Compatible Concept Geometry for Cross-Modal Compositional Generalization

Anonymous authors
Paper under double-blind review

Abstract

Vision-language dual encoders must transfer familiar concepts to combinations absent from training. Prior work relates compositional generalization to additive and orthogonal representations within each modality, but this does not determine whether image and text factors match on unseen combinations. We study that cross-modal condition. We show that two encoders can each have additive, cross-factor orthogonal representations and still fail because their factor correspondence is wrong. We characterize the compatibility and identifiability conditions that make the correspondence recoverable. We use order-wise geometric diagnostics and Relative Orthogonal Concept Alignment (ROCA) to test the resulting predictions. Across three compositional datasets, text representations are usually closer to first-order additivity, whereas corresponding image and text factors remain misaligned. In the local retrieval evaluation, ROCA-1 obtains Top-1 accuracy of 23.60%, 3.97%, and 5.99% on dSprites, PUG-Animal, and MPI3D, respectively. It obtains 76.0% accuracy on a two-candidate CLEVR-2D attribute-swap task. Theoretical and controlled analyses show why first-order correspondence alone cannot resolve attribute-object binding: a representation must distinguish factor ownership. Together, the results separate within-modality structure, cross-modal compatibility, and binding as distinct requirements for cross-modal compositional generalization.

1 Introduction

Given a scene containing a red square and a blue circle, a model must identify the colors, shapes, and which object owns each color. This combines compositional generalization with concept binding (Treisman & Gelade, 1980; Treisman, 1998). Vision-language models such as CLIP transfer well in many zero-shot settings (Radford et al., 2021), but often behave like "bags of concepts" on unseen combinations, relation reversals, and multi-object binding (Thrush et al., 2022; Yuksekgonul et al., 2022). Recent work links compositional generalization to additive and approximately orthogonal concept representations, and links robust binding to reusable low-complexity interactions (Uselis et al., 2026a;b). These results describe each modality and the binding function separately. They do not determine whether an image factor and its textual counterpart use compatible directions on held-out combinations. The encoders may represent the same change along different directions or use different interactions for ownership. **Within-modality structure alone is therefore insufficient.**

We study the cross-modal conditions required for a dual encoder to transfer from seen concepts to unseen combinations. We distinguish factor matching between image and text from ownership decisions in multi-object scenes. Factor matching requires compatible relational structure; ownership decisions require interactions that break attribute-swap symmetry. We first show that additive, cross-factor orthogonal representations can still fail every held-out combination under an incorrect correspondence. We then characterize orthogonal compatibility, identifiability from training support, and task-level retrieval margins, and show that a sum of first-order factors cannot distinguish attribute-swapped scenes.

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.

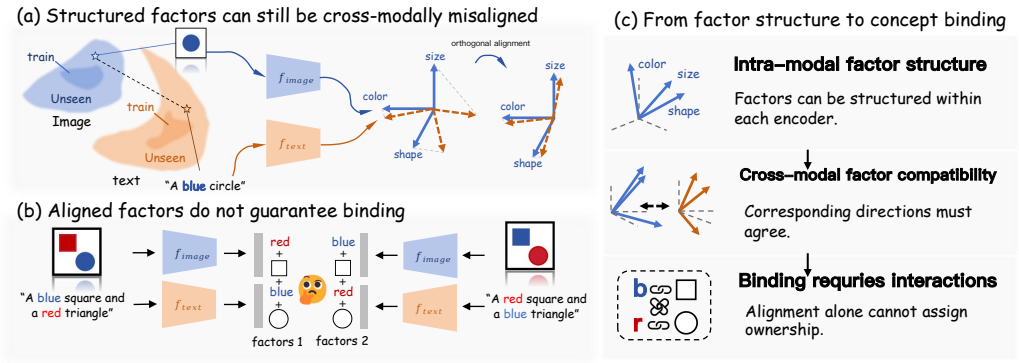


Figure 1: From factor structure to cross-modal compatibility and concept binding. (a) Structured factors may remain misaligned across modalities. (b) First-order alignment cannot distinguish attribute-swapped scenes. (c) Compositional generalization requires within-modality structure, cross-modal compatibility, and binding interactions.

We test these conditions with order-wise finite differences and held-out mapping comparisons. ROCA isolates the part of cross-modal misalignment that a relative rotation can explain. Across vision-language models and compositional datasets, text representations are generally closer to first-order additivity, while corresponding factors do not form a shared coordinate system without intervention. The local evaluation measures whether an unseen tuple outranks alternatives that change one factor. ROCA-1 reaches 23.60% with 18 candidates on dSprites, 3.97% with 141 on PUG-Animal, and 5.99% with 34 on MPI3D. Its attribute-swap accuracy on the two-candidate CLEVR-2D task is 76.0%. Controlled symbolic experiments separately examine how explicit factor-swap supervision affects binding.

Our contributions are threefold. **First,** we formalize cross-modal factor compatibility and show that within-modality additivity and orthogonality do not suffice for dual-encoder compositional generalization. **Second,** we give Gram and support conditions for compatibility and identifiability, bound the effect of alignment error on retrieval margins, and state why first-order correspondence cannot resolve multi-object binding. **Third,** order-wise diagnostics, the ROCA intervention, and controlled binding experiments distinguish orthogonal misalignment, non-isometric differences, higher-order mismatch, and missing binding structure.

2 Related Work

Compositional generalization. Controlled benchmarks evaluate vision-language models on objects, attributes, and relations (Zhao et al., 2022); word order and relation reversals (Thrush et al., 2022; Yuksekgonul et al., 2022); unseen combinations and compositional complexity (Ma et al., 2022); and hard negative examples (Hsieh et al., 2023). These studies show that models often rely on concept co-occurrence and overlook structural relations (Yamada et al., 2022), struggle to recover attribute-object compositions (Lewis et al., 2022), and may lose relation, negation, and attribute-ownership information in their text encoders (Kamath et al., 2023). Compositional hard negatives and structured image-text data can mitigate some of these failures (Yuksekgonul et al., 2022; Doveh et al., 2022). We focus on the representational conditions underlying these behavioral errors.

Geometry of learned representations. Prior theory gives conditions for compositional generalization under specific data-generating processes, training supports, or model assumptions (Wiedemer et al., 2023b;a; Uselis et al., 2025). Empirical studies find linear concept subspaces in vision-language models (Trager et al., 2023), propose the linear representation hypothesis for language models (Park et al., 2023), and observe feature superposition and approximately orthogonal encoding (Elhage et al., 2022). Disentangled and object-centric representations seek to promote out-of-distribution transfer by separating generative fac-

tors, but their effect depends on the task and training conditions (Eastwood & Williams, 2018; Montero et al., 2021; Dittadi et al., 2020). Uselis et al. (2026a) prove that, under their requirements for divisibility, transferability, and stability, compositional generalization requires linear factor decomposition and cross-factor orthogonality. We turn to dual encoders and ask what enables corresponding factors to transfer across modalities when each modality has this structure.

Cross-modal alignment. Contrastive learning maps images and text into a shared space, while the two modalities can form different global distributions and local geometries, creating a modality gap (Liang et al., 2022). Parameter sharing, within-modality regularization, and post-training have all been used to reduce this gap (Eslami & de Melo, 2024; Yamaguchi et al., 2025). Orthogonal transformations support representation comparison and structure-preserving post-processing because they preserve inner products, distances, and Gram structure (Schönemann, 1966; Kornblith et al., 2019; Li et al., 2026). Existing work focuses on global distributions, sample-level correspondence, or shared representation transforms. We examine factor directions under fixed concept correspondences and test whether alignment estimated from training support transfers to held-out contexts and unseen combinations.

Concept binding. Binding requires a model to determine which object owns each attribute or relation, a longstanding problem in cognitive science and neural representations (Treisman & Gelade, 1980; Treisman, 1998; Greff et al., 2020). CLIP can recognize individual concepts in a scene yet often fails to recover their ownership relations (Yamada et al., 2022; Lewis et al., 2022). Attribute and object information can be decoded within a single modality without being fully exploited by standard cross-modal similarity (Koishigarina et al., 2025); generalizable binding also depends on low-complexity, reusable multiplicative interactions (Uselis et al., 2026b). We treat binding as a boundary of first-order factor correspondence. Factor compatibility supports concept recombination, while multi-object scenes require interactions that break attribute-swap symmetry.

3 The Cross-Modal Composition Problem

This section formalizes the concept space, training support, and dual-encoder matching problem. We use a simple counterexample to show that regular factor structure in each modality does not guarantee correct matching on unseen combinations.

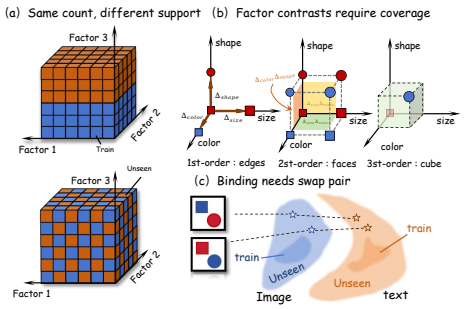


Figure 2: Training support determines which factor contrasts are identifiable. Equal sample counts can yield different design-matrix ranks and cross-modal identifiability.

3.1 Preliminaries

Concept space and training support. Suppose that an object is described by K factors, where the value set of factor i is $\mathcal{C}_i$. The complete concept space is

$$\mathcal{C} = \mathcal{C}_1 \times \cdots \times \mathcal{C}_K, \qquad \mathbf{c} = (c_1, \ldots, c_K) \in \mathcal{C}. \tag{1}$$

A scene $s = (\mathbf{c}^{(1)}, \ldots, \mathbf{c}^{(M)})$ can contain one or more objects. The same concept tuple can correspond to different backgrounds, positions, rendering instances, or text templates; we collectively denote these variations by a context ξ_m. In unseen-composition evaluation, the training and test supports contain disjoint complete tuples, while every single-factor value in the test support appears during training:

$$\Omega_{\text{tr}} \cap \Omega_{\text{te}} = \varnothing, \qquad \pi_i(\Omega_{\text{te}}) \subseteq \pi_i(\Omega_{\text{tr}}), \quad \forall i. \tag{2}$$

Dual encoders and factor structure. For modality $m \in \{I, T\}$, let the encoder output and its unit-normalized representation be

$$h_m(x) = f_m(x), \qquad z_m(x) = \frac{h_m(x)}{\|h_m(x)\|_2}. \tag{3}$$

Factor analysis uses the unnormalized conditional mean

$$\bar{h}_m(\mathbf{c}) = \mathbb{E}_{\xi_m}\left[h_m\left(x^m_{\mathbf{c},\xi_m}\right)\right], \tag{4}$$

Retrieval uses

$$\mathrm{sim}(s, s') = z_I\left(x^I_s\right)^\top z_T\left(x^T_{s'}\right). \tag{5}$$

To distinguish single-factor and interaction structure, we use a functional ANOVA decomposition:

$$\bar{h}_m(\mathbf{c}) = \mu_m + \sum_{\varnothing \neq A \subseteq [K]} \phi_{m,A}(\mathbf{c}_A), \tag{6}$$

where $|A|$ is the interaction order. The main experiments analyze orders one through three; higher-order variation is included in the residual term.

Factor compatibility and binding. For a set of semantically corresponding concept changes, we stack the unit difference directions from the two modalities as

$$D_I = [d_{I,1}, \ldots, d_{I,P}], \qquad D_T = [d_{T,1}, \ldots, d_{T,P}]. \tag{7}$$

If there exists an isometric map on the concept subspace, $Q_{\mathcal{U}}$, such that

$$D_I = Q_{\mathcal{U}} D_T, \qquad Q_{\mathcal{U}}^\top Q_{\mathcal{U}} = I, \tag{8}$$

we say that the two factor frames are orthogonally compatible. For a multi-object scene, $\mathrm{Swap}(s)$ preserves the values present for each factor and changes only their object ownership. The binding margin is

$$\gamma_{\mathrm{bind}}(s) = \mathrm{sim}(s, s) - \max_{\widetilde{s} \in \mathrm{Swap}(s)} \mathrm{sim}(s, \widetilde{s}). \tag{9}$$

Factor compatibility describes whether the same concept change can be matched across modalities; $\gamma_{\mathrm{bind}}(s) > 0$ additionally requires the model to distinguish the correct ownership relation from a factor-swapped alternative.

3.2 Datasets and Evaluation

Datasets and models. We evaluate colored dSprites, a position-downsampled MPI3D subset, and PUG-Animal. The datasets vary shape, color, size, position, camera and background factors, together with character, world, scale, texture, and viewpoint. MPI3D text descriptions use dataset factor indices for color and shape, measuring correspondence between indexed factors. The three datasets contain 3,072, 64,800, and 215,040 tuples, respectively. Each is evaluated with eight pretrained models; results are averaged over three random splits and then equally across models.

Evaluation protocols. We use tuple, context, and downstream task holdouts to test reconstruction, direction transfer, and unseen-combination generalization. Detailed split rules, candidate pools, and chance accuracies are given in Appendix B.

3.3 Failure Without Cross-Modal Correspondence

Recognizing each individual factor does not determine which complete text description matches an image. If corresponding image and text factors occupy incompatible directions, a candidate with the wrong combination can outrank the correct one even when each modality has additive, orthogonal factor structure. Figure 3 illustrates this failure mode; the construction below makes it exact.

Proposition 1 (Within-modality structure is insufficient). There exist two encoders that each use additive, cross-factor orthogonal representations and match perfectly on the training combinations, yet mis-rank all unseen combinations.

The construction is direct. Let the image encoder represent two binary factors on two orthogonal axes, while the text encoder swaps these axes. A training support containing only synchronized changes cannot identify this permutation, so the training pairs still match perfectly. On held-out combinations in which the two factors change in opposite directions, the ranking of the correct and swapped texts is exactly reversed. This construction shows that within-modality additivity and orthogonality do not determine the cross-modal correspondence.

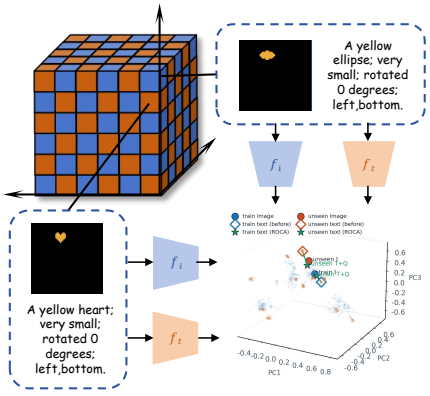


Figure 3: Cross-modal failure on an unseen composition.

4 Experiments

4.1 Intra-Modal Factor Structure

Setup and metrics. For each modality, we estimate first-, second-, and third-order reconstructions from the training support. The held-out explained variance is

$$R^2_{\le r} = 1 - \frac{\sum_{\mathbf{c}\in\Omega_{\text{te}}} \left\| \bar{h}_m(\mathbf{c}) - \widehat{h}^{(\le r)}_m(\mathbf{c}) \right\|_2^2}{\sum_{\mathbf{c}\in\Omega_{\text{te}}} \left\| \bar{h}_m(\mathbf{c}) - \widehat{\mu}_m \right\|_2^2}, \qquad \Delta R^2_r = R^2_{\le r} - R^2_{\le r-1}. \tag{10}$$

Results. Table 1 shows that text is more first-order additive on dSprites and MPI3D, while image representations retain larger second- and third-order increments. PUG-Animal follows the same image-side pattern, with text additivity varying by model family. Appendix B reports the complete order-wise spectra.

Takeaway. Both modalities have transferable low-order structure with distinct order profiles. Higher-order terms improve held-out reconstruction.

4.2 Cross-Modal Factor Misalignment

Proposition 2 (Gram condition for orthogonal compatibility). For semantically indexed direction frames D_I and D_T, an orthogonal Q satisfies $D_I = QD_T$ if and only if

$$D_I^\top D_I = D_T^\top D_T. \tag{11}$$

Orthogonal alignment preserves the pairwise angles of corresponding factor directions. **Matching Gram matrices characterize orthogonal compatibility.** The proof is in Appendix A.

Table 1: Held-out explained variance by order.

Data.	Mod.	R_1^2	R_2^2	R_3^2	Δ_2	Δ_3
dSprites	Image	0.657	0.817	0.866	0.160	0.048
dSprites	Text	0.946	0.993	0.996	0.046	0.004
PUG-Animal	Image	0.688	0.859	0.891	0.172	0.031
PUG-Animal	Text	0.799	0.942	0.980	0.142	0.038
MPI3D	Image	0.541	0.723	0.824	0.182	0.101
MPI3D	Text	0.934	0.985	0.996	0.052	0.010

Setup and metrics. For orders one through three, we fit identity, global and factor-specific orthogonal, general linear, and ridge maps. We evaluate direction cosines on held-out contexts and transfer the first-order map to higher-order responses, reporting $\tau_{1\to2}$ and $\tau_{1\to3}$.

Results. Global Q raises the held-out cosine from 0.132 to 0.653 on dSprites, from 0.070 to 0.323 on PUG-Animal, and from nearly zero to 0.452 on MPI3D. Factor-specific maps are usually higher. Linear and ridge maps fit some dSprites and MPI3D orders better, but fall below orthogonal Q on PUG-Animal at first order. Full results are in Appendix B, Table 5. Each cosine measures agreement between semantically matched factor directions across held-out contexts. Factor-specific maps allow each factor its own rotation; their larger gains reveal factor-dependent variation in the relative geometry.

Shared-geometry boundary. Gram compatibility falls with order. Transferring Q_1 gives second-order cosines of 0.369, 0.090, and 0.131, and third-order cosines of 0.191, 0.003, and 0.021, for dSprites, PUG-Animal, and MPI3D. These scores quantify the order dependence of cross-modal geometry. The first-order map transfers more strongly to second-order than to third-order responses on all three datasets. Higher response orders show weaker correspondence under the same transformation.

Takeaway. Pretrained dual encoders have a transferable relative orthogonal component. Its strength varies by dataset, factor, and response order.

4.3 When Is Factor Alignment Recoverable?

Recovering factor alignment requires informative training combinations. Let X_{tr} be the centered first-order design matrix and p its factor degrees of freedom. Under exact additivity, coefficients are unique when $\mathrm{rank}(X_{\mathrm{tr}}) = p$. For the cross-modal map, training text directions must also span the target concept subspace:

$$\mathrm{span}(D_T^{\mathrm{tr}}) = \mathcal{U}_T^{\mathrm{target}}. \tag{12}$$

When both conditions hold, the map is determined on the target subspace. Rank-deficient designs admit multiple decompositions; an incomplete training span admits orthogonal extensions with different predictions on unseen directions. With noise, stability depends on the smallest nonzero singular value. Formal statements and proofs are in Appendix A, Section A.3. Independent factor contrasts determine support identifiability. Repeated measurements of a contrast reduce noise while leaving the training span unchanged. A split covering every color, shape, or viewpoint can still confound factors when they occur only in restricted combinations. **Rank and span checks use the observed training contrasts.**

5 Improving Cross-Modal Compositional Generalization

Building on Section 4, we introduce ROCA to align image and text factor directions and evaluate its effect on unseen-composition retrieval and attribute binding.

5.1 Relative Orthogonal Concept Alignment

Factor responses. Let $A \subseteq [K]$ contain the simultaneously changed factors, with $|A| = r$. Holding the remaining context c_{-A} fixed, we define the mixed finite-difference response and its normalized direction as

$$\Delta_{m,A|c_{-A}} = \sum_{b\in\{0,1\}^r} (-1)^{r-\|b\|_1} \bar{h}_m(c_A^b, c_{-A}), \qquad d_{m,A|c_{-A}} = \frac{\Delta_{m,A|c_{-A}}}{\|\Delta_{m,A|c_{-A}}\|_2 + \varepsilon}. \tag{13}$$

First-, second-, and third-order responses change one, two, and three factors. Higher-order finite differences measure the total response to a joint change and include contributions from higher orders. Directions with the same semantic indices form $D_I^{(r)}$ and $D_T^{(r)}$. We estimate maps from training differences with all endpoints observed, then evaluate them on held-out contexts or tuples.

Relative orthogonal alignment. ROCA-1 solves the first-order unit-direction Procrustes problem

$$Q_1^* = \arg\min_{Q^\top Q = I} \left\| D_I^{(1)} - Q D_T^{(1)} \right\|_F^2 . \tag{14}$$

If $D_I^{(1)} D_T^{(1)\top} = U\Sigma V^\top$, then $Q_1^* = UV^\top$. Using the training-support centers μ_I and μ_T, downstream matching uses

$$s_Q(x,t) = \cos(h_I(x) - \mu_I, Q\,[h_T(t) - \mu_T]) . \tag{15}$$

The centered identity map isolates the effect of centering. ROCA-1+2 and ROCA-1+2+3 fit one shared Procrustes map to multiple response orders. General linear, ridge, factor-specific orthogonal, random orthogonal, and semantic-mismatch maps provide controls for non-isometric correction, factor dependence, and erroneous correspondence. Full objectives, weight selection, and metric definitions are given in Appendix B.

5.2 Results on Unseen Compositions

Setup and metrics. We fit maps on factor differences within the outer-training support and report seen/unseen H-mean and local unseen-tuple Top-1. Local pools contain 18 candidates on dSprites, 141 on PUG-Animal, and 34 on MPI3D; each negative changes one factor value.

ROCA-1 leads both metrics on all three datasets. Its local Top-1 is about 4.2, 5.6, and 2.0 times chance on dSprites, PUG-Animal, and MPI3D. Higher-order variants score below ROCA-1, with the largest decline on MPI3D. The linear and ridge controls also score below ROCA-1. Its H-means are 0.284, 0.236, and 0.212, compared with 0.006, 0.069, and 0.148 for ridge. The gains therefore extend across global retrieval and local factor discrimination. Centering alone yields small changes from raw matching, whereas aligning corresponding factor directions gives substantial gains. Shuffled and random maps remain close to zero, highlighting the role of semantic correspondence in fitting the map.

Table 2: Unseen-composition retrieval. H-mean is a fraction; local Top-1 is in percent.

	dSprites		PUG-Animal		MPI3D	
Method	H-mean	Local	H-mean	Local	H-mean	Local
Local chance	–	5.56	–	0.71	–	2.94
Raw matching	0.004	0.54	0.022	0.24	0.0004	0.01
Centered identity	0.005	0.78	0.033	0.32	0.0002	0.01
ROCA-1	**0.284**	**23.60**	**0.236**	**3.97**	**0.212**	**5.99**
ROCA-1+2	0.272	23.10	0.222	3.70	0.165	4.20
ROCA-1+2+3	0.268	23.00	0.217	3.60	0.121	3.40
General linear	0.002	0.23	0.062	1.37	0.147	3.47
Ridge linear	0.006	0.61	0.069	1.46	0.148	3.49
Equal order weights	0.174	15.01	0.085	1.20	0.090	2.52
Nested-val. weights	0.187	15.75	0.108	1.74	0.099	2.74
Globally shuffled Q	0.0001	0.02	0.00001	0.01	0.0001	0.01
Random Q	0.0004	0.07	0.0002	0.01	0.0001	0.01

Fitting and generalization. ROCA-1 improves H-mean on all three datasets and exceeds both linear controls in global and local retrieval. Linear and ridge maps can fit directions better, so **fitting accuracy does not determine candidate ranking.**

Higher orders. Adding second- and third-order responses gives small local declines on dSprites (23.60%, 23.10%, 23.00%) and larger declines on MPI3D (5.99%, 4.20%, 3.40%). Equal and nested-validation weights also score below ROCA-1. First-order alignment gives the strongest retrieval results across all three datasets.

Figure 4 illustrates the geometric correction on one held-out query. Its ranks and cosines use the original embedding space.

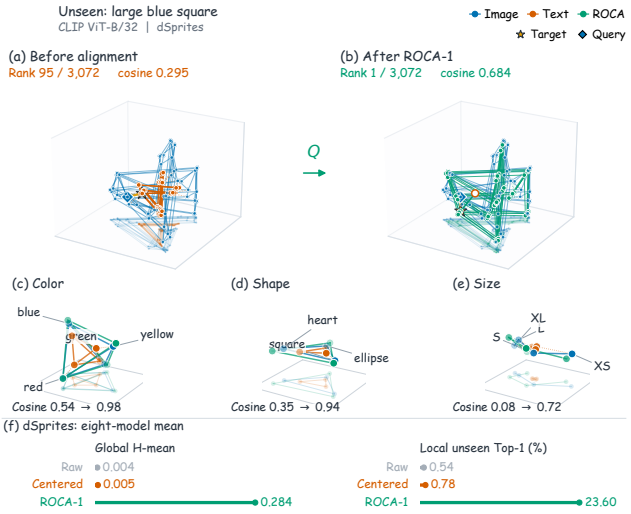


Figure 4: Relative alignment on a held-out dSprites example; retrieval means are from Table 2.

Takeaway. ROCA-1 gives the best unseen-composition retrieval on all three datasets. Better direction fit and higher-order alignment do not improve these rankings.

5.3 What Does Alignment Fix?

Improved directional compatibility does not necessarily improve accuracy. Let the minimum norm of the ideal centered representation be $b > 0$, and let the image residual and aligned-text residual be at most ϵ_I and ϵ_T, respectively. If the ideal margin between the correct candidate and the hardest negative is $\gamma^\star(s)$, then the actual margin satisfies

$$\gamma_{\widehat{Q}}(s) \geq \gamma^\star(s) - \frac{4(\epsilon_I + \epsilon_T)}{b}. \tag{16}$$

Correct ranking is guaranteed only when the ideal margin exceeds the perturbation term on the right. Average direction cosine controls neither the worst residual over all candidates nor higher-order residuals and center-estimation error, so it need not vary monotonically with accuracy. The complete proof is in Appendix A.4.

Synthetic diagnostic validation. Controlled injections separate these failure modes. Orthogonal alignment recovers rotations on the supported subspace to numerical precision, but a first-order map transfers to an independently rotated second-order subspace with cosine only 0.247 ± 0.087. A linear map exactly recovers a non-isometric shear, whereas the orthogonal fit retains an MSE of 0.0357; with rank-two support, its test cosine falls to 0.183. Direction fit is also insufficient: for composition-specific residuals, linear alignment reaches 0.945 on held-out second-order directions but only 0.473 on held-out tuples, below 0.758 for identity. Appendix B.9 gives the protocol and complete results.

5.4 The Binding Boundary

Factor alignment has a more basic boundary. If a multi-object scene is represented only as an ownership-free sum of first-order factors, a canonical scene and its factor-swapped counterpart contain the same terms and therefore have the same representation.

Even perfect cross-modal compatibility cannot recover which attribute belongs to which object. **Object-internal interactions or an equivalent binding mechanism must break this symmetry**; Appendix A.5 gives the derivation.

Table 3: CLEVR-2D hard binding subset (chance: 0.5).

Method	Swap accuracy	Std.	Binding margin
Raw matching	0.539	0.049	0.0006
Centered identity	0.549	0.060	0.0027
ROCA-1	**0.760**	0.050	**0.0220**
ROCA-1+2	0.746	0.050	**0.0220**
ROCA-1+2+3	0.746	0.050	0.0210
General linear	0.387	0.207	-0.1095
Ridge linear	0.361	0.013	-0.0312
Globally shuffled orthogonal	0.505	0.064	-0.0008
Random orthogonal	0.524	0.036	0.0012

On this two-candidate color and shape ownership task, ROCA-1 reaches 0.760 ± 0.050 accuracy and a 0.0220 binding margin. Raw matching and centered identity score 0.539 and 0.549, with margins 0.0006 and 0.0027. ROCA-1+2 and ROCA-1+2+3 both reach 0.746 accuracy, with margins 0.0220 and 0.0210. First-order alignment therefore improves both ownership accuracy and the score separation between correct and swapped descriptions.

Binding supervision in controlled dual encoders. We separate scene recognition from attribute ownership with small symbolic dual Transformers. Scene-retrieval Top-1 remains near 0.995 in every condition, while the baseline and slot-only binding margins are 0.0009 and 0.0010. Adding factor-swap negatives yields perfect swap accuracy and margins from 0.662 to 0.855 on unseen object pairs. Explicit swap supervision stabilizes ownership decisions in this controlled setting. Architecture, splits, and complete results are in Appendix B.11.

Takeaway. Relative alignment improves attribute ownership scores. In controlled dual encoders, swap supervision gives stable binding on unseen object pairs.

6 Conclusion

Main findings. We study the cross-modal representation conditions required for dual encoders to transfer familiar concepts to unseen combinations. Theory and geometric diagnostics show that within-modality additivity and cross-factor orthogonality are insufficient: corresponding image and text factors must also have identifiable, transferable relative geometry. Text representations are generally closer to first-order additive structure, while cross-modal compatibility weakens at higher response orders. ROCA-1 reaches H-mean 0.284, 0.236, and 0.212 and local unseen-tuple Top-1 of 23.60%, 3.97%, and 5.99% on dSprites, PUG-Animal, and MPI3D. Its CLEVR-2D attribute-swap accuracy is 76.0%, with a 0.0220 binding margin. First-order alignment gives the strongest retrieval and swap accuracies among the evaluated maps.

Scope and boundaries. Factor compatibility is not sufficient for correct retrieval. Factor amplitudes, within-modality residuals, mapping error, and hard negatives also determine task rankings. The MPI3D local accuracy is about twice its task-specific chance level and exceeds the linear and ridge controls. Factor correspondence also does not replace binding: first-order alignment uses existing ownership evidence, while object-sensitive interactions distinguish attribute swaps. In the controlled experiment, explicit swap supervision produces stable positive binding margins. Higher-order responses contribute according to their within-modality stability, cross-modal correspondence, and relevance to task negatives.

Future work. Future studies should test ROCA on natural multi-object scenes, use unified candidate protocols across datasets, and measure retention of zero-shot capabilities. They should also study orthogonal-map identifiability beyond the observed concept subspace and separate finite-support ANOVA estimation error from higher-order interactions.

AI use statement

Generative AI tools were used for language polishing.

Ethics statement

We use public benchmark data and frozen pretrained models. The work involves no human subjects, personal information, or deployment decisions. We state the limits of the controlled datasets and do not claim transfer to unrestricted natural scenes.

Reproducibility statement

The paper and appendices specify the datasets, splits, preprocessing, candidate pools, fitting procedures, validation rules, metrics, proofs, and reported aggregates. The supplementary code package contains the core ROCA implementation, frozen-feature evaluation, and synthetic diagnostics. The full codebase will be publicly released upon acceptance.

References

Andrea Dittadi, Frederik Träuble, Francesco Locatello, Manuel Wüthrich, Vaibhav Agrawal, Ole Winther, Stefan Bauer, and Bernhard Schölkopf. On the transfer of disentangled representations in realistic settings. *arXiv preprint arXiv:2010.14407*, 2020. URL https://arxiv.org/abs/2010.14407.

Sivan Doveh, Assaf Arbelle, Sivan Harary, Rameswar Panda, Roei Herzig, Eli Schwartz, Donghyun Kim, Raja Giryes, Rogerio Feris, Shimon Ullman, and Leonid Karlinsky. Teaching structured vision&language concepts to vision&language models. *arXiv preprint arXiv:2211.11733*, 2022. URL https://arxiv.org/abs/2211.11733.

Cian Eastwood and Christopher K. I. Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *arXiv preprint arXiv:2209.10652*, 2022. URL https://arxiv.org/abs/2209.10652.

Sedigheh Eslami and Gerard de Melo. Mitigate the gap: Investigating approaches for improving cross-modal alignment in CLIP. *arXiv preprint arXiv:2406.17639*, 2024. URL https://arxiv.org/abs/2406.17639.

Klaus Greff, Sjoerd van Steenkiste, and Jürgen Schmidhuber. On the binding problem in artificial neural networks. *arXiv preprint arXiv:2012.05208*, 2020. URL https://arxiv.org/abs/2012.05208.

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

Amita Kamath, Jack Hessel, and Kai-Wei Chang. Text encoders bottleneck compositionality in contrastive vision-language models. *arXiv preprint arXiv:2305.14897*, 2023. URL https://arxiv.org/abs/2305.14897.

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

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. *arXiv preprint arXiv:1905.00414*, 2019. URL https://arxiv.org/abs/1905.00414.

Martha Lewis, Nihal V. Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H. Bach, and Ellie Pavlick. Does CLIP bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*, 2022. URL https://arxiv.org/abs/2212.10537.

Zesheng Li, Chengchang Pan, and Honggang Qi. GraSP-VL: Length as a semantic granularity interface for vision-language representations. *arXiv preprint arXiv:2605.17727*, 2026. URL https://arxiv.org/abs/2605.17727.

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

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? *arXiv preprint arXiv:2212.07796*, 2022. URL https://arxiv.org/abs/2212.07796.

Milton L. Montero, Casimir J. H. Ludwig, Rui Ponte Costa, Gaurav Malhotra, and Jeffrey S. Bowers. The role of disentanglement in generalisation. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=qbH974jKUVy.

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

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. *arXiv preprint arXiv:2103.00020*, 2021. URL https://arxiv.org/abs/2103.00020.

Peter H. Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966. doi: 10.1007/BF02289451. URL https://doi.org/10.1007/BF02289451.

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. *arXiv preprint arXiv:2204.03162*, 2022. URL https://arxiv.org/abs/2204.03162.

Matthew Trager, Pramuditha Perera, Luca Zancato, Alessandro Achille, Parminder Bhatia, and Stefano Soatto. Linear spaces of meanings: Compositional structures in vision-language models. *arXiv preprint arXiv:2302.14383*, 2023. URL https://arxiv.org/abs/2302.14383.

Anne Treisman. Feature binding, attention and object perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353(1373):1295–1306, 1998. doi: 10.1098/rstb.1998.0284. URL https://doi.org/10.1098/rstb.1998.0284.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. doi: 10.1016/0010-0285(80)90005-5. URL https://doi.org/10.1016/0010-0285(80)90005-5.

Arnas Uselis, Andrea Dittadi, and Seong Joon Oh. Does data scaling lead to visual compositional generalization? *arXiv preprint arXiv:2507.07102*, 2025. URL https://arxiv.org/abs/2507.07102.

Arnas Uselis, Andrea Dittadi, and Seong Joon Oh. Compositional generalization requires linear, orthogonal representations in vision embedding models. *arXiv preprint arXiv:2602.24264*, 2026a. URL https://arxiv.org/abs/2602.24264.

Arnas Uselis, Darina Koishigarina, and Seong Joon Oh. How can embedding models bind concepts? *arXiv preprint arXiv:2605.31503*, 2026b. URL https://arxiv.org/abs/2605.31503.

Thaddäus Wiedemer, Jack Brady, Alexander Panfilov, Attila Juhos, Matthias Bethge, and Wieland Brendel. Provable compositional generalization for object-centric learning. *arXiv preprint arXiv:2310.05327*, 2023a. URL https://arxiv.org/abs/2310.05327.

Thaddäus Wiedemer, Prasanna Mayilvahanan, Matthias Bethge, and Wieland Brendel. Compositional generalization from first principles. *arXiv preprint arXiv:2307.05596*, 2023b. URL https://arxiv.org/abs/2307.05596.

Yutaro Yamada, Yingtian Tang, Yoyo Zhang, and Ilker Yildirim. When are lemons purple? the concept association bias of vision-language models. *arXiv preprint arXiv:2212.12043*, 2022. URL https://arxiv.org/abs/2212.12043.

Shin'ya Yamaguchi, Dewei Feng, Sekitoshi Kanai, Kazuki Adachi, and Daiki Chijiwa. Post-pre-training for modality alignment in vision-language foundation models. *arXiv preprint arXiv:2504.12717*, 2025. URL https://arxiv.org/abs/2504.12717.

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? *arXiv preprint arXiv:2210.01936*, 2022. URL https://arxiv.org/abs/2210.01936.

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

Supplementary Material

Contents

A Proofs 13
A.1 Within-modality counterexample 13
A.2 Gram characterization 14
A.3 Support identifiability 14
A.4 Retrieval-margin bound 15
A.5 Binding boundary 16
B Experimental details 16
B.1 Datasets and evaluation protocols 17
B.2 Order-wise reconstruction 17
B.3 Finite-difference responses 19
B.4 Relative orthogonal intervention 19
B.5 Multi-order alignment 20
B.6 Baseline mappings 20
B.7 Geometry and task metrics 21
B.8 Unseen-composition evaluation 21
B.9 Synthetic error injection and recovery 22
B.10 Binding swaps in pretrained models 22
B.11 Binding supervision in controlled dual encoders 23
C Per-model results 24
C.1 dSprites 24
C.2 PUG-Animal 25
C.3 MPI3D 27

A Proofs

The main text states the propositions and their proof ideas. This appendix gives the complete constructions and derivations.

A.1 Proof of Proposition 1

Proposition 1 (Within-modality structure is insufficient). There exist two encoders that each use additive, cross-factor orthogonal representations and match perfectly on the training combinations, yet mis-rank all unseen combinations.

Proof. Consider two binary factors $c_1, c_2 \in \{-1, +1\}$. Let e_1, e_2 be unit orthogonal basis vectors, and define the unnormalized image and text representations by

$$h_I(c_1, c_2) = c_1 e_1 + c_2 e_2, \qquad e_1^\top e_2 = 0, \tag{A.1}$$

$$h_T(c_1, c_2) = c_1 f_1 + c_2 f_2, \qquad f_1^\top f_2 = 0. \tag{A.2}$$

Both encoders satisfy first-order additivity and cross-factor orthogonality. Let $f_1 = e_2$ and $f_2 = e_1$, so that the text coordinates for the two factors are permuted. For $c = (+1, -1)$, the image representation is $e_1 - e_2$, the correct text representation is $e_2 - e_1$, and the swapped text representation is $e_1 - e_2$. Hence, under identity matching in the shared space,

$$h_I(+1, -1)^\top h_T(+1, -1) = -2, \qquad h_I(+1, -1)^\top h_T(-1, +1) = 2. \tag{A.3}$$

Take the training support to be $\{(+1, +1), (-1, -1)\}$ and the test support to be $\{(+1, -1), (-1, +1)\}$. The two encoders match exactly on the training combinations, while every test combination is matched to the other, incorrect combination. All representations have norm $\sqrt{2}$, so normalization does not change the ranking. □

The two factor frames remain orthogonally compatible: a permutation matrix recovers their correspondence. The example shows that the existence of an orthogonal map does not imply that the original shared space already uses the correct correspondence; the training support must also identify the map. Here the support contains only synchronized factor changes, exactly the unidentifiable case discussed in Section 4.3.

A.2 Proof of the Gram Characterization

Proposition 2 (Gram condition for orthogonal congruence). Let image and text factor directions with the same semantic indices be

$$D_I = [d_{I,1}, \ldots, d_{I,P}], \qquad D_T = [d_{T,1}, \ldots, d_{T,P}] \in \mathbb{R}^{d \times P}. \tag{A.4}$$

There exists an orthogonal matrix $Q \in \mathbb{R}^{d \times d}$ such that

$$D_I = Q D_T, \qquad Q^\top Q = I, \tag{A.5}$$

if and only if $D_I^\top D_I = D_T^\top D_T$; neither matrix needs to have full rank in $\mathbb{R}^d$.

Proof. Necessity follows from orthogonal invariance:

$$D_I^\top D_I = D_T^\top Q^\top Q D_T = D_T^\top D_T. \tag{A.6}$$

Conversely, if

$$D_I^\top D_I = D_T^\top D_T, \tag{A.7}$$

then $\|D_T a\|^2 = \|D_I a\|^2$ for every a. Thus the map $D_T a \mapsto D_I a$ is well-defined and preserves inner products. Completing orthonormal bases of the two image spaces extends this map to an orthogonal transformation on all of $\mathbb{R}^d$. □

In practice, the condition is approximate. We measure the Gram discrepancy by

$$\delta_{\mathrm{Gram}} = \frac{\left\| D_I^\top D_I - D_T^\top D_T \right\|_F}{\left\| D_I^\top D_I \right\|_F + \left\| D_T^\top D_T \right\|_F}. \tag{A.8}$$

We also report Gram cosine and principal angles between the factor subspaces. Because the columns of each frame are normalized, this test describes their relative angles and does not constrain the amplitudes of the original factor vectors. Gram agreement of direction frames therefore does not imply that the unnormalized factor vectors can be exactly aligned by a single orthogonal map.

A.3 Proof of the Identifiability Result

Proposition 3 (Support identifiability). Under an exact additive model, let Ω_{tr} be the training combination set and let the centered first-order factor design matrix be

$$X_{\mathrm{tr}} \in \mathbb{R}^{N_{\mathrm{tr}} \times p}, \tag{A.9}$$

where p is the number of first-order degrees of freedom under the zero-mean constraints. If

$$\mathrm{rank}(X_{\mathrm{tr}}) = p, \tag{A.10}$$

then the first-order factor coefficients are unique; otherwise, distinct factor coefficients produce the same representations on the training support.

Proof. Full column rank gives a unique solution to the linear system. Otherwise there is a nonzero a with $X_{\mathrm{tr}}a = 0$. For any embedding vector b, adding $ab^\top$ to the factor coefficients leaves all training predictions unchanged while potentially changing predictions on unseen combinations. Repeatedly sampling the same combinations cannot remove this ambiguity when the factors always co-occur. □

Cross-modal mapping also requires the training directions to cover the target text concept subspace. Let

$$\mathcal{U}_T^{\mathrm{target}} = \mathrm{span}(D_T^{\mathrm{target}}). \tag{A.11}$$

If the training direction matrix is D_T^{tr}, estimating the map on the target subspace requires

$$\mathrm{span}(D_T^{\mathrm{tr}}) = \mathcal{U}_T^{\mathrm{target}}. \tag{A.12}$$

Under an exact orthogonal relationship, the action of the map on this subspace is uniquely determined when the training directions span it; its extension on the orthogonal complement may remain non-unique. If the target contains a direction outside the training span, two orthogonal extensions can agree on every training direction and disagree on that target direction. With finite noise, the smallest nonzero singular value of the design and direction matrices also matters: a full-rank but ill-conditioned support can still amplify estimation error.

A.4 Proof of the Retrieval-Margin Bound

Alignment error and retrieval margin. Direction compatibility does not constrain factor amplitudes and does not by itself guarantee that the correct candidate outranks every negative. We connect unnormalized factor error to cosine retrieval.

Let the ideal centered image representation be $a(s) = \sum_i u_{I,i}(c_i)$, and assume $\|a(s)\| \geq b > 0$ for every candidate. Write the actual centered image and aligned text representations as

$$h_I(s) - \mu_I = a(s) + e_I(s), \qquad \widehat{Q}[h_T(s) - \mu_T] = a(s) + e_T(s). \tag{A.13}$$

For a reference map $Q^\star$, the text error decomposes as

$$\begin{aligned} e_T(s) = &\sum_i [Q^\star u_{T,i}(c_i) - u_{I,i}(c_i)] \\ &+ (\widehat{Q} - Q^\star) \sum_i u_{T,i}(c_i) + \widehat{Q}\, r_T^{(>1)}(s). \end{aligned} \tag{A.14}$$

The three terms are unnormalized factor alignment error, finite-support map estimation error, and the text higher-order residual; center estimation error is included in e_T. Thus, even if unit directions align, a mismatch in factor amplitudes can remain large.

Proposition 4 (Perturbation bound for the retrieval margin). Suppose $\|e_I(s)\| \leq \epsilon_I$ and $\|e_T(s)\| \leq \epsilon_T$ for the test query and all candidates, and the centered representations are nonzero. Let $z^\star(s) = a(s)/\|a(s)\|$ and define the ideal candidate margin

$$\gamma^\star(s) = 1 - \max_{s' \neq s} \langle z^\star(s), z^\star(s') \rangle. \tag{A.15}$$

Then the actual margin after alignment satisfies

$$\gamma_{\widehat{Q}}(s) \geq \gamma^\star(s) - \frac{4(\epsilon_I + \epsilon_T)}{b}. \tag{A.16}$$

Proof. For nonzero a and $a + e$, normalized perturbation obeys

$$\left\| \frac{a+e}{\|a+e\|} - \frac{a}{\|a\|} \right\| \leq \frac{2\|e\|}{\|a\|}. \tag{A.17}$$

By Cauchy–Schwarz, the deviation of any image-text score from its ideal score is at most $2(\epsilon_I + \epsilon_T)/b$. The correct and incorrect scores each incur one such perturbation, which gives (A.16). □

Consequently,

$$\gamma^\star(s) > \frac{4(\epsilon_I + \epsilon_T)}{b} \tag{A.18}$$

is sufficient to preserve the correct ranking. This bound explains the role of compatibility: reducing factor misalignment tightens the representation error, but correct ranking additionally depends on higher-order residuals, estimation error, and candidate separation. Average direction cosine is not a uniform error bound in (A.18), so improving it need not improve accuracy monotonically.

A.5 Proof of the Binding Boundary

All statements below concern the pre-normalized encoding space. Consider the two-object scenes

$$s = (\text{red square}, \text{blue circle}), \qquad s_{\text{swap}} = (\text{blue square}, \text{red circle}). \tag{A.19}$$

The scenes contain the same color and shape values and differ only in object ownership. If a scene representation contains only a sum of first-order factors, then

$$h_m^{(1)}(s) = \mu_m + \sum_{r=1}^{2} \left[u_{m,\text{color}}\left(c_{\text{color}}^{(r)}\right) + u_{m,\text{shape}}\left(c_{\text{shape}}^{(r)}\right) \right]. \tag{A.20}$$

Swapping the objects does not change this sum, so

$$h_m^{(1)}(s) = h_m^{(1)}(s_{\text{swap}}). \tag{A.21}$$

First-order representations cannot distinguish the two scenes, even when the image and text factors are perfectly compatible. Breaking this symmetry requires interactions sensitive to factors co-occurring within the same object, for example

$$h_m(s) = h_m^{(1)}(s) + \sum_{r=1}^{2} v_{m,\text{color,shape}}\left(c_{\text{color}}^{(r)}, c_{\text{shape}}^{(r)}\right) + \cdots. \tag{A.22}$$

The binding margin is

$$\gamma_{\text{bind}}(s) = \text{sim}(s, s) - \text{sim}(s, s_{\text{swap}}). \tag{A.23}$$

Positive margin means that the model prefers the correct ownership; zero margin means that it cannot distinguish the correct and swapped scenes. ROCA can correct relative misalignment of first-order factors, but it cannot create object-internal interactions that are absent from the representation.

B Experimental Details

This appendix gives the multi-order objectives, baseline mappings, weight selection procedure, candidate pools, and geometry metrics omitted from the main text.

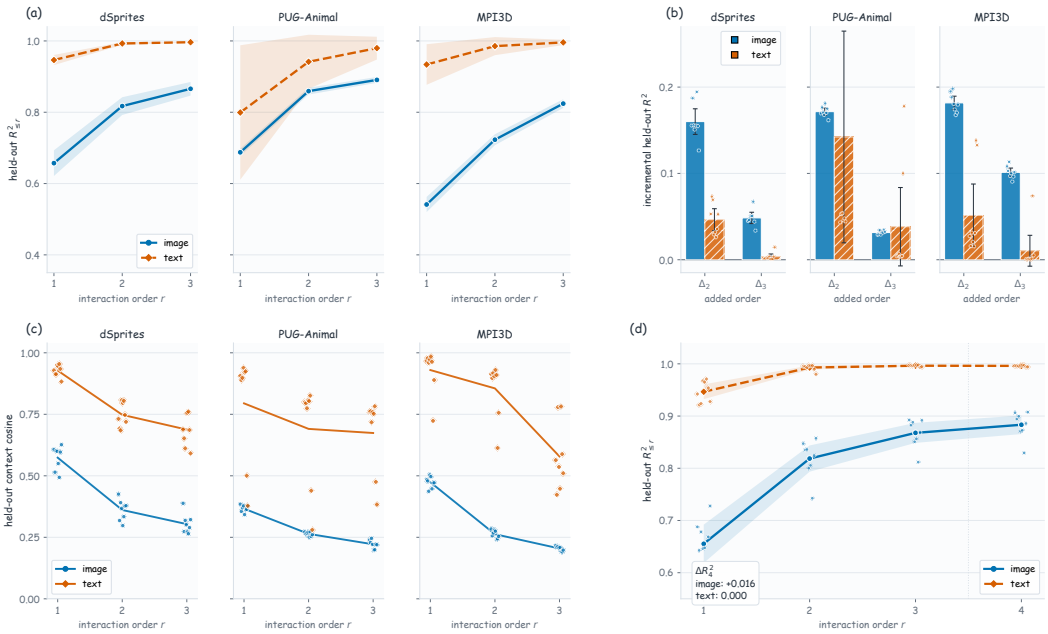


Figure 5: Order-wise reconstruction and context stability. (a) Held-out explained variance through order three. (b) Second- and third-order increments. (c) Context stability of factor directions. (d) Fourth-order sensitivity on an independent dSprites split.

B.1 Datasets and evaluation protocols

Splits. Within-modality structure analysis uses an 80/20 tuple holdout, with each order-wise reconstruction estimated from the training support. Direction analysis holds out the remaining contexts within each factor set to measure contextual transfer of the maps. For downstream interventions, tuples are split first, and differences are constructed from training tuples whose two endpoints are both available; order weights are selected only by inner validation. These protocols test representation reconstruction, direction transfer, and task-level generalization, respectively, and do not treat context holdout as complete global tuple disjointness.

Candidate sets. We report global-candidate retrieval and local concept discrimination separately. Local candidates contain the correct tuple and Hamming-1 negatives that change exactly one factor value. The number of local candidates is fixed within each dataset, whereas PUG-Animal and MPI3D use different global screening-pool sizes. Global results for these two datasets compare methods only within the same model and candidate pool. Figure 5 shows held-out reconstruction and order increments for image and text representations, context stability of factor directions, and the fourth-order dSprites check.

Table 4: Candidate counts and chance accuracies for downstream evaluation.

Dataset	Model setting	Global N	G. chance	Local N	L. chance
dSprites	All 8	3,072	0.0326%	18	5.56%
PUG-Animal	7, excl. OpenCLIP-L/14	8,192	0.0122%	141	0.71%
PUG-Animal	OpenCLIP-L/14	2,048	0.0488%	141	0.71%
MPI3D	OpenAI-B/16 + OpenCLIP-B/16	8,192	0.0122%	34	2.94%
MPI3D	Other 6	2,048	0.0488%	34	2.94%

B.2 Order-wise reconstruction details

On the complete balanced grid, each order is obtained by subtracting the lower-order components from the conditional means and imposing zero marginal means along each factor axis.

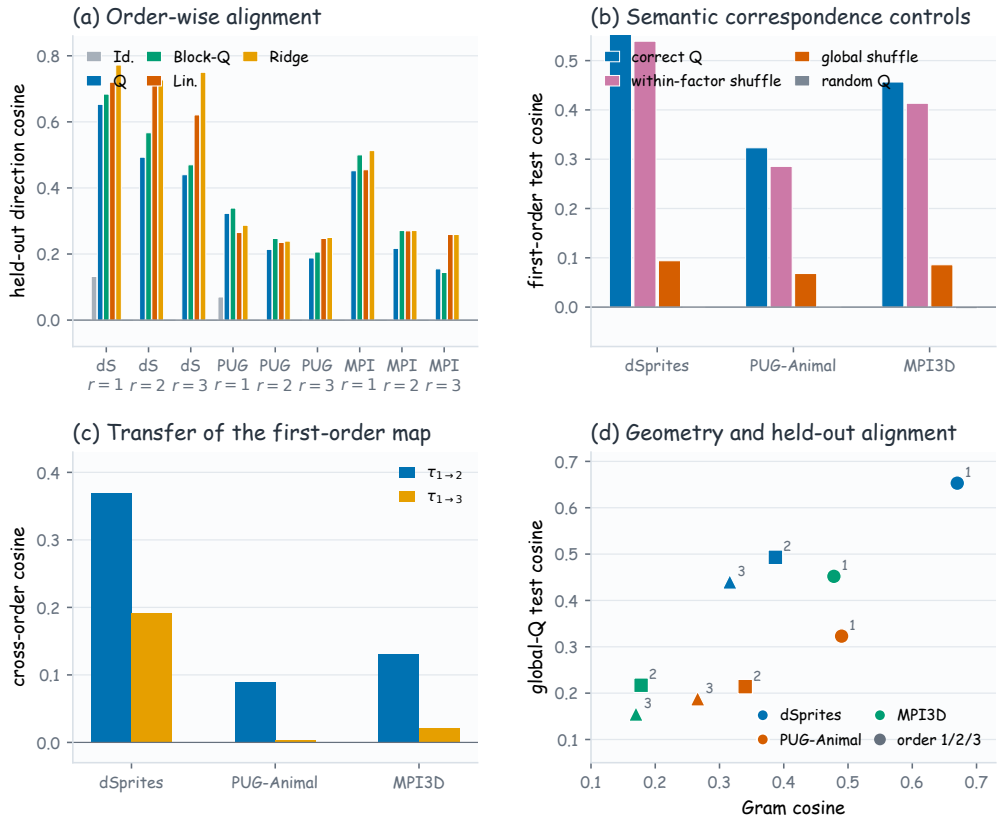


Figure 6: Order-wise cross-modal factor geometry.

Table 5: Held-out cross-modal geometry by interaction order.

Data.	r	Id.	Q	Fact.-Q	Lin.	Ridge	Gram	Ang.	MSE	ρ_r
dSprites	1	0.132	0.653	0.684	0.720	0.772	0.670	38.6	0.694	0.601
dSprites	2	-0.001	0.493	0.567	0.733	0.727	0.387	46.0	1.014	0.494
dSprites	3	0.001	0.440	0.470	0.621	0.750	0.316	49.0	1.121	0.439
PUG-Animal	1	0.070	0.323	0.339	0.265	0.287	0.490	61.1	1.354	0.274
PUG-Animal	2	0.002	0.214	0.247	0.235	0.239	0.340	58.0	1.573	0.212
PUG-Animal	3	-0.001	0.188	0.206	0.247	0.250	0.266	57.4	1.624	0.189
MPI3D	1	0.000	0.452	0.500	0.455	0.513	0.478	38.8	1.095	0.452
MPI3D	2	0.000	0.217	0.271	0.270	0.271	0.178	54.9	1.566	0.217
MPI3D	3	-0.001	0.155	0.144	0.259	0.259	0.170	56.8	1.689	0.156

Fact.-Q denotes separate factor-specific orthogonal fits ("Block-Q" in Figure 6). The order-2 and order-3 transfer cosines $\tau_{1\to r}$ are 0.369/0.191 for dSprites, 0.090/0.003 for PUG-Animal, and 0.131/0.021 for MPI3D. Angles in degrees.

On incomplete training support, we use the same sequential estimator: missing condition cells are first set to zero and then projected; this is not a jointly constrained least-squares estimator. All reconstruction coefficients and centers are estimated from the training support. Effective rank and component energy are computed separately on the complete grid as descriptive measures of representation structure, not as held-out generalization evidence.

Context stability uses the mean unit difference direction over 80% of the contexts as a reference and reports the absolute cosine with the remaining 20%. It measures axis consistency and is insensitive to sign flips. A complete ANOVA decomposition can extend to K factors;

third order is the main-experiment truncation. On an independent dSprites split, image held-out R^2 at orders one through four is 0.655, 0.819, 0.868, and 0.883. The fourth-order increment is 0.016, smaller than the second- and third-order increments of 0.163 and 0.049, while the text fourth-order increment is close to zero. Fourth-order image responses are present, but their additional explanatory value and context stability (0.295) are limited.

B.3 Finite-difference responses

Let $A \subseteq [K]$ be a set of simultaneously changed factors, with $|A| = r$. For each factor, we pre-specify two values and use $b \in \{0,1\}^r$ to index the vertices of the resulting hypercube. Holding the remaining context c_{-A} fixed, the order-r mixed finite difference is

$$\Delta_{m,A|c_{-A}} = \sum_{b\in\{0,1\}^r} (-1)^{r-\|b\|_1} \bar{h}_m(c_A^b, c_{-A}). \tag{B.1}$$

The normalized direction is

$$d_{m,A|c_{-A}} = \frac{\Delta_{m,A|c_{-A}}}{\|\Delta_{m,A|c_{-A}}\|_2 + \varepsilon}. \tag{B.2}$$

First-order responses change one factor, while second- and third-order responses change two and three factors simultaneously. Experiment 2 uses pre-specified endpoint values for higher-order factors; Experiment 3 uses all available value pairs in the training support for first-order directions. All semantic correspondences are fixed before fitting.

Finite-difference responses differ from isolated ANOVA components. If the complete representation is $\bar{h}_m = \sum_B \phi_{m,B}$, then

$$\Delta_A \bar{h}_m = \sum_{B\supseteq A} \Delta_A \phi_{m,B}. \tag{B.3}$$

Thus, a second-order difference measures the total response to jointly changing two factors and may include contributions from third- and higher-order components. We call it a second-order contrast response; an isolated second-order ANOVA effect removes those contributions.

For each order, we stack directions with the same semantic indices as

$$D_m^{(r)} = [d_{m,1}^{(r)}, \ldots, d_{m,M_r}^{(r)}] \in \mathbb{R}^{d\times M_r}, \qquad m \in \{I, T\}. \tag{B.4}$$

Experiment 2 splits training and test directions by remaining contexts within each factor set and therefore tests context holdout; endpoint representations may overlap across factor sets. Experiment 3 uses the stricter tuple split and includes a direction in map estimation only when all of its endpoints belong to the training support.

B.4 Relative orthogonal intervention

For first-order responses, we solve the orthogonal Procrustes problem on unit directions:

$$Q_1^* = \arg\min_{Q^\top Q = I} \left\| D_I^{(1)} - Q D_T^{(1)} \right\|_F^2. \tag{B.5}$$

If

$$D_I^{(1)} D_T^{(1)\top} = U\Sigma V^\top, \tag{B.6}$$

the closed-form solution is

$$Q_1^* = UV^\top. \tag{B.7}$$

The intervention is applied to unnormalized encoder outputs; difference directions are unit-normalized before fitting. For downstream matching, we first estimate the two training-support centers

$$\mu_I = \frac{1}{N_{\text{tr}}} \sum_{s \in \Omega_{\text{tr}}} h_I(s), \qquad \mu_T = \frac{1}{N_{\text{tr}}} \sum_{s \in \Omega_{\text{tr}}} h_T(s), \tag{B.8}$$

and then use the centered cross-modal score

$$s_Q(x,t) = \cos\left(h_I(x) - \mu_I, Q\left[h_T(t) - \mu_T\right]\right). \tag{B.9}$$

ROCA uses a complete $d \times d$ orthogonal matrix, without an additional low-rank projection or adding the image mean to the text representation. The centered identity baseline sets $Q = I$ and isolates the effect of centering from the relative rotation.

B.5 Multi-order relative alignment

To test whether one transform can explain responses of multiple orders, we solve

$$Q^*_{\le R} = \arg\min_{Q^\top Q = I} \sum_{r=1}^{R} \frac{\lambda_r}{\nu_r} \left\| D_I^{(r)} - Q D_T^{(r)} \right\|_F^2, \tag{B.10}$$

where $\lambda_r \ge 0$ and $\sum_r \lambda_r = 1$. The normalization ν_r is the number of direction columns M_r, or the average squared Frobenius norm of the two matrices. Concatenating each matrix after multiplying it by $\sqrt{\lambda_r/\nu_r}$ reduces the problem to one Procrustes solve.

The equal-weight version fixes the order weights. For validation, we select λ on a three-dimensional simplex grid with step size 0.25. The selection metric is bidirectional local Hamming-1 accuracy on the inner validation split; ties are broken by validation margin and then by higher-order weight. After selecting the weights, we re-estimate the map on the outer training support. We compare ROCA-1, ROCA-1+2, and ROCA-1+2+3; the latter two are predefined controls for testing shared higher-order geometry, not an assumption that higher-order responses are reliable.

B.6 Baseline mappings

In addition to the shared orthogonal map, we include the following controls. The unconstrained linear upper bound is

$$A^* = \arg\min_{A} \sum_{r=1}^{3} \left\| D_I^{(r)} - A D_T^{(r)} \right\|_F^2, \tag{B.11}$$

and the ridge version adds $\lambda \|A\|_F^2$. Both use the concatenation of first- through third-order directions, making them capacity-matched to ROCA-1+2+3 as closely as possible.

Factor-specific orthogonal maps fit a separate complete orthogonal matrix for each factor or interaction set. This is a diagnostic upper bound for factor-dependent misalignment, not a block-diagonal transform in one common environment space. Random orthogonal maps, within-factor mismatches, and global semantic mismatches test whether the gains depend on correct concept correspondence.

Table 6: Mapping baselines and their roles.

Method	Geometric constraint	Role
Centered identity	$Q = I$	Isolate centering
ROCA-1	Single $Q^\top Q = I$	Main first-order intervention
ROCA-1+2 / 1+2+3	Shared multi-order Q	Test shared geometry
Factor-specific orthogonal	Independent full Q_A	Diagnostic upper bound
Ridge linear	L_2-regularized matrix	Constrained non-isometric control
General linear	Unconstrained matrix	Linear fitting upper bound
Random / shuffled	Random map or wrong pairing	Negative controls

B.7 Geometry and task metrics

For each held-out order, we report unit-direction cosine, squared Euclidean error per direction, and the orthogonal explainable fraction

$$\rho_r = 1 - \frac{E^{(r)}_{\text{orth,test}}}{E^{(r)}_{\text{id,test}} + \varepsilon}. \tag{B.12}$$

$\rho_r < 0$ means that the orthogonal map is worse than identity on held-out directions. Gram compatibility compares $D_m^{(r)\top} D_m^{(r)}$ for unit direction frames and therefore primarily reflects angular relations, not uniform scaling. Principal angles are computed using at most 256 test directions: each side retains the 95%-energy dimension and the smaller retained rank is used.

The downstream tables report local Hamming-1 Top-1, seen/unseen H-mean, binding-swap accuracy, and binding margin. For orthogonal maps, the cosine Gram drift of centered text residuals is zero up to numerical precision because $Q^\top Q = I$; this does not imply that the cosine Gram matrix of the original uncentered text representations is unchanged.

If the same orthogonal transformation is applied to both modalities, then

$$(Qz_I)^\top (Qz_T) = z_I^\top z_T. \tag{B.13}$$

This invariance check shows that ROCA changes the relative factor coordinates of image and text while preserving the absolute orientation of the shared space.

B.8 Unseen-composition evaluation and additional results

Table 2 reports global H-mean and local Hamming-1 Top-1 for the evaluated mappings. Test images are excluded from direction construction and center estimation. The correct tuple is compared with descriptions formed by changing one factor value, and the task is scored by the fraction of held-out queries for which the correct description ranks first. The reported results use the main text template. dSprites, PUG-Animal, and MPI3D use 18, 141, and 34 local candidates, with chance levels of 5.56%, 0.71%, and 2.94%.

Direction fitting and retrieval. Table 2 reports global H-mean and local accuracy for the same mapping families. ROCA-1 scores highest on both metrics across all three datasets. ROCA-1+2 and ROCA-1+2+3 yield H-means of 0.272 and 0.268 on dSprites, 0.222 and 0.217 on PUG-Animal, and 0.165 and 0.121 on MPI3D. Their local accuracies follow the same ordering. This comparison separates direction fitting from complete-candidate ranking and shows that adding response orders changes task transfer.

Order weights and negative controls. Both multi-order versions use one shared orthogonal matrix. The equal-weight baseline normalizes each order before concatenation; the nested-validation version selects order weights on the inner support and refits on the outer training support. Shuffled and random orthogonal maps test whether correct factor correspondence matters.

Residual geometry. Gram drift is measured relative to centered text residuals. For any such residual matrix R_T, $(QR_T)^\top(QR_T) = R_T^\top R_T$; unit normalization also preserves its pairwise cosines under Q. A general linear map need not preserve these relations. The original uncentered representation provides a different reference: subtracting the training mean can already change its cosine Gram matrix before rotation.

B.9 Synthetic Error Injection and Recovery

We inject first-order orthogonal rotations, order-specific rotations, non-isometric shears, composition-specific higher-order residuals, and rank-deficient training support into known factor representations. Maps are estimated on training directions and evaluated separately on held-out directions and complete held-out tuples. Order-specific rotations act on independent component subspaces, so a shared block-orthogonal transform may still exist globally; the experiment tests whether first-order support identifies its higher-order action. Pure rotations and shears use the same coordinate projection rather than separately fitted PCA projections.

Table 7: Representative held-out direction results under known error injections. Errors are MSE; numerical values near zero for pure rotations are reported as $< 10^{-18}$.

Injection	Order	Identity MSE	Orthogonal MSE	Linear MSE	Orthogonal cosine	Linear cosine
None	1	0	$< 10^{-18}$	$< 10^{-18}$	1.000	1.000
First-order rotation	1	1.403	$< 10^{-18}$	$< 10^{-18}$	1.000	1.000
Independent second-order rotation	2	1.135	$< 10^{-18}$	$< 10^{-18}$	1.000	1.000
Non-isometric shear	1	0.0575	0.0357	$< 10^{-18}$	0.987	1.000
Composition-specific higher-order residual	2	3.272	2.618	0.0773	0.613	0.945
Rank-deficient support	1	1.494	1.196	0.5784	0.183	0.384

The first-order rotation is recovered to numerical precision. When the second-order subspace uses another rotation, a separately fitted second-order map also recovers it exactly, but transfer of the first-order map to second order reaches only 0.247 ± 0.087 cosine. A general linear map reduces the first-order shear error to numerical precision, while the orthogonal map retains an MSE of 0.0357. Restricting the training directions to rank two reduces the first-order orthogonal test cosine to 0.183.

Direction fitting still does not imply composition recovery. Under the higher-order residual, the general linear map reaches 0.945 cosine on second-order directions but only 0.473 on held-out complete tuples, below the identity value of 0.758. True rotations and shears are recovered at the tuple level when support is sufficient.

Table 8: Paired cosine on held-out complete tuples.

Injection	Identity	Orthogonal	Linear
None	1.000	1.000	1.000
First-order rotation	0.287	1.000	1.000
Independent first- and second-order rotations	0.065	1.000	1.000
Non-isometric shear	0.975	0.988	1.000
Composition-specific higher-order residual	0.758	0.588	0.473
Rank-deficient support	−0.015	0.148	0.380

These synthetic tests distinguish identifiable rotations, non-isometric corrections, composition-specific residuals, and incomplete training support under known ground truth. Evaluating both held-out directions and complete tuples separates direction recovery from composition transfer.

B.10 Binding Swaps in Pretrained Models

We construct a two-candidate CLEVR-2D task that changes only the color and shape ownership of the left and right objects, using frozen OpenAI CLIP ViT-B/32 representations. The 576 scene tuples use a 20% holdout. Only test views in which both color and shape

differ across the two objects enter the hard binding subset. Each query compares the correct description with a factor-swapped description, giving a 50% chance level.

ROCA-1 obtains 0.760±0.050 attribute-swap accuracy and a 0.0220 binding margin. ROCA-1+2 and ROCA-1+2+3 both obtain 0.746 ± 0.050, with margins 0.0220 and 0.0210. Raw matching scores 0.539 with a 0.0006 margin, and centered identity scores 0.549 with a 0.0027 margin. The linear controls produce negative margins, while shuffled and random orthogonal controls remain near chance. Frozen representations contain interactions that can enter conditional first-order differences, allowing relative alignment to improve the use of existing binding evidence.

B.11 Binding Supervision in Controlled Dual Encoders

To separate concept recognition from object ownership, we train small symbolic dual Transformer encoders. Each side has two layers and four attention heads, with hidden width and output dimension 64; together the encoders have 214,017 parameters. The concept space has $C = 3$ factors with $V = 8$ values each. All 512 single objects occur in training, while 2,048 object pairs are split into 1,229 training, 410 validation, and 409 test pairs. The two modalities receive separate encodings of the same symbolic scene. At test time, the correct scene is compared with a negative that swaps only the first factor. We evaluate five conditions spanning ordinary scene matching, slot markers, factor-swap negatives, and their combinations, using three random seeds per condition, batch size 128, learning rate 3×10^{-4}, and 2,000 training steps.

Table 9: Final results for controlled binding training. Swap accuracy is the mean and standard deviation over three seeds, with chance 0.5; other columns report seed means.

Training condition	Binding-swap accuracy	Binding margin	Scene retrieval Top-1	Whole-object permutation cosine
Standard scene matching	0.777 ± 0.144	0.0009	0.995	0.999
Slot markers, no swap supervision	0.783 ± 0.136	0.0010	0.995	0.999
Swap supervision, weight 0.5	1.000 ± 0.000	0.6620	0.995	0.996
Slot markers and swap supervision	1.000 ± 0.000	0.8448	0.995	0.998
Swap supervision	1.000 ± 0.000	0.8554	0.995	0.998

Ordinary scene-retrieval Top-1 is approximately 0.995 in all conditions, but standard matching and slot markers alone yield binding margins of only 0.0009 and 0.0010. Their swap accuracies exceed chance but remain sensitive to very small score perturbations. All three conditions with swap negatives attain 1.000 swap accuracy and positive margins from 0.662 to 0.855.

This control tests whether a specific signal is sufficient for stable binding; it does not claim that explicit swap negatives are the only way to learn binding. Relational descriptions, object localization, and implicit hard negatives in natural image-text data may also provide supervision that breaks swap symmetry.

C Per-Model Results

This section disaggregates the eight-model macro-averages in Tables 1 and 5. Unless stated otherwise, each model entry is first averaged over the three predefined repeats or outer splits used by the corresponding experiment; the reported macro-average then gives each model equal weight. The mean rows below reproduce those geometric aggregate tables up to rounding. "OpenAI" denotes the official CLIP checkpoints, and "OpenCLIP" denotes the LAION-2B checkpoints. Local retrieval is summarized at the task level in Table 2.

C.1 dSprites

Table 10: Per-model held-out explained variance on dSprites.

	Image					Text				
Model	R_1^2	R_2^2	R_3^2	Δ_2	Δ_3	R_1^2	R_2^2	R_3^2	Δ_2	Δ_3
OpenAI B/16	0.691	0.847	0.892	0.155	0.045	0.942	0.995	0.997	0.053	0.002
OpenAI B/32	0.646	0.833	0.880	0.187	0.046	0.922	0.995	0.997	0.073	0.002
OpenAI L/14	0.681	0.837	0.887	0.156	0.050	0.923	0.993	0.997	0.069	0.004
OpenCLIP B/16	0.647	0.798	0.848	0.151	0.051	0.967	0.997	0.998	0.029	0.001
OpenCLIP B/32	0.647	0.802	0.853	0.155	0.051	0.965	0.997	0.998	0.032	0.001
OpenCLIP L/14	0.549	0.743	0.811	0.194	0.067	0.971	0.997	0.998	0.026	0.001
SigLIP B/16	0.669	0.823	0.867	0.155	0.044	0.954	0.990	0.993	0.036	0.004
SigLIP2 B/16	0.728	0.855	0.888	0.127	0.034	0.927	0.980	0.994	0.053	0.014
Mean	**0.657**	**0.817**	**0.866**	**0.160**	**0.048**	**0.946**	**0.993**	**0.996**	**0.046**	**0.004**

The cosine tables compare identity, global orthogonal (Q), factor-specific orthogonal (Factor-Q), general linear, and ridge maps. In the diagnostic tables, Gram is the cosine between the two direction Gram matrices, Ang. is the mean principal angle in degrees, MSE and ρ_r evaluate the global orthogonal map, and $\tau_{1\to r}$ transfers the first-order map to order r.

Table 11: Per-model held-out direction cosine on dSprites.

Model	r	Identity	Q	Factor-Q	Linear	Ridge
OpenAI B/16	1	0.107	0.688	0.728	0.797	0.823
	2	0.000	0.556	0.645	0.778	0.779
	3	-0.004	0.487	0.542	0.747	0.773
OpenAI B/32	1	0.104	0.607	0.642	0.695	0.720
	2	-0.005	0.436	0.516	0.674	0.674
	3	0.007	0.366	0.405	0.664	0.688
OpenAI L/14	1	0.122	0.703	0.738	0.732	0.804
	2	0.003	0.574	0.653	0.794	0.786
	3	-0.001	0.535	0.550	0.524	0.787
OpenCLIP B/16	1	0.157	0.681	0.711	0.762	0.794
	2	0.003	0.519	0.596	0.732	0.730
	3	-0.002	0.444	0.487	0.740	0.753
OpenCLIP B/32	1	0.147	0.629	0.651	0.714	0.743
	2	-0.006	0.441	0.502	0.673	0.674
	3	-0.002	0.398	0.437	0.693	0.712
OpenCLIP L/14	1	0.093	0.580	0.605	0.656	0.735
	2	-0.003	0.482	0.546	0.739	0.713
	3	0.000	0.461	0.483	0.533	0.770
SigLIP B/16	1	0.145	0.674	0.700	0.700	0.770
	2	0.003	0.509	0.575	0.746	0.740
	3	0.004	0.468	0.473	0.535	0.786
SigLIP2 B/16	1	0.178	0.663	0.695	0.704	0.785
	2	-0.001	0.426	0.501	0.730	0.720
	3	0.001	0.358	0.386	0.529	0.730
Mean	1	0.132	0.653	0.684	0.720	0.772
	2	-0.001	0.493	0.567	0.733	0.727
	3	0.001	0.440	0.470	0.621	0.750

Table 12: Per-model held-out cross-modal diagnostics on dSprites.

Model	r	Gram	Ang.	MSE	ρ_r	$\tau_{1\to r}$
OpenAI B/16	1	0.659	36.4	0.624	0.651	–
	2	0.419	42.4	0.888	0.556	0.405
	3	0.347	45.7	1.026	0.489	0.187
OpenAI B/32	1	0.592	38.2	0.786	0.561	–
	2	0.287	46.2	1.127	0.439	0.325
	3	0.183	49.5	1.268	0.362	0.157
OpenAI L/14	1	0.685	36.2	0.595	0.661	–
	2	0.465	43.4	0.851	0.573	0.457
	3	0.440	45.9	0.929	0.536	0.262
OpenCLIP B/16	1	0.726	34.3	0.639	0.621	–
	2	0.427	43.8	0.962	0.518	0.394
	3	0.343	49.2	1.113	0.445	0.189
OpenCLIP B/32	1	0.679	39.0	0.741	0.566	–
	2	0.343	46.9	1.118	0.444	0.336
	3	0.258	49.2	1.204	0.399	0.171
OpenCLIP L/14	1	0.561	40.7	0.840	0.537	–
	2	0.393	47.8	1.035	0.484	0.370
	3	0.370	50.2	1.078	0.461	0.212
SigLIP B/16	1	0.721	40.8	0.651	0.619	–
	2	0.449	48.5	0.982	0.507	0.375
	3	0.368	50.4	1.064	0.466	0.212
SigLIP2 B/16	1	0.737	43.4	0.674	0.590	–
	2	0.315	49.2	1.148	0.426	0.288
	3	0.219	51.9	1.283	0.357	0.139
Mean	1	0.670	38.6	0.694	0.601	–
	2	0.387	46.0	1.014	0.494	0.369
	3	0.316	49.0	1.121	0.439	0.191

C.2 PUG-Animal

Table 13: Per-model held-out explained variance on PUG-Animal.

	Image					Text				
Model	R_1^2	R_2^2	R_3^2	Δ_2	Δ_3	R_1^2	R_2^2	R_3^2	Δ_2	Δ_3
OpenAI B/16	0.698	0.868	0.897	0.170	0.029	0.951	0.994	0.998	0.043	0.004
OpenAI B/32	0.694	0.871	0.900	0.176	0.029	0.939	0.992	0.998	0.053	0.006
OpenAI L/14	0.681	0.851	0.886	0.171	0.034	0.937	0.991	0.997	0.054	0.007
OpenCLIP B/16	0.702	0.870	0.899	0.168	0.029	0.950	0.995	0.999	0.045	0.004
OpenCLIP B/32	0.679	0.860	0.891	0.181	0.032	0.948	0.995	0.999	0.048	0.004
OpenCLIP L/14	0.703	0.873	0.904	0.170	0.031	0.950	0.995	0.999	0.044	0.004
SigLIP B/16	0.659	0.834	0.869	0.175	0.035	0.402	0.882	0.982	0.479	0.100
SigLIP2 B/16	0.685	0.847	0.879	0.162	0.032	0.318	0.688	0.866	0.370	0.178
Mean	**0.688**	**0.859**	**0.891**	**0.172**	**0.031**	**0.799**	**0.942**	**0.980**	**0.142**	**0.038**

Table 14: Per-model held-out direction cosine on PUG-Animal.

Model	r	Identity	Q	Factor-Q	Linear	Ridge
OpenAI B/16	1	0.102	0.381	0.396	0.331	0.335
	2	0.000	0.261	0.294	0.268	0.269
	3	-0.005	0.208	0.236	0.260	0.260
OpenAI B/32	1	0.074	0.340	0.355	0.307	0.310
	2	0.008	0.231	0.264	0.249	0.249
	3	0.009	0.192	0.222	0.257	0.257
OpenAI L/14	1	0.105	0.366	0.385	0.252	0.295
	2	0.000	0.242	0.280	0.240	0.249
	3	0.001	0.214	0.225	0.273	0.280
OpenCLIP B/16	1	0.098	0.377	0.393	0.324	0.334
	2	0.003	0.238	0.272	0.247	0.248
	3	0.000	0.203	0.221	0.247	0.248
OpenCLIP B/32	1	0.072	0.341	0.356	0.288	0.291
	2	0.000	0.229	0.264	0.240	0.240
	3	-0.003	0.187	0.210	0.227	0.227
OpenCLIP L/14	1	0.092	0.369	0.385	0.239	0.313
	2	0.006	0.247	0.283	0.234	0.250
	3	-0.009	0.206	0.229	0.249	0.258
SigLIP B/16	1	-0.004	0.217	0.236	0.199	0.214
	2	0.000	0.151	0.185	0.213	0.216
	3	-0.003	0.160	0.179	0.262	0.264
SigLIP2 B/16	1	0.021	0.193	0.203	0.177	0.205
	2	0.000	0.108	0.130	0.189	0.192
	3	0.001	0.133	0.128	0.204	0.207
Mean	1	0.070	0.323	0.339	0.265	0.287
	2	0.002	0.214	0.247	0.235	0.239
	3	-0.001	0.188	0.206	0.247	0.250

Table 15: Per-model held-out cross-modal diagnostics on PUG-Animal.

Model	r	Gram	Ang.	MSE	ρ_r	$\tau_{1\to r}$
OpenAI B/16	1	0.529	60.3	1.239	0.310	–
	2	0.376	57.1	1.477	0.262	0.098
	3	0.259	55.9	1.584	0.212	0.003
OpenAI B/32	1	0.478	61.0	1.320	0.287	–
	2	0.318	57.6	1.537	0.225	0.108
	3	0.223	56.3	1.616	0.184	-0.010
OpenAI L/14	1	0.550	63.3	1.268	0.292	–
	2	0.360	59.7	1.515	0.242	0.124
	3	0.297	58.8	1.572	0.213	0.004
OpenCLIP B/16	1	0.508	62.5	1.245	0.310	–
	2	0.310	57.7	1.524	0.236	0.106
	3	0.221	58.2	1.594	0.203	0.007
OpenCLIP B/32	1	0.486	62.2	1.317	0.290	–
	2	0.324	57.7	1.541	0.230	0.101
	3	0.240	57.6	1.627	0.189	-0.001
OpenCLIP L/14	1	0.528	62.9	1.262	0.305	–
	2	0.344	59.2	1.506	0.242	0.094
	3	0.254	59.0	1.588	0.213	0.004
SigLIP B/16	1	0.410	58.6	1.565	0.221	–
	2	0.365	58.2	1.699	0.151	0.046
	3	0.335	56.7	1.680	0.162	0.007
SigLIP2 B/16	1	0.433	57.5	1.615	0.176	–
	2	0.320	56.8	1.784	0.108	0.039
	3	0.300	56.3	1.735	0.132	0.006
Mean	1	0.490	61.1	1.354	0.274	–
	2	0.340	58.0	1.573	0.212	0.090
	3	0.266	57.4	1.624	0.189	0.003

C.3 MPI3D

Table 16: Per-model held-out explained variance on MPI3D.

	Image					Text				
Model	R_1^2	R_2^2	R_3^2	Δ_2	Δ_3	R_1^2	R_2^2	R_3^2	Δ_2	Δ_3
OpenAI B/16	0.520	0.715	0.823	0.195	0.108	0.972	0.999	1.000	0.027	0.001
OpenAI B/32	0.524	0.712	0.815	0.188	0.103	0.983	0.999	1.000	0.016	0.000
OpenAI L/14	0.486	0.684	0.797	0.198	0.113	0.967	0.999	1.000	0.031	0.001
OpenCLIP B/16	0.575	0.756	0.853	0.181	0.097	0.979	0.999	1.000	0.021	0.001
OpenCLIP B/32	0.572	0.746	0.846	0.174	0.100	0.984	1.000	1.000	0.016	0.000
OpenCLIP L/14	0.566	0.733	0.824	0.168	0.090	0.968	0.999	1.000	0.031	0.001
SigLIP B/16	0.551	0.721	0.818	0.170	0.096	0.757	0.895	0.969	0.138	0.074
SigLIP2 B/16	0.537	0.716	0.818	0.180	0.102	0.862	0.994	1.000	0.133	0.005
Mean	**0.541**	**0.723**	**0.824**	**0.182**	**0.101**	**0.934**	**0.985**	**0.996**	**0.052**	**0.010**

Table 17: Per-model held-out direction cosine on MPI3D.

Model	r	Identity	Q	Factor-Q	Linear	Ridge
OpenAI B/16	1	-0.004	0.466	0.525	0.504	0.539
	2	0.001	0.237	0.299	0.280	0.280
	3	0.001	0.152	0.138	0.249	0.249
OpenAI B/32	1	0.004	0.450	0.502	0.467	0.488
	2	0.003	0.214	0.264	0.253	0.253
	3	-0.001	0.136	0.098	0.237	0.237
OpenAI L/14	1	0.003	0.444	0.490	0.406	0.507
	2	0.003	0.244	0.302	0.309	0.310
	3	0.000	0.202	0.212	0.310	0.310
OpenCLIP B/16	1	-0.001	0.491	0.544	0.530	0.568
	2	-0.003	0.242	0.301	0.275	0.275
	3	-0.001	0.157	0.136	0.244	0.244
OpenCLIP B/32	1	-0.002	0.490	0.536	0.521	0.556
	2	-0.004	0.241	0.283	0.265	0.266
	3	0.000	0.153	0.109	0.234	0.234
OpenCLIP L/14	1	0.001	0.461	0.501	0.408	0.519
	2	0.005	0.219	0.278	0.271	0.273
	3	-0.001	0.168	0.182	0.282	0.282
SigLIP B/16	1	0.001	0.353	0.398	0.385	0.432
	2	-0.001	0.144	0.187	0.253	0.255
	3	-0.001	0.123	0.139	0.258	0.258
SigLIP2 B/16	1	0.002	0.463	0.505	0.422	0.494
	2	0.000	0.197	0.253	0.253	0.254
	3	-0.001	0.152	0.138	0.258	0.258
Mean	1	0.000	0.452	0.500	0.455	0.513
	2	0.000	0.217	0.271	0.270	0.271
	3	-0.001	0.155	0.144	0.259	0.259

Table 18: Per-model held-out cross-modal diagnostics on MPI3D.

Model	r	Gram	Ang.	MSE	ρ_r	$\tau_{1\to r}$
OpenAI B/16	1	0.467	33.6	1.067	0.468	–
	2	0.170	52.3	1.525	0.237	0.143
	3	0.147	54.9	1.695	0.152	0.015
OpenAI B/32	1	0.432	33.0	1.101	0.448	–
	2	0.166	55.2	1.571	0.212	0.135
	3	0.162	55.6	1.727	0.137	0.020
OpenAI L/14	1	0.457	40.1	1.111	0.443	–
	2	0.195	54.3	1.512	0.242	0.155
	3	0.196	58.3	1.596	0.202	0.029
OpenCLIP B/16	1	0.541	30.5	1.019	0.491	–
	2	0.198	52.5	1.516	0.244	0.146
	3	0.189	54.7	1.687	0.158	0.022
OpenCLIP B/32	1	0.535	38.2	1.021	0.491	–
	2	0.210	54.4	1.518	0.244	0.167
	3	0.214	55.0	1.695	0.152	0.022
OpenCLIP L/14	1	0.513	39.5	1.078	0.461	–
	2	0.174	55.6	1.563	0.215	0.122
	3	0.153	58.8	1.665	0.168	0.026
SigLIP B/16	1	0.366	52.5	1.293	0.353	–
	2	0.137	57.7	1.712	0.145	0.080
	3	0.123	58.3	1.753	0.125	0.024
SigLIP2 B/16	1	0.510	42.7	1.074	0.462	–
	2	0.171	57.2	1.607	0.197	0.098
	3	0.175	58.5	1.695	0.153	0.014
Mean	1	0.478	38.8	1.095	0.452	–
	2	0.178	54.9	1.566	0.217	0.131
	3	0.170	56.8	1.689	0.156	0.021