How Far Do Simple Transformations Translate Across Heterogeneous Text and Image Embedding Models?

Anonymous authors
Paper under double-blind review

Abstract

We investigate whether simple transformations can translate representations across heterogeneous text and image embedding models. Understanding how independently trained models organize semantic information is an enabler for AI-to-AI latent communication without decoding into human-readable text, and for reusing components across models. A recurring hypothesis in the literature is that independently trained models converge to a shared geometry over semantic content, so that a simple, low-supervision transformation should relate their internal representations; we test this hypothesis of latent universality in a realistic setting beyond simplified benchmarks. We study 14 text embedding models and 4 image/multimodal models that differ in architecture, pooling strategy, training objective, and dimension, and evaluate compatibility with linear CKA, downstream transfer, representation fidelity, and k-NN retrieval, on full evaluation sets and three anchor protocols. Simple translators recover meaningful shared structure and support transfer for some compatible pairs, but fail sharply for others; same-design models cluster tightly while cross-design pairs degrade in structured ways. Compatibility depends jointly on architecture, training objective, readout/normalization, and data distribution. Overall, the results show that heterogeneous embedding spaces are not universally related by simple mappings as often suggested in the literature.

1 Introduction

Modern learning systems often train an expensive downstream component against one encoder's representations. A classifier, retrieval index, frozen prediction head, or generative conditioning stack then expects that encoder's dimension and geometry. Deployment, however, offers many heterogeneous encoders with different accuracy, latency, modality, language, and licensing trade-offs. These encoders are not interchangeable even when they describe the same input. A translator from model A to model B would let a component trained for B consume representations produced by A, avoiding retraining the component whenever the encoder changes.

The same interface has broader uses. A text-conditioned image generator is normally coupled to the text representation used during its training; aligning another text or multimodal encoder to that conditioning space could extend the frozen generator without retraining it, as demonstrated by adapter-based systems such as GlueGen (Qin et al., 2023). In multi-model LLM serving, translating key–value (KV) caches could let one model reuse context processed by another instead of repeating prefill (Fu et al., 2026). In multi-agent systems, agents could exchange continuous hidden states or embeddings rather than repeatedly decoding and re-encoding text (Zou et al., 2025; Du et al., 2026; Yang & Eshraghian, 2025). Text remains a portable and interpretable interface, but token-by-token serialization adds latency and can discard uncertainty or task-relevant distinctions not expressed by the selected token sequence. A latent channel can preserve such continuous information only if the receiver can interpret the sender's representation.

This requirement exposes our scientific question: *how universal is representation geometry across independently trained models?* Prior work develops model stitching, direct alignment, relative representations (RR), which encode each vector by its cosine similarities to shared anchors, and

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.

inverse relative projection (IRP), which maps those relative coordinates back into a target space (Lenc & Vedaldi, 2015; Bansal et al., 2021; Lähner & Moeller, 2024; Moschella et al., 2023; Maiorca et al., 2024), while the Platonic Representation Hypothesis proposes broader convergence as models scale (Huh et al., 2024). We test an operational consequence across 14 text and four image/multimodal encoders: whether lightweight translators preserve global geometry, item identity, pointwise fidelity, and downstream decisions across every ordered pair. Some same-design pairs transfer well, showing that reusable latent interfaces are plausible and worth pursuing. The sharp, structured failures across other pairs show that the broader goals above are difficult, especially because generator conditioning and KV caches are structured sequences rather than the single vectors tested here. Compatibility depends on the pair, translator, anchor protocol, direction, and task; sentence-anchor IRP nearly matches the high-supervision text map, but no tested method provides a universal interface.

Contributions.

- A systematic empirical study of simple translation approaches across a heterogeneous suite of 14 text and four image/multimodal embedding models, covering every ordered pair, full evaluation splits, and three anchor protocols.
- A multi-diagnostic evaluation combining geometric similarity (CKA), k-NN retrieval, pointwise fidelity, and downstream transfer, rather than a single metric or a single benchmark.
- Evidence that simple transformations are useful but strongly pair-dependent, with translator, anchor, direction, and task changing the ranking and with same-family size variants forming the tightest clusters.

2 Related Work

Prior work has used linear maps to compare independently trained image representations (Lenc & Vedaldi, 2015) and word embeddings (Lample et al., 2018). Model stitching tests whether a representation from one network can replace a layer in another (Bansal et al., 2021); more recent work studies direct latent alignment and transfer across language models (Lähner & Moeller, 2024; Chen et al., 2025; Gorbett & Jana, 2026). Unconstrained linear or affine maps often outperform isometric maps, but reported success depends on models, layers, data, and evaluation. Relative representations (RR) instead encode a vector by its cosine similarities to shared semantic anchors, giving coordinates invariant to rotations and reflections under corresponding anchors (Moschella et al., 2023). Inverse relative projection (IRP) maps those coordinates back into a target representation space (Maiorca et al., 2024); direct least-squares translation (Maiorca et al., 2023), products of invariances (Cannistraci et al., 2024), and learned or whitened anchors (Svendsen et al., 2026) extend the same view. We compare these approaches under matched anchor counts and expose their sensitivity to the anchor protocol. Our focus is not a new optimizer, but the behavior of the same simple map family across a deliberately heterogeneous suite.

CKA measures agreement between representation geometries (Kornblith et al., 2019), while similarity-matching analyses emphasize that no single invariance class fits every comparison (Csiszárik et al., 2021). Recent analyses further show that representational similarity can be confounded by scale or fail to predict behavior (Gröger et al., 2026; Armstrong et al., 2026). Representation translation also supports modular systems: GlueGen aligns alternative text and multimodal encoders to an existing image generator (Qin et al., 2023); Cache-to-Cache projects and fuses KV states between heterogeneous LLMs (Fu et al., 2026); and latent-agent methods exchange hidden representations without an intervening textual message (Zou et al., 2025; Du et al., 2026; Yang & Eshraghian, 2025; Zhu et al., 2025). These applications require more structure than our pooled-vector setting, but they share its central prerequisite: a representation produced by one frozen model must become readable to another frozen component. We test how often simple maps satisfy that prerequisite and whether geometric agreement forecasts functional reuse.

3 Problem Formulation and Translators

Let $f_A : \mathcal{Z} \to \mathbb{R}^{d_A}$ and $f_B : \mathcal{Z} \to \mathbb{R}^{d_B}$ be frozen encoders. For paired inputs z, write $x = f_A(z)$ and $y = f_B(z)$. A translator $T_{A\to B}$ should make $\hat{y} = T_{A\to B}(x)$ usable wherever y is expected. This definition is directional: $A \to B$ and $B \to A$ are fitted and evaluated separately.

3.1 LINEAR ESTIMATORS

A linear translator is a natural starting point because embeddings often encode semantic relations as approximately linear directions. The canonical *king–queen–man–woman* parallelogram (Mikolov et al., 2013a;b), bilingual lexicon induction (Lample et al., 2018), and the analysis of Allen & Hospedales (2019) all use the premise that meaning can compose locally as linear combinations of reference points. Appendix A tests this structure across heterogeneous encoders.

Let $X \in \mathbb{R}^{d_A \times m}$ and $Y \in \mathbb{R}^{d_B \times m}$ collect the embeddings of m shared anchors in models A and B. Reconstructing a source vector $x \in \mathbb{R}^{d_A}$ as a least-squares combination of the source anchors yields

$$c^*(x) = \arg\min_{c \in \mathbb{R}^m} \|x - Xc\|_2^2 = X^+ x, \tag{1}$$

which is the multi-anchor generalization of the analogy coefficients above: *king* $\approx \alpha$ *queen* $+ \beta$ *man* $+$ γ *woman* is the same equation written with three anchors instead of m. Transporting the coefficients to the target side, under the assumption that the local linear geometry is preserved, gives

$$\hat{y} = Yc^*(x) = YX^+x = T_{A\to B}x, \qquad T_{A\to B} = YX^+. \tag{2}$$

This is exactly the closed-form Linear (pinv) estimator used for direct anchor-to-anchor alignment (Maiorca et al., 2023). Coefficient-based translation through shared anchors and direct pseudo-inverse alignment are therefore not two different algorithms but two readings of the same operation: one expresses the source point in an anchor basis, while the other applies that same expression on the target side. The experiments implement the direct anchor-to-anchor solve. Appendix A develops the analogy connection and fits the coefficients in modern encoders.

Linear (SGD) optimizes the same squared-error map family on a much larger paired alignment corpus using mini-batches. It is therefore a *high-supervision linear reference*, not a mathematical upper bound: its advantage comes from anchor-independent fitting and greater supervision, and it more closely approximates the best performance attainable under the tested global linear objective.

3.2 RELATIVE TRANSLATORS

Let $\bar{x} = x/\|x\|_2$ and let $\bar{a}_i^A$ be the normalized source anchor embeddings. RR computes

$$r_A(x) = \left[\langle \bar{x}, \bar{a}_1^A \rangle, \ldots, \langle \bar{x}, \bar{a}_m^A \rangle\right]^\top \in \mathbb{R}^m. \tag{3}$$

A target head is trained in B's relative space and receives $r_A(x)$ at test time. IRP instead treats the relative vector as constraints and reconstructs a vector in the native target space using the inverse relative-projection procedure of Maiorca et al. (2024). Viewed together, the three low-supervision methods differ mainly in how they encode relationships to shared anchors: Linear (pinv) uses ordinary least-squares reconstruction coefficients, RR uses cosine affinities, and IRP reconstructs a native target vector from those affinities. All three test whether an input's relationship to corresponding semantic anchors survives the crossing between models.

3.3 EVALUATION CRITERIA

For n paired evaluation items, let X and Y contain the native source and target vectors as columns, and let $\hat{Y}$ contain the translated source vectors, after centering when required. We report four criteria:

1. **Linear CKA** (Centered Kernel Alignment with a linear kernel (Kornblith et al., 2019)): compares the two native representation matrices of models A and B on the same items, X and Y, with no fitted map involved. After centering the item representations, form the Gram matrices $K = X^\top X$ and $L = Y^\top Y$ and define $\mathrm{CKA}(X, Y) = \langle K, L \rangle_F / (\|K\|_F \|L\|_F)$. This normalized Frobenius similarity is 1 when the two models relate the shared items identically up to rotation and isotropic scaling, and lower when their pairwise relations disagree. The resulting matrix is symmetric and has 1 on the diagonal.
2. **Identity retrieval**: top-k accuracy when each $\hat{y}_i$ retrieves its paired y_i among all target items by cosine similarity.
3. **Pointwise fidelity**: mean row-wise cosine and mean normalized error $\|y_i - \hat{y}_i\|_2 / \|y_i\|_2$. Appendix E reports the detailed pair-wise fidelity matrices.

4. **Functional transfer**: accuracy and macro-F1 of a decision rule trained on the native target embeddings Y and evaluated on the translated source embeddings $\hat{Y} = T_{A\to B}X$.

For RR, these criteria compare source- and target-relative profiles, not a relative vector with a raw target vector. Their absolute values are therefore meaningful within RR but not directly comparable in scale to raw-space errors.

For summaries, $s_{A\to B}$ denotes transfer accuracy averaged over tasks. We define symmetric pair compatibility

$$\mathrm{Compat}(A, B) = \tfrac{1}{2}(s_{A\to B} + s_{B\to A}), \tag{4}$$

and use distance $1 - \mathrm{Compat}$ for average-linkage visualization. We retain directional matrices for analysis; symmetrization is only a summary.

4 Experimental Setup

Models and selection logic. We evaluate directional stitching between fourteen text models and four image/multimodal models. The text suite includes Qwen3-Embedding-4B/8B, NV-Embed-v2, stella_en_1.5B_v5, jina-embeddings-v3, GTR-T5-XXL, INSTRUCTOR-XL, e5-large-v2, mxbai-embed-large-v1, all-roberta-large-v1, bge-m3, gtr-t5-large, all-mpnet-base-v2, and nomic-embed-text-v1.5. The image suite contains Qwen3-VL-Embedding-2B/8B, siglip-so400m-patch14-384, and clip-vit-large-patch14. They span BERT-like, T5-style, causal-decoder, and multimodal architectures with dimensions from 768 to 4096 (Appendix Tables 5 and 6), including controlled same-family comparisons among Qwen3-Embedding, T5-family, and Qwen3-VL models.

Text transfer covers SST-2, DBpedia-14, Emotions, AG News, and IMDB. Image transfer covers CIFAR-10 and CIFAR-100 zero-shot classification, CIFAR-10 and CIFAR-100 frozen-embedding classification, and ImageNet-1k zero-shot classification. Every ordered source-to-target pair, including self-pairs, is evaluated on the complete evaluation split.

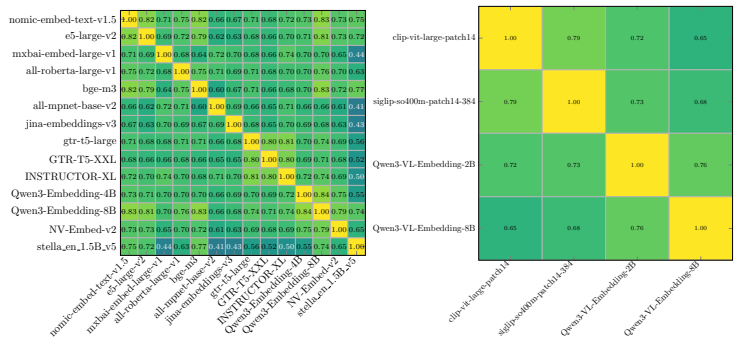
Anchor protocols. Text uses two independent 850-example anchor sets: Basic English words and WikiText sentences. The sentences come from the wikitext-2-raw-v1 training split after whitespace normalization and sentence splitting, retaining sentences with at least three words. Image uses 850 ImageNet training images. Anchor ordering and preprocessing are fixed across models, and the text protocols are reported separately.

Translator fitting. We evaluate four translators. **Linear (pinv)** fits the Moore–Penrose pseudo-inverse solution on paired anchor embeddings. **Relative representations (RR)** map an embedding to its vector of cosine similarities with the anchor set; target downstream models are trained in target-relative space and source examples are evaluated in source-relative space. **Inverse relative projection (IRP)** uses the anchor-based inverse projection. **Linear (SGD)** fits one directional linear map per source-to-target pair from paired alignment embeddings.

Text alignment uses 36,718 processed examples from the wikitext-2-raw-v1 training split; image alignment uses a 50,000-image ImageNet subset. Linear (SGD) uses Adam with learning rate 10^{-3}, batch size 64, three epochs, mean-squared error, and at most 2^{13} examples per map.

Downstream protocol. For frozen-embedding classification, a two-layer MLP with hidden sizes $[256, 128]$ is trained on native target embeddings Y and evaluated on translated source embeddings $\hat{Y} = T_{A\to B}X$. Training uses early stopping for at most 200 iterations with validation fraction 0.1. RR uses an MLP in target-relative space. Tables report mean accuracy over all ordered pairs, diagonal accuracy, and mean macro-F1.

Image zero-shot classification is performed as image-to-target-text evaluation. An image is encoded by source model A, transformed into target model B's representation protocol, and compared by cosine similarity with class names encoded by B. The highest-scoring class is compared with the ground-truth label. For RR, both images and class names are converted to their respective anchor-affinity profiles before comparison.



(a) Text (14 models). (b) Image (4 models).

Figure 1: Linear CKA between native representations on the 5,000-item alignment sample. Rows are sources and columns are targets; higher values indicate more similar item relations.

Alignment diagnostics. Alignment diagnostics use paired embeddings from the same examples: 5,000 unique, non-empty preprocessed WikiText examples for text and a fixed paired ImageNet sample for image. We report cosine self-retrieval at top-1, top-5, and top-10, native Linear CKA, normalized error $\|y_{\text{target}} - y_{\text{stitched}}\|_2 / \|y_{\text{target}}\|_2$, and mean cosine similarity. Native CKA compares source and target embeddings without stitching; for RR, the metrics compare source- and target-relative profiles. Appendix E gives the full matrices.

5 Results

5.1 Geometry, identity, and fidelity disagree

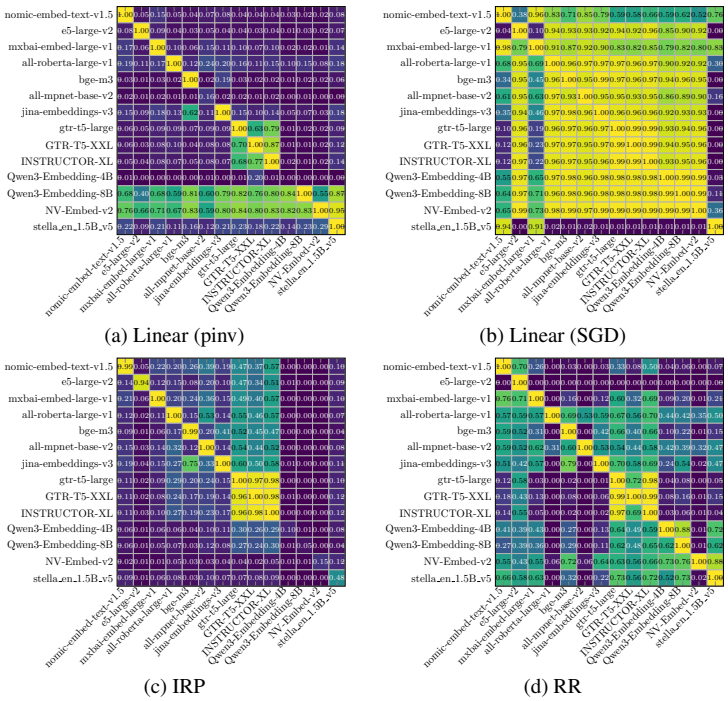
Figure 1 reports native Linear CKA for every ordered model pair. Text agreement is moderate and variable (off-diagonal mean 0.691, range 0.41–0.84), whereas image agreement is higher and more uniform (mean 0.722, range 0.65–0.79). The strongest values are family-local: the Qwen3-Embedding sizes (0.84), the T5 triple (0.80–0.81), CLIP–SigLIP (0.79), and the Qwen3-VL sizes (0.76). The weakest text pair is all-mpnet/stella (0.41), and the lowest image pair is CLIP/Qwen3-VL (0.65). Native CKA therefore reveals family-structured geometry, not interchangeability.

Figures 2 and 3 report top-1 paired-item retrieval with source rows, target columns, and one panel per translator. Linear (SGD) reaches 0.756 on cross-model text pairs, compared with 0.176 for IRP, 0.194 for pinv, and 0.329 for RR. On image, the corresponding values are 0.878, 0.617, 0.610, and 0.342. High-retrieval blocks remain family-local: Linear (SGD) reaches 0.99 for the Qwen3-Embedding sizes and T5 triple, while CLIP–SigLIP and Qwen3-VL reach 0.93–1.00 on image.

Native geometry and translation also diverge by pair. The Qwen3-Embedding sizes have the highest text CKA (0.84), yet word-anchor pinv retrieves only 0.00 from 4B to 8B and IRP reaches 0.01 in both directions. The slightly lower-CKA T5 triple translates almost perfectly under IRP (0.96–0.98). On image, CLIP–SigLIP has the highest CKA (0.79), but lower pinv retrieval (0.50–0.52) than SigLIP–Qwen3-VL (0.58–0.72). Thus native agreement does not certify translation, and the gap depends on the translator.

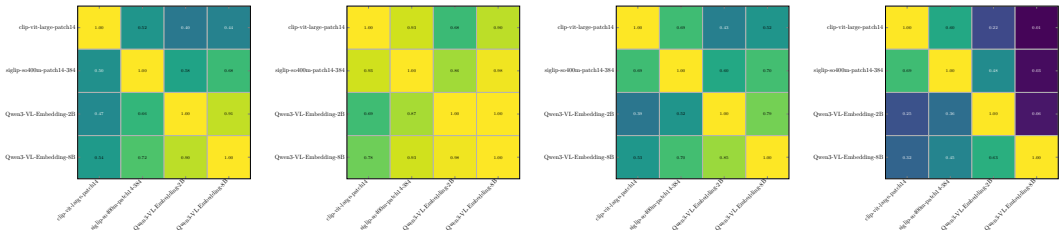
Fidelity produces another ordering (Appendix Table 15). RR has the highest cosine (0.820 text, 0.970 image), while pinv has the lowest text cosine (0.272); normalized error is highest for pinv and lowest for IRP on text and Linear (SGD) on image. Neither ordering reproduces top-1 retrieval. RR illustrates why: image cosine is 0.970, but top-1 retrieval is 0.342 and the Qwen3-VL 2B→8B pair fails at 0.06. Similarity between smooth anchor-affinity profiles need not preserve item rank or decision boundaries. Appendix E reports the full matrices.

The diagnostics agree on family structure but rank individual pairs differently (Appendix D). Compatibility is therefore family-structured, translator-dependent, and not certified by any single metric. Downstream transfer tests whether the pattern survives a functional decision interface.



(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 2: Word-anchor top-1 paired-item retrieval after translation. Rows are source models and columns are targets; panels are translators. Higher values mean the correct paired target is ranked first.



(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 3: ImageNet-anchor top-1 paired-item retrieval after translation. Rows are source models and columns are targets; panels are translators. Higher values mean the correct paired target is ranked first.

5.2 Downstream transfer is pair- and task-dependent

The preceding diagnostics establish partial geometric and item-level agreement. Functional transfer asks whether a decision rule trained on target model B can act on a representation translated from source model A. We use the four translators as complementary probes of this universality question, not as competitors.

Tables 1 and 2 report all-pair and self-translation accuracy ($T_{i\to j}/T_{i\to i}$). Self-pairs are not native-model evaluations: they still apply a potentially lossy translator, so their values vary by method. Appendix Table 3 averages the results over tasks.

Across tasks, modalities, anchor protocols, and translators, performance drops when source and target models differ.

Figures 4, 5, and 6 show pairwise transfer for AG News, DBpedia-14, and ImageNet-1k. Brighter blocks concentrate among related designs, while cross-family pairs transfer less reliably. The same pattern appears across tasks and translators, although absolute accuracy changes. Because architecture,

Table 1: Sentence anchors (850 WikiText sentences), full test sets. Accuracy is all ordered pairs / self-pairs; macro-F1 averages all ordered pairs.

Task	Linear (pinv)		Linear (SGD)		IRP		RR	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
SST-2	0.761 / **0.905**	0.758	0.770 / 0.889	0.752	**0.803** / 0.900	**0.799**	0.800 / 0.903	0.778
DBpedia-14	0.566 / 0.958	0.555	**0.769** / 0.979	**0.761**	0.711 / 0.963	0.698	0.680 / **0.983**	0.660
Emotions	0.406 / **0.713**	0.293	0.461 / 0.656	**0.328**	0.447 / 0.654	0.296	**0.468** / 0.653	0.281
AG News	0.621 / **0.918**	0.618	**0.754** / 0.902	**0.744**	0.728 / 0.894	0.721	0.712 / 0.908	0.679
IMDB	0.724 / 0.901	0.704	0.729 / 0.891	0.694	**0.788** / 0.911	**0.775**	0.786 / **0.913**	0.758

Table 2: ImageNet anchors, full evaluation sets. Zero-shot matches translated images to target-encoded labels; frozen-embedding tasks train only a classifier on target embeddings. Accuracy is all ordered pairs / self-pairs; macro-F1 averages all ordered pairs.

Task	Linear (pinv)		Linear (SGD)		IRP		RR	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
CIFAR-10 (zero-shot)	0.720 / 0.891	0.703	**0.867 / 0.895**	**0.856**	0.791 / 0.873	0.770	0.453 / 0.566	0.395
CIFAR-100 (zero-shot)	0.418 / 0.668	0.401	**0.595 / 0.678**	**0.579**	0.436 / 0.607	0.421	0.178 / 0.330	0.159
CIFAR-10 (frozen embedding)	0.755 / **0.978**	0.748	**0.960** / 0.977	**0.960**	0.814 / 0.970	0.808	0.757 / 0.973	0.736
CIFAR-100 (frozen embedding)	0.465 / **0.838**	0.453	**0.687** / 0.833	**0.680**	0.464 / 0.778	0.447	0.392 / 0.779	0.368
ImageNet-1k (zero-shot)	0.356 / **0.599**	0.329	**0.486** / 0.596	**0.456**	0.328 / 0.507	0.299	0.084 / 0.168	0.073

objective, readout, and data vary jointly, these matrices cannot identify a single cause; they do show that preserving the target decision function is pair-dependent.

5.3 Compatibility is family-structured but not universal

The average-linkage dendrograms in Figure 7 summarize compatibility across translators and tasks. Same-family pairs generally form the tightest clusters, but some cross-family pairs are closer than models from the same family. For example, e5-large-v2 is closer to GTR-T5-XXL than to all-roberta-large-v1, while mxbai-embed-large-v1 is closer to INSTRUCTOR-XL. The late-joining e5 and stella clusters also show that training objective and embedding scale can separate otherwise related models. Compatibility is therefore family-structured but not determined by architecture alone.

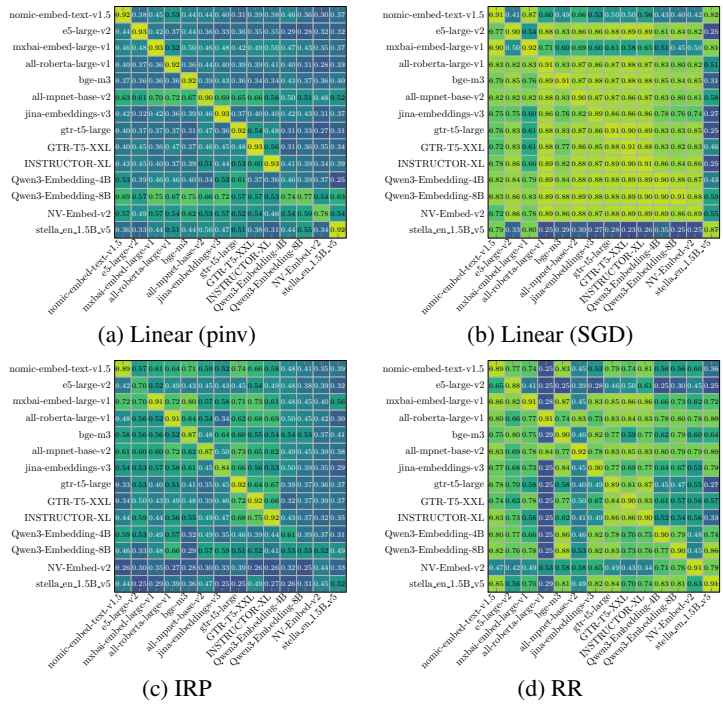
6 Discussion

What is shared, and what is not universal. The experiments support a qualified shared-geometry hypothesis. Native CKA, same-family retrieval, and transfer blocks show that independently trained encoders recover common relational structure. They do not imply a model-independent coordinate system: every translator retains a cross-model transfer gap, including Linear (SGD), and failures concentrate across families. The evidence supports family-conditional compatibility rather than universal interchangeability.

Implications for latent interfaces. The distinction also matters for image-generator conditioning, cross-model KV-cache reuse, and latent communication. Compatible pairs motivate learned latent interfaces, while structured failures argue for validated, pair-specific adapters. These applications are harder than the present pooled-vector benchmark because they expose sequence- and layer-level states. Our results do not rule out other objectives, data regimes, or nonlinear maps for particular pairs, but show that simple linear maps are not universally sufficient.

7 Limitations and Broader Impact

Scope and causal interpretation. The benchmark is heterogeneous rather than factorial. Architecture, objective, data, tokenizer, readout, normalization, dimension, and scale vary jointly, so family associations are not causal estimates of any one factor. Controlled same-family size comparisons remain an important direction.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

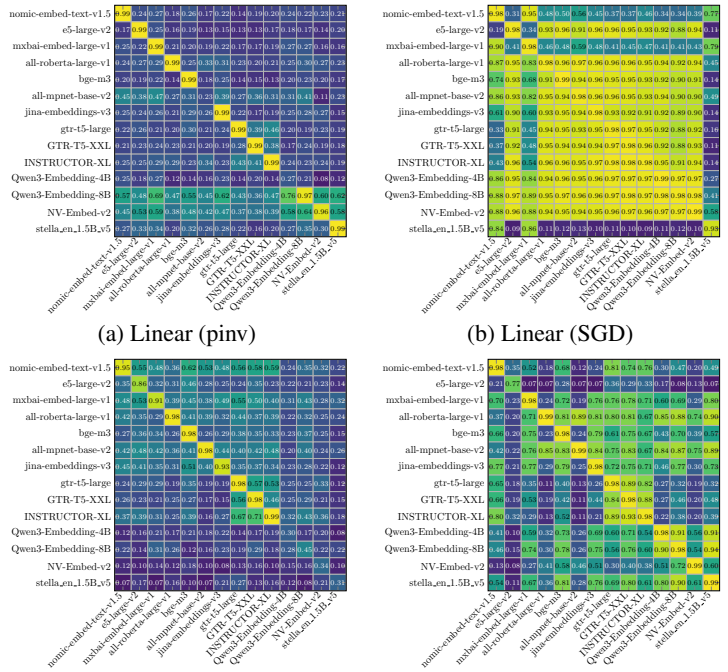
Figure 4: Word-anchor AG News transfer accuracy. Rows are source models, columns are target models, and panels are translators.

Limitation of our translators. We did not optimize translator hyperparameters, search for pair-specific anchors, or tune anchor selection. Instead, we used fixed 850-word and 850-sentence sets, while also testing Linear (SGD) with substantially more alignment data. The comparison is therefore a test of simple, declared protocols rather than a search for each method's best configuration.

8 Conclusion

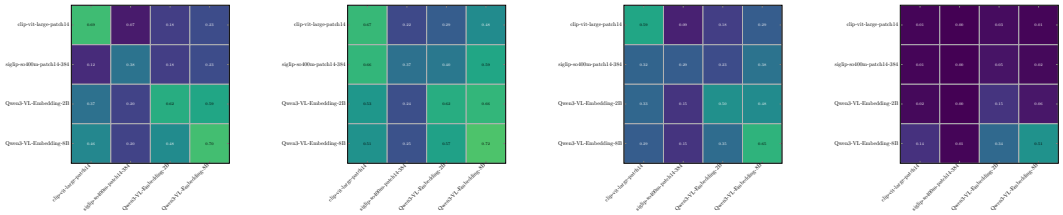
Across 14 text and four image/multimodal encoders, the experiments reveal both shared structure and a clear boundary to its universality. Simple translators recover meaningful information for selected pairs, especially same-family size variants.

These successes do not extend uniformly across heterogeneous encoders. Cross-model transfer remains below self-translation for every translator and protocol, the largest losses occur across model families, and high native CKA does not guarantee preservation of the target decision interface. The tested representations therefore do not form a model-independent geometry that a simple translator can expose universally. They support useful but conditional interfaces whose validity depends on the source–target pair, direction, translator, anchor protocol, and receiving task; each of these choices must be validated before deployment.



(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 5: Word-anchor DBpedia-14 transfer accuracy. Rows are source models, columns are target models, and panels are translators.



(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 6: ImageNet-1k zero-shot transfer accuracy with ImageNet anchors. Rows are source models, columns are target models, and panels are translators.

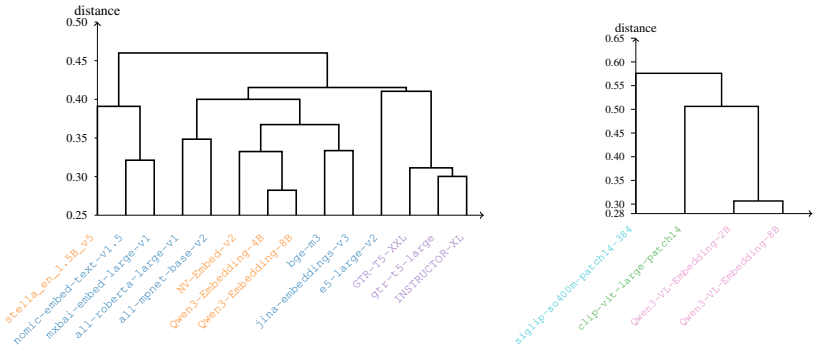


Figure 7: Transfer-compatibility dendrograms (distance $= 1 - \mathrm{Compat}$): text with word anchors (left) and image (right). Colors denote backbone family; scores average four translators and all tasks.

AI USE STATEMENT

Generative AI tools were used to assist manuscript restructuring, language editing, LaTeX preparation, and the derivation and checking of descriptive summaries from existing result artifacts. They were not used to generate experimental observations. The authors checked the resulting prose, equations, numerical claims, citations, and compiled document against the source artifacts and take responsibility for the final content.

ETHICS STATEMENT

This work uses existing pretrained models and benchmark datasets to study representation compatibility. It involves no new data collection or interaction with human participants. Because representation translation could be misapplied as a universal interface, we report pair-specific failures and limitations rather than claiming reliable latent communication across arbitrary models.

REPRODUCIBILITY STATEMENT

Section 4 summarizes the models, anchors, tasks, fitting protocol, and evaluation interfaces. Appendix B provides the corresponding setup details, including the software environment, data splits, preprocessing, classifier protocols, pairwise evaluation, and execution notes; Appendix Tables 5 and 6 list the model metadata. The paper reports the fixed anchor protocols, alignment sample sizes, preprocessing, fitting settings, and evaluation metrics used for the experiments.

REFERENCES

Carl Allen and Timothy Hospedales. Analogies Explained: Towards Understanding Word Embeddings. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 223–231. PMLR, 2019. URL https://proceedings.mlr.press/v97/allen19a.html.

Marcus Armstrong, Navid Ayoobi, and Arjun Mukherjee. Thinking in Different Spaces: Domain-Specific Latent Geometry Survives Cross-Architecture Translation, 2026. URL https://arxiv.org/abs/2603.20406.

Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting Model Stitching to Compare Neural Representations, 2021. URL http://arxiv.org/abs/2106.07682.

Irene Cannistraci, Luca Moschella, Marco Fumero, Valentino Maiorca, and Emanuele Rodolà. From Bricks to Bridges: Product of Invariances to Enhance Latent Space Communication, 2024. URL http://arxiv.org/abs/2310.01211.

Alan Chen, Jack Merullo, Alessandro Stolfo, and Ellie Pavlick. Transferring Linear Features Across Language Models With Model Stitching, 2025. URL http://arxiv.org/abs/2506.06609.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-Linguality, Multi-Functionality, Multi-Granularity Text Embeddings Through Self-Knowledge Distillation, 2024. URL https://arxiv.org/abs/2402.03216.

Adrián Csiszárik, Péter Kőrősi-Szabó, Ákos Matszangosz, Gergely Papp, and Dániel Varga. Similarity and Matching of Neural Network Representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 5656–5668. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/hash/2cb274e6ce940f47beb8011d8ecb1462-Abstract.html.

Zhuoyun Du, Runze Wang, Huiyu Bai, Zouying Cao, Xiaoyong Zhu, Yu Cheng, Bo Zheng, Wei Chen, and Haochao Ying. Enabling Agents to Communicate Entirely in Latent Space, 2026. URL http://arxiv.org/abs/2511.09149.

Tianyu Fu, Zihan Min, Hanling Zhang, Jichao Yan, Guohao Dai, Wanli Ouyang, and Yu Wang. Cache-to-cache: Direct semantic communication between large language models. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2510.03215.

Matt Gorbett and Suman Jana. Characterizing Linear Alignment Across Language Models, 2026. URL http://arxiv.org/abs/2603.18908.

Fabian Gröger, Shuo Wen, and Maria Brbić. Revisiting the Platonic Representation Hypothesis: An Aristotelian View, 2026. URL https://arxiv.org/abs/2602.14486. ICML 2026.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The Platonic Representation Hypothesis, 2024. URL http://arxiv.org/abs/2405.07987.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of Neural Network Representations Revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529. PMLR, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Zorah Lähner and Michael Moeller. On the Direct Alignment of Latent Spaces. In *Proceedings of UniReps: The First Workshop on Unifying Representations in Neural Models*, pp. 158–169. PMLR, 2024. URL https://proceedings.mlr.press/v243/lahner24a.html.

Guillaume Lample, Alexis Conneau, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=H196sainb.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models, 2024. URL https://arxiv.org/abs/2405.17428. ICLR 2025 (Spotlight).

Karel Lenc and Andrea Vedaldi. Understanding Image Representations by Measuring Their Equivariance and Equivalence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 991–999, 2015. URL https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Lenc_Understanding_Image_Representations_2015_CVPR_paper.html.

Mingxin Li, Yanzhao Zhang, Dingkun Long, Keqin Chen, Sibo Song, Shuai Bai, Zhibo Yang, Pengjun Xie, An Yang, Dayiheng Liu, Jingren Zhou, and Junyang Lin. Qwen3-VL-Embedding and Qwen3-VL-Reranker: A Unified Framework for State-of-the-Art Multimodal Retrieval and Ranking, 2026. URL https://arxiv.org/abs/2601.04720.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent Space Translation via Semantic Alignment. *Advances in Neural Information Processing Systems*, 36:55394–55414, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/ad5fa03c906ca15905144ca3fbf2a768-Abstract-Conference.html.

Valentino Maiorca, Luca Moschella, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent Space Translation via Inverse Relative Projection, 2024. URL http://arxiv.org/abs/2406.15057.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, 2013a. URL http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013b. URL https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html.

Mixedbread AI. mxbai-embed-large-v1: A State-of-the-Art English Embedding Model, 2024. URL https://huggingface.co/mixedbread-ai/mxbai-embed-large-v1.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication, 2023. URL http://arxiv.org/abs/2209.15430.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyu Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. Large Dual Encoders Are Generalizable Retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 9844–9855. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.669. URL https://aclanthology.org/2022.emnlp-main.669/.

Zach Nussbaum, John X. Morris, Brandon Duderstadt, and Andriy Mulyar. Nomic Embed: Training a Reproducible Long Context Text Embedder, 2024. URL https://arxiv.org/abs/2402.01613. TMLR.

Can Qin, Ning Yu, Chen Xing, Shu Zhang, Zeyuan Chen, Stefano Ermon, Yun Fu, Caiming Xiong, and Ran Xu. Gluegen: Plug and play multi-modal encoders for x-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23028–23039, 2023. doi: 10.1109/ICCV51070.2023.02110. URL https://arxiv.org/abs/2303.10056.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410.

Sentence Transformers. all-mpnet-base-v2, 2021a. URL https://huggingface.co/sentence-transformers/all-mpnet-base-v2.

Sentence Transformers. all-roberta-large-v1, 2021b. URL https://huggingface.co/sentence-transformers/all-roberta-large-v1.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MPNet: Masked and Permuted Pre-training for Language Understanding. In *Advances in Neural Information Processing Systems*, volume 33, pp. 16857–16867, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c3a690be93aa602ee2dc0ccab5b7b67e-Abstract.html.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Nan Wang, and Han Xiao. jina-embeddings-v3: Multilingual Embeddings with Task LoRA, 2024. URL https://arxiv.org/abs/2409.10173.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One Embedder, Any Task: Instruction-Finetuned Text Embeddings, 2023. URL https://arxiv.org/abs/2212.09741. ACL 2023 (Findings).

Oscar Thorsted Svendsen, Nikolaj Holst Jakobsen, Fabian Mager, and Hiba Nassar. Improving Relative Representations with Learned Anchors and Whitened Inner Products, 2026. URL https://arxiv.org/abs/2605.30596.

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.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text Embeddings by Weakly-Supervised Contrastive Pre-training, 2024. URL http://arxiv.org/abs/2212.03533.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davador, Sylvain Gugger, et al. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, 2020. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6.

Fu-Chun Yang and Jason Eshraghian. Direct Semantic Communication Between Large Language Models via Vector Translation, 2025. URL https://arxiv.org/abs/2511.03945.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 5760–5770, 2023. doi: 10.48550/arXiv.2303.15343.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and Stella: Distillation of SOTA Embedding Models, 2024. URL https://arxiv.org/abs/2412.19048.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. Qwen3 Embedding: Advancing Text Embedding and Reranking Through Foundation Models, 2025. URL https://arxiv.org/abs/2506.05176.

Rui-Jie Zhu, Tianhao Peng, Tianhao Cheng, Xingwei Qu, Jinfa Huang, Dawei Zhu, Hao Wang, Kaiwen Xue, Xuanliang Zhang, Yong Shan, Tianle Cai, Taylor Kergan, Assel Kembay, Andrew Smith, Chenghua Lin, Binh Nguyen, Yuqi Pan, Yuhong Chou, Zefan Cai, Zhenhe Wu, Yongchi Zhao, Tianyu Liu, Jian Yang, Wangchunshu Zhou, Chujie Zheng, Chongxuan Li, Yuyin Zhou, Zhoujun Li, Zhaoxiang Zhang, Jiaheng Liu, Ge Zhang, Wenhao Huang, and Jason Eshraghian. A Survey on Latent Reasoning, 2025. URL http://arxiv.org/abs/2507.06203.

Jiaru Zou, Xiyuan Yang, Ruizhong Qiu, Gaotang Li, Katherine Tieu, Pan Lu, Ke Shen, Hanghang Tong, Yejin Choi, Jingrui He, James Zou, Mengdi Wang, and Ling Yang. Latent Collaboration in Multi-Agent Systems, 2025. URL http://arxiv.org/abs/2511.20639.

A From Word Analogies to an Anchor-Coordinate View of Linear Translation

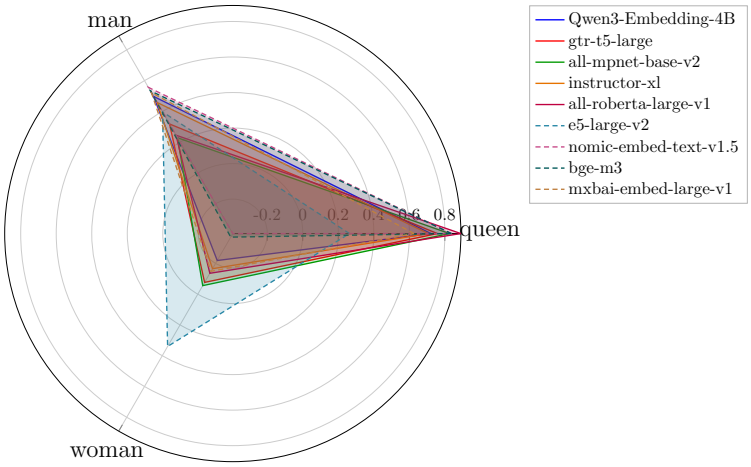
This appendix illustrates the anchor-coordinate interpretation of Equation 1. For each encoder in an initial nine-model text subset, we solve

$$(\alpha_A, \beta_A, \gamma_A) = \arg\min_{\alpha,\beta,\gamma} \left\| x^A_{\text{king}} - \alpha x^A_{\text{queen}} - \beta x^A_{\text{man}} - \gamma x^A_{\text{woman}} \right\|_2^2, \tag{5}$$

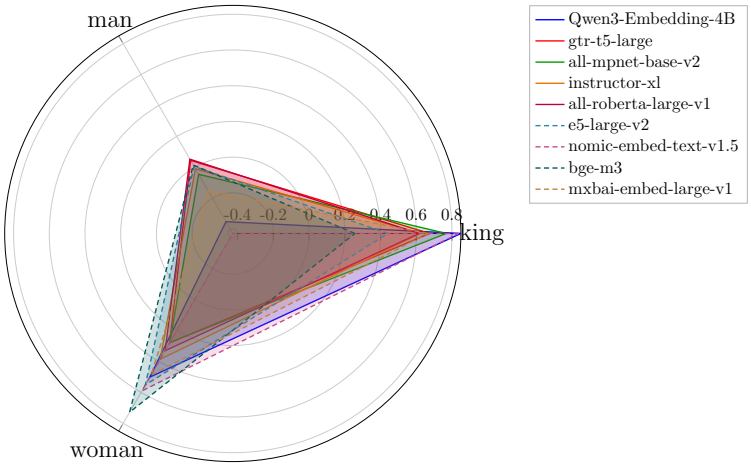
We repeat the fit with *queen* as the reconstructed concept and [*king*, *man*, *woman*] as the anchor dictionary. In matrix notation, Equation 5 is $c^*_A = X_A^+ x_A$. Applying the direct least-squares map of Maiorca et al. (2023) then gives $\hat{y}_B = X_B X_A^+ x_A = X_B c^*_A$. Thus, the map transports the decomposition computed in source model A and evaluates it in target model B. The three-concept analogy and the m-anchor estimator use the same two-stage algebra: compute coefficients in the source dictionary, then evaluate those coefficients in the target dictionary.

Figure 8 shows that the coarse sign pattern persists across the tested encoders, whereas the fitted magnitudes vary. The shared sign pattern helps explain why a small semantic decomposition can transfer at all; the variation in magnitude explains why that transfer need not be model-independent. The relevant condition is not that every encoder realize the exact vector $(1, 1, -1)^\top$, but that source-dictionary coefficients reconstruct the corresponding point accurately when evaluated in the target dictionary.

Figure 9 extends the same construction toward the 850-anchor estimator used in the benchmark. Additional anchors enlarge the reconstructible span, but they can also introduce small singular values and destabilize the least-squares coefficients. Anchor count therefore creates a coverage–conditioning trade-off; increasing it does not guarantee monotonic improvement across models. The main benchmark evaluates the 850-anchor endpoint of this construction.

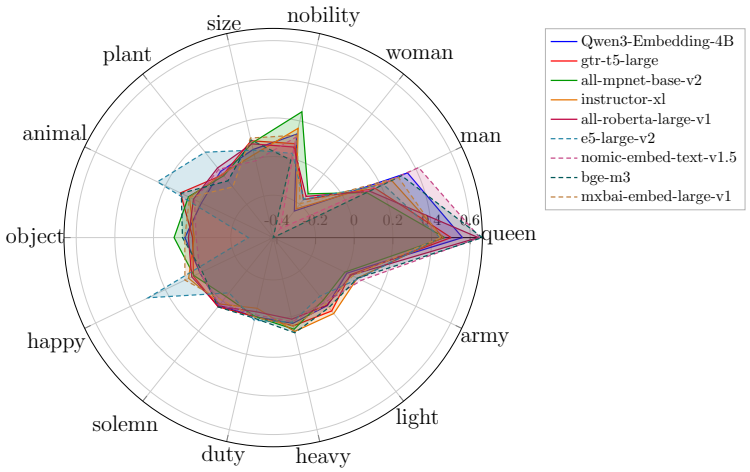


(a) *king* $\approx \alpha$ *queen* $+ \beta$ *man* $+ \gamma$ *woman*.

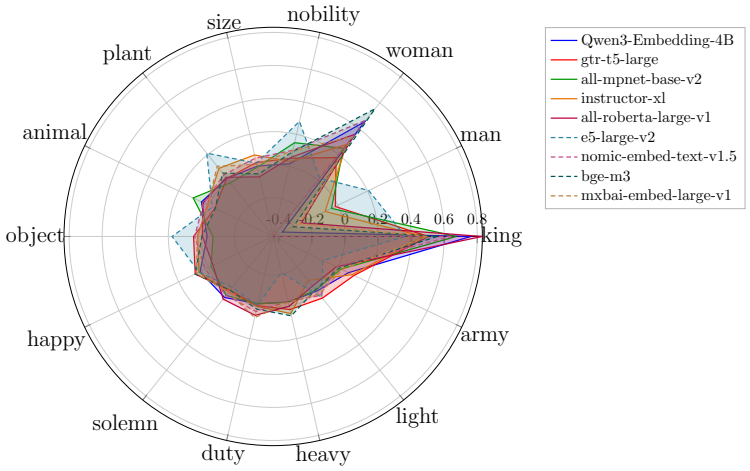


(b) *queen* $\approx \alpha$ *king* $+ \beta$ *man* $+ \gamma$ *woman*.

Figure 8: Least-squares coefficients in a three-concept anchor dictionary. Across the nine encoders, the dominant sign pattern matches the classical analogy: positive *queen* and *man* with negative *woman* for reconstructing *king*, and the mirrored pattern for *queen*. Coefficient magnitudes remain model-dependent.



(a) Reconstructing *king* with a 14-concept dictionary.



(b) Reconstructing *queen* with a 14-concept dictionary.

Figure 9: Expanded-dictionary coefficients. The royalty–gender pattern remains visible, while secondary coefficients vary more across encoders. Increasing the anchor dictionary therefore improves expressivity but also exposes more opportunities for cross-model coefficient mismatch.

B EXPERIMENTAL SETUP DETAILS

This appendix records the settings underlying the reported tables, transfer matrices, and compatibility plots. Each experimental job runs on one machine with sufficient CPU, GPU, and storage, while independent jobs for different models, tasks, and ordered pairs may be distributed across machines (nodes on a cluster computer). No reported result depends on machine identity or machine-local settings.

B.1 SOFTWARE AND EXECUTION ENVIRONMENT

The experiments use Python 3.12, PyTorch 2.13.0, and CUDA 13.2. Text and image models are loaded through Transformers (Wolf et al., 2020) and Sentence-Transformers (Reimers & Gurevych, 2019), together with their model-specific processors and tokenizers. Dataset loading uses datasets==5.0.1. Scikit-learn and SciPy provide the classifiers, least-squares solvers, and projection routines; h5py reads HDF5 embeddings; and pandas and Parquet tooling export tables. Qwen vision-language preprocessing uses qwen-vl-utils, while SigLIP processing requires SentencePiece.

Independent model, task, and ordered-pair evaluations run in parallel. Each job loads and releases its model independently, so changing the allocation of jobs across machines does not change the reported metrics or figures.

B.2 ANCHOR SETS

All anchor examples are paired across models and kept in the same order.

Text words. The word protocol uses 850 entries from the fixed Basic English vocabulary.

Text sentences. The sentence protocol draws candidates from the wikitext-2-raw-v1 training split. Rows are whitespace-normalized and split into sentences; candidates shorter than three words are discarded. The first 850 entries of the fixed selection are used for every model, and word and sentence protocols are reported separately.

Image anchors. The image protocol uses 850 ImageNet training images, and the same image identities and preprocessing are used for every model.

B.3 DOWNSTREAM TASKS AND SPLITS

The text tasks are SST-2, DBpedia-14, Emotions, AG News, and IMDB. Where the dataset provides them, the standard training, validation, and test splits are used. The downstream classifier is a two-layer MLP with hidden sizes $[256, 128]$, at most 200 iterations, early stopping, and a validation fraction of 0.1.

The image tasks are CIFAR-10 and CIFAR-100 zero-shot classification, CIFAR-10 and CIFAR-100 frozen-embedding classification, and ImageNet-1k zero-shot classification. CIFAR training data trains the downstream heads, and the original test split is divided into a validation portion and an evaluation portion following the benchmark data loader. ImageNet uses a local 50,000-image validation set, with 5,000 images reserved for validation and 45,000 for evaluation.

For image zero-shot classification, class names are embedded with the target model's text protocol. A source image embedding is transformed into the target protocol and compared with every target class-name embedding by cosine similarity, and the highest-scoring class is predicted.

B.4 PAIRWISE EVALUATION

For every task and every ordered source-to-target pair, including self-pairs, the following procedure is used:

1. obtain source and target embeddings for the same examples;
2. fit or load the selected source-to-target transformation $T_{A \to B}$;

3. transform the source embeddings;
4. evaluate with the target downstream protocol or the target text-label protocol;
5. record accuracy, macro-F1, timing, alignment statistics, and resource usage.

For each task and translator, the benchmark table reports the mean accuracy over all ordered pairs, the mean accuracy over diagonal pairs, and the mean macro-F1.

B.5 Alignment-quality diagnostics

The alignment diagnostics use paired examples drawn from the alignment data. Text uses 5,000 unique non-empty preprocessed WikiText examples, and image uses a fixed paired sample of 5,000 ImageNet examples. The same item identities are used across all models and translators.

For each ordered source-to-target pair, we report:

Top-k retrieval. Each transformed source embedding is compared by cosine similarity with all target embeddings, and top-1, top-5, and top-10 accuracy record whether the paired target example is retrieved.

Linear CKA. Centered Linear CKA is computed between the transformed source and target embeddings. A separate native Linear CKA measurement compares the native source and target embeddings directly, without stitching.

Normalized error. The reported value is the mean of

$$\frac{\|y_{\text{target}} - y_{\text{stitched}}\|_2}{\|y_{\text{target}}\|_2},$$

and lower values indicate better reconstruction.

Cosine similarity. The reported value is the mean row-wise cosine similarity between the transformed source and target embeddings, and higher values indicate better alignment.

For Relative Representation (RR), these diagnostics compare source-relative and target-relative anchor-affinity profiles. The native Linear CKA baseline is symmetric, whereas the post-translation metrics are directional because the map $T_{A\to B}$ differs from its reverse $T_{B\to A}$.

B.6 Parallel execution and reproducibility

The independent computations are parallelized over models, tasks, anchor protocols, and ordered model pairs. Embedding extraction, downstream training, learned-linear training, benchmark evaluation, and alignment diagnostics are separate stages, so each stage can be reproduced independently on a single machine or distributed across machines. The exact model configurations, anchor files, sample identities, preprocessing settings, and per-pair JSON results are retained with the experiment outputs.

C Full Transfer Results and Model Inventory

C.1 Aggregate and per-task transfer

Table 3: Transfer aggregated over tasks. "All" averages all ordered pairs, "cross" excludes self-pairs, "self" averages $A \to A$, and the gap is self minus cross. Self is a method-specific calibration, not native-model accuracy. Macro-F1 averages all ordered pairs.

Modality	Method	Acc. all	Acc. cross	Acc. self	Gap	Macro-F1
Text (words)	Linear (pinv)	0.504	0.478	0.847	0.370	0.462
Text (words)	Linear (SGD)	**0.697**	**0.684**	0.863	**0.180**	**0.656**
Text (words)	IRP	0.542	0.522	0.799	0.277	0.490
Text (words)	RR	0.666	0.651	**0.865**	0.214	0.608
Text (sentences)	Linear (pinv)	0.616	0.595	**0.879**	0.284	0.586
Text (sentences)	Linear (SGD)	**0.697**	**0.684**	0.863	**0.180**	0.656
Text (sentences)	IRP	0.695	0.682	0.864	0.182	**0.658**
Text (sentences)	RR	0.689	0.675	0.872	0.197	0.631
Image	Linear (pinv)	0.543	0.459	0.795	0.336	0.527
Image	Linear (SGD)	**0.719**	**0.693**	**0.796**	**0.102**	**0.706**
Image	IRP	0.567	0.507	0.747	0.240	0.549
Image	RR	0.373	0.309	0.563	0.254	0.346

Table 4: Word anchors (850 Basic English words), full test sets. Accuracy is all ordered pairs / self-pairs; macro-F1 averages all ordered pairs.

Task	Linear (pinv)		Linear (SGD)		IRP		RR	
	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
SST-2	0.690 / 0.880	0.670	0.770 / 0.889	0.752	0.759 / 0.877	0.746	**0.807 / 0.895**	**0.792**
DBpedia-14	0.316 / 0.879	0.293	**0.769 / 0.979**	**0.761**	0.326 / 0.781	0.291	0.545 / 0.967	0.513
Emotions	0.391 / **0.715**	0.290	0.461 / 0.656	0.328	0.427 / 0.673	0.299	**0.525** / 0.650	**0.339**
AG News	0.475 / 0.869	0.447	**0.754 / 0.902**	**0.744**	0.500 / 0.774	0.450	0.666 / **0.902**	0.634
IMDB	0.649 / 0.894	0.612	0.729 / 0.891	0.694	0.698 / 0.892	0.663	**0.787 / 0.911**	**0.763**

Table 3 aggregates transfer over tasks and separates cross-model pairs from the diagonal calibration $T_{i \to i}$. Table 4 provides the word-anchor task breakdown that complements the sentence-anchor results in the main paper. The method-specific values reveal when shared structure is recoverable rather than serving as a leaderboard. The high-supervision Linear (SGD) reference reaches cross-model accuracy of 0.684 on text and 0.693 on image. On text, sentence-anchor IRP reaches 0.682, sentence-anchor RR 0.675, word-anchor RR 0.651, and word-anchor IRP 0.522. The word-anchor task results localize the largest losses: on DBpedia-14, pinv and IRP obtain 0.316 and 0.326, respectively, compared with 0.769 for Linear (SGD); AG News shows the same separation, at 0.475, 0.500, and 0.754. RR behaves differently from the reconstructive anchor methods because it compares relative profiles, reaching 0.545 on DBpedia-14 and 0.787 on IMDB. Finally, the diagonal calibration shows that cross-model loss is not a constant property of a translator: on word-anchor text, pinv falls 0.370 below its self-translation accuracy, whereas the Linear (SGD) gap is 0.180.

C.2 Model inventory

Table 5 lists the 14 text encoders. The suite spans BERT-like, T5-style, and causal-decoder backbones, parameter counts from 109M to 11B, four readout and normalization conventions, and embedding dimensions from 768 to 4096, so architecture, training objective, tokenizer, and scale vary together rather than independently. Table 6 lists the four image and multimodal encoders, which cover CLIP-like and SigLIP-like contrastive encoders and unified VLMs with dimensions from 768 to 4096. Table 7 reports the pair-level agreement between the word-anchor and sentence-anchor compatibility matrices.

Table 5: Text embedding models. The suite spans BERT-like encoders, T5-style encoders, and causal-decoder backbones, with parameter counts from 109M to 11B, different tokenizers, readouts, normalization conventions, and dimensions (768–4096). "Family" groups models by backbone; parameter counts are reported model sizes, while readout and normalization follow the defaults used by our extraction pipeline.

Model	Parameters	Family	Readout	Norm.	Dim.
`Qwen3-Embedding-8B` (Zhang et al., 2025)	8B	causal decoder	last token	yes	4096
`Qwen3-Embedding-4B` (Zhang et al., 2025)	4B	causal decoder	last token	yes	2560
`NV-Embed-v2` (Lee et al., 2024)	7.8B	causal decoder	last token	yes	4096
`stella_en_1.5B_v5` (Zhang et al., 2024)	1.5B	causal decoder	mean	no	1024
`jina-embeddings-v3` (Sturua et al., 2024)	570M	BERT-like	mean	yes	1024
`e5-large-v2` (Wang et al., 2024)	335M	BERT-like	mean	yes	1024
`mxbai-embed-large-v1` (Mixedbread AI, 2024)	335M	BERT-like	mean	no	1024
`all-roberta-large-v1` (Sentence Transformers, 2021b)	355M	BERT-like	mean	yes	1024
`bge-m3` (Chen et al., 2024)	568M	BERT-like	mean	yes	1024
`all-mpnet-base-v2` (Sentence Transformers, 2021a; Song et al., 2020)	109M	BERT-like	mean	yes	768
`nomic-embed-text-v1.5` (Nussbaum et al., 2024)	137M	BERT-like	mean	no	768
`GTR-T5-XXL` (Ni et al., 2022)	11B	T5-like	mean	yes	768
`gtr-t5-large` (Ni et al., 2022)	335M	T5-like	mean	yes	768
`INSTRUCTOR-XL` (Su et al., 2023)	1.5B	T5-like	mean	yes	768

Table 6: Image/multimodal embedding models. The suite spans CLIP-like and SigLIP-like image–text encoders and unified multimodal VLMs, with parameter counts from 400M to 8B and different dimensions.

Model	Parameters	Family	Readout	Dim.
`clip-vit-large-patch14` (Radford et al., 2021)	428M	CLIP-like	mean	768
`siglip-so400m-patch14-384` (Zhai et al., 2023; Tschannen et al., 2025)	400M	SigLIP-like	mean	1152
`Qwen3-VL-Embedding-2B` (Li et al., 2026)	2B	multimodal VLM	mean	2048
`Qwen3-VL-Embedding-8B` (Li et al., 2026)	8B	multimodal VLM	mean	4096

Table 7: Pair-level agreement between the word-anchor and sentence-anchor compatibility matrices (downstream accuracy, averaged over tasks). Spearman and Kendall correlations are computed over all ordered cross-model pairs.

Method	Spearman ρ	Kendall τ
RR	0.820	0.648
IRP	0.491	0.349
Linear (pinv)	0.656	0.489
Linear (SGD)	1.000	1.000
Average across methods	0.912	0.758

C.3 PER-MODEL AND PER-TASK SUMMARIES

Table 8: Per-model mean symmetric compatibility (off-diagonal, so the self-pair is excluded) under each translator, for word and sentence anchors. "Mean" averages the four methods. Rows are ordered by the mean over both anchor sets.

	Word anchors					Sentence anchors				
Model	RR	IRP	pinv	SGD	Mean	RR	IRP	pinv	SGD	Mean
Qwen3-Embedding-8B	0.690	0.532	0.571	0.737	0.633	0.702	0.699	0.684	0.737	0.706
Qwen3-Embedding-4B	0.670	0.514	0.462	0.741	0.597	0.666	0.686	0.675	0.741	0.692
bge-m3	0.698	0.537	0.462	0.729	0.606	0.724	0.686	0.573	0.729	0.678
INSTRUCTOR-XL	0.660	0.537	0.466	0.726	0.597	0.684	0.714	0.580	0.726	0.676
NV-Embed-v2	0.621	0.467	0.550	0.747	0.596	0.598	0.663	0.675	0.747	0.671
jina-embeddings-v3	0.669	0.518	0.462	0.712	0.591	0.701	0.647	0.547	0.712	0.652
gtr-t5-large	0.639	0.525	0.452	0.708	0.581	0.675	0.686	0.572	0.708	0.660
all-mpnet-base-v2	0.610	0.528	0.490	0.730	0.589	0.625	0.659	0.590	0.730	0.651
GTR-T5-XXL	0.656	0.519	0.451	0.710	0.584	0.681	0.669	0.549	0.710	0.652
mxbai-embed-large-v1	0.700	0.576	0.476	0.597	0.587	0.706	0.704	0.584	0.597	0.648
all-roberta-large-v1	0.575	0.518	0.455	0.735	0.571	0.610	0.682	0.557	0.735	0.646
e5-large-v2	0.550	0.538	0.442	0.715	0.561	0.627	0.680	0.586	0.715	0.652
nomic-embed-text-v1.5	0.672	0.556	0.469	0.562	0.565	0.698	0.701	0.573	0.562	0.633
stella_en_1.5B_v5	0.697	0.446	0.479	0.419	0.510	0.757	0.677	0.586	0.419	0.610

Table 9: Per-model mean symmetric compatibility on image (ImageNet anchors), off-diagonal, averaged over the five image tasks. Rows are ordered by the mean over methods.

Model	RR	IRP	Linear (pinv)	Linear (SGD)	Mean
Qwen3-VL-Embedding-8B	0.353	0.568	0.565	0.730	0.554
Qwen3-VL-Embedding-2B	0.355	0.540	0.536	0.702	0.534
clip-vit-large-patch14	0.278	0.468	0.396	0.686	0.457
siglip-so400m-patch14-384	0.251	0.451	0.339	0.655	0.424

Table 10: Per-model mean compatibility (off-diagonal, averaged over the four translators) broken down by word-anchor task. Rows are ordered by the mean over tasks.

Model	SST-2	DBpedia-14	Emotions	AG News	IMDB	Mean
Qwen3-Embedding-8B	0.797	0.530	0.494	0.610	0.733	0.633
bge-m3	0.779	0.493	0.434	0.601	0.725	0.606
INSTRUCTOR-XL	0.758	0.492	0.434	0.602	0.702	0.597
Qwen3-Embedding-4B	0.767	0.454	0.468	0.578	0.717	0.597
NV-Embed-v2	0.755	0.450	0.503	0.538	0.734	0.596
jina-embeddings-v3	0.737	0.486	0.443	0.586	0.701	0.591
all-mpnet-base-v2	0.723	0.489	0.444	0.627	0.664	0.589
mxbai-embed-large-v1	0.741	0.449	0.413	0.596	0.738	0.587
GTR-T5-XXL	0.723	0.486	0.427	0.603	0.681	0.584
gtr-t5-large	0.726	0.476	0.424	0.594	0.686	0.581
all-roberta-large-v1	0.743	0.476	0.388	0.581	0.665	0.571
nomic-embed-text-v1.5	0.733	0.422	0.388	0.579	0.701	0.565
e5-large-v2	0.765	0.372	0.409	0.541	0.719	0.561
stella_en_1.5B_v5	0.704	0.330	0.401	0.460	0.657	0.510

Table 8 exposes two forms of heterogeneity hidden by the global mean. First, sentence anchors increase the four-translator mean for every text model, but the gains are uneven; models that transfer poorly under the word-anchor reconstructive methods often improve the most. Second, the method-specific columns can reverse the ordering induced by the average. The dendrograms therefore visualize one declared aggregation, not a unique taxonomy of models.

The model ordering also changes with the task. Under word anchors, the per-model rankings for the two topic-classification tasks, DBpedia-14 and AG News, correlate at $\rho = 0.811$, while the

Table 11: Mean off-diagonal compatibility per task under each translator (the diagonal, i.e. self-translation, is excluded). Lower values mean the task is harder to transfer across models.

Modality	Task	RR	IRP	Linear (pinv)	Linear (SGD)
Text (words)	SST-2	**0.800**	0.750	0.675	0.761
Text (words)	DBpedia-14	0.513	0.291	0.272	**0.753**
Text (words)	Emotions	**0.515**	0.408	0.366	0.445
Text (words)	AG News	0.648	0.479	0.445	**0.742**
Text (words)	IMDB	**0.777**	0.683	0.630	0.716
Text (sentences)	SST-2	0.792	**0.795**	0.750	0.761
Text (sentences)	DBpedia-14	0.657	0.692	0.535	**0.753**
Text (sentences)	Emotions	**0.454**	0.431	0.382	0.445
Text (sentences)	AG News	0.697	0.715	0.598	**0.742**
Text (sentences)	IMDB	0.777	**0.778**	0.710	0.716
Image	CIFAR-10 (zero-shot)	0.416	0.763	0.663	**0.858**
Image	CIFAR-100 (zero-shot)	0.127	0.379	0.334	**0.568**
Image	CIFAR-10 (finetune)	0.685	0.763	0.681	**0.954**
Image	CIFAR-100 (finetune)	0.263	0.359	0.341	**0.638**
Image	ImageNet-1k (zero-shot)	0.056	0.269	0.276	**0.449**

two sentiment tasks, SST-2 and IMDB, correlate at 0.741. By contrast, AG News and IMDB have correlation -0.015. A model that translates well for topic classification need not translate equally well for sentiment. Table 11 adds a translator-dependent view: Emotions is the hardest task for Linear (SGD), whereas DBpedia-14 produces the largest word-anchor gap between SGD and pinv.

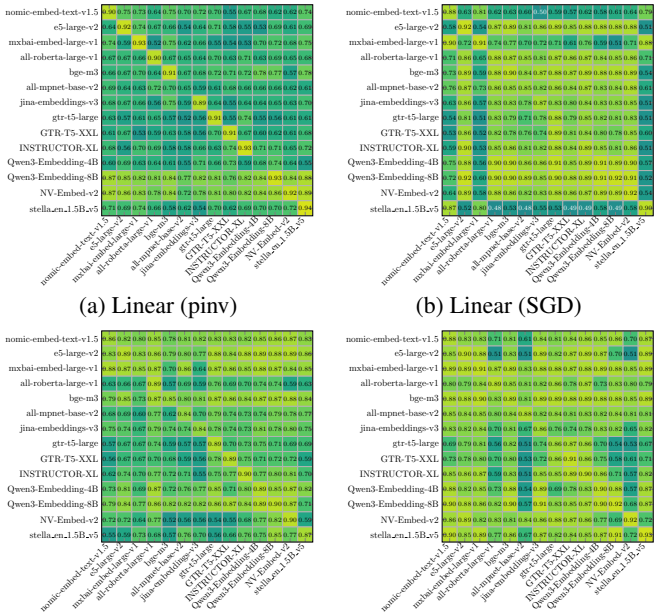
C.4 Remaining transfer matrices

Figures 10–18 complete the pair-level view of Tables 1 and 2 for the tasks not shown in the main paper. In every matrix a row is a source model, a column is a target model, and a panel is a translator, with the same layout as Figure 1; a high value means that the translated source is classified correctly by the decision rule trained on the target model.

Word-anchor text. Task difficulty and pair structure vary together. On SST-2, all four translators exceed 0.69, with RR at 0.807. IMDB is similarly tolerant of anchor-based translation: RR reaches 0.787 and IRP 0.698, compared with 0.729 for Linear (SGD). Emotions is harder under every translator, spanning 0.391–0.525, and distinguishes the translators less sharply. Across all three matrices, the high-accuracy cells still concentrate within model families, while cross-family pairs remain weak under the word-anchor reconstructive methods.

Sentence-anchor text. Sentence anchors substantially improve the methods that reconstruct a target vector. IRP reaches 0.728 on AG News and 0.711 on DBpedia-14, compared with word-anchor values of 0.500 and 0.326; pinv rises to 0.621 and 0.566 from 0.475 and 0.316. As expected for an anchor-independent fit, Linear (SGD) remains at 0.754 and 0.769. RR changes less, reaching 0.712 and 0.680, because it compares relative profiles instead of reconstructing a native target vector. The protocol therefore changes the amount of recoverable transfer most strongly for the reconstructive anchor methods.

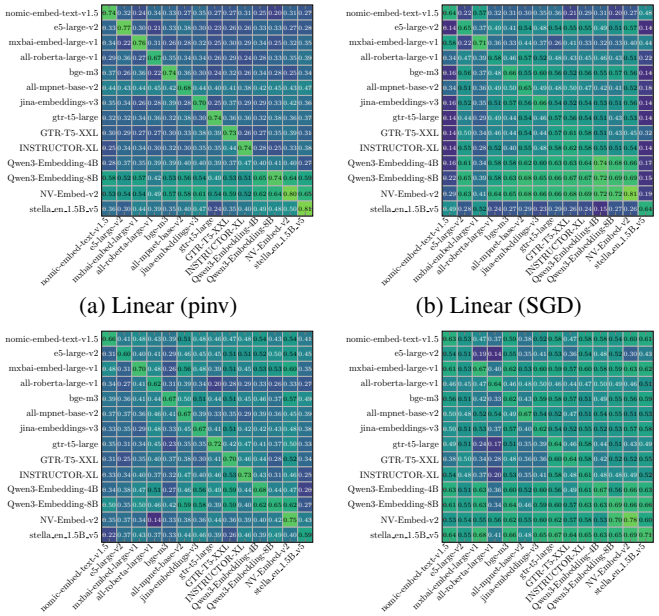
Image tasks. The receiving interface strongly affects image transfer. Linear (SGD) reaches 0.867 on CIFAR-10 and 0.595 on CIFAR-100 under zero-shot matching, compared with 0.960 and 0.687 when a classifier is trained on native target embeddings. The learned target classifier is more tolerant of translation error than direct cosine matching to target-encoded labels. RR shows the largest CIFAR-10 change, from 0.453 zero-shot accuracy to 0.757 frozen-embedding accuracy. Its relative profiles therefore retain information that a trained classifier can use, but they do not preserve an equally effective zero-shot label interface.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

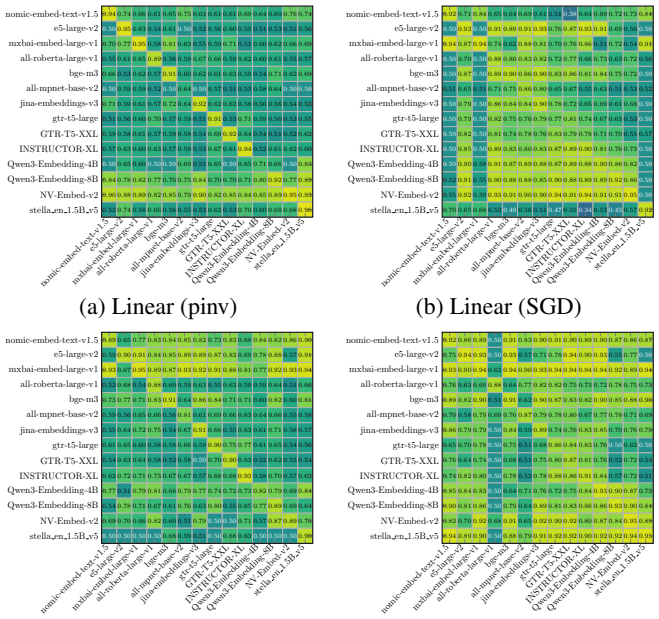
Figure 10: Word-anchor SST-2 transfer accuracy. Rows are source models and columns are targets; panels are translators. Higher values mean the translated source is classified correctly by the head trained on the target.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

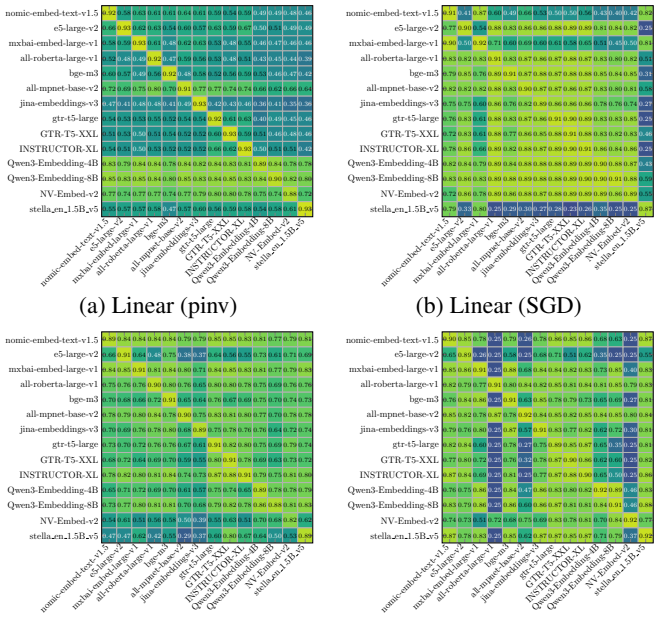
Figure 11: Word-anchor Emotions transfer accuracy. Rows are source models and columns are targets; panels are translators.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

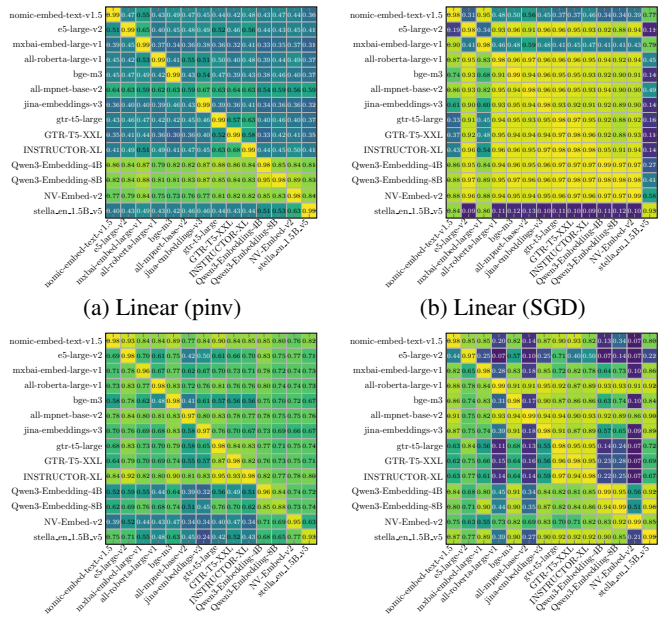
Figure 12: Word-anchor IMDB transfer accuracy. Rows are source models and columns are targets; panels are translators.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

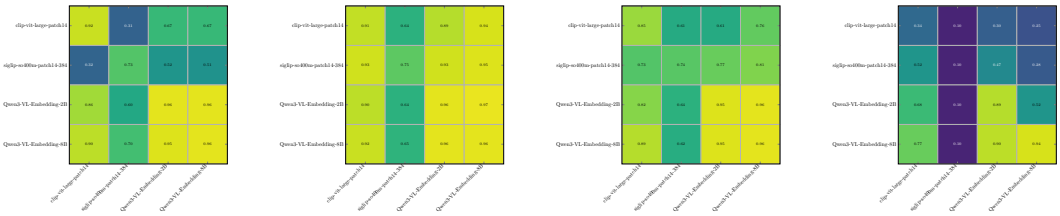
Figure 13: Sentence-anchor AG News transfer accuracy. Rows are source models and columns are targets; panels are translators.



(a) Linear (pinv) (b) Linear (SGD)

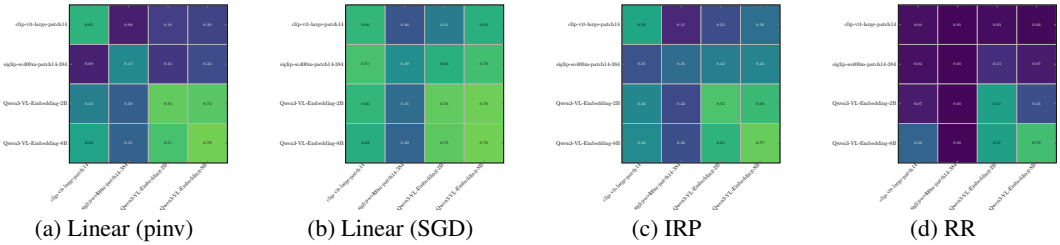
(c) IRP (d) RR

Figure 14: Sentence-anchor DBpedia-14 transfer accuracy. Rows are source models and columns are targets; panels are translators.



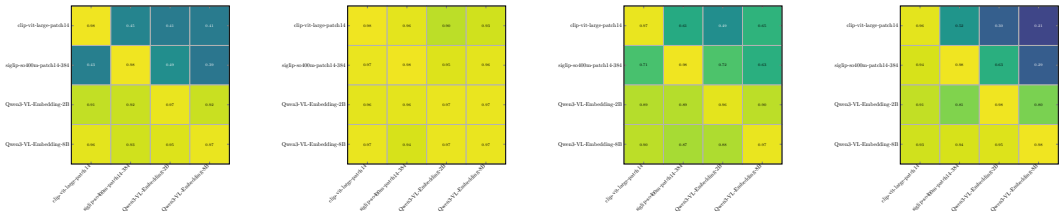
(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 15: CIFAR-10 zero-shot transfer accuracy. Rows are source models and columns are targets; panels are translators. Higher values mean the translated image is matched to the target-encoded class label.



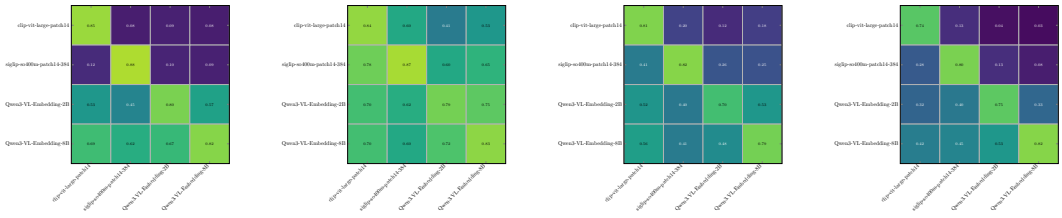
(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 16: CIFAR-100 zero-shot transfer accuracy. Rows are source models and columns are targets; panels are translators.



(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 17: CIFAR-10 frozen-embedding transfer accuracy. Rows are source models and columns are targets; panels are translators. Higher values mean the translated source is classified correctly by the head trained on native target embeddings.



(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 18: CIFAR-100 frozen-embedding transfer accuracy. Rows are source models and columns are targets; panels are translators.

Anchor-protocol comparison. Under sentence anchors, the same families remain tightly grouped and the same-design pairs retain their relative position. The family structure is therefore robust to this protocol change even though absolute accuracies and method-specific orderings change.

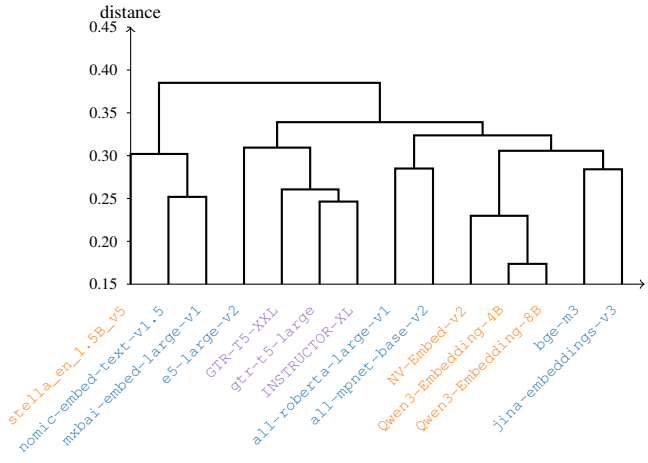


Figure 19: Text compatibility with sentence anchors (distance $= 1 - \mathrm{Compat}$). Colors denote backbone family; scores average four translators and all tasks.

D Cross-diagnostic Agreement

Table 12: Spearman ρ between pairwise compatibility matrices from different diagnostics on text word anchors (each averaged over the four translation methods). Normalized error is negated so that higher is better.

	downstream	CKA	cosine	kNN top-1	norm. err
downstream	–	0.504	0.319	0.500	0.412
CKA	–	–	0.097	0.187	0.177
cosine	–	–	–	0.408	0.804
kNN top-1	–	–	–	–	0.587
norm. err	–	–	–	–	–

Table 13: Spearman ρ between pairwise compatibility matrices from different diagnostics on image ImageNet anchors (each averaged over the four translation methods). Normalized error is negated so that higher is better.

	downstream	CKA	cosine	kNN top-1	norm. err
downstream	–	0.283	-0.240	-0.141	-0.311
CKA	–	–	0.615	0.566	0.434
cosine	–	–	–	0.406	0.909
kNN top-1	–	–	–	–	0.154
norm. err	–	–	–	–	–

Table 14: Method-specific Spearman ρ between downstream transfer and each representation diagnostic over ordered cross-model pairs. Normalized error is negated so that higher is better. The image values are descriptive because four models provide only 12 ordered cross-model pairs.

Modality	Translator	CKA	Cosine	k-NN top-1	$-$Norm. error
Text (words)	RR	0.116	0.529	0.300	0.725
	IRP	0.377	0.419	0.336	0.481
	Linear (pinv)	0.359	0.169	0.281	0.298
	Linear (SGD)	0.749	0.589	0.846	0.606
Image	RR	-0.116	-0.293	-0.127	-0.384
	IRP	-0.121	-0.601	0.469	-0.484
	Linear (pinv)	0.474	0.340	0.240	0.396
	Linear (SGD)	0.255	0.184	0.371	0.170

These tables test whether any representation diagnostic provides a translator-independent proxy for functional compatibility. Tables 12 and 13 correlate pairwise compatibility matrices after averaging over the four translators, whereas Table 14 computes the correlations within each translator. The aggregate view asks whether a broad notion of compatibility persists across diagnostics; the method-specific view asks whether a diagnostic predicts downstream transfer for a fixed interface. The distinction matters. On text, top-1 retrieval correlates strongly with downstream transfer for Linear (SGD) ($\rho = 0.846$), but much more weakly for pinv (0.281) and RR (0.300). Thus, even the relationship between two diagnostics is translator-dependent. For image, four models yield only 12 ordered cross-model pairs, and the signs vary by translator, so the results do not support a stable hierarchy of diagnostics.

E Alignment-quality Matrices and Retrieval Depth

Table 15: Alignment-quality metrics over paired items (WikiText for text, ImageNet for image; 5,000 aligned pairs). Each cell is the mean over all ordered cross-model pairs; the self-pair (diagonal) value is shown in parentheses. Lower normalized error is better; all other metrics are higher-is-better.

Modality	Method	kNN top-1	kNN top-5	kNN top-10	Linear CKA	Cosine	Norm. error
Text (words)	RR	0.329 (1.000)	0.445 (1.000)	0.484 (1.000)	0.419 (1.000)	**0.820 (1.000)**	1.541 (0.000)
Text (words)	IRP	0.176 (0.764)	0.274 (0.797)	0.315 (0.813)	0.530 (0.949)	0.393 (0.822)	**1.386 (0.512)**
Text (words)	Linear (pinv)	0.194 (0.869)	0.285 (0.884)	0.328 (0.893)	0.455 (0.935)	0.272 (0.869)	2.272 (0.295)
Text (words)	Linear (SGD)	**0.756 (1.000)**	**0.828 (1.000)**	**0.845 (1.000)**	**0.701 (0.991)**	0.679 (0.947)	1.447 (0.306)
Image	RR	0.342 (1.000)	0.522 (1.000)	0.590 (1.000)	0.435 (1.000)	**0.970 (1.000)**	0.670 (0.000)
Image	IRP	0.617 (1.000)	0.861 (1.000)	0.910 (1.000)	0.492 (0.980)	0.606 (0.922)	0.939 (0.367)
Image	Linear (pinv)	0.610 (1.000)	0.846 (1.000)	0.890 (1.000)	0.478 (1.000)	0.643 (0.978)	1.023 (0.138)
Image	Linear (SGD)	**0.878 (1.000)**	**0.974 (1.000)**	**0.985 (1.000)**	**0.821 (0.990)**	0.798 (0.970)	**0.615 (0.233)**

Table 16: Sentence-anchor alignment-quality metrics over WikiText paired items (5,000 aligned pairs). Each cell is the mean over all ordered cross-model pairs; the self-pair value is in parentheses. Lower normalized error is better; all other metrics are higher-is-better.

Method	kNN top-1	kNN top-5	kNN top-10	Linear CKA	Cosine	Norm. error
RR	0.341 (1.000)	0.479 (1.000)	0.530 (1.000)	0.518 (1.000)	**0.828 (1.000)**	1.321 (0.000)
IRP	0.740 (0.999)	**0.893 (1.000)**	**0.927 (1.000)**	0.576 (0.960)	0.515 (0.896)	**1.078 (0.419)**
Linear (pinv)	0.558 (0.943)	0.724 (0.959)	0.777 (0.968)	0.370 (0.972)	0.455 (0.920)	1.475 (0.229)
Linear (SGD)	**0.756 (1.000)**	0.828 (1.000)	0.845 (1.000)	**0.701 (0.991)**	0.679 (0.947)	1.447 (0.306)

Table 17: Where the correct paired target ranks among 5,000 candidates, averaged over ordered cross-model pairs. The mutually exclusive rank bands are obtained from top-k retrieval: 2–5 $=$ top5$-$top1, 6–10 $=$ top10 $-$ top5, and $> 10 = 1 -$ top10.

Modality/protocol	Translator	Rank 1	Ranks 2–5	Ranks 6–10	Rank > 10
Text, words	RR	0.329	0.117	0.039	0.516
	IRP	0.176	0.098	0.041	0.685
	Linear (pinv)	0.194	0.091	0.043	0.672
	Linear (SGD)	0.756	0.073	0.017	0.155
Text, sentences	RR	0.341	0.138	0.052	0.470
	IRP	0.740	0.153	0.034	0.073
	Linear (pinv)	0.558	0.167	0.053	0.223
	Linear (SGD)	0.756	0.073	0.017	0.155
Image, ImageNet	RR	0.342	0.180	0.068	0.410
	IRP	0.617	0.243	0.049	0.090
	Linear (pinv)	0.610	0.236	0.044	0.110
	Linear (SGD)	0.878	0.097	0.011	0.015

Table 15 summarizes the word-anchor text and ImageNet-anchor image diagnostics, and Table 16 gives the corresponding text results with sentence anchors. Each entry reports the mean over ordered cross-model pairs, followed in parentheses by the self-pair mean. Lower normalized error is better; all other diagnostics are higher-is-better. Table 17 decomposes retrieval errors by rank, while Figures 20–29 retain the full pairwise structure.

Linear CKA before and after translation. Figure 1 in the main paper compares the two native representations. Its feature matrices are the models' own embeddings of the same items, X for model A and Y for model B; no fitted map is involved. Native CKA is therefore symmetric, with a diagonal of 1. The matrices below instead compute CKA after translation. They replace the source matrix with the translated embeddings $\hat{Y}$ from the fitted directional map and compare $\hat{Y}$ with the native target embeddings Y. Post-translation CKA asks whether a translator reproduces the target model's pairwise geometry, not whether the two models agree before translation. Because each direction has its own map, this diagnostic is directional, and it can differ sharply from native CKA.

The sentence-anchor summary complements the word-anchor results in Table 15. The change is concentrated in the anchor-fitted reconstructive methods: IRP top-1 retrieval rises from 0.176 to 0.740, and pinv rises from 0.194 to 0.558. Anchor-independent Linear (SGD) remains at 0.756, while RR changes little because it does not reconstruct a vector in the native target space.

What top-5 and top-10 reveal. Top-1 retrieval conflates two different failures: ranking the paired target just below first place and losing it from the local neighborhood altogether. Table 17 separates them. With word anchors, the paired text target falls outside the top 10 for 68.5% of IRP queries and 67.2% of pinv queries. Their low top-1 accuracy therefore reflects a loss of neighborhood correspondence, not merely swaps among a few close candidates. Linear (SGD) reduces the outside-top-10 mass to 15.5%.

Sentence anchors change the character of the error. Under IRP, 74.0% of paired targets rank first, a further 15.3% rank from 2 through 5, and only 7.3% remain outside the top 10. IRP therefore exceeds Linear (SGD) at top-5 (0.893 versus 0.828) and top-10 (0.927 versus 0.845), despite trailing slightly at top-1 (0.740 versus 0.756). Sentence-anchor IRP preserves the broader local neighborhood particularly well, whereas Linear (SGD) places a slightly larger share of targets at rank 1. Pinv also improves, although 22.3% of paired targets remain outside its top 10, indicating residual alignment error that the sentence anchors do not remove.

Image retrieval exhibits a third regime. IRP and pinv place approximately one quarter of the paired targets at ranks 2–5 (0.243 and 0.236), making many of their top-1 errors near misses; both exceed 0.89 by top-10. Linear (SGD) is nearly saturated at top-10 (0.985), with only 1.5% of paired images outside the first ten candidates. RR rises from 0.342 at top-1 to 0.590 at top-10, but still loses the paired image from the local neighborhood in 41.0% of cases. This is consistent with the main result: similar anchor-affinity profiles do not guarantee preservation of item-level neighborhoods.

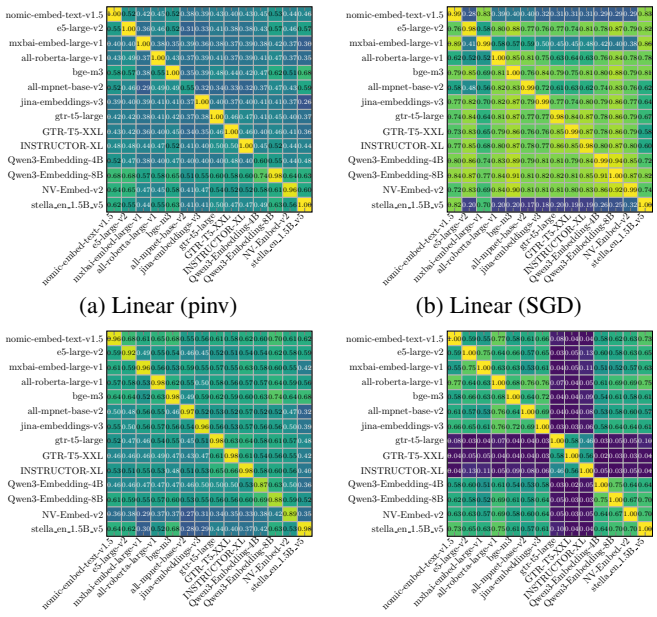
The main-paper top-1 matrices (Figures 2 and 3) and the top-5 and top-10 matrices below (Figures 21–22 and 26–27) show where these gains occur. Increasing k expands the compatible blocks instead of brightening all pairs uniformly. Deeper retrieval therefore does not erase the source–target structure already visible at top-1.

Directionality. Native Linear CKA is symmetric, but post-translation CKA need not be because each ordered direction has its own fitted map. For word-anchor text, the mean absolute difference between opposite directions is 0.174 for Linear (SGD), 0.079 for IRP, 0.067 for pinv, and zero for RR. On image, it is 0.042 for Linear (SGD) and 0.333 for pinv. We therefore use the symmetric quantity in Equation 4 only for summary visualizations such as clustering, while retaining directional measurements for the analysis.

E.1 TEXT MATRICES

Figures 20–24 give the five word-anchor text diagnostics, with one panel per translator and a common source-row, target-column layout. The post-translation CKA and retrieval matrices reproduce the family structure from Section 5.1: same-family blocks remain bright, while many cross-family pairs remain dim. RR illustrates why fidelity must be read separately from identity. Its cosine fidelity is

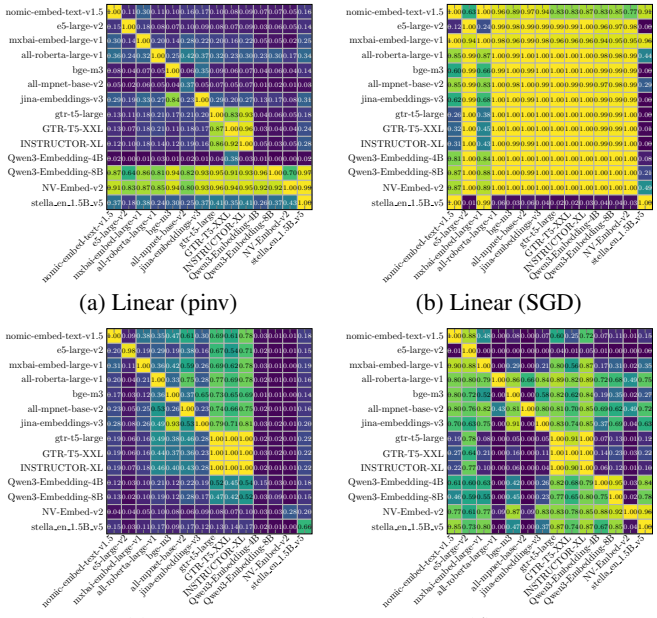
uniformly high, averaging 0.820 over cross-model pairs, even where retrieval falls to 0.329 at top-1 and 0.484 at top-10. Smooth agreement between relative profiles does not ensure preservation of item identity.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

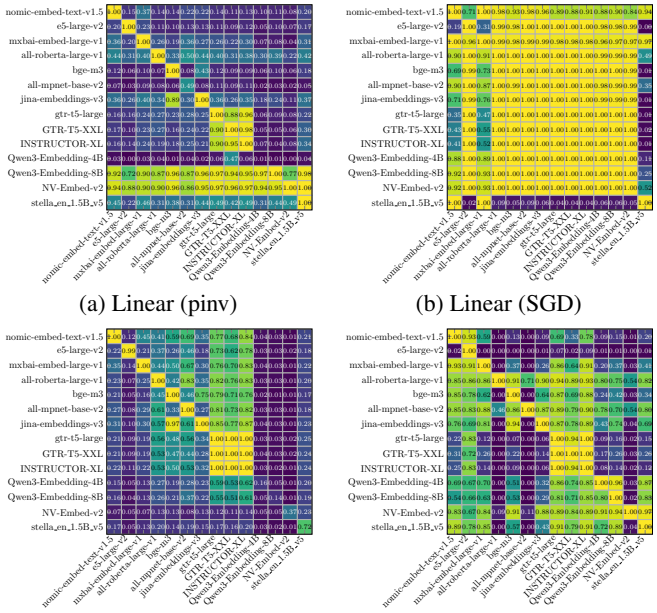
Figure 20: Text, word anchors: centered Linear CKA after translation. Rows are source models and columns are targets; panels are translators. Higher values mean the translated source reproduces the target's pairwise geometry more closely.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

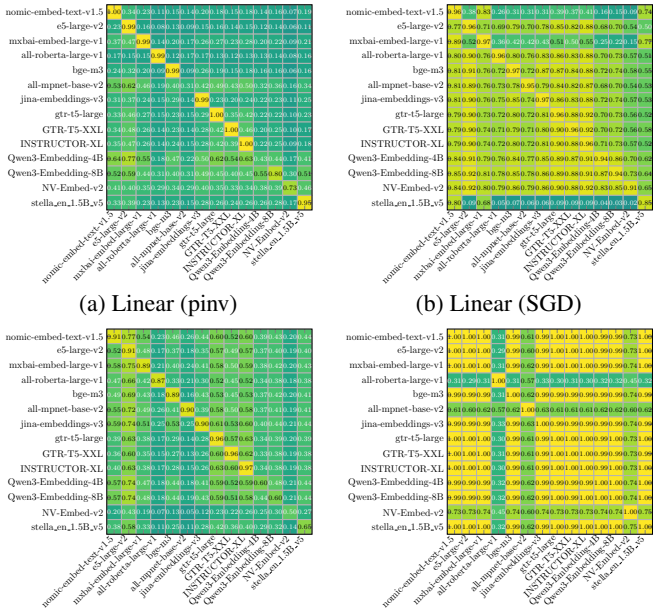
Figure 21: Text, word anchors: k-NN top-5 paired-item retrieval. Rows are source models and columns are targets; panels are translators. Higher values mean the correct paired target appears in the first five candidates.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

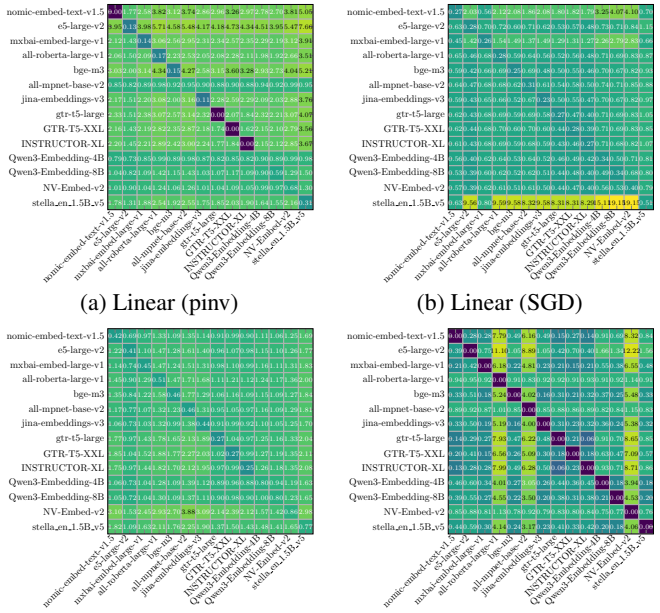
Figure 22: Text, word anchors: k-NN top-10 paired-item retrieval. Rows are source models and columns are targets; panels are translators.



(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

Figure 23: Text, word anchors: mean cosine fidelity between translated and native target embeddings. Rows are source models and columns are targets; panels are translators. Higher values mean closer pointwise agreement.



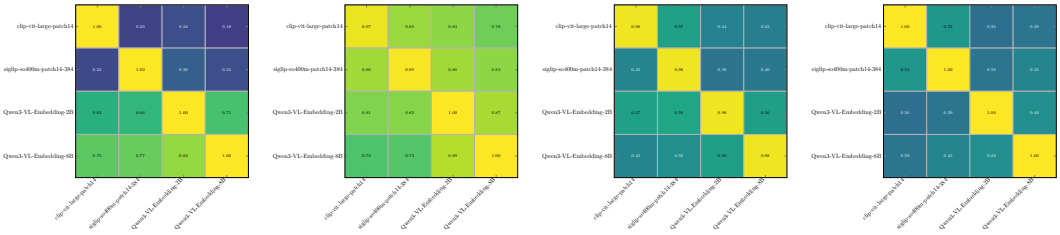
(a) Linear (pinv) (b) Linear (SGD)

(c) IRP (d) RR

Figure 24: Text, word anchors: normalized pointwise error between translated and native target embeddings. Rows are source models and columns are targets; panels are translators. Lower values are better.

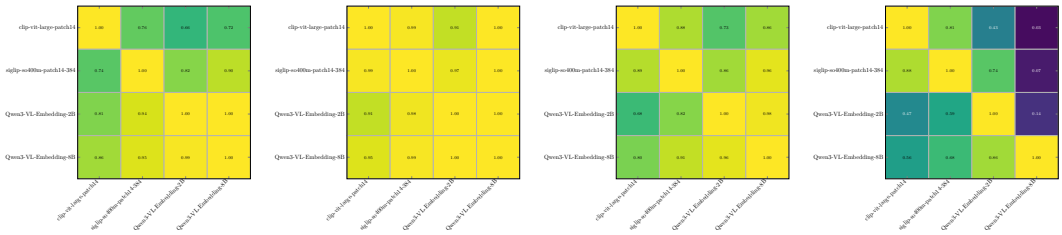
E.2 IMAGE MATRICES

Figures 25–29 repeat the five diagnostics with ImageNet anchors. Because the image suite contains only four models, each pair can be inspected directly. The brightest blocks occur for `clip-vit-large-patch14/siglip-so400m-patch14-384` and for the two `Qwen3-VL` sizes. As in text, fidelity and identity diverge: RR has uniformly high cosine fidelity (mean 0.970) but much lower retrieval (0.342 at top-1 and 0.590 at top-10). Linear (SGD) has the lowest mean normalized error (0.615) and the highest top-1 retrieval (0.878), illustrating that these two diagnostics can agree for one translator without becoming interchangeable in general.



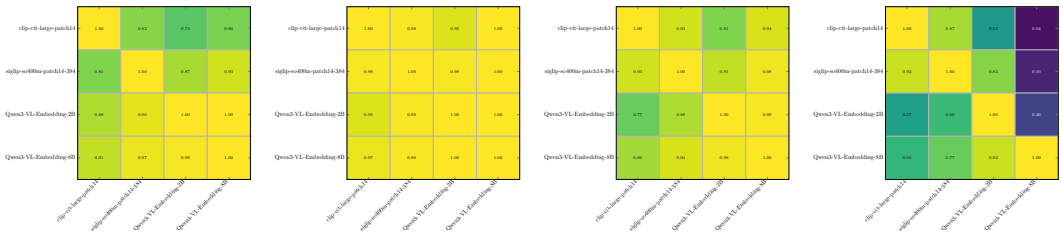
(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 25: Image, ImageNet anchors: centered Linear CKA after translation. Rows are source models and columns are targets; panels are translators. Higher values mean the translated source reproduces the target's pairwise geometry more closely.



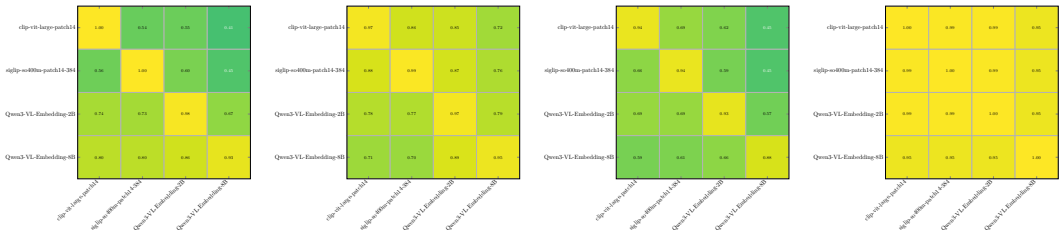
(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 26: Image, ImageNet anchors: k-NN top-5 paired-item retrieval. Rows are source models and columns are targets; panels are translators. Higher values mean the correct paired target appears in the first five candidates.



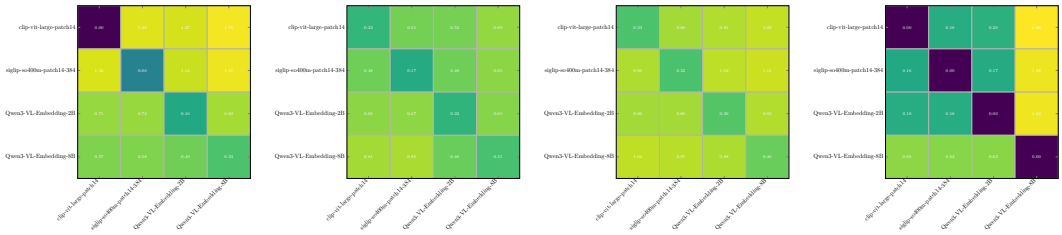
(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 27: Image, ImageNet anchors: k-NN top-10 paired-item retrieval. Rows are source models and columns are targets; panels are translators.



(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 28: Image, ImageNet anchors: mean cosine fidelity between translated and native target embeddings. Rows are source models and columns are targets; panels are translators. Higher values mean closer pointwise agreement.



(a) Linear (pinv) (b) Linear (SGD) (c) IRP (d) RR

Figure 29: Image, ImageNet anchors: normalized pointwise error between translated and native target embeddings. Rows are source models and columns are targets; panels are translators. Lower values are better.

F NOTATION AND ABBREVIATIONS

Table 18: Abbreviations, notation, and table conventions used in the paper and its appendix.

Symbol	Meaning
Abbreviations	
CKA	Centered Kernel Alignment; *Linear CKA* is its centered linear-kernel form (Section 3.3).
RR	Relative Representations: a vector is replaced by its cosine similarities to the shared anchors.
IRP	Inverse Relative Projection: maps relative coordinates back into a target representation space.
SGD	Stochastic gradient descent; *Linear (SGD)* is the high-supervision linear estimator (Section 3).
pinv	Moore–Penrose pseudo-inverse; *Linear (pinv)* is the closed-form anchor-based linear estimator (Section 3).
MLP	Multilayer perceptron.
top-k	Paired-item retrieval accuracy at rank k among all target items.
macro-F1	F1 score averaged over classes.
`cls`, `finetune`	Internal names of the image zero-shot and frozen-embedding protocols (Section 4).
KV	Key–value cache.
Notation	
A, B	Source and target models.
f_A, f_B	Frozen encoders of models A and B.
x, y	Native source and target embeddings of a paired input z.
$\hat{y}$	Translated source embedding, $\hat{y} = T_{A\to B}(x)$.
$T_{A\to B}$	Directional translator from A to B; $T_{i\to i}$ is the same-model translation.
X, Y	Matrices whose columns are the native source and target embeddings of the n paired evaluation items.
$\hat{Y}$	Matrix of translated source embeddings.
$K = X^\top X$, $L = Y^\top Y$	Centered Gram matrices compared by Linear CKA.
$\langle K, L\rangle_F$, $\|K\|_F$	Frobenius inner product and Frobenius norm.
X^+	Moore–Penrose pseudo-inverse of the anchor matrix X.
$c^*(x) = X^+x$	Source anchor coefficients of x.
m	Number of shared anchors.
d_A, d_B	Embedding dimensions of models A and B.
ρ, τ	Spearman and Kendall rank correlations.
$s_{A\to B}$	Transfer accuracy averaged over tasks for direction $A \to B$.
$\mathrm{Compat}(A, B)$	Symmetric pair compatibility, $\frac{1}{2}(s_{A\to B} + s_{B\to A})$.
Table conventions	
all / self	In accuracy cells, the value before the slash averages all ordered pairs and the value after it averages the diagonal $T_{i\to i}$.
all, cross, self	Rows of Table 3: *all* averages every ordered pair, *cross* excludes the diagonal, and *self* averages the diagonal.
gap	*self* minus *cross*.
heat-map cell	Row is a source model, column is a target model, and panel is a translator; higher values mean the translated source works better with the head trained on the target.
direction	Linear CKA on native representations is symmetric, while CKA after translation is directional because each ordered pair has its own fitted map.
higher / lower is better	Linear CKA, retrieval, cosine, accuracy, and macro-F1 are higher-is-better; normalized error is lower-is-better.