EQUIVALENT GEOMETRY DOES NOT IMPLY EQUIVALENT DISTILLATION TARGETS

Anonymous authors
Paper under double-blind review

ABSTRACT

Two distillation targets can preserve exactly the same relational geometry yet train the same student to different outcomes. We distinguish two decisions in cross-dimensional embedding distillation: *representation reduction* determines what teacher information the student retains at its dimension, whereas *coordinate resolution* determines how that geometry is presented to the pretrained student. We show that relational objectives are indifferent to orthogonally equivalent reduced targets and derive an exact *initialization-headroom identity* for the endpoint loss recoverable by selecting an orthogonal realization. Under a unique Procrustes solution, the resulting student-conditioned construction is independent of the reducer's arbitrary orthonormal basis and exactly preserves the retained teacher geometry. These results motivate **Orthogonal Resolution of Basis-Indeterminate Targets for Knowledge Distillation (ORBIT-KD)**, which uses a corpus-level closed-form alignment to provide direct endpoint supervision without a trainable cross-space interface. In a matched fixed-target comparison, a single resolution improves a 22M student's nine-task average by 2.22 points over the unreoriented target. A non-orthogonal least-squares target fits the pretrained student more closely but performs 4.40 points worse than fixed Procrustes, showing that proximity alone is insufficient. Across six main model–data settings, ORBIT-KD improves over the strongest baseline by up to 1.66 average points (2.27% relative), with up to a $3.9\times$ reduction in measured optimization-step time. Our work aims to motivate further research on coordinate resolution as a distinct design axis for student-aware, geometry-preserving transfer across heterogeneous representation spaces.

1 INTRODUCTION

Knowledge distillation transfers the behavior of a strong teacher to a compact student. For text embedding models, the transferred representation is itself the deployed output used for retrieval and semantic matching (Li et al., 2025; Truong et al., 2025; Dao et al., 2026; Vujanic & Rückstieß, 2026), rather than an intermediate feature consumed by later layers. When teacher and student embeddings have different dimensions, direct supervision must therefore both preserve useful teacher geometry and specify how that geometry is expressed in the student's representation space.

Figure 1 illustrates our key insight: this interface contains two distinct decisions. *Representation reduction* determines what teacher information survives at the student's width. Yet the reduced embeddings can be rotated together without changing any pairwise similarities, distances, or retained information. *Coordinate resolution* determines which member of this orthogonal orbit is presented to the pretrained student. These realizations are relationally equivalent, but they assign different target vectors—and hence different initial losses and gradients—to the same examples. Relational objectives are intentionally indifferent to this choice (Tian et al., 2020; Truong et al., 2025), while a learned projector may absorb it into an auxiliary interface (Zhang et al., 2024b; Dao et al., 2026). Directly supervising the deployed student embedding instead requires the choice to be explicit.

We propose *Orthogonal Resolution of Basis-Indeterminate Targets for Knowledge Distillation* (ORBIT-KD), a student-conditioned target construction that resolves this coordinate ambiguity without learning a cross-space interface. Our theory formalizes the orthogonal target orbit, shows why relational objectives cannot select a preferred realization, and, under a unique Procrustes solution,

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

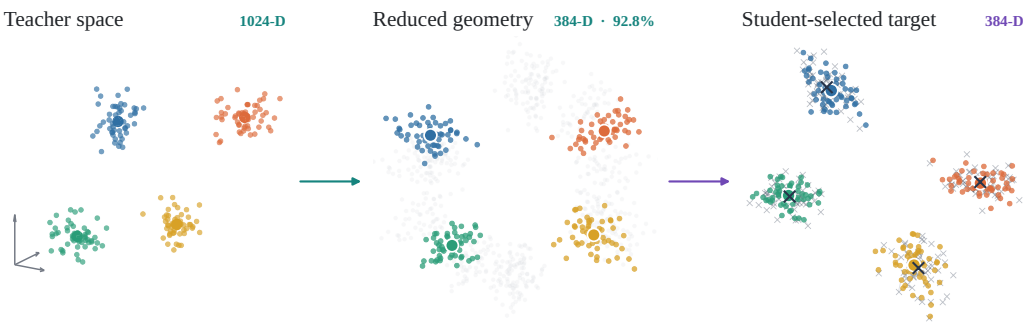


Figure 1: Cross-dimensional embedding distillation contains two decisions. **Left:** teacher embeddings at $d_T = 1024$. **Middle:** representation reduction determines what survives at $d_S = 384$, here 92.8% of the measured teacher energy. **Right:** the retained representation admits an orthogonal orbit of coordinate realizations relative to the student embeddings (crosses). Every realization has the same Gram matrix and relational information, but does not provide the same endpoint supervision.

establishes reducer-basis invariance, student-rotation equivariance, and exact preservation of the retained teacher geometry. It also derives an exact *initialization-headroom identity* for the endpoint loss recoverable by coordinate resolution. Guided by this analysis, ORBIT-KD reduces cached teacher embeddings, aligns them to the pretrained student at the corpus level through closed-form orthogonal Procrustes, and trains the native student output using the resolved endpoint together with coordinate-invariant structural support.

Our contributions are:

1. **We expose a hidden design decision in cross-dimensional distillation.** By separating representation reduction from coordinate resolution, we derive the exact initialization headroom for endpoint supervision and characterize the invariance, equivariance, and geometry preservation of the resolved target.
2. **We resolve the ambiguity without learning a cross-space interface.** ORBIT-KD uses closed-form, corpus-level orthogonal alignment to construct direct targets for native student embeddings while exactly preserving the retained teacher geometry.
3. **We stress-test coordinate resolution across scales, data sources, and training horizons.** Extensive experiments disentangle coordinate resolution from target proximity and structural support, spanning six main settings, teachers above 7B parameters, task-independent corpus, longer training, and optimization efficiency.

2 Related Work

Cross-Dimensional Embedding Distillation. Representation distillation commonly uses bottlenecks or learned projections to connect teacher and student features of different architectures and widths (Sun et al., 2020; Zhao et al., 2022; Zhang et al., 2024b). Sentence-embedding methods have transferred direct embedding targets, contrastive signals, intra-model relations, and teacher-compatible representations (Reimers & Gurevych, 2020; Xu et al., 2023; Truong et al., 2025; Dao et al., 2026; Vujanic & Rückstieß, 2026). A related line learns representations that support multiple output dimensions for resource-adaptive deployment (Kusupati et al., 2022; Li et al., 2025). These approaches address complementary aspects of representation compression. In linear projection-based objectives, the interface can absorb rotations of the native student representation. ORBIT-KD instead fixes the teacher geometry retained at student width and separately resolves its coordinates against the receiving student.

Relational and Contrastive Distillation. A different strategy avoids coordinate-wise matching by transferring attention relations, pairwise similarities, contrastive structure, or global topology (Wang et al., 2020; 2021; Tian et al., 2020; Xu et al., 2023; Truong et al., 2025; Kim et al., 2024).

Objectives built from these relational quantities can preserve geometry across heterogeneous spaces without fixing coordinate-wise correspondence. However, when supervision depends only on pairwise inner products, distances, or a Gram matrix, it is invariant to a shared orthogonal transformation and therefore cannot distinguish two global orthogonal realizations of the same representation. Relational supervision and coordinate resolution thus serve complementary roles: the former specifies the structure to preserve, while the latter selects a concrete realization that can directly supervise the student.

Representation Equivalence and Coordinate Alignment. Representation equivalence motivates CKA and shape metrics (Kornblith et al., 2019; Williams et al., 2021), studies of geometric convergence across models (Huh et al., 2024), and orthogonal alignment of independently trained embedding spaces (Smith et al., 2017). In distillation, geometry-aware methods optimize Procrustes or Gram-matrix losses (Bhattarai et al., 2025), while VkD constrains a trainable projector to preserve within-batch similarities (Miles et al., 2024). ORBIT-KD instead uses Procrustes for explicit target construction after dimensionality reduction: each student-conditioned, corpus-global realization directly supervises the native student output while preserving the retained Gram geometry, remains fixed during gradient updates, and may be refreshed only between epochs.

3 Two Decisions Behind One Target

Let $T \in \mathbb{R}^{N \times d_T}$ and $Z \in \mathbb{R}^{N \times d_S}$ contain teacher and student embeddings for paired corpus examples, with row-normalized Z. Constructing a direct student-width endpoint target requires two distinct decisions: *representation reduction* selects which teacher information survives at width d_S, while *coordinate resolution* selects how the retained geometry is realized relative to the current student. The analysis below concerns fixed matrices Y and Z; Section 4 applies the same construction on a paired fit set $\mathcal{G}$ and holds each resolved target fixed during subsequent differentiated updates.

3.1 Coordinate Ambiguity and Objective Invariance

For a fixed reduction operator ρ, define the student-width teacher representation

$$Y = \mathrm{norm}_{\mathrm{row}}(\rho(T)) \in \mathbb{R}^{N \times d_S}. \tag{1}$$

Its returned coordinates are one realization of the retained geometry. For any orthogonal $R \in O(d_S)$,

$$[Y] = \{YR : R \in O(d_S)\}, \qquad (YR)(YR)^\top = YY^\top. \tag{2}$$

All members of $[Y]$ have identical pairwise inner products, cosine similarities, and Euclidean distances. For an orthonormal subspace reducer, different bases of the *same* retained subspace generate precisely this ambiguity (Appendix A.1); different retained subspaces need not belong to the same orbit. Thus reduction fixes the geometry but provides no student-dependent criterion for choosing an endpoint realization.

Relational objectives are invariant. Let $\mathcal{S}$ be any statistic satisfying $\mathcal{S}(YR) = \mathcal{S}(Y)$ and consider

$$\mathcal{L}_{\mathrm{rel}}(Y, Z) = \Phi(\mathcal{S}(Y), \mathcal{S}(Z)). \tag{3}$$

Gram matrices, pairwise cosines, and distances meet this condition and underlie standard relational objectives (Park et al., 2019; Tung & Mori, 2019; Tian et al., 2020).

Proposition 1 (Relational invariance)**.** *For every $R \in O(d_S)$, $\mathcal{L}_{\mathrm{rel}}(YR, Z) = \mathcal{L}_{\mathrm{rel}}(Y, Z)$.*

Such supervision constrains the geometry represented by $[Y]$ but cannot distinguish which realization should serve as a direct endpoint target. This is an intended invariance of relational objectives, not a deficiency in their stated task.

Linear interfaces may absorb the choice. For a student-to-teacher linear map $W \in \mathbb{R}^{d_S \times d_T}$, consider

$$\mathcal{J}(Z, W) = F(ZW, T, \mathcal{S}(Z), \mathcal{S}(T)), \tag{4}$$

where the native student representation enters outside ZW only through orthogonally invariant quantities.

Proposition 2 (Interface reparameterization). *For every $Q \in O(d_S)$, $\mathcal{J}(ZQ, Q^\top W) = \mathcal{J}(Z, W)$.*

At the representation level the interface can absorb a rotation of Z without changing this objective. The proposition does not claim a symmetry of arbitrary neural-network parameterizations or extend automatically to nonlinear, biased, or coordinate-wise interfaces (Appendix A.3).

3.2 Endpoint Supervision and Initialization Headroom

Direct supervision instead evaluates a concrete member YR of the orbit. For fixed row-normalized Y, Z, define

$$\mathcal{L}_Z(R) = 1 - \frac{1}{N}\operatorname{tr}\big(R^\top Y^\top Z\big), \qquad C = Y^\top Z. \tag{5}$$

Write $C = R^\star H$, where $H = (C^\top C)^{1/2} \succeq 0$ and $R^\star$ is an orthogonal Procrustes maximizer (the unique orthogonal polar factor when C is nonsingular). We use $\|\cdot\|_*$ for the nuclear norm, i.e., the sum of singular values. For this fixed pair Y, Z, how much endpoint loss can an orthogonal choice recover relative to the reducer's returned coordinates $R = I$?

Theorem 1 (Initialization headroom under coordinate resolution). *For fixed row-normalized Y, Z, the improvement in initial endpoint loss available by resolving the particular returned basis Y is exactly*

$$\begin{aligned}\Delta_{\text{init}}(Y, Z) &:= \mathcal{L}_Z(I) - \min_{R \in O(d_S)} \mathcal{L}_Z(R) \\ &= \frac{\|C\|_* - \operatorname{tr}(C)}{N} = \frac{\operatorname{tr}((I - R^\star)H)}{N} \geq 0.\end{aligned} \tag{6}$$

Equality holds if and only if C is symmetric positive semidefinite.

The identity follows from the orthogonal Procrustes optimum (Appendix A.4). It is basis-relative: replacing the unreoriented Y by YQ generally changes $\mathcal{L}_Z(I)$ and Δ_{init}, despite preserving $YY^\top$. Zero headroom means that this particular basis is already endpoint-optimal for the fixed Z.

Different realizations can also change local gradients. For this calculation, let $h_i \in \mathbb{R}^{d_S}$ be a column-vector student output, $z_i = h_i/\|h_i\|_2$ the transpose of row i of Z, and $\tau_i = (y_i R)^\top$ for row y_i of Y. With per-example loss $\ell_i(R) = 1 - z_i^\top \tau_i$, the chain rule gives

$$\nabla_{h_i} \ell_i(R) = -\frac{1}{\|h_i\|_2}(I - z_i z_i^\top)\tau_i. \tag{7}$$

Hence, relationally equivalent targets generally induce different local update directions at a fixed student state. Appendix A.4 additionally characterizes the full best-to-worst loss range over the orbit.

Corollary 1 (No uniformly endpoint-optimal teacher-only realization). *For any two distinct realized targets $YR_1, YR_2 \in [Y]$, there exist row-normalized student states at which either realization has strictly lower endpoint loss than the other.*

Thus a teacher-only canonical basis *may* be defined, but no such fixed realization minimizes endpoint loss for every possible receiving student. This does not imply that the student has an intrinsically defined unique coordinate system (Appendix A.6).

3.3 Student-Conditioned Resolution of the Orbit

Define the student-conditioned Procrustes target construction

$$\Psi(Y, Z) := YR^\star(Y, Z), \qquad R^\star(Y, Z) \in \operatorname*{arg\,min}_{R \in O(d_S)} \|YR - Z\|_F^2. \tag{8}$$

When $C = Y^\top Z = U\Sigma V^\top$ is nonsingular, its unique orthogonal factor is $R^\star = UV^\top = \operatorname{polar}(C)$.

Proposition 3 (Basis invariance, student equivariance, and geometry preservation). *Suppose $Y^\top Z$ is nonsingular. For all $Q \in O(d_S)$,*

$$\begin{aligned}\Psi(YQ, Z) &= \Psi(Y, Z), \\ \Psi(Y, ZQ) &= \Psi(Y, Z)Q, \\ \Psi(Y, Z)\Psi(Y, Z)^\top &= YY^\top.\end{aligned} \tag{9}$$

In particular, the resolved target is independent of the arbitrary orthonormal basis returned by a reducer when the retained subspace is fixed. This says nothing about the equivalence of *different* retained subspaces. When $Y^\top Z$ is singular, the optimal rotation may be non-unique; the *set* of Procrustes-resolved targets has the corresponding invariance/equivariance, but an arbitrary SVD completion need not select the same target across bases. The shared transformation still preserves $YY^\top$ exactly for every selection (Appendix A.5).

Global consistency. One orthogonal matrix should be shared across the fit corpus. Example-specific or independently fitted minibatch rotations preserve within-group geometry but generally change cross-group inner products: $\langle y_i R_i, y_j R_j \rangle = y_i R_i R_j^\top y_j^\top$ need not equal $y_i y_j^\top$. A common R preserves the whole corpus Gram matrix (Appendix A.7). These results concern fixed representations and do not guarantee training convergence or downstream gains. Section 4 turns the shared target construction into a training procedure, whose longer-term behavior is evaluated in Section 5.3.

4 Proposed Method

Overview ORBIT-KD turns cached teacher embeddings into direct targets for the student's native output. Let $T \in \mathbb{R}^{N \times d_T}$ and $Z \in \mathbb{R}^{N \times d_S}$ denote row-normalized teacher and student embeddings for a shared corpus. The method reduces T to width d_S, resolves the remaining orthogonal ambiguity against Z with a corpus-level Procrustes fit, and trains the student with a cosine endpoint loss and coordinate-invariant structural support. Target construction requires no trainable cross-space interface.

Teacher Representation Reduction The first stage selects a student-width teacher subspace independently of the student. We fit rank-d_S PCA to the centered teacher cache:

$$\bar{\mathbf{t}} = \frac{1}{N} T^\top \mathbf{1}, \qquad \widetilde{T} = T - \mathbf{1}\bar{\mathbf{t}}^\top. \tag{10}$$

With $\widetilde{T} = U_T \Sigma_T V_T^\top$, the leading d_S right singular vectors maximize projected variance (Hotelling, 1933):

$$P_{\mathrm{PCA}} \in \underset{P^\top P = I_{d_S}}{\arg\max} \|\widetilde{T} P\|_F^2, \qquad P_{\mathrm{PCA}} = [\mathbf{v}_1, \ldots, \mathbf{v}_{d_S}]. \tag{11}$$

We apply this basis to the original uncentered embeddings, without whitening, and row-normalize:

$$Y = \mathrm{norm}_{\mathrm{row}}(T P_{\mathrm{PCA}}) \in \mathbb{R}^{N \times d_S}. \tag{12}$$

PCA is the default reducer; the identity can be used when $d_T = d_S$, and other fixed reducers are compatible. For every $R \in O(d_S)$,

$$(YR)(YR)^\top = YY^\top. \tag{13}$$

The next stage selects coordinates relative to the student while preserving this retained geometry.

Student-Conditioned Coordinate Resolution The second stage resolves the degree of freedom left by reduction using the receiving student as a reference. The classical orthogonal Procrustes problem finds the rotation that minimizes the squared discrepancy between two paired matrices (Schönemann, 1966). For matrices A and B, its objective and one closed-form minimizer are

$$\begin{aligned} R_{\mathrm{Proc}}(A, B) &\in \underset{R \in O(d_S)}{\arg\min} \|AR - B\|_F^2, \\ A^\top B = U \Sigma V^\top \quad &\Longrightarrow \quad R_{\mathrm{Proc}}(A, B) := UV^\top. \end{aligned} \tag{14}$$

In our setting, both fit matrices have unit-norm rows, so minimizing this Frobenius discrepancy is equivalent, up to an affine rescaling, to minimizing their aggregate cosine endpoint loss. ORBIT-KD specializes the result to a fixed corpus-level fit set $\mathcal{G}$, using its reduced teacher embeddings $Y_{\mathcal{G}}$ and the student's current embeddings $Z_{\mathcal{G}}$. The fit-set cross-covariance $C_{\mathcal{G}} = Y_{\mathcal{G}}^\top Z_{\mathcal{G}}$ is the counterpart of $C = Y^\top Z$ in Section 3.2; its fitted rotation attains the headroom optimum of Theorem 1 on this

set (replace N by $|\mathcal{G}|$). The selected rotation is shared by the entire corpus; with $\mathbf{y}_i$ denoting row i of Y, it defines a column-vector target for each example:

$$R^\star = R_{\mathrm{Proc}}(Y_\mathcal{G}, Z_\mathcal{G}), \qquad \boldsymbol{\tau}_i = (\mathbf{y}_i R^\star)^\top. \tag{15}$$

Using one global transformation is essential: independent sample-wise rotations could reduce individual discrepancies but would destroy the coherence of the representation. In contrast, the shared orthogonal $R^\star$ selects an endpoint-optimal member of $[Y]$ on the paired fit set while preserving $YY^\top$ exactly. When $C_\mathcal{G}$ is nonsingular, Proposition 3 guarantees that the resulting target does not depend on the particular orthonormal basis of the retained teacher subspace. If $C_\mathcal{G}$ is singular, the SVD selects one optimal orthogonal completion; basis invariance then applies to the set of optimal targets, not necessarily to a particular numerical completion. The resulting $\boldsymbol{\tau}_i$ can therefore supervise the native student embedding directly, without an intervening projector. With $\mathbf{z}_i$ the column-vector form of row i of Z, the cosine endpoint loss for a minibatch $\mathcal{B}$ is

$$\mathcal{L}_{\mathrm{end}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \left(1 - \langle \mathbf{z}_i, \boldsymbol{\tau}_i \rangle\right). \tag{16}$$

Coordinate resolution therefore changes each endpoint direction in a student-referenced frame while preserving every pairwise relation in the retained teacher geometry.

Coordinate-Invariant Structural Support Endpoint supervision is complemented by a structural loss computed in the native teacher and student spaces. Hofer et al. formulate topological regularization through the finite H_0 death times: for a minibatch partitioned into n sub-batches, let $\dagger(\mathcal{B}_i)$ denote the persistence barcode of $\mathcal{B}_i$ with its infinite interval omitted. Their connectivity penalty pulls every death time toward a prescribed scale β (Hofer et al., 2020):

$$\mathcal{L}_{\mathrm{topo}}(\mathcal{B}; \beta) = \sum_{i=1}^{n} \sum_{d \in \dagger(\mathcal{B}_i)} |d - \beta|. \tag{17}$$

For distillation, a fixed β would impose a generic connectivity scale rather than transfer the teacher's structure. We instead use the teacher barcode as a data-dependent reference. Within a minibatch of size B, we compute chord distances $d(\mathbf{u}, \mathbf{v}) = \sqrt{2 - 2\langle \mathbf{u}, \mathbf{v} \rangle}$ and build separate minimum spanning trees in the native teacher and student spaces. Their sorted edge weights $q^T_{(j)}$ and $q^S_{(j)}$ are exactly the finite H_0 death times. Motivated by Equation 17, we replace the fixed scale β with the corresponding teacher death time and use the mean squared discrepancy between the two sorted barcodes:

$$\mathcal{L}_{H_0} = \frac{1}{B-1} \sum_{j=1}^{B-1} \left(q^S_{(j)} - q^T_{(j)}\right)^2. \tag{18}$$

This loss matches the sorted multiscale connectivity merge times of the two clouds without requiring equal ambient dimensions or a shared coordinate frame. It does not enforce the identities of particular nearest-neighbor pairs or MST edges; endpoint supervision supplies per-example correspondence.

Overall Optimization Objective. With the resolved targets held fixed between coordinate refits, we optimize the student parameters θ over minibatches $\mathcal{B}$:

$$\min_\theta \; \mathbb{E}_{\mathcal{B} \sim \mathcal{P}_{\mathrm{batch}}} \left[\mathcal{L}_{\mathrm{end}}(\theta; \mathcal{B}) + \lambda_{H_0} \mathcal{L}_{H_0}(\theta; \mathcal{B})\right]. \tag{19}$$

Here, $\mathcal{P}_{\mathrm{batch}}$ is the training minibatch distribution, and λ_{H_0} weights structural support relative to direct endpoint supervision.

5 Experiments

5.1 Setup

Baselines. We re-implement RKD (Park et al., 2019), Jasper (Zhang et al., 2024a), CDM (Chen et al., 2025), DSKD (Zhang et al., 2024b), EMO (Truong et al., 2025), and TALAS (Dao et al., 2026). All methods share teacher–student pairs, data, evaluation, batch size, and training budget while retaining their original objectives and optimizers; learning rates are selected on validation data.

Training Data. The main corpus contains 14,760 unlabeled texts drawn roughly equally from the training splits of Emotion, WiC, and STS-B; we repeat the comparison on approximately 100K unlabeled texts from all nine task training splits. No labels or evaluation examples enter distillation. We treat both corpora as task-adapted.

Evaluation. We evaluate frozen embeddings on Banking77, Tweet, and Emotion using macro-F1; MRPC, SciTail, and WiC using average precision from cosine similarity; and SICK, STS12, and STS-B using Spearman correlation. Metrics are multiplied by 100, and their unweighted mean is reported as a descriptive overall score.

Implementation Details. We test Qwen3-Embedding-4B $\rightarrow$ BERT-base, BGE-M3 $\rightarrow$ MiniLMv2-L6-H768, and Qwen3-Embedding-0.6B $\rightarrow$ MiniLMv2-L6-H384 (Zhang et al., 2025; Devlin et al., 2019; Chen et al., 2024; Wang et al., 2021). We train for five epochs with batch size 128 and report mean $\pm$ sample standard deviation over seeds 42, 43, and 44. ORBIT-KD uses PCA and fits a corpus-level Procrustes rotation to the pretrained student before training, then refreshes it after each non-final epoch using the same paired fit set; no gradient passes through the fit. We set $\lambda_{H_0} = 0.5$ and the learning rate to 7×10^{-5}; Appendix B gives full details.

5.2 Main Results

Table 1: Overall performance across nine sentence benchmarks. Scores are the unweighted mean of task-appropriate macro-F1, average precision, and Spearman correlation, scaled by 100. **Bold** and underlined values denote the best and second-best distilled methods.

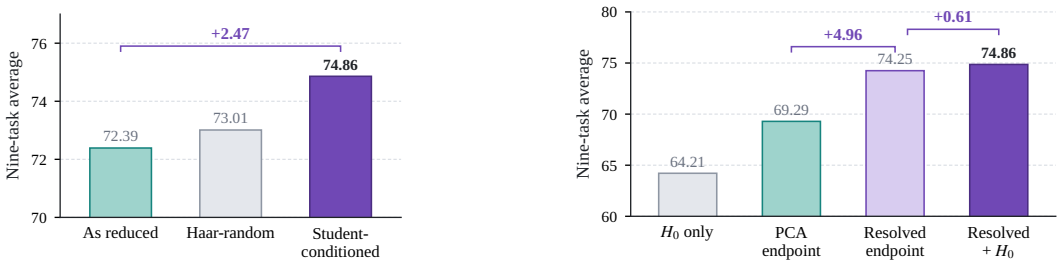
	15K corpus			100K corpus		
Method	(a) 109M	(b) 66M	(c) 22M	(a) 109M	(b) 66M	(c) 22M
Teacher	81.07	79.80	79.14	81.07	79.80	79.14
Student (pre-KD)	54.20	56.36	52.62	54.20	56.36	52.62
RKD	73.52 $\pm$ 0.06	73.66 $\pm$ 0.04	71.44 $\pm$ 0.10	76.48 $\pm$ 0.04	75.54 $\pm$ 0.10	73.56 $\pm$ 0.09
Jasper	71.95 $\pm$ 0.06	71.13 $\pm$ 0.11	68.57 $\pm$ 0.10	75.59 $\pm$ 0.10	73.95 $\pm$ 0.04	70.34 $\pm$ 0.02
CDM	71.12 $\pm$ 0.18	70.77 $\pm$ 0.02	68.69 $\pm$ 0.27	72.36 $\pm$ 0.13	71.27 $\pm$ 0.10	69.36 $\pm$ 0.08
DSKD	71.14 $\pm$ 0.10	70.83 $\pm$ 0.11	68.59 $\pm$ 0.04	72.07 $\pm$ 0.03	70.58 $\pm$ 0.10	69.29 $\pm$ 0.09
EMO	69.17 $\pm$ 0.24	69.59 $\pm$ 0.12	67.62 $\pm$ 0.02	69.07 $\pm$ 0.89	70.40 $\pm$ 0.35	68.47 $\pm$ 0.15
TALAS	77.29 $\pm$ 0.04	75.55 $\pm$ 0.06	73.20 $\pm$ 0.06	79.95 $\pm$ 0.08	77.97 $\pm$ 0.03	76.35 $\pm$ 0.08
ORBIT-KD	**77.67** $\pm$ 0.03	**76.03** $\pm$ 0.11	**74.86** $\pm$ 0.06	**80.02** $\pm$ 0.03	**78.24** $\pm$ 0.05	**77.48** $\pm$ 0.04

ORBIT-KD establishes a new performance frontier. Table 1 shows the highest overall average in all six model–data settings. ORBIT-KD improves over TALAS by up to 1.66 points, over RKD by 2.37–4.15, and over Jasper by 4.29–7.14. The margin over TALAS is larger for the more compact student in our tested configurations: it grows from 0.38 to 1.66 points on the 15K corpus and from 0.07 to 1.13 on the 100K corpus, peaking for the compact 22M student.

5.3 Dissecting Coordinate Resolution

The target frame, not a change in geometry, drives the gain. Figure 2(a) shows a 2.47-point spread among orthogonal realizations of one rank-384 PCA target, despite identical Gram matrices and retained energy. Across reducers, student-conditioned coordinates add 1.40–2.47 points, whereas resolved full-rank reducers differ by at most 0.40 (Appendix C.1). Figure 2(b) then isolates 4.96 points from resolving the endpoint and only 0.61 more from H_0: ORBIT-KD's primary gain is direct supervision in the student's frame, with structural support playing a secondary role (Appendix C.4).

Equal geometry does not mean equally useful starting targets. In Figure 3(a,b), initial loss spans 0.35 at $R^\star$ to 1.65 at $-R^\star$ without changing geometry. Among 3,000 Haar draws, even



(a) Coordinate realizations **(b)** Objective decomposition

Figure 2: Coordinate resolution in configuration (c). **(a)** Orthogonal realizations of one rank-384 PCA geometry differ by 2.47 points. **(b)** Resolving the endpoint adds 4.96 points; H_0 adds 0.61 more. Full results are in Appendices C.1 and C.4.

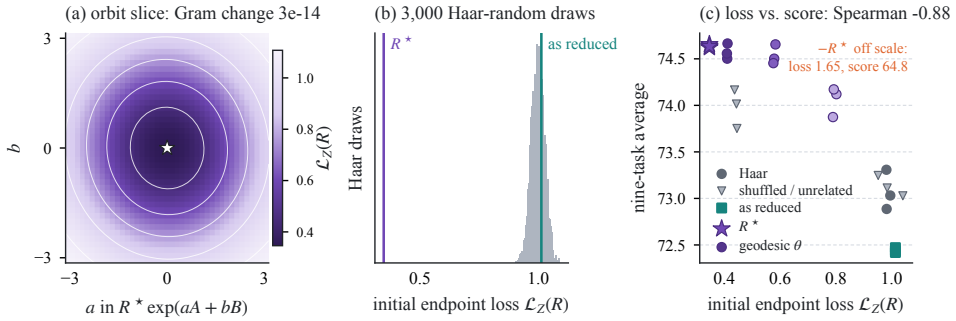


Figure 3: Initialization loss and trained outcomes across one orthogonal orbit (configuration (c)). **(a)** Loss on an orbit slice; Gram change is below 4×10^{-14}. **(b)** Loss of 3,000 Haar draws versus the reduced frame and $R^\star$. **(c)** Initial loss versus final nine-task average for fixed-target realizations, including a geodesic from $R^\star$ to a Haar draw (Appendix C.6).

the best incurs $2.6\times$ the loss of $R^\star$; ORBIT-KD computes that favorable realization rather than accepting an arbitrary reducer basis.

The initial advantage survives training. Figure 3(c) extends the theorem's starting-loss result empirically: initial loss and final score correlate at Spearman -0.88 (excluding the antipode). Along a path from $R^\star$ to a Haar draw, the nine-task average falls from 74.64 to 73.08; $-R^\star$ scores 64.84. This dose–response evidence does not guarantee downstream optima (Appendix C.6).

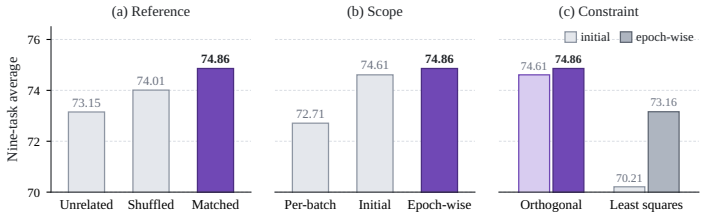


Figure 4: Coordinate-resolution controls in configuration (c). Matched references, corpus-level fitting, and orthogonal maps perform best in the left, center, and right panels, respectively. Full results are in Appendices C.2 and C.3.

The reference must be matched and globally coherent. Figure 4 shows 74.86 with the matched student versus 73.15 with an unrelated reference and 74.01 with shuffled pairings. One corpus-level fit reaches 74.61 versus 72.71 for independent minibatch fits; refreshing adds 0.25. ORBIT-KD needs paired student information and a shared frame, not just a rotation.

A closer target can be worse if it changes geometry. In the matched fixed-map comparison in Figure 4's right panel, least squares raises initial cosine agreement from 0.65 to 0.99 yet trails fixed Procrustes by 4.40 points. It changes the target Gram matrix by 186% and reduces effective minibatch rank from 104.6 to 18.8 (Appendix C.3). Orthogonal resolution adapts the frame while retaining geometry.

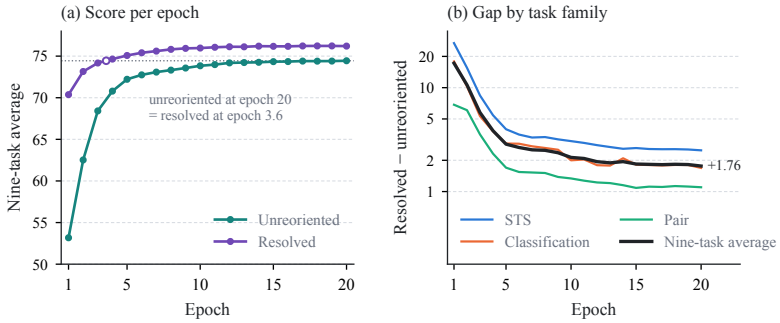


Figure 5: Twenty-epoch training with endpoint loss only. **(a)** Nine-task averages for unreoriented and coordinate-resolved targets. **(b)** Resolved-minus-unreoriented gains by task family.

The head start lasts beyond early training. With endpoint loss alone, Figure 5 shows the resolved student reaching the unreoriented student's epoch-20 score by epoch 3.6 and still leading by 1.76 points at epoch 20 across all three task families. The finite-budget gain needs neither H_0 nor a short run; it does not establish different asymptotic optima.

The first resolution does most of the work. Under the complete objective and matched three-seed settings, a single initial fit raises the epoch-20 score by 1.00 point over unreoriented PCA (Figure 9). Epoch-wise refitting adds only about 0.05 point. In this setting, most of the observed benefit comes from choosing the student's coordinate frame before training.

5.4 Training Efficiency

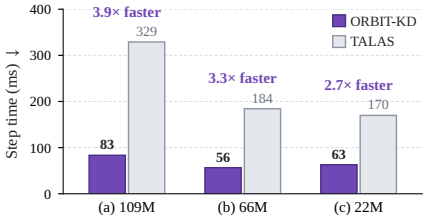


Figure 6: Optimization-step time on one NVIDIA H200 GPU per run (mean ± standard deviation). ORBIT-KD is up to 3.9× faster than TALAS; full results are in Appendix D.5.

ORBIT-KD nearly quadruples optimization in the best setting. Figure 6 shows 3.9×, 3.3×, and 2.7× lower step time than TALAS without an online teacher or trainable projector; refresh takes under 0.7 minutes per run.

6 Conclusion

ORBIT-KD resolves the student's coordinate frame after reducing teacher representations, enabling direct supervision without a trainable cross-space interface. It achieves the highest nine-task average in all six main settings. Controlled studies isolate a 4.96-point gain from coordinate-resolved endpoints and up to 3.9× faster optimization than the strongest baseline. Coordinate resolution is therefore a distinct design decision for embedding distillation.

AI USE STATEMENT

Generative AI tools were used to improve the clarity and readability of the manuscript and to survey related literature for potential overlap. All AI-assisted suggestions were reviewed by the authors, who take full responsibility for the theoretical claims, experimental results, code, and written content of this work.

REPRODUCIBILITY STATEMENT

Section 4 specifies the representation reduction, coordinate resolution, and training objectives; Section 5.1 gives the training schedule. Appendix A gives the assumptions and proofs, while Appendix B records the data construction, model checkpoints and pooling choices, training configuration, and implementation details. Appendices C and D provide controlled analyses and task-level results. We report the seeds and matched settings used for controlled comparisons. Every experimental run used a single NVIDIA H200 GPU.

ETHICS STATEMENT

This work studies knowledge distillation for text embedding models using existing research datasets and does not collect new human-subject data. As with other distillation approaches, the student may inherit biases present in the teacher representations and training data. We therefore encourage further evaluation of fairness and robustness before deployment in sensitive applications.

REFERENCES

Prajjwal Bhattarai, Mohammad Amjad, Dmytro Zhylko, and Tuka Alhanai. Knowledge distillation through geometry-aware representational alignment. *arXiv preprint arXiv:2509.25253*, 2025. doi:10.48550/arXiv.2509.25253. URL https://arxiv.org/abs/2509.25253.

Jianlyu 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. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024. doi:10.18653/v1/2024.findings-acl.137. URL https://aclanthology.org/2024.findings-acl.137/.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Enhancing cross-tokenizer knowledge distillation with contextual dynamical mapping. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8005–8018, 2025. doi:10.18653/v1/2025.findings-acl.419. URL https://aclanthology.org/2025.findings-acl.419/.

Quoc Phong Dao, Hoang Son Nguyen, Pham Khanh Chi, Linh Ngo Van, Nguyen Thi Ngoc Diep, Thien Huu Nguyen, and Trung Le. TALAS: Teacher-anchored layer alignment with adaptive sharpness-aware minimization for embedding distillation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 32705–32723, 2026. doi:10.18653/v1/2026.acl-long.1509. URL https://aclanthology.org/2026.acl-long.1509/.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186, 2019. doi:10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423/.

Christoph D. Hofer, Florian Graf, Marc Niethammer, and Roland Kwitt. Topologically densified distributions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4304–4313, 2020. URL https://proceedings.mlr.press/v119/hofer20a.html.

Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933. doi:10.1037/h0071325.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 20617–20642, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

Jungeun Kim, Junwon You, Dongjin Lee, Ha Young Kim, and Jae-Hun Jung. Do topological characteristics help in knowledge distillation? In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 24674–24693, 2024. URL https://proceedings.mlr.press/v235/kim24aj.html.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey E. Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529, 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham M. Kakade, Prateek Jain, and Ali Farhadi. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pp. 30233–30249, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/c32319f4868da7613d78af9993100e42-Abstract-Conference.html.

Xianming Li, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. ESE: Espresso sentence embeddings. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=plgLA2YBLH.

Roy Miles, Ismail Elezi, and Jiankang Deng. VkD: Improving knowledge distillation using orthogonal projections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15720–15730, June 2024. doi:10.1109/CVPR52733.2024.01488. URL https://openaccess.thecvf.com/content/CVPR2024/html/Miles_VkD_Improving_Knowledge_Distillation_using_Orthogonal_Projections_CVPR_2024_paper.html.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3967–3976, 2019. doi:10.1109/CVPR.2019.00409. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Park_Relational_Knowledge_Distillation_CVPR_2019_paper.html.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 4512–4525, 2020. doi:10.18653/v1/2020.emnlp-main.365. URL https://aclanthology.org/2020.emnlp-main.365/.

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

Samuel L. Smith, David H. P. Turban, Steven Hamblin, and Nils Y. Hammerla. Offline bilingual word vectors, orthogonal transformations and the inverted softmax. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=r1Aab85gg.

Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, and Denny Zhou. MobileBERT: A compact task-agnostic BERT for resource-limited devices. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2158–2170, 2020. doi:10.18653/v1/2020.acl-main.195. URL https://aclanthology.org/2020.acl-main.195/.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SkgpBJrtvS.

Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. EMO: Embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 7594–7606, 2025. doi:10.18653/v1/2025.emnlp-main.385. URL https://aclanthology.org/2025.emnlp-main.385/.

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019. doi:10.1109/ICCV.2019.00145. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Tung_Similarity-Preserving_Knowledge_Distillation_ICCV_2019_paper.html.

Robin Vujanic and Thomas Rückstieß. LEAF: Knowledge distillation of text embedding models with teacher-aligned representations. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 43362–43383, 2026. doi:10.18653/v1/2026.acl-long.2008. URL https://aclanthology.org/2026.acl-long.2008/.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Wenhui Wang, Hangbo Bao, Shaohan Huang, Li Dong, and Furu Wei. MiniLMv2: Multi-head self-attention relation distillation for compressing pretrained transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2140–2151, 2021. doi:10.18653/v1/2021.findings-acl.188. URL https://aclanthology.org/2021.findings-acl.188/.

Alex H. Williams, Erin Kunz, Simon Kornblith, and Scott W. Linderman. Generalized shape metrics on neural representations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 4738–4750, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/252a3dbaeb32e7690242ad3b556e626b-Abstract.html.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. DistillCSE: Distilled contrastive learning for sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8153–8165, 2023. doi:10.18653/v1/2023.findings-emnlp.547. URL https://aclanthology.org/2023.findings-emnlp.547/.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. Jasper and Stella: Distillation of SOTA embedding models. *arXiv preprint arXiv:2412.19048*, 2024a. doi:10.48550/arXiv.2412.19048. URL https://arxiv.org/abs/2412.19048.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181, 2024b. doi:10.18653/v1/2024.emnlp-main.1010. URL https://aclanthology.org/2024.emnlp-main.1010/.

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. *arXiv preprint arXiv:2506.05176*, 2025. doi:10.48550/arXiv.2506.05176. URL https://arxiv.org/abs/2506.05176.

Xuandong Zhao, Zhiguo Yu, Ming Wu, and Lei Li. Compressing sentence representation for semantic retrieval via homomorphic projective distillation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 774–781, 2022. doi:10.18653/v1/2022.findings-acl.64. URL https://aclanthology.org/2022.findings-acl.64/.

A Theoretical Proofs

This appendix provides the derivations supporting Section 3. Throughout, $Y, Z \in \mathbb{R}^{N \times d_S}$ denote row-normalized teacher targets and student embeddings, respectively, and

$$[Y] = \{YR : R \in O(d_S)\} \tag{A.1}$$

denotes the orthogonal orbit of a fixed student-width teacher representation. The results concern a fixed representation Y, a fixed student state Z, and the objectives specified in Section 3. They characterize coordinate identifiability and endpoint supervision in this setting; they are not guarantees about the final solution of non-convex student training.

A.1 Basis Non-Uniqueness after Representation Reduction

We first make explicit why retaining a teacher subspace does not privilege one coordinate frame within that subspace. Let $P, P' \in \mathbb{R}^{d_T \times d_S}$ have orthonormal columns and span the same d_S-dimensional subspace. Define

$$R = P^\top P'. \tag{A.2}$$

Because every column of P' lies in $\mathrm{span}(P)$,

$$PR = PP^\top P' = P'. \tag{A.3}$$

Moreover,

$$R^\top R = {P'}^\top PP^\top P' = {P'}^\top P' = I, \tag{A.4}$$

so $R \in O(d_S)$. Thus, two orthonormal coordinate systems for the same retained subspace differ only by an orthogonal transformation.

Assume the projected rows are nonzero, as required by row normalization, and let

$$Y = \mathrm{norm}_{\mathrm{row}}(TP), \qquad Y' = \mathrm{norm}_{\mathrm{row}}(TP'). \tag{A.5}$$

For any row vector x and orthogonal R, $\|xR\|_2 = \|x\|_2$. Row normalization therefore commutes with the change of basis:

$$\begin{aligned} Y' &= \mathrm{norm}_{\mathrm{row}}(TPR) \\ &= \mathrm{norm}_{\mathrm{row}}(TP)R \\ &= YR. \end{aligned} \tag{A.6}$$

This establishes the orthogonal orbit in Equation A.1 directly from the non-uniqueness of a basis for the retained teacher subspace. A particular reduction algorithm may return one numerical basis, but the retained subspace itself supplies no criterion for preferring that basis over another.

The transformation changes only the coordinate realization. In addition to the Gram identity $(YR)(YR)^\top = YY^\top$, it preserves row norms and pairwise Euclidean distances. It also preserves rank and singular values because

$$(YR)^\top (YR) = R^\top (Y^\top Y)R, \tag{A.7}$$

which is an orthogonal similarity transformation. Before row normalization, $\|TPR\|_F = \|TP\|_F$, so the retained-energy statistic is unchanged as well. Selecting a member of $[Y]$ therefore resolves a coordinate degree of freedom without revising which teacher subspace or geometry was retained.

A.2 Proof of Proposition 1: Relational Invariance

We first show that the standard relational quantities considered in Section 3.1 are invariant over the orbit $[Y]$. For any $R \in O(d_S)$,

$$(YR)(YR)^\top = YRR^\top Y^\top = YY^\top. \tag{A.8}$$

Thus, every member of $[Y]$ has the same Gram matrix and therefore the same pairwise inner products.

Because right multiplication by an orthogonal matrix preserves row norms, $\|y_i R\|_2 = \|y_i\|_2$, so pairwise cosine similarities are also unchanged. Likewise,

$$\begin{aligned} \|y_i R - y_j R\|_2^2 &= \|(y_i - y_j)R\|_2^2 \\ &= (y_i - y_j)RR^\top (y_i - y_j)^\top \\ &= \|y_i - y_j\|_2^2. \end{aligned} \tag{A.9}$$

Hence, any statistic $\mathcal{S}$ constructed from these orthogonally invariant quantities satisfies

$$\mathcal{S}(YR) = \mathcal{S}(Y). \tag{A.10}$$

For the class of relational objectives considered in Section 3.1,

$$\mathcal{L}_{\mathrm{rel}}(Y, Z) = \Phi(\mathcal{S}(Y), \mathcal{S}(Z)), \tag{A.11}$$

we therefore have

$$\begin{aligned} \mathcal{L}_{\mathrm{rel}}(YR, Z) &= \Phi(\mathcal{S}(YR), \mathcal{S}(Z)) \\ &= \Phi(\mathcal{S}(Y), \mathcal{S}(Z)) \\ &= \mathcal{L}_{\mathrm{rel}}(Y, Z). \end{aligned} \tag{A.12}$$

This proves Proposition 1. Such an objective may constrain the relational geometry represented by $[Y]$, but it cannot distinguish which member of that orbit is used as a concrete coordinate realization.

A.3 Proof of Proposition 2: Interface Reparameterization

Consider the linear student-to-teacher interface $W \in \mathbb{R}^{d_S \times d_T}$ and the objective

$$\mathcal{J}(Z, W) = F(ZW, T, \mathcal{S}(Z), \mathcal{S}(T)), \tag{A.13}$$

where every dependence on the native student representation outside the projected embedding ZW occurs through an orthogonally invariant statistic $\mathcal{S}$.

Let $Q \in O(d_S)$ and define

$$Z' = ZQ, \qquad W' = Q^\top W. \tag{A.14}$$

The projected student representation is unchanged:

$$Z'W' = ZQQ^\top W = ZW. \tag{A.15}$$

By orthogonal invariance,

$$\mathcal{S}(Z') = \mathcal{S}(ZQ) = \mathcal{S}(Z). \tag{A.16}$$

Substitution into Equation A.13 gives

$$\begin{aligned} \mathcal{J}(ZQ, Q^\top W) &= F\big(ZQQ^\top W, T, \mathcal{S}(ZQ), \mathcal{S}(T)\big) \\ &= F(ZW, T, \mathcal{S}(Z), \mathcal{S}(T)) \\ &= \mathcal{J}(Z, W). \end{aligned} \tag{A.17}$$

This proves Proposition 2.

The proposition is deliberately scoped to the linear interface and objective class in Equation A.13. It does not assert an exact symmetry for arbitrary nonlinear projectors. Rather, it shows that in a common projector formulation, native coordinate changes can be absorbed through the interface parameterization while leaving both the projected representation and orthogonally invariant relational terms unchanged. The coordinate decision is therefore handled inside the auxiliary interface rather than exposed as a separate target-construction decision.

A.4 Proof of Theorem 1: Initialization Headroom

For $R \in O(d_S)$, the mean cosine endpoint loss is

$$\mathcal{L}_Z(R) = 1 - \frac{1}{N}\operatorname{tr}(R^\top C), \qquad C = Y^\top Z. \tag{A.18}$$

Let $C = U\Sigma V^\top$ be a full singular-value decomposition. For any orthogonal R, put $Q = U^\top RV \in O(d_S)$. Then

$$\operatorname{tr}(R^\top C) = \operatorname{tr}(Q^\top \Sigma) = \sum_{j=1}^{d_S} \sigma_j Q_{jj} \le \sum_{j=1}^{d_S} \sigma_j = \|C\|_*. \tag{A.19}$$

The bound is attained by $R^\star = UV^\top$, a classical orthogonal Procrustes solution (Schönemann, 1966). Consequently,

$$\min_{R \in O(d_S)} \mathcal{L}_Z(R) = 1 - \frac{\|C\|_*}{N}. \tag{A.20}$$

Subtracting from the loss of the particular returned realization $R = I$ gives the initialization-headroom identity

$$\Delta_{\text{init}}(Y, Z) = \mathcal{L}_Z(I) - \min_R \mathcal{L}_Z(R) = \frac{\|C\|_* - \operatorname{tr}(C)}{N} \ge 0. \tag{A.21}$$

For a (possibly singular) polar decomposition $C = R^\star H$ with $H = (C^\top C)^{1/2} \succeq 0$, one may choose an orthogonal extension $R^\star$ of the polar partial isometry. Then $\|C\|_* = \operatorname{tr}(H)$ and $\operatorname{tr}(C) = \operatorname{tr}(R^\star H)$, so

$$\Delta_{\text{init}}(Y, Z) = \frac{\operatorname{tr}((I - R^\star)H)}{N}. \tag{A.22}$$

When C is singular, the orthogonal extension need not be unique, but $R^\star H = C$ makes the displayed quantity independent of the extension.

Equality characterization. If $C \succeq 0$ is symmetric, its nuclear norm equals its trace and $\Delta_{\text{init}} = 0$. Conversely, suppose equality holds. For an SVD $C = U\Sigma V^\top$,

$$\operatorname{tr}(C) = \sum_j \sigma_j (V^\top U)_{jj} = \sum_j \sigma_j. \tag{A.23}$$

Each diagonal entry of an orthogonal matrix is at most one. Thus $(V^\top U)_{jj} = 1$ for every j with $\sigma_j > 0$. Such a diagonal entry forces the corresponding basis vector to be fixed; hence the orthogonal polar factor acts as identity on the support of Σ. Therefore $U\Sigma V^\top = V\Sigma V^\top$, which is symmetric positive semidefinite. This proves the "if and only if" statement, including the rank-deficient case. The result measures endpoint improvement at a fixed student state; it asserts nothing about the optimizer's final parameter state.

Classical best-to-worst range (context only). Since $-R \in O(d_S)$ whenever $R \in O(d_S)$, the worst orientation is $-R^\star$, giving

$$\min_R \mathcal{L}_Z(R) = 1 - \frac{\|C\|_*}{N}, \qquad \max_R \mathcal{L}_Z(R) = 1 + \frac{\|C\|_*}{N}, \tag{A.24}$$

$$\operatorname{range}_{R \in O(d_S)} \mathcal{L}_Z(R) = \frac{2}{N}\|C\|_*. \tag{A.25}$$

This range is the exact classical Procrustes best-to-worst spread, not the headroom measured relative to the reducer's actual returned basis. It does not predict either downstream score or optimization speed.

A.5 Proof of Proposition 3: Basis Invariance and Student Equivariance

Assume $C = Y^\top Z$ is nonsingular. Its orthogonal polar factor is unique:

$$R^\star = \operatorname{polar}(C) = C(C^\top C)^{-1/2}. \tag{A.26}$$

For an orthogonal Q, the cross-covariance of (YQ, Z) equals $Q^\top C$, and uniqueness of the polar factor implies $\operatorname{polar}(Q^\top C) = Q^\top R^\star$. Similarly, $\operatorname{polar}(CQ) = R^\star Q$ because $CQ = (R^\star Q)(Q^\top H Q)$ with $Q^\top H Q \succeq 0$. Thus

$$\begin{aligned} \Psi(YQ, Z) &= YQ \operatorname{polar}(Q^\top C) = YR^\star = \Psi(Y, Z), \\ \Psi(Y, ZQ) &= Y \operatorname{polar}(CQ) = YR^\star Q = \Psi(Y, Z)Q. \end{aligned} \tag{A.27}$$

As $R^\star$ is orthogonal, the Gram matrix is unchanged:

$$\Psi(Y,Z)\Psi(Y,Z)^\top = YR^\star(R^\star)^\top Y^\top = YY^\top. \tag{A.28}$$

If C is singular, define the *set* of optimal rotations $\mathcal{R}(Y,Z) = \arg\min_{R\in O(d_S)} \|YR - Z\|_F^2$ and resolved targets $\mathfrak{P}(Y,Z) = \{YR : R \in \mathcal{R}(Y,Z)\}$. Left/right multiplication by Q bijects the optimizer sets:

$$\mathcal{R}(YQ,Z) = Q^\top\mathcal{R}(Y,Z), \qquad \mathcal{R}(Y,ZQ) = \mathcal{R}(Y,Z)Q. \tag{A.29}$$

Consequently, $\mathfrak{P}(YQ,Z) = \mathfrak{P}(Y,Z)$ and $\mathfrak{P}(Y,ZQ) = \{PQ : P \in \mathfrak{P}(Y,Z)\}$. For every selected $R \in \mathcal{R}(Y,Z)$, the Gram identity still holds. An arbitrary numerical completion of null singular directions need not respect the single-valued equivariances: those are guaranteed as stated only when the polar factor is unique.

A.6 Endpoint Gradients and Student Dependence

Theorem 1 quantifies the headroom in initial endpoint loss. We next note that equivalent realizations can also produce different local optimization directions.

Let $h_i \in \mathbb{R}^{d_S}$ denote the pre-normalized student output as a column vector, so z_i is the transpose of row i of Z:

$$z_i = \frac{h_i}{\|h_i\|_2}. \tag{A.30}$$

For a target realization $\tau_i = (y_iR)^\top$, with y_i denoting row i of Y, define

$$\ell_i(R) = 1 - z_i^\top\tau_i. \tag{A.31}$$

The Jacobian of normalization is

$$\frac{\partial z_i}{\partial h_i} = \frac{1}{\|h_i\|_2}\left(I - z_iz_i^\top\right). \tag{A.32}$$

Since $\nabla_{z_i}\ell_i = -\tau_i$, the chain rule gives

$$\nabla_{h_i}\ell_i(R) = -\frac{1}{\|h_i\|_2}\left(I - z_iz_i^\top\right)\tau_i. \tag{A.33}$$

Different R generally produce different τ_i, and hence different projected update directions, even though $(YR)(YR)^\top = YY^\top$. The qualifier *generally* is necessary: special target and student configurations may yield the same projected gradient for distinct realizations.

Proof of Corollary 1. Let YR_1 and YR_2 be two distinct members of $[Y]$:

$$YR_1 \neq YR_2. \tag{A.34}$$

Consider the valid normalized student state

$$Z_1 = YR_1. \tag{A.35}$$

For this state,

$$\mathcal{L}_{Z_1}(R_1) = 0. \tag{A.36}$$

Because $YR_1 \neq YR_2$, there exists at least one row i for which $y_iR_1 \neq y_iR_2$. Both rows have unit norm, so $\langle y_iR_1, y_iR_2\rangle < 1$, and therefore

$$\mathcal{L}_{Z_1}(R_2) > 0. \tag{A.37}$$

Thus, R_1 is preferred to R_2 for the student state Z_1.

Conversely, choose

$$Z_2 = YR_2. \tag{A.38}$$

Then

$$\mathcal{L}_{Z_2}(R_2) = 0 < \mathcal{L}_{Z_2}(R_1). \tag{A.39}$$

Hence, neither representative is uniformly preferred across student states. Since the argument applies to any two distinct members of the orbit, there is no universally optimal realization determined independently of the student.

The conclusion is intentionally narrower than claiming that coordinates are intrinsically or uniquely defined by the student. It states that if coordinate resolution is defined by minimizing direct endpoint mismatch, the preferred representative necessarily depends on the student state.

A.7 Corpus-Level versus Minibatch Coordinate Resolution

ORBIT-KD resolves one shared realization for the fit corpus. We show why independently resolving different minibatches does not, in general, define a single member of the original corpus-level orbit. Partition Y into K blocks,

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_K \end{bmatrix}, \tag{A.40}$$

and suppose each block is independently transformed by $R_k \in O(d_S)$. The resulting representation is

$$\widetilde{Y} = \begin{bmatrix} Y_1 R_1 \\ Y_2 R_2 \\ \vdots \\ Y_K R_K \end{bmatrix}. \tag{A.41}$$

For every block k,

$$(Y_k R_k)(Y_k R_k)^\top = Y_k Y_k^\top, \tag{A.42}$$

so all within-block pairwise relations are preserved.

For two different blocks $a \neq b$, however,

$$(\widetilde{Y}\widetilde{Y}^\top)_{ab} = Y_a R_a R_b^\top Y_b^\top, \tag{A.43}$$

whereas

$$(YY^\top)_{ab} = Y_a Y_b^\top. \tag{A.44}$$

Their difference is

$$\Delta_{ab} = Y_a \left(R_a R_b^\top - I\right) Y_b^\top. \tag{A.45}$$

The diagonal blocks of $\widetilde{Y}\widetilde{Y}^\top - YY^\top$ are zero, while the off-diagonal blocks occur in transpose pairs. Therefore,

$$\left\|\widetilde{Y}\widetilde{Y}^\top - YY^\top\right\|_F^2 = 2\sum_{a<b} \left\|Y_a \left(R_a R_b^\top - I\right) Y_b^\top\right\|_F^2. \tag{A.46}$$

Thus, independent minibatch transformations preserve geometry inside each batch but generally alter relations between examples belonging to different batches. They correspond to a single global representative YR only when their relative transformations are consistent on the relevant data, including the special case in which the same effective transformation is shared across blocks. A single corpus-level transformation R instead gives $\widetilde{Y} = YR$ and therefore

$$\widetilde{Y}\widetilde{Y}^\top = YY^\top \tag{A.47}$$

for the complete corpus.

This establishes a geometric distinction between local and corpus-level coordinate resolution. It does not prove that a global fit must always yield better downstream generalization, since minibatch fitting also changes target stationarity, estimation noise, and optimization dynamics. The corresponding comparison is therefore evaluated empirically in Section 5.3.

A related limitation arises from fit rank. For a minibatch $\mathcal{B}$,

$$C_{\mathcal{B}} = Y_{\mathcal{B}}^\top Z_{\mathcal{B}} \in \mathbb{R}^{d_S \times d_S}$$

satisfies

$$\operatorname{rank}(C_{\mathcal{B}}) \leq |\mathcal{B}|. \tag{A.48}$$

Hence, when $|\mathcal{B}| < d_S$, the cross-covariance is necessarily rank deficient, leaving some orthogonal directions unconstrained by the minibatch Procrustes problem. A larger corpus-level fit need not suffer this dimensional rank limitation. This provides an additional reason to use one global coordinate reference, although the empirical comparison in Section 5.3 remains the relevant test of downstream effect.

B Reproducibility Details

This appendix gives the settings needed to reproduce the reported runs: the distillation corpora and evaluation benchmarks, the training configuration shared by all methods, and the ORBIT-KD procedure in the notation of Section 4.

B.1 Training Data

Table 2 summarizes the three unlabeled distillation corpora: 15K and 100K use training-split texts from three and nine benchmarks, respectively, while Wiki20K uses Wikipedia sentences only. Pair datasets contribute both sentences and repeated strings are retained; labels and evaluation examples never enter distillation.

Table 2: Distillation corpora. Source counts include both sentences from pair datasets.

Corpus	Source contributions	Total	Distinct
15K (main)	Emotion 5,000; WiC 5,000; STS-B 4,760	14,760	13,553
100K	Banking77 9,003; Tweet 23,726; Emotion 15,956; MRPC 7,052; SciTail 23,795; WiC 6,333; SICK 4,802; STS12 2,498; STS-B 9,196	102,361	102,361
Wiki20K	Wikipedia 20,000	20,000	19,882

B.2 Benchmarks

Table 3 lists nine benchmarks evaluated with frozen student embeddings. Classification uses a logistic-regression probe (scikit-learn, `max_iter` 200, fixed random state) fitted on training embeddings and scored by test macro-F1; pair classification uses cosine-similarity average precision, and STS uses cosine-similarity Spearman correlation. Scores are scaled by 100 and averaged without weighting. *Source* averages Emotion, WiC, and STS-B (marked *); *non-source* averages the other six. These names indicate 15K corpus membership, not distribution shift.

Table 3: Evaluation benchmarks. Sizes count examples for classification and sentence pairs otherwise. Classification training labels are used only to fit the logistic-regression probes; training texts may also appear in the task-adapted distillation corpora. Validation splits select learning rates, and all reported scores use the test splits.

Benchmark	Family	Metric	Labels	Train	Validation	Test
Banking77	Classification	Macro-F1	77 intents	9,003	1,000	3,076
Tweet	Classification	Macro-F1	3 sentiments	23,726	3,006	3,432
Emotion*	Classification	Macro-F1	6 emotions	15,956	1,988	1,986
MRPC	Pair classification	AP	paraphrase	—	408	1,725
SciTail	Pair classification	AP	entailment	—	1,304	2,126
WiC*	Pair classification	AP	same sense	—	638	1,400
SICK	STS	Spearman	relatedness	—	500	4,927
STS12	STS	Spearman	similarity	—	734	3,108
STS-B*	STS	Spearman	similarity	—	1,500	1,379

B.3 Training Configuration

Models. The three teacher–student pairs (Hugging Face checkpoints) are (a) Qwen/Qwen3-Embedding-4B $\rightarrow$ google-bert/bert-base-uncased ($2560 \rightarrow 768$; 109M-parameter student), (b) BAAI/bge-m3 $\rightarrow$ nreimers/MiniLMv2-L6-H768-distilled-from-BERT-Base ($1024 \rightarrow 768$; 66M), and (c) Qwen/Qwen3-Embedding-0.6B $\rightarrow$ nreimers/MiniLMv2-L6-H384-distilled-from-BERT-Base ($1024 \rightarrow 384$; 22M). The Qwen teachers use last-token pooling and BGE-M3 uses `[CLS]`;

every student uses its final-layer [CLS] state for training and evaluation. Inputs are truncated to 256 tokens, and all sentence embeddings are row-normalized before target construction or cosine-based evaluation.

Optimization. All main comparisons use seeds 42, 43, and 44, batch size 128, and five epochs. Students are loaded in float32 and trained with CUDA automatic mixed precision. Except for TALAS, every method uses AdamW with a cosine schedule, a method-specific warm-up ratio, and a minimum learning rate of 2×10^{-6}; TALAS uses adaptive SAM ($\rho = 0.05$) over AdamW, and Jasper keeps its two-stage schedule of two plus three epochs. Learning rates are selected on validation data and then held fixed across the three configurations (Table 4).

Table 4: Training hyperparameters of the main experiments.

Method	Epochs	Batch size	Learning rate	Optimizer	Warm-up
RKD	5	128	7×10^{-5}	AdamW	0.06
Jasper	$2 + 3$	128	5×10^{-5}	AdamW	0.06
CDM	5	128	2×10^{-5}	AdamW	0.06
DSKD	5	128	2×10^{-5}	AdamW	0.06
EMO	5	128	1×10^{-5}	AdamW	0.10
TALAS	5	128	2×10^{-5}	AdamW + adaptive SAM	0.06
ORBIT-KD	5	128	7×10^{-5}	AdamW	0.06

B.4 Pseudo-Algorithm

Algorithm 1 states the complete procedure of Section 4. Target construction and refreshes lie outside the differentiated update, and only the student encoder is kept after training.

Algorithm 1 ORBIT-KD training

Require: corpus $\mathcal{C}$ of size N, frozen teacher f_T, student $f_S(\cdot;\theta)$, fit set $\mathcal{G} \subseteq \mathcal{C}$, epochs E, weight λ_{H_0}

$T \leftarrow \mathrm{norm}_{\mathrm{row}}(f_T(\mathcal{C}))$ ▷ cached once
$\bar{\mathbf{t}} \leftarrow \frac{1}{N} T^\top \mathbf{1}$; $\widetilde{T} \leftarrow T - \mathbf{1}\bar{\mathbf{t}}^\top$ ▷ Eq. 10
$P_{\mathrm{PCA}} \leftarrow$ leading d_S right singular vectors of $\widetilde{T}$ ▷ Eq. 11
$Y \leftarrow \mathrm{norm}_{\mathrm{row}}(T P_{\mathrm{PCA}})$ ▷ Eq. 12
procedure RESOLVE(θ)
    $Z_{\mathcal{G}} \leftarrow \mathrm{norm}_{\mathrm{row}}(f_S(\mathcal{G};\theta))$ ▷ no gradient
    $U\Sigma V^\top \leftarrow Y_{\mathcal{G}}^\top Z_{\mathcal{G}}$; $R^\star \leftarrow UV^\top$ ▷ Eq. 15
    $\boldsymbol{\tau}_i \leftarrow (\mathbf{y}_i R^\star)^\top$ for all $i \in \mathcal{C}$
end procedure
RESOLVE(θ) ▷ pretrained student
for $e = 1, \ldots, E$ **do**
    for minibatch $\mathcal{B} \subset \mathcal{C}$ of size B **do**
        $\mathbf{z}_i \leftarrow \mathrm{norm}(f_S(x_i;\theta))$ for $i \in \mathcal{B}$
        $\mathcal{L}_{\mathrm{end}} \leftarrow \frac{1}{B}\sum_{i\in\mathcal{B}}(1 - \langle \mathbf{z}_i, \boldsymbol{\tau}_i \rangle)$ ▷ Eq. 16
        $q^T \leftarrow$ SORTEDMSTWEIGHTS($\{\mathbf{t}_i\}_{i\in\mathcal{B}}$); $q^S \leftarrow$ SORTEDMSTWEIGHTS($\{\mathbf{z}_i\}_{i\in\mathcal{B}}$)
        $\mathcal{L}_{H_0} \leftarrow \frac{1}{B-1}\sum_{j=1}^{B-1}(q^S_{(j)} - q^T_{(j)})^2$ ▷ Eq. 18
        Update θ with AdamW on $\mathcal{L}_{\mathrm{end}} + \lambda_{H_0}\mathcal{L}_{H_0}$ ▷ Eq. 19
    end for
    if $e < E$ **then**
        RESOLVE(θ) ▷ refresh with the current student
    end if
end for
return $f_S(\cdot;\theta)$

SORTEDMSTWEIGHTS builds a minimum spanning tree over chord distances $d(\mathbf{u}, \mathbf{v}) = \sqrt{2 - 2\langle \mathbf{u}, \mathbf{v} \rangle}$ in the native space of its input and returns the $B - 1$ edge weights in ascending order. Every minimum spanning tree has the same sorted weights, so ties do not change the loss.

C ADDITIONAL COORDINATE-RESOLUTION ANALYSIS

This appendix expands the controlled evidence of Section 5.3 with six experiments. We cross reduction operators with coordinate treatments (Appendix C.1); vary the student reference and the scope of the coordinate fit (Appendix C.2); replace the fixed orthogonal target with learned or unconstrained interfaces (Appendix C.3); decompose the objective into endpoint and structural terms (Appendix C.4); sweep the structural weight and fit-set size (Appendix C.5); and probe the mechanism behind the gain (Appendix C.6). Unless stated otherwise, every run uses configuration (c), the 15K corpus, batch size 128, five epochs, a rank-384 PCA reduction, the default objective with $\lambda_{H_0} = 0.5$, and seeds 42, 43, and 44. We report the nine-task overall average together with the source and non-source averages defined in Appendix B.2; error bars and $\pm$ give the sample standard deviation over seeds.

C.1 REDUCTION AND COORDINATES

Implementation Details. Figure 7 crosses the two decisions isolated in Section 3. The reduction operator is rank-k PCA with $k \in \{64, 128, 256, 384\}$, one of three independent Haar-random 384-dimensional subspaces, or rank-384 uncentered SVD; rank-k targets are zero-padded to width 384. Each reduction is then used as reduced, rotated by a fixed Haar-random matrix, or resolved by the student-conditioned Procrustes fit with epoch-wise refresh. Retained energy is measured on the uncentered normalized cache as $\mathcal{E}(P) = \|TB\|_F^2 / \|T\|_F^2$, where B is an orthonormal basis of $\mathrm{span}(P)$.

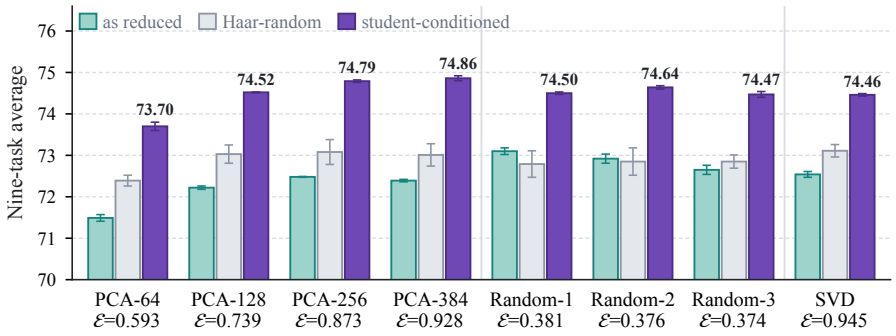


Figure 7: Reduction operators crossed with coordinate treatments. Tick labels give each reducer's retained energy $\mathcal{E}$; values are printed for student-conditioned targets. Separators group the PCA ranks, the random subspaces, and uncentered SVD.

Coordinates matter more than the reducer. Student-conditioned coordinates are best for every reducer and add 1.40–2.47 points over the as-reduced target. Once coordinates are resolved, the five full-rank reducers span only 0.40 points (74.46–74.86). A Haar-random rotation, by contrast, is not a reliable improvement: it helps the PCA targets by up to 0.90 points but lowers Random-1 by 0.31. The gain therefore comes from the student-matched frame, not from rotating the reduced basis as such.

Retained energy does not predict the resolved score. The random subspaces retain only 0.374–0.381 of teacher energy yet reach 74.47–74.64 after resolution, while uncentered SVD retains 0.945 and reaches 74.46. Within PCA, lowering the rank from 384 to 64 costs 1.16 points after resolution. Rank-d_S PCA remains the default because it leads the resolved column, but its 0.22-point margin over the next-best full-rank reducer is far smaller than the coordinate effect.

C.2 Coordinate Reference and Scope

Implementation Details. Table 5 holds the rank-384 PCA representation fixed. Panel (a) changes only the student embeddings used as the Procrustes reference: the matched student being trained, the matched student with its rows randomly permuted against the targets (shuffled correspondence), or an isotropic random point cloud of the same shape that carries no student information (unrelated). Panel (b) keeps the matched reference and varies the scope and schedule of the fit: an independent fit on every minibatch, one corpus-level fit before training, the same fit with one refresh, or the default epoch-wise refresh. Figure 8 measures how each scope changes the target Gram matrix on eight random minibatches of the pre-KD student.

Table 5: Coordinate-reference and resolution-scope controls in configuration (c), all on the same rank-384 PCA representation. Δ is the overall gain over the as-reduced target in points. †The matched-student and epoch-wise rows are the same ORBIT-KD run.

Treatment	Source ↑	Non-source ↑	Overall ↑	Δ
Reference-free baselines				
As reduced	65.64 ± 0.04	75.77 ± 0.06	72.39 ± 0.03	—
Haar-random	66.82 ± 0.44	76.11 ± 0.18	73.01 ± 0.27	+0.62
(a) Coordinate reference (corpus-level, epoch-wise)				
Unrelated student	67.28 ± 0.47	76.08 ± 0.06	73.15 ± 0.20	+0.76
Shuffled correspondence	67.57 ± 0.36	77.23 ± 0.20	74.01 ± 0.21	+1.62
Matched student†	**68.80** ± 0.16	**77.89** ± 0.01	**74.86** ± 0.06	**+2.47**
(b) Resolution scope and schedule (matched student)				
Per-minibatch fits	65.75 ± 0.20	76.19 ± 0.09	72.71 ± 0.07	+0.32
Corpus-level, initial only	68.24 ± 0.10	77.80 ± 0.04	74.61 ± 0.04	+2.22
Corpus-level, one refresh	68.34 ± 0.02	77.85 ± 0.03	74.68 ± 0.01	+2.29
Corpus-level, epoch-wise†	**68.80** ± 0.16	**77.89** ± 0.01	**74.86** ± 0.06	**+2.47**

A matched reference gives the largest gain. The matched student adds 2.47 points over the reduced target, shuffled correspondence adds 1.62, and the unrelated reference adds 0.76, barely above the 0.62 of a Haar rotation: a Procrustes fit without student information is just another arbitrary orientation. Shuffling keeps the student's population-level statistics but breaks example pairing, so its intermediate score indicates that both population structure and example-level correspondence contribute. The gain is larger on source tasks (+3.16) than on non-source tasks (+2.12), but it is positive on both.

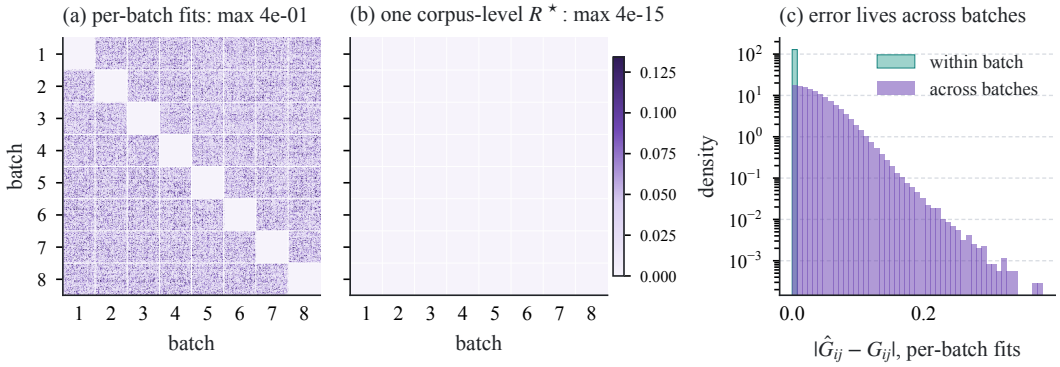


Figure 8: Gram-matrix change $|\hat{G}_{ij} - G_{ij}|$ on eight random minibatches of 128 examples. **(a)** Each minibatch receives its own Procrustes fit. **(b)** One corpus-level $R^\star$ is shared by all examples. **(c)** Distribution of the per-minibatch errors within and across minibatches.

One global frame matters more than frequent refresh. Independent minibatch fits reach only 72.71, 1.90 points below a single corpus-level fit (74.61); one refresh and the epoch-wise schedule add a further 0.07 and 0.25. Figure 8 shows why scope matters, as proved in Appendix A.7: minibatch fits preserve within-minibatch similarities to rounding error (4×10^{-15}) but rotate minibatches differently, changing cross-minibatch similarities by a median of 0.03 and up to 0.38, whereas one shared $R^\star$ leaves the whole Gram matrix unchanged. Because later training minibatches mix examples across these boundaries, this inconsistency is a plausible cause of the deficit.

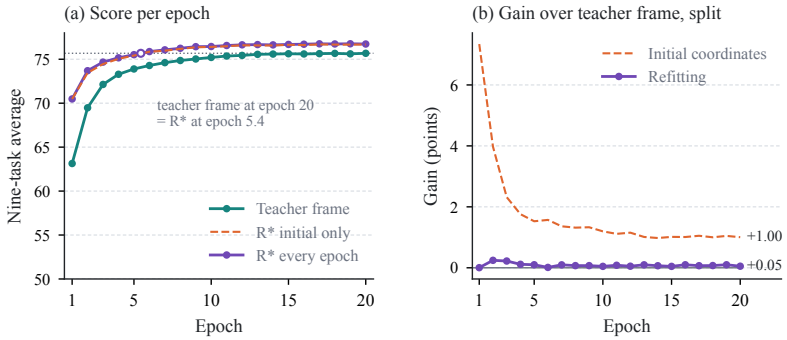
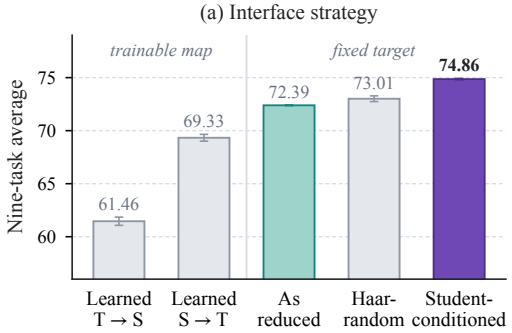


Figure 9: Twenty-epoch schedule ablation with the complete ORBIT-KD objective. **(a)** Nine-task averages for unreoriented, initial-only, and epoch-wise targets. **(b)** Gains from the initial fit and epoch-wise refits.

C.3 Interface Constraints

Implementation Details. Figure 10 replaces the fixed orthogonal target with less constrained interfaces. Panel (a) compares two trainable linear maps, teacher-to-student ($d_T \to d_S$) and student-to-teacher ($d_S \to d_T$), trained jointly with the student, against three fixed targets built from the same PCA reduction. Panel (b) fits the fixed map to the matched student either as an orthogonal Procrustes rotation or as an unconstrained least-squares map, under initial-only and epoch-wise schedules. Initial cosine is the mean cosine between targets and pre-KD student embeddings; Gram change is the relative Frobenius change of the target Gram matrix; 10-NN overlap is the fraction of ten-nearest-neighbor sets preserved; and effective rank is measured on training minibatches. The learned-interface and least-squares rows come from separate matched sweeps with the same model pair, corpus, seeds, and budget, so we treat them as supporting evidence.



(b) Map constraint

	Orthogonal	Least squares
Initial cosine	0.65	**0.99**
Gram change	**0**	1.86 → 0.98
10-NN overlap	**100%**	24% → 43%
Effective rank	**104.6**	18.8
Overall, initial fit	**74.61**	70.21
Overall, epoch-wise	**74.86**	73.16

Figure 10: Interface controls. **(a)** Trainable maps versus fixed endpoint targets. **(b)** Orthogonal versus least-squares fixed maps fitted to the matched student; arrows give the value before and after epoch-wise refits.

Trainable interfaces underperform fixed targets. Both learned maps fall below even the unrotated PCA target: teacher-to-student reaches 61.46 and student-to-teacher 69.33, respectively 13.40

and 5.53 points below ORBIT-KD. A plausible reason is that a learned map lets the target drift with the student and absorbs part of the supervision, whereas a fixed target supervises the native output directly.

Geometry preservation matters more than proximity. Least squares starts much closer to the student (initial cosine 0.99 versus 0.65), yet trails the orthogonal map by 4.40 points with an initial fit and 1.70 after epoch-wise refits. It achieves that proximity by distorting the target: the Gram matrix changes by 1.86 in relative norm, only 24% of ten-nearest-neighbor sets survive, and effective minibatch rank falls from 104.6 to 18.8. Refitting partly restores the geometry but not the score, which supports the design choice of resolving the frame while keeping the retained teacher geometry exact.

C.4 Objective Decomposition

Implementation Details. Table 6 crosses the endpoint target (none, unrotated PCA, or student-conditioned) with a coordinate-invariant structural objective computed in the native spaces: the default H_0 loss or Gram matching (mean squared error between the off-diagonal within-batch cosine Gram matrices). Each structural term uses weight 0.5, and all other settings follow the default configuration. The row without any loss keeps the training loop but provides no distillation signal, so it essentially recovers the pre-KD student.

Table 6: Endpoint-target and structural-objective ablation in configuration (c). Δ is the overall change in points relative to the resolved endpoint without structural support (marked ref.). The default ORBIT-KD objective is marked *.

Endpoint target	Structural objective	Source ↑	Non-source ↑	Overall ↑	Δ
None	None	41.61 ± 0.00	58.19 ± 0.03	52.66 ± 0.02	-21.59
	H_0 persistence	57.55 ± 0.47	67.53 ± 0.60	64.21 ± 0.48	-10.04
PCA-reduced	None	62.35 ± 0.22	72.76 ± 0.04	69.29 ± 0.07	-4.96
Student-conditioned	None	67.61 ± 0.09	77.57 ± 0.06	74.25 ± 0.01	ref.
	Gram matching	67.64 ± 0.15	76.94 ± 0.05	73.84 ± 0.03	-0.41
	H_0 persistence*	**68.80** ± 0.16	**77.89** ± 0.01	**74.86** ± 0.06	$\mathbf{+0.61}$

The resolved endpoint carries most of the gain. H_0 alone reaches 64.21, and the unrotated PCA endpoint reaches 69.29. Resolving the same endpoint raises the score to 74.25 without structural support, a 4.96-point gain, and adding H_0 contributes a further 0.61. The dominant improvement therefore comes from coordinate-resolved endpoint supervision rather than from the auxiliary term.

C.5 Sensitivity Analysis

Implementation Details. Figure 11 varies one factor at a time around the default. Panel (a) sweeps the structural weight $\lambda_{H_0} \in \{0, 0.1, 0.3, 0.5, 1.0\}$. Panel (b) fits the coordinate transformation on 2,048, 4,096, or 8,192 rows evenly spaced over the corpus, or on all 14,760; the same fit rows are reused by every refresh.

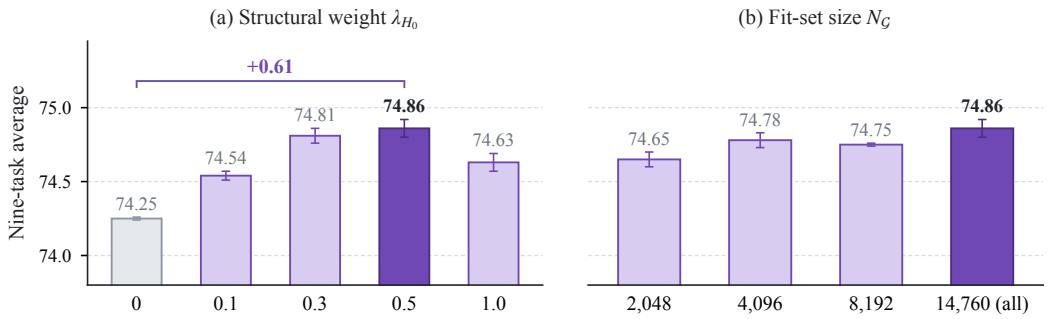


Figure 11: Sensitivity to **(a)** the structural weight λ_{H_0} and **(b)** the coordinate fit-set size $N_{\mathcal{G}}$. Dark bars mark the default; $\lambda_{H_0} = 0$ is the endpoint-only objective.

ORBIT-KD is stable across practical settings. Every positive structural weight from 0.1 to 1.0 improves over the endpoint-only objective, giving 74.54–74.86, with the default 0.5 at the top. The largest weight costs 0.23 points, with the larger drop on source tasks (-0.42 versus -0.14), so structural support helps as a complement but should not dominate the endpoint loss.

A small fit set is enough. A 2,048-row fit, 14% of the corpus, already reaches 74.65, and the four fit sizes span only 0.21 points. Because the fit solves for a single $d_S \times d_S$ rotation, a few thousand paired rows determine it well. The experiment does not establish invariance to arbitrary weights or much smaller fit sets.

C.6 Mechanism Diagnostics

Implementation Details. Figure 3 in Section 5.3 and Figures 12 and 13 probe why coordinate resolution helps, using the rank-384 PCA representation Y and the pre-KD student embeddings Z of the fit corpus. Panels computed from Y and Z alone involve no training. Panels reporting nine-task averages come from a separate five-epoch sweep in which every orthogonal target is fixed at initialization (no refresh), with seeds 42, 43, and 44; its anchors were retrained for this analysis and agree with retrained copies to within 0.04 points, so they may differ from the tables above by run-to-run noise. Figure 12 uses a twenty-epoch run with the complete objective and seed 42, and Figure 13 adds BERT-base from configuration (a).

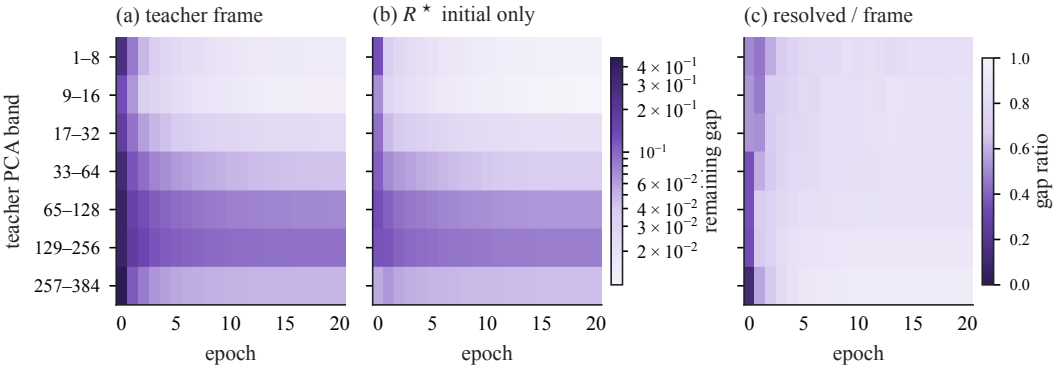


Figure 12: Remaining endpoint gap by teacher-PCA band over twenty epochs, with each student measured against the fixed targets it was trained on. The residual $\boldsymbol{\tau}_i - \mathbf{z}_i$ is expressed in the teacher's PCA coordinates and its squared norm summed within each band. **(a)** Unreoriented targets. **(b)** Targets resolved once by $R^\star$. **(c)** Ratio of (b) to (a); values below one mean that the resolved student is closer to its own targets in that band.

The head start is spread across the spectrum. Resolution lowers the initial gap in every band, by factors of 0.12 to 0.53, and most strongly in the low-variance bands where the unreoriented frame is furthest from the student (Figure 12). Training closes much of both gaps within the first few epochs, but the resolved student stays closer to its targets in every band through epoch 20, with ratios between 0.83 and 0.93; its total gap falls from 0.69 to 0.27, compared with 2.03 to 0.31 without resolution. For both targets, the largest remaining residual lies in bands 65–256, so neither schedule fits the mid-spectrum completely within this budget.

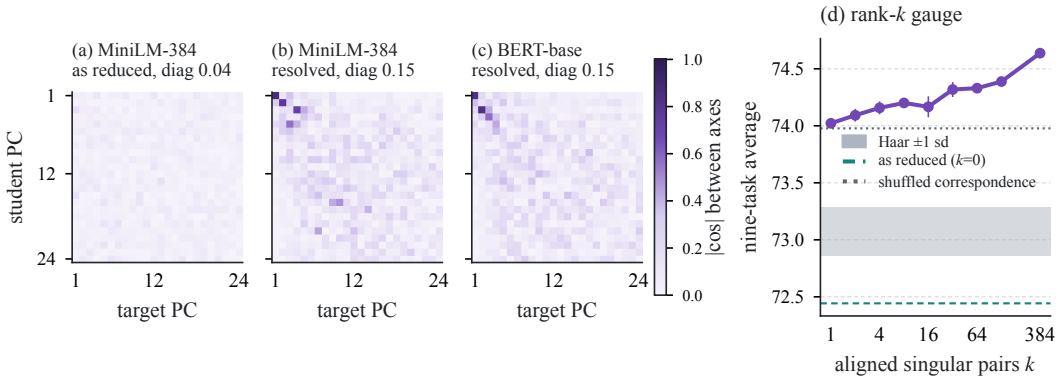


Figure 13: Alignment between principal axes and the rank-k gauge. **(a–c)** Absolute cosine between the top 24 principal axes of the pre-KD student and those of its targets, before and after resolution, for configuration (c) and for BERT-base in configuration (a). Titles report the mean diagonal. **(d)** Final nine-task average when only the top k singular pairs of $C = Y^\top Z$ are aligned and the complement stays as close to the reduced frame as orthogonality allows; $k = 0$ is the reduced frame and $k = 384$ is $R^\star$. Error bars show the standard deviation over three seeds.

Resolution aligns the dominant axes first. Before resolution, the target's principal axes are essentially unrelated to the student's; after resolution, the leading axes line up, raising the mean diagonal cosine from 0.04 to 0.15 for MiniLM-384 and from 0.03 to 0.15 for BERT-base (Figure 13). In the rank-k sweep, aligning a single singular pair already raises the score from 72.44 to 74.02, matching shuffled correspondence (73.98), which likewise retains only population-level structure; the remaining 0.62 points accumulate gradually up to full rank. On this anisotropic student the leading singular pair is close to the shared mean direction, so the largest single gain comes from matching that dominant direction, while the example-level correspondence exploited by the full $R^\star$ adds a smaller, distributed improvement.

D Additional Evaluation and Efficiency

This appendix reports five sets of experiments beyond the compact main-paper summaries. We give the full task-level results behind Table 1 for the 15K corpus (Appendix D.1) and the 100K corpus (Appendix D.2); distill multi-billion-parameter teachers into a much smaller student (Appendix D.3); replace the task-adapted corpus with general Wikipedia text (Appendix D.4); and give the complete optimization-cost measurements (Appendix D.5). All experiments use the frozen-embedding protocol and metrics of Appendix B.2 and report mean ± sample standard deviation over seeds 42, 43, and 44.

D.1 Task-Level Results on the 15K Corpus

Implementation Details. Table 7 reports all nine metrics for the three teacher–student pairs specified in Appendix B.3, trained on the 14,760-text corpus with the same settings. Each block also lists the undistilled teacher and the pre-KD student as reference points; source, non-source, and overall averages follow Appendix B.2.

Table 7: Nine-task results on the 15K distillation corpus. Distilled students report mean ± sample standard deviation over three seeds; best and second-best results are bold and underlined. ORBIT-KD uses $\lambda_{H_0} = 0.5$.

	Classification (F1)			Pair Classification (AP)			STS (Spearman)			Averages		
Method	Banking77	Tweet	Emotion	MRPC	SciTail	WiC	SICK	STS12	STS-B	Src	Non-src	All
(a) Qwen3-Embedding-4B → BERT-base (109M)												
Teacher	93.26	71.04	69.08	84.80	91.77	66.79	83.43	82.55	86.87	74.25	84.48	81.07
Student (pre-KD)	84.33	67.08	47.79	77.36	67.74	59.22	42.43	21.54	20.29	42.43	60.08	54.20
RKD	89.76 ± 0.11	70.90 ± 0.31	59.98 ± 0.05	81.94 ± 0.26	75.59 ± 0.13	69.02 ± 0.18	72.93 ± 0.07	67.84 ± 0.18	73.72 ± 0.08	67.57 ± 0.05	76.49 ± 0.07	73.52 ± 0.06
Jasper	88.48 ± 0.30	68.94 ± 0.17	53.69 ± 0.50	83.95 ± 0.10	75.39 ± 0.54	**71.93** ± 0.09	69.87 ± 0.64	64.39 ± 0.21	70.86 ± 0.36	65.49 ± 0.17	75.17 ± 0.13	71.95 ± 0.06
CDM	88.13 ± 0.23	69.03 ± 0.20	52.83 ± 0.50	83.00 ± 0.06	77.10 ± 0.27	71.84 ± 0.09	67.89 ± 0.37	61.06 ± 0.37	69.19 ± 0.42	64.62 ± 0.28	74.37 ± 0.12	71.12 ± 0.18
DSKD	87.97 ± 0.16	68.90 ± 0.28	53.07 ± 0.31	83.03 ± 0.15	76.98 ± 0.59	71.86 ± 0.08	67.77 ± 0.66	61.11 ± 0.26	69.55 ± 0.12	64.83 ± 0.12	74.29 ± 0.14	71.14 ± 0.10
EMO	86.89 ± 0.07	68.99 ± 0.11	52.02 ± 1.09	79.74 ± 0.12	76.37 ± 0.35	69.62 ± 0.16	67.89 ± 0.09	56.46 ± 0.47	64.54 ± 0.50	62.06 ± 0.49	72.72 ± 0.15	69.17 ± 0.24
TALAS	90.53 ± 0.09	73.77 ± 0.21	68.47 ± 0.34	**85.38** ± 0.05	**81.83** ± 0.21	69.59 ± 0.14	76.79 ± 0.06	70.79 ± 0.19	**78.50** ± 0.08	**72.19** ± 0.07	79.85 ± 0.07	77.29 ± 0.04
ORBIT-KD	**90.60** ± 0.21	**73.86** ± 0.10	**69.48** ± 0.22	85.31 ± 0.11	80.23 ± 0.13	68.19 ± 0.12	**77.80** ± 0.14	**75.15** ± 0.07	78.44 ± 0.16	72.04 ± 0.08	**80.49** ± 0.03	**77.67** ± 0.03
(b) BGE-M3 → MiniLMv2-H768 (66M)												
Teacher	93.49	73.75	69.03	85.80	91.87	61.46	79.18	78.74	84.86	71.78	83.81	79.80
Student (pre-KD)	81.18	71.30	48.85	79.66	65.06	60.89	49.46	26.76	24.12	44.62	62.24	56.36
RKD	89.14 ± 0.09	73.29 ± 0.09	62.23 ± 0.27	84.35 ± 0.06	78.74 ± 0.11	65.02 ± 0.10	73.20 ± 0.19	65.02 ± 0.29	71.97 ± 0.38	66.41 ± 0.07	77.29 ± 0.06	73.66 ± 0.04
Jasper	86.82 ± 0.28	70.06 ± 0.12	56.61 ± 0.74	83.02 ± 0.12	74.63 ± 0.12	**69.72** ± 0.05	69.08 ± 0.20	63.14 ± 0.28	67.10 ± 0.10	64.47 ± 0.27	74.46 ± 0.05	71.13 ± 0.11
CDM	87.82 ± 0.19	69.99 ± 0.09	58.32 ± 0.51	83.44 ± 0.07	75.56 ± 0.40	69.70 ± 0.02	65.93 ± 0.16	60.84 ± 0.34	65.31 ± 0.15	64.45 ± 0.13	73.93 ± 0.06	70.77 ± 0.02
DSKD	87.70 ± 0.16	70.35 ± 0.32	59.71 ± 0.61	82.99 ± 0.18	75.58 ± 0.12	69.59 ± 0.04	65.36 ± 0.37	61.21 ± 0.31	65.02 ± 0.18	64.77 ± 0.27	73.87 ± 0.10	70.83 ± 0.11
EMO	87.54 ± 0.18	71.03 ± 0.14	58.52 ± 0.30	82.46 ± 0.18	74.73 ± 0.58	69.20 ± 0.22	64.71 ± 0.55	56.89 ± 0.27	61.25 ± 0.03	62.99 ± 0.19	72.89 ± 0.10	69.59 ± 0.12
TALAS	89.83 ± 0.09	**74.70** ± 0.09	61.65 ± 0.23	**86.62** ± 0.05	80.85 ± 0.22	66.72 ± 0.07	75.40 ± 0.10	67.08 ± 0.31	77.08 ± 0.01	68.48 ± 0.07	79.08 ± 0.08	75.55 ± 0.06
ORBIT-KD	**90.17** ± 0.05	73.74 ± 0.24	**64.83** ± 0.72	86.42 ± 0.09	**80.97** ± 0.08	65.47 ± 0.09	**76.77** ± 0.03	**68.27** ± 0.04	**77.63** ± 0.12	**69.31** ± 0.21	**79.39** ± 0.06	**76.03** ± 0.11
(c) Qwen3-Embedding-0.6B → MiniLMv2-H384 (22M)												
Teacher	92.51	71.28	65.80	83.38	90.09	66.84	80.65	77.11	84.59	72.41	82.50	79.14
Student (pre-KD)	67.03	69.45	42.90	78.39	64.55	59.58	47.60	21.96	22.08	41.52	58.16	52.62
RKD	86.76 ± 0.07	69.87 ± 0.21	56.31 ± 0.35	81.70 ± 0.03	76.03 ± 0.20	68.56 ± 0.22	69.21 ± 0.18	64.26 ± 0.19	70.25 ± 0.21	65.04 ± 0.20	74.64 ± 0.09	71.44 ± 0.10
Jasper	84.04 ± 0.04	68.73 ± 0.44	52.41 ± 0.28	81.50 ± 0.12	72.26 ± 0.32	**68.65** ± 0.16	64.64 ± 0.19	60.59 ± 0.12	64.27 ± 0.20	61.78 ± 0.13	71.96 ± 0.13	68.57 ± 0.10
CDM	84.74 ± 0.21	69.52 ± 0.20	53.76 ± 0.95	81.15 ± 0.14	72.86 ± 0.30	68.29 ± 0.11	63.98 ± 0.36	60.43 ± 0.22	63.52 ± 0.24	61.86 ± 0.43	72.11 ± 0.19	68.69 ± 0.27
DSKD	84.76 ± 0.10	69.56 ± 0.20	53.14 ± 0.29	81.15 ± 0.01	72.81 ± 0.09	68.33 ± 0.04	63.78 ± 0.06	60.24 ± 0.19	63.53 ± 0.18	61.67 ± 0.03	72.05 ± 0.06	68.59 ± 0.04
EMO	83.54 ± 0.31	69.31 ± 0.06	54.48 ± 0.27	80.87 ± 0.13	72.00 ± 0.16	67.89 ± 0.09	62.19 ± 0.16	57.04 ± 0.17	61.24 ± 0.23	61.20 ± 0.09	70.82 ± 0.07	67.62 ± 0.02
TALAS	86.39 ± 0.13	71.81 ± 0.11	58.40 ± 0.30	**84.25** ± 0.03	79.29 ± 0.11	66.82 ± 0.02	70.42 ± 0.10	68.08 ± 0.22	73.34 ± 0.10	66.19 ± 0.11	76.71 ± 0.04	73.20 ± 0.06
ORBIT-KD	**87.77** ± 0.07	**72.30** ± 0.19	**63.45** ± 0.43	84.19 ± 0.03	**81.56** ± 0.16	68.17 ± 0.05	**72.59** ± 0.03	**68.91** ± 0.03	**74.79** ± 0.05	**68.80** ± 0.16	**77.89** ± 0.01	**74.86** ± 0.06

ORBIT-KD leads in every configuration. ORBIT-KD attains the highest overall average for all three pairs, ahead of TALAS by 0.38, 0.48, and 1.66 points in configurations (a)–(c). Measured against the distance from the pre-KD student to the teacher, it closes 87%, 84%, and 84% of the gap, respectively, so the improvement holds whether the student is a 109M BERT-base or a 22M MiniLM.

The tightest bottleneck shows the broadest gain. For the 22M student in configuration (c), ORBIT-KD outperforms TALAS on eight of nine tasks; TALAS is higher only on MRPC. The task-level ordering varies for the other pairs: TALAS is higher on MRPC, SciTail, WiC, and STS-B in configuration (a), and on Tweet, MRPC, and WiC in configuration (b). Nevertheless, ORBIT-KD has the highest nine-task average in all three configurations.

D.2 Task-Level Results on the 100K Corpus

Implementation Details. Table 8 repeats the comparison on the 102,361-text corpus drawn from the training splits of all nine benchmarks (Table 2). Models, hyperparameters, and evaluation are unchanged; ORBIT-KD fits its coordinates on a fixed subset of 16,384 evenly spaced rows. Because this corpus changes both size and task coverage, it is not an isolated data-scaling test.

The overall lead persists with broader task coverage. ORBIT-KD again has the highest overall average for all three pairs, although its margin over TALAS narrows to 0.07, 0.27, and 1.13 points. In configuration (c), the larger corpus raises ORBIT-KD from 74.86 to 77.48 ($+2.62$) and TALAS from 73.20 to 76.35 ($+3.15$). Additional task-covering data thus benefits both methods and slightly narrows the gap, while the ordering remains unchanged.

D.3 Large Capacity Gaps

Implementation Details. Three teachers above 7B parameters with 4096-dimensional outputs, Qwen3-Embedding-8B, WeMM-Embedding-9B, and NV-Embed-v2, are each distilled into the same four-layer TinyBERT student with a 312-dimensional `[CLS]` embedding. Holding the student fixed isolates whether ORBIT-KD remains useful when capacity and width differ sharply. Qwen uses last-token pooling, and WeMM and NV-Embed-v2 use their native embedding interfaces. All methods train on the 15K corpus for five epochs with batch size 128, maximum length 256, and learning rate 5×10^{-5}. The four distillation methods (RKD, Jasper, TALAS, and ORBIT-KD) read cached sentence-level teacher embeddings, and SimCSE is a no-teacher control whose row is therefore shared across the three teachers. ORBIT-KD uses $\lambda_{H_0} = 0.5$ with epoch-wise refresh.

ORBIT-KD remains effective across extreme capacity gaps. Table 9 shows the highest nine-task average for all three teachers. Relative to the strongest competitor, TALAS, ORBIT-KD improves by 1.49 points with Qwen3-Embedding-8B, 1.49 with WeMM-Embedding-9B, and 0.43 with NV-Embed-v2, and it leads on seven, seven, and four of the nine tasks. Coordinate-resolved targets therefore stay competitive even when a 4096-dimensional, multi-billion-parameter teacher supervises a 312-dimensional student.

The distribution of the gain depends on the teacher. With NV-Embed-v2 the margin is narrower, and TALAS remains stronger on MRPC and the three STS benchmarks; with the Qwen and WeMM teachers, ORBIT-KD wins most tasks. The consistent average ordering supports robustness to teacher scale, while the task-level variation cautions against treating the capacity gap alone as an explanation of the gain.

D.4 General-Purpose Transfer

Implementation Details. We replace the task-adapted corpus with the Wiki20K corpus of 20,000 unlabeled Wikipedia sentences (Table 2), so no text from the benchmark training splits enters distillation. Qwen3-Embedding-0.6B is distilled into MiniLMv2-H384 for five epochs with batch size 128; baselines keep their main-experiment learning rates, and ORBIT-KD uses $\lambda_{H_0} = 0.5$, and epoch-wise refresh.

Table 8: Nine-task results with the 100K distillation corpus. Distilled students report mean ± standard deviation over three seeds; best and second-best results are bold and underlined.

Method	Classification (F1)			Pair Classification (AP)			STS (Spearman)			Avg.
	Banking77	Tweet	Emotion	MRPC	SciTail	WiC	SICK	STS12	STS-B	
(a) Qwen3-Embedding-4B → BERT-base (109M)										
Teacher	93.26	71.04	69.08	84.80	91.77	66.79	83.43	82.55	86.87	81.07
Student (pre-KD)	84.33	67.08	47.79	77.36	67.74	59.22	42.43	21.54	20.29	54.20
RKD	92.23 ± 0.15	71.96 ± 0.18	63.79 ± 0.40	85.00 ± 0.04	80.27 ± 0.38	66.31 ± 0.10	79.32 ± 0.02	71.72 ± 0.16	77.70 ± 0.20	76.48 ± 0.04
Jasper	90.02 ± 0.15	71.90 ± 0.22	60.14 ± 0.45	85.96 ± 0.05	78.82 ± 0.27	69.93 ± 0.13	76.53 ± 0.23	69.79 ± 0.12	77.22 ± 0.16	75.59 ± 0.10
CDM	89.16 ± 0.04	69.63 ± 0.23	53.74 ± 0.13	85.57 ± 0.08	71.22 ± 0.19	**70.04** ± 0.05	71.26 ± 0.29	66.97 ± 0.21	73.69 ± 0.38	72.36 ± 0.13
DSKD	88.87 ± 0.07	69.05 ± 0.09	54.02 ± 0.56	85.90 ± 0.05	70.02 ± 0.31	69.63 ± 0.21	71.07 ± 0.10	66.52 ± 0.17	73.52 ± 0.01	72.07 ± 0.03
EMO	84.96 ± 0.84	68.93 ± 0.65	54.80 ± 1.52	85.05 ± 0.04	65.75 ± 0.22	68.80 ± 0.13	62.00 ± 2.71	61.75 ± 1.85	69.54 ± 0.37	69.07 ± 0.89
TALAS	93.12 ± 0.13	**74.67** ± 0.38	71.42 ± 0.26	**86.30** ± 0.01	87.06 ± 0.12	68.09 ± 0.02	80.29 ± 0.03	76.34 ± 0.01	82.26 ± 0.05	79.95 ± 0.08
ORBIT-KD	**93.47** ± 0.12	74.18 ± 0.25	**72.50** ± 0.47	85.45 ± 0.08	**87.24** ± 0.12	66.57 ± 0.08	**81.85** ± 0.06	**76.40** ± 0.07	**82.52** ± 0.08	**80.02** ± 0.03
(b) BGE-M3 → MiniLMv2-H768 (66M)										
Teacher	93.49	73.75	69.03	85.80	91.87	61.46	79.18	78.74	84.86	79.80
Student (pre-KD)	81.18	71.30	48.85	79.66	65.06	60.89	49.46	26.76	24.12	56.36
RKD	91.68 ± 0.06	74.57 ± 0.40	62.45 ± 0.75	85.81 ± 0.08	79.81 ± 0.17	63.57 ± 0.13	76.91 ± 0.05	69.54 ± 0.11	75.56 ± 0.09	75.54 ± 0.10
Jasper	88.72 ± 0.41	73.43 ± 0.23	61.63 ± 0.29	86.39 ± 0.08	74.13 ± 0.31	66.98 ± 0.27	73.32 ± 0.07	68.30 ± 0.12	72.66 ± 0.14	73.95 ± 0.04
CDM	87.64 ± 0.23	70.82 ± 0.14	56.37 ± 0.57	85.67 ± 0.08	67.69 ± 0.23	66.89 ± 0.20	70.15 ± 0.48	66.68 ± 0.26	69.49 ± 0.15	71.27 ± 0.10
DSKD	86.69 ± 0.38	69.76 ± 0.22	56.98 ± 0.39	84.92 ± 0.07	66.75 ± 0.17	**67.23** ± 0.10	68.26 ± 0.45	66.28 ± 0.02	68.31 ± 0.35	70.58 ± 0.10
EMO	87.17 ± 0.59	71.07 ± 0.22	59.32 ± 0.02	84.25 ± 0.41	67.89 ± 0.92	66.09 ± 0.18	69.04 ± 0.38	62.26 ± 1.77	66.47 ± 0.40	70.40 ± 0.35
TALAS	92.44 ± 0.10	75.58 ± 0.15	65.13 ± 0.21	**87.09** ± 0.09	85.39 ± 0.14	64.39 ± 0.12	78.14 ± 0.06	72.56 ± 0.13	**81.03** ± 0.05	77.97 ± 0.03
ORBIT-KD	**92.56** ± 0.04	**76.30** ± 0.31	**65.62** ± 0.29	86.58 ± 0.05	**87.20** ± 0.06	63.43 ± 0.02	**79.23** ± 0.03	**72.69** ± 0.05	80.60 ± 0.13	**78.24** ± 0.05
(c) Qwen3-Embedding-0.6B → MiniLMv2-H384 (22M)										
Teacher	92.51	71.28	65.80	83.38	90.09	66.84	80.65	77.11	84.59	79.14
Student (pre-KD)	67.03	69.45	42.90	78.39	64.55	59.58	47.60	21.96	22.08	52.62
RKD	89.74 ± 0.04	72.34 ± 0.22	57.99 ± 0.52	84.13 ± 0.12	75.20 ± 0.14	**67.49** ± 0.06	74.48 ± 0.03	68.12 ± 0.15	72.55 ± 0.14	73.56 ± 0.09
Jasper	85.74 ± 0.22	70.25 ± 0.07	53.48 ± 0.28	84.43 ± 0.04	71.23 ± 0.05	66.51 ± 0.14	70.96 ± 0.28	63.02 ± 0.04	67.42 ± 0.24	70.34 ± 0.02
CDM	85.18 ± 0.08	69.14 ± 0.07	52.30 ± 0.26	84.02 ± 0.04	68.24 ± 0.23	66.57 ± 0.29	69.15 ± 0.13	63.05 ± 0.06	66.55 ± 0.10	69.36 ± 0.08
DSKD	85.04 ± 0.14	69.21 ± 0.13	52.85 ± 0.36	84.34 ± 0.10	67.33 ± 0.24	65.98 ± 0.12	69.01 ± 0.13	63.46 ± 0.21	66.42 ± 0.09	69.29 ± 0.09
EMO	85.08 ± 0.17	68.11 ± 0.10	53.96 ± 0.60	84.38 ± 0.16	65.48 ± 0.16	65.23 ± 0.17	68.61 ± 0.07	60.00 ± 0.82	65.38 ± 0.38	68.47 ± 0.15
TALAS	89.79 ± 0.03	73.87 ± 0.17	61.75 ± 0.61	**85.99** ± 0.03	82.07 ± 0.09	65.79 ± 0.04	75.67 ± 0.05	**73.91** ± 0.03	78.31 ± 0.04	76.35 ± 0.08
ORBIT-KD	**91.68** ± 0.09	**74.43** ± 0.21	**66.88** ± 0.36	85.00 ± 0.06	**83.46** ± 0.02	67.44 ± 0.04	**77.94** ± 0.06	72.17 ± 0.02	**78.36** ± 0.07	**77.48** ± 0.04

Table 9: Distillation from teachers above 7B parameters into TinyBERT-4L-312D. Results are mean ± sample standard deviation over three seeds. **Bold** and underlined values denote the best and second-best methods within each teacher.

Method	Banking77	Tweet	Emotion	MRPC	SciTail	WiC	SICK	STS12	STS-B	Avg
(a) Qwen3-Embedding-8B → TinyBERT										
SimCSE	83.25 ± 0.12	67.81 ± 0.13	47.81 ± 0.94	81.61 ± 0.23	71.89 ± 0.71	65.70 ± 0.15	66.89 ± 0.51	62.66 ± 0.25	65.33 ± 0.06	68.10 ± 0.18
RKD	84.15 ± 0.25	69.13 ± 0.03	51.12 ± 0.40	80.33 ± 0.20	71.16 ± 0.11	65.25 ± 0.12	66.47 ± 0.23	63.91 ± 0.21	66.39 ± 0.11	68.66 ± 0.06
Jasper	82.28 ± 0.08	67.41 ± 0.28	44.81 ± 0.20	80.30 ± 0.22	71.19 ± 0.16	**66.43** ± 0.06	66.85 ± 0.18	61.65 ± 0.42	64.00 ± 0.24	67.21 ± 0.09
TALAS	81.48 ± 0.25	72.00 ± 0.10	57.47 ± 0.50	**85.44** ± 0.06	73.61 ± 0.13	63.73 ± 0.15	69.60 ± 0.09	66.07 ± 0.22	70.41 ± 0.15	71.09 ± 0.02
ORBIT-KD	**86.02** ± 0.15	**72.32** ± 0.08	**58.38** ± 0.54	84.94 ± 0.19	**76.80** ± 0.17	65.62 ± 0.03	**70.91** ± 0.11	**66.96** ± 0.11	**71.26** ± 0.09	**72.58** ± 0.07
(b) WeMM-Embedding-9B → TinyBERT										
SimCSE	83.25 ± 0.12	67.81 ± 0.13	47.81 ± 0.94	81.61 ± 0.23	71.89 ± 0.71	65.70 ± 0.15	66.89 ± 0.51	62.66 ± 0.25	65.33 ± 0.06	68.10 ± 0.18
RKD	82.95 ± 0.05	69.16 ± 0.08	53.73 ± 0.52	80.46 ± 0.13	72.40 ± 0.05	64.48 ± 0.05	65.34 ± 0.20	65.01 ± 0.09	66.09 ± 0.04	68.85 ± 0.06
Jasper	82.27 ± 0.35	67.34 ± 0.32	45.14 ± 0.58	80.39 ± 0.25	71.41 ± 0.08	**66.35** ± 0.06	66.48 ± 0.19	62.10 ± 0.45	64.07 ± 0.33	67.28 ± 0.19
TALAS	80.09 ± 0.32	71.90 ± 0.16	64.45 ± 0.25	**83.17** ± 0.17	71.77 ± 0.19	63.88 ± 0.13	67.67 ± 0.07	63.89 ± 0.50	68.45 ± 0.09	70.58 ± 0.07
ORBIT-KD	**84.54** ± 0.04	**72.25** ± 0.06	**65.66** ± 0.28	82.80 ± 0.09	**75.16** ± 0.11	65.41 ± 0.03	**68.47** ± 0.07	**65.31** ± 0.09	**69.00** ± 0.14	**72.07** ± 0.05
(c) NV-Embed-v2 → TinyBERT										
SimCSE	83.25 ± 0.12	67.81 ± 0.13	47.81 ± 0.94	81.61 ± 0.23	71.89 ± 0.71	65.70 ± 0.15	66.89 ± 0.51	62.66 ± 0.25	65.33 ± 0.06	68.10 ± 0.18
RKD	82.94 ± 0.13	68.16 ± 0.13	48.62 ± 0.47	80.18 ± 0.24	71.09 ± 0.26	64.21 ± 0.13	66.33 ± 0.11	63.42 ± 0.07	64.59 ± 0.24	67.73 ± 0.08
Jasper	82.17 ± 0.16	67.98 ± 0.21	44.86 ± 0.25	79.87 ± 0.03	71.26 ± 0.11	**66.34** ± 0.18	66.89 ± 0.09	62.27 ± 0.23	64.15 ± 0.14	67.31 ± 0.07
TALAS	80.76 ± 0.17	72.06 ± 0.14	58.16 ± 0.48	**85.03** ± 0.08	73.80 ± 0.18	62.01 ± 0.02	**68.63** ± 0.05	**66.85** ± 0.22	**68.00** ± 0.19	70.59 ± 0.02
ORBIT-KD	**85.07** ± 0.13	**72.52** ± 0.27	**58.59** ± 0.06	82.99 ± 0.03	**75.70** ± 0.12	62.74 ± 0.11	68.55 ± 0.09	66.33 ± 0.14	66.67 ± 0.18	**71.02** ± 0.02

Table 10: General-purpose embedding transfer from Qwen3-Embedding-0.6B to MiniLMv2-H384 using 20K unlabeled Wikipedia sentences. Results are mean ± sample standard deviation over three seeds. **Bold** and underlined values denote the best and second-best methods.

Method	Banking77	Tweet	Emotion	MRPC	SciTail	WiC	SICK	STS12	STS-B	Avg
SimCSE	83.97 ± 0.30	69.96 ± 0.23	52.75 ± 0.75	80.14 ± 0.06	72.02 ± 0.22	64.43 ± 0.33	61.42 ± 0.16	47.20 ± 2.20	53.22 ± 0.71	65.01 ± 0.40
RKD	86.97 ± 0.17	69.85 ± 0.09	52.98 ± 0.70	83.07 ± 0.04	76.14 ± 0.20	66.50 ± 0.01	63.28 ± 0.33	51.86 ± 0.48	58.60 ± 0.42	67.69 ± 0.11
Jasper	83.53 ± 0.04	68.16 ± 0.34	50.61 ± 0.36	80.85 ± 0.34	71.51 ± 0.47	**66.97** ± 0.27	63.04 ± 0.11	51.34 ± 0.27	57.67 ± 0.46	65.96 ± 0.09
TALAS	86.22 ± 0.25	70.88 ± 0.08	51.66 ± 0.52	84.77 ± 0.08	79.85 ± 0.12	66.64 ± 0.06	69.53 ± 0.02	**64.32** ± 0.26	69.82 ± 0.18	71.52 ± 0.02
ORBIT-KD	**88.12** ± 0.11	**71.35** ± 0.26	**55.44** ± 0.78	**84.80** ± 0.05	**81.76** ± 0.05	66.78 ± 0.06	**69.66** ± 0.17	63.74 ± 0.27	**70.53** ± 0.22	**72.46** ± 0.16

Coordinate resolution transfers beyond task-adapted text. ORBIT-KD reaches a 72.46 nine-task average in Table 10, 0.94 points above TALAS, and leads on seven of nine benchmarks across classification, pair classification, and semantic similarity. The aggregate result is therefore not carried by one task family.

D.5 Training Efficiency

Implementation Details. Table 11 gives the measurements behind Figure 6 for the 15K runs, each on one NVIDIA H200 GPU with batch size 128. Step time excludes the first ten warm-up steps; throughput and optimization minutes are derived from the same differentiated updates, and peak memory is the maximum allocated device memory. The measurements exclude teacher-cache construction, preprocessing, loading, coordinate refresh, checkpointing, and evaluation. We time the four no-gradient refreshes of the default schedule separately.

Table 11: Full optimization-loop measurements on one NVIDIA H200 GPU per run, excluding non-training overhead. Results are mean ± standard deviation over three seeds; best and second-best values are bold and underlined.

Method	Overall ↑	ms/step ↓	samples/s ↑	Minutes ↓	Peak GiB ↓
(a) Qwen3-Embedding-4B → BERT-base (109M)					
RKD	73.52 ± 0.06	<u>119.7</u> ± 18.3	<u>1081.0</u> ± 179.8	<u>1.20</u> ± 0.18	<u>1.72</u> ± 0.00
Jasper	71.95 ± 0.06	130.7 ± 8.1	975.9 ± 58.8	1.27 ± 0.08	1.78 ± 0.01
CDM	71.12 ± 0.18	340.4 ± 13.6	374.2 ± 15.3	3.38 ± 0.12	9.31 ± 0.01
DSKD	71.14 ± 0.10	290.9 ± 9.0	437.7 ± 13.3	2.88 ± 0.09	9.45 ± 0.00
EMO	69.17 ± 0.24	721.1 ± 12.9	176.5 ± 3.1	7.02 ± 0.12	9.30 ± 0.00
TALAS	<u>77.29</u> ± 0.04	329.0 ± 12.5	387.1 ± 15.0	3.23 ± 0.11	**1.43** ± 0.00
ORBIT-KD	**77.67** ± 0.03	**83.4** ± 3.8	**1527.2** ± 70.8	**0.84** ± 0.04	<u>1.72</u> ± 0.00
(b) BGE-M3 → MiniLMv2-H768 (66M)					
RKD	73.66 ± 0.04	<u>77.5</u> ± 8.8	<u>1656.4</u> ± 201.2	<u>0.79</u> ± 0.09	<u>1.07</u> ± 0.00
Jasper	71.13 ± 0.11	152.8 ± 30.8	859.1 ± 195.6	1.49 ± 0.30	1.09 ± 0.00
CDM	70.77 ± 0.02	127.5 ± 13.8	1006.4 ± 115.8	1.30 ± 0.12	2.15 ± 0.01
DSKD	70.83 ± 0.11	115.0 ± 21.9	1137.2 ± 241.0	1.14 ± 0.20	2.22 ± 0.01
EMO	69.59 ± 0.12	394.6 ± 33.4	324.0 ± 28.3	3.87 ± 0.35	2.15 ± 0.01
TALAS	<u>75.55</u> ± 0.06	184.4 ± 4.6	690.2 ± 17.3	1.82 ± 0.03	**0.85** ± 0.00
ORBIT-KD	**76.03** ± 0.11	**56.5** ± 7.7	**2281.6** ± 310.0	**0.57** ± 0.07	<u>1.07</u> ± 0.00
(c) Qwen3-Embedding-0.6B → MiniLMv2-H384 (22M)					
RKD	71.44 ± 0.10	<u>64.8</u> ± 13.7	<u>2012.3</u> ± 488.6	<u>0.68</u> ± 0.14	<u>0.41</u> ± 0.00
Jasper	68.57 ± 0.10	150.7 ± 15.3	850.0 ± 81.5	1.47 ± 0.13	0.42 ± 0.00
CDM	68.69 ± 0.27	161.8 ± 21.6	796.3 ± 113.1	1.65 ± 0.20	1.53 ± 0.00
DSKD	68.59 ± 0.04	175.4 ± 8.2	726.4 ± 33.5	1.78 ± 0.06	1.56 ± 0.00
EMO	67.62 ± 0.02	499.3 ± 1.4	254.8 ± 0.7	4.90 ± 0.01	1.53 ± 0.00
TALAS	<u>73.20</u> ± 0.06	169.8 ± 15.4	753.7 ± 71.4	1.69 ± 0.14	**0.34** ± 0.00
ORBIT-KD	**74.86** ± 0.06	**62.9** ± 3.1	**2025.2** ± 98.3	**0.64** ± 0.04	<u>0.41</u> ± 0.00

ORBIT-KD has the fastest optimization loop. ORBIT-KD needs 83.4, 56.5, and 62.9 ms per step in configurations (a)–(c), compared with 329.0, 184.4, and 169.8 ms for TALAS, a 3.9×, 3.3×, and 2.7× speedup. It is also faster than RKD, a cache-based baseline with a lightweight relational objective (119.7, 77.5, and 64.8 ms), while scoring 4.15, 2.37, and 3.42 points higher. The complete five-epoch optimization takes 0.84, 0.57, and 0.64 minutes.