Beyond Pairwise Overlap: Shared-Value Observations for Compositional Text Embeddings

Anonymous Authors

Anonymous Institution

Abstract. A city directory can change through a few swaps of the cities assigned to people. Observing every combination quickly becomes expensive. Which joint edits best reconstruct the embeddings of unseen records? We study two triples that form the same overlap triangle: three swaps can meet at one shared city, or each pair can meet at a different city. Two observation designs each encode 23 states and feed the same quadratic reconstruction. Across 2,048 test worlds, shared-value triples reduce normalized reconstruction error in both BGE and E5 under two renderings. On atomic records, the reductions are 53.4% and 36.0%; on compressed records, target identity retrieval improves by 15.5–15.8 percentage points. Geometry-based calibration selects the shared-value set in all 24 cells of the triangle family. Qwen3-Embedding-0.6B retains a 6.58% error reduction on 512 preselected atomic worlds. Fact ranking reveals a different preference: same-entity queries improve in the two-encoder atomic study, while different-entity queries decline, and query-selected alternatives lead on several settings. These results identify shared-value incidence as a useful cue for choosing reconstruction observations and show how the downstream query objective changes that choice.

Keywords: Text embeddings · Compositional edits · Observation selection · Dense retrieval

1 Introduction

A city directory records where each person lives, works, and studies. Correcting a field can swap the cities assigned to two people. Six such edits produce 64 versions of the directory, each with its own text embedding. Encoding a small set of versions and reconstructing the rest turns this simple update into an observation-selection problem: which combinations of edits are most informative?

Consider three swaps. The first exchanges Dublin and Accra for residence, the second exchanges Dublin and Seoul for work, and the third exchanges Dublin and Athens for study. Dublin connects all three. Now replace the study swap with one between Accra and Seoul. Every pair still overlaps, but the three overlaps refer to Dublin, Accra, and Seoul. The two triples have the same triangle-shaped overlap graph. Retaining the city on each edge exposes the difference (Fig. 1).

We ask whether this shared-value structure helps choose the states to observe. Our comparison gives both designs the same origin, six single-edit states, and

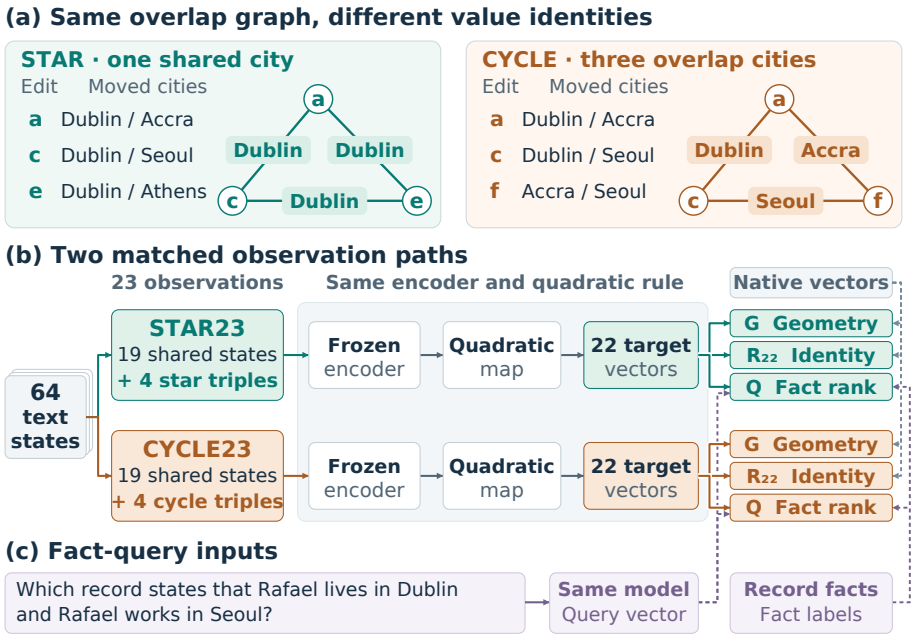


Fig. 1. Value identities distinguish joint observations with the same overlap graph. (a) The star triple *ace* shares Dublin on all three edges; the cycle triple *acf* connects through Dublin, Accra, and Seoul. (b) Each arm adds four triples to 19 shared states, encodes its 23 observations, and reconstructs 22 targets per origin with the same quadratic rule. Native target vectors enter the geometry and identity readouts. (c) An encoded fact query ranks the reconstructed documents, with relevance determined by the record facts. Dashed arrows supply evaluation inputs; the two arms retain separate predictions and scores.

twelve overlapping pair states. One design adds four triples linked by a common city; the other adds four triples linked cyclically by different cities. Both therefore observe 23 states. Exchanging edit indices and rotating relation roles lets us follow the value structure through several equivalent presentations of the same pair graph.

The selected observations feed a common quadratic estimator. Each triple supplies an interaction residual whose degree-two projection adjusts the intercept, single-edit coefficients, and pair coefficients. This gives a direct comparison of observation choices at a fixed representation degree. It also leads to three concrete questions. How closely does the reconstructed vector match its native embedding? Can it retrieve the correct state? And can a textual query retrieve records satisfying a conjunction of facts?

The answers separate observation structure from evaluation objective. On 2,048 test worlds, shared-value triples improve geometric reconstruction for BGE and E5 under both atomic and compressed renderings. Atomic error falls by

53.4% and 36.0%, respectively. A matched comparison with Qwen3-Embedding-0.6B on 512 preselected atomic worlds yields a 6.58% reduction. Fact queries reveal finer differences: gains can concentrate on one entity group, and broader query-calibrated observation sets can achieve higher ranking scores.

The study makes two contributions. First, it introduces a controlled design for observation selection that preserves pairwise overlap while exchanging shared-value incidence. The matched results identify that incidence as a useful reconstruction cue across the tested encoders and renderings. Second, it connects observation selection to three distinct outcomes—geometry, state identity, and conjunctive fact ranking. Entity-group analyses and objective-specific calibration explain when these outcomes favor different sets of joint observations.

2 Related Work

Compositionality through reconstruction. Reconstructing unseen combinations offers a direct way to study how embeddings compose. Guo et al. [3] reconstruct embeddings for unseen attribute combinations and measure distance, cosine similarity, and retrieval. Uselis et al. [8] study concept binding in vision–language embeddings using edits involving shared and distinct concepts, relating generalization to binding-function complexity in controlled models. Our study focuses on the observations supplied to a text-embedding reconstruction. Holding the pair graph, observation count, and estimator fixed exposes the information carried by the value identities on overlapping edits.

Interactions and low-degree representations. Anchored and ANOVA decompositions express a multivariate function through components indexed by subsets of its variables [6]. We use anchored differences to extract local edit interactions, then project selected triple terms into a common degree-two representation. This algebra makes observation choice explicit: different joint states contribute different residuals to the same set of quadratic coefficients.

Approximation for retrieval. The usefulness of a vector approximation depends on the task it serves. Anisotropic quantization weights approximation error according to its effect on inner-product search [2]. We examine this objective dependence at the observation-selection stage. Relevance comes directly from the facts in each record, and nDCG evaluates the resulting rankings [5]. Comparing geometry-selected and query-selected observations reveals how a retrieval objective changes which composed records are worth encoding.

3 Shared-Value Observation Design

The Dublin example suggests a simple principle: retain the identity of an overlapping value when choosing joint edits to observe. We turn this intuition into a matched experiment. First we construct edit families with the same pair graph. We then select four triples by their value structure and reconstruct unseen states with a common quadratic map.

3.1 From records to an overlap graph

A world assigns four names to entity slots and four cities to a local vocabulary. Each of three relations—lives in, works in, and studies in—assigns the four cities bijectively to the four entities. Within each relation, two binary edits swap the values at slots $(0, 1)$ and $(2, 3)$. We denote the residence edits by (a, b), work edits by (c, d), and study edits by (e, f). All six edits commute, giving a Boolean cube of 64 record states.

Let V_i contain the two cities moved by edit i. The three relations use the three distinct perfect matchings of the four cities. Edits within a relation therefore have disjoint value sets, and edits from different relations share exactly one city. The resulting graph is $K_{2,2,2}$, with edge set E_{12} and eight triangles. Each triangle contains one edit from each relation.

Four triangles are *stars*: a single city lies in all three value sets. The other four are *cycles*: their three pairwise intersections contain distinct cities. For the example in Fig. 1, $V_a = \{\text{Dublin, Accra}\}$, $V_c = \{\text{Dublin, Seoul}\}$, and $V_e = \{\text{Dublin, Athens}\}$ form a star. Substituting $V_f = \{\text{Accra, Seoul}\}$ gives a cycle. Swapping the positions of the two study blocks exchanges the roles of ace and acf while preserving E_{12}.

We apply this even–odd exchange to three cyclic assignments of the matchings to relation roles, yielding six conditions. Four cyclic rotations of entity order provide the textual views of each condition. The construction thus varies the indices and surface positions associated with a shared city while keeping the pairwise overlap pattern fixed.

3.2 Two sets of 23 observations

Both arms start with 19 states around an origin: the origin itself, the six single edits, and the twelve pairs in E_{12}. Star23 adds the four star triples, and Cycle23 adds the four cycle triples. Each selected triple brings its complete set of lower-order faces through the shared states. Each arm therefore observes 23 states and supplies all the quantities needed by the reconstruction below.

Let $h(s) \in \mathbb{R}^d$ be the raw pooled embedding of state s for a fixed world, encoder, rendering, condition, and view. Applying a relative edit mask b at origin o gives state $o \oplus b$. We use 16 balanced origins and reconstruct the 22 relative masks with Hamming weight at least four. Observed states have relative weight zero through three. Appendix A lists the origins and construction details.

3.3 A common quadratic reconstruction

An edit's effect is the change from the origin embedding. A pair residual captures the additional change after accounting for its two individual effects. The same subtraction extends to a triple. For an edit set U with mask $m(U)$, the anchored difference is

$$\mu_U(o) = \sum_{V \subseteq U} (-1)^{|U|-|V|} h(o \oplus m(V)). \tag{1}$$

The selected triples contribute their residuals through a shared degree-two map:

$$q_{E,T}(o,b) = \mu_{\varnothing}(o) + \sum_i b_i \mu_i(o) + \sum_{\{i,j\}\in E} b_i b_j \mu_{ij}(o) + \sum_{t\in T} \phi_t(b)\mu_t(o), \tag{2}$$

$$\phi_t(b) = \frac{1}{8} - \frac{1}{4}\sum_{i\in t} b_i + \frac{1}{2}\sum_{\{i,j\}\subseteq t} b_i b_j. \tag{3}$$

Here $E = E_{12}$ for both arms, and T is their respective set of four triples. To obtain ϕ_t, write each bit as $b_i = (1-\chi_i)/2$, where $\chi_i = (-1)^{b_i}$, expand the triple monomial, and retain Walsh terms of degree at most two. A selected triple consequently adjusts the constant, linear, and pairwise coefficients of the reconstruction. The observation structure changes the information available to these coefficients; the output degree and reconstruction rule remain shared.

4 Evaluation

4.1 Worlds, text, and embedding models

The main study uses 256 calibration worlds and 2,048 test worlds. Each world selects four names and four cities and assigns them to slots. Calibration draws from pools of eight names and eight cities; test pools contain twelve of each and use disjoint strings. Test queries also use a different wording. Every world contributes all six conditions, four views, and 16 origins for each method and encoder.

ATOMIC renders the twelve facts as separate sentences grouped by relation. Swaps preserve its word multiset. COMPRESSED groups facts by entity and merges relations that share a city. For example, *Rafael lives, works, and studies in Dublin* expresses three facts in one clause. Both renderings express the same underlying states and relevance labels.

We use BGE-small-en-v1.5 [10,1] and E5-small-v2 [9,4]. Both produce 384-dimensional vectors. BGE pools the CLS token and adds a retrieval instruction to queries. E5 uses masked mean pooling with its passage and query prefixes. Raw pooled outputs are computed in FP32; reconstruction and scoring use FP64. Cosine scores use L2-normalized vectors.

A decoder-based transfer comparison uses Qwen3-Embedding-0.6B [11,7] on 512 preselected worlds from the same test pool. It retains atomic text, all six conditions, and the matched STAR23 and CYCLE23 designs. Last-valid-token pooling gives 1024 dimensions. Documents use the record text, and queries use the checkpoint's retrieval instruction. Appendix B specifies the frozen inference configuration.

4.2 Three uses of the reconstructed embeddings

Geometric reconstruction. Normalized mean squared error (NMSE, denoted G) measures the squared distance from a reconstructed target vector to its native

embedding, relative to the world's within-cube variation. Scores average over target states, origins, views, and conditions. Lower NMSE means closer reconstruction; the exact expression appears in Appendix A.

State identity retrieval. Each reconstructed vector queries the corresponding gallery of 22 native target vectors using cosine similarity. R22 is top-1 accuracy for recovering the same state identity. This readout tests whether the approximation preserves enough detail to distinguish a complete record from its other edited versions.

Conjunctive fact ranking. Text queries retrieve from the 22 reconstructed target documents. A same-entity query can ask for *Rafael lives in Dublin and Rafael works in Seoul*; a different-entity query can ask for *Rafael lives in Dublin and Nora works in Seoul*. Every two-fact query spans two relations. A state is relevant precisely when both requested facts hold. Each world and condition has 48 same-entity and 144 different-entity queries, together with 24 single-fact queries for a separate readout.

We rank by query–document cosine similarity, with absolute state ID breaking exact ties. Binary nDCG@5 scores each query. The balanced score Q gives equal weight to the two entity groups, whose individual means reveal their respective contributions. Balanced AUC measures relevant–irrelevant ordering accuracy, assigning half credit to ties. Uniform random rankings yield $Q = 27.214\%$ and AUC $= 50\%$ in this gallery.

4.3 Observation families and calibration

Two pair-only designs establish the value of joint observations. StructPair19 observes the origin, singles, and twelve overlap pairs, for 19 states. FullPair22 includes all fifteen pairs, for 22 states. The full-information references characterize the embedding space: Native uses original target vectors, while Quad64 uses the uniform degree-two projection of the entire 64-state cube.

Five calibrated families broaden the observation choices. Tri70 selects four of eight overlap triangles. Part21 uses two three-edit blocks and six cross-block pairs; Part23 uses the same block structure and all but one cross-block pair. Pair21 selects fourteen pairs, and PairPlus23 selects all pairs and one triple. Each family uses the map in Eq. (2). Their names indicate the candidate count for Tri70 and the state budget for the other families.

Calibration selects a minimum-G candidate and a maximum-Q candidate for each encoder, rendering, and condition, giving 24 cells. Part21 searches all 840 candidates for geometry and a shortlist of 80 for queries: the eight geometric leaders in each of ten partitions. The other menus contain 70, 90, 15, and 20 candidates, respectively. We then score every selected candidate on test worlds, using geometry winners for reconstruction comparisons and query winners for ranking comparisons.

4.4 Paired comparisons

All measurements are aggregated within worlds before comparing methods. Positive geometric effects use comparator NMSE minus STAR23 NMSE; positive ranking effects use STAR23 Q minus comparator Q. The main study resamples the 2,048 paired worlds 10,000 times. Its eight STAR23–CYCLE23 contrasts receive simultaneous 95% intervals, as do the separate family of 48 strong-control contrasts. Entity-group and full-cube diagnostics use paired percentile intervals.

For Qwen, the primary endpoint is the relative reduction in mean NMSE, $\rho = 1 - \overline{G}_{\text{Star23}}/\overline{G}_{\text{Cycle23}}$. Its 95% percentile interval uses 10,000 paired-world bootstrap samples. Ranking and condition-wise intervals describe the accompanying outcomes. Appendix A details the resampling rule.

5 Results

5.1 Shared values guide geometric reconstruction

The value carried by an overlap changes which triples best support reconstruction. On atomic text, STAR23 lowers BGE's NMSE from 0.02657 to 0.01238 and E5's from 0.02367 to 0.01514 (Fig. 2), giving relative reductions of 53.4% and 36.0%. The advantage also appears with compressed clauses: the paired NMSE reductions are 0.40545 for BGE and 0.39781 for E5, with simultaneous 95% intervals $[0.40290, 0.40801]$ and $[0.39592, 0.39971]$.

The preference follows the shared city through the index exchange. In atomic records, BGE's NMSE reduction is approximately 0.0142 for each parity group, while E5's ranges from 0.0085 to 0.0086. All three relation rotations preserve the direction. State identity retrieval improves alongside geometry (Table 1). On compressed text, BGE's R22 rises from 78.10% to 93.55%, and E5's rises from 77.83% to 93.63%. The selected triples thus help retain the identity of the complete edited record.

5.2 Fact queries reveal an entity-dependent preference

Fact ranking asks a more targeted question: which states satisfy the requested conjunction? On atomic records, the balanced nDCG@5 gain is 0.04427 percentage points for BGE and 0.01064 for E5. Their simultaneous 95% intervals are $[0.04189, 0.04665]$ and $[0.00867, 0.01262]$. Splitting queries by entity shows where these small gains arise (Fig. 3). BGE gains 0.12424 points for same-entity queries and loses 0.03569 for different-entity queries. E5 gains 0.03309 and loses 0.01181 points in the same groups.

Clause compression brings the distinction into the balanced score. BGE improves in both groups, yielding a 0.03527-point balanced gain. E5 improves by 0.12377 points on same-entity queries, while its different-entity score falls by 0.19461 points. Its balanced score consequently falls by 0.03542 points, with simultaneous interval $[-0.04166, -0.02917]$. In this setting, the same observations improve complete-state identity retrieval by 15.80 points while favoring a different trade-off for conjunctive fact queries.

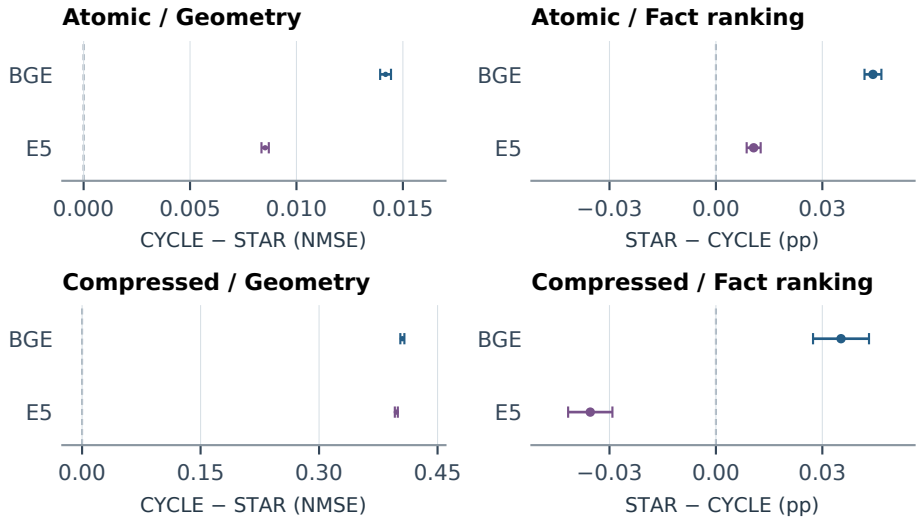


Fig. 2. Shared-value observations improve reconstruction in all four encoder–rendering settings. Points show paired effects on NMSE and balanced fact ranking; positive values favor Star23. Bars are simultaneous 95% intervals for the eight contrasts across 2,048 test worlds.

Table 1. Star23 improves geometry and state identity across all four settings, while balanced fact ranking varies. Each cell reports Star23 / Cycle23 means over 2,048 worlds. R22 and Q are percentages.

Encoder	Text	NMSE	R22 (%)	Balanced Q (%)
BGE	Atomic	0.01238 / 0.02657	99.02 / 96.00	29.019 / 28.975
BGE	Compressed	0.31746 / 0.72291	93.55 / 78.10	28.621 / 28.585
E5	Atomic	0.01514 / 0.02367	98.15 / 96.80	28.592 / 28.581
E5	Compressed	0.29736 / 0.69517	93.63 / 77.83	29.277 / 29.312

5.3 Calibration adapts observations to the objective

A direct search over the triangle family recovers the geometric preference. Among its 70 four-triangle subsets, geometry calibration chooses the shared-value set in all 24 encoder–rendering–condition cells. Query calibration chooses it in three. The geometric winner thus reproduces Star23 on test worlds. The query objective often chooses another mixture of triangles.

Broader menus let calibration allocate observations across pairs and triples (Table 2). Star23 has lower NMSE on compressed text and in most atomic comparisons. For E5 atomic text, geometry-selected Part23 and PairPlus23 achieve lower NMSE at the same 23-state budget. Shared-value incidence supplies an effective rule within the matched triangle family, and richer menus give calibration further ways to use the observation budget.

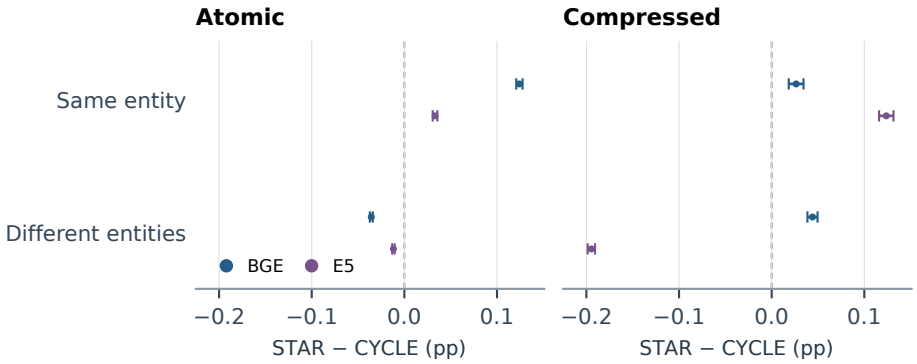


Fig. 3. Same-entity queries favor shared-value observations across the four settings; different-entity queries depend on encoder and rendering. Points show Star23 minus Cycle23 nDCG@5 in percentage points, with paired percentile 95% intervals across test worlds.

The query-selected ordering changes more substantially (Table 3). On compressed text, Part23 exceeds Star23 by 0.55355 percentage points for BGE and 0.49502 for E5. For BGE, the simultaneous interval for the Star23 difference is $[-0.56626, -0.54085]$; for E5, it is $[-0.50451, -0.48552]$. On E5 atomic text, query-selected Part21 and Part23 also lead. Optimizing the ranking objective can therefore select observations with stronger fact retrieval even when their geometric approximation is worse.

5.4 Full-cube projection strengthens the fact-ranking signal

Native embeddings offer a modest fact-ranking signal in these shared-vocabulary galleries. Balanced nDCG@5 ranges from 28.61% to 29.27%, compared with 27.21% for uniform random ranking; balanced AUC ranges from 50.26% to 51.41% (Table 4). These absolute scores place the observation-selection effects on the scale of the underlying retrieval task.

The full-cube degree-two projection improves balanced nDCG@5 in all four settings. E5 on compressed text has the largest gain, 0.22076 percentage points. The result complements the calibrated comparisons: retaining a particular low-degree representation can improve conjunctive ranking even as it changes the native vectors. Together, the two analyses motivate evaluating observation choice against the intended query objective.

Table 2. Shared-value observations lead most geometry comparisons; E5 atomic text favors two broader 23-state designs. Entries are comparator minus STAR23 NMSE, so positive values favor STAR23. Brackets give simultaneous 95% intervals for the strong-control family. Calibrated comparators are selected for geometry.

	BGE		E5	
Comparator	Atomic	Compressed	Atomic	Compressed
TRI70	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]	+0.0000 [0.0000, 0.0000]
PART21	+0.0080 [0.0076, 0.0083]	+0.3443 [0.3414, 0.3472]	+0.0021 [0.0017, 0.0024]	+0.3160 [0.3135, 0.3186]
PART23	+0.0068 [0.0064, 0.0071]	+0.7627 [0.7563, 0.7691]	−0.0011 [−0.0015, −0.0008]	+0.6817 [0.6762, 0.6873]
PAIR21	+0.0124 [0.0120, 0.0128]	+1.4436 [1.4323, 1.4549]	+0.0024 [0.0020, 0.0027]	+1.3043 [1.2955, 1.3131]
PAIRPLUS23	+0.0067 [0.0063, 0.0070]	+1.0731 [1.0641, 1.0821]	−0.0012 [−0.0015, −0.0008]	+0.9707 [0.9631, 0.9783]
FULLPAIR22	+0.0119 [0.0115, 0.0123]	+1.7497 [1.7350, 1.7644]	+0.0015 [0.0011, 0.0019]	+1.5715 [1.5610, 1.5821]

Table 3. Query-selected observations change the ranking order, with PART23 leading on compressed text. Entries are STAR23 minus comparator balanced nDCG@5 in percentage points. Brackets give simultaneous 95% intervals for the strong-control family. Calibrated comparators are selected for Q.

	BGE		E5	
Comparator	Atomic	Compressed	Atomic	Compressed
TRI70	+0.0030 [0.0021, 0.0040]	−0.0558 [−0.0597, −0.0519]	+0.0014 [−0.0001, 0.0028]	−0.0732 [−0.0778, −0.0686]
PART21	+0.0317 [0.0293, 0.0341]	−0.0617 [−0.0665, −0.0569]	−0.0124 [−0.0160, −0.0089]	−0.1198 [−0.1253, −0.1144]
PART23	+0.0250 [0.0208, 0.0291]	−0.5536 [−0.5663, −0.5408]	−0.0051 [−0.0087, −0.0015]	−0.4950 [−0.5045, −0.4855]
PAIR21	+0.0421 [0.0380, 0.0462]	−0.5294 [−0.5417, −0.5172]	−0.0017 [−0.0054, 0.0019]	−0.3903 [−0.3996, −0.3810]
PAIRPLUS23	+0.0376 [0.0345, 0.0408]	−0.0847 [−0.0906, −0.0789]	−0.0014 [−0.0048, 0.0021]	+0.0222 [0.0163, 0.0282]
FULLPAIR22	+0.0522 [0.0488, 0.0556]	−0.0212 [−0.0308, −0.0117]	+0.0034 [−0.0002, 0.0070]	+0.1010 [0.0937, 0.1082]

Table 4. Full-cube quadratic projection improves balanced fact ranking in every setting. Native columns give absolute scores; the final column gives the change in Q from projecting all 64 states. Random ranking yields $Q = 27.214\%$ and AUC = 50%. Brackets give paired percentile 95% intervals.

Encoder	Text	Native Q (%)	AUC (%)	QUAD64 − native (pp)
BGE	Atomic	29.026	50.257	+0.00599 [0.00468, 0.00730]
BGE	Compressed	28.643	51.263	+0.05390 [0.04984, 0.05805]
E5	Atomic	28.606	50.365	+0.00196 [0.00084, 0.00309]
E5	Compressed	29.271	51.406	+0.22076 [0.21657, 0.22495]

5.5 The geometric preference extends to Qwen embeddings

The matched atomic comparison retains its direction with Qwen3-Embedding-0.6B. On 512 preselected test worlds, mean NMSE falls from 0.09871 with CYCLE23 to 0.09222 with STAR23 (Table 5). The relative reduction is 6.58%, with paired 95% interval [6.42%, 6.75%]. Each of the six conditions favors the shared-value set, with relative reductions from 5.35% to 7.68%.

The ranking outcomes have a smaller scale. R22 improves by 0.2501 percentage points, with interval [0.2248, 0.2754]. Balanced fact nDCG@5 improves by 0.0126 percentage points, with interval [0.0097, 0.0155], and both entity groups gain. The original Qwen target vectors achieve 29.4668% balanced nDCG@5 and 50.2004% balanced AUC. Across the two encoder families, shared-value observations thus retain a geometric advantage while the magnitude and ranking behavior vary with the embedding endpoint.

Table 5. Shared-value triples lower Qwen's reconstruction error on 512 atomic-text worlds. Rows give world-level means. R22, balanced Q, and balanced AUC are percentages. Native scores the original target embeddings.

Representation	NMSE ↓	R22 ↑	Q ↑	AUC ↑
CYCLE23	0.09871	95.7897	29.2780	50.2214
STAR23	0.09222	96.0398	29.2906	50.2231
NATIVE	0.00000	100.0000	29.4668	50.2004

6 Discussion

Value identities enrich an overlap graph. An overlap edge identifies two edits that touch a common value. Keeping that value's identity reveals whether several edges refer to the same city. Our matched comparison shows that this extra information guides observation choice: four shared-value triples consistently improve reconstruction over their cyclic counterparts across the tested settings. The preference follows value incidence when edit indices and relation roles change. For structured records, a value-labeled overlap graph is therefore a useful starting point for selecting composed states to encode.

Joint observations shape low-degree coefficients. The reconstruction illustrates how higher-order observations can support a lower-degree representation. A triple residual contributes to the constant, linear, and pair coefficients through Eq. (3). Choosing its state changes the evidence available for those coefficients. Observation selection and representation degree are thus two complementary design decisions, and the calibrated families explore several allocations of the same small budget.

Query margins connect geometry to ranking. A state-identity query targets one complete record, while a fact query targets every record satisfying a conjunction. Their ranking margins can differ. For a unit query u and normalized document vectors $z_i, \hat{z}_i$, define the score error $e_i = |u^\mathsf{T}(\hat{z}_i - z_i)| \le \|\hat{z}_i - z_i\|_2$. The native ordering of i and j is preserved when its positive margin $u^\mathsf{T}(z_i - z_j)$ exceeds $e_i + e_j$. This relation connects approximation to the particular query and candidate pair. The entity-group results and query-calibrated controls show why observation policies benefit from measuring the rankings they are intended to serve.

7 Conclusion

Shared-value incidence helps choose joint observations for compositional text-embedding reconstruction. With the pair graph, state budget, and quadratic rule held fixed, shared-value triples improve geometry across two encoders and two renderings, and retain their advantage in a matched Qwen atomic comparison. State identity, conjunctive ranking, and calibration expose different uses of the resulting vectors. The practical design principle is to retain value identity when proposing observations and to select among them using the downstream objective.

8 Limitations

Study scope. The main study uses synthetic four-entity, four-city records, two fixed small encoders, and two renderings. The Qwen comparison uses one checkpoint and a preselected 512-world subset of the same test pool, restricted to atomic text. The three endpoints differ jointly in architecture, scale, training, pooling, and query instructions. Isolating the contribution of any one factor requires additional controlled comparisons. Natural documents, larger edit spaces, independent world pools, and checkpoint variation remain important extensions.
Rendering and attribution. Atomic edits preserve word counts. Compression changes length, word frequency, and clause structure. Each rendering's NMSE is scaled by its own native variation. The index exchange and relation rotations establish a value-incidence preference within this design, while multiple representational and surface effects may contribute to it. Same-entity queries also repeat a name more often than different-entity queries. These comparisons leave the encoder's internal binding mechanism unresolved. Other reconstruction families may use the observations differently.
Evaluation and deployment. Calibration and test change lexical strings and query phrasing together. Intervals describe sampled worlds conditional on the fixed checkpoints and the selected calibration candidates. The native AUC values indicate weak fact-query separation, so the ranking effects should be examined in galleries with stronger native signal. QUAD64 has access to all 64 native states, including the evaluated targets; it characterizes a full-information projection.

The 23-state budget measures observations per origin. End-to-end update latency, shared computation across origins, and deployment cost require separate measurement.

AI Usage. ChatGPT (OpenAI) was used strictly for language copy-editing and LaTeX formatting of author-written text and figures.

A Design and Measurement Details

A cube with a fixed overlap graph. For city IDs $0, 1, 2, 3$, the even base uses the permutations $(0, 1, 2, 3)$, $(0, 2, 1, 3)$, and $(0, 3, 1, 2)$ for residence, work, and study. We rotate these assignments cyclically for the other two relation orders. In each odd variant, the two blocks of the study permutation exchange slots. The even shared-value triangles are ace, adf, bcf, bde. Their complements among the eight graph triangles, acf, ade, bce, bdf, carry the shared values in the odd variant.

Origins and targets. The origin masks, with a the most significant bit, are

```
000000 010101 001100 011001
000011 010110 001111 011010
111111 101010 110011 100110
111100 101001 110000 100101
```

The patterns on any one, two, or three bits are balanced across these origins. The gallery at o is $\{o \oplus b : \|b\|_0 \geq 4\}$. An exact match for two facts fixes two bits. Depending on their values relative to the origin, the remaining four bits yield 1, 5, or 11 positive target states. Their frequencies are $1/4, 1/2, 1/4$ in each entity group. For r positives the random-ranking expected DCG@5 is $(r/22)\sum_{k=1}^{5} 1/\log_2(k+1)$, which we divide by the ideal DCG for that query.

Calibration search. The geometry menus have 70, 840, 90, 15, and 20 candidates for Tri70, Part21, Part23, Pair21, and PairPlus23. The query menus have 70, 80, 90, 15, and 20. With the fixed pair-only designs, geometry evaluates 1,037 choices per cell. Adding the fixed pair-only and full-information references gives 279 scored choices for queries. Geometry ties use the candidate identifier after rounding the objective to 12 decimal places. Query ties compare geometry and then use the candidate identifier. The rounded keys serve selection, and the reported endpoints retain their stored precision.

Geometry normalization. For a fixed encoder, world, and renderer, let h_{cvs} be the native vector for condition c, view v, and state s, and let $\bar{h}_{cv} = 64^{-1}\sum_s h_{cvs}$. Define

$$D = \mathrm{mean}_{c,v,s}\,\|h_{cvs} - \bar{h}_{cv}\|_2^2, \qquad G = \frac{\mathrm{mean}_{c,v,o,b:\|b\|_0\geq 4}\,\|q_{cv}(o,b) - h_{cv,o\oplus b}\|_2^2}{D}. \tag{4}$$

The same D scales all methods within a world. Each condition has four views, 16 origins, and 22 targets per origin.

Full-cube quadratic projection. With $\chi_U(s) = \prod_{i \in U} (-1)^{s_i}$, the full-information representation is

$$\widetilde{h}_{\leq 2}(s) = \sum_{|U| \leq 2} \chi_U(s) \left[\frac{1}{64} \sum_{t \in \{0,1\}^6} \chi_U(t) h(t) \right]. \tag{5}$$

We evaluate each target gallery using the corresponding vectors from the projected cube.

Resampling. Worlds are ordered by their fixed identifiers and sampled with replacement using PCG64 seed 2026110905. Let $\hat{\theta}_j$ be the paired point estimate, $\hat{\theta}_j^{*(r)}$ a resampled estimate, and s_j the sample standard deviation of the 10,000 resampled estimates. For each family we take the 95th percentile of $\max_{j:s_j>0} |(\hat{\theta}_j^{*(r)} - \hat{\theta}_j)/s_j|$ and multiply it by s_j for the interval half-width. Constant contrasts have point intervals. The main and strong-control families each use all 2,048 worlds and their own critical values, 2.7053 and 3.0404.

Experimental workload. The evaluation encodes the complete cubes and their queries, then reuses these frozen model outputs across reconstruction methods. Table 6 gives the measured workload.

Table 6. Workload of the paired evaluation across both encoders and renderings.

Quantity	Observed
Encoded document rows	14,155,776
Encoded query rows	150,230
Frozen calibration choices	240
Total GPU execution wall time (s)	1789.139

B Decoder-Based Transfer Protocol

We use Qwen3-Embedding-0.6B at revision `f57bff507cba` [7]. Ordering the 2,048 test identifiers by the SHA-256 digest of a fixed study prefix followed by each identifier gives the 512-world subset. The accompanying protocol records the prefix and the ordered identifiers. A separate 8-world calibration set supports input and throughput checks. We compute scientific endpoints on the selected test worlds.

The tokenizer uses left padding and default special tokens. We pool the last valid token, keep the raw vector for reconstruction, and apply L2 normalization for cosine scores. The model runs in FP32 with eager attention and disabled TF32 and autocasting. Reconstruction and scoring use FP64. Identical full query strings reuse their embeddings.

For each of 10,000 resamples, we draw 512 worlds with replacement using PCG64 seed 2026111006, keeping all conditions and methods paired. We recompute ρ from the two mean NMSEs and take its 2.5th and 97.5th percentiles for the interval.

The calibration stage encodes 12,288 documents and 3,297 unique queries. The test stage encodes 786,432 documents and 60,089 unique queries. The experiment uses 1712.156 seconds of GPU execution wall time, including loading and scoring, on one RTX 5090. The artifact includes complete world and condition statistics.

References

1. BAAI: BGE-small-en-v1.5 model card. Hugging Face, `https://huggingface.co/BAAI/bge-small-en-v1.5`, revision `5c38ec7`
2. Guo, R., Sun, P., Lindgren, E., Geng, Q., Simcha, D., Chern, F., Kumar, S.: Accelerating large-scale inference with anisotropic vector quantization. In: Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 3887–3896. PMLR (2020), `https://proceedings.mlr.press/v119/guo20h.html`
3. Guo, Z., Xue, C., Xu, Z., Bo, H., Ye, Y., Pierrehumbert, J.B., Lewis, M.: Quantifying compositionality of classic and state-of-the-art embeddings. In: Findings of the Association for Computational Linguistics: EMNLP 2025. pp. 22130–22146. Association for Computational Linguistics (2025). `https://doi.org/10.18653/v1/2025.findings-emnlp.1206`
4. intfloat: E5-small-v2 model card. Hugging Face, `https://huggingface.co/intfloat/e5-small-v2`, revision `ffb93f3`
5. Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems **20**(4), 422–446 (2002). `https://doi.org/10.1145/582415.582418`
6. Kuo, F.Y., Sloan, I.H., Wasilkowski, G.W., Woźniakowski, H.: On decompositions of multivariate functions. Mathematics of Computation **79**(270), 953–966 (2010). `https://doi.org/10.1090/S0025-5718-09-02319-9`
7. Qwen Team: Qwen3-Embedding-0.6B model card. Hugging Face (2025), revision f57bff507cba2367b3afbfbc9847736d3f2479cb
8. Uselis, A., Koishigarina, D., Oh, S.J.: How can embedding models bind concepts? In: Proceedings of the 43rd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 306, pp. 123108–123135. PMLR (2026), `https://proceedings.mlr.press/v306/uselis26b.html`
9. Wang, L., Yang, N., Huang, X., Jiao, B., Yang, L., Jiang, D., Majumder, R., Wei, F.: Text embeddings by weakly-supervised contrastive pre-training. arXiv:2212.03533 (2022), `https://arxiv.org/abs/2212.03533`
10. Xiao, S., Liu, Z., Zhang, P., Muennighoff, N., Lian, D., Nie, J.Y.: C-Pack: Packed resources for general Chinese embeddings. In: Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM (2024). `https://doi.org/10.1145/3626772.3657878`
11. Zhang, Y., Li, M., Long, D., Zhang, X., Lin, H., Yang, B., Xie, P., Yang, A., Liu, D., Lin, J., Huang, F., Zhou, J.: Qwen3 embedding: Advancing text embedding and reranking through foundation models. arXiv:2506.05176 (2025). `https://doi.org/10.48550/arXiv.2506.05176`