathcal{L}_{+} + 2.5 \times 10^{-4}\mathcal{L}_{GH} + 0.1\mathcal{L}_{cross}$. Masking is an operator, not an additive loss.

Term	Weight	Status
Self reconstruction $\mathcal{L}_{self}$	1	Active; masked dense codes
Positive-only paired reconstruction $\mathcal{L}_{+}$	0.1	Active; support $\{j : z_{v,j} > 0 \wedge z_{t,j} > 0\}$
Group-Hoyer $\mathcal{L}_{GH}$	2.5×10^{-4}	Active; same-index group magnitudes
Cross-decoding $\mathcal{L}_{cross}$	0.1	Active; independently inferred code
Shared latent masking	$p_{mask} = 0.05$	Active stochastic operator; no inverted rescaling
Internal balance / mismatch	0 / 0	Inactive in the final checkpoint
Code/support alignment	0 / 0	Inactive (mode off)
Contrastive / selected balance	0 / 0	Inactive
Decoder alignment / orthogonality	0 / 0	Inactive for the shared decoder
Group-$L_{1,\infty}$	0	Inactive

C.1 Self-Reconstruction and Cross-Decoding

The primary term reconstructs each modality from its own independently inferred code,

$$\mathcal{L}_{\text{self}} = \mathbb{E}[\ell(x_v, Dz_v + \mu_v) + \ell(x_t, Dz_t + \mu_t)]\,. \tag{67}$$

Cross-decoding decodes each code against the opposite modality's target while substituting the target modality's center,

$$\hat{x}_{t\leftarrow v} = Dz_v + \mu_t, \qquad \hat{x}_{v\leftarrow t} = Dz_t + \mu_v, \tag{68}$$

$$\mathcal{L}_{\text{cross}} = \mathbb{E}[\ell(x_t, \hat{x}_{t\leftarrow v}) + \ell(x_v, \hat{x}_{v\leftarrow t})]\,. \tag{69}$$

The opposite-modality target appears only in the training loss. It does not alter the modality-local inference map in Eq. 9.

Remark 3 (Cross-decoding is a regularizer)**.** *Cross-decoding encourages each modality-local code to retain information useful for reconstructing the other modality. It is therefore weighted rather than used as the primary reconstruction objective. In the completed one-seed tuning slice, increasing* λ_{cross} *from 0 to 0.2 increases automatic broad-unit coverage while worsening FVU and GPI. This is an observed fidelity-coverage trade-off; the tested range does not identify a collapse threshold.*

C.2 Strictly-Positive Paired Auxiliary

The primary reconstruction path remains dense and independently inferred. A second, training-only path identifies coordinates that are strictly positive in both codes,

$$S_{+}(x_v, x_t) = \{j : z_{v,j} > 0 \ \wedge \ z_{t,j} > 0\}\,, \qquad m_j^{+} = \mathbf{1}[j \in S_{+}], \tag{70}$$

and forms $z_v^{+} = m^{+} \odot z_v$ and $z_t^{+} = m^{+} \odot z_t$. The auxiliary loss is

$$\mathcal{L}_{+} = \mathbb{E}\big[\ell(x_v, Dz_v^{+} + \mu_v) + \ell(x_t, Dz_t^{+} + \mu_t)\big]\,. \tag{71}$$

There is no ranking operation and no prescribed support size. Every strictly positive bilateral product is retained and every zero product is excluded. Consequently, the auxiliary cannot fill a fixed budget with arbitrary zero-score coordinates.

C.3 Group-Hoyer Concentration

To concentrate paired activation energy in a smaller effective support, we form the same-index group magnitudes

$$g_j = \left(z_{v,j}^2 + z_{t,j}^2 + \epsilon\right)^{1/2} \tag{72}$$

and apply the normalized Hoyer score

$$\text{Hoyer}(g) = \text{clip}_{[0,1]}\left(\frac{\sqrt{p} - \|g\|_1 / \max(\|g\|_2, \epsilon)}{\sqrt{p} - 1}\right), \qquad \mathcal{L}_{\text{group}} = \mathbb{E}\left[1 - \text{Hoyer}(g)\right]. \tag{73}$$

Table 5: **Implementation concordance for statements that are easy to misread from the conceptual objective.** This table describes the executed trainer and final checkpoint.

Item	Executed Behavior	Consequence
Mask scaling	$\tilde{z} = r \odot z$	The trainer does not divide by $1-p_{mask}$.
Mask scope	self, cross, Hoyer, and paired auxiliary	The shared mask is applied before all active loss paths.
Balance/mismatch	weights 0/0	No margin γ is operative in the final checkpoint.
Positive auxiliary	all strictly bilateral-positive coordinates	Variable support; no ranking, cap, or zero-score fill.
Primary code	dense ReLU	Metadata $K = 768$ does not truncate the primary forward pass.

The score approaches one when energy is concentrated in few groups and zero when it is spread uniformly. Unlike hard TopK, this loss does not prescribe an exact number of nonzero activations. The dense ReLU code can therefore contain many small positive coordinates while concentrating most of its energy in a smaller participation-ratio support.

C.4 Inactive Balance and Mismatch Terms

Pair-level magnitude-balance and mismatch penalties were explored during model development, but the selected checkpoint sets both coefficients to zero:

$$\lambda_{\text{bal}} = \lambda_{\text{mis}} = 0. \tag{74}$$

They are not part of the reported final objective, and no mismatch margin γ is operative. We report this negative design decision explicitly rather than presenting an inactive term as part of the method.

C.5 Shared Latent Masking

For each paired training example, we draw $r_j \sim \text{Bernoulli}(1 - p_{\text{mask}})$ and apply the same mask to the two independently inferred codes,

$$\widetilde{z}_m = r \odot z_m, \qquad m \in \{v, t\}. \tag{75}$$

The executed trainer does not divide by $1 - p_{\text{mask}}$. Masks are independent across paired examples. The masked codes enter self-reconstruction, cross-decoding, and Group-Hoyer, and the same availability mask restricts the strictly-positive paired auxiliary. Masking is disabled at evaluation. Table 5 records these implementation details together with the interpretation of the nominal K and inactive alignment terms.

C.6 Complete Objective

The final optimized objective is

$$\mathcal{L} = \mathcal{L}_{\text{self}} + 0.1\,\mathcal{L}_{+} + 2.5 \times 10^{-4}\,\mathcal{L}_{\text{group}} + 0.1\,\mathcal{L}_{\text{cross}}. \tag{76}$$

Shared latent masking is a stochastic training operator rather than an additive loss term. Code alignment, support alignment, contrastive alignment, selected magnitude balance, decoder alignment, decoder orthogonality, and Group-$L_{1,\infty}$ are all inactive in the selected checkpoint.

C.7 Auxiliary Support Statistics

Because $K_+ = |S_+|$ is data-dependent rather than prescribed, we audit its distribution across validation pairs. Figure 2 shows the full distribution and its training trajectory, while Table 6 reports the exact seedwise summary statistics.

The pooled final mean is $K_+ = 1667.6$, and none of the 15,000 seed-example evaluations falls below 768. This shows that the auxiliary uses a large data-dependent support in this checkpoint. Its selector validity does not depend on being smaller than 768: it follows from retaining all and only strictly positive bilateral products, so the selected zero-score fraction is identically zero.

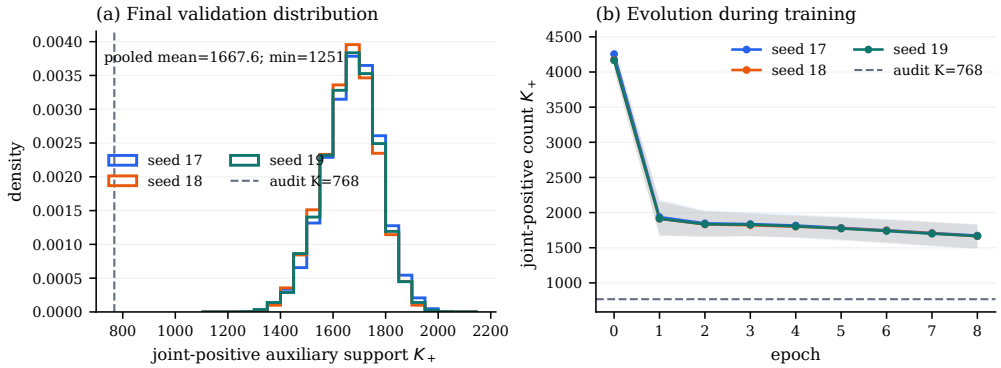

Figure 2: **Support of the strictly-positive paired auxiliary for the canonical SigLIP2 checkpoint.** Left: exact final validation distributions over 5,000 pairs for seeds 17-19. Right: mean support during training; translucent bands span the per-seed 5th-95th percentiles. The dashed $K = 768$ line is an audit reference inherited from the run metadata, not a primary or auxiliary selection budget.

Table 6: **Validation distribution of the variable-size strictly-positive paired auxiliary support.** K_+ counts coordinates for which both independently inferred ReLU activations are positive. The nominal $K = 768$ is an audit reference inherited by the run metadata; it is not a primary or auxiliary support cap.

Seed	Mean ± SD	Min	P05	Median	P95	Max	$\Pr(K_+ < 768)$
17	1673.0 ± 105.0	1251	1494	1678	1838	1995	0.0000
18	1663.8 ± 102.3	1315	1485	1669	1825	2020	0.0000
19	1666.1 ± 104.4	1265	1485	1672	1830	2002	0.0000
Pooled	1667.6 ± 104.0	1251	1487	1673	1832	2020	0.0000

C.8 Optimization and Hyperparameters

Table 7 gives the complete optimization, data, architecture, and evaluation settings for the canonical three-seed SigLIP2 experiment.

The cross-backbone experiments freeze separate operating configurations before compositional evaluation. Table 8 lists their widths, active support constraints, learning rates, masking probabilities, and cross-modal loss weights.

Figure 3 shows the corresponding optimization trajectory and frozen-validation preservation across epochs.

C.9 Training-Weight Sensitivity

The available sensitivity experiment changes one coefficient at a time in the earlier seed-17 width-10,240 SigLIP2 reference with $\lambda_{\text{group}} = 5 \times 10^{-4}$. It is a model-selection diagnostic, not a three-seed uncertainty estimate for the final checkpoint. Figure 4 visualizes the main trade-offs, and Table 9 provides the complete numerical results.

Across the tested range, stronger cross-decoding increases broad-unit coverage from 138 to 198 while FVU worsens from 0.0371 to 0.0587. Increasing the positive-only auxiliary from 0 to 0.2 changes FVU from 0.0433 to 0.0465 and broad-unit coverage from 188 to 171. These results justify treating both terms as weighted regularizers, but they do not establish a collapse threshold.

D Exact Margin Decomposition and Causal Analysis Protocol

This section derives the exact feature-level decomposition of reconstructed compositional margins and specifies the protocol used to discover and intervene on sparse circuits. The frozen five-seed causal experiment uses the earlier dense tie-free shared-decoder endpoint at width 10,240 and $\lambda_{\text{group}} = 5 \times 10^{-4}$; it is not the later 2.5×10^{-4} three-seed canonical checkpoint or a two-decoder

Table 7: **Optimization and implementation details for the canonical three-seed SigLIP2 checkpoint.** Unlogged hardware and timing fields are reported as unavailable rather than reconstructed from the current machine.

Setting	Frozen Value
Training data	COCO paired pooled embeddings (118,287 train, 5,000 validation)
Backbone / input dimension	SigLIP2 ViT-B/16 / 768
Dictionary	shared affine encoder, shared unit-norm decoder, width 10,240
Centering / activation	mean-initialized, learnable modality-specific biases / ReLU
Primary inference	dense and modality-local, no primary TopK
Nominal K	768 (metadata/audit reference only)
Optimizer	AdamW
Learning rate / schedule	5×10^{-4} / constant (no scheduler)
Weight decay / gradient clip	10^{-6} / global norm 1.0
Batch / epochs	1,024 / 8 (116 updates per epoch, 928 total)
Loader	shuffled each epoch, 2 workers, final partial batch retained
Precision	float32 embeddings and model training
Seeds / checkpoint	17, 18, 19 / final epoch
Evaluation	dense modality-local codes, semantic audit uses post-hoc independent absolute TopK-64
Hardware	CUDA requested, accelerator model was not stored in the run manifest
Wall-clock	not instrumented in the frozen run

Table 8: **Frozen per-backbone operating configurations used in the cross-backbone experiments.** All use modality-specific centering, a shared encoder and decoder, Group-Hoyer weight 2.5×10^{-4}, and eight epochs. Dense rows have no operative primary K.

Backbone	Family	Width	Operative K	LR	p_{mask}	λ_{cross}	λ_{+}
CLIP ViT-B/16	dense tie-free	10,240	–	0.00075	0.05	0.10	0.05
CLIP ViT-B/16	independent TopK	20,480	256	0.00050	0.10	0.10	0.10
SigLIP ViT-B/16	dense tie-free	20,480	–	0.00075	0.05	0.10	0.10
SigLIP ViT-B/16	independent TopK	20,480	256	0.00050	0.05	0.20	0.10
SigLIP2 ViT-B/16	dense tie-free	10,240	–	0.00050	0.05	0.20	0.10
SigLIP2 ViT-B/16	independent TopK	10,240	256	0.00050	0.05	0.20	0.10

run. All image and caption codes are inferred independently. Circuit coordinates are selected on a deterministic discovery half of each benchmark and evaluated on its disjoint complement. The complete executed protocol is summarized in Table 10 and Figure 5.

D.1 Caption-Side Margin Decomposition

Consider an example $e = (I, T^+, T^-)$. Let

$$u_v = \frac{\hat{x}_v}{\|\hat{x}_v\|_2}, \qquad r_+ = \|\hat{x}_{t,+}\|_2, \qquad r_- = \|\hat{x}_{t,-}\|_2, \tag{77}$$

where the reconstructed captions are

$$\hat{x}_{t,+} = Dz_{t,+} + \mu_t, \qquad \hat{x}_{t,-} = Dz_{t,-} + \mu_t. \tag{78}$$

The reconstructed compositional margin is

$$\begin{aligned} \hat{\Delta}(e) &= u_v^\top \left(\frac{\hat{x}_{t,+}}{r_+} - \frac{\hat{x}_{t,-}}{r_-} \right) \\ &= u_v^\top \left(\frac{Dz_{t,+} + \mu_t}{r_+} - \frac{Dz_{t,-} + \mu_t}{r_-} \right). \end{aligned} \tag{79}$$

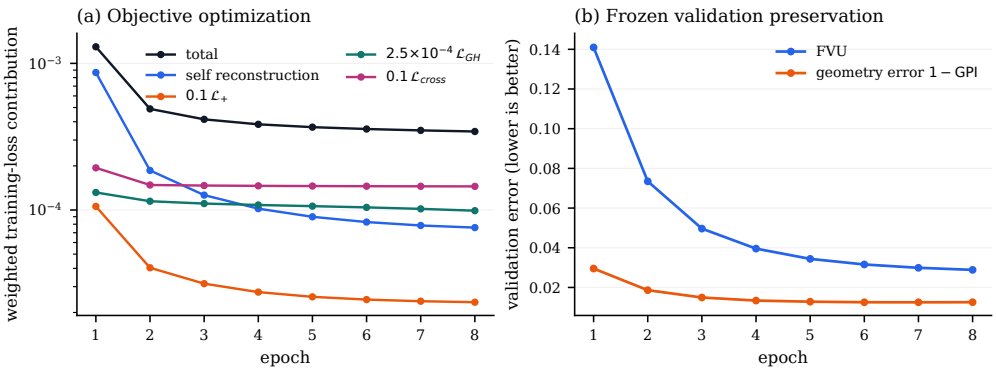

Figure 3: **Optimization of the canonical SigLIP2 objective over eight epochs.** Curves and bands show mean and SD over seeds 17-19. Left: actual weighted contributions to the optimized loss, with their total in black. Right: validation FVU and geometry error $1 - \mathrm{GPI}$.

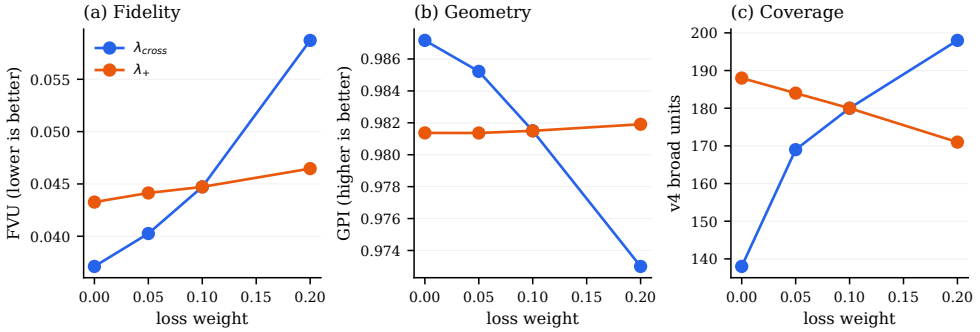

Figure 4: **One-at-a-time sensitivity to cross-decoding and positive-only paired-auxiliary weights.** The other coefficient remains fixed at 0.1. Cross-decoding exposes a clear fidelity-coverage trade-off, the paired-auxiliary sweep is comparatively flat.

Expanding the decoder column-wise gives

$$\hat{\Delta}(e) = \sum_{j=1}^{p} u_v^\top D_j \left(\frac{z_{t,+,j}}{r_+} - \frac{z_{t,-,j}}{r_-} \right) + u_v^\top \mu_t \left(\frac{1}{r_+} - \frac{1}{r_-} \right). \tag{80}$$

We therefore define the contribution of coordinate j as

$$c_j(e) = u_v^\top D_j \left(\frac{z_{t,+,j}}{r_+} - \frac{z_{t,-,j}}{r_-} \right), \tag{81}$$

and the contribution of the text center as

$$c_\mu(e) = u_v^\top \mu_t \left(\frac{1}{r_+} - \frac{1}{r_-} \right). \tag{82}$$

The margin then decomposes exactly as

$$\hat{\Delta}(e) = \sum_{j=1}^{p} c_j(e) + c_\mu(e). \tag{83}$$

No first-order approximation is involved: the normalization factors are those of the observed forward pass, and the identity follows from the linear decoder and linearity of the inner product. Table 11 verifies this identity numerically on every audited seed and benchmark.

Each coordinate contribution factors into

$$c_j(e) = \underbrace{u_v^\top D_j}_{\text{decoder alignment}} \underbrace{\left(\frac{z_{t,+,j}}{r_+} - \frac{z_{t,-,j}}{r_-} \right)}_{\text{normalized activation contrast}}. \tag{84}$$

Table 9: **One-seed training-weight sensitivity on the earlier width-10,240** SigLIP2 reference ($\lambda_{GH} = 5 \times 10^{-4}$). One coefficient changes at a time; the other remains 0.1. This diagnostic supports a fidelity–coverage trade-off over the tested range, not a demonstrated collapse threshold.

Sweep	Weight	FVU↓	GPI↑	Decoded R@1↑	PR-L0	Emitted	Dead↓	Strict / Broad / Unique
λ_{cross}	0.00	0.0371	0.9872	0.4782	578.7	1735.2	0.307	7 / 138 / 7
	0.05	0.0403	0.9852	0.4739	584.1	1664.7	0.308	11 / 169 / 9
	0.10	0.0447	0.9815	0.4751	584.5	1646.9	0.320	16 / 180 / 9
	0.20	0.0587	0.9730	0.4611	565.3	1564.3	0.304	19 / 198 / 11
λ_{+}	0.00	0.0433	0.9814	0.4728	565.5	1775.0	0.323	13 / 188 / 9
	0.05	0.0441	0.9814	0.4731	568.0	1710.6	0.319	13 / 184 / 10
	0.10	0.0447	0.9815	0.4751	584.5	1646.9	0.320	16 / 180 / 9
	0.20	0.0465	0.9819	0.4685	588.6	1535.8	0.305	16 / 171 / 10

Table 10: **Executed causal-analysis protocol.** The table records the frozen implementation used for the reported top-32 circuit results; it intentionally distinguishes completed rank matching from the unexecuted cross-seed intervention-transfer experiment.

Stage	Executed Implementation	Validity Boundary
Model endpoint	SigLIP2 ViT-B/16, COCO, width 10,240, shared encoder/decoder, dense ReLU; $\lambda_{\text{group}} = 5 \times 10^{-4}$, $\lambda_{+} = \lambda_{\text{cross}} = 0.1$; seeds 17–21.	This is the frozen five-seed causal endpoint, not the later 2.5×10^{-4} three-seed canonical checkpoint or a two-decoder model.
Inference	Image and caption codes are inferred independently; the audited tie-free model uses dense ReLU codes.	No paired support is available to an evaluated query.
Split	Two deterministic halves from SHA-256(example ID) parity: $\mathcal{E}_{\text{disc}}$ and $\mathcal{E}_{\text{eval}}$.	Discovery coordinates never use evaluation examples.
Discovery	Rank by discovery mean $\lvert c_j(e)\rvert$; freeze the top 32 coordinates separately for each benchmark.	Only the frozen circuit is intervened on evaluation examples.
Controls	8 size- and local-activity-matched random circuits; each selected coordinate is matched in $\log(1 + E_j)$ among unused coordinates.	Same circuit size and comparable activation level.
Interventions	Zero ablation and positive/negative caption-activation exchange; decode and renormalize after each edit.	Pre-intervention norms are never reused.
Seed stability	Seed-17 top-32 text-decoder atoms are matched one-to-one to each target dictionary by cosine similarity and Hungarian assignment.	Reports rank recovery, not transferred causal effects.

Activation magnitude alone is therefore not a behavioral attribution. A strongly active feature contributes little when its decoder direction is nearly orthogonal to the normalized image representation, whereas a smaller activation contrast can matter strongly when its decoder direction is aligned with the decision. For a modality-specific-decoder SAE, the same derivation holds after replacing D by the text decoder D_t; the frozen results in Appendix J use the shared-decoder model.

D.2 Image-Side Margin Decomposition

The same reconstructed margin can be decomposed over image coordinates. Define

$$r_v = \|\hat{x}_v\|_2, \qquad w_t = \frac{\hat{x}_{t,+}}{r_+} - \frac{\hat{x}_{t,-}}{r_-}. \tag{85}$$

Using $\hat{x}_v = Dz_v + \mu_v$,

$$\begin{aligned} \hat{\Delta}(e) &= \frac{(Dz_v + \mu_v)^\top w_t}{r_v} \\ &= \sum_{j=1}^{p} \frac{z_{v,j} D_j^\top w_t}{r_v} + \frac{\mu_v^\top w_t}{r_v}. \end{aligned} \tag{86}$$

Thus

$$c_j^{(v)}(e) = \frac{z_{v,j} D_j^\top w_t}{r_v}, \qquad c_\mu^{(v)}(e) = \frac{\mu_v^\top w_t}{r_v}, \tag{87}$$

with

$$\hat{\Delta}(e) = \sum_{j=1}^{p} c_j^{(v)}(e) + c_\mu^{(v)}(e). \tag{88}$$

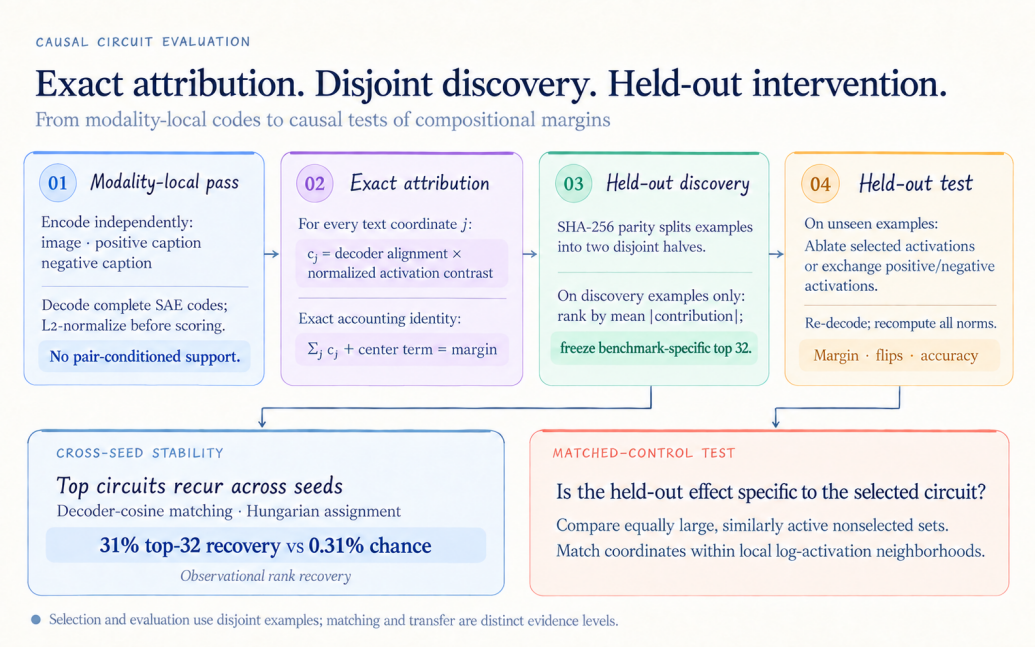

Figure 5: Exact-attribution and causal-analysis pipeline. Image, positive caption, and negative caption are encoded independently. The reconstructed margin is decomposed exactly over caption-side coordinates; coordinates are ranked only on the deterministic discovery half; and the frozen top-32 circuit is ablated or activation-exchanged on the held-out evaluation half. Eight non-selected circuits matched for size and local activation level form the control distribution. Cross-seed decoder matching measures rank recovery only, mapped target-seed interventions were not executed.

Table 11: **Numerical verification of the exact caption-side margin decomposition.** The last column is the range across SAE seeds of the largest per-example absolute residual $\left|\sum_j c_j(e) + c_\mu(e) - \hat{\Delta}(e)\right|$. Residuals below 10^{-7} are consistent with single-precision floating-point arithmetic.

Benchmark	N	Discovery	Evaluation	Seeds	Maximum-residual range
SugarCrepe	7,511	3,696	3,815	5	7.08×10^{-8}–8.94×10^{-8}
ARO	52,685	26,217	26,468	5	5.68×10^{-8}–6.38×10^{-8}

The main analysis uses the caption-side decomposition because SugarCrepe and ARO hold the image fixed while contrasting two captions. The image-side form is the corresponding exact attribution when the evaluated contrast is defined on the image side. For modality-specific decoders, D_v replaces D in this image-side expression.

D.3 Why We Use a One-Sided Decomposition

Expanding the image and caption codes simultaneously produces bilinear interactions:

$$\begin{aligned}\hat{x}_v^\top \hat{x}_t &= (Dz_v + \mu_v)^\top (Dz_t + \mu_t) \\ &= \sum_{j=1}^{p}\sum_{k=1}^{p} z_{v,j} z_{t,k} D_j^\top D_k \\ &\quad + \sum_{j=1}^{p} z_{v,j} D_j^\top \mu_t + \sum_{k=1}^{p} z_{t,k} \mu_v^\top D_k + \mu_v^\top \mu_t. \end{aligned} \tag{89}$$

The off-diagonal terms indexed by (j, k) are interactions between different image and text coordinates. Assigning each term uniquely to only j or only k requires an additional attribution convention.

The one-sided decomposition avoids that ambiguity by treating the normalized image representation as fixed while decomposing the caption side. This is exact for the observed margin, but its normalization constants are not held fixed under intervention. We therefore decode the complete edited code and recompute every embedding norm after each ablation or activation exchange.

D.4 Discovery and Evaluation Split

The executed protocol uses two, not three, deterministic subsets:

$$\mathcal{E} = \mathcal{E}_{\text{disc}} \dot{\cup} \mathcal{E}_{\text{eval}}. \tag{90}$$

An example enters the discovery half when the first byte of $\mathrm{SHA256}(\text{example ID})$ is even, the evaluation half is its complement. The rule is deterministic, approximately balanced, and independent of model outputs. It yields 3,696 discovery and 3,815 evaluation examples for SugarCrepe, and 26,217 discovery and 26,468 evaluation examples for ARO (Table 11).

$\mathcal{E}_{\text{disc}}$ is used to compute contribution rankings and the activation statistic used by the random-control sampler. $\mathcal{E}_{\text{eval}}$ is used only after the selected circuit and all controls are frozen. The reported run discovers one circuit per benchmark; perturbation-category summaries are descriptive and are not used to select a different circuit for each category. Cross-seed matching uses decoder geometry and therefore does not introduce a third example split.

D.5 Circuit Discovery and Held-Out Concentration

For coordinate j, define its discovery score as

$$a_j = \frac{1}{|\mathcal{E}_{\text{disc}}|} \sum_{e \in \mathcal{E}_{\text{disc}}} |c_j(e)|. \tag{91}$$

The reported circuit has the single preregistered size $L = 32$:

$$C_{32} = \mathrm{Top32}_j(a_j). \tag{92}$$

Absolute contributions prevent effects with opposite signs on different examples from cancelling during ranking. Once selected, C_{32} is fixed for all held-out interventions. We use *circuit* operationally for this compact set of SAE coordinates with large contributions to one reconstructed decision family. The term does not imply a contiguous transformer pathway or a symbolic program.

The held-out concentration statistic is

$$R_{32} = \frac{\sum_{e \in \mathcal{E}_{\text{eval}}} \sum_{j \in C_{32}} |c_j(e)|}{\sum_{e \in \mathcal{E}_{\text{eval}}} \sum_{j=1}^{p} |c_j(e)|}. \tag{93}$$

This measures the fraction of absolute latent contribution captured by the discovery-selected coordinates out of sample. The frozen experiment contains only $L = 32$, so we do not infer a circuit-size curve or a minimum sufficient circuit size.

D.6 Circuit Ablation and Activation Exchange

Ablation. For a coordinate set C, define

$$\mathcal{A}_C(z)_j = \begin{cases} 0, & j \in C, \\ z_j, & j \notin C. \end{cases} \tag{94}$$

For the caption-side intervention, $\mathcal{A}_C$ is applied to both $z_{t,+}$ and $z_{t,-}$. The complete codes are decoded and normalized again before recomputing the margin. The per-example signed effect is

$$\delta_{\text{abl}}(e; C) = \hat{\Delta}_{\text{abl}}(e; C) - \hat{\Delta}(e). \tag{95}$$

Ablation asks whether the selected coordinates are necessary for the reconstructed decision, but it does not by itself show that they encode the specific positive-negative distinction.

Activation exchange. To test contrast-specific control, selected positive and negative caption activations are exchanged:

$$\widetilde{z}_{t,+,j} = \begin{cases} z_{t,-,j}, & j \in C, \\ z_{t,+,j}, & j \notin C, \end{cases} \qquad \widetilde{z}_{t,-,j} = \begin{cases} z_{t,+,j}, & j \in C, \\ z_{t,-,j}, & j \notin C. \end{cases} \tag{96}$$

and

$$\delta_{\mathrm{ex}}(e; C) = \hat{\Delta}_{\mathrm{ex}}(e; C) - \hat{\Delta}(e). \tag{97}$$

Unlike ablation, exchange replaces the selected evidence with activations observed for the competing caption. A useful contrast circuit should therefore move the margin toward the opposite composition more often than a matched random circuit. This is an intervention on the SAE reconstruction, not on the internal layers of the frozen VLM.

D.7 Size- and Activation-Matched Controls

Contribution-ranked coordinates can simply be more active than uniform random coordinates. The executed control matches each selected coordinate to a non-selected coordinate with similar discovery-set caption activity. The saved implementation uses

$$E_j = \frac{1}{2|\mathcal{E}_{\mathrm{disc}}|} \sum_{e \in \mathcal{E}_{\mathrm{disc}}} \left(|z_{t,+,j}(e)| + |z_{t,-,j}(e)| \right). \tag{98}$$

Although the project artifacts call this quantity "activation energy," it is a mean absolute-activation statistic rather than a mean squared activation.

For each of $B = 8$ control circuits, the selected coordinates are visited in a seeded random order. For selected coordinate j, the sampler finds the 64 unused, non-selected candidates closest to j in $|\log(1 + E_k) - \log(1 + E_j)|$, samples one candidate uniformly, and removes it from the available pool. This yields a size-32, one-to-one local-activity match. It does not impose a global relative-energy tolerance, and the frozen analysis does not report a control-tail p-value. Selected and control sets receive exactly the same ablation and activation-exchange operators.

D.8 Permutation Non-Identifiability and Executed Seed Matching

Literal coordinate indices cannot be compared across independently trained SAEs. Let $P \in \mathbb{R}^{p \times p}$ be a permutation matrix and define

$$z' = Pz, \qquad D' = DP^{\top}. \tag{99}$$

Then

$$D'z' = DP^{\top}Pz = Dz. \tag{100}$$

Lemma 1 (Permutation non-identifiability)**.** *Simultaneously permuting latent coordinates and decoder columns leaves every reconstruction unchanged. Consequently, a coordinate's numerical index has no intrinsic meaning across independently trained dictionaries.*

The reported causal-circuit stability audit uses a deliberately narrower matching rule than a combined activation/contribution matcher. Let seed 17 be the reference seed and let $C_{32}^{(17)}$ be its discovered circuit. For each target seed $b \in \{18, 19, 20, 21\}$, form the cosine-similarity matrix between the 32 unit-normalized reference decoder atoms and all $p = 10{,}240$ target decoder atoms,

$$M_{jk}^{17 \to b} = \left(D_j^{(17)} \right)^{\top} D_k^{(b)}, \qquad j \in C_{32}^{(17)}, \quad k \in \{1, \ldots, p\}. \tag{101}$$

A one-to-one assignment is obtained by Hungarian maximum-weight matching:

$$\pi_{17 \to b} = \arg\max_{\pi} \sum_{j \in C_{32}^{(17)}} M_{j,\pi(j)}^{17 \to b}. \tag{102}$$

We then report mean matched decoder cosine and the fraction of mapped units that fall within the target seed's independently discovered top-32 or top-128 contribution ranking. At width 10,240, literal random top-32 recovery is $32/10{,}240 = 0.003125$. This audit removes permutation freedom at the decoder direction level; it does not establish full semantic equivalence of matched atoms.

Table 12: **Frozen pooled-embedding backbones.** Model and pretrained identifiers are the OpenCLIP identifiers used to construct the cached representations. Every SAE and local baseline for a backbone reads the same arrays.

Backbone	Model identifier	Pretrained tag	Image dim.	Text dim.	Train / validation pairs
CLIP ViT-B/16	`ViT-B-16`	`openai`	512	512	118,287 / 5,000
SigLIP ViT-B/16	`ViT-B-16-SigLIP`	`webli`	768	768	118,287 / 5,000
SigLIP2 ViT-B/16	`ViT-B-16-SigLIP2`	`webli`	768	768	118,287 / 5,000

D.9 Cross-Seed Transfer: Defined but Not Executed

The natural stronger test would map a source circuit into a target seed,

$$C_{32}^{17\to b} = \left\{ \pi_{17\to b}(j) : j \in C_{32}^{(17)} \right\}, \tag{103}$$

and then ablate or exchange those mapped target-seed coordinates without reranking them on target evaluation examples. The frozen project does not contain this mapped intervention. We therefore report cross-seed *rank recovery*, not cross-seed *causal transfer*. This distinction prevents an observational decoder match from being described as replicated causal function.

D.10 Effect Aggregation and Uncertainty

For each independently trained SAE seed and benchmark, intervention statistics are computed on $\mathcal{E}_{\text{eval}}$ after circuit membership and all eight control circuits have been frozen. The saved causal artifact reports:

1. mean reconstructed margin before intervention;
2. signed change in mean reconstructed margin;
3. change in reconstructed accuracy;
4. activation-exchange decision-flip rate;
5. semantic intervention accuracy $|\mathcal{E}_{\text{eval}}|^{-1} \sum_e \mathbf{1}[\hat{\Delta}(e) > \hat{\Delta}_{\text{ex}}(e)]$; and
6. the mean of each corresponding statistic over the eight matched random circuits when that statistic was saved.

The paper summarizes the five independently trained SAE seeds using the mean and sample standard deviation and also exposes every seedwise result. The frozen causal artifact does not retain the full per-example intervention arrays, so this section does not claim paired-bootstrap confidence intervals for the intervention effects. It likewise contains no circuit-size sweep and no mapped cross-seed intervention transfer. The complete top-32 contribution, intervention, matched-control, rank-recovery, and per-seed results are reported in Appendix J.

E Additional Experimental Details

E.1 Data and Split Construction

We train the primary sparse autoencoders on paired image-caption representations extracted from the MS COCO training split. The vision-language backbones remain frozen throughout SAE optimization: gradients are never propagated into either the image or text encoder. Embeddings are computed once and cached, so every SAE and matched local baseline for a given backbone receives the same input arrays. Table 12 gives the exact model identifiers, pretraining tags, representation dimensions, and cache sizes.

The primary experiments use the normalized pooled image and text embeddings consumed by the dual encoder's contrastive matching head. Patch- and token-level representations appear only in explicitly identified grounding and relation diagnostics; they are not used to infer the primary pooled codes, choose their supports, or select the circuits in the pooled compositional analysis.

Table 13: **Experimental data usage and held-out boundaries.** Model selection, core preservation, and automatic semanticity-v4 reuse the frozen 5,000-pair COCO validation pool; causal discovery and intervention evaluation are disjoint.

Purpose	Source	Size	Separation / Reuse
SAE optimization	MS COCO train pooled pairs	118,287	Disjoint from COCO validation.
Model selection and core preservation	MS COCO validation	5,000	Shared for FVU, retrieval, GPI, and activity diagnostics.
Automatic semanticity-v4	Same COCO validation pool	5,000	Reused; not an independent held-out semantic set.
SugarCrepe behavior	Official examples	7,511	Final compositional evaluation.
SC++ official ITT behavior	Official examples	4,757	Final compositional evaluation.
SugarCrepe circuit discovery / intervention	SHA-256 parity halves	3,696 / 3,815	Strictly disjoint discovery and intervention examples.
ARO circuit diagnostic	SHA-256 parity halves	26,217 / 26,468	Strictly disjoint; ARO remains plausibility-confounded.
Cross-seed circuit matching	Decoder columns only	No examples	Cosine/Hungarian rank recovery; no matching set.
Mapped cross-seed causal transfer	—	Not executed	No transfer-effect claim is made.

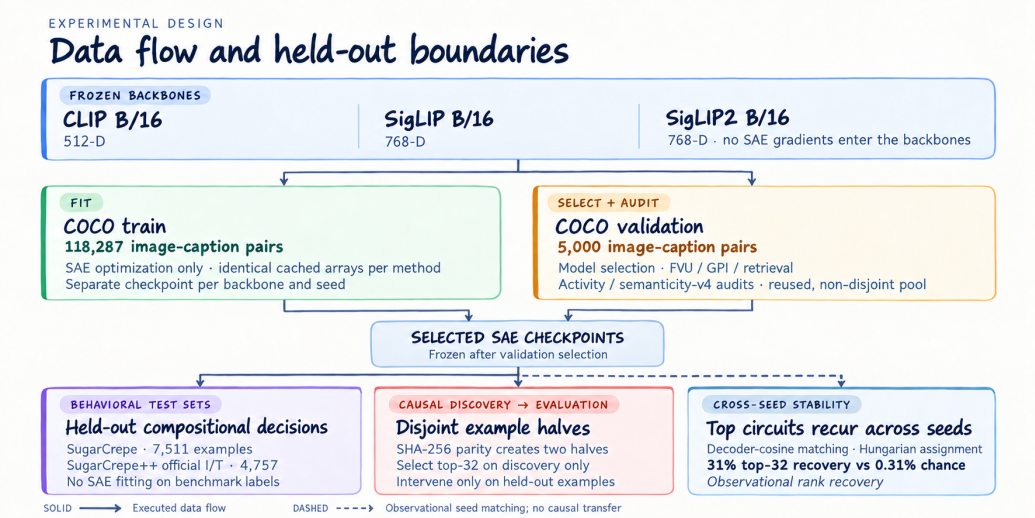

Figure 6: **Experimental data flow and held-out boundaries.** Frozen backbone representations feed COCO training and validation caches. SAE checkpoints are optimized on COCO train, selected without compositional benchmark labels, and then evaluated on the behavioral datasets. For causal experiments, deterministic SHA-256 parity divides each benchmark into disjoint discovery and intervention halves. The dashed arrow denotes observational cross-seed rank matching; mapped cross-seed causal transfer was not executed.

The experimental partitions are summarized in Table 13 and visualized in Figure 6. The 118,287-pair COCO training cache is disjoint from the 5,000-pair COCO validation cache. Within that validation cache, however, model selection, FVU, retrieval, GPI, activity statistics, and semanticity-v4 reuse the same examples. We therefore do *not* describe those measurements as mutually held out.

SugarCrepe and SugarCrepe++ (SC++) are used only after checkpoints and evaluation rules are frozen. The causal analyses impose an additional discovery/evaluation separation: top-ranked units are selected on one deterministic half and intervened on only in the other half. Cross-seed rank recovery uses decoder-column cosine similarity and Hungarian assignment, it does not consume an additional example-matching set. Accordingly, we report cross-seed recovery as observational stability and make no claim of mapped cross-seed intervention transfer.

E.2 Compared Methods

Table 14 distinguishes competitive methods from validity controls and records the inference rule used for every reported performance row.

Table 14: **Methods in the matched comparisons.** Every sparse performance row uses modality-local inference. Pair-conditioned supports, frozen random dictionaries, and SPARC's Global-support path are validity diagnostics rather than model-quality leaderboard rows.

Method	Evaluation inference	Support	Role
Dense VLM	Original normalized pooled embeddings	Not sparse	No-SAE behavior and geometry reference
Reconstruction-only SAE	Modality-local SAE code	Architecture-dependent	Reconstruction ceiling without multimodal regularizers
Ordinary TopK SAE	Independent per-modality TopK	Exact K	Generic modality-local fidelity baseline
MGSAE	Independent per-modality TopK	Exact K	Matched Group-$L_{2,1}$ multimodal baseline
Ours: dense tie-free	Independent dense ReLU primary code	No primary K	Behavioral-preservation endpoint; paired path is training-only
Ours: independent TopK	Independent per-modality TopK	$K \in \{64, 128, 256\}$	Matched-support semantic-coverage endpoint

Dense VLM. The original normalized image and text embeddings provide the no-SAE reference. This row defines the behavior and pairwise geometry that a reconstruction should preserve; it has neither a sparse support nor SAE reconstruction error.

Reconstruction-only SAE. This architecture is trained to reconstruct the two embedding streams without the multimodal regularizers of the proposed objective. It estimates the fidelity ceiling attainable by the corresponding SAE parameterization and tests whether vector reconstruction alone yields shared semantic features.

Ordinary TopK SAE. This baseline applies TopK independently to the image and text encoder activations,

$$z_v = \mathrm{TopK}_K(f(x_v)), \qquad z_t = \mathrm{TopK}_K(f(x_t)). \tag{104}$$

Neither modality can influence the other's selected support. The matched experiments use $K \in \{64, 128, 256\}$; the $K = 768$ row is retained as a higher-fidelity architecture control rather than a matched low-support comparison.

MGSAE. MGSAE (Kaushik et al., 2026) uses independent per-modality TopK inference, a shared encoder/decoder, Group-$L_{2,1}$ regularization, modality-specific pre-biases, and shared random masking. It is the closest completed matched baseline because it explicitly targets shared concepts in aligned pooled multimodal embeddings. We compare methods at the same emitted support whenever a semanticity-v4 claim is made.

Proposed dense tie-free SAE. The behavioral-preservation endpoint uses modality-specific centering, a shared affine encoder and concept dictionary, independently inferred dense ReLU codes, self-reconstruction, cross-decoding, shared latent masking, Group-Hoyer concentration, and a variable-size strictly-positive paired auxiliary. The auxiliary is used only during training; it never constructs a performance-evaluation support.

Proposed independent-TopK SAE. The semantic-coverage endpoint uses the same multimodal design but imposes modality-local TopK on each primary code. Its $K = 64, 128, 256$ checkpoints are compared with MGSAE and ordinary TopK at exactly matched emitted support. This separates the contribution of the multimodal objective from the effect of allowing a denser primary representation.

Validity controls and SPARC. Shared random supports, frozen random dictionaries, and pair-conditioned supports test whether the evaluation can encode pair identity without semantic features. They are diagnostics, not competitive models. The local two-stream SPARC adaptation is likewise excluded from headline model-quality tables: training uses Global support whereas valid deployment uses Local support, and our adaptation does not reproduce SPARC's native three-stream OpenCLIP/DINO configuration. Its pair-informed Global path is retained as direct mechanism evidence for the modality-local inference requirement.

Table 15: **Statistical protocol by result family.** The resampling unit is the benchmark example, not an individual seed–example duplicate.

Result family	Replicates	Uncertainty / test	Boundary
Core FVU, GPI, retrieval, deadness	Seeds 17–19	Mean ± sample SD	Whole 5,000-pair validation cache
Semanticity-v4 counts	Seeds 17–19 within matched K	Mean ± sample SD	Compare within matched support; v4 retains at most 64 coordinates
Primary compositional deltas	Three common seeds	10,000 paired example-cluster bootstrap draws	Seed outcomes retained within each resampled example
Primary category tests	Three common seeds	20,000 paired sign flips; global Holm correction	Dense tie-free versus MGSAE family
Matched-support TopK deltas	Three common seeds	20,000 paired bootstrap draws; 10,000 sign flips	Holm family stated with each result table
Top-32 causal circuits	Seeds 17–21	Mean ± sample SD; eight controls per seed	No saved per-example intervention bootstrap
Single-seed diagnostics	Seed 17 unless stated	Descriptive only	Not promoted to replicated comparisons

E.3 Optimization and Model Selection

All methods are trained separately for each backbone. Primary matched results aggregate independently initialized checkpoints with seeds 17, 18, and 19. Hyperparameters are selected using the frozen COCO validation cache using prespecified criteria-reconstruction, geometry preservation, selector validity, deadness, and semantic coverage-without access to SugarCrepe or SC++ labels.

The complete canonical SigLIP2 optimization configuration is reported in Table 7, backbone-specific operating points are listed in Table 8; and the status and coefficient of every objective term are given in Table 4. These tables are provided in Appendix C and are referenced here rather than duplicated with a second, potentially inconsistent set of hyperparameters.

For compactness, the selected dense objective can be written as

$$\mathcal{L} = \mathcal{L}_{\text{self}} + \lambda_{\text{cross}}\mathcal{L}_{\text{cross}} + \lambda_{\text{GH}}\mathcal{L}_{\text{group}} + \lambda_{+}\mathcal{L}_{+}, \tag{105}$$

with $\lambda_{\text{cross}} = 0.1$, $\lambda_{\text{GH}} = 2.5 \times 10^{-4}$, and $\lambda_{+} = 0.1$ for the canonical dense SigLIP2 checkpoint. Shared latent masking with probability 0.05 is a stochastic training operator, not a fifth additive loss. Appendix C defines each term and records the configurations in which a coefficient or support rule differs.

E.4 Statistical Protocol

The uncertainty procedure depends on the result family and is summarized in Table 15. Unless stated otherwise, core and semantic results are means over seeds 17-19, and the displayed variation is the sample standard deviation across independently trained checkpoints.

For primary compositional method differences, we use 10,000 paired example-cluster bootstrap draws. An example is the resampling unit; its outcomes for the common seeds remain together inside a draw. Category-level paired tests use 20,000 sign flips with Holm correction over the declared hypothesis family. The matched-support TopK comparisons use 20,000 paired bootstrap draws and 10,000 sign flips. Thus the reported intervals respect the paired benchmark structure instead of treating seed-example duplicates as independent observations.

For the top-32 causal analyses, a circuit is selected once on the discovery half and evaluated on the untouched intervention half. Reported causal summaries aggregate seeds 17-21 and compare each selected circuit with eight activation-energy-matched control circuits per seed. The frozen artifact does not contain a per-example intervention bootstrap, so these results are reported as across-seed mean and sample standard deviation rather than with an invented example-level confidence interval. Single-seed mechanism audits are explicitly labeled descriptive and are not promoted to replicated method comparisons. Further causal-protocol details appear in Appendix D, with uncertainty boundaries stated in Appendix D.10.

F Evaluation Metrics and Protocols

F.1 Representational Fidelity

We evaluate whether the sparse reconstruction retains the information and geometry of the original VLM embeddings.

Fraction of variance unexplained. For embeddings x_i and reconstructions $\hat{x}_i$, FVU measures normalized squared reconstruction error:

$$\mathrm{FVU} = \frac{\sum_i \|x_i - \hat{x}_i\|_2^2}{\sum_i \|x_i - \bar{x}\|_2^2}, \tag{106}$$

where $\bar{x}$ is the empirical mean of the corresponding modality. Lower values indicate better reconstruction.

Cosine reconstruction. We compute the mean cosine similarity between each original embedding and its reconstruction:

$$\frac{1}{N}\sum_{i=1}^{N} \cos(x_i, \hat{x}_i). \tag{107}$$

This complements FVU by measuring directional preservation in the normalized embedding space used for retrieval.

Image-text retrieval. We report image-to-text and text-to-image Recall@1 using the original dense embeddings and the reconstructed embeddings. All SAE retrieval results use independently inferred modality-local codes. Pair-conditioned retrieval is reported only in the validity audit.

Embedding geometry. We compare pairwise similarities before and after reconstruction to determine whether the SAE preserves the global organization of the embedding space. This analysis complements pointwise reconstruction because low reconstruction error does not necessarily preserve relative neighborhoods or small contrastive margins.

F.2 Compositional Behavior

We evaluate compositional behavior on SugarCrepe, SugarCrepe++, ARO, and Winoground. These benchmarks probe complementary aspects of structured understanding, including attribute assignment, relation direction, word order, semantic roles, and robustness to lexical reformulation.

For a dense VLM decision y and reconstructed decision $\hat{y}$, examples are partitioned into:

- **Preserved:** $\hat{y} = y$;
- **Corrected:** the dense VLM is incorrect but the reconstruction is correct;
- **Degraded:** the dense VLM is correct but the reconstruction is incorrect.

Decision preservation is therefore

$$\mathrm{DP} = 1 - r_{\mathrm{corrected}} - r_{\mathrm{degraded}}. \tag{108}$$

This accounting distinguishes faithful reconstruction from changes that happen to improve aggregate accuracy.

For a correct caption t^+ and compositional foil t^-, the dense and reconstructed margins are

$$m = s(v, t^+) - s(v, t^-), \qquad \hat{m} = s(\hat{v}, \hat{t}^+) - s(\hat{v}, \hat{t}^-). \tag{109}$$

We evaluate sign preservation and the distortion between m and $\hat{m}$. Small-margin examples are analyzed separately because they are most susceptible to decision reversal under reconstruction.

F.3 Semantic Structure and Grounding

Semantic evaluation asks whether active coordinates correspond to coherent and reusable concepts rather than only reconstructing the input vector.

Feature purity and label concentration. For each coordinate, we inspect its most highly activating examples and measure the concentration of semantic labels within this set. High purity indicates that a coordinate responds consistently to a recognizable concept.

Lift. Lift compares the frequency of a semantic label among a feature's highly activating examples with its background frequency. This reduces the tendency of common concepts to appear semantically meaningful merely because they occur frequently in the dataset.

Shared-concept coverage. We measure how many semantic concepts are recoverable from both independently inferred image and text codes at a given sparsity level. Shared coverage is evaluated without using paired captions to select image supports or paired images to select text supports.

Dense patch/token circuit diagnostic. We examine where a selected compositional circuit affects the frozen model using image-region occlusion and caption-token deletion. The circuit is selected from pooled SAE margin contributions; the patch/token analysis diagnoses its effects rather than assigning a region or span to each pooled feature. These diagnostics do not change the pooled codes used in the primary experiments.

F.4 Causal Structure

Causal evaluation uses circuits selected by mean absolute margin contribution on a discovery split. We test their effects through ablation and activation exchange on a disjoint evaluation split. Each circuit is compared with random coordinate sets matched for size and aggregate activation energy. Cross-seed decoder matching evaluates whether corresponding circuits recur in independently trained dictionaries (rank recovery). The complete protocol appears in Appendix D.

G Q1: Detailed Validity Audit

This section provides the complete evidence for the two validity requirements introduced in Section 3.2: modality-local inference and selector-score identifiability. We distinguish three questions that can otherwise be conflated. First, do the conclusions survive an exact, matched sparsity constraint? Second, can pair-conditioned supports transmit the known image-caption correspondence into the representation? Third, does every coordinate chosen by a fixed bilateral selector have positive evidence under the selector's stated score?

G.1 Matched Exact-Sparsity Audit

The native dense model uses nonnegative ReLU codes and does not constrain the primary representation to activate exactly K coordinates. A formally dense code can nevertheless concentrate most of its activation energy in a much smaller participation-ratio support. To ensure that the reported behavior is not explained only by a larger emitted support, we additionally train and evaluate exact-sparse checkpoints whose image and text supports are selected independently:

$$z_v^{(K)} = \mathrm{TopK}_K(f(x_v)), \qquad z_t^{(K)} = \mathrm{TopK}_K(f(x_t)). \tag{110}$$

The image support depends only on x_v, and the text support depends only on x_t. No paired observation, bilateral product, or joint mask participates in either inference map.

The primary exact-sparsity comparison uses $K \in \{64, 128, 256\}$. Within each K block, our objective, MGSAE, and ordinary reconstruction-only TopK emit the same number of coordinates. This matched-support sweep uses a separately selected SigLIP2 checkpoint grid from the cross-backbone operating-point comparison in the main paper. Its TopK-256 results should therefore be interpreted alongside the $K = 256$ MGSAE and ordinary TopK rows in this grid. Table 16 reports reconstruction, geometry, retrieval, deadness, and automatic semantic coverage. Table 17 reports reconstructed behavioral accuracy and the full correction/degradation accounting on SugarCrepe and SC++.

The matched-support results define a Pareto frontier rather than a universal winner. Our objective preserves SugarCrepe and SC++ decisions better than matched MGSAE at every tested K, while MGSAE has lower FVU and higher GPI at every K. Against ordinary TopK, the behavioral results

Table 16: **Matched exact-sparsity reconstruction, retrieval, and semantic coverage on SigLIP2 ViT-B/16.** Image and text are encoded independently. R@1 is computed from reconstructed embeddings. Semantic counts are strict/broad/unique under semanticity-v4; they are comparable within an exact-K block because the v4 truncation fraction is matched. Values are mean $\pm$ SD over seeds 17–19.

K	Method	FVU↓	GPI↑	I2T R@1↑	T2I R@1↑	Dead Fraction↓	V4 Strict/Broad/Unique↑
64	Ours (full objective)	0.1919 ± 0.0004	0.9313 ± 0.0006	0.3805 ± 0.0046	0.3933 ± 0.0030	0.710 ± 0.002	49.3 ± 3.2 / 250.3 ± 8.5 / 23.3 ± 1.2
64	MGSAE $L_{2,1}$	0.1356 ± 0.0003	0.9466 ± 0.0006	0.3311 ± 0.0021	0.3774 ± 0.0014	0.628 ± 0.003	44.3 ± 0.6 / 249.0 ± 2.6 / 21.3 ± 1.5
64	Ordinary TopK	0.1811 ± 0.0009	0.9368 ± 0.0004	0.3601 ± 0.0053	0.3869 ± 0.0027	0.849 ± 0.005	53.7 ± 2.1 / 230.0 ± 2.0 / 25.7 ± 0.6
128	Ours (full objective)	0.1507 ± 0.0002	0.9453 ± 0.0007	0.4031 ± 0.0041	0.4083 ± 0.0050	0.638 ± 0.002	55.3 ± 3.5 / 267.0 ± 12.5 / 24.7 ± 1.2
128	MGSAE $L_{2,1}$	0.1042 ± 0.0003	0.9606 ± 0.0005	0.3733 ± 0.0013	0.4110 ± 0.0036	0.516 ± 0.002	48.3 ± 2.9 / 268.3 ± 7.8 / 23.7 ± 1.5
128	Ordinary TopK	0.1391 ± 0.0003	0.9517 ± 0.0003	0.3947 ± 0.0017	0.4090 ± 0.0007	0.849 ± 0.002	49.3 ± 3.5 / 214.0 ± 1.0 / 23.3 ± 0.6
256	Ours (full objective)	0.1007 ± 0.0005	0.9621 ± 0.0005	0.4271 ± 0.0007	0.4413 ± 0.0086	0.608 ± 0.003	56.0 ± 4.6 / 247.0 ± 11.5 / 25.3 ± 2.3
256	MGSAE $L_{2,1}$	0.0791 ± 0.0003	0.9698 ± 0.0002	0.4069 ± 0.0033	0.4347 ± 0.0015	0.441 ± 0.002	60.7 ± 4.5 / 307.7 ± 8.5 / 24.3 ± 2.3
256	Ordinary TopK	0.0820 ± 0.0002	0.9762 ± 0.0004	0.4404 ± 0.0031	0.4546 ± 0.0009	0.884 ± 0.003	23.0 ± 1.0 / 128.3 ± 1.2 / 14.7 ± 1.2

Table 17: **Matched exact-K compositional behavior.** DP is the fraction of dense-backbone decisions whose label is unchanged after reconstruction; corrected and degraded are directed disagreement rates. Values are mean $\pm$ SD over seeds 17–19. The frozen SigLIP2 references are 0.8667 on SugarCrepe and 0.7538 on SC++ ITT.

Benchmark	K	Method	Accuracy↑	DP↑	Corrected	Degraded↓	GPI↑
SugarCrepe	64	Ours (full objective)	0.8134 ± 0.0012	0.8564 ± 0.0012	0.0451 ± 0.0009	0.0985 ± 0.0009	0.9247 ± 0.0003
SugarCrepe	64	MGSAE $L_{2,1}$	0.7627 ± 0.0012	0.8051 ± 0.0064	0.0454 ± 0.0033	0.1495 ± 0.0033	0.9435 ± 0.0005
SugarCrepe	64	Ordinary TopK	0.8151 ± 0.0022	0.8611 ± 0.0015	0.0436 ± 0.0018	0.0953 ± 0.0003	0.9313 ± 0.0003
SugarCrepe	128	Ours (full objective)	0.8244 ± 0.0041	0.8763 ± 0.0014	0.0407 ± 0.0024	0.0830 ± 0.0019	0.9403 ± 0.0001
SugarCrepe	128	MGSAE $L_{2,1}$	0.7810 ± 0.0024	0.8292 ± 0.0019	0.0425 ± 0.0020	0.1283 ± 0.0009	0.9581 ± 0.0006
SugarCrepe	128	Ordinary TopK	0.8239 ± 0.0022	0.8820 ± 0.0026	0.0376 ± 0.0011	0.0804 ± 0.0021	0.9473 ± 0.0001
SugarCrepe	256	Ours (full objective)	0.8359 ± 0.0035	0.8917 ± 0.0025	0.0387 ± 0.0008	0.0696 ± 0.0029	0.9591 ± 0.0006
SugarCrepe	256	MGSAE $L_{2,1}$	0.8017 ± 0.0063	0.8618 ± 0.0060	0.0366 ± 0.0002	0.1016 ± 0.0061	0.9678 ± 0.0003
SugarCrepe	256	Ordinary TopK	0.8421 ± 0.0018	0.9085 ± 0.0030	0.0334 ± 0.0016	0.0580 ± 0.0019	0.9743 ± 0.0005
SC++ ITT	64	Ours (full objective)	0.7075 ± 0.0050	0.8333 ± 0.0036	0.0602 ± 0.0007	0.1065 ± 0.0043	0.9254 ± 0.0001
SC++ ITT	64	MGSAE $L_{2,1}$	0.6676 ± 0.0034	0.7852 ± 0.0074	0.0643 ± 0.0044	0.1505 ± 0.0036	0.9429 ± 0.0003
SC++ ITT	64	Ordinary TopK	0.7042 ± 0.0032	0.8358 ± 0.0035	0.0572 ± 0.0016	0.1069 ± 0.0029	0.9312 ± 0.0008
SC++ ITT	128	Ours (full objective)	0.7146 ± 0.0052	0.8500 ± 0.0021	0.0554 ± 0.0030	0.0946 ± 0.0025	0.9409 ± 0.0004
SC++ ITT	128	MGSAE $L_{2,1}$	0.6843 ± 0.0034	0.8074 ± 0.0038	0.0615 ± 0.0026	0.1310 ± 0.0024	0.9579 ± 0.0002
SC++ ITT	128	Ordinary TopK	0.7156 ± 0.0025	0.8568 ± 0.0028	0.0525 ± 0.0014	0.0907 ± 0.0023	0.9473 ± 0.0001
SC++ ITT	256	Ours (full objective)	0.7220 ± 0.0040	0.8702 ± 0.0033	0.0490 ± 0.0032	0.0808 ± 0.0017	0.9594 ± 0.0007
SC++ ITT	256	MGSAE $L_{2,1}$	0.6914 ± 0.0049	0.8341 ± 0.0063	0.0517 ± 0.0013	0.1141 ± 0.0055	0.9677 ± 0.0003
SC++ ITT	256	Ordinary TopK	0.7292 ± 0.0041	0.8864 ± 0.0076	0.0445 ± 0.0031	0.0691 ± 0.0053	0.9745 ± 0.0006

are statistically close at $K = 64$ and $K = 128$, and ordinary TopK is modestly higher at $K = 256$. Our advantage over ordinary TopK is instead the much larger active dictionary and shared semantic coverage at $K = 128$ and $K = 256$. At $K = 64$, ordinary TopK has slightly more strict and unique semanticity-v4 units. We therefore use the native dense checkpoint as the behavioral-preservation endpoint and the exact-K checkpoints as controlled semantic-fidelity operating points.

G.2 Pair-Conditioned Support Leakage

Let $z_v, z_t \in \mathbb{R}^p_{\geq 0}$ denote the image and text codes for a matched example. Under modality-local inference, each representation is a function only of its own input:

$$z_v = f_v(x_v), \qquad z_t = f_t(x_t). \tag{111}$$

Under pair-conditioned inference, a joint selector $S(x_v, x_t) \subseteq [p]$ instead determines the support of both codes:

$$\widetilde{z}_{v,j} = z_{v,j}\mathbf{1}[j \in S(x_v, x_t)], \qquad \widetilde{z}_{t,j} = z_{t,j}\mathbf{1}[j \in S(x_v, x_t)]. \tag{112}$$

The same pair-dependent support is now embedded in both representations. Retrieval can consequently compare support fingerprints even when the selected coordinates contain no independently recognized semantic content. High latent retrieval under this protocol does not establish cross-modal concept alignment.

We test this channel with three controls. A *shared-random-support* control assigns each true pair an independently sampled support shared only within that pair. A *frozen-random-dictionary* control leaves randomly initialized encoders and decoders untrained but constructs the support from the known pair. A *learned pair-conditioned* control uses trained dictionaries while retaining the same

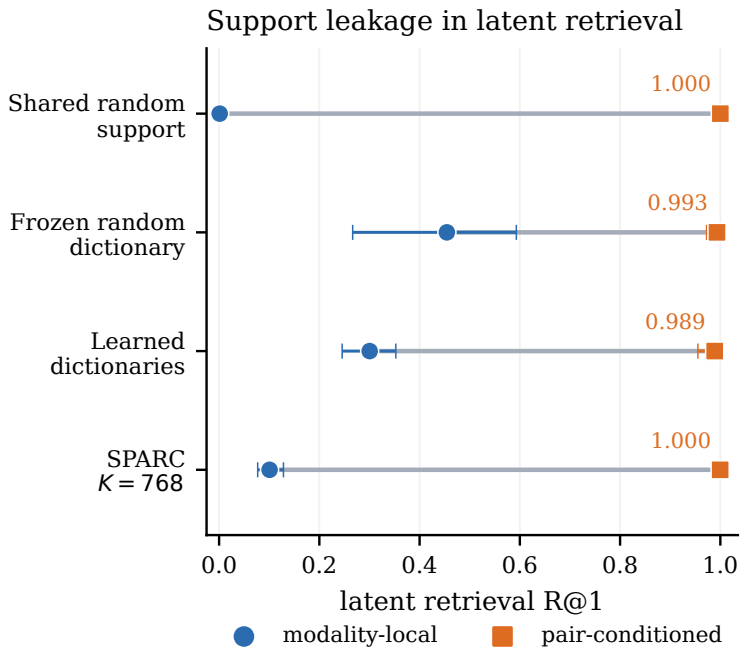

Figure 7: **Pair-conditioned versus modality-local latent retrieval.** Points show the mean and whiskers show the full range over widths and support sizes for the constructed random controls, over checkpoints for the learned dictionaries, and over seed-direction combinations for SPARC. Pair-conditioned retrieval remains near one even for shared random supports and frozen random dictionaries, whereas valid modality-local retrieval falls sharply.

invalid access to the matched partner. Each is compared with a modality-local representation in which the two sides are encoded independently.

Figure 7 summarizes the resulting collapse, and Table 18 gives every retained value. For the constructed controls and four legacy learned checkpoints, the frozen audit retains the mean of image-to-text and text-to-image R@1 rather than the two directions separately. SPARC is reported direction by direction.

Shared random supports reach R@1 equal to 1.000 despite containing no semantic information. Pair-conditioned frozen random dictionaries reach 0.9725-0.9980, and learned pair-conditioned dictionaries reach 0.9552-1.0000. Their modality-local counterparts are substantially lower, showing that the pair-conditioned score largely reflects the inference procedure rather than independently recovered concepts.

The same fingerprint appears in an externally motivated architecture. For the three-seed SPARC $K = 768$ reproduction, pair-informed Global support gives image-to-text and text-to-image latent R@1 of 0.9997 ± 0.0001 and 0.9996 ± 0.0000. Applying SPARC's Local-inference safeguard reduces them to 0.0795 ± 0.0030 and 0.1220 ± 0.0076. This is evidence for the necessity of the safeguard, not a claim that SPARC reports the invalid Global score as deployment performance.

Pairwise information may still be used as a training signal. Our strictly-positive paired auxiliary uses matched examples to shape the dictionary, but it is removed at evaluation. The primary image code depends only on the image embedding, and the primary text code depends only on the text embedding. Pair identity therefore cannot be transmitted through the evaluated supports.

G.3 Score Indeterminacy Under Fixed Product-TopK

We next audit fixed product-TopK selection over nonnegative codes. For each matched example, define the bilateral score and its strictly-positive count:

$$s_j = z_{v,j} z_{t,j}, \qquad N_+ = \#\{j : s_j > 0\} = \#\{j : z_{v,j} z_{t,j} > 0\}. \tag{113}$$

Table 18: **Pair-conditioned support leakage.** For the constructed controls and the four legacy learned checkpoints, R@1 is the mean of image-to-text and text-to-image retrieval, which is the directional information retained by the frozen audit. SPARC is reported direction by direction over three seeds. Pair-conditioned codes are not valid deployment representations because construction of either side uses the known matched pair.

Codes/Supports	W	K	N	Direction	Pair-Conditioned R@1	Modality-Local R@1
Shared random support	2048	64	1000	mean	1.0000	0.0015
Shared random support	2048	128	1000	mean	1.0000	0.0010
Shared random support	2048	256	1000	mean	1.0000	0.0005
Shared random support	2048	512	1000	mean	1.0000	0.0015
Shared random support	5120	64	1000	mean	1.0000	0.0015
Shared random support	5120	128	1000	mean	1.0000	0.0010
Shared random support	5120	256	1000	mean	1.0000	0.0025
Shared random support	5120	512	1000	mean	1.0000	0.0010
Shared random support	10240	64	1000	mean	1.0000	0.0010
Shared random support	10240	128	1000	mean	1.0000	0.0010
Shared random support	10240	256	1000	mean	1.0000	0.0010
Shared random support	10240	512	1000	mean	1.0000	0.0010
Frozen random dictionary	2048	64	1000	mean	0.9930	0.3190
Frozen random dictionary	2048	128	1000	mean	0.9935	0.4480
Frozen random dictionary	2048	256	1000	mean	0.9865	0.5475
Frozen random dictionary	2048	512	1000	mean	0.9725	0.5930
Frozen random dictionary	5120	64	1000	mean	0.9975	0.2665
Frozen random dictionary	5120	128	1000	mean	0.9975	0.4325
Frozen random dictionary	5120	256	1000	mean	0.9955	0.5280
Frozen random dictionary	5120	512	1000	mean	0.9940	0.5815
Frozen random dictionary	10240	64	1000	mean	0.9980	0.2790
Frozen random dictionary	10240	128	1000	mean	0.9980	0.3935
Frozen random dictionary	10240	256	1000	mean	0.9970	0.5115
Frozen random dictionary	10240	512	1000	mean	0.9965	0.5535
Learned: alignment_loss_null	–	–	5000	mean	0.9552	0.3357
Learned: crossdecode_selbal0	–	–	5000	mean	0.9997	0.3526
Learned: crossdecode_selbal0p003	–	–	5000	mean	1.0000	0.2678
Learned: crossdecode_selbal0p01	–	–	5000	mean	1.0000	0.2455
SPARC ($K = 768$)	10240	768	5000	I2T	0.9997 ± 0.0001	0.0795 ± 0.0030
SPARC ($K = 768$)	10240	768	5000	T2I	0.9996 ± 0.0000	0.1220 ± 0.0076

A fixed product-TopK selector requests the K largest bilateral scores. If $K \leq N_+$, each selected coordinate has strictly positive bilateral evidence, aside from possible equal-score ties among positive values. If $K > N_+$, the selector must fill exactly $K - N_+$ positions with coordinates whose scores are all zero. We measure this forced fill by

$$\rho_{\text{zero}} = \frac{\max(0, K - N_+)}{K}. \tag{114}$$

The zero-score coordinates are not preferred by the stated product score: their identities depend on sorting, index order, and tie-breaking behavior. Deterministic code can therefore return a reproducible support even when part of that support is not identified by the selection objective. We call a selected coordinate *score-identifiable* only when its inclusion is determined by the declared score rather than an unstated zero-tie rule.

Table 19 summarizes the exhaustive manifest-backed corpus, and Table 20 gives all 29 compatible checkpoint-level results. Forced zero-product selection occurs in 26 of 29 checkpoints, 25 have at least 10% tied selected slots. The median tied-slot fraction is 0.3495, and the maximum is 0.5859.

Figure 8 shows that this indeterminacy is not merely a missing-score count. In the representative deterministic per-example selector, every forced zero-score choice falls in the first dictionary-index decile. Across the historical deterministic per-example checkpoints, the first decile receives approximately 6.9-9.1 times its uniform activation-energy share. Randomizing the zero-tie choice removes the index preference, confirming that it is a selector artifact rather than a learned semantic allocation. Random tie-breaking is used only as a diagnostic; it does not make zero-score coordinates positively justified.

Table 19: **Prevalence of score-indeterminate fixed product-TopK support** in the compatible historical checkpoint corpus. The census is exhaustive for checkpoints with reconstructable manifest metadata, not for arbitrary checkpoint files without a protocol record.

Prevalence statistic	Result
Audited compatible checkpoints	29
Any forced zero-product selection	26/29
At least 1% tied selected slots	26/29
At least 10% tied selected slots	25/29
Median tied selected-slot fraction	0.3495
Maximum tied selected-slot fraction	0.5859
Median examples/blocks containing a tie	1.0000

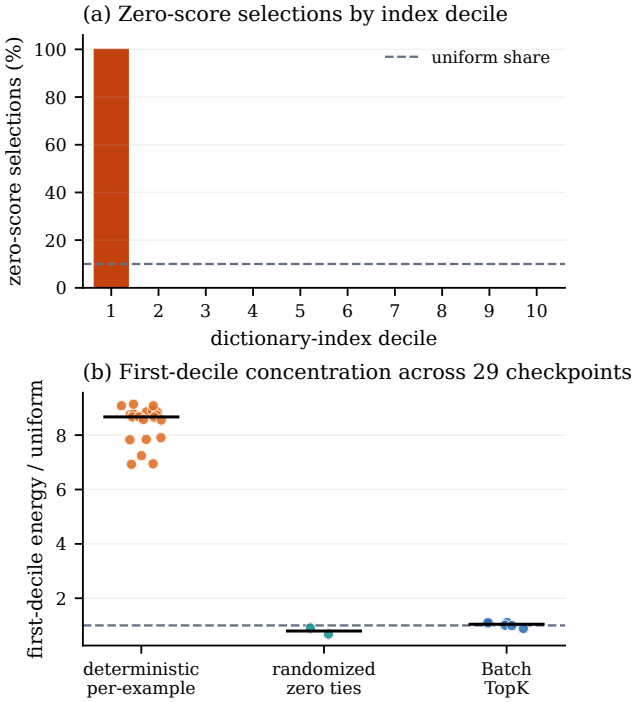

Figure 8: **Selector indeterminacy produces coordinate artifacts.** (a) In a representative fixed product-TopK checkpoint, every forced zero-score selection lies in the first dictionary-index decile; the dashed line shows the 10% uniform share. (b) Deterministic per-example checkpoints concentrate 6.9-9.1 times the uniform activation energy in the first decile, whereas randomized zero-tie controls and BatchTopK checkpoints are near uniform. Random tie-breaking diagnoses the artifact but is not used as a performance repair.

G.4 Selector Validity During Training

Conventional quality metrics do not reliably reveal the transition into the score-indeterminate regime. We retrain matched $K = 256$ and $K = 1024$ checkpoints with the same data, initialization, architecture, optimizer, and training duration, changing only the support budget. Figure 9 tracks N_+/K, forced zero-score fill, FVU, decoded retrieval retention, and GPI.

Both runs start with more joint-positive coordinates than their requested budgets. After one epoch, the $K = 1024$ model has mean $N_+ = 691.9 < K$ and requires 32.4% zero-score fill. By epoch eight, mean N_+ falls to 424.1 and forced fill reaches 58.6%. The $K = 256$ run remains effectively tie-free. Nevertheless, the invalid high-K support has better FVU, decoded retrieval, and GPI. Reconstruction quality and selector-score identifiability are therefore separate axes: a model can preserve vectors better while its nominal selected support becomes less determinate.

Table 20: **Checkpoint-level fixed-product selector audit.** N_+ is reported as mean [5th, 95th percentile]. Tie slots/selected equals ρ_{zero}; rows tied is the fraction of examples or BatchTopK blocks containing forced zero-score fill. Overall, 26/29 checkpoints contain forced ties. The final column is activation energy in the first dictionary-index decile divided by its uniform share.

ID	Backbone	K	Scope	Configuration	N_+ Mean [p05, p95]	Tie Slots	Rows Tied	Index Conc.
P01	clip_b16	768	example	cross-backbone, s17	467.2 [436, 499]	0.3917	1.0000	9.08
P02	clip_b16	768	example	cross-backbone, s18	471.4 [441, 503]	0.3862	1.0000	9.08
P03	clip_b16	768	example	cross-backbone, s19	472.4 [442, 505]	0.3849	1.0000	9.14
P04	siglip	768	example	cross-backbone, s17	501.3 [455, 551]	0.3473	1.0000	8.66
P05	siglip	768	example	cross-backbone, s18	499.5 [453, 550]	0.3495	1.0000	8.66
P06	siglip	768	example	cross-backbone, s19	495.9 [450, 546]	0.3543	1.0000	8.67
P07	siglip2	64	batch	BatchTopK, s17	253.8 [150, 463]	0.0000	0.0000	1.10
P08	siglip2	128	batch	BatchTopK, s17	215.9 [178, 266]	0.0000	0.0000	1.11
P09	siglip2	256	batch	BatchTopK, s17	316.0 [274, 364]	0.0000	0.0000	1.01
P10	siglip2	512	batch	BatchTopK, s17	461.8 [405, 526]	0.0979	1.0000	0.89
P11	siglip2	768	example	common center, s17	523.5 [470, 576]	0.3184	1.0000	8.85
P12	siglip2	768	example	common center, s18	513.9 [461, 566]	0.3309	1.0000	8.84
P13	siglip2	768	example	common center, s19	522.0 [468, 576]	0.3202	1.0000	8.90
P14	siglip2	768	example	no center, s17	487.1 [439, 534]	0.3658	0.9998	8.76
P15	siglip2	768	example	no center, s18	484.8 [438, 531]	0.3688	0.9998	8.77
P16	siglip2	768	example	no center, s19	484.2 [438, 529]	0.3695	0.9998	8.74
P17	siglip2	768	example	two decoder, s18	537.3 [473, 609]	0.3004	1.0000	6.93
P18	siglip2	768	example	shared decoder, s18	481.9 [429, 538]	0.3726	1.0000	8.67
P19	siglip2	768	example	two decoder, s19	529.3 [468, 601]	0.3108	1.0000	7.24
P20	siglip2	768	example	shared decoder, s19	477.9 [425, 534]	0.3778	1.0000	8.58
P21	siglip2	768	example	shared decoder, s17	479.6 [429, 535]	0.3755	1.0000	8.56
P22	siglip2	768	example	random ties, s17	483.3 [422, 554]	0.3707	1.0000	0.91
P23	siglip2	768	example	random exact ties, s17	338.9 [266, 415]	0.5587	1.0000	0.68
P24	siglip2	768	batch	BatchTopK, s17	443.5 [387, 508]	0.4225	1.0000	0.99
P25	siglip2	768	example	simple product TopK, s17	534.2 [470, 604]	0.3044	1.0000	6.95
P26	siglip2	768	example	group L1-infinity, s17	633.5 [588, 683]	0.1752	0.9994	7.84
P27	siglip2	768	example	group L1-infinity, s18	637.8 [593, 688]	0.1696	0.9996	7.91
P28	siglip2	768	example	group L1-infinity, s19	627.0 [581, 679]	0.1836	0.9996	7.83
P29	siglip2	1024	batch	BatchTopK, s17	424.1 [366, 487]	0.5859	1.0000	1.08

G.5 Implications for the Remaining Experiments

The two validity failures require different safeguards. Pair-conditioned support construction can manufacture cross-modal agreement by encoding pair identity, modality-local inference prevents this. Fixed product-TopK can return unsupported coordinates whenever $K > N_+$; positive-only variable support prevents forced zero-score fill.

Every primary performance result in this paper therefore satisfies:

1. the image code is inferred without access to its paired caption,
2. the text code is inferred without access to its paired image, and
3. every coordinate retained by a bilateral training selector has strictly positive evidence under its declared score.

The positive-only paired auxiliary implements the third condition by allowing its support size to equal the number of strictly-positive bilateral products instead of padding to a fixed K. Because the auxiliary is training-only, it does not alter the modality-local primary codes used for reconstruction, retrieval, semantic evaluation, or causal intervention.

H Q2: Detailed Behavioral-Preservation Analysis

This section provides the complete evidence supporting Section 5.3. We distinguish three questions that are often conflated: whether an SAE reconstructs the embedding vector, whether it preserves the frozen backbone's decisions, and whether the reconstructed representation is correct on a benchmark. A reconstruction can have low vector error while reversing a small compositional margin, and it can improve aggregate accuracy while changing many originally correct and incorrect decisions. We therefore report vector fidelity, behavioral accounting, margin geometry, retrieval geometry, and semantic coverage as separate axes.

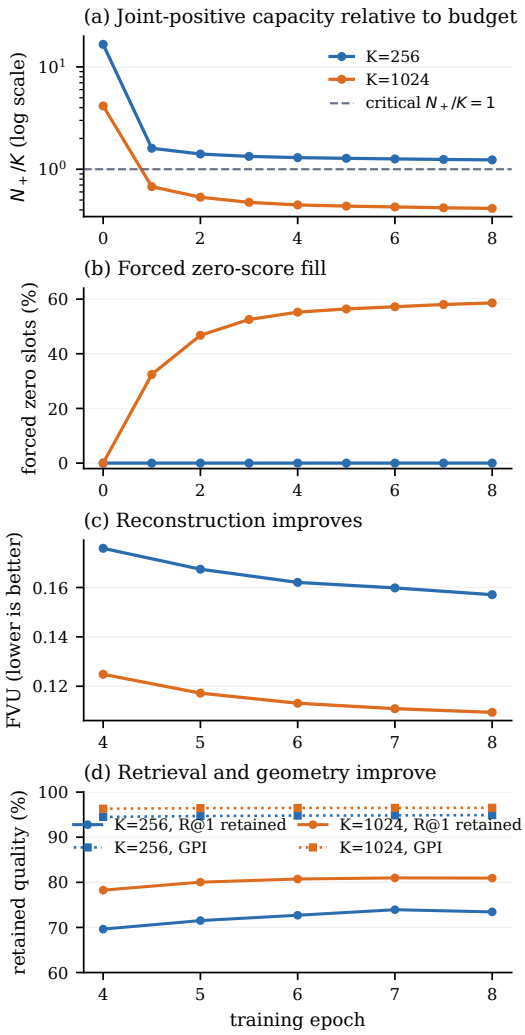

Figure 9: **Quality and selector validity are distinct axes.** The matched $K = 1024$ run crosses below $N_+/K = 1$ after one epoch and reaches 58.6% forced zero-score fill, while $K = 256$ remains effectively tie-free. Nevertheless, the high-K run has lower FVU and higher retained decoded R@1 and image-text GPI. Selector statistics are shown for epochs 0-8; the stored per-epoch quality trace covers epochs 4-8.

The experiments below also belong to three explicitly separated regimes. The matched-support analysis uses independently trained, modality-local TopK checkpoints at $K \in \{64, 128, 256\}$. The correction-enrichment analysis uses the canonical dense tie-free checkpoint and matched MGSAE. The final control table summarizes the frozen dense-family and architecture controls available in the common evaluator. Results from one regime are not silently substituted for another.

H.1 Behavioral Accounting

For each example, we classify the reconstructed decision as preserved, corrected, or degraded relative to the frozen dense backbone. Let Corr be the fraction for which the backbone is wrong and the reconstruction is right, and let Deg be the fraction for which the backbone is right and the reconstruction is wrong. As established in Proposition 1,

$$\mathrm{DP} = 1 - \mathrm{Corr} - \mathrm{Deg}, \qquad A_{\mathrm{SAE}} - A_{\mathrm{dense}} = \mathrm{Corr} - \mathrm{Deg}\,. \tag{115}$$

Thus reconstructed accuracy reports the *net* balance of corrections and degradations, whereas decision preservation (DP) reports how often the SAE retains the backbone's original binary decision.

Table 21: **Behavioral accounting under matched, modality-local exact sparsity.** Values are mean ± SD over seeds 17–19. Corrections are dense-wrong/SAE-right; degradations are dense-right/SAE-wrong. SC++ uses the official image-to-text (ITT) comparison.

Benchmark	Method	K	Dense Acc.	Recon. Acc.	DP	Corrected	Degraded
SugarCrepe	Frozen SigLIP2	–	0.8667	0.8667	1.0000	0.0000	0.0000
SugarCrepe	Ours	64	0.8667	0.8134 ± 0.0012	0.8564 ± 0.0012	0.0451 ± 0.0009	0.0985 ± 0.0009
SugarCrepe	MGSAE $L_{2,1}$	64	0.8667	0.7627 ± 0.0012	0.8051 ± 0.0064	0.0454 ± 0.0033	0.1495 ± 0.0033
SugarCrepe	Ordinary TopK	64	0.8667	0.8151 ± 0.0022	0.8611 ± 0.0015	0.0436 ± 0.0018	0.0953 ± 0.0003
SugarCrepe	Ours	128	0.8667	0.8244 ± 0.0041	0.8763 ± 0.0014	0.0407 ± 0.0024	0.0830 ± 0.0019
SugarCrepe	MGSAE $L_{2,1}$	128	0.8667	0.7810 ± 0.0024	0.8292 ± 0.0019	0.0425 ± 0.0020	0.1283 ± 0.0009
SugarCrepe	Ordinary TopK	128	0.8667	0.8239 ± 0.0022	0.8820 ± 0.0026	0.0376 ± 0.0011	0.0804 ± 0.0021
SugarCrepe	Ours	256	0.8667	0.8359 ± 0.0035	0.8917 ± 0.0025	0.0387 ± 0.0008	0.0696 ± 0.0029
SugarCrepe	MGSAE $L_{2,1}$	256	0.8667	0.8017 ± 0.0063	0.8618 ± 0.0060	0.0366 ± 0.0002	0.1016 ± 0.0061
SugarCrepe	Ordinary TopK	256	0.8667	0.8421 ± 0.0018	0.9085 ± 0.0030	0.0334 ± 0.0016	0.0580 ± 0.0019
SC++ ITT	Frozen SigLIP2	–	0.7538	0.7538	1.0000	0.0000	0.0000
SC++ ITT	Ours	64	0.7538	0.7075 ± 0.0050	0.8333 ± 0.0036	0.0602 ± 0.0007	0.1065 ± 0.0043
SC++ ITT	MGSAE $L_{2,1}$	64	0.7538	0.6676 ± 0.0034	0.7852 ± 0.0074	0.0643 ± 0.0044	0.1505 ± 0.0036
SC++ ITT	Ordinary TopK	64	0.7538	0.7042 ± 0.0032	0.8358 ± 0.0035	0.0572 ± 0.0016	0.1069 ± 0.0029
SC++ ITT	Ours	128	0.7538	0.7146 ± 0.0052	0.8500 ± 0.0021	0.0554 ± 0.0030	0.0946 ± 0.0025
SC++ ITT	MGSAE $L_{2,1}$	128	0.7538	0.6843 ± 0.0034	0.8074 ± 0.0038	0.0615 ± 0.0026	0.1310 ± 0.0024
SC++ ITT	Ordinary TopK	128	0.7538	0.7156 ± 0.0025	0.8568 ± 0.0028	0.0525 ± 0.0014	0.0907 ± 0.0023
SC++ ITT	Ours	256	0.7538	0.7220 ± 0.0040	0.8702 ± 0.0033	0.0490 ± 0.0032	0.0808 ± 0.0017
SC++ ITT	MGSAE $L_{2,1}$	256	0.7538	0.6914 ± 0.0049	0.8341 ± 0.0063	0.0517 ± 0.0013	0.1141 ± 0.0055
SC++ ITT	Ordinary TopK	256	0.7538	0.7292 ± 0.0041	0.8864 ± 0.0076	0.0445 ± 0.0031	0.0691 ± 0.0053

Table 22: **Correction enrichment for the canonical dense tie-free model and matched MGSAE.** The benchmark-chance redraw uses $c = 1/2$ for SugarCrepe and ARO and the official group-score reference $c = 1/6$ for Winoground. Confidence intervals jointly resample examples across seeds. The replicated p-value is the largest seedwise exact binomial p-value followed by Holm correction over six benchmark–method tests.

Benchmark	Method	Dense Acc.	Corrected	Degraded	Observed C/D	Null C/D	Lift [95% CI]	Holm p
SugarCrepe	MGSAE $L_{2,1}$	0.8667	842	2000	0.421	0.154	2.74 [2.50, 2.98]	4.31e-29
SugarCrepe	Tie-free dense	0.8667	620	896	0.692	0.154	4.50 [3.98, 5.09]	2.50e-47
ARO	MGSAE $L_{2,1}$	0.5144	21326	21308	1.001	0.944	1.06 [1.04, 1.08]	1.000
ARO	Tie-free dense	0.5144	13829	9688	1.427	0.944	1.51 [1.47, 1.56]	2.55e-54
Winoground	MGSAE $L_{2,1}$	0.1075	25	60	0.417	1.660	0.25 [0.14, 0.40]	1.000
Winoground	Tie-free dense	0.1075	25	39	0.641	1.660	0.39 [0.20, 0.69]	1.000

A lift of one is the null. These rows belong to the canonical dense-model audit and are not the matched exact-K rows in Table 21.

Table 21 gives the complete matched-support accounting for SugarCrepe and the official SugarCrepe++ image-to-text (ITT) decision. At every matched K, our independently inferred TopK model has higher reconstructed accuracy and DP than MGSAE, primarily because it produces fewer degradations. This is a comparison with MGSAE, not a universal architecture win: ordinary reconstruction-only TopK is statistically tied with ours at $K = 64$ and $K = 128$ on the aggregate benchmarks and is modestly better at $K = 256$.

Benchmark-chance correction-enrichment control. The descriptive accounting above does not establish that changed decisions preferentially correct backbone errors. For the canonical dense comparison, we therefore condition on the observed number of changes and calibrate an independent redraw to each benchmark's official chance level c. If the dense accuracy is a, the null corrected-to-degraded ratio is

$$R_0 = \frac{(1-a)c}{a(1-c)}, \qquad \text{correction lift} = \frac{C/D}{R_0}. \tag{116}$$

SugarCrepe and ARO use $c = 1/2$, whereas Winoground group score uses the official $c = 1/6$ reference. Table 22 shows that dense tie-free changes are correction-enriched on SugarCrepe and ARO, and more strongly so than MGSAE in the paired audit. Neither method is correction-enriched on Winoground. ARO is retained as a diagnostic but not as a clean role-binding leaderboard because its aggregate score is known to be plausibility-confounded.

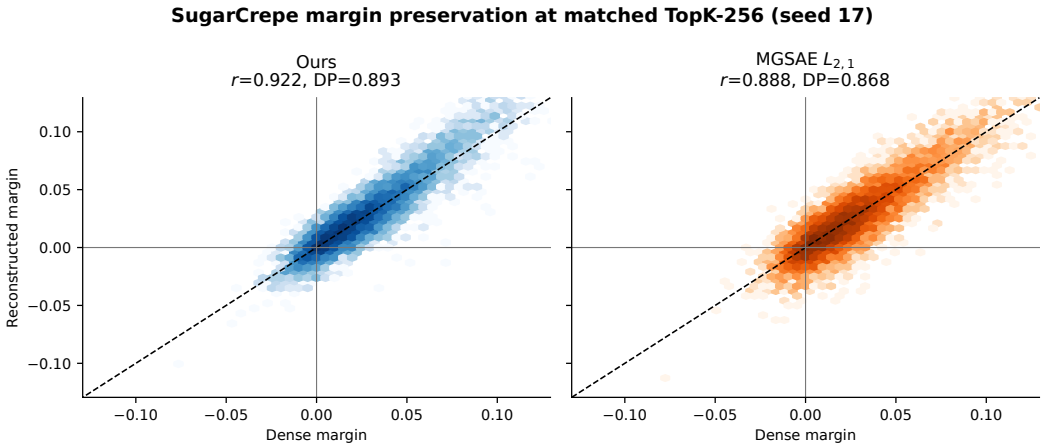

Figure 10: **SugarCrepe dense versus reconstructed margins at matched TopK-256 for seed 17.** The dashed diagonal denotes exact margin preservation. The horizontal and vertical zero lines divide preserved decisions from corrections and degradations. The panel annotations report Pearson correlation and decision preservation.

H.2 Compositional-Margin Preservation

Decision preservation records whether a benchmark decision is retained, but not how faithfully its confidence is reconstructed. For dense margin $\Delta(e)$ and reconstructed margin $\widehat{\Delta}(e)$, we report

$$\text{MarginMAE} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \left| \Delta(e) - \widehat{\Delta}(e) \right|, \tag{117}$$

$$\text{SignAgree} = \frac{1}{|\mathcal{E}|} \sum_{e \in \mathcal{E}} \mathbf{1}\left[\text{sign}(\Delta(e)) = \text{sign}(\widehat{\Delta}(e)) \right]. \tag{118}$$

For binary comparisons, sign agreement is the corresponding decision- preservation event. Pearson correlation additionally measures preservation of the linear margin geometry.

Figure 10 visualizes all SugarCrepe margins for the seed-17 TopK-256 checkpoints. Points in opposite-sign quadrants are decision changes. Our reconstruction is more tightly aligned with the dense margin than MGSAE in this comparison ($r = 0.922$ versus 0.888; DP $= 0.893$ versus 0.868).

To isolate sensitivity to the dense decision boundary, we partition examples into quartiles of $|\Delta(e)|$, with boundaries computed from the dense backbone separately for each seed. Table 23 shows that almost all decision instability is concentrated in the smallest- margin quartile. In that quartile, our model has higher DP and lower margin MAE than MGSAE. Ordinary TopK remains the strongest pure-preservation control, which is why the paper treats the result as a fidelity-semantic-coverage frontier rather than claiming universal superiority.

H.3 Exact Sparsity and the Fidelity-Sparsity Frontier

The canonical dense tie-free encoder emits a variable-size nonnegative ReLU code. To answer the separate objection that this native code is not literally sparse enough, we trained exact-support variants in which image and text are encoded independently and each modality retains its own K largest activations. These are independently trained TopK checkpoints, not a post-hoc truncation of the canonical dense checkpoint.

For a code z, we distinguish literal emitted support from concentration via the participation-ratio effective support

$$K_{\text{PR}}(z) = \frac{\|z\|_1^2}{\|z\|_2^2}. \tag{119}$$

For the hard TopK models, emitted support is exactly K (up to zero-valued edge cases), while K_{PR} can be smaller when activation mass is concentrated. The semanticity-v4 evaluator retains at most

Table 23: **SugarCrepe margin preservation by quartile of the absolute dense margin at matched TopK-256.** Quartile boundaries are computed from the dense backbone separately within each seed (the dense margins themselves are identical across methods).

Method	Bin	$\lvert\Delta\rvert$ Range	DP	Corrected	Degraded	Margin MAE
Ours	Q1	[0.0000, 0.0088]	0.6622 ± 0.0067	0.1368 ± 0.0037	0.2009 ± 0.0082	0.00715 ± 0.00012
Ours	Q2	[0.0088, 0.0209]	0.9148 ± 0.0051	0.0179 ± 0.0011	0.0673 ± 0.0043	0.00804 ± 0.00012
Ours	Q3	[0.0209, 0.0386]	0.9913 ± 0.0017	0.0002 ± 0.0003	0.0085 ± 0.0014	0.00890 ± 0.00009
Ours	Q4	[0.0386, 0.1738]	0.9984 ± 0.0000	0.0000 ± 0.0000	0.0016 ± 0.0000	0.01289 ± 0.00003
MGSAE $L_{2,1}$	Q1	[0.0000, 0.0088]	0.6439 ± 0.0164	0.1258 ± 0.0050	0.2302 ± 0.0168	0.00905 ± 0.00009
MGSAE $L_{2,1}$	Q2	[0.0088, 0.0209]	0.8448 ± 0.0054	0.0199 ± 0.0048	0.1353 ± 0.0030	0.01043 ± 0.00005
MGSAE $L_{2,1}$	Q3	[0.0209, 0.0386]	0.9610 ± 0.0063	0.0005 ± 0.0005	0.0385 ± 0.0059	0.01117 ± 0.00005
MGSAE $L_{2,1}$	Q4	[0.0386, 0.1738]	0.9975 ± 0.0008	0.0000 ± 0.0000	0.0025 ± 0.0008	0.01254 ± 0.00017
Ordinary TopK	Q1	[0.0000, 0.0088]	0.6960 ± 0.0073	0.1225 ± 0.0060	0.1816 ± 0.0033	0.00610 ± 0.00013
Ordinary TopK	Q2	[0.0088, 0.0209]	0.9444 ± 0.0041	0.0110 ± 0.0012	0.0446 ± 0.0029	0.00670 ± 0.00007
Ordinary TopK	Q3	[0.0209, 0.0386]	0.9957 ± 0.0016	0.0002 ± 0.0003	0.0041 ± 0.0016	0.00725 ± 0.00017
Ordinary TopK	Q4	[0.0386, 0.1738]	0.9980 ± 0.0003	0.0000 ± 0.0000	0.0020 ± 0.0003	0.00849 ± 0.00013

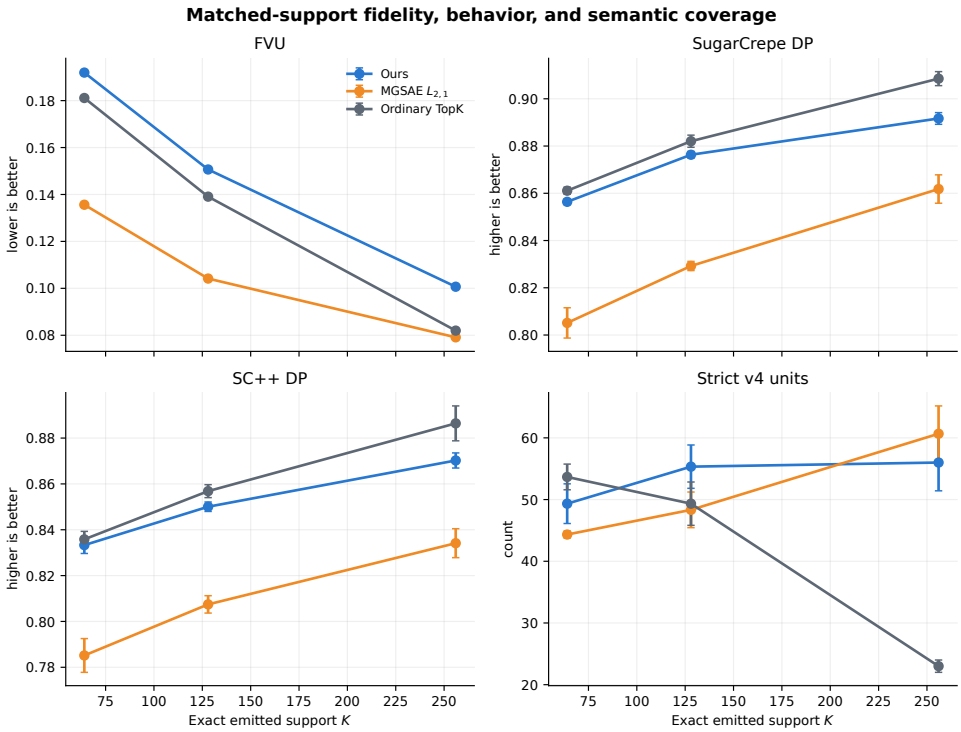

Figure 11: **Matched-support fidelity, behavior, and automatic semantic coverage on SigLIP2 ViT-B/16.** Every point is independently inferred at the displayed exact emitted support. Error bars show SD across seeds 17-19. Semanticity-v4 counts should be compared within a fixed-K block because the evaluator retains at most 64 coordinates.

64 coordinates; the reported truncation fraction therefore makes explicit how much of an emitted support is omitted before semantic scoring.

Figure 11 shows the resulting multi-objective frontier, and Table 24 reports the full core metrics. Increasing K improves FVU and behavior preservation for all three families. Our objective consistently improves SugarCrepe and SC++ DP over matched MGSAE, while MGSAE generally has better FVU and GPI. Ordinary TopK has the strongest preservation at $K = 256$ but its active dictionary and strict-unit coverage collapse sharply at larger support. Conversely, our model retains substantially more strict shared units than ordinary TopK at $K = 128$ and $K = 256$. No method dominates all axes.

Table 24: **Complete SigLIP2 width-10,240 matched-support sweep.** v4 semantic counts are directly comparable only within a fixed-K block because the evaluator retains at most 64 coordinates. Values are mean ± SD across seeds 17–19.

K	Method	Emitted Support	PR-L0	V4 Trunc.	FVU	GPI	Dead	Effective Width	Strict/Broad/Unique
64	Ours independent TopK-64	64.0 ± 0.0	37.7 ± 0.2	0.000 ± 0.000	0.1919 ± 0.0004	0.9313 ± 0.0006	0.710 ± 0.002	2972 ± 23	49.3 ± 3.2 / 250.3 ± 8.5 / 23.3 ± 1.2
64	MGSAE L2,1 TopK-64	64.0 ± 0.0	40.3 ± 0.1	0.000 ± 0.000	0.1356 ± 0.0003	0.9466 ± 0.0006	0.628 ± 0.003	3813 ± 32	44.3 ± 0.6 / 249.0 ± 2.6 / 21.3 ± 1.5
64	Ordinary TopK-64	64.0 ± 0.0	41.5 ± 0.5	0.000 ± 0.000	0.1811 ± 0.0009	0.9368 ± 0.0004	0.849 ± 0.005	1542 ± 56	53.7 ± 2.1 / 230.0 ± 2.0 / 25.7 ± 0.6
128	Ours independent TopK-128	128.0 ± 0.0	72.8 ± 0.2	0.500 ± 0.000	0.1507 ± 0.0002	0.9453 ± 0.0007	0.638 ± 0.002	3708 ± 21	55.3 ± 3.5 / 267.0 ± 12.5 / 24.7 ± 1.2
128	MGSAE L2,1 TopK-128	128.0 ± 0.0	74.6 ± 0.1	0.500 ± 0.000	0.1042 ± 0.0003	0.9606 ± 0.0005	0.516 ± 0.002	4952 ± 19	48.3 ± 2.9 / 268.3 ± 7.8 / 23.7 ± 1.5
128	Ordinary TopK-128	128.0 ± 0.0	84.7 ± 0.3	0.500 ± 0.000	0.1391 ± 0.0003	0.9517 ± 0.0003	0.849 ± 0.002	1550 ± 16	49.3 ± 3.5 / 214.0 ± 1.0 / 23.3 ± 0.6
256	Ours independent TopK-256	256.0 ± 0.0	150.3 ± 0.4	0.750 ± 0.000	0.1007 ± 0.0005	0.9621 ± 0.0005	0.608 ± 0.003	4018 ± 36	56.0 ± 4.6 / 247.0 ± 11.5 / 25.3 ± 2.3
256	MGSAE L2,1 TopK-256	256.0 ± 0.0	143.0 ± 0.5	0.750 ± 0.000	0.0791 ± 0.0003	0.9698 ± 0.0002	0.441 ± 0.002	5724 ± 21	60.7 ± 4.5 / 307.7 ± 8.5 / 24.3 ± 2.3
256	Ordinary TopK-256	256.0 ± 0.0	181.1 ± 0.7	0.750 ± 0.000	0.0820 ± 0.0002	0.9762 ± 0.0004	0.884 ± 0.003	1191 ± 33	23.0 ± 1.0 / 128.3 ± 1.2 / 14.7 ± 1.2
768	Ordinary TopK-768	768.0 ± 0.0	598.2 ± 4.5	0.917 ± 0.000	0.0107 ± 0.0004	0.9985 ± 0.0001	0.843 ± 0.002	1603 ± 22	1.7 ± 0.6 / 39.0 ± 8.9 / 1.7 ± 0.6

Table 25: **Matched-support comparisons and nearest-FVU pairs available in the frozen grid.** The nearest-FVU rows are not exact matches; FVU is shown explicitly so the residual mismatch cannot be hidden.

Match	K	Method	FVU	SugarCrepe DP	SugarCrepe Acc.	V4 Strict
support	64	Ours	0.1919	0.8564	0.8134	49.3
support	64	MGSAE $L_{2,1}$	0.1356	0.8051	0.7627	44.3
support	64	Ordinary TopK	0.1811	0.8611	0.8151	53.7
support	128	Ours	0.1507	0.8763	0.8244	55.3
support	128	MGSAE $L_{2,1}$	0.1042	0.8292	0.7810	48.3
support	128	Ordinary TopK	0.1391	0.8820	0.8239	49.3
support	256	Ours	0.1007	0.8917	0.8359	56.0
support	256	MGSAE $L_{2,1}$	0.0791	0.8618	0.8017	60.7
support	256	Ordinary TopK	0.0820	0.9085	0.8421	23.0
nearest FVU	64	Ours independent TopK-64	0.1919	0.8564	0.8134	49.3
nearest FVU	64	Ordinary TopK-64	0.1811	0.8611	0.8151	53.7
nearest FVU	128	Ours independent TopK-128	0.1507	0.8763	0.8244	55.3
nearest FVU	128	Ordinary TopK-128	0.1391	0.8820	0.8239	49.3
nearest FVU	256	Ours independent TopK-256	0.1007	0.8917	0.8359	56.0
nearest FVU	128	MGSAE L2,1 TopK-128	0.1042	0.8292	0.7810	48.3

H.4 Matched-Support and Nearest-Fidelity Comparisons

Equal K controls emitted support but not reconstruction quality. Conversely, the frozen grid does not contain exact cross-method FVU matches. We therefore report two complementary views in Table 25: (i) exact matched-support rows and (ii) the closest-FVU pair available for each of our operating points. The second view is a nearest-neighbor diagnostic, not an exact matched-fidelity experiment; both FVU values remain visible.

The closest comparison for our TopK-256 checkpoint is MGSAE TopK-128 (FVU 0.1007 versus 0.1042). Despite this small residual mismatch, our SugarCrepe DP is 0.8917 rather than 0.8292. This makes it unlikely that the entire MGSAE gap is explained by vector reconstruction alone, while avoiding the stronger and unsupported claim of exact fidelity matching.

H.5 Benchmark- and Perturbation-Level Results

Aggregate scores can conceal whether preservation is driven by easy object changes rather than the attribute, relation, and role distinctions motivating this work. Table 26 therefore reports every stored SugarCrepe and SugarCrepe++ category at the interpretable TopK-256 operating point. These include additions, replacements, and swaps of objects, attributes, and relations where provided by the benchmark.

Our model has higher DP than MGSAE in every displayed category. The remaining difficulty is also visible: dense accuracy and reconstructed accuracy are lower for relation and swap categories than for straightforward object replacement. Ordinary TopK frequently preserves these decisions even more closely, again separating behavior preservation from shared semantic coverage. ARO and Winoground are not included in this category table: ARO is retained as a plausibility-sensitive diagnostic, and the frozen SigLIP2 backbone is below the official Winoground group-score chance reference.

Table 26: **Category-level behavior at matched TopK-256.** Values are mean ± SD across seeds 17–19. Category labels are those stored by each benchmark evaluator.

Benchmark	Category	Method	Dense Acc.	Recon. Acc.	DP	Margin MAE
SugarCrepe	add_att	Ours	0.8497	0.8121 ± 0.0233	0.8478 ± 0.0096	0.00832 ± 0.00026
SugarCrepe	add_att	MGSAE $L_{2,1}$	0.8497	0.7168 ± 0.0151	0.7919 ± 0.0115	0.01098 ± 0.00054
SugarCrepe	add_att	Ordinary TopK	0.8497	0.8304 ± 0.0110	0.8776 ± 0.0008	0.00711 ± 0.00017
SugarCrepe	add_obj	Ours	0.9306	0.8885 ± 0.0015	0.9190 ± 0.0078	0.00912 ± 0.00017
SugarCrepe	add_obj	MGSAE $L_{2,1}$	0.9306	0.8624 ± 0.0061	0.8852 ± 0.0067	0.01150 ± 0.00026
SugarCrepe	add_obj	Ordinary TopK	0.9306	0.9101 ± 0.0059	0.9371 ± 0.0057	0.00743 ± 0.00005
SugarCrepe	replace_att	Ours	0.8693	0.8350 ± 0.0044	0.8913 ± 0.0015	0.00900 ± 0.00016
SugarCrepe	replace_att	MGSAE $L_{2,1}$	0.8693	0.8033 ± 0.0111	0.8604 ± 0.0154	0.01111 ± 0.00029
SugarCrepe	replace_att	Ordinary TopK	0.8693	0.8384 ± 0.0136	0.9124 ± 0.0058	0.00741 ± 0.00011
SugarCrepe	replace_obj	Ours	0.9697	0.9625 ± 0.0016	0.9726 ± 0.0003	0.01229 ± 0.00017
SugarCrepe	replace_obj	MGSAE $L_{2,1}$	0.9697	0.9455 ± 0.0034	0.9625 ± 0.0037	0.01161 ± 0.00027
SugarCrepe	replace_obj	Ordinary TopK	0.9697	0.9655 ± 0.0016	0.9796 ± 0.0035	0.00789 ± 0.00009
SugarCrepe	replace_rel	Ours	0.7404	0.7162 ± 0.0109	0.8307 ± 0.0064	0.00752 ± 0.00004
SugarCrepe	replace_rel	MGSAE $L_{2,1}$	0.7404	0.6873 ± 0.0125	0.8103 ± 0.0206	0.00893 ± 0.00009
SugarCrepe	replace_rel	Ordinary TopK	0.7404	0.7074 ± 0.0067	0.8514 ± 0.0100	0.00619 ± 0.00025
SugarCrepe	swap_att	Ours	0.7673	0.7142 ± 0.0031	0.8188 ± 0.0111	0.00809 ± 0.00021
SugarCrepe	swap_att	MGSAE $L_{2,1}$	0.7673	0.6587 ± 0.0231	0.7613 ± 0.0128	0.01067 ± 0.00016
SugarCrepe	swap_att	Ordinary TopK	0.7673	0.7022 ± 0.0057	0.8258 ± 0.0113	0.00664 ± 0.00028
SugarCrepe	swap_obj	Ours	0.6694	0.6272 ± 0.0309	0.7891 ± 0.0205	0.00620 ± 0.00021
SugarCrepe	swap_obj	MGSAE $L_{2,1}$	0.6694	0.6000 ± 0.0367	0.7565 ± 0.0379	0.00887 ± 0.00050
SugarCrepe	swap_obj	Ordinary TopK	0.6694	0.6354 ± 0.0125	0.8163 ± 0.0216	0.00550 ± 0.00036
SC++ ITT	replace_att	Ours	0.7754	0.7276 ± 0.0039	0.8735 ± 0.0015	0.01115 ± 0.00007
SC++ ITT	replace_att	MGSAE $L_{2,1}$	0.7754	0.7035 ± 0.0015	0.8232 ± 0.0078	0.01211 ± 0.00004
SC++ ITT	replace_att	Ordinary TopK	0.7754	0.7453 ± 0.0114	0.8921 ± 0.0116	0.00799 ± 0.00016
SC++ ITT	replace_obj	Ours	0.9467	0.9340 ± 0.0031	0.9619 ± 0.0006	0.01308 ± 0.00009
SC++ ITT	replace_obj	MGSAE $L_{2,1}$	0.9467	0.9098 ± 0.0068	0.9433 ± 0.0073	0.01247 ± 0.00017
SC++ ITT	replace_obj	Ordinary TopK	0.9467	0.9354 ± 0.0045	0.9637 ± 0.0050	0.00835 ± 0.00002
SC++ ITT	replace_rel	Ours	0.6053	0.5858 ± 0.0057	0.8165 ± 0.0079	0.00884 ± 0.00009
SC++ ITT	replace_rel	MGSAE $L_{2,1}$	0.6053	0.5522 ± 0.0146	0.7790 ± 0.0112	0.00961 ± 0.00004
SC++ ITT	replace_rel	Ordinary TopK	0.6053	0.5875 ± 0.0081	0.8466 ± 0.0161	0.00614 ± 0.00023
SC++ ITT	swap_att	Ours	0.6577	0.5776 ± 0.0250	0.7898 ± 0.0105	0.01015 ± 0.00014
SC++ ITT	swap_att	MGSAE $L_{2,1}$	0.6577	0.5155 ± 0.0122	0.7307 ± 0.0146	0.01110 ± 0.00011
SC++ ITT	swap_att	Ordinary TopK	0.6577	0.5906 ± 0.0202	0.8048 ± 0.0128	0.00687 ± 0.00012
SC++ ITT	swap_obj	Ours	0.4980	0.4490 ± 0.0108	0.7687 ± 0.0143	0.00731 ± 0.00012
SC++ ITT	swap_obj	MGSAE $L_{2,1}$	0.4980	0.4571 ± 0.0327	0.7306 ± 0.0245	0.00837 ± 0.00030
SC++ ITT	swap_obj	Ordinary TopK	0.4980	0.4776 ± 0.0082	0.7973 ± 0.0330	0.00527 ± 0.00003

H.6 Representation Geometry and Retrieval

All retrieval results use modality-local inference: an image code is computed without its paired caption, and a text code without its paired image. Table 27 reports reconstructed image-to-text and text-to-image R@1. R@5 and R@10 are not reported because the matched MGSAE artifacts did not save them; we do not impute unavailable values.

GPI-vt is the Pearson correlation between the flattened, mean-centered image-text cosine-similarity matrices before and after reconstruction. Support Jaccard is

$$J(z_v, z_t) = \frac{|\operatorname{supp}(z_v) \cap \operatorname{supp}(z_t)|}{|\operatorname{supp}(z_v) \cup \operatorname{supp}(z_t)|}, \tag{120}$$

averaged over independently encoded image-text pairs. Jaccard measures same-index overlap, not semantic identity, and is therefore not treated as an unconditional quality score.

Our model has higher image-to-text R@1 than MGSAE at every matched K, while text-to-image R@1 is close and reverses slightly at $K = 128$. MGSAE has higher GPI at every K. The behavioral result is therefore not explained by a uniform win on global geometry: the methods preserve different aspects of the representation.

Table 27: **Modality-local reconstructed retrieval and geometry** under matched exact sparsity. R@5/R@10 are omitted because the matched MGSAE evaluator saved only reconstructed R@1; no missing values are imputed.

K	Method	I2T R@1	T2I R@1	GPI-vt	Support Jaccard
64	Ours	0.3805 ± 0.0046	0.3933 ± 0.0030	0.9313 ± 0.0006	0.1746 ± 0.0019
64	MGSAE $L_{2,1}$	0.3311 ± 0.0021	0.3774 ± 0.0014	0.9466 ± 0.0006	0.0606 ± 0.0009
64	Ordinary TopK	0.3601 ± 0.0053	0.3869 ± 0.0027	0.9368 ± 0.0004	0.1375 ± 0.0027
128	Ours	0.4031 ± 0.0041	0.4083 ± 0.0050	0.9453 ± 0.0007	0.2201 ± 0.0017
128	MGSAE $L_{2,1}$	0.3733 ± 0.0013	0.4110 ± 0.0036	0.9606 ± 0.0005	0.0589 ± 0.0005
128	Ordinary TopK	0.3947 ± 0.0017	0.4090 ± 0.0007	0.9517 ± 0.0003	0.2486 ± 0.0041
256	Ours	0.4271 ± 0.0007	0.4413 ± 0.0086	0.9621 ± 0.0005	0.3321 ± 0.0039
256	MGSAE $L_{2,1}$	0.4069 ± 0.0033	0.4347 ± 0.0015	0.9698 ± 0.0002	0.0699 ± 0.0015
256	Ordinary TopK	0.4404 ± 0.0031	0.4546 ± 0.0009	0.9762 ± 0.0004	0.4877 ± 0.0057

Table 28: **Frozen objective and architecture controls available in the common compositional evaluator (three seeds).** This is the completed ablation set; it does not claim an ordinary-Hoyer versus Group-Hoyer comparison, which was not run in this locked shared-decoder protocol.

Variant	Sugar Acc.	Sugar DP	Sugar Margin r	ARO Acc.	ARO DP	Sugar GPI
Reconstruction Only	0.8648 ± 0.0042	0.9868	0.9982	0.5132	0.9659	0.9995
Ordinary TopK	0.8622 ± 0.0016	0.9715	0.9951	0.5092	0.9301	0.9984
Ordinary BatchTopK	0.8516 ± 0.0029	0.9617	0.9930	0.5081	0.9214	0.9981
MGSAE L2,1	0.8153 ± 0.0082	0.8739	0.9061	0.5145	0.7303	0.9793
Tie-free dense Group-Hoyer (full)	0.8545 ± 0.0025	0.9327	0.9669	0.5406	0.8512	0.9815
Tie-free dense Group-Hoyer (COCO+CC3M+VG+DOCCI)	0.8494 ± 0.0060	0.9299	0.9661	0.5440	0.8553	0.9833

H.7 Dense-Family Architecture and Objective Controls

Table 28 reports the variants actually available in the frozen common evaluator: reconstruction-only, ordinary TopK, ordinary BatchTopK, MGSAE, the full dense tie-free Group-Hoyer model trained on COCO, and the same dense objective trained on the COCO+CC3M+Visual Genome+DOCCI mixture. This is a control table across completed systems, not an exhaustive one-factor ablation. In particular, the locked artifact does not contain a direct ordinary-Hoyer versus Group-Hoyer comparison, so no such comparison is claimed here.

The table separates three outcomes. Reconstruction-only and ordinary TopK are closest to the dense behavior but do not provide the same usable shared- concept dictionary. The dense tie-free model preserves substantially more SugarCrepe behavior than MGSAE while improving ARO accuracy, although ARO is not treated as a clean compositional leaderboard. Finally, the mixed-data run does not uniformly improve the COCO-trained model: it slightly improves ARO and GPI but slightly reduces SugarCrepe accuracy and DP. Dataset scale alone is therefore not presented as the mechanism.

H.8 Objective-Ladder and AuxK Controls

We test whether a conventional AuxK dead-feature loss reproduces the dictionary utilization obtained with our complete objective. All runs in this study use SigLIP2 ViT-B/16, COCO training data, dictionary width $p = 10{,}240$, modality-local inference, and three seeds (17-19). Comparisons within each table use the same exact emitted support K. Effective width is the number of dictionary atoms that activate at least once on the evaluation set, deadness is the complementary fraction. The $K = 256$ complete-objective reference is the checkpoint from the matched-support SigLIP2 grid introduced in Appendix G.1, rather than the cross-backbone operating point in main Table 2.

High-support comparison. Table 29 compares ordinary TopK, TopK+AuxK, and our complete objective at $K = 768$. AuxK leaves effective width near that of ordinary TopK ($1{,}589$ versus $1{,}603$ atoms), whereas the complete objective activates $5{,}121$ atoms on average. The increase comes with higher FVU (0.0421 versus 0.0105 for TopK+AuxK). Because the complete objective changes several training terms, this comparison establishes a utilization-fidelity trade-off but does not identify the effect of an individual term.

Table 29: **Matched objective controls at exact support $K = 768$.** Effective width counts dictionary atoms active at least once. Strict counts use the full-support matched-K semanticity-v5 evaluator. Values are mean ± SD over seeds 17–19.

Method	FVU↓	GPI↑	Dead↓	Eff. width↑	Strict↑
Ordinary TopK	0.0107 ± 0.0004	0.9985 ± 0.0001	0.843 ± 0.002	$1{,}603 \pm 22$	3.0 ± 1.0
TopK + AuxK	$\mathbf{0.0105 \pm 0.0001}$	$\mathbf{0.9985 \pm 0.0001}$	0.845 ± 0.004	$1{,}589 \pm 38$	4.7 ± 2.5
Complete objective	0.0421 ± 0.0002	0.9807 ± 0.0002	$\mathbf{0.500 \pm 0.010}$	$\mathbf{5{,}121 \pm 100}$	$\mathbf{62.0 \pm 6.6}$

Table 30: **Objective controls at exact support $K = 256$.** Rows 0–3 and rows 2, 4–6 are separate branches: the TopK-bar hinge in row 4 replaces the positive paired auxiliary in row 3. "Ref." denotes the complete objective, not another single-term step. Strict/broad/unique counts use the full-support matched-K semanticity-v5 evaluator. Values are mean ± SD over seeds 17–19.

ID	Terms beyond reconstruction	FVU↓	GPI↑	Dead↓	Eff. width↑	Strict / broad / unique↑
0	None: ordinary TopK	0.0820 ± 0.0002	0.9748 ± 0.0003	0.884 ± 0.003	$1{,}191 \pm 33$	39.0 ± 4.4 / 200.7 ± 10.1 / 21.0 ± 1.0
1	AuxK	0.0817 ± 0.0005	0.9751 ± 0.0003	0.890 ± 0.001	$1{,}126 \pm 10$	38.7 ± 2.1 / 195.7 ± 16.6 / 20.3 ± 1.5
2	AuxK + cross-decoding	0.0881 ± 0.0003	0.9742 ± 0.0003	0.880 ± 0.001	$1{,}228 \pm 9$	56.7 ± 0.6 / 251.7 ± 19.0 / 26.0 ± 1.0
3	Row 2 + positive paired auxiliary	0.0883 ± 0.0003	0.9723 ± 0.0001	0.800 ± 0.003	$2{,}046 \pm 32$	56.3 ± 5.5 / 257.0 ± 22.3 / 23.7 ± 0.6
4	Row 2 + TopK-bar hinge	0.0887 ± 0.0002	0.9731 ± 0.0001	0.878 ± 0.001	$1{,}250 \pm 9$	60.7 ± 3.8 / 267.0 ± 18.1 / 24.7 ± 1.5
5	Row 4 + shared group mask	0.1024 ± 0.0003	0.9541 ± 0.0002	0.746 ± 0.005	$2{,}597 \pm 49$	72.0 ± 4.0 / 391.7 ± 21.1 / 26.7 ± 1.5
6	Row 5 + Group-Hoyer	0.0990 ± 0.0003	0.9601 ± 0.0003	0.750 ± 0.005	$2{,}561 \pm 52$	69.0 ± 2.0 / 380.3 ± 22.0 / 24.3 ± 0.6
Ref.	Complete objective	0.1007 ± 0.0005	0.9601 ± 0.0004	$\mathbf{0.608 \pm 0.003}$	$\mathbf{4{,}018 \pm 36}$	59.0 ± 5.6 / 342.3 ± 12.7 / 24.3 ± 2.3

Controlled objective branches. Table 30 examines exact-$K = 256$ models. Rows 0-3 form a branch adding AuxK, cross-decoding, and then the strictly positive paired auxiliary. Rows 2 and 4-6 form a separate branch: the TopK-bar hinge *replaces* the positive paired auxiliary, after which shared group masking and Group-Hoyer are added. Thus row 4 is not an addition to row 3, and the complete-objective reference is not the final step of either single-term ladder.

AuxK alone does not improve utilization. On the first branch, the positive paired auxiliary increases effective width from $1{,}228$ to $2{,}046$. On the second branch, shared group masking increases effective width from $1{,}250$ to $2{,}597$, with a corresponding fidelity cost. Adding Group-Hoyer then improves FVU from 0.1024 to 0.0990 and GPI from 0.9541 to 0.9601, without a further deadness reduction. These comparisons are not an exhaustive factorial ablation.

For this study, strict/broad/unique counts use the full-support matched-K semanticity-v5 evaluator. They are compared only within an exact-K block and are not substituted for semanticity-v4 counts reported elsewhere.

H.9 CLIP ViT-L/14 Scale Replication

We repeat the dense-versus-exact-sparse comparison on CLIP ViT-L/14 to assess whether the fidelity-automatic-coverage trade-off observed with B/16 backbones persists at a larger backbone scale. Both models use dictionary width $20{,}480$, modality-local inference, and three independently trained seeds. The dense tie-free model has no exact-K constraint; the sparse model independently selects $K = 256$ coordinates in each modality.

Table 31 shows that dense inference gives lower FVU (0.0211 versus 0.1298), higher GPI (0.9899 versus 0.9254), and higher SugarCrepe/SC++ decision preservation. The TopK-256 model exposes more automatically identified strict shared units (62.3 versus 25.0) and unique strict labels (23.0 versus 9.7). The result replicates the trade-off at this backbone scale; it does not establish that the exact-sparse model preserves behavior better.

Table 31: **Three-seed CLIP ViT-L/14 replication with dictionary width** $20{,}480$**.** DP is decision preservation relative to the frozen backbone. Strict/broad/unique counts use semanticity-v4. Values are mean ± SD where seedwise variation is available; DP entries are shown at the precision available in the saved results.

Model	FVU↓	GPI↑	Decoded R@1↑	SC DP↑	SC++ DP↑	Dead↓	Strict / broad / unique↑
Dense tie-free	$\mathbf{0.0211 \pm 0.0001}$	$\mathbf{0.9899 \pm 0.0001}$	$\mathbf{0.3534 \pm 0.0011}$	**0.9429**	**0.9364**	$\mathbf{0.3658 \pm 0.0058}$	25.0 ± 3.0 / 231.3 ± 9.5 / 9.7 ± 2.1
Independent TopK-256	0.1298 ± 0.0002	0.9254 ± 0.0004	0.3050 ± 0.0025	0.8417	0.8389	0.6630 ± 0.0026	$\mathbf{62.3 \pm 2.5}$ / $\mathbf{266.3 \pm 13.9}$ / $\mathbf{23.0 \pm 1.0}$

H.10 Per-Seed Results and Uncertainty

All primary matched-support models are trained with seeds 17, 18, and 19. Table 32 reports every TopK-256 run rather than only the across-seed mean. Hyperparameters and operating points were selected without access to compositional benchmark test labels.

For headline method differences, all methods are evaluated on the same examples and the confidence intervals use paired example-level resampling, seed-to-seed variation is reported separately. The per-seed table shows that the advantage over MGSAE is not caused by one favorable initialization. It also shows the limitation consistently: ordinary TopK remains the stronger pure-preservation reference at TopK-256.

I Q3: Detailed Semantic and Compositional Structure Analysis

This section expands the semantic analysis in the main paper. We keep three evidence classes separate. First, semanticity-v4 measures automatic cross-modal coherence and coverage on a fixed COCO validation pool. Second, split-based probes test whether relation information or individual relation-selective units generalize to disjoint examples. Third, patch/token interventions diagnose where a discovered *circuit*'s effect is expressed. These tests answer different questions: automatic label coherence does not establish context-stable semantics, decodability is not localization, and localization is not causal control.

I.1 Feature-Labeling Protocol

All automatic semantic results use the frozen evaluator *semanticity-v4-abs-topk64-coactive10*. Image and text embeddings are encoded independently. For semantic evaluation only, each code is reduced to its 64 largest-magnitude coordinates,

$$\widetilde{z}_{m,i} = \mathrm{AbsTopK}_{64}(z_{m,i}), \qquad m \in \{v, t\}. \tag{121}$$

This operation neither creates a common image-text support nor changes the code used for reconstruction. For feature j and modality m, the labeler then retains its $M = 50$ highest positive evaluation activations,

$$\mathcal{A}_j^m = \underset{M,\, i}{\arg\operatorname{top}}\, \widetilde{z}_{m,i,j}, \qquad \widetilde{z}_{m,i,j} > 0. \tag{122}$$

For each modality, the independently selected top examples are scored against both COCO object annotations and the caption-derived keyword vocabulary. The better object-or-keyword candidate under Eq. 126 is retained. The image and text sides are evaluated separately before their labels are compared.

For concept c, define

$$p_m(c \mid j) = \frac{1}{|\mathcal{A}_j^m|} \sum_{i \in \mathcal{A}_j^m} \mathbf{1}[c \in \mathcal{Y}_i^m], \tag{123}$$

where $\mathcal{Y}_i^m$ is the set of labels available for example i on side m. Feature purity and lift are

$$\mathrm{Purity}_m(j) = \max_c p_m(c \mid j), \tag{124}$$

$$\mathrm{Lift}_m(j, c) = \frac{p_m(c \mid j)}{p_m(c) + \epsilon}, \tag{125}$$

where $p_m(c)$ is the modality-specific dataset prevalence. Purity asks how often the candidate concept occurs among the top examples; lift asks how much more often it occurs than in the background

Table 32: **Per-seed behavioral results** for the matched TopK-256 comparison.

Benchmark	Method	Seed	Acc.	DP	Corrected	Degraded	Margin MAE
SugarCrepe	Ours	17	0.8394	0.8934	0.0397	0.0670	0.00922
SugarCrepe	Ours	18	0.8357	0.8928	0.0381	0.0691	0.00926
SugarCrepe	Ours	19	0.8325	0.8888	0.0385	0.0727	0.00925
SugarCrepe	MGSAE $L_{2,1}$	17	0.8077	0.8675	0.0367	0.0957	0.01076
SugarCrepe	MGSAE $L_{2,1}$	18	0.8020	0.8623	0.0365	0.1012	0.01078
SugarCrepe	MGSAE $L_{2,1}$	19	0.7952	0.8555	0.0365	0.1080	0.01085
SugarCrepe	Ordinary TopK	17	0.8432	0.9059	0.0353	0.0588	0.00719
SugarCrepe	Ordinary TopK	18	0.8432	0.9117	0.0324	0.0559	0.00717
SugarCrepe	Ordinary TopK	19	0.8400	0.9080	0.0326	0.0594	0.00704
SC++ ITT	Ours	17	0.7177	0.8713	0.0462	0.0824	0.01087
SC++ ITT	Ours	18	0.7255	0.8665	0.0526	0.0809	0.01077
SC++ ITT	Ours	19	0.7229	0.8728	0.0481	0.0790	0.01076
SC++ ITT	MGSAE $L_{2,1}$	17	0.6971	0.8407	0.0513	0.1081	0.01111
SC++ ITT	MGSAE $L_{2,1}$	18	0.6889	0.8337	0.0507	0.1156	0.01117
SC++ ITT	MGSAE $L_{2,1}$	19	0.6882	0.8280	0.0532	0.1188	0.01120
SC++ ITT	Ordinary TopK	17	0.7278	0.8781	0.0479	0.0740	0.00733
SC++ ITT	Ordinary TopK	18	0.7261	0.8882	0.0420	0.0698	0.00731
SC++ ITT	Ordinary TopK	19	0.7339	0.8930	0.0435	0.0635	0.00716

dataset. The primary concept is selected by

$$c_m^*(j) = \arg\max_c \{p_m(c \mid j)\ \log(1 + \mathrm{Lift}_m(j, c))\}\,. \tag{126}$$

This score prevents a ubiquitous but only mildly enriched concept from winning solely through purity, while also preventing a very rare one-example coincidence from winning solely through lift.

For valid labeled features, monosemanticity is one minus the normalized entropy of label occurrences in the top-activating set. If $q_{m,j}(c)$ is the normalized distribution of those occurrences and $\mathcal{C}_{m,j}$ is the set of observed labels, then

$$\mathrm{Mono}_m(j) = \begin{cases} 0, & |\mathcal{C}_{m,j}| = 0, \\ 1, & |\mathcal{C}_{m,j}| = 1, \\ 1 - \dfrac{-\sum_{c \in \mathcal{C}_{m,j}} q_{m,j}(c) \log q_{m,j}(c)}{\log |\mathcal{C}_{m,j}|}, & |\mathcal{C}_{m,j}| > 1. \end{cases} \tag{127}$$

Thus a single supported label receives one, while a unit with no usable label evidence receives zero and is excluded from conditional label summaries.

A side is *clear* only if it has at least 10 active examples, purity at least 0.35, and lift at least 2. A coordinate is *broad shared* when both sides are clear, their independently selected primary labels are equal or compatible, and the image and text codes coactivate on at least 10 paired examples. It is *strict shared* when it is broad and

$$\min\{\mathrm{Mono}_v(j), \mathrm{Mono}_t(j)\} \geq 0.25. \tag{128}$$

Finally, *unique* is the number of distinct labels represented by the strict set. It is a count, not a fraction of an arbitrarily chosen vocabulary. These definitions expose an important denominator: mean purity is conditional on units with usable evidence, whereas strict, broad, and unique counts report how much of the dictionary passes the corresponding filters.

I.2 Semantic Structure at Matched Exact Support

Table 33 reports the complete automatic semanticity-v4 comparison for independently trained SigLIP2 ViT-B/16 checkpoints at emitted supports $K \in \{64, 128, 256\}$. Because the evaluator retains at most 64 coordinates, counts are compared only within an exact-K block. In particular, the truncation fractions are 0, 0.5, and 0.75 at $K = 64$, 128, and 256, respectively; cross-K count differences combine model behavior with a changing evaluation bottleneck.

Table 33: **Automatic semanticity-v4 summary under independent evaluation-only absolute TopK-64.** Purity and label agreement are conditional diagnostics; strict, broad, and unique counts expose coverage. Values are mean ± SD over seeds 17–19 and should be compared within an exact-K block.

K	Method	Image Purity	Text Purity	Image Lift	Text Lift	Label Agreement	Active Purity	Strict	Broad	Unique
64	Ours independent TopK-64	0.5812 ± 0.0099	0.5820 ± 0.0063	20.62 ± 0.26	34.12 ± 1.09	0.6344 ± 0.0350	0.5816 ± 0.0020	49.3 ± 3.2	250.3 ± 8.5	23.3 ± 1.2
64	MGSAE L2,1 TopK-64	0.5519 ± 0.0036	0.5492 ± 0.0005	17.42 ± 0.40	38.32 ± 1.67	0.5839 ± 0.0208	0.5505 ± 0.0020	44.3 ± 0.6	249.0 ± 2.6	21.3 ± 1.5
64	Ordinary TopK-64	0.6107 ± 0.0034	0.6277 ± 0.0066	23.00 ± 0.86	37.92 ± 0.41	0.5940 ± 0.0115	0.6192 ± 0.0046	53.7 ± 2.1	230.0 ± 2.0	25.7 ± 0.6
128	Ours independent TopK-128	0.6034 ± 0.0076	0.6144 ± 0.0059	21.89 ± 0.29	37.70 ± 0.68	0.6094 ± 0.0378	0.6089 ± 0.0017	55.3 ± 3.5	267.0 ± 12.5	24.7 ± 1.2
128	MGSAE L2,1 TopK-128	0.5565 ± 0.0027	0.5660 ± 0.0018	17.37 ± 0.30	40.99 ± 0.32	0.5876 ± 0.0227	0.5613 ± 0.0005	48.3 ± 2.9	268.3 ± 7.8	23.7 ± 1.5
128	Ordinary TopK-128	0.6315 ± 0.0061	0.6531 ± 0.0020	24.89 ± 0.44	44.14 ± 0.97	0.5752 ± 0.0168	0.6423 ± 0.0038	49.3 ± 3.5	214.0 ± 1.0	23.3 ± 0.6
256	Ours independent TopK-256	0.6035 ± 0.0051	0.6070 ± 0.0072	22.69 ± 0.81	38.22 ± 0.43	0.6166 ± 0.0111	0.6053 ± 0.0060	56.0 ± 4.6	247.0 ± 11.5	25.3 ± 2.3
256	MGSAE L2,1 TopK-256	0.5544 ± 0.0026	0.5574 ± 0.0024	16.53 ± 0.49	36.70 ± 0.59	0.6044 ± 0.0108	0.5559 ± 0.0008	60.7 ± 4.5	307.7 ± 8.5	24.3 ± 2.3
256	Ordinary TopK-256	0.5839 ± 0.0031	0.5963 ± 0.0040	20.65 ± 0.84	31.66 ± 0.90	0.5314 ± 0.0359	0.5901 ± 0.0028	23.0 ± 1.0	128.3 ± 1.2	14.7 ± 1.2

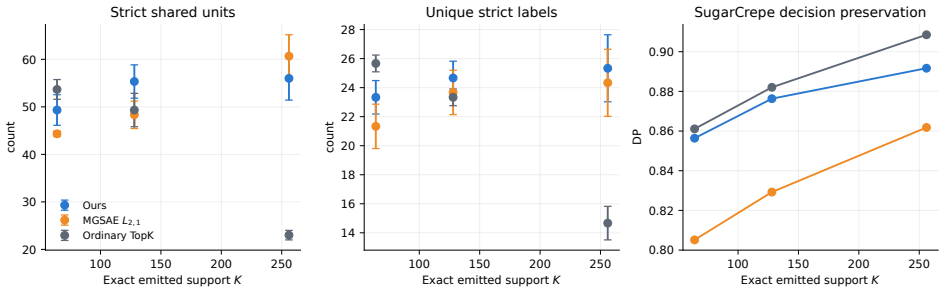

Figure 12: **Automatic semanticity-v4 coverage and SugarCrepe decision preservation under matched exact emitted support.** Error bars show SD over seeds 17-19. Strict and unique counts should be compared within a fixed K block because v4 retains at most 64 coordinates; the decision-preservation values use the full reconstructed code.

The result is a Pareto frontier, not a single-method sweep. At $K = 64$, ordinary TopK has the largest strict and unique counts. At $K = 128$, our model has the largest strict count (55.3) and unique count (24.7), while its broad coverage (267.0) is essentially tied with MGSAE (268.3) and ordinary TopK has the highest conditional purity. At $K = 256$, MGSAE has more strict and broad units, whereas our model retains one more unique label on average, higher conditional purity, and substantially more coverage than ordinary TopK. Thus our objective prevents the active-dictionary collapse seen for reconstruction-only TopK at larger support, but it does not dominate MGSAE or ordinary TopK on every semantic statistic.

Figure 12 places these conditional semantic counts beside SugarCrepe decision preservation. The semantic-count panels do not connect points across K, precisely because the v4 truncation fraction changes. The decision-preservation panel may be read as a support sweep because it is evaluated on the full reconstructed code.

I.3 Qualitative Image-Text Coordinate Matches

Figure 13 shows, for three COCO validation pairs, which semanticity-v4 labels occur at the same latent index in the independently inferred image and text codes. The examples were chosen to illustrate what a matched labeled coordinate looks like; they are not representative of aggregate counts, which favor MGSAE at K=256 (Table 24). Labels are automatic dictionary annotations rather than verified detections: in the first example, our model matches *toilet*, which is barely visible, while MGSAE matches *sink*, which our model does not.

I.4 What Has and Has Not Been Tested Out of Sample

A coherent top-activating set does not establish that a coordinate represents a reusable semantic variable. Apparent coherence can arise from feature splitting, absorption, conjunctive coding, or recurring dataset context. For example, a coherent *red car* unit need not provide separately reusable *red* and *car* variables.

The automatic v4 labels above are derived and scored on the same 5,000-example COCO validation pool. They are therefore evidence of conditional coherence and coverage, not a completed

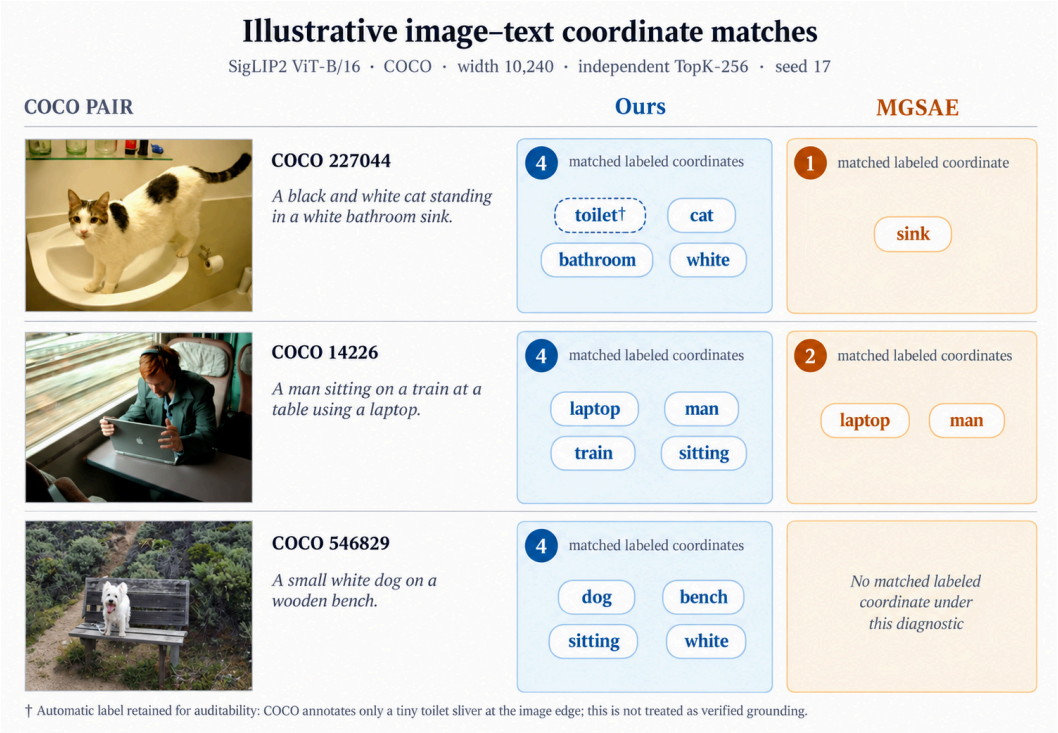

Figure 13: **Illustrative image-text coordinate matches** (SigLIP2 ViT-B/16, width 10,240, independent TopK-256, seed 17). A matched labeled coordinate is a latent index carrying the same semanticity-v4 label in both independently inferred codes. Examples illustrate the diagnostic and do not reflect aggregate counts (Table 24). †: automatic label without verified grounding.

Table 34: **Status of held-out semantic generalization tests.** The automatic v4 labeler is not a discovery/test generalization experiment; the only frozen split feature test targets low-energy relation-edit selectivity.

Analysis	Status	Supported Conclusion
Automatic semanticity-v4 labels	Same 5k COCO validation pool	Measures conditional coherence and coverage; it does not establish held-out context reuse.
Low-energy relation-unit audit	Discovery/evaluation split; 3 seeds	Zero units satisfy the prespecified strong replication rule for ours, MGSAE, or the token SAE; relation-category information is distributed.

discovery/test demonstration of context-stable object, attribute, action, or scene semantics. Table 34 records this boundary and the completed split-based feature test.

The completed split-based feature test instead asks a narrower question: whether low-energy units that discriminate relation edits on a discovery split retain the same directional selectivity on disjoint examples. The full result is reported in Table 35. Individual relation-edit associations exist-the Benjamini-Hochberg significant counts are nonzero, and the selected 32-unit sets have held-out circuit AUROC above chance-but no unit satisfies the prespecified strong cross-split replication rule for any evaluated method. The appropriate conclusion is distributed, context-dependent relation information, not stable symbolic relation neurons.

I.5 Patch/Token Causal Diagnostic

The primary SAE is trained on pooled image and text embeddings. The frozen project does *not* contain a per-SAE-feature region/span benchmark that would justify treating every pooled coordinate as a localized detector. Instead, the available experiment diagnoses where the effect of a discovered compositional circuit is expressed in the frozen dense model.

Table 35: **Held-out low-energy relation-edit selectivity on SugarCrepe.** The top 32 units are selected on the discovery split from the lower half of each method's eligible activation-energy distribution and evaluated on disjoint examples. A replicated unit must have discovery effect at least 0.5 and same-direction held-out effect at least 0.25. Values are mean ± SD over seeds 17–19.

Method	Eligible	Low-Energy Pool	Replicated	Mean AUROC	BH Sig.	Circuit AUROC	Energy/Median
Ours canonical dense tie-free	10163.3 ± 8.5	5082.0 ± 4.0	0.0 ± 0.0	0.5258 ± 0.0032	16.3 ± 3.8	0.5930 ± 0.0109	0.5692 ± 0.0677
MGSAE L2,1 K=768 (matched local reproduction)	3481.7 ± 79.5	1741.0 ± 40.0	0.0 ± 0.0	0.5105 ± 0.0037	8.7 ± 3.5	0.6098 ± 0.0191	0.4859 ± 0.0815
Ours token SAE, width 10,240	10240.0 ± 0.0	5120.0 ± 0.0	0.0 ± 0.0	0.5598 ± 0.0033	31.0 ± 1.0	0.6189 ± 0.0123	0.8431 ± 0.0266

The token-SAE row is a within-project diagnostic, not a matched reproduction of MGSAE. Relation-edit selectivity is caption-only and does not by itself establish visually grounded ordered-role binding.



Figure 14: **Representative dense patch/token causal diagnostics for an attribute swap, relation replacement, and role swap (seed 17).** The panels show where the frozen model's contrast changes under the selected intervention. They illustrate localized and distributed evidence, but do not establish that an individual pooled SAE coordinate is a grounded detector.

The circuit is selected on a discovery split using exact reconstructed-margin contributions. On held-out examples, visual evidence is perturbed by localized image-region occlusion, and textual evidence is perturbed by exact token deletion. Final-block gradient-activation scores provide an additional token localization diagnostic. In every case, the selected circuit is compared with activation-energy-matched random circuits. This design asks two separate questions: whether the selected circuit changes the decision, and whether its effect is more spatially or lexically concentrated than an equally energetic control.

The circuit produces a much larger decision-flip rate than the matched random control (0.1481 versus 0.0471), establishing non-random causal sensitivity. However, its intervention does not improve accuracy, its maximum normalized patch drop is smaller than the control, and its contrast-token hit rates are not better. Causal relevance and localization therefore separate in this experiment: the selected circuit matters to decisions, but the saved evidence does not show that it is more localized than an energy-matched circuit.

I.6 Relation Presence, Decodability, and Ordered Binding

We distinguish three increasingly demanding properties:

1. *relation presence*: information about a predicate is present somewhere in the representation,
2. *relation-category decodability*: a held-out classifier can distinguish categories such as spatial, action, or interaction relations, and
3. *ordered-role binding*: the representation identifies which entity fills each semantic role.

The first two do not imply the third. For example, *person riding horse* and *horse riding person* contain the same entities and predicate but assign the entities to different roles.

Table 37 evaluates these levels with frozen discovery/evaluation splits. The pooled relation probe is repeated in both rows because it is the same pooled-checkpoint control for the two token-SAE widths. Relation-category accuracy is approximately 0.80-0.82 for pooled and token representations, so predicate-category information is readily decodable. In contrast, Winoground ordered-pair decoding is approximately 0.505, near its binary chance level, and the strong held-out selective-unit count is zero or nearly zero.

Table 36: **Dense patch/token causal diagnostic on the 520-example held-out set (seed 17).** The discovered circuit changes decisions more than energy-matched controls, but it is not more localized in patch or token evidence.

Metric	Discovered	Matched Random	Paired Effect / Interpretation
Circuit intervention accuracy	0.5135	0.5322	-0.0187 $[-0.0680, +0.0310]$
Decision-flip rate	0.1481	0.0471	$+0.1010$ $[+0.0700, +0.1320]$
Max. normalized patch drop	0.0317	0.0909	not better than control
Contrast-token hit@1	0.3442	0.3772	-0.0329 $[-0.0582, -0.0076]$
Contrast-token hit@3	0.7048	0.7338	-0.0290 $[-0.0502, -0.0075]$

Table 37: **Relation-category decodability versus ordered-role evidence.** Probes and selective-unit tests use frozen splits; values are mean $\pm$ SD across seeds 17–19.

Token-SAE Width	Pooled Relation Probe	Token Relation Probe	Winoground Pair Probe	Held-Out Selective Units
10,240	0.8180 ± 0.0043	0.8058 ± 0.0009	0.5049 ± 0.0210	0.0 ± 0.0
40,960	0.8180 ± 0.0043	0.8008 ± 0.0031	0.5056 ± 0.0067	0.3 ± 0.6

Together with Table 35, this result supports a distributed account. Relation information can be read from population-level codes and compact circuits can affect compositional margins, but no stable single coordinate behaves as a context-independent subject-relation-object variable under the prespecified tests.

I.7 Interpretation Boundary

The supported semantic conclusion is deliberately narrower than "the SAE discovers symbolic concepts." Independently inferred codes contain many coordinates with automatically coherent, compatible image/text labels; at matched support, our objective preserves substantially more of this coverage than ordinary TopK at $K = 128$ and $K = 256$, while comparison with MGSAE remains metric- and K-dependent. Relation-category information is recoverable from pooled and token representations, and selected circuits affect held-out decisions more than energy-matched random controls.

Three stronger claims are not supported by the frozen evidence. First, v4 does not by itself establish held-out context reuse. Second, the patch/token analysis does not establish per-feature region or span grounding. Third, neither decodability nor causal circuit involvement yields clean symbolic ordered-role neurons. The exact margin decomposition and causal intervention definitions are given in Appendix D.

J Q4: Detailed Causal-Circuit Results

This section reports the complete frozen causal-circuit results summarized in the main paper. The exact reconstructed-margin decomposition, deterministic discovery/evaluation split, intervention operators, local-activity-matched controls, and cross-seed decoder matching are defined in Appendix D. The experiment uses the five dense tie-free shared-decoder SAEs trained with seeds 17-21. For each benchmark and seed, one top-32 caption-side circuit is selected using only the discovery split and is then frozen before evaluation. No held-out example is used to select a coordinate or a random control.

J.1 Concentration of Margin Contributions

For a held-out example e, let $c_j(e)$ be the exact contribution of caption coordinate j to the reconstructed image-caption contrastive margin, as defined in Eq. 81. If C_{32} is the top-32 circuit selected by discovery-set mean absolute contribution, its held-out concentration is

$$R_{32} = \frac{\sum_{e \in \mathcal{E}_{\text{eval}}} \sum_{j \in C_{32}} |c_j(e)|}{\sum_{e \in \mathcal{E}_{\text{eval}}} \sum_{j=1}^{p} |c_j(e)|}. \tag{129}$$

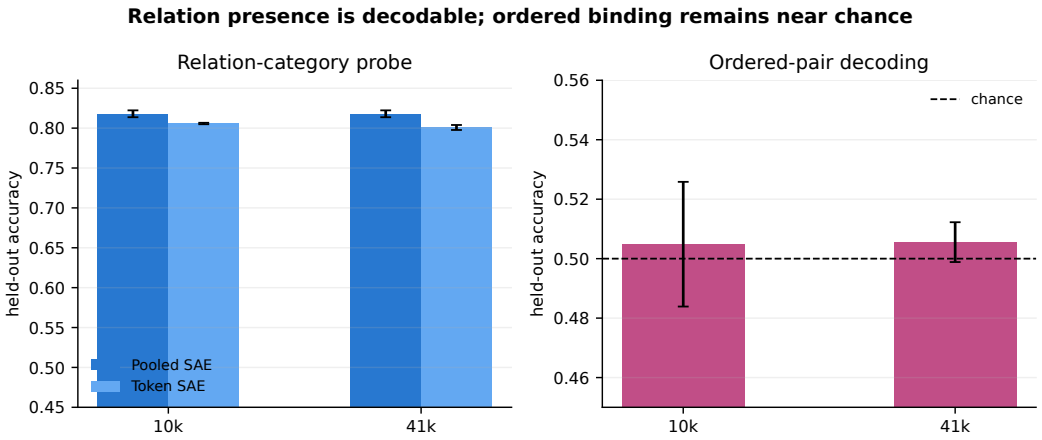

Figure 15: **Relation-category information is recoverable from pooled and token-SAE representations, while matched ordered-pair decoding remains close to chance.** Error bars show SD over seeds 17-19. The contrast demonstrates that knowing which relation is present is easier than binding its arguments in the correct order.

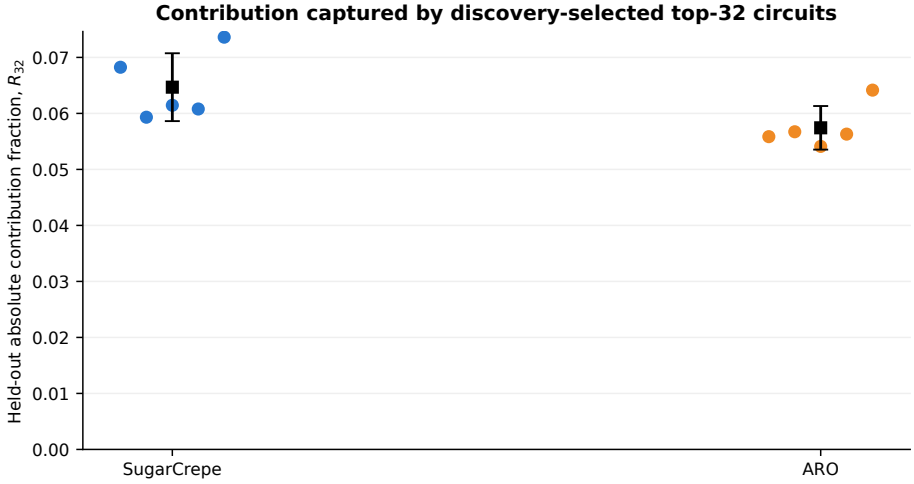

Figure 16: **Held-out absolute contribution fraction captured by the discovery-selected top-32 circuit.** Colored points are the five independently trained SAE seeds; black squares and error bars show the across-seed mean and sample SD.

This statistic asks how much absolute decision contribution is carried out of sample by the coordinates selected on the disjoint discovery split. It does not assert that all selected contributions share the same sign.

Figure 16 shows $R_{32} = 0.0647 \pm 0.0060$ on SugarCrepe and $R_{32} = 0.0574 \pm 0.0039$ on ARO. The circuit contains only $32/10{,}240 = 0.3125\%$ of dictionary coordinates but carries 5.7-6.5% of held-out absolute contribution on average. Thus its contribution share is approximately 18-21 times its coordinate share. This is evidence of concentration, not evidence that 32 coordinates are a minimal or sufficient circuit. The frozen experiment contains only $L = 32$, it does not support a cumulative circuit-size curve or category-specific minimum circuit widths.

J.2 Held-Out Circuit Ablation

Ablation sets the selected coordinates to zero in both caption codes, decodes the complete edited codes, renormalizes the reconstructed embeddings, and recomputes the margin. Negative changes in accuracy or margin mean that removing the circuit harms the reconstructed decision.

Table 38: **Held-out top-32 circuit ablation.** Random controls are matched for circuit size and discovery-set mean absolute caption activation (called *activation energy* in the frozen artifact). Values are mean ± SD over five independently trained SAE seeds.

Benchmark	Selected △acc.	Random △acc.	Selected △margin	Random △margin
SugarCrepe	-0.0115 ± 0.0126	-0.0025 ± 0.0018	-0.0025 ± 0.0010	-0.0006 ± 0.0001
ARO	0.0030 ± 0.0055	-0.0005 ± 0.0016	0.0001 ± 0.0001	-0.0000 ± 0.0001

Table 39: **Held-out top-32 activation exchange.** Semantic accuracy is the fraction of evaluation examples for which exchanging the selected positive/negative activations lowers the original compositional margin. Values are mean ± SD over five SAE seeds.

Benchmark	Sel. △acc.	Rand. △acc.	Sel. △margin	Rand. △margin	Sel. Flip	Rand. Flip	Sel. Semantic	Rand. Semantic
SugarCrepe	-0.0534 ± 0.0295	-0.0040 ± 0.0035	-0.0053 ± 0.0019	-0.0014 ± 0.0002	0.1298 ± 0.0298	0.0358 ± 0.0020	0.6366 ± 0.0319	0.6176 ± 0.0150
ARO	0.0015 ± 0.0138	0.0015 ± 0.0041	-0.0000 ± 0.0004	0.0000 ± 0.0001	0.1949 ± 0.0190	0.0630 ± 0.0010	0.4765 ± 0.0191	0.4974 ± 0.0222

On SugarCrepe, selected-circuit ablation changes reconstructed accuracy by -0.0115 ± 0.0126, compared with -0.0025 ± 0.0018 for the matched random controls (Table 38). The corresponding mean-margin changes are -0.0025 ± 0.0010 and -0.0006 ± 0.0001. Four of the five selected-circuit seeds have a negative accuracy change and the fifth is numerically zero. This supports a held-out necessity claim for the selected SugarCrepe circuit, while the across-seed variability shows that the magnitude depends on the learned dictionary.

ARO does not support the same directional conclusion. Its selected ablation changes accuracy by $+0.0030 \pm 0.0055$, and the sign differs across seeds. Because ARO aggregate accuracy is also plausibility-confounded in our benchmark audit, we treat this row as a sensitivity diagnostic rather than evidence that the circuit is uniformly necessary or beneficial.

J.3 Activation Exchange

Activation exchange replaces each selected positive-caption activation with the corresponding negative-caption activation and vice versa. It therefore tests whether the circuit carries information specific to the evaluated caption contrast, rather than merely contributing generic activation magnitude. The *flip* column reports the fraction of decisions whose sign changes after exchange. The *semantic* column reports the fraction of examples for which exchange lowers the original positive-minus-negative margin.

The clearest causal result is on SugarCrepe. Exchanging the selected circuit changes accuracy by -0.0534 ± 0.0295, versus -0.0040 ± 0.0035 for the matched controls, and changes the mean margin by -0.0053 ± 0.0019, versus -0.0014 ± 0.0002. It flips 0.1298 ± 0.0298 of decisions, compared with 0.0358 ± 0.0020 for the matched controls. Because the edited values come from the competing caption, this result is stronger than zero ablation: the selected coordinates mediate contrast-specific information in the reconstructed decision.

ARO again shows sensitivity without a stable signed effect. Selected exchange flips 0.1949 ± 0.0190 of decisions, compared with 0.0630 ± 0.0010 for controls, but its mean accuracy and margin changes are approximately zero and vary in sign across seeds. Large flip rates can therefore coexist with cancellation between corrected and degraded examples; flip rate alone is not evidence of improved compositional reasoning.

J.4 Comparison with Local-Activity-Matched Controls

Contribution-ranked coordinates can have large intervention effects simply because they are unusually active. Each selected coordinate is therefore matched to a non-selected coordinate with similar discovery-set *mean absolute caption activation*; the historical artifact calls this quantity activation energy, but it is not a squared-energy statistic. Eight size-32 control circuits are sampled per benchmark and seed. Selection and matching use only $\mathcal{E}_{\text{disc}}$, and selected and control circuits receive identical held-out interventions.

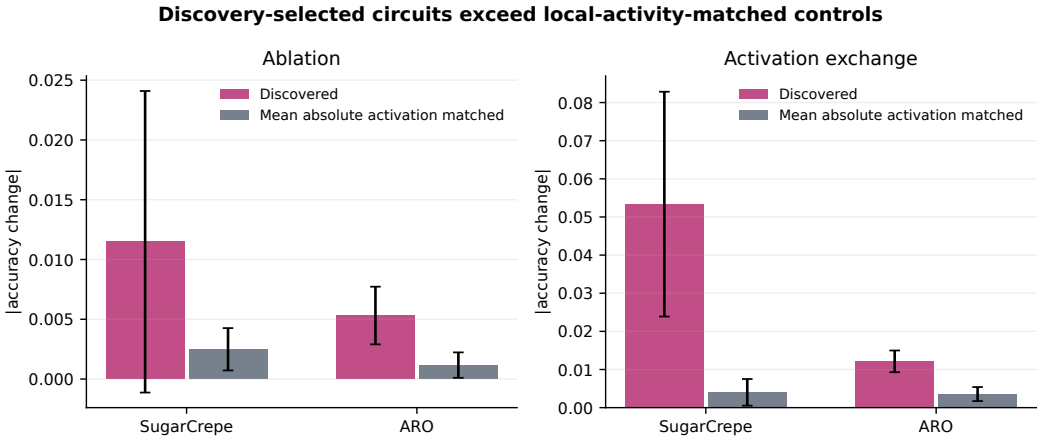

Figure 17: **Absolute held-out accuracy-change magnitude produced by discovery-selected top-32 circuits and random controls matched for size and discovery-set mean absolute caption activation.** Bars and error bars show the mean and sample SD across five SAE seeds. Absolute magnitude is plotted to compare intervention sensitivity; direction is reported in Tables 38 and 39.

Table 40: **Cross-seed recovery of reference-seed circuit rankings.** Reference seed 17 is compared with seeds 18–21. This is rank recovery, not a transferred intervention.

Benchmark	Quantity	Mean ± SD
SugarCrepe	top-32 recovery	0.3125 ± 0.0920
SugarCrepe	top-128 recovery	0.3750 ± 0.0675
SugarCrepe	matched decoder cosine	0.2190 ± 0.0250
SugarCrepe	chance top-32 recovery	0.0031
ARO	top-32 recovery	0.2578 ± 0.0861
ARO	top-128 recovery	0.3125 ± 0.0675
ARO	matched decoder cosine	0.2092 ± 0.0224
ARO	chance top-32 recovery	0.0031

Figure 17 shows that the discovered circuits produce larger absolute changes than the locally matched controls, especially for SugarCrepe activation exchange. This makes the simplest explanation-that the observed effect is solely a consequence of choosing highly active coordinates-less plausible. It does not supply a control-tail p-value: the frozen artifact retains the mean over the eight controls, not a complete inferential null distribution. The signed tables remain essential because an intervention can be large while helping some examples and harming others.

J.5 Functional Stability Across Seeds

Literal coordinate indices are not comparable across independently trained dictionaries. We therefore match seed-17 circuit atoms one-to-one to atoms in seeds 18-21 by maximum-weight Hungarian assignment on decoder cosine similarity, then ask whether the mapped atoms fall in each target seed's independently discovered contribution ranking.

For SugarCrepe, $31.25\% \pm 9.20\%$ of the mapped seed-17 top-32 atoms appear in the target seed's top 32 and $37.50\% \pm 6.75\%$ appear in its top 128. For ARO the corresponding recoveries are $25.78\% \pm 8.61\%$ and $31.25\% \pm 6.75\%$. These top-32 recoveries are far above the literal index-chance rate of $32/10{,}240 = 0.31\%$. The matched decoder cosine is nevertheless modest (0.219 ± 0.025 for SugarCrepe and 0.209 ± 0.022 for ARO), consistent with substantial rotation, splitting, or merging across learned dictionaries. The result establishes gauge-aware rank recurrence; decoder matching alone does not prove semantic identity or transferred causal function.

Table 41: Status of the proposed cross-seed intervention-transfer experiment.

Requested Analysis	Status	Supported Statement
Functional rank recovery	Completed	Top circuits recur above index-chance after gauge-aware matching.
Mapped ablation/exchange in target seed	Not executed	The frozen artifacts do not contain target-seed interventions on source-mapped coordinates; no transfer-effect claim is made.

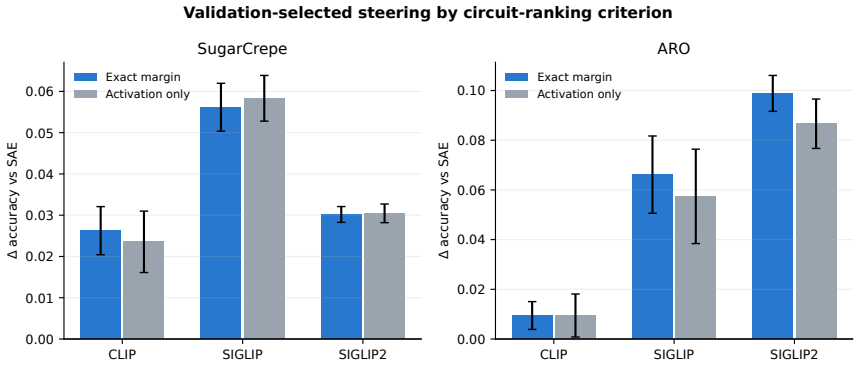

Figure 18: **Validation-selected activation-strength steering for exact-margin and activation-only circuit rankings across CLIP, SigLIP, and SigLIP2.** Bars and error bars show the mean and sample SD across three seeds of the accuracy change relative to the unmodified SAE.

J.6 Cross-Seed Circuit Transfer

The stronger experiment would ablate or exchange the source-mapped coordinates in the target seed without reranking them on the target evaluation examples. That experiment was defined in Appendix D.9 but was not executed in the frozen project. We therefore report cross-seed rank recovery, not cross-seed causal transfer. No object-, attribute-, or relation-specific transfer claim is made.

J.7 Ranking-Criterion Steering Extension

A separate three-backbone extension compares circuits ranked by the exact reconstructed-margin contribution with circuits ranked only by caption-code activation change. Intervention strength and locus are selected on validation, and the displayed accuracy changes are evaluated on the untouched test split. This experiment studies activation-strength steering; it is not the top-32 zero-ablation/exchange experiment above and does not vary circuit size.

Figure 18 gives a qualified robustness result. On SugarCrepe, exact-margin and activation-only rankings are similar: exact-minus-activation accuracy is $+0.0027 \pm 0.0019$ for CLIP, -0.0022 ± 0.0035 for SigLIP, and -0.0003 ± 0.0010 for SigLIP2. On ARO, the differences are approximately zero for CLIP and favor exact-margin ranking by 0.0088 ± 0.0038 for SigLIP and 0.0122 ± 0.0090 for SigLIP2. Exact decomposition is therefore useful for mechanistic attribution and sometimes for selecting a stronger steering set, but activation-only selection can be competitive. The experiment does not support a universal superiority claim for exact-margin ranking.

J.8 Per-Seed Results and Uncertainty

Table 42 exposes every seedwise result behind the five-seed aggregates. The principal SugarCrepe exchange effect is negative in all five seeds and larger in magnitude than the corresponding mean matched control in every seed. ARO ablation and exchange change sign across seeds, which is why only sensitivity, not a stable directional effect, is claimed. All aggregate uncertainties in this section are mean $\pm$ sample SD across independently trained SAE seeds. The frozen causal artifact does not retain the per-example intervention arrays, so we do not report paired-bootstrap confidence intervals for these particular effects.

Table 42: **Per-seed held-out top-32 circuit effects.** Random entries are means over eight control circuits matched for size and discovery-set mean absolute caption activation.

Benchmark	Seed	R_{32}	Ablate Δacc.	Random Ablate	Swap Δacc.	Random Swap	Swap Flip
SugarCrepe	17	0.0682	-0.0128	-0.0040	-0.0383	-0.0097	0.1227
SugarCrepe	18	0.0593	-0.0052	-0.0006	-0.0443	-0.0015	0.1177
SugarCrepe	19	0.0615	-0.0068	-0.0010	-0.0511	-0.0010	0.1193
SugarCrepe	20	0.0608	+0.0000	-0.0046	-0.0291	-0.0047	0.1072
SugarCrepe	21	0.0736	-0.0325	-0.0023	-0.1041	-0.0032	0.1822
ARO	17	0.0558	-0.0025	-0.0030	-0.0128	-0.0052	0.2004
ARO	18	0.0567	+0.0079	+0.0008	+0.0128	+0.0048	0.1941
ARO	19	0.0541	+0.0053	-0.0011	+0.0071	+0.0005	0.1740
ARO	20	0.0563	+0.0075	+0.0003	+0.0141	+0.0036	0.1823
ARO	21	0.0642	-0.0034	+0.0006	-0.0138	+0.0037	0.2235

J.9 Scope of the Causal Claim

These interventions act on SAE caption codes and then recompute the complete decoded and normalized embedding. They establish that the selected SAE coordinates mediate part of the *reconstructed* model's SugarCrepe behavior and that this effect is not reproduced by equally sized, local-activity-matched random circuits. They do not trace a causal pathway through the internal layers of the frozen VLM, establish that 32 coordinates are sufficient, demonstrate mapped cross-seed causal transfer, or isolate a single symbolic relation neuron. The supported conclusion is consequently narrow but substantive: exact margin decomposition discovers compact, cross-seed-recoverable SAE feature sets with non-random held-out causal effects on SugarCrepe, while ordered relation binding and robust ARO directionality remain distributed and unresolved.

K Discussion and Limitations

Our results distinguish three forms of evidence that are often conflated in multimodal SAE evaluation: whether a feature is semantically interpretable, whether SAE reconstruction preserves the backbone's behavior, and whether intervening on the feature causally changes that behavior. Reconstruction, retrieval, semanticity, and interventions evaluate different parts of this hierarchy; no single metric establishes all three. The validity audit further shows that apparent multimodal structure may be introduced by inference rather than learned by the dictionary. Pair-conditioned supports can encode sample identity, while fixed product-TopK can select coordinates unsupported by positive bilateral evidence. Modality-local inference and selector-score identifiability are therefore prerequisites for interpreting retrieval, support overlap, or shared features as evidence of cross-modal semantics. Our model satisfies these requirements while maintaining a useful trade-off among reconstruction, sparsity, semantic coverage, and compositional faithfulness. Modality-specific centering removes the dominant image-text offset, while a shared dictionary, cross-decoding, Group-Hoyer regularization, and the positive-only paired auxiliary promote transferable sparse structure without altering the modality-local inference path. The resulting features expose compact object- and attribute-sensitive circuits, but do not support a fully symbolic account of composition: ordered relations remain distributed and context-dependent. Several limitations remain. Our main experiments use frozen pooled embeddings from a limited set of contrastive VLM families, and the training data and several primary evaluations are COCO-based. Pooling may discard binding information retained in intermediate patch or token representations. Semantic measurements depend on the available annotation vocabulary and remain sensitive to feature splitting, absorption, and conjunctive coding. Moreover, our exact margin decomposition attributes one side of the reconstructed contrastive computation; jointly allocating bilinear image-text interactions requires an additional attribution convention. Finally, the interventions establish causal influence within the SAE reconstruction pathway, not a complete mechanistic circuit inside the original VLM. Extending identifiable sparse analysis to intermediate representations, additional VLM families, and generative multimodal models is an important direction for future work.