SUGAR: Spectral and Geometry-Aware Alignment for Matryoshka Representation Distillation

Anonymous ACL submission

Abstract

Matryoshka Representation Learning (MRL) produces nested embeddings whose prefixes remain semantically meaningful, allowing representations to be truncated at inference time and enabling a single model to operate across multiple dimensional budgets. This paradigm has been adopted in several modern embedding models such as Qwen3-Embedding and OpenAI text-embedding-3. However, distilling MRL teachers into compact students remains largely unexplored, as conventional knowledge distillation methods focus on aligning single-scale embeddings and do not account for the multi-scale structure of Matryoshka representations. We introduce Matryoshka Representation Distillation and propose SUGAR, a new framework that preserves the structural properties of MRL embeddings during knowledge distillation. SUGAR aligns point-wise semantic fidelity, spectral capacity via effective-rank matching, and global subspace geometry through Grassmannian alignment. Our extensive experiments show that SUGAR produces compact students that retain strong performance across multiple embedding dimensions while preserving deployment flexibility.

1 Introduction

Dense representations are a core component of modern NLP systems, supporting tasks such as retrieval and semantic search (Mikolov et al., 2013; Pennington et al., 2014; Devlin et al., 2019). Matryoshka Representation Learning (MRL) (Kusupati et al., 2022) introduces a flexible paradigm for embedding generation by producing nested, prefix-able representations. Instead of a single fixed-dimensional vector, an MRL model yields embeddings whose prefixes remain semantically meaningful, enabling representations to be truncated at inference time to satisfy different computational budgets without retraining. This paradigm has been adopted in modern embedding systems (e.g., Qwen and OpenAI text-embedding models) to support variable-dimensional representations in real-world deployments.

Despite its flexibility, compressing MRL teachers into lightweight students introduces challenges that existing Knowledge Distillation (KD) frameworks do not address. Traditional KD assumes a single-scale representation space and aligns individual teacher and student embeddings (Zhang et al., 2025; Xu et al., 2023; Truong et al., 2025). In contrast, Matryoshka representations are inherently multi-scale, organizing embeddings as nested prefixes under different dimensional budgets. Therefore, preserving only vector-level similarity is insufficient to transfer their multi-scale behavior. A key advantage of MRL is adaptive inference, enabling a single model to operate under diverse computational budgets, from high-precision server retrieval to low-latency on-device inference. However, most KD pipelines assume a fixed-dimensional student representation, making it difficult to distill compact models that retain this flexibility.

To the best of our knowledge, we are the first to formalize the problem of Matryoshka Representation Distillation. We propose **SUGAR** (Spectral and Geometry-Aware Alignment for Representation Distillation), a principled framework that preserves multi-scale geometry during unsupervised distillation. The core idea is to orchestrate the distillation process through two complementary levels of geometric alignment: (i) **Spectral and Subspace Geometry Alignment**, which jointly preserves intrinsic dimensionality via effective-rank (eRank) matching and global orientations via Grassmannian chordal-distance minimization; and (ii) **Local Fidelity and Incremental Novelty**, which anchors individual semantic coordinates via point-wise losses while preventing dimensional collapse through orthogonal band regularization. Together, these unified objectives encourage the student not only to mimic individual embeddings but also to

faithfully reproduce the nested subspace hierarchy of Matryoshka representations, enabling the distilled model to inherit their multi-scale truncation property and deployment elasticity.

Our contributions are as follows:

- We introduce Matryoshka Representation Distillation and propose **SUGAR**, a hierarchical distillation framework that integrates spectral and subspace geometry alignment (eRank and Grassmannian) with local fidelity and incremental novelty constraints (point-wise anchors and orthogonal band regularization).
- We empirically validate SUGAR across multiple teacher-student pairs and benchmarks (STS, NLI, and classification), showing that SUGAR produces compact students that maintain strong representational quality from low to high dimensions, fully preserving deployment elasticity for real-world applications.

2 Related Work and Background

2.1 Related Works

Learning Sentence Embeddings Early sentence representations relied on static word embeddings (Pennington et al., 2014; Mikolov et al., 2013) and later contextualized models (Peters et al., 2018; Devlin et al., 2019), which were not explicitly optimized for sentence-level semantics. Sentence-BERT (Reimers and Gurevych, 2019) addressed this by introducing a siamese fine-tuning framework for efficient semantic similarity and retrieval. Later work improved embedding quality through contrastive learning (Gao et al., 2022; Nishikawa et al., 2022). More recently, Large Language Models have enabled new approaches where LLMs act as strong backbones or teachers for producing high-quality sentence embeddings (He et al., 2025; BehnamGhader et al., 2024; Tao et al., 2025).

Matryoshka Representation Learning Methods Although sentence representation learning has advanced significantly, fixed embedding dimensionality still limits large-scale deployment due to computational and storage costs. Matryoshka Representation Learning (MRL) addresses this by learning nested embeddings in a coarse-to-fine manner, where larger prefixes yield more expressive representations. Low-dimensional prefixes can serve as standalone embeddings while requiring supervision at only $O(\log d)$ checkpoints and no extra forward passes (Kusupati et al., 2022). This enables strong compression with competitive performance and faster inference. Building on this idea, Espresso Sentence Embeddings (ESE) introduce learn-to-express and learn-to-compress mechanisms to scale representations across both dimensionality and network depth (LI et al., 2025). The Matryoshka principle has also been extended to other domains, including image generation (Gu et al., 2024), multimodal representation learning (Cai et al., 2024), and LLMs (Hu et al., 2024).

2.2 Background

Knowledge Distillation. Knowledge Distillation transfers knowledge from a teacher model T to a student model S by training the student to mimic the teacher under a specific supervision signal. For an input x, the teacher generates an output $z_T(x)$ while the student produces $z_S(x)$, where $z(\cdot)$ may denote logits, hidden states, or embeddings depending on the task. The student is optimized by minimizing the distillation objective

$$\mathcal{L}_{\text{KD}} = \mathcal{D}\big(z_S(x), z_T(x)\big), \tag{1}$$

where $\mathcal{D}(\cdot,\cdot)$ is a discrepancy function such as cross-entropy or a similarity-based distance, etc.

2.2.1 Matryoshka Representation Learning

Formal formulation. Let an encoder f_θ map an input sentence x to a dense embedding $z = f_\theta(x) \in \mathbb{R}^D$. In MRL (Kusupati et al., 2022), a predefined set of increasing prefix dimensions is specified as $\mathcal{D} = \{d_1, d_2, \ldots, d_n\}, \quad 0 < d_1 < d_2 < \cdots < d_n = D$. For each $d_i \in \mathcal{D}$, the corresponding representation is obtained by truncating the full embedding to its first d_i components, denoted by $z^{(d_i)} := z_{1:d_i} \in \mathbb{R}^{d_i}$. Each prefix $z^{(d_i)}$ is treated as a valid embedding and is fed into a task-specific head $g_{\phi_i} : \mathbb{R}^{d_i} \to \mathcal{Y}$. Given a labeled pair (x, y), the prediction from this prefix is evaluated using a task loss

$$\mathcal{L}_{\text{task}}^{(d_i)}(x, y) = \ell\big(g_{\phi_i}(z^{(d_i)}), y\big).$$

Training in MRL jointly supervises all prefixes of the embedding. A typical objective aggregates the losses across all prefix dimensions:

$$\mathcal{L}_{\text{MRL}}(\theta, \{\phi_i\}) = \sum\nolimits_{i=1}^{n} \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{train}}}\big[\mathcal{L}_{\text{task}}^{(d_i)}(x, y)\big], \tag{2}$$

although a weighted variant $\sum_{i=1}^{n} \alpha_i \, \mathbb{E}[\mathcal{L}_{\text{task}}^{(d_i)}]$ with $\alpha_i \geq 0$ can also be used. Optimizing this objective encourages the encoder to distribute useful semantic information across the embedding

hierarchy, such that lower-dimensional prefixes $z^{(d_1)}, \ldots, z^{(d_{n-1})}$ remain directly usable as independent embeddings at inference time without requiring additional forward passes.

Practical Impact and Deployment Elasticity. Beyond its theoretical formulation, MRL offers significant practical advantages for large-scale retrieval systems (Kusupati et al., 2022). In real-world deployments, storing and searching over massive vector databases introduces substantial memory and latency costs. MRL alleviates this issue through *adaptive retrieval*, where compressed prefixes (e.g., $d = 16$ or 32) are used for fast and memory-efficient coarse search, followed by re-ranking the top-K candidates with full embeddings. This cascading strategy greatly reduces storage requirements, memory footprint, and I/O overhead while maintaining retrieval accuracy. Moreover, MRL provides strong deployment flexibility: a single trained model can serve both resource-constrained devices using truncated prefixes and high-performance servers using full embeddings, eliminating the need to train and maintain multiple models for different computational budgets.

2.2.2 Grassmann Manifold and Subspace Geometry

Many representation learning problems are naturally described by linear subspaces rather than individual vectors. The set of all k-dimensional subspaces of $\mathbb{R}^d$ forms the *Grassmann manifold*, denoted $\mathrm{Gr}(k, d)$ (Edelman et al., 1998). Each point on this manifold represents a k-dimensional subspace and can be expressed by an orthonormal basis matrix $U \in \mathbb{R}^{d\times k}$ satisfying:

$$U^\top U = I_k. \tag{3}$$

Different orthonormal bases spanning the same column space correspond to the same point on the Grassmann manifold. A fundamental tool for measuring the relationship between two subspaces is the concept of *principal angles*. Given two subspaces $\mathcal{U}$ and $\mathcal{V}$ with orthonormal basis matrices $U \in \mathbb{R}^{d\times k}$ and $V \in \mathbb{R}^{d\times k}$, the principal angles $\{\theta_i\}_{i=1}^k$ are defined through the singular values of $U^\top V$ (Åke Björck and Golub, 1973):

$$\cos\theta_i = \sigma_i(U^\top V), \qquad i = 1, \ldots, k, \tag{4}$$

where $\sigma_i(\cdot)$ denotes the i-th singular value. These angles measure the alignment between two subspaces: small angles indicate close alignment, while larger angles indicate greater misalignment. Based on the principal angles, several distances can be defined on the Grassmann manifold, with the *chordal distance* being a commonly used metric:

$$d_c^2(\mathcal{U}, \mathcal{V}) = \sum\nolimits_{i=1}^k \sin^2\theta_i. \tag{5}$$

This distance admits a convenient matrix form when orthonormal bases are used:

$$d_c^2(\mathcal{U}, \mathcal{V}) = k - \|U^\top V\|_F^2, \tag{6}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. The chordal distance provides a computationally efficient way to measure subspace discrepancy and has been widely used in representation learning and computer vision applications on Grassmann manifolds (Harandi et al., 2013).

In representation distillation, this framework provides a natural way to compare the geometric structure of embedding spaces. Instead of aligning individual vectors only, one can compare the subspaces spanned by collections of embeddings, allowing the distillation process to preserve higher-level geometric relationships between representations.

3 Methodology

The goal of SUGAR is to compress a massive MRL teacher into a lightweight student without compromising the topological structure of its nested subspaces. To achieve this, we first map the student's embeddings into the teacher's ambient space via a trainable linear projection layer. Subsequently, the student is optimized end-to-end using a unified, multi-objective distillation strategy computed across a predefined sequence of prefix dimensions. This strategy orchestrates two complementary levels of geometric alignment: (i) **Spectral and Subspace Geometry Alignment**, which preserves the spectral capacity of each prefix via *effective-rank (eRank) matching* and aligns the geometric orientations of the nested spaces via *Grassmannian subspace distillation*; and (ii) **Local Fidelity and Incremental Novelty**, which maintains local vector-level fidelity via *point-wise alignment* and ensures that progressively expanding prefix bands capture genuinely novel, non-redundant semantic information via *orthogonal band regularization*.

3.1 Problem Setup

Let f_T and f_S denote a frozen teacher and a trainable student encoder that map an input x to dense

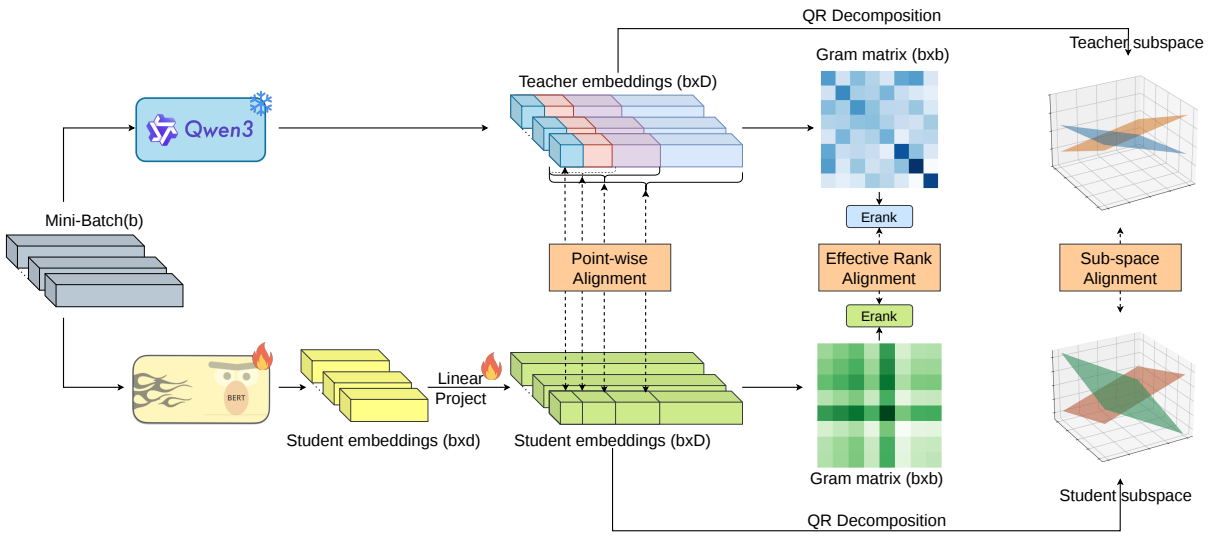

Figure 1: Overview of the SUGAR framework.

representations $h_T \in \mathbb{R}^{D_T}$ and $z_S \in \mathbb{R}^{D_S}$. Since $D_S \leq D_T$, we project the student representation into the teacher space using a learnable linear mapping $W \in \mathbb{R}^{D_S \times D_T}$, yielding $h_S = z_S W$. To distill the multi-scale hierarchy of the Matryoshka teacher, we define an ordered set of target dimensions $\mathcal{D} = \{d_1, d_2, \ldots, d_K\}$, where $d_1 < \cdots < d_K = D_T$. For any dimension $d \in \mathcal{D}$, we extract the corresponding d-dimensional prefix by slicing the ambient representations:

$$s_d = h_S[:d], \qquad t_d = h_T[:d]. \tag{7}$$

All distillation objectives in the SUGAR framework are computed over these matched pairs of prefix slices (s_d, t_d) at every dimension $d \in \mathcal{D}$ to ensure the student faithfully inherits the teacher's nested multi-scale geometry.

3.2 Spectral and Subspace Geometry Alignment

To faithfully transfer the Matryoshka structure, the student must accurately mirror the teacher at the macroscopic level of information density and global spatial orientation. We achieve this through a dual-alignment strategy combining intrinsic dimensionality matching and Grassmannian subspace alignment.

3.2.1 Effective-Rank Matching for Spectral Capacity

In addition to structural alignment, it is desirable to preserve the intrinsic dimensionality of the representation space. In practice, student representations may concentrate variance along a limited number of dominant directions, resulting in an underutilization of the available embedding subspace. To encourage a more balanced distribution of information across dimensions, we introduce an intrinsic dimensionality regularization that aligns the effective rank (eRank) of the teacher and student batch representations. For a batch matrix $H \in \mathbb{R}^{b \times d}$, let $\{\sigma_k\}$ denote the eigenvalues of its normalized Gram matrix $G = \frac{1}{b} H H^\top$. The normalized spectrum is given by $p_k = \frac{\sigma_k}{\sum_j \sigma_j}$. The effective rank is defined as the Shannon entropy of this spectrum:

$$\mathrm{eRank}(H) = \exp\Big(-\sum\nolimits_k p_k \log p_k\Big). \tag{8}$$

This metric provides a continuous, robust estimation of variance dispersion, capturing the true intrinsic dimensionality of the embedding space. Geometrically, an isotropic distribution spans all available dimensions uniformly (analogous to a high-dimensional sphere, where eRank $\approx d$), whereas highly anisotropic representations concentrate their energy along fewer dominant directions (collapsing into a plane or a line, yielding eRank $\ll d$). Unlike point-wise regression, eRank matching is inherently rotation-invariant and dimension-agnostic, naturally accommodating the architectural and geometric mismatches between the student and teacher networks. By acting as a structural regularizer, this objective forces the student to explicitly replicate the teacher's specific spectral profile and calibrated intrinsic dimensionality. Consequently, the compact student optimally allocates its limited representational capacity, maintaining high information density without semantic redundancy across

all nested prefixes. We formulate the eRank distillation loss using L1 penalty on the dimension-normalized ranks:

$$\mathcal{L}_{\text{erank}}^{(d)} = \text{L1}\Big(\frac{\text{eRank}(S_d)}{d}, \frac{\text{eRank}(T_d)}{d}\Big) \quad (9)$$

where $S_d, T_d \in \mathbb{R}^{b\times d}$ are the batch matrices of the student and teacher prefixes, respectively.

3.2.2 Geometry Alignment via Grassmannian Manifold

While spectral capacity constraints ensure that the student correctly populates the representational volume, they are fundamentally oblivious to the global orientation of the underlying embedding manifold. To truly replicate the Matryoshka structure, the student's subspace at every prefix dimension perfectly aligns with the teacher's geometric orientation. We formalize this by minimizing the discrepancy between the subspaces spanned by their respective batch representations on the Grassmann manifold.

Orthonormal Basis Extraction. To represent the subspace spanned by a batch matrix $H \in \mathbb{R}^{b\times d}$ as a point on the Grassmannian $\text{Gr}(r, d)$, we must extract an orthonormal basis. We achieve this by performing a thin QR decomposition on the transposed batch matrix:

$$H^\top = QR, \text{ where } Q \in \mathbb{R}^{d\times \min(d,b)},\ R \in \mathbb{R}^{\min(d,b)\times b}$$

To isolate the true informational subspace from computational noise, we retain only the columns of Q whose corresponding diagonal entries in R satisfy $|R_{jj}| > \epsilon$ (with $\epsilon = 10^{-6}$). This yields a robust orthonormal basis $U \in \mathbb{R}^{d\times r}$, where $r \leq \min(b, d)$ represents the effective rank of the batch.

Generalized Chordal Distance and Geometric Interpretation. A critical challenge in representation distillation is that the student and teacher batches may span subspaces of different intrinsic dimensionalities during training ($r_S \neq r_T$). To mathematically resolve this, we employ a generalized squared chordal distance. For the teacher's pre-computed basis $U_T \in \mathbb{R}^{d\times r_T}$ and the student's basis $U_S \in \mathbb{R}^{d\times r_S}$, the Grassmannian distillation loss for a prefix dimension d :

$$\mathcal{L}_{\text{grass}}^{(d)} = r - \left\| U_S^\top U_T \right\|_F^2 \quad (10)$$

where $r = \min(r_S, r_T)$. This formulation has profound geometric implications. Mathematically, the Frobenius norm term $\left\|U_S^\top U_T\right\|_F^2$ is exactly equivalent to the sum of squared cosines of the principal angles between the two subspaces ($\sum_{i=1}^{r} \cos^2\theta_i$). Consequently, minimizing this loss is analytically equivalent to minimizing $\sum_{i=1}^{r} \sin^2\theta_i$.

While eRank regulates the intrinsic variance within the embedding space, the Grassmannian loss aligns the global d-dimensional subspace spanned by the student's batch with that of the teacher. By establishing this macroscopic alignment, we avoid prematurely forcing rigid coordinate-level matching, which can overwhelm low-capacity student models. Instead, the student is first required to orient its overall representation space with the teacher's global subspace. This grants the student the flexibility to organically organize its embeddings within a valid semantic volume before fine-grained local constraints are applied. Furthermore, extracting the orthonormal basis U_T provides a low-rank approximation of the underlying data manifold, capturing the principal directions of semantic variation while filtering coordinate-level noise in individual teacher embeddings. This denoised geometric alignment is particularly important for Matryoshka distillation, where nested embeddings rely on consistent semantic directions across different prefixes. By aligning these subspaces, the student preserves the hierarchical structure of the teacher and avoids geometric distortion when embedding dimensions are truncated.

3.3 Local Fidelity and Incremental Novelty

Having aligned the macroscopic geometric structure, we now introduce microscopic constraints to anchor individual semantic coordinates and regulate the incremental information captured as prefix dimensions expand.

3.3.1 Point-wise Alignment for Local Fidelity

At the most granular level, the student should preserve the semantic similarities in the teacher's embeddings. We enforce this by aligning the student's projected views with the teacher's representations using a cosine distance objective. For a prefix dimension d, the point-wise alignment loss over a mini-batch of size b is:

$$\mathcal{L}_{\text{point-wise}}^{(d)} = \frac{1}{b}\sum_{i=1}^{b} \left(1 - \cos(s_{d,i}, t_{d,i})\right). \quad (11)$$

Where $s_{d,i}$ and $t_{d,i}$ are the d-dimensional prefix slices of the projected student and teacher embeddings for the i-th sample. This objective acts as

a local coordinate anchor, ensuring that individual data points are pinned to their correct semantic locations within the subspace.

3.3.2 Orthogonal Band Regularization

Previous objectives do not explicitly govern the *incremental* information added when expanding from a smaller prefix d_i to a larger d_{i+1}. To avoid redundancy, we introduce an orthogonal band regularization that encourages new dimensions to capture information not present in smaller prefixes.

For each consecutive dimension pair (d_i, d_{i+1}) in $\mathcal{D}$, we partition the embedding $z \in \mathbb{R}^{d_K}$ into a base prefix and an extension band:

$$z_i^{\text{prefix}} = z_{1:d_i} \in \mathbb{R}^{d_i}, \; z_i^{\text{band}} = z_{d_i+1:d_{i+1}} \in \mathbb{R}^{d_{i+1}-d_i}.$$

Let $P_i \in \mathbb{R}^{b \times d_i}$ and $B_i \in \mathbb{R}^{b \times (d_{i+1}-d_i)}$ denote their respective mini-batch matrices. To compute cross-correlation, we mean-center and column-wise L_2-normalize both matrices:

$$\overline{P}_i = \left(P_i - \frac{1}{b}\mathbf{1}\mathbf{1}^\top P_i\right)\Lambda_{P_i}^{-1}, \tag{12}$$

$$\overline{B}_i = \left(B_i - \frac{1}{b}\mathbf{1}\mathbf{1}^\top B_i\right)\Lambda_{B_i}^{-1}, \tag{13}$$

where $\mathbf{1} \in \mathbb{R}^b$ is a vector of ones, and $\Lambda_{P_i}, \Lambda_{B_i}$ are diagonal matrices containing the L_2-norms of the mean-centered columns. The cross-correlation matrix is given by $C_i = \overline{P}_i^\top \overline{B}_i \in \mathbb{R}^{d_i \times (d_{i+1}-d_i)}$. We explicitly minimize its squared Frobenius norm across all progressive expansions:

$$\mathcal{L}_{\text{ortho}} = \frac{1}{K-1} \sum_{i=1}^{K-1} \frac{\|C_i\|_F^2}{d_i \cdot (d_{i+1} - d_i)}. \tag{14}$$

Normalizing by the spatial dimensions ensures scale invariance. By driving $C_i \approx 0$, we force new extension bands to be orthogonal to the prefix space, improving representational density across the hierarchy.

3.4 Final Objective Function

With the multi-scale geometric forces defined, we now aggregate them to form the unified SUGAR objective. As established in our problem formulation, the Matryoshka representation space is structured by an ordered set of nested dimensions $\mathcal{D} = \{d_1, d_2, \ldots, d_K\}$, where $0 < d_1 < d_2 < \cdots < d_K = D$. The point-wise fidelity, effective-rank matching, and Grassmannian subspace alignment objectives are initially computed independently for each prefix dimension $d \in \mathcal{D}$. To optimize the entire nested hierarchy simultaneously, we aggregate these scale-specific terms by averaging them across all K target dimensions: $\mathcal{L}_{\text{point-wise}} = \frac{1}{K}\sum_{d\in\mathcal{D}} \mathcal{L}_{\text{point-wise}}^{(d)}$, $\mathcal{L}_{\text{erank}} = \frac{1}{K}\sum_{d\in\mathcal{D}} \mathcal{L}_{\text{erank}}^{(d)}$, $\mathcal{L}_{\text{grass}} = \frac{1}{K}\sum_{d\in\mathcal{D}} \mathcal{L}_{\text{grass}}^{(d)}$.

Finally, the SUGAR framework unifies the multi-scale geometric distillation objectives with the standard Matryoshka formulation ($\mathcal{L}_{\text{MRL}}$):

$$\mathcal{L} = \alpha\mathcal{L}_{\text{MRL}} + (1-\alpha)\left(\mathcal{L}_{\text{point-wise}} + \mathcal{L}_{\text{erank}} + \mathcal{L}_{\text{grass}}\right) + \mathcal{L}_{\text{ortho}} \tag{15}$$

where $\alpha \in [0, 1]$ balances the base MRL objective and the distillation objectives

4 Experiments

4.1 Experimental Setup

We evaluate embeddings across in-domain and out-of-domain settings to assess both task performance and generalization.

Task categories. We report results across three representative task families, all of which rely critically on embedding quality. First, for **Text Classification**, which requires models to capture the overall semantics of a single text input, we evaluate on TweetEval (Barbieri et al., 2020), as well as Emotion and Banking77 from the MTEB benchmark (Muennighoff et al., 2023). Second, regarding **NLI**, which focuses on modeling semantic relationships between two inputs, we consider MRPC from GLUE (Wang et al., 2018), WiC (Pilehvar and Camacho-Collados, 2019), and SciTail (Khot et al., 2018). Finally, for **Semantic Textual Similarity (STS)**, which measures the ability to capture fine-grained semantic similarity, we evaluate on STS-B and SICK (Marelli et al., 2014) for in-domain testing, and extend to STS12–16 and SickR (Muennighoff et al., 2023) for OOD evaluation. Additional details regarding model architectures, training procedures, datasets, and evaluation protocols are provided in Appendix A.

Baselines. We compare against SOTA Matryoshka baselines: **MRL** (Kusupati et al., 2022), which learns nested prefixes via multi-scale supervision, and **ESE** (LI et al., 2025), which scales across both dimensionality and model depth. These provide a strong benchmark for evaluating low-dimensional representation quality.

Dim.	Data	TinyBERT 6L				BERT				Data	TinyBERT 6L				BERT			
		MRL	ESE	SUGAR$_{0.6B}$	SUGAR$_{4B}$	MRL	ESE	SUGAR$_{0.6B}$	SUGAR$_{4B}$		MRL	ESE	SUGAR$_{0.6B}$	SUGAR$_{4B}$	MRL	ESE	SUGAR$_{0.6B}$	SUGAR$_{4B}$
16	Banking77	40.64	**43.30**	40.38	41.80	46.39	45.86	48.96	**50.34**	WiC	58.07	**58.12**	58.07	57.79	60.13	61.14	**62.34**	62.22
32		63.48	63.78	64.41	**65.96**	64.90	66.39	68.38	**69.19**		59.85	59.35	**60.79**	60.07	62.42	62.64	63.43	**64.77**
64		74.89	73.96	77.44	**78.68**	76.84	79.04	80.48	**80.98**		61.35	59.95	**61.93**	61.79	64.28	63.57	64.64	**64.71**
128		81.97	81.53	**85.22**	85.16	83.49	85.39	86.01	**86.40**		62.78	60.42	62.86	**63.21**	65.24	66.14	**66.29**	65.71
256		85.51	85.76	88.20	**88.22**	86.45	87.59	87.85	**88.23**		**63.71**	60.76	61.36	62.29	66.42	66.92	**67.14**	66.93
512		87.20	87.04	**89.95**	89.57	87.85	88.93	89.15	**89.23**		63.14	61.50	62.93	**63.64**	65.85	67.77	68.23	**68.29**
768		87.70	88.89	**90.01**	89.93	88.43	89.83	89.62	**90.23**		63.07	61.85	63.50	**63.93**	65.91	67.42	**69.33**	69.07
16	TweetEval	53.66	52.71	**57.33**	57.06	**55.96**	50.73	50.76	52.41	SICK	59.07	58.42	64.04	**64.49**	59.35	61.50	**65.11**	63.47
32		59.08	59.17	**61.49**	59.99	**60.51**	54.75	55.84	57.15		65.08	64.98	68.00	**68.66**	63.34	62.36	**68.03**	66.48
64		60.22	65.16	**65.66**	65.13	**65.68**	60.55	62.74	63.21		68.47	66.92	70.30	**71.08**	66.11	62.07	**70.56**	68.94
128		66.71	66.75	69.48	**69.54**	**68.14**	65.51	67.07	67.17		69.84	68.73	71.97	**72.09**	66.57	66.74	**71.51**	70.49
256		68.03	69.93	71.44	**71.94**	69.38	68.73	69.13	**69.59**		69.85	69.03	71.53	**72.00**	67.23	66.43	**72.22**	71.23
512		68.25	69.32	**73.31**	72.73	70.86	70.22	70.46	**71.51**		70.09	70.57	72.23	**72.39**	68.33	67.04	**72.60**	71.61
768		69.10	70.02	73.20	**73.51**	70.72	70.75	71.06	**71.92**		70.00	70.67	72.34	**72.47**	68.94	68.19	**72.50**	71.59
16	MRPC	61.12	65.12	**71.77**	71.59	65.93	66.73	70.78	**71.19**	STSB	60.00	57.45	59.78	**61.93**	61.65	59.39	63.21	**63.34**
32		69.87	69.38	**74.03**	73.10	70.11	71.07	72.06	**73.16**		63.51	60.12	67.83	**68.48**	65.09	64.39	68.29	**68.46**
64		72.28	70.12	**74.55**	73.74	70.14	72.57	73.22	**73.68**		65.45	65.78	71.95	**72.28**	67.32	67.51	70.65	**71.39**
128		72.34	72.89	**74.38**	**74.38**	72.96	71.71	74.09	**74.32**		67.16	68.52	73.72	**73.92**	69.59	70.12	72.75	**73.39**
256		72.52	73.97	**74.09**	74.03	72.98	72.29	73.86	**74.67**		67.79	68.36	73.00	**73.17**	70.41	70.91	73.52	**74.15**
512		72.76	74.31	**74.78**	74.20	73.16	73.27	74.38	**74.49**		68.85	68.92	**74.60**	**74.60**	71.33	70.83	74.67	**75.23**
768		72.87	74.02	**74.78**	74.61	73.22	72.46	73.80	**74.14**		68.93	69.02	74.55	**74.85**	71.30	71.99	74.40	**75.14**

Table 1: Results on in-domain datasets with TinyBERT 6L and BERT backbone embedding models.

Dim.	Data	TinyBERT 6L				BERT				Data	TinyBERT 6L				BERT			
		MRL	ESE	SUGAR$_{0.6B}$	SUGAR$_{4B}$	MRL	ESE	SUGAR$_{0.6B}$	SUGAR$_{4B}$		MRL	ESE	SUGAR$_{0.6B}$	SUGAR$_{4B}$	MRL	ESE	SUGAR$_{0.6B}$	SUGAR$_{4B}$
16	STS12	51.48	50.90	58.75	**59.29**	55.13	51.34	59.10	**59.29**	STS16	59.21	60.26	**62.27**	62.25	54.78	59.67	60.81	**61.47**
32		55.25	56.02	64.41	**64.71**	59.78	53.78	62.84	**63.10**		63.45	65.63	68.10	**68.40**	60.96	64.54	**67.10**	**67.10**
64		55.67	58.32	**67.45**	67.30	61.24	55.23	64.38	**64.55**		66.29	66.04	72.25	**72.91**	63.45	68.15	70.07	**71.10**
128		59.80	60.38	67.82	**68.48**	62.34	59.97	65.78	**66.18**		67.90	68.12	**74.50**	74.42	66.78	70.14	72.06	**73.07**
256		60.17	61.23	66.85	**68.17**	64.13	60.33	65.47	**66.13**		68.10	70.55	73.22	**73.47**	69.89	70.82	72.66	**73.37**
512		61.52	60.82	66.79	**68.44**	65.09	61.06	65.98	**66.66**		68.60	70.87	73.74	**74.50**	70.60	71.47	72.98	**73.73**
768		61.78	64.16	67.90	**68.53**	64.84	60.43	64.86	65.70		68.69	70.25	73.75	**74.63**	70.53	70.54	73.10	**73.97**
16	STS13	55.03	53.51	**62.40**	61.45	62.13	60.48	63.60	**64.58**	SickR	60.43	60.79	63.92	**64.60**	63.45	64.06	**65.53**	64.11
32		63.48	60.46	**69.54**	69.51	63.44	65.78	69.40	**70.31**		66.71	66.88	67.45	**68.07**	65.92	66.77	**68.27**	66.76
64		62.96	64.56	73.76	**74.13**	68.98	68.97	73.20	**74.40**		68.45	68.30	68.89	**70.77**	68.25	69.09	**70.58**	68.93
128		67.88	66.40	75.44	**75.92**	70.97	71.68	74.80	**75.79**		69.82	69.39	71.71	**72.00**	69.61	70.30	**71.41**	70.88
256		67.80	68.53	74.60	**75.60**	71.24	72.32	75.57	**76.42**		69.87	69.63	71.47	**71.93**	70.49	71.08	**72.09**	71.10
512		68.96	69.65	**76.73**	76.31	72.35	73.63	75.94	**76.54**		70.16	70.38	72.19	**72.34**	70.83	71.48	**72.60**	71.62
768		69.16	69.94	**76.84**	76.39	73.56	73.81	76.51	**77.32**		70.10	70.49	72.28	**72.44**	70.98	71.58	**72.56**	71.61
16	STS14	50.42	50.12	**54.38**	53.91	51.76	54.20	54.84	**55.71**	Emotion	**31.30**	29.76	27.72	28.54	26.98	27.95	27.60	**29.10**
32		53.67	53.27	**60.99**	59.35	54.78	56.89	60.13	**61.41**		**37.94**	36.11	36.41	36.53	36.87	**37.49**	34.21	35.26
64		59.57	57.64	**65.35**	65.05	60.88	60.44	63.29	**64.29**		43.53	44.29	**45.89**	44.79	42.52	44.61	43.19	**44.76**
128		60.84	61.42	67.13	**67.21**	62.05	61.45	64.44	**65.39**		49.50	**52.37**	50.77	51.62	49.15	**50.37**	49.47	49.33
256		60.91	62.36	66.85	**66.98**	63.68	63.81	65.54	**66.60**		53.92	56.12	**57.57**	57.02	54.67	54.33	**56.59**	56.09
512		62.22	63.47	67.50	**67.56**	64.56	64.63	66.59	**67.04**		56.39	56.82	**60.39**	59.38	55.98	56.63	**58.56**	57.75
768		62.32	63.85	**67.56**	67.41	64.65	64.85	66.59	**67.56**		58.53	60.43	60.27	**60.83**	54.08	57.42	58.65	**58.70**
16	STS15	64.17	63.84	**66.40**	65.95	61.99	64.78	67.79	**67.95**	SciTail	72.25	71.87	72.62	**73.80**	68.45	69.61	70.60	**70.84**
32		69.88	69.23	73.81	**74.03**	70.62	69.46	71.67	**72.64**		74.92	73.32	**76.58**	76.48	68.78	73.51	**73.71**	73.37
64		72.51	72.08	76.45	**77.68**	73.26	72.64	75.36	**76.78**		75.40	73.75	77.33	**77.85**	71.34	75.02	74.60	**75.82**
128		73.24	73.75	**78.57**	78.22	75.28	75.63	77.65	**78.96**		75.68	74.90	77.42	**78.41**	72.56	75.77	76.42	**76.72**
256		73.29	76.12	**78.52**	78.40	76.24	76.93	79.14	**80.10**		75.92	75.53	76.06	**77.01**	75.77	76.29	77.03	**77.14**
512		73.87	75.24	**79.19**	79.11	76.73	76.12	79.12	**80.22**		75.82	75.77	77.93	**78.45**	75.72	76.38	77.00	**77.42**
768		73.90	76.24	79.40	**79.81**	76.62	76.94	79.25	**80.13**		75.90	75.72	77.80	**78.08**	76.34	76.85	**78.65**	78.50

Table 2: Results on out-domain datasets with TinyBERT 6L and BERT backbone embedding models.

4.2 Results

Tables 1 and 2 report results on both in-domain and out-of-domain datasets. Using **TinyBERT-6L** and **BERT-base** as student embedding models, SUGAR consistently outperforms or remains competitive with strong baselines (MRL and ESE) across all dimensional budgets, with the largest gains at low dimensions ($d = 16, 32$), highlighting the effectiveness of geometric distillation under severe truncation. SUGAR also generalizes well to unseen benchmarks, and distillation from a larger teacher (SUGAR$4B$ distilled from Qwen3-Embedding-4B versus SUGAR$0.6B$ from Qwen3-Embedding-0.6B) further improves performance, demonstrating effective transfer of multi-scale geometry into compact encoders. Additional results on larger student models (e.g., BGEM3) are provided in Section B.

5 Ablation study

Comparision with Distillation methods As shown in Table 3, the SUGAR framework demonstrates a significant advantage in representational efficiency. Notably, the performance of SUGAR

	Classification (F1)			NLI (Accuracy)			STS (Spearman)			Domain Avg		
	Emotion	**Tweet⋆**	**Banking77⋆**	**Scitail**	**WiC⋆**	**MRPC⋆**	**SickR**	**STS15**	**STSB⋆**	**Avg-In**	**Avg-Out**	**Avg.**
Qwen3-Embedding 0.6B → Bert-base 109M												
Teacher	64.13	71.30	92.08	84.76	70.50	74.36	72.06	79.42	75.90	76.83	75.09	76.06
Student base	51.08	67.13	86.21	73.84	63.29	68.49	58.65	60.29	47.29	66.48	60.97	64.03
SimCSE-unsup	56.02	69.42	89.06	77.94	63.50	73.91	67.24	77.06	69.95	73.17	69.57	72.78
CDM	56.20	68.45	88.10	76.80	64.90	73.15	69.40	77.75	69.30	71.58	70.04	71.58
DSKD	55.10	68.90	87.65	76.20	65.10	71.05	68.10	72.40	65.85	71.71	67.95	70.04
Jasper and Stella	57.40	70.15	88.90	77.85	68.80	73.20	71.95	78.60	73.10	74.83	71.45	73.33
DistillCSE	56.80	69.70	88.45	77.10	67.95	72.40	70.25	76.80	70.95	73.89	70.24	72.27
EMO	57.85	70.80	89.10	78.20	68.15	73.45	72.40	78.95	73.80	75.06	71.85	73.63
SUGAR (256)	56.59	69.13	87.85	77.03	67.14	73.86	72.09	79.14	73.52	74.30	71.21	72.93
SUGAR (512)	58.56	70.46	89.15	77.00	68.23	74.38	72.60	79.12	74.67	75.38	71.82	73.80
SUGAR (768)	**58.65**	**71.06**	**89.62**	**78.65**	**69.33**	**73.80**	**72.56**	**79.25**	**74.40**	**75.64**	**72.28**	**74.15**

Table 3: Experimental results for Qwen3-Embedding 0.6B to Bert-base distillation. Datasets marked with ⋆ are in-domain. All baselines are evaluated using full-dimensional embeddings.

at lower dimensional budgets is highly competitive with traditional KD baselines that utilize their full dimensionality. Furthermore, when compared at the same maximum capacity, SUGAR consistently surpasses all state-of-the-art KD methods across both in-domain and out-domain benchmarks. This highlights its superior capability in preserving the nested subspace structure and deployment elasticity of the teacher model, ensuring high-quality embeddings even under tight computational constraints. Training configurations and details of KD baselines are provided in Section A.3.2.

Impact of each component in SUGAR Table 4 highlights that all components of SUGAR are essential, with the full model performing best. Removing point-wise alignment ($\mathcal{L}$point-wise) leads to the largest drop, underscoring the role of local semantic anchoring. Excluding macroscopic constraints ($\mathcal{L}$erank, $\mathcal{L}$grass) degrades performance, confirming the importance of global subspace alignment and avoiding collapse. Omitting orthogonal regularization ($\mathcal{L}$ortho) further reduces performance, showing its role in minimizing redundancy.

$\mathcal{L}_{\text{point-wise}}$	$\mathcal{L}_{\text{erank}}$	$\mathcal{L}_{\text{grass}}$	$\mathcal{L}_{\text{ortho}}$	**MRPC**	**Scitail**	**STS15**
Qwen3-Embedding 4B → Bert-base 109M						
	✓	✓	✓	71.85	73.92	73.54
✓		✓	✓	72.14	74.43	75.12
✓	✓	✓		72.45	74.68	75.31
✓	✓		✓	72.82	74.95	75.73
✓	✓	✓	✓	**73.66**	**75.69**	**76.68**

Table 4: Ablation study of the SUGAR framework showing the impact of each distillation component (averaged across all dimensions).

Method	Time (ms/step)	Mean Mem (MB)	Peak Mem (MB)	Std Mem (MB)
SimCSE-unsup	151	2662	3690	235
DistillCSE	182	3373	3674	124
Jasper and Stella	204	2932	3884	71
SUGAR	485	4852	5680	168
EMO	1223	11203	13831	492
DSKD	1002	10991	12393	336
CDM	1078	10814	12156	314

Table 5: Runtime and GPU memory while training comparison for Qwen3-Embedding-4B → BERT-base.

Computational Efficiency. In Table 5, SUGAR strikes a highly practical balance between computational cost and representational quality. While it introduces moderate overhead compared to standard contrastive baselines, due to the calculation of geometric constraints across multiple prefix dimensions, it remains significantly more efficient than complex token-level distillation methods. By operating on dense subspace representations rather than relying on expensive cross-layer attention or optimal transport mechanisms, SUGAR substantially reduces both training latency and memory footprint, making it highly scalable for distilling massive foundation models.

6 Conclusion

We identify the limitations of single-scale distillation for nested embeddings and formalize *Matryoshka Representation Distillation*. We propose SUGAR, a multi-objective framework combining point-wise alignment, effective-rank matching, Grassmannian alignment, and orthogonal regularization to preserve both semantics and hierarchy. Experiments show that SUGAR transfers large models into compact students with strong performance across dimensions.

7 Limitations

While SUGAR successfully distills Matryoshka representations, we acknowledge a few limitations. First, the framework inherently relies on a pre-trained MRL teacher to transfer multi-scale geometry, meaning the student's representational quality is upper-bounded by the teacher's structural properties. Second, computing the macro-geometric alignment objectives, specifically QR decomposition and cross-correlation matrices, introduces moderate computational overhead and memory usage during training, though it crucially adds zero latency during inference. Finally, our current evaluations are primarily conducted on English text representation benchmarks, leaving the exploration of SUGAR's applicability in cross-lingual, multimodal, and generative contexts to future work.

References

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2Vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. 2024. Matryoshka multimodal models.

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2025. Enhancing cross-tokenizer knowledge distillation with contextual dynamical mapping. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8005–8018, Vienna, Austria. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. Senteval: An evaluation toolkit for universal sentence representations.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding.

Alan Edelman, T. A. Arias, and Steven T. Smith. 1998. The geometry of algorithms with orthogonality constraints.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2022. Simcse: Simple contrastive learning of sentence embeddings.

Jiatao Gu, Shuangfei Zhai, Yizhe Zhang, Josh Susskind, and Navdeep Jaitly. 2024. Matryoshka diffusion models.

Mehrtash Harandi, Conrad Sanderson, Chunhua Shen, and Brian C. Lovell. 2013. Dictionary learning and sparse coding on grassmann manifolds: An extrinsic solution.

Liyang He, Chenglong Liu, Rui Li, Zhenya Huang, Shulan Ruan, Jun Zhou, and Enhong Chen. 2025. Refining sentence embedding model through ranking sentences generation with large language models.

Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. 2024. Matryoshka query transformer for large vision-language models.

Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.

Aditya Kusupati, Gantavya Bhatt, Aniket Rege, Matthew Wallingford, Aditya Sinha, Vivek Ramanujan, William Howard-Snyder, Kaifeng Chen, Sham Kakade, Prateek Jain, and Ali Farhadi. 2022. Matryoshka representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 30233–30249. Curran Associates, Inc.

Xianming LI, Zongxi Li, Jing Li, Haoran Xie, and Qing Li. 2025. ESE: Espresso sentence embeddings. In *The Thirteenth International Conference on Learning Representations*.

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality.

Yongyu Mu, Yuzhang Wu, Yuchun Fan, Chenglong Wang, Hengyu Li, Jiali Zeng, Qiaozhi He, Murun Yang, Fandong Meng, Jie Zhou, Tong Xiao, and Jingbo Zhu. 2025. Cross-layer attention sharing for pre-trained large language models.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark.

Sosuke Nishikawa, Ryokan Ri, Ikuya Yamada, Yoshimasa Tsuruoka, and Isao Echizen. 2022. EASE: Entity-aware contrastive learning of sentence embedding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for*

Computational Linguistics: Human Language Technologies, pages 3870–3885, Seattle, United States. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. 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)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Chongyang Tao, Tao Shen, Shen Gao, Junshuo Zhang, Zhen Li, Kai Hua, Wenpeng Hu, Zhengwei Tao, and Shuai Ma. 2025. Llms are also effective embedding models: An in-depth overview.

Minh-Phuc Truong, Hai An Vu, Tu Vu, Nguyen Thi Ngoc Diep, Linh Ngo Van, Thien Huu Nguyen, and Trung Le. 2025. 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*, pages 7594–7606, Suzhou, China. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Jiahao Xu, Wei Shao, Lihui Chen, and Lemao Liu. 2023. DistillCSE: Distilled contrastive learning for sentence embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8153–8165, Singapore. Association for Computational Linguistics.

Dun Zhang, Jiacheng Li, Ziyang Zeng, and Fulong Wang. 2025. Jasper and stella: distillation of sota embedding models.

Songming Zhang, Xue Zhang, Zengkui Sun, Yufeng Chen, and Jinan Xu. 2024. Dual-space knowledge distillation for large language models.

Åke Björck and Gene H. Golub. 1973. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594.

A Experimental Details

A.1 Model Architectures

To evaluate the scalability and generalizability of the SUGAR framework, we conduct distillation experiments using a diverse range of student backbone architectures varying in size and capacity. These student models include the compact **TinyBERT 6L**, the standard 12-layer **BERT-base** encoder, and the high-performance **BGE-M3**. As our knowledge sources, we utilize massive, open-source embedding models that are inherently pre-trained with Matryoshka representations, specifically **Qwen3-Embedding-0.6B** and **Qwen3-Embedding-4B**. By transferring the rich, multi-scale geometry of these large-scale teacher models into compact student encoders, we demonstrate that SUGAR consistently optimizes the internal organization of semantic information regardless of the target architecture's depth or parameter count.

A.2 Detailed Dataset Statistics

Table 6: Dataset Statistics for training and test set

Dataset	Train (Sampled)	Test Size
Banking77	3,000	3,080
TweetEval	3,000	3434
Emotion (OOD)	-	1,990
MRPC	1,500	1,730
WiC	1,500	1,400
SciTail (OOD)	-	2,130
SICK	1,500	4,823
STS-B	1,500	1,390
STS12 (OOD)	-	3,108
STS13 (OOD)	-	1,500
STS14 (OOD)	-	3,750
STS15 (OOD)	-	3,000
STS16 (OOD)	-	1,186
SickR (OOD)	-	9,927

We use a diverse collection of datasets for both training and evaluation, covering text classification, natural language inference (NLI), and semantic textual similarity (STS). To build training data for contrastive sentence representation learning, we sample sentences from multiple task categories to promote domain and objective diversity. Specifically, we collect 6,000 sentences from classification datasets (3,000 per dataset), 3,000 sentence pairs from STS datasets (1,500 per dataset), and 3,000 sentence pairs from pair classification datasets (1,500 per dataset). All sentence pairs are flattened into individual sentences, resulting in 24,000 unique sentences, which are then used to train the backbone encoder with unsupervised SimCSE-style contrastive learning under a unified training framework. For evaluation, we test the trained models on both in-domain test sets and unseen out-of-domain datasets, including Emotion, SciTail, and multiple STS benchmarks (STS12–STS16 and SickR), which differ from the training data in domain, style, and annotation protocol. Dataset statistics are reported in Table 6, and all baseline methods (MRL, ESE) are trained on the same training corpus and evaluated on the same set of test datasets for fair comparison

A.3 Training Configurations

A.3.1 Matryoshka experiment

The detailed training configurations for the SUGAR framework across our various backbones are summarized in Table 8. We maintain a consistent setup to ensure a fair comparison across different scales, applying specific hyperparameters like α to trade off MRL and SUGAR losses. We explored the loss

balancing hyperparameter α over the set $\{0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7\}$. The optimal configurations for each backbone are also reported in Table 8.

Table 7: Detailed training configurations for SUGAR across different backbones.

Configuration	TinyBERT-6L	BERT-base	BGE-M3
Epochs	5	5	5
Learning Rate	2×10^{-5}	2×10^{-5}	2×10^{-5}
Max Length	256	256	256
Batch Size	16	16	16
LR Scheduler	Cosine	Cosine	Cosine
Optimizer	AdamW	AdamW	AdamW
Temperature (τ)	0.05	0.05	0.05
α (MRL weight)	0.4	0.4	0.5

Task Objective Specification. In our experimental setting, the task-specific loss component within the MRL objective (Eq. 2) is instantiated as $\mathcal{L}_{\text{SimCSE}}$, the unsupervised contrastive loss adopted from (Gao et al., 2022) . Given a batch of input sentences, we feed them through the student encoder twice with different standard dropout masks z, z' to obtain two views of embeddings e_i^z and $e_i^{z'}$. The loss is formulated as:

$$\mathcal{L}_{\text{SimCSE}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{e^{\text{sim}(e_i^z, e_i^{z'})/\tau}}{\sum_{j=1}^{N} e^{\text{sim}(e_i^z, e_j^{z'})/\tau}}, \tag{16}$$

where N is the batch size, τ is the temperature hyperparameter, and $\text{sim}(\cdot)$ denotes cosine similarity.

A.3.2 Distillation experiments

Baselines. To thoroughly evaluate the proposed framework, we compare it with representative knowledge distillation methods across multiple levels of granularity. For **token-level KD**, we consider DSKD (Zhang et al., 2024), CDM (Chen et al., 2025), and EMO (Truong et al., 2025), which transfer knowledge by aligning fine-grained token representations or relational structures between teacher and student models. For **sentence-level KD**, we benchmark against JASPER & STELLA (Zhang et al., 2025) and DISTILLCSE (Xu et al., 2023), strong baselines that distill knowledge solely through global sentence embeddings without token-level supervision.

Training Configuration The detailed training configurations for all methods are summarized in Table 8. All knowledge distillation (KD) experiments are conducted on the same dataset constructed as described in Section A.2, ensuring a fair comparison across methods. Evaluation is the same as in Section A.4.

Settings	DSKD	CDM	EMO	Stella & Jasper	DistillCSE	SUGAR
Epoch	5	5	5	2 + 5 (two-stage)	5	5
LR	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}	2×10^{-5}
Batch Size	32	32	32	32	32	32
LR Scheduler	Cosine	Cosine	Cosine	Cosine	Cosine	Cosine

Table 8: Training configurations for different distillation and sentence embedding methods.

Baseline Implementation Details To ensure a rigorous evaluation, we adapt state-of-the-art generative distillation methods to the discriminative sentence embedding setting.

Contextual Dynamic Mapping (CDM). We apply CDM to align hidden representations $\mathbf{H} \in \mathbb{R}^{B \times L \times D}$. The original formulation relies on the teacher's predictive entropy to modulate the Dynamic Time Warping (DTW) alignment cost. However, since standard sentence encoders do not readily output token-level vocabulary distributions, we modify the cost function to utilize *token edit distance* (Levenshtein distance).

Specifically, we compute a pairwise distance matrix between the student and teacher token sequences and employ DTW to identify the optimal monotonic alignment path that minimizes the cumulative cost. Based on this path, teacher hidden states are aggregated (via averaging) to match the student's sequence length, establishing a synchronized one-to-one correspondence for the final Mean Squared Error (MSE) loss.

Dual Space Knowledge Distillation (DSKD). We extend DSKD to enforce geometric consistency through a symmetric, bidirectional process. Unlike standard unidirectional approaches, DSKD projects representations into each model's respective latent space. We implement this at two granularities: (1) *Sentence-level:* Global representations (e.g., `[CLS]` tokens) are distilled via learnable linear projection heads to align the global semantic spaces. (2) *Token-level:* To address the misalignment caused by differing vocabularies and sequence lengths, we incorporate Cross-Layer Attention (CLA) (Mu et al., 2025). CLA computes an attention mechanism between the student's and teacher's intermediate layers, allowing the student to reconstruct fine-grained teacher features (and vice versa) without requiring rigid token-to-token matching.

A.4 Evaluation

To assess the efficacy of the learned sentence embeddings, we conduct evaluations across three distinct categories of downstream tasks. First, for classification tasks, we follow the standard protocol established by (Conneau and Kiela, 2018), which involves training a Logistic Regression classifier on top of frozen sentence representations. Second, in pair classification, we generate predictions by applying an optimal threshold to the cosine similarity of the sentence pairs, measuring performance via Accuracy. Finally, for Semantic Textual Similarity (STS), we evaluate the alignment between the embeddings' cosine similarity and human-annotated gold standards using Spearman correlation. For **Matryoshka experiments**, we evaluate the quality of representations across a fixed set of nested dimensions $\mathcal{D} = \{16, 32, 64, 128, 256, 512, 768, 1024\}$. For **KD experiments**, baseline methods are evaluated using their full embedding dimensionality. Across tasks, we report F1 Score for Text Classification, Accuracy for NLI and pair classification tasks, and Spearman Correlation Coefficient for STS benchmarks

B Additional Results

Dim.	Data	BGE-M3 (568M)			Data	BGE-M3 (568M)		
		MRL	ESE	SUGAR		MRL	ESE	SUGAR
16	**Banking77**	73.06	72.07	**74.35**	**WiC**	61.35	61.53	**63.00**
32		86.09	84.91	**86.38**		61.78	61.23	**62.57**
64		90.53	**90.67**	90.34		62.14	**63.12**	63.07
128		92.10	92.02	**92.77**		63.85	64.32	**64.87**
256		92.71	92.06	**92.98**		64.21	**64.28**	63.71
512		92.93	92.13	**93.82**		64.71	64.21	**65.21**
1024		93.11	92.25	**93.67**		65.38	64.42	**65.43**
16	**TweetEval**	56.91	55.32	**57.53**	**SICK**	69.66	68.35	**72.91**
32		61.72	59.95	**62.30**		71.03	69.48	**74.00**
64		**66.80**	65.95	66.70		71.63	70.27	**75.09**
128		70.06	69.67	**70.28**		71.93	70.83	**75.64**
256		**72.66**	71.90	71.86		72.46	71.08	**75.91**
512		72.61	72.69	**73.83**		72.78	71.24	**76.02**
1024		72.95	74.13	**75.87**		72.77	71.23	**76.23**
16	**MRPC**	67.21	67.43	**69.22**	**STSB**	75.21	73.96	**76.46**
32		67.36	67.65	**69.57**		77.71	76.13	**79.44**
64		67.47	67.69	**68.99**		79.50	77.61	**81.05**
128		67.39	67.47	**69.39**		80.59	78.55	**81.92**
256		67.42	67.88	**69.51**		82.44	79.40	**82.76**
512		67.71	67.59	**69.39**		81.28	80.94	**82.88**
1024		67.59	67.53	**70.09**		82.52	80.52	**83.18**

Table 9: Results on in-domain datasets comparing SUGAR with MRL and Espresso (ESE) using the BGE-M3 (568M) backbone.

To rigorously evaluate the scalability and architectural generalizability of our framework, we extend our distillation experiments to the larger and more complex BGE-M3 (568M) backbone. Unlike compact

Dim.	Data	BGE-M3 (568M)			Data	BGE-M3 (568M)		
		MRL	Espresso	SUGAR		MRL	Espresso	SUGAR
16		62.95	61.97	**70.59**		73.16	72.77	**75.22**
32		67.02	64.78	**71.87**		74.67	75.54	**79.21**
64		67.99	65.85	**73.94**		76.96	77.10	**80.38**
128	**STS12**	69.38	66.83	**74.54**	**STS16**	78.88	77.75	**81.65**
256		70.82	67.73	**75.30**		80.91	78.48	**81.97**
512		71.64	68.80	**75.59**		**82.13**	80.20	81.99
1024		72.11	68.56	**75.60**		82.31	79.97	**83.26**
16		**75.32**	73.32	73.89		70.49	69.87	**73.07**
32		77.20	75.60	**78.84**		73.43	72.06	**74.49**
64		79.66	77.88	**80.90**		75.43	72.76	**75.65**
128	**STS13**	80.96	78.98	**82.17**	**SickR**	**76.52**	75.12	76.35
256		82.42	80.30	**83.34**		77.03	76.41	**77.66**
512		83.54	81.76	**84.03**		77.56	76.55	**77.89**
1024		84.03	81.79	**84.32**		77.68	76.48	**77.91**
16		67.07	65.03	**67.85**		29.92	28.46	**30.32**
32		69.43	67.28	**70.15**		36.94	33.94	**37.54**
64		72.46	68.96	**72.96**		43.45	44.53	**45.13**
128	**STS14**	73.38	70.18	**73.88**	**Emotion**	49.84	50.05	**51.84**
256		74.45	71.39	**74.95**		54.07	53.41	**56.91**
512		74.93	72.45	**75.43**		58.73	55.59	**60.35**
1024		75.33	72.35	**75.83**		61.39	60.12	**62.42**
16		74.57	75.44	**75.94**		80.62	80.47	**82.78**
32		78.46	77.98	**78.96**		81.74	81.32	**85.32**
64		81.39	80.14	**81.89**		83.58	81.60	**85.89**
128	**STS15**	82.29	80.46	**82.79**	**SciTail**	84.43	82.03	**87.25**
256		82.90	81.10	**83.40**		84.19	82.03	**87.77**
512		83.19	81.45	**83.69**		85.03	82.30	**88.19**
1024		83.62	81.69	**84.12**		85.46	82.69	**88.24**

Table 10: Results on out-domain datasets comparing SUGAR with MRL and Espresso using the BGE-M3 (568M) backbone.

encoders such as TinyBERT, BGE-M3 possesses a significantly larger parameter space and natively captures highly sophisticated semantic representations, making it a challenging testbed for representation distillation. As demonstrated in Tables 9 and 10, SUGAR consistently outpaces both the standard MRL formulation and the competitive Espresso (ESE) baseline across all evaluated prefix dimensions. This consistent superiority is evident not only on in-domain datasets—where the student successfully replicates the teacher's localized semantic knowledge—but also broadly across out-domain benchmarks, showcasing robust zero-shot transferability and strong generalization.

Crucially, the performance advantages of our approach are maintained seamlessly from the most aggressively compressed dimensions (e.g., $d = 16$ or 32) up to the full capacity of the 1024-dimensional space. This observation validates that our dual-level geometric alignment, operating at both the macroscopic subspace level and the microscopic point-wise level,effectively prevents dimensional collapse and preserves the nested information hierarchy, even when the student model has a vast representational capacity. Overall, these findings confirm that the effectiveness of SUGAR is highly robust to the depth, size, and architectural complexity of the target backbone. By successfully transferring the multi-scale geometry of massive LLM-based teachers into a high-performance student encoder, our framework proves to be a highly scalable solution for modern embedding deployment.

C Use of Large Language Models

We acknowledge the use of Large Language Models (LLMs) during the preparation of this manuscript. These tools were utilized strictly for stylistic refinement, grammatical editing, and improving the overall flow of the presentation. At no point were LLMs used to conceptualize research ideas, formulate hypotheses, or interpret experimental findings. The authors remain fully accountable for the integrity and accuracy of the final content.