Toward a Unified Statistical Theory of Multi-View Pretraining for Neural Knowledge Graph Learning

Jifan Zhang
Department of Statistics and Data Science
Northwestern University

Miklós Z. Rácz
Department of Statistics and Data Science
Department of Computer Science
Northwestern University

Suqi Liu
Department of Statistics
University of California, Riverside

Abstract

Knowledge graph learning offers a powerful framework for representing and inferring structured knowledge, with broad practical applications. Two obstacles limit current methods: labeled triples are scarce for each entity and relation, which makes expressive models hard to train, and scoring functions are designed ad hoc, which limits their generalizability and leaves them without theoretical grounding. We address both obstacles with a theoretically grounded framework, with end-to-end guarantees, that extends and subsumes existing methods. The framework proceeds in two stages: multi-view unsupervised pretraining that integrates heterogeneous sources of entity side information into a shared representation, followed by supervised learning across multiple relation types. We establish a unified nonasymptotic risk bound that combines the subspace recovery error of pretraining and the sample complexity of the labeled data in a single guarantee, quantifying formally how large-scale unlabeled data from multiple views benefits downstream knowledge prediction. Synthetic experiments validate each component of the theory, and experiments on large-scale real-world knowledge graph benchmarks confirm the effectiveness of our approach.

1 Introduction

Knowledge graphs (KGs) organize information as a network of entities and the relations between them. They have become the standard representation of relational knowledge in domains ranging from question answering and recommendation to information extraction and biomedical discovery [Yang et al., 2022; Zhang et al., 2016; Huang et al., 2019; Bonner et al., 2022]. Large-scale KGs such as the Google Knowledge Graph and the Unified Medical Language System (UMLS) contain millions to billions of entities spanning dozens of heterogeneous relation types, making scalable and accurate KG prediction a central problem in machine learning [Ji et al., 2022]. Despite their scale, real-world KGs are sparse: the observed triples cover only a small fraction of all possible entity pairs, and this sparsity hinders the learning of KG prediction models. At the same time, entities in these graphs usually carry rich unlabeled side information—textual descriptions, ontology memberships, database annotations—that is readily available even when labeled triples are scarce.

The dominant approach to KG prediction is embedding-based: learn low-dimensional entity and relation vectors so that a translational, bilinear, or rotational scoring function ranks observed triples above negatives [Bordes et al., 2013; Yang et al., 2015; Trouillon et al., 2016; Sun et al., 2019]. A substantial empirical literature further augments these representations with textual descriptions,

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

pretrained language models, or other entity attributes [Xie et al., 2016; Yao et al., 2019; Wang et al., 2021b, 2022], demonstrating that side information is valuable in practice, especially when labeled triples are scarce. However, these methods are typically tailored to specific applications, rely on domain-specific inductive biases, and largely lack theoretical guarantees. Recently, Liu et al. [2024] established nonasymptotic risk bounds for neural KG learning with fixed and trainable entity representations, and Ge et al. [2024] showed that unsupervised pretraining reduces labeled-sample complexity in generic latent-variable models. We build on these developments to establish a unified statistical framework connecting unsupervised pretraining to supervised KG prediction.

A fundamental challenge in KG learning is the tension between data scarcity and model capacity: labeled triples are typically limited, yet flexible representation-based models require many parameters. How to systematically exploit diverse unlabeled datasets with heterogeneous side information, and what provable benefits doing so confers, therefore remain critical open questions. We address these questions with a theoretically grounded two-stage framework, which we call the Pretrained Neural Knowledge Graph (PNKG) framework. In the first stage, side information on entities from multiple views is mapped through view-specific positive-semidefinite kernels, and kernel principal component analysis (KPCA) on a weighted aggregate of their Gram matrices yields low-dimensional entity coordinates. The kernel formulation also offers a natural interface to pretrained language encoders: each frozen encoder induces a positive-semidefinite kernel on entity descriptions through the inner product of its feature map, consistent with the kernel perspective on deep architectures of Cho and Saul [2009]. The PNKG framework aggregates these encoder-induced Gram matrices, so that views with different architectures and output dimensions can be combined on their common entity index set. In the second stage, these embeddings serve as inputs to ReLU networks trained on labeled triples. The weighted direct sum of the view-specific reproducing kernel Hilbert spaces (RKHSs) provides a unified representation space for heterogeneous side information, such as text descriptions and metadata, drawn from diverse datasets, while the ReLU networks model the scoring function flexibly without predefined domain-specific priors.

We establish two main theoretical results for this framework. Theorem 1 gives a nonasymptotic perturbation bound for the KPCA pretraining stage, showing that the learned entity subspace is accurate when the multi-view kernel residuals are small relative to the effective signal separation of the weighted population geometry. The result covers both inner-product and Gaussian distance kernels in admissible information-plus-noise regimes, and it separates finite-sample kernel error from the population bias that a misspecified kernel–noise geometry may induce. It also motivates a signal-adjusted inverse-variance principle for weighting views: views with strong effective signal, small deterministic bias, and low fluctuation variance should receive larger weights, and noisy or systematically biased views should be downweighted. This subspace recovery result supplies the pretraining component of our end-to-end statistical theory and may be of independent interest. Our main result, Theorem 2, is a unified end-to-end risk bound that decomposes the downstream KG prediction error additively into neural approximation error, supervised estimation error, optimization error, and pretraining error. It makes explicit which term is the bottleneck as the unlabeled and labeled sample sizes and the model capacity vary, and it attributes each source of error to a specific stage of the pipeline or a specific modeling or algorithmic choice.

Simulation studies on synthetic data further support the theoretical findings. We also evaluate the framework on two real-world KGs: WordNet [Miller, 1995], a lexical KG whose synset definitions provide textual side information, and RENKI-KG [Liu et al., 2024], a large-scale medical KG that integrates multiple curated clinical ontologies and drug databases. On both, multi-view pretrained neural KG models outperform graph-only baselines and improve on their single-view counterparts.[1] These results corroborate the theory and demonstrate the practical utility of the framework.

In summary, our contributions are threefold. (i) We propose a generic two-stage framework that combines multi-view KPCA for unsupervised pretraining with ReLU networks for supervised KG learning. (ii) We prove an end-to-end transfer theorem with nonasymptotic bounds that connects pretraining error to supervised KG prediction risk. (iii) We validate the theory with controlled simulations that isolate each error component and with real-world experiments on benchmark KGs.

[1] Code is available at https://github.com/jifanzhang999/PNKG.

2 Related Work

Knowledge graph embeddings. Most KG completion methods learn entity and relation embeddings together with a relation-specific scoring function. TransE [Bordes et al., 2013] models relations as translations; DistMult [Yang et al., 2015] and ComplEx [Trouillon et al., 2016] use bilinear interactions; and RotatE [Sun et al., 2019] represents relations as rotations in complex space. Nickel et al. [2016] and Ji et al. [2022] survey these approaches. Although they are effective graph-only baselines, they rely mainly on observed triples and do not directly exploit multiple unlabeled sources of entity side information, which can be valuable when labels for the target relations are scarce or entity coverage is uneven.

Text- and side-information-enhanced KG completion. Several approaches enrich KG prediction with entity descriptions or pretrained language models. DKRL [Xie et al., 2016] encodes entity descriptions alongside KG embeddings, KG-BERT [Yao et al., 2019] scores triples written as text, and KEPLER [Wang et al., 2021b] learns textual and KG representations jointly. BLP [Daza et al., 2021] learns inductive representations, StAR [Wang et al., 2021a] couples textual encoding with structure-aware scoring, and SimKGC [Wang et al., 2022] applies contrastive learning to text-based KG completion. Rather than proposing another scoring architecture, we study how representations estimated from *unlabeled, heterogeneous views* affect the risk of a downstream relation-specific predictor. Our guarantees cover this two-stage setting with fixed pretrained coordinates, not end-to-end fine-tuning of the encoders; this separation isolates the statistical contribution of unsupervised representation estimation.

Statistical theory for KG and network learning. Latent-position network models [Hoff et al., 2002] and shared-structure models for multilayer networks [MacDonald et al., 2022] provide theoretical foundations for estimating latent geometry, typically under specified link functions or structured observation models. Such assumptions can be restrictive for KGs with heterogeneous relations and unknown scoring functions. Most closely related, Liu et al. [2024] formulate neural KG prediction as nonparametric regression over triples and derive nonasymptotic in-sample and out-of-sample MSE bounds. We complement that supervised analysis: instead of treating the entity coordinates as observed features, we quantify the error in recovering them from unlabeled multi-view data and how this error propagates to the downstream KG risk.

Theory of unsupervised pretraining. Theories of contrastive representation learning [Saunshi et al., 2019; HaoChen et al., 2021] establish downstream benefits under latent-class or augmentation assumptions, and prediction-based pretraining [Lee et al., 2021] can reduce the number of labeled samples needed under suitable structural conditions. Closest in spirit, Ge et al. [2024] separate the unlabeled and labeled contributions to excess risk under an informative-representation condition. Whereas such generic analyses abstract away the representation estimator, we analyze a specific one, weighted kernel PCA over multiple views, in a KG setting with heterogeneous views and relation-labeled triples: Theorem 2 tracks the multi-view pretraining error through a neural KG regression bound. Because the models and assumptions differ, rates are not directly comparable across these analyses.

Deep feature kernels and spectral aggregation. A frozen encoder defines a feature map ϕ_s and an induced kernel $k_s(u, u') = \langle \phi_s(u), \phi_s(u') \rangle$, so inner products of frozen language-model embeddings evaluate an encoder-induced kernel. This links our procedure to the deep-kernel perspectives of Cho and Saul [2009] and Wilson et al. [2016]. Our aggregation also shares with distributed PCA [Fan et al., 2019] the goal of recovering a common spectral subspace, but the two operate differently: distributed PCA averages local eigenspace projectors, whereas PNKG aggregates the view-specific entity-kernel matrices *before* the eigendecomposition. This distinction motivates our analysis of the view weights and the signal eigengap.

3 The PNKG Framework

We now formalize the two-stage PNKG framework. The first stage uses unlabeled side information on the entities from multiple sources to construct low-dimensional entity coordinates. The second stage takes these coordinates as inputs and learns relation-specific ReLU networks from labeled triples.

3.1 Multi-view pretraining via weighted KPCA

Let $[N] := \{1, \dots, N\}$ denote the common entity set. Side information is collected from several sources, called views, and $[S] := \{1, \dots, S\}$ indexes them. For each view $s \in [S]$, the side information is organized into a matrix

$$E_s = \begin{pmatrix} u_{s,1}^\top \\ \vdots \\ u_{s,N}^\top \end{pmatrix} \in \mathbb{R}^{N \times m_s},$$

where m_s is the feature dimension of view s and $u_{s,j} \in \mathbb{R}^{m_s}$ is the data vector of entity j in view s. For example, view s may be a hospital with m_s patients and j may index diagnoses, so that $u_{s,j}$ records information on all patients with diagnosis j in the medical records of hospital s.

Our theoretical analysis is motivated by an information-plus-noise structure: for each view s, we model the observed entity vector as

$$u_{s,j} = v_{s,j} + \sigma_{s,j} \frac{\varepsilon_{s,j}}{\sqrt{m_s}}, \qquad j = 1, \dots, N, \tag{3.1}$$

where $v_{s,j} \in \mathbb{R}^{m_s}$ is the latent signal vector of entity j, $\varepsilon_{s,j} \in \mathbb{R}^{m_s}$ is high-dimensional noise, and $\sigma_{s,j}$ is a view-specific noise scale. The normalization $m_s^{-1/2}$ keeps the noise energy bounded as the feature dimension grows.

A linear kernel example. We first specialize (3.1) to the simplest, linear isotropic setting. Assume that the noise scale is constant within each view, $\sigma_{s,j} = \sigma_s$, and that

$$\varepsilon_{s,j} \sim N(0, I_{m_s}), \qquad j = 1, \dots, N.$$

Then

$$E_s = V_s + \frac{\sigma_s}{\sqrt{m_s}} \xi_s, \qquad \xi_s(j, \cdot) \overset{\text{iid}}{\sim} N(0, I_{m_s}),$$

where the jth row of V_s is $v_{s,j}^\top$. The corresponding linear entity-similarity matrix is

$$K_s^{\text{lin}} = \frac{1}{N} E_s E_s^\top \in \mathbb{R}^{N \times N}. \tag{3.2}$$

With multiple views, we aggregate the entity-similarity matrices as

$$K^{\text{lin}}(w) = \sum_{s=1}^{S} w_s K_s^{\text{lin}}, \qquad w_s \geq 0, \quad \sum_{s=1}^{S} w_s = 1.$$

The leading eigenvectors of $K^{\text{lin}}(w)$ then provide low-dimensional entity coordinates.

General kernels. The linear construction extends to any positive-semidefinite kernel. For each view $s \in [S]$, let

$$k_s : \mathbb{R}^{m_s} \times \mathbb{R}^{m_s} \to \mathbb{R}$$

be a positive-semidefinite kernel. There then exist a Hilbert space $\mathcal{H}_s$ and a feature map

$$\phi_s : \mathbb{R}^{m_s} \to \mathcal{H}_s$$

such that

$$k_s(u, u') = \langle \phi_s(u), \phi_s(u') \rangle_{\mathcal{H}_s}.$$

We define the view-specific entity-kernel matrix by

$$K_s^{\text{obs}}(j, j') := \frac{1}{N} k_s(u_{s,j}, u_{s,j'}) = \frac{1}{N} \langle \phi_s(u_{s,j}), \phi_s(u_{s,j'}) \rangle_{\mathcal{H}_s}, \qquad j, j' \in [N].$$

Thus K_s^{obs} captures the (possibly nonlinear) geometry that view s induces on the common entity set. Examples include inner-product kernels of the form $k_s(u, u') = f_s(u^\top u')$ and distance kernels of the form $k_s(u, u') = f_s(\|u - u'\|_2^2)$. The linear construction above is the special case $k_s(u, u') = u^\top u'$, $\phi_s(u) = u$.

Algorithm 1 Multi-view kernel pretraining

Require: Entity-feature observations $\{u_{s,j} \in \mathbb{R}^{m_s}, j \in [N]\}_{s=1}^{S}$, kernels $\{k_s\}_{s=1}^{S}$, rank d, and weights $w \in \Delta_S$.

Ensure: Frozen entity coordinates $\widehat{Z} \in \mathbb{R}^{N \times d}$.

1: **for** $s = 1, \ldots, S$ **do**
2: $\quad K_s^{\mathrm{obs}}(j, j') \leftarrow \frac{1}{N} k_s(u_{s,j}, u_{s,j'})$ for all $j, j' \in [N]$.
3: **end for**
4: $K^{\mathrm{obs}}(w) \leftarrow \sum_{s=1}^{S} w_s K_s^{\mathrm{obs}}$.
5: $\widehat{Z} \leftarrow \mathrm{TopEig}_d\big(K^{\mathrm{obs}}(w)\big)$.
6: **return** $\widehat{Z}$.

For weights $w = (w_1, \ldots, w_S)$ in the probability simplex Δ_S, define the weighted multi-view entity-kernel matrix

$$K^{\mathrm{obs}}(w) := \sum_{s=1}^{S} w_s K_s^{\mathrm{obs}}, \qquad w_s \geq 0, \quad \sum_{s=1}^{S} w_s = 1. \tag{3.3}$$

Because the aggregation is performed on the common entity index set, it is well defined even when the input spaces and feature dimensions of the views differ, and it is itself a positive-semidefinite kernel on the common entity set. Appendix A.2 gives the weighted direct-sum construction and interprets frozen encoders as feature maps (Remark 1); Appendix A.1 revisits the linear example.

Spectral entity coordinates. The pretrained entity coordinates come from the leading eigenspace of $K^{\mathrm{obs}}(w)$. For a symmetric matrix G, let $\mathrm{TopEig}_d(G)$ denote any $N \times d$ matrix whose columns form an orthonormal basis of the eigenspace associated with the d largest eigenvalues of G. Let

$$\widehat{Z} := \mathrm{TopEig}_d\big(K^{\mathrm{obs}}(w)\big) \in \mathbb{R}^{N \times d}, \qquad \widehat{Z}^\top \widehat{Z} = I_d.$$

The supervised stage uses $\widehat{Z}$ as the frozen entity-coordinate matrix, whose jth row is the pretrained coordinate of entity j; the choice of d is discussed in Remark 2. Our analysis assumes that the population signal kernels share a rank-d eigenspace Z_{sig}, with view-specific signal strengths Λ_s and residuals R_s^{sig} on its orthogonal complement; see (A.1) in Appendix A.3. Algorithm 1 summarizes the multi-view kernel pretraining procedure.

3.2 Supervised learning with relation-wise networks

The supervised stage uses labeled (head, relation, tail) triples, abbreviated as (h, r, t) where h and t are entities and r is the relation between them. Let $[K] = \{1, \ldots, K\}$ be the relation set and $\mathcal{X} = [N] \times [K] \times [N]$ be the space of all possible triples. We observe labeled samples $\mathcal{D}_{\mathrm{sup}} = \{(X_i, Y_i)\}_{i=1}^{n}$, where

$$X_i = (h_i, r_i, t_i) \in \mathcal{X} \quad \text{and} \quad Y_i = f^\star(X_i) + \varepsilon_i. \tag{3.4}$$

We assume that $\mathbb{E}[\varepsilon_i \mid X_i] = 0$. The true target function $f^\star : \mathcal{X} \to \mathbb{R}$ is the KG scoring function. For binary link prediction, $f^\star(X_i)$ can be interpreted as the conditional probability that triple X_i is true.

Given the pretrained embeddings $\widehat{Z} = (\widehat{z}_1, \ldots, \widehat{z}_N)^\top$, we learn one bounded ReLU network for each relation. Let $\Pi_B : \mathbb{R} \to [-B, B]$ be the clipping map. For relation r, let $F_r : \mathbb{R}^{2d} \to \mathbb{R}$ be a ReLU network with depth at most D and at most W trainable parameters, and define

$$f_r(a, b) := \Pi_B\{F_r([a^\top, b^\top]^\top)\}. \tag{3.5}$$

The neural knowledge graph model is

$$f(\widehat{Z}; h, r, t) = f_r(\widehat{z}_h, \widehat{z}_t). \tag{3.6}$$

Let $\mathcal{F}_{\widehat{Z}}(D, W, B)$ denote the neural knowledge graph function class given $\widehat{Z}$. We work with a predictor $\widehat{f}(\widehat{Z}; \cdot) \in \mathcal{F}_{\widehat{Z}}(D, W, B)$ whose empirical risk is within δ_{opt} of the best empirical risk over the class

$$\frac{1}{n} \sum_{i=1}^{n} \{Y_i - \widehat{f}(\widehat{Z}; X_i)\}^2 \leq \inf_{f \in \mathcal{F}_{\widehat{Z}}(D,W,B)} \frac{1}{n} \sum_{i=1}^{n} \{Y_i - f(\widehat{Z}; X_i)\}^2 + \delta_{\mathrm{opt}}. \tag{3.7}$$

In practice, $\widehat{f}$ is computed with standard optimizers such as stochastic gradient descent.

4 Theoretical Results

Building on the two-stage construction in Section 3, we first bound the subspace error of the weighted KPCA coordinates $\widehat{Z}$ and then propagate this error to the downstream relation-specific predictor. We introduce only the quantities needed to state the guarantees; the precise kernel-dependent definitions and the proofs are in Appendices B and C.

4.1 Subspace recovery for multi-view pretraining

Assumptions. We work under the information-plus-noise model (3.1) with the weighted kernel (3.3). The central structural condition is a *shared signal eigenspace*: the noiseless kernel matrices $K_s^{\mathrm{sig}}(j,j') := N^{-1}k_s(v_{s,j}, v_{s,j'})$ decompose as

$$K_s^{\mathrm{sig}} = Z_{\mathrm{sig}}\Lambda_s Z_{\mathrm{sig}}^\top + R_s^{\mathrm{sig}}, \qquad s \in [S],$$

where $Z_{\mathrm{sig}} \in \mathbb{R}^{N\times d}$ with $Z_{\mathrm{sig}}^\top Z_{\mathrm{sig}} = I_d$ spans the common d-dimensional entity eigenspace, the positive-semidefinite matrix $\Lambda_s \in \mathbb{R}^{d\times d}$ gives the signal strength of view s along it, and the symmetric residual R_s^{sig}, which satisfies $R_s^{\mathrm{sig}} Z_{\mathrm{sig}} = 0$, captures view-specific variation in its orthogonal complement. The views need not be equally informative; they need only share this leading eigenspace. In addition, the recovery guarantee requires the noise in each view to satisfy suitable concentration or Gaussian conditions, the shared signal to be separated from the residual directions, and the views to be independent. The precise statements are Assumptions 1–5 in Appendix B. Throughout this subsection, the weights $w \in \Delta_S$ are fixed independently of the pretraining data.

Kernel-dependent errors and effective eigengap. Noise affects each kernel family differently. Let $\mathcal{S}_{\mathrm{lin}}$, $\mathcal{S}_{\mathrm{gauss}}$, and $\mathcal{S}_{\mathrm{ip}}$ denote the sets of linear, Gaussian distance-kernel, and nonlinear inner-product views, respectively. Four view-dependent quantities enter the bound:

- $\alpha_s > 0$ is a *signal multiplier* that captures the attenuation caused by the kernel and the noise. It is 1 for linear and nonlinear inner-product kernels, and $e^{-2\tau_s\sigma_s^2}$ for a Gaussian kernel with bandwidth τ_s and homogeneous noise scale σ_s.
- $b_{s,N,m_s} \geq 0$ measures the *deterministic kernel bias* of the population approximation to the noisy kernel.
- $d_{s,N,m_s} \geq 0$ measures a *nonconstant diagonal correction* for nonlinear inner-product kernels; it vanishes in the other two regimes.
- $v_{s,N,m_s} \geq 0$ measures the *random kernel fluctuation* around its population approximation; v_{s,N,m_s}^2 is the variance scale that enters the bound.

The kernel-specific formulas, including the required noise parameters, are given in Appendix B.3.

The separation between the *weighted* signal and residual parts is measured by

$$\gamma(w) := \lambda_{\min}\left(\sum_{s=1}^{S} w_s\alpha_s\Lambda_s\right) - \left\|\sum_{s=1}^{S} w_s\alpha_s R_s^{\mathrm{sig}}\right\|_{\mathrm{op}}. \tag{4.1}$$

Here $\lambda_{\min}$ is the smallest eigenvalue of the $d\times d$ weighted signal matrix and $\|\cdot\|_{\mathrm{op}}$ is the spectral norm. Thus $\gamma(w)$ is a *lower bound on the effective eigengap*: the larger it is, the more clearly the common signal subspace stands out from the residual directions. The separation assumption guarantees $\gamma(w) > 0$ for the weights under consideration.

We collect the two types of error into

$$\mathcal{B}(w) := \sum_{s=1}^{S} w_s b_{s,N,m_s} + \sum_{s\in\mathcal{S}_{\mathrm{ip}}} w_s d_{s,N,m_s}, \qquad \mathcal{V}(w) := \sum_{s=1}^{S} w_s^2 v_{s,N,m_s}^2, \tag{4.2}$$

and define the normalized pretraining-error quantity

$$\mathcal{E}_{\mathrm{pt}}(w,\delta) := \frac{\mathcal{B}(w)^2 + \delta^{-1}\mathcal{V}(w)}{\gamma(w)^2}, \qquad 0 < \delta < 1. \tag{4.3}$$

Here δ controls the failure probability. The squared weights in $\mathcal{V}(w)$ reflect the independence of the fluctuations across views, whereas the deterministic errors accumulate linearly before being squared.

Theorem 1 (Subspace recovery for weighted entity-kernel PCA)**.** *Suppose Assumptions 1–5 hold and $w \in \Delta_S$ is deterministic. For every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\left\| \sin\Theta(\widehat{Z}, Z_{\mathrm{sig}}) \right\|_F^2 \le C\, \mathcal{E}_{\mathrm{pt}}(w,\delta) = \frac{C}{\gamma(w)^2}\left[\mathcal{B}(w)^2 + \frac{\mathcal{V}(w)}{\delta}\right]. \tag{4.4}$$

Here $C > 0$ depends only on the view-level noise constants, not on N, m_s, w, or δ.

The matrix $\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})$ has as its singular values the sines of the principal angles between the estimated and target d-dimensional subspaces. Theorem 1 separates (i) the deterministic bias $\mathcal{B}(w)^2$, (ii) the random fluctuation $\mathcal{V}(w)/\delta$, and (iii) the amplification of both by the inverse squared eigengap $\gamma(w)^{-2}$. A view is useful not simply because it has many features or low noise, but because its noise is small relative to the signal separation it contributes.

Rates and view weighting. Let $M = \sum_{s=1}^{S} m_s$ be the total feature dimension across views. Under the bounded-parameter conditions specified in Appendix B.4, the fluctuation scales typically satisfy $v_{s,N,m_s}^2 = O(m_s^{-1})$ and $b_{s,N,m_s} = O(m_s^{-1})$. With the simple sample-size weights $w_s = m_s/M$, at fixed confidence and for an effective gap bounded away from zero, the squared subspace error is $O(M^{-1})$ when no nonconstant diagonal correction is present. Nonlinear inner-product kernels may contribute an additional $O(N^{-1})$ term. More generally, let

$$\gamma_s := \lambda_{\min}(\Lambda_s) - \|R_s^{\mathrm{sig}}\|_{\mathrm{op}}, \qquad g_s := \alpha_s \gamma_s, \tag{4.5}$$

where γ_s (assumed positive) is the signal–residual separation of *view s alone* and g_s also accounts for the noise-induced signal multiplier. Writing $v_s = v_{s,N,m_s}$ for brevity, an idealized signal-adjusted inverse-variance rule sets

$$w_s^{\mathrm{IV}} \propto \frac{g_s}{v_s^2}. \tag{4.6}$$

This rule optimizes a variance-only surrogate of the error bound, ignoring the lower-order bias term. The quantities Λ_s, R_s^{sig}, and the fluctuation scales are unknown population quantities, but they can be estimated from the observed data under suitable assumptions. Thus (4.6) is an oracle principle that also motivates practical weighting estimators, as our simulations illustrate; sample-size weighting, in contrast, uses only the observable m_s. Appendix B.4 gives the surrogate optimization, the rates, and kernel-specific qualifications. Both rules favor reliable, well-separated views.

4.2 End-to-end risk bound

Assumptions. We retain the supervised triple model (3.4) and the clipped relation-wise ReLU heads from Section 3.2. To obtain a population risk bound, we impose standard conditions on the sampling mechanism and the target function. The labeled triples are i.i.d. from a distribution P whose head and tail marginals are uniform on $[N]$, and the label noise is sub-Gaussian with variance proxy σ_y^2. The true scores are bounded by B and are smooth functions of latent entity coordinates $Z^\star \in \mathbb{R}^{N\times d}$. The key condition linking the two stages is an *alignment condition*: there exists an invertible $d \times d$ matrix $A_\star$ such that

$$\sqrt{N}\, Z_{\mathrm{sig}} = Z^\star A_\star, \qquad \max\{\|A_\star\|_{\mathrm{op}}, \|A_\star^{-1}\|_{\mathrm{op}}\} \le \kappa_A,$$

where the factor $\sqrt{N}$ keeps the latent coordinates of order one. This transfer assumption requires the latent subspace relevant for KG prediction to agree with the population subspace recovered by pretraining, up to a well-conditioned change of coordinates; reducing the pretraining noise alone does not imply it. The precise statements are Assumptions 6–8 in Appendix B.2.

Risk and complexity. We measure the prediction error of a predictor $q : \mathcal{X} \to \mathbb{R}$ on a new triple by its population risk

$$\|q - f^\star\|_P^2 := \mathbb{E}_{X\sim P}\{q(X) - f^\star(X)\}^2.$$

The statistical complexity of the frozen-embedding predictor class is measured by its *pseudo-dimension*; for the relation-wise networks of Section 3.2, it is at most $p := C_{\mathrm{pd}} KDW \log W$ for an absolute constant $C_{\mathrm{pd}} \ge 1$ (Lemma 7). Let $L_\star$ be a common Lipschitz constant of the true scores as functions of the latent coordinates; it is controlled by the smoothness norm in Assumption 7. Finally, the pretraining error enters through $\mathcal{E}_{\mathrm{pt}}(w,\delta)$ defined in (4.3).

Theorem 2 (End-to-end oracle inequality for pretrained neural KG learning)**.** *Suppose Assumptions 1–8 hold,* $w \in \Delta_S$ *is deterministic,* $p = C_{\mathrm{pd}} KDW \log W \le n$*,* $W \ge W_0$*, and* $D \gtrsim \log W$*. With probability at least* $1 - 10(p/en)^p - \delta$ *over the unlabeled pretraining and labeled supervised samples,*

$$\|\widehat{f}(\widehat{Z};\cdot) - f^\star\|_P^2 \le \underbrace{C_1 \left(\frac{W}{\log W}\right)^{-\beta/d}}_{\text{neural approximation error}} + \underbrace{C_2(\sigma_y^2 + B^2)\frac{KDW\log W}{n}\log\frac{en}{KDW\log W}}_{\text{supervised estimation error}} + \underbrace{C_3(L_\star^2 + B^2)\kappa_A^2 \mathcal{E}_{\mathrm{pt}}(w,\delta)}_{\text{pretraining error}} + \underbrace{C_4\delta_{\mathrm{opt}}}_{\text{optimization error}} . \tag{4.7}$$

The constants $C_1, \dots, C_4$ *and* W_0 *depend only on the fixed quantities specified in Appendix C.2.*

Interpretation of the four errors. The *approximation term* decreases as each relation head becomes more expressive (larger W) and as the true scoring functions become smoother (larger β). Its exponent involves the latent dimension d, not the dimensions m_s of the raw side information. The *supervised estimation term* is the cost of learning K relation-specific networks from only n labeled triples; it grows with network size, depth, and label noise σ_y^2. The *pretraining term* carries the subspace recovery error from Theorem 1 into prediction risk, scaled by $L_\star^2$ (sensitivity of the true score to latent-coordinate changes), B^2 (the score range, covering poorly recovered entities), and κ_A^2 (the alignment cost). The *optimization term* vanishes when $\delta_{\mathrm{opt}} = 0$.

In particular, choosing $D \asymp \log W$ makes the supervised estimation rate of order KW/n up to logarithms, while the approximation rate is $(W/\log W)^{-\beta/d}$. Balancing the two gives $W \asymp (n/K)^{d/(\beta+d)}$ and the risk bound $(n/K)^{-\beta/(\beta+d)} + (L_\star^2 + B^2)\kappa_A^2 \mathcal{E}_{\mathrm{pt}}(w,\delta) + \delta_{\mathrm{opt}}$, up to logarithmic factors. The first term is the minimax rate for a β-smooth function of the $2d$ latent coordinates, independent of the raw dimensions m_s. This requires a small pretraining error: more labeled data cannot remove a persistent representation mismatch, such as a failure of Assumption 8.

5 Experiments

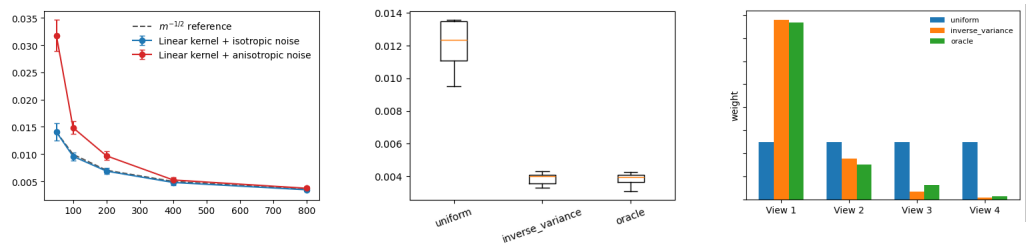
Our empirical evaluation has two parts. Section 5.1 uses controlled simulations with known ground truth to directly check the rates and qualitative predictions of Theorems 1 and 2: subspace recovery, view weighting, and the end-to-end risk decomposition. Section 5.2 presents the main real-data results on WordNet and RENKI-KG, testing whether multi-view pretraining improves link prediction under realistic textual side information and limited labeled triples.

5.1 Simulation study

We use three controlled simulations to check these predictions. Experiments 1 and 2 test the pretraining component of the theory: Experiment 1 verifies the predicted subspace recovery rate, while Experiment 2 evaluates inverse-variance view weighting under heterogeneous view quality. Experiment 3 tests the end-to-end risk decomposition by varying the labeled sample size n, the pretraining sample size m, and the number of parameters W, thereby isolating the supervised estimation, pretraining, and approximation components of the bound.

Figure 1 summarizes the pretraining behavior. In Experiment 1, the subspace error follows the predicted $m^{-1/2}$ rate; the fitted log–log slope is -0.506. In Experiment 2, the theory-guided inverse-variance weighting rule shifts mass from the noisier views to the cleaner ones, consistent with its near-oracle performance.

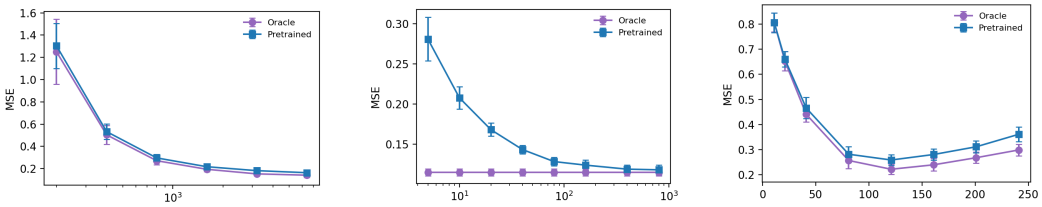
End-to-end supervised risk. Experiment 3 uses pretrained embeddings and learns bounded relation-wise ReLU networks on a nonlinear triple-score teacher with relation-specific bilinear interactions and Gaussian label noise $N(0, 0.3^2)$. We vary the labeled sample size n, pretraining sample size m, and number of parameters W; the results are shown in Figure 2. The oracle baseline uses the true latent entity coordinates from the simulator while keeping the same supervised training procedure. As the



(a) Subspace error vs. m (Exp. 1). (b) Error by weighting rule (Exp. 2). (c) View weights by rule (Exp. 2).

Figure 1: Pretraining simulations.

number of labeled triples increases, test MSE decreases consistently with the supervised-estimation term (Figure 2a). Multi-view KPCA remains at or above the oracle, with the gap representing the downstream cost of imperfect coordinate recovery; this gap shrinks as the pretraining sample grows (Figure 2b). The parameter sweep reveals a bias–variance trade-off in the relation heads: larger networks improve expressiveness but raise estimation variance when labeled data are limited (Figure 2c). All three trends match the risk decomposition in Theorem 2, where supervised sample complexity, approximation capacity, and pretraining error jointly determine the final risk.



(a) MSE vs. n. (b) MSE vs. m. (c) MSE vs. W.

Figure 2: End-to-end risk decomposition (Experiment 3). The n-, m-, and W-sweeps probe the supervised estimation, pretraining, and approximation terms of Theorem 2; the oracle uses the true latent coordinates in place of $\widehat{Z}$.

5.2 Real-data experiments

We evaluate multi-view pretraining on two KGs: WordNet, whose synset definitions provide textual side information, and RENKI-KG [Liu et al., 2024], a large-scale medical KG that links clinical concepts, diagnoses, and medications across curated ontologies and drug databases. Full protocols and additional results are given in Appendix E.

Setup and metrics. WordNet contains 116,566 entities, 33 relations, and 373,330 positive triples. RENKI-KG contains 118,554 entities of three types, 9 relations, and 698,368 positive triples. Both use random edge splits, corrupted negative triples, and relation-balanced AUROC, accuracy, and F1. Baselines are TransE, DistMult, ComplEx, and RotatE. In `PNKG`, we compare random coordinates, single-encoder representations, and multi-view KPCA, using the same relation-specific head architecture across variants. During supervised learning with text-derived coordinates, only the heads are trained, while the entity coordinates remain frozen. On WordNet, we use $d = 128$ and 50 epochs with four pretrained encoders (BERT, RoBERTa, ELECTRA, DeBERTa-v3); the view weights are selected on validation AUROC from a grid with step 0.25 on the simplex, which yields $(1 : 2 : 1 : 0)$. On RENKI-KG, we use $d = 128$, uniform weights over seven BERT-family views (BERT, RoBERTa, NER-BERT, MentalBERT, SciBERT, BioBERT, ClinicalBERT), and 100 epochs. Results are averaged over 10 seeds.

Fixed-embedding results. Figures 3 and 4 show a consistent pattern across both KGs. Frozen text-derived coordinates clearly outperform random coordinates with trainable embeddings (except DeBERTa-v3 on WordNet), which highlights the value of pretraining for knowledge graph learning. Multi-view KPCA is best overall; in AUROC/accuracy/F1 (WordNet; RENKI-KG), it improves over the best graph-only baseline by 0.1300/0.0981/0.1617 and 0.0242/0.0544/0.0702, and over the

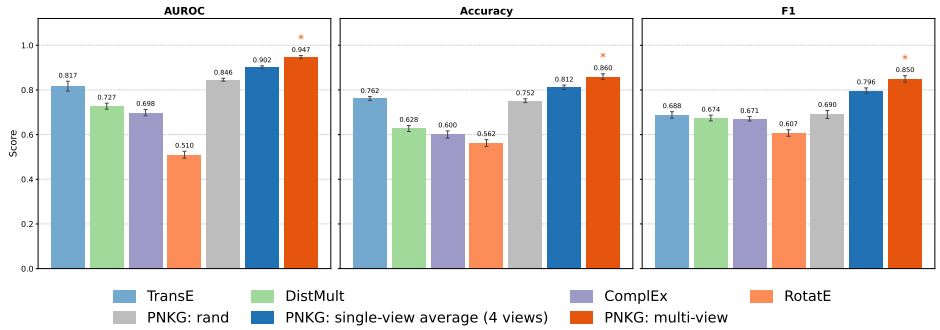


Figure 3: Random edge-split results on WordNet with relation-balanced training and evaluation; text-derived PNKG coordinates are fixed, and the random variant trains its embeddings. Results are mean $\pm$ standard deviation over 10 seeds.

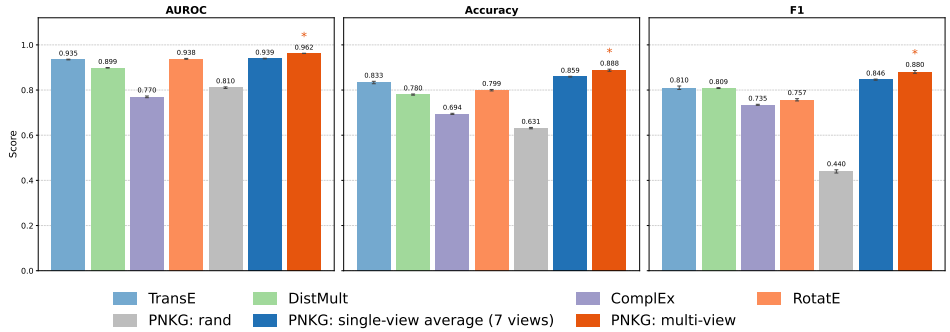


Figure 4: Random edge-split results on RENKI-KG with relation-balanced training and evaluation; text-derived PNKG coordinates are fixed, and the random variant trains its embeddings. Results are mean $\pm$ standard deviation over 10 seeds.

average single view by $0.0453/0.0481/0.0531$ and $0.0229/0.0285/0.0344$, winning on all 10 seeds ($p < 10^{-7}$). Against the strongest single view (BERT on WordNet, SciBERT on RENKI-KG), multi-view KPCA gains AUROC on every seed ($+0.0028$ and $+0.0017$, $p < 0.01$) and matches it in accuracy and F1. Thus aggregation recovers the best encoder's performance without having to identify that encoder in advance. These modest but consistent gains over single views suggest that the encoders carry complementary information.

6 Discussion

Interpreting the error decomposition. The bound (4.7) separates four sources of error: neural approximation, supervised estimation, pretraining, and optimization. The pretraining term becomes negligible once enough unlabeled data drive it below the approximation–estimation trade-off. The factor K in the supervised complexity reflects the cost of learning a separate head for each relation. The n-, m-, and W-sweeps in Experiment 3 illustrate these sources of error separately.

View weighting. The theory favors views with strong effective signal $\alpha_s\gamma_s$ and small fluctuation scale v_{s,N,m_s}. Experiment 2 shows the benefit of downweighting noisy views, and WordNet confirms it on real data: uniform weights are pulled down by the weak DeBERTa-v3 view, whereas validation-selected weights that exclude it give the best results. On RENKI-KG, where the strongest views are close in quality, uniform weights already perform best.

Limitations and future directions. Our analysis assumes frozen pretrained coordinates, an exactly shared signal subspace (A.1), and i.i.d. labeled triples. Allowing fine-tuning, approximately shared subspaces, or unseen test entities would broaden its scope. Our weighting rule depends on unknown kernel variances and effective eigengaps, motivating data-driven weighting with guarantees. Other directions include KG structure as a pretraining view and other relational tasks.

Funding Disclosure

This work received no external funding. It was conducted through academic collaboration between researchers at Northwestern University and the University of California, Riverside.

Competing Interests

The authors declare no relevant competing financial interests.

References

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

Stephen Bonner, Ian P Barrett, Cheng Ye, Rowan Swiers, Ola Engkvist, Andreas Bender, Charles Tapley Hoyt, and William L Hamilton. A review of biomedical datasets relating to drug discovery: a knowledge graph perspective. *Briefings in Bioinformatics*, 23(6):bbac404, 2022.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 2013.

Youngmin Cho and Lawrence Saul. Kernel methods for deep learning. *Advances in Neural Information Processing Systems*, 22, 2009.

Daniel Daza, Michael Cochez, and Paul Groth. Inductive entity representations from text via link prediction. In *Proceedings of The Web Conference*, pages 798–808, 2021.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1): 1–50, 2010.

Jianqing Fan, Dong Wang, Kaizheng Wang, and Ziwei Zhu. Distributed estimation of principal eigenspaces. *The Annals of Statistics*, 47(6):3009, 2019.

William J Firey. Remainder formulae in Taylor's theorem. *The American Mathematical Monthly*, 67 (9):903–905, 1960.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. In *The Twelfth International Conference on Learning Representations*, 2024.

Jeff Z. HaoChen, Colin Wei, Adrien Gaidon, and Tengyu Ma. Provable guarantees for self-supervised deep learning with spectral contrastive loss. In *Advances in Neural Information Processing Systems*, volume 34, pages 5000–5011, 2021.

Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock. Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460):1090–1098, 2002.

Xiao Huang, Jingyuan Zhang, Dingcheng Li, and Ping Li. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 105–113, 2019.

Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Neural Networks and Learning Systems*, 33(2):494–514, 2022.

Jason D. Lee, Qi Lei, Nikunj Saunshi, and Jiacheng Zhuo. Predicting what you already know helps: Provable self-supervised learning. In *Advances in Neural Information Processing Systems*, volume 34, pages 309–323, 2021.

Suqi Liu, Tianxi Cai, and Xiaoou Li. Representation-enhanced neural knowledge integration with application to large-scale medical ontology learning. *arXiv preprint arXiv:2410.07454*, 2024.

Peter W. MacDonald, Elizaveta Levina, and Ji Zhu. Latent space models for multiplex networks with shared structure. *Biometrika*, 109(3):683–706, 2022.

George A Miller. WordNet: a lexical database for English. *Communications of the ACM*, 38(11): 39–41, 1995.

Maximilian Nickel, Kevin Murphy, Volker Tresp, and Evgeniy Gabrilovich. A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33, 2016.

Ivan Nourdin, Giovanni Peccati, and Gesine Reinert. Invariance principles for homogeneous sums: universality of Gaussian Wiener chaos. *The Annals of Probability*, 38(5):1947–1985, 2010. doi: 10.1214/10-AOP531.

Nikunj Saunshi, Orestis Plevrakis, Sanjeev Arora, Mikhail Khodak, and Hrishikesh Khandeparkar. A theoretical analysis of contrastive unsupervised representation learning. In *International Conference on Machine Learning*, pages 5628–5637. PMLR, 2019.

Zhiqing Sun, Zhi-Hong Deng, Jian-Yun Nie, and Jian Tang. RotatE: Knowledge graph embedding by relational rotation in complex space. In *International Conference on Learning Representations*, 2019.

Théo Trouillon, Johannes Welbl, Sebastian Riedel, Éric Gaussier, and Guillaume Bouchard. Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080. PMLR, 2016.

Bo Wang, Tao Shen, Guodong Long, Tianyi Zhou, Ying Wang, and Yi Chang. Structure-augmented text representation learning for efficient knowledge graph completion. In *Proceedings of The Web Conference*, pages 1737–1748, 2021a.

Liang Wang, Wei Zhao, Zhuoyu Wei, and Jingming Liu. SimKGC: Simple contrastive knowledge graph completion with pre-trained language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4281–4294, 2022.

Xiaozhi Wang, Tianyu Gao, Zhaocheng Zhu, Zhengyan Zhang, Zhiyuan Liu, Juanzi Li, and Jian Tang. KEPLER: A unified model for knowledge embedding and pre-trained language representation. *Transactions of the Association for Computational Linguistics*, 9:176–194, 2021b.

Andrew Gordon Wilson, Zhiting Hu, Ruslan Salakhutdinov, and Eric P Xing. Deep kernel learning. In *Artificial Intelligence and Statistics*, pages 370–378. PMLR, 2016.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. Representation learning of knowledge graphs with entity descriptions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30, 2016.

Yuling Yan, Yuxin Chen, and Jianqing Fan. Inference for heteroskedastic PCA with missing data. *The Annals of Statistics*, 52(2):729–756, 2024.

Bishan Yang, Wen-tau Yih, Xiaodong He, Jianfeng Gao, and Li Deng. Embedding entities and relations for learning and inference in knowledge bases. In *International Conference on Learning Representations*, 2015.

Yang Yang, Zhilei Wu, Yuexiang Yang, Shuangshuang Lian, Fengjie Guo, and Zhiwei Wang. A survey of information extraction based on deep learning. *Applied Sciences*, 12(19):9691, 2022.

Liang Yao, Chengsheng Mao, and Yuan Luo. KG-BERT: BERT for knowledge graph completion. *arXiv preprint arXiv:1909.03193*, 2019.

Dmitry Yarotsky. Error bounds for approximations with deep ReLU networks. *Neural Networks*, 94: 103–114, 2017.

Yi Yu, Tengyao Wang, and Richard J Samworth. A useful variant of the Davis–Kahan theorem for statisticians. *Biometrika*, 102(2):315–323, 2015.

Fuzheng Zhang, Nicholas Jing Yuan, Defu Lian, Xing Xie, and Wei-Ying Ma. Collaborative knowledge base embedding for recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 353–362, 2016.

A Additional Framework Details

A.1 A linear kernel example

We first specialize (3.1) to the simplest, linear isotropic setting. Assume that the noise scale is constant within each view, $\sigma_{s,j} = \sigma_s$, and that

$$\varepsilon_{s,j} \sim N(0, I_{m_s}), \qquad j = 1, \dots, N.$$

Then

$$E_s = V_s + \frac{\sigma_s}{\sqrt{m_s}} \xi_s,$$

where the jth rows of V_s and ξ_s are $v_{s,j}^\top$ and $\varepsilon_{s,j}^\top$, respectively. The corresponding linear entity-kernel matrix is

$$K_s^{\mathrm{lin}} = \frac{1}{N} E_s E_s^\top \in \mathbb{R}^{N \times N}.$$

With multiple views, we aggregate the entity-kernel matrices as

$$K^{\mathrm{lin}}(w) = \sum_{s=1}^{S} w_s K_s^{\mathrm{lin}}, \qquad w \in \Delta_S.$$

The leading eigenvectors of $K^{\mathrm{lin}}(w)$ then provide low-dimensional coordinates for the common entity set. This is the special case of the general construction with $k_s(u, u') = u^\top u'$ and $\phi_s(u) = u$, for which $K_s^{\mathrm{lin}} = K_s^{\mathrm{obs}}$ and $K^{\mathrm{lin}}(w) = K^{\mathrm{obs}}(w)$.

A.2 Weighted direct-sum representation and encoder kernels

We now show that the weighted multi-view matrix $K^{\mathrm{obs}}(w)$ in (3.3) is itself the Gram matrix of a single kernel. Write $\mathbf{u}_j := (u_{1,j}, \dots, u_{S,j})$ for the collection of side-information observations of entity j, and define the weighted direct-sum feature map

$$\Phi_w(\mathbf{u}_j) := \left(\sqrt{w_1}\phi_1(u_{1,j}), \dots, \sqrt{w_S}\phi_S(u_{S,j})\right) \in \bigoplus_{s=1}^{S} \mathcal{H}_s.$$

Here $\bigoplus_{s=1}^{S} \mathcal{H}_s$ is the Hilbert direct sum of the view-specific feature spaces. Since S is finite, it consists of the tuples $(h_1, \dots, h_S)$ with $h_s \in \mathcal{H}_s$, equipped with the inner product

$$\langle (h_1, \dots, h_S), (h_1', \dots, h_S') \rangle_{\oplus_s \mathcal{H}_s} := \sum_{s=1}^{S} \langle h_s, h_s' \rangle_{\mathcal{H}_s}.$$

The weighted views therefore induce the aggregated kernel

$$k_w(\mathbf{u}, \mathbf{u}') := \langle \Phi_w(\mathbf{u}), \Phi_w(\mathbf{u}') \rangle_{\oplus_s \mathcal{H}_s} = \sum_{s=1}^{S} w_s k_s(u_s, u_s')$$

for $\mathbf{u} = (u_1, \dots, u_S)$ and $\mathbf{u}' = (u_1', \dots, u_S')$, which is positive semidefinite because it is defined by an inner product (equivalently, because $w_s \geq 0$ and each k_s is positive semidefinite). Moreover, $K^{\mathrm{obs}}(w)$ is precisely the normalized Gram matrix of k_w on the entity set:

$$K^{\mathrm{obs}}(w)(j, j') = \frac{1}{N} k_w(\mathbf{u}_j, \mathbf{u}_{j'}) = \frac{1}{N} \langle \Phi_w(\mathbf{u}_j), \Phi_w(\mathbf{u}_{j'}) \rangle_{\oplus_s \mathcal{H}_s} = \sum_{s=1}^{S} w_s K_s^{\mathrm{obs}}(j, j').$$

Remark 1 (Pretrained encoder-induced kernels). The kernel formulation naturally accommodates frozen pretrained encoders. In the language-model application, $u_{s,j} \in \mathbb{R}^{m_s}$ is a vectorized representation of the textual description of entity j, and the frozen encoder of view s supplies the feature map $\phi_s : \mathbb{R}^{m_s} \to \mathbb{R}^{m_s'}$, so that $\mathcal{H}_s = \mathbb{R}^{m_s'}$. The encoder thus induces the positive-semidefinite kernel $k_s(u, u') = \langle \phi_s(u), \phi_s(u') \rangle_{\mathbb{R}^{m_s'}}$, and the dot product between language-model embeddings is exactly the Hilbert-space inner product realizing this kernel. Because the procedure operates only on the $N \times N$ kernel matrices, different encoders may have different feature dimensions m_s'. The framework applies to arbitrary encoder-induced positive-semidefinite kernels, whereas our theory derives explicit rates only for the linear, Gaussian distance, and nonlinear inner-product kernel regimes of Appendix B.3.

A.3 Shared signal structure and embedding dimension

Define the pure-signal kernel matrix of view s by

$$K_s^{\mathrm{sig}}(j,j') := \frac{1}{N} k_s(v_{s,j}, v_{s,j'}), \qquad j,j' \in [N],$$

where $v_{s,j}$ is the latent signal vector in the information-plus-noise model (3.1). The central structural condition behind the PNKG framework is that the pure-signal kernels share a common rank-d entity eigenspace, while allowing view-specific variation outside that subspace:

$$K_s^{\mathrm{sig}} = Z_{\mathrm{sig}} \Lambda_s Z_{\mathrm{sig}}^\top + R_s^{\mathrm{sig}}, \qquad Z_{\mathrm{sig}}^\top Z_{\mathrm{sig}} = I_d, \qquad \Lambda_s \in \mathbb{R}^{d\times d},\ \Lambda_s \succeq 0. \tag{A.1}$$

Here $Z_{\mathrm{sig}} \in \mathbb{R}^{N\times d}$ spans the shared population entity eigenspace, Λ_s captures the strength of view s along this subspace, and R_s^{sig} is a symmetric view-specific residual matrix supported on the orthogonal complement of the column space of Z_{sig}:

$$(R_s^{\mathrm{sig}})^\top = R_s^{\mathrm{sig}}, \qquad R_s^{\mathrm{sig}} Z_{\mathrm{sig}} = 0.$$

This is Assumption 3. For a weight vector $w \in \Delta_S$, the weighted population signal kernel matrix is then

$$K^{\mathrm{sig}}(w) := \sum_{s=1}^S w_s \alpha_s K_s^{\mathrm{sig}} = Z_{\mathrm{sig}} \left(\sum_{s=1}^S w_s \alpha_s \Lambda_s \right) Z_{\mathrm{sig}}^\top + \sum_{s=1}^S w_s \alpha_s R_s^{\mathrm{sig}}, \tag{A.2}$$

where $\alpha_s > 0$ is the signal multiplier of Appendix B.3 ($\alpha_s = 1$ for linear and inner-product kernels). The views need not be equally informative; they need only agree on the leading entity-level signal eigenspace. The observed kernels K_s^{obs}, computed from the noisy vectors $u_{s,j}$, are information-plus-noise perturbations of the rescaled signal kernels $\alpha_s K_s^{\mathrm{sig}}$, where the multiplier α_s and the bias and fluctuation scales are given in Appendix B.3. Since $\alpha_s > 0$, this rescaling does not change the shared eigenspace Z_{sig}.

Remark 2 (Choice of d). The embedding dimension d is the number of leading eigenvectors retained by Algorithm 1. The theory takes d to be the rank of the shared signal eigenspace in (A.1). The pretrained coordinates are identifiable only up to an orthogonal change of basis, since any such rotation spans the same eigenspace; what matters is the subspace spanned by the selected eigenvectors, which is why Theorem 1 is stated in terms of principal angles. In practice, d is treated as a tuning parameter: choosing it too small may discard signal directions, whereas choosing it too large adds residual directions and unnecessary model complexity.

B Assumptions, Kernel Regimes, and View Weighting

B.1 Pretraining assumptions

Assumption 1 (Entity-level information-plus-noise model)**.** For each view $s \in [S]$, the observed entity vectors satisfy

$$u_{s,j} = v_{s,j} + \sigma_{s,j} \frac{\varepsilon_{s,j}}{\sqrt{m_s}}, \qquad j = 1, \dots, N,$$

where $v_{s,j} \in \mathbb{R}^{m_s}$ is the latent signal vector of entity j, $\varepsilon_{s,j} \in \mathbb{R}^{m_s}$ is high-dimensional noise, and $\sigma_{s,j}$ is a view-specific noise scale. The normalization $m_s^{-1/2}$ keeps the noise energy bounded as the feature dimension grows.

We next require the noise to satisfy conditions matched to the kernel class of each view.

Assumption 2 (Admissible noise regimes)**.** For each view $s \in [S]$, the noise vectors $\{\varepsilon_{s,j}\}_{j=1}^N$ are independent of the signal vectors $\{v_{s,j}\}_{j=1}^N$, and one of the following two regimes holds.

1. **Inner-product kernel regime.** The noise vectors $\varepsilon_{s,j}$ are i.i.d. across j, with $\mathbb{E}[\varepsilon_{s,j}] = 0$, and satisfy the concentration inequality

$$\mathbb{P}\{|F(\varepsilon_{s,j}) - \mathbb{E}F(\varepsilon_{s,j})| > r\} \le C_s \exp(-c_s r^{q_s}), \qquad r > 0, \tag{B.1}$$

for every 1-Lipschitz function $F : \mathbb{R}^{m_s} \to \mathbb{R}$, where $C_s, c_s, q_s > 0$ do not depend on N or m_s. Let

$$\nu_s := \frac{\mathbb{E}\|\varepsilon_{s,j}\|_2^2}{m_s},$$

and assume $\nu_s = O(1)$. The noise scales are uniformly bounded, $|\sigma_{s,j}| \le \sigma_{s,\infty}$, and the signal inner products are controlled: there is a deterministic sequence $B_{s,N}$ such that

$$\max_{j,j' \in [N]} |v_{s,j}^\top v_{s,j'}| \le B_{s,N}.$$

2. **Gaussian distance-kernel regime.** The noise is Gaussian with a homogeneous scale within the view:

$$\varepsilon_{s,j} \overset{\text{iid}}{\sim} N(0, \Psi_{s,m_s}), \qquad \sigma_{s,j} = \sigma_s,$$

where $\Psi_{s,m_s} \succeq 0$ does not depend on j and is normalized so that $\operatorname{tr}(\Psi_{s,m_s})/m_s = 1$. The signal energy is controlled: there are a deterministic center $a_s \in \mathbb{R}^{m_s}$ and a deterministic sequence $B_{s,N}$ such that

$$\frac{1}{N}\sum_{j=1}^{N} \|v_{s,j} - a_s\|_2^{2q} \le B_{s,N}^q, \qquad q = 1, 2, 3.$$

Assumption 3 (Shared effective signal eigenspace)**.** The pure-signal kernel matrices K_s^{sig} of Appendix A.3 share a common rank-d entity eigenspace: there exist $Z_{\text{sig}} \in \mathbb{R}^{N \times d}$ with $Z_{\text{sig}}^\top Z_{\text{sig}} = I_d$, positive-semidefinite matrices $\Lambda_s \in \mathbb{R}^{d \times d}$, and symmetric residual matrices $R_s^{\text{sig}} \in \mathbb{R}^{N \times N}$ such that

$$K_s^{\text{sig}} = Z_{\text{sig}} \Lambda_s Z_{\text{sig}}^\top + R_s^{\text{sig}}, \qquad R_s^{\text{sig}} Z_{\text{sig}} = 0, \qquad s \in [S].$$

Assumption 4 (Effective signal separation)**.** Every view has a positive separation margin:

$$\gamma_s := \lambda_{\min}(\Lambda_s) - \|R_s^{\text{sig}}\|_{\text{op}} > 0, \qquad s \in [S].$$

Since $\lambda_{\min}$ is superadditive and $\|\cdot\|_{\text{op}}$ is subadditive, the weighted gap (4.1) satisfies

$$\gamma(w) \ge \sum_{s=1}^{S} w_s \alpha_s \gamma_s > 0 \qquad \text{for every } w \in \Delta_S. \tag{B.2}$$

Assumption 5 (Sample independence)**.** The signal vectors $\{v_{s,j}\}$ are deterministic (equivalently, all pretraining statements are conditional on them), and the noise arrays $\{\varepsilon_{1,j}\}_{j=1}^N, \ldots, \{\varepsilon_{S,j}\}_{j=1}^N$ are independent across views, so that the entity matrices $E_1, \ldots, E_S$ used for pretraining are independent. Whenever a supervised sample $\mathcal{D}_{\text{sup}} = \{(X_i, Y_i)\}_{i=1}^n$ is also considered, it is independent of the pretraining matrices.

Together, Assumptions 1–5 ensure that the weighted population kernel $\sum_{s=1}^S w_s \alpha_s K_s^{\text{sig}}$ has a well-separated rank-d eigenspace, and that each observed kernel K_s^{obs} is a controlled information-plus-noise perturbation of its population counterpart $\alpha_s K_s^{\text{sig}}$, for any mixture of linear, Gaussian distance, and nonlinear inner-product kernels across views. Theorem 1 turns this structure into a nonasymptotic Davis–Kahan-type bound on the principal-angle distance between the aggregated empirical eigenspace $\widehat{Z}$ and the signal eigenspace Z_{sig}. The bound is expressed through view-specific bias scales b_{s,N,m_s}, diagonal-correction scales d_{s,N,m_s}, and fluctuation scales v_{s,N,m_s}, which we calibrate separately for each kernel regime in Appendix B.3. Although we use the theorem to control the pretraining error of the PNKG pipeline, its statement makes no reference to the downstream supervised task. It is a general finite-sample subspace-recovery guarantee for weighted multi-view (equivalently, multi-kernel) PCA under heterogeneous information-plus-noise models, and may be of independent interest beyond knowledge graphs.

B.2 Supervised-stage assumptions

Theorem 2 adds three conditions to the pretraining assumptions above: a sampling condition on the labeled triples, a smooth latent-score model for the KG scoring function, and an alignment condition that links the pretrained eigenspace to the latent coordinates of that model. These conditions control, respectively, supervised generalization, neural approximation, and coordinate transfer together with identifiability. Throughout, σ_y^2 denotes the sub-Gaussian variance proxy of the *label noise*.

Assumption 6 (Supervised sampling)**.** Conditionally on $X_1, \dots, X_n$, the noises ε_i in (3.4) are independent and sub-Gaussian with variance proxy σ_y^2. The target is bounded: $\sup_{x \in \mathcal{X}} |f^\star(x)| \le B$. For out-of-sample bounds, $X_1, \dots, X_n$ are i.i.d. from a distribution P on $\mathcal{X}$ that is uniform on entities: if $X = (h, r, t) \sim P$, then h and t are each uniformly distributed on $[N]$.

Assumption 7 (Smooth latent KG model)**.** There exist latent entity coordinates $Z^\star = (z_1^\star, \dots, z_N^\star)^\top \in \mathbb{R}^{N \times d}$, a radius $R_\star > 1$ with $\max_{j \in [N]} \|z_j^\star\|_2 \le R_\star - 1$, an integer smoothness index $\beta \ge 1$, and bounded relation-specific functions $g_r^\star : \mathcal{D}_\star \to [-B, B]$, where $\mathcal{D}_\star = [-R_\star, R_\star]^d \times [-R_\star, R_\star]^d$, such that

$$f^\star(h, r, t) = g_r^\star(z_h^\star, z_t^\star), \qquad \|g_r^\star\|_{W^{\beta,\infty}(\mathcal{D}_\star)} \le S_\beta. \tag{B.3}$$

Here $W^{\beta,\infty}(\mathcal{D}_\star)$ denotes the Sobolev space of functions on $\mathcal{D}_\star$ whose weak derivatives up to order β are essentially bounded, with norm $\|g\|_{W^{\beta,\infty}(\mathcal{D}_\star)} := \max_{|\alpha| \le \beta} \|D^\alpha g\|_{L^\infty(\mathcal{D}_\star)}$, and S_β is a common bound on the norms.

The smoothness index β governs the ReLU approximation rate. The unit margin between the latent coordinates and the boundary of $\mathcal{D}_\star$ lets the proof evaluate $g_r^\star$ at slightly perturbed coordinates; see Appendix C.2.

Assumption 8 (KPCA alignment with the true latent span)**.** There exists an invertible $d \times d$ matrix $A_\star$ such that

$$\sqrt{N}\, Z_{\mathrm{sig}} = Z^\star A_\star, \qquad \|A_\star\|_{\mathrm{op}} \le \kappa_A, \qquad \|A_\star^{-1}\|_{\mathrm{op}} \le \kappa_A. \tag{B.4}$$

The factor $\sqrt{N}$ fixes the scale. Because $Z_{\mathrm{sig}}^\top Z_{\mathrm{sig}} = I_d$, the rows of Z_{sig} have average squared norm d/N, whereas the rows of $\sqrt{N} Z_{\mathrm{sig}}$ have average squared norm d. Hence the latent coordinates $z_j^\star$ remain of order one as N grows, and the scores $g_r^\star(z_h^\star, z_t^\star)$ do not degenerate to a relation-specific constant. This *transfer assumption* requires the latent subspace relevant for KG prediction to agree with the population subspace recovered by pretraining, up to a well-conditioned change of coordinates; reducing the pretraining noise alone does not imply it. The condition is nevertheless broad: it allows different feature mappings across datasets, and it requires only subspace-level compatibility between the population pretraining geometry and the downstream KG geometry, not an exact semantic match between views and relations or a correctly specified parametric KG scoring function. The matrix $A_\star$ accounts for the usual change-of-basis non-identifiability of spectral representations; each relation-wise ReLU head can absorb it, together with the factor $\sqrt{N}$, into its first affine layer, as shown in Appendix C.2. In particular, Algorithm 1 needs no rescaling of $\widehat{Z}$.

B.3 Kernel-specific scales for Theorem 1

The sets $\mathcal{S}_{\mathrm{lin}}$, $\mathcal{S}_{\mathrm{gauss}}$, and $\mathcal{S}_{\mathrm{ip}}$ of linear, Gaussian distance-kernel, and nonlinear inner-product views partition $[S]$. For each view $s \in [S]$, we define the signal multiplier α_s, the deterministic Taylor-bias scale b_{s,N,m_s}, the nonconstant diagonal-correction scale d_{s,N,m_s}, and the centered fluctuation scale v_{s,N,m_s} according to the regime of the view. Throughout, $C_{s,b}, C_{s,v} > 0$ denote constants that depend only on the view-level noise constants in Assumption 2, and not on N or m_s.

1. **Linear kernel regime.** Let $s \in \mathcal{S}_{\mathrm{lin}}$, with

   $$k_s(u, u') = u^\top u',$$

   and suppose that the noise satisfies Assumption 2(1) with a homogeneous noise scale, $\sigma_{s,j} \equiv \sigma_s$ for $j \in [N]$. Set

   $$\alpha_s := 1, \qquad b_{s,N,m_s}^{\mathrm{lin}} := 0, \qquad d_{s,N,m_s}^{\mathrm{lin}} := 0,$$

   $$\left(v_{s,N,m_s}^{\mathrm{lin}}\right)^2 := C_{s,v} \left[\frac{\sigma_s^2 B_{s,N}}{m_s} + \frac{\sigma_s^4 (1 + \nu_s)}{m_s} \right].$$

2. **Gaussian distance-kernel regime.** Let $s \in \mathcal{S}_{\mathrm{gauss}}$, with

   $$k_s(u, u') = \exp\{-\tau_s \|u - u'\|_2^2\}, \qquad \tau_s > 0,$$

and suppose that the noise satisfies Assumption 2(2). Define

$$\alpha_s := \exp(-2\tau_s\sigma_s^2), \qquad \eta_s^2 := \frac{\sigma_s^4 \operatorname{tr}(\Psi_{s,m_s}^2)}{m_s^2} + \frac{\sigma_s^2 \|\Psi_{s,m_s}\|_{\mathrm{op}} B_{s,N}}{m_s},$$

and set

$$b_{s,N,m_s}^{\mathrm{gauss}} := C_{s,b}\left(\tau_s^2\eta_s^2 + \tau_s^3\eta_s^3\right), \qquad d_{s,N,m_s}^{\mathrm{gauss}} := 0,$$
$$\left(v_{s,N,m_s}^{\mathrm{gauss}}\right)^2 := C_{s,v}\left(\tau_s^2\eta_s^2 + \tau_s^4\eta_s^4 + \tau_s^6\eta_s^6\right).$$

3. **Nonlinear inner-product kernel regime.** Let $s \in \mathcal{S}_{\mathrm{ip}}$, with

$$k_s(u,u') = f_s(u^\top u'),$$

where $f_s \in C^3(\mathbb{R})$ has bounded derivatives, $L_{s,r} := \sup_{t\in\mathbb{R}} |f_s^{(r)}(t)| < \infty$ for $r = 1,2,3$, and suppose that the noise satisfies Assumption 2(1). Set

$$\alpha_s := 1, \qquad \eta_s := \sigma_{s,\infty}\left(\frac{B_{s,N}}{m_s}\right)^{1/2} + \sigma_{s,\infty}^2\left(\frac{1+\nu_s}{m_s}\right)^{1/2}.$$

The Taylor-bias, diagonal-correction, and fluctuation scales are

$$b_{s,N,m_s}^{\mathrm{ip}} := C_{s,b}\left(L_{s,2}\eta_s^2 + L_{s,3}\eta_s^3\right), \qquad d_{s,N,m_s}^{\mathrm{ip}} := \frac{C_{s,b}L_{s,1}\nu_s\sigma_{s,\infty}^2}{\sqrt{N}},$$
$$\left(v_{s,N,m_s}^{\mathrm{ip}}\right)^2 := C_{s,v}\left(L_{s,1}^2\eta_s^2 + L_{s,2}^2\eta_s^4 + L_{s,3}^2\eta_s^6\right).$$

For each $s \in [S]$, the scales entering Theorem 1 are those of the regime of view s:

$$(b_{s,N,m_s}, d_{s,N,m_s}, v_{s,N,m_s}) := \begin{cases} \left(b_{s,N,m_s}^{\mathrm{lin}}, d_{s,N,m_s}^{\mathrm{lin}}, v_{s,N,m_s}^{\mathrm{lin}}\right), & s \in \mathcal{S}_{\mathrm{lin}}, \\ \left(b_{s,N,m_s}^{\mathrm{gauss}}, d_{s,N,m_s}^{\mathrm{gauss}}, v_{s,N,m_s}^{\mathrm{gauss}}\right), & s \in \mathcal{S}_{\mathrm{gauss}}, \\ \left(b_{s,N,m_s}^{\mathrm{ip}}, d_{s,N,m_s}^{\mathrm{ip}}, v_{s,N,m_s}^{\mathrm{ip}}\right), & s \in \mathcal{S}_{\mathrm{ip}}. \end{cases}$$

B.4 Rates and view weighting

Let $M := \sum_{s=1}^S m_s$, and abbreviate $b_s := b_{s,N,m_s}$ and $v_s := v_{s,N,m_s}$. Throughout this subsection we treat S and the view-level parameters as fixed, assume $B_{s,N} = O(1)$ and, for Gaussian distance-kernel views, $\|\Psi_{s,m_s}\|_{\mathrm{op}} = O(1)$, and assume that the effective gap $\gamma(w)$ is bounded away from zero.

Sample-size weighting. We first set aside the diagonal correction d_{s,N,m_s}, which arises only for nonlinear inner-product kernels. Under the above conditions, the scales of Appendix B.3 satisfy

$$v_s^2 \le \frac{c_s}{m_s}, \qquad b_s = O(m_s^{-1}),$$

for view-dependent constants $c_s > 0$, so the squared bias is of lower order than the variance. A natural rule weights each view in proportion to its sample size, $w_s^{\mathrm{size}} := m_s/M$. Under this choice,

$$\sum_{s=1}^S (w_s^{\mathrm{size}})^2 v_s^2 = O(M^{-1}), \qquad \left(\sum_{s=1}^S w_s^{\mathrm{size}} b_s\right)^2 = O(M^{-2}).$$

The variance term therefore dominates, and at a fixed confidence level Theorem 1 gives $\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2 = O(M^{-1})$.

Signal-adjusted inverse-variance weighting. A more refined rule accounts for both the noise level and the effective signal strength of each view through $g_s = \alpha_s\gamma_s$ in (4.5). For a Gaussian distance-kernel view, $g_s = e^{-2\tau_s\sigma_s^2}\gamma_s$, while v_s^2 contains terms proportional to τ_s^2, τ_s^4, and τ_s^6. An overly sharp Gaussian kernel thus both attenuates the effective signal and inflates the stochastic error.

Ignoring the lower-order bias term and replacing $\gamma(w)$ by its lower bound $\sum_s w_s g_s$ from (B.2) leads to the variance-only surrogate

$$\frac{\sum_{s=1}^S w_s^2 v_s^2}{\left(\sum_{s=1}^S w_s g_s\right)^2} \geq \left(\sum_{s=1}^S \frac{g_s^2}{v_s^2}\right)^{-1},$$

where the inequality is Cauchy–Schwarz. Equality holds for the *signal-adjusted inverse-variance weights*

$$w_s^{\mathrm{IV}} := \frac{g_s/v_s^2}{\sum_{t=1}^S g_t/v_t^2},$$

which therefore minimize the surrogate over Δ_S and give

$$\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2 \lesssim \frac{1}{\delta}\left(\sum_{s=1}^S \frac{g_s^2}{v_s^2}\right)^{-1}$$

up to the lower-order bias term. The quantity $\sum_{s=1}^S g_s^2/v_s^2$ thus acts as the aggregate effective information across views. If $v_s^2 \asymp c_s/m_s$ and the ratios g_s/c_s are comparable across views, then $w_s^{\mathrm{IV}} \asymp m_s/M$. Both rules therefore achieve the $O(M^{-1})$ rate, and the inverse-variance rule has a leading constant that is no larger, since it minimizes the surrogate.

Diagonal correction. For nonlinear inner-product kernels, the nonconstant diagonal correction contributes $O(N^{-1})$ to the squared subspace error, so the overall rate becomes $O(M^{-1} + N^{-1})$. This term is absent in the linear regime with a homogeneous noise scale, because the diagonal correction is then a multiple of the identity, which can be absorbed into the target without changing its eigenspace; it is likewise absent for Gaussian distance kernels, whose diagonal is constant. Diagonal-imputation methods for heteroskedastic PCA [Yan et al., 2024] suggest a possible route for estimating and removing the nonconstant correction.

Why the linear regime matters. We highlight the linear regime because it is a natural leading-order approximation to high-dimensional random kernel matrices: El Karoui [2010] shows that the leading spectral behavior of many commonly used kernel matrices in high dimensions is governed by covariance-type linear matrices.

C Proofs

C.1 Proof of Theorem 1 (subspace recovery)

We first prove view-level kernel approximation bounds for nonlinear inner-product kernels, then for linear kernels, and finally for Gaussian distance kernels. The linear kernel corresponds to $f_s(t) = t$, but we treat it separately because, under a homogeneous noise scale, its diagonal correction is a multiple of I_N and can be absorbed into the target without changing its eigenspace. Throughout, kernel matrices are normalized by $1/N$, as in the main text. For each view s, recall that

$$\alpha_s := \begin{cases} 1, & s \in \mathcal{S}_{\mathrm{ip}} \cup \mathcal{S}_{\mathrm{lin}}, \\ \exp(-2\tau_s\sigma_s^2), & s \in \mathcal{S}_{\mathrm{gauss}}, \end{cases}$$

where $\mathcal{S}_{\mathrm{ip}}$, $\mathcal{S}_{\mathrm{lin}}$, and $\mathcal{S}_{\mathrm{gauss}}$ denote the sets of nonlinear inner-product, linear, and Gaussian distance-kernel views, respectively; these three sets partition $[S]$. We also define the "harmless identity-shift" coefficient

$$\chi_s := \begin{cases} \nu_s\sigma_s^2/N, & s \in \mathcal{S}_{\mathrm{lin}}, \\ 0, & s \in \mathcal{S}_{\mathrm{ip}}, \\ (1-\alpha_s)/N, & s \in \mathcal{S}_{\mathrm{gauss}}. \end{cases}$$

Lemma 3 (Inner-product kernel decomposition)**.** *Let $s \in \mathcal{S}_{\mathrm{ip}}$, with $k_s(u,u') = f_s(u^\top u')$, and suppose that f_s satisfies the derivative bounds of the nonlinear inner-product regime in Appendix B.3 and that the noise satisfies Assumption 2(1). Let*

$$\eta_s := \sigma_{s,\infty}\left(\frac{B_{s,N}}{m_s}\right)^{1/2} + \sigma_{s,\infty}^2\sqrt{\frac{1+\nu_s}{m_s}}.$$

Then

$$K_s^{\mathrm{obs}} - K_s^{\mathrm{sig}} = \Upsilon_s + \Xi_s, \qquad \mathbb{E}\Xi_s = 0,$$

where

$$\|\Upsilon_s\|_F \le b_s + d_s, \qquad \mathbb{E}\|\Xi_s\|_F^2 \le v_s^2,$$

with

$$b_s := C_{s,b}\left(L_{s,2}\eta_s^2 + L_{s,3}\eta_s^3\right), \qquad d_s := C_{s,b}\frac{L_{s,1}\nu_s\sigma_{s,\infty}^2}{\sqrt{N}},$$

and

$$v_s^2 := C_{s,v}\left(L_{s,1}^2\eta_s^2 + L_{s,2}^2\eta_s^4 + L_{s,3}^2\eta_s^6\right).$$

The constants $C_{s,b}$ and $C_{s,v}$ depend only on (C_s, c_s, q_s) and a uniform upper bound on ν_s, and not on N or m_s.

Proof. We suppress the view index s and write

$$m = m_s, \qquad \nu = \nu_s, \qquad \sigma_j = \sigma_{s,j}, \qquad \sigma_\infty = \sigma_{s,\infty},$$
$$B_N = B_{s,N}, \qquad f = f_s, \qquad L_r = L_{s,r},$$
$$u_j = u_{s,j}, \qquad v_j = v_{s,j}, \qquad \varepsilon_j = \varepsilon_{s,j}.$$

The signal vectors are fixed throughout, and all expectations and probabilities are taken over the noise only. In this proof, C denotes a constant that depends only on (C_s, c_s, q_s) and an upper bound on ν, and whose value may change from line to line; $C_{s,p}$ denotes such a constant that may also depend on p.

For $j, k \in [N]$, define

$$x_{jk} := v_j^\top v_k + \mathbf{1}_{\{j=k\}}\nu\sigma_j^2, \qquad \Delta_{jk} := u_j^\top u_k - x_{jk}.$$

Since $u_j = v_j + \sigma_j\varepsilon_j/\sqrt{m}$, expanding the inner product gives

$$\Delta_{jk} = \frac{\sigma_j}{\sqrt{m}}\varepsilon_j^\top v_k + \frac{\sigma_k}{\sqrt{m}}\varepsilon_k^\top v_j + \frac{\sigma_j\sigma_k}{m}\varepsilon_j^\top\varepsilon_k - \mathbf{1}_{\{j=k\}}\nu\sigma_j^2.$$

For $j \neq k$, every term is centered because the noise vectors are centered and independent across entities; for $j = k$, $\mathbb{E}\|\varepsilon_j\|_2^2 = m\nu$ by the definition of ν, so the last two terms cancel in expectation. Hence $\mathbb{E}\Delta_{jk} = 0$.

Moment bounds for Δ_{jk}. The noise assumption implies the following standard fact: if $G : \mathbb{R}^m \to \mathbb{R}$ is L-Lipschitz, then for every fixed $p \ge 1$,

$$\|G(\varepsilon_j) - \mathbb{E}G(\varepsilon_j)\|_{L^p} \le C_{s,p}L.$$

Indeed, let $T := G(\varepsilon_j) - \mathbb{E}G(\varepsilon_j)$. Applying (B.1) to the 1-Lipschitz function G/L gives

$$\mathbb{P}(|T| > t) \le C_s \exp\left\{-c_s\left(\frac{t}{L}\right)^{q_s}\right\},$$

and the tail-integral identity yields

$$\begin{aligned}\mathbb{E}|T|^p &= p\int_0^\infty t^{p-1}\mathbb{P}(|T| > t)\,dt \\ &\le pC_s\int_0^\infty t^{p-1}\exp\left\{-c_s\left(\frac{t}{L}\right)^{q_s}\right\}dt \\ &= \frac{pC_s}{q_s}c_s^{-p/q_s}\Gamma\left(\frac{p}{q_s}\right)L^p.\end{aligned}$$

Hence $\|T\|_{L^p} \le C_{s,p}L$, where $C_{s,p}$ depends only on (p, C_s, c_s, q_s) and not on N or m.

The same concentration property tensorizes, up to constants, to an independent pair $(\varepsilon_j, \varepsilon_k)$. To see this, let $G : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ be 1-Lipschitz under the Euclidean product metric, and define $g(x) := \mathbb{E}_{\varepsilon_k}G(x, \varepsilon_k)$. Then g is 1-Lipschitz, and conditioning on ε_j gives

$$\begin{aligned}\mathbb{P}\{|G - \mathbb{E}G| > r\} &\le \mathbb{P}\{|G - g(\varepsilon_j)| > r/2\} + \mathbb{P}\{|g(\varepsilon_j) - \mathbb{E}g(\varepsilon_j)| > r/2\} \\ &\le 2C_s\exp(-c_s2^{-q_s}r^{q_s}).\end{aligned}$$

Consider first the signal–noise term $\sigma_j \varepsilon_j^\top v_k/\sqrt{m}$. As a function of ε_j, it is centered and Lipschitz with constant

$$\frac{|\sigma_j|\|v_k\|_2}{\sqrt{m}} \le \sigma_\infty \left(\frac{B_N}{m}\right)^{1/2},$$

so for every fixed $p \ge 1$,

$$\left\| \frac{\sigma_j}{\sqrt{m}} \varepsilon_j^\top v_k \right\|_{L^p} \le C_{s,p} \sigma_\infty \left(\frac{B_N}{m}\right)^{1/2}.$$

The same bound holds for the other signal–noise term.

We next control the noise–noise term. For $j = k$, let $R_j := |\sigma_j|\|\varepsilon_j\|_2$. Since R_j is $|\sigma_j|$-Lipschitz in ε_j,

$$\|R_j - \mathbb{E}R_j\|_{L^p} \le C_{s,p}|\sigma_j|.$$

Moreover,

$$\mathbb{E}R_j^2 = m\nu\sigma_j^2, \qquad \mathbb{E}R_j \le |\sigma_j|\sqrt{m\nu}.$$

Using the identity

$$R_j^2 - \mathbb{E}R_j^2 = (R_j - \mathbb{E}R_j)(R_j + \mathbb{E}R_j) - \mathrm{Var}(R_j),$$

Hölder's inequality, and $\mathrm{Var}(R_j) \le C_{s,2}^2\sigma_j^2$, we obtain for every fixed $p \ge 1$

$$\begin{aligned}
\|R_j^2 - \mathbb{E}R_j^2\|_{L^p} &\le \|(R_j - \mathbb{E}R_j)(R_j + \mathbb{E}R_j)\|_{L^p} + \mathrm{Var}(R_j) \\
&\le \|R_j - \mathbb{E}R_j\|_{L^{2p}} \|R_j + \mathbb{E}R_j\|_{L^{2p}} + \mathrm{Var}(R_j) \\
&\le C_{s,p}|\sigma_j|\big(C_{s,p}|\sigma_j| + 2|\sigma_j|\sqrt{m\nu}\big) \\
&\le C_{s,p}\sigma_\infty^2(1 + \sqrt{m\nu}).
\end{aligned}$$

Consequently,

$$\left\| \sigma_j^2 \left(\frac{\|\varepsilon_j\|_2^2}{m} - \nu \right) \right\|_{L^p} \le C_{s,p}\sigma_\infty^2 \left(\frac{1}{m} + \sqrt{\frac{\nu}{m}} \right).$$

For $j \ne k$, define

$$R_{jk}^{\pm} := \|\sigma_j\varepsilon_j \pm \sigma_k\varepsilon_k\|_2.$$

As functions of the independent pair $(\varepsilon_j, \varepsilon_k)$, these are Lipschitz with constants at most $(\sigma_j^2 + \sigma_k^2)^{1/2} \le \sqrt{2}\,\sigma_\infty$, and

$$\mathbb{E}(R_{jk}^+)^2 = \mathbb{E}(R_{jk}^-)^2 = m\nu(\sigma_j^2 + \sigma_k^2).$$

The preceding squared-norm argument, combined with the tensorized concentration bound, therefore gives

$$\left\| (R_{jk}^{\pm})^2 - \mathbb{E}(R_{jk}^{\pm})^2 \right\|_{L^p} \le C_{s,p}\sigma_\infty^2(1 + \sqrt{m\nu}).$$

By the polarization identity $4\sigma_j\sigma_k\varepsilon_j^\top\varepsilon_k = (R_{jk}^+)^2 - (R_{jk}^-)^2$ and the equality of the two expectations,

$$\left\| \frac{\sigma_j\sigma_k}{m} \varepsilon_j^\top\varepsilon_k \right\|_{L^p} \le C_{s,p}\sigma_\infty^2 \left(\frac{1}{m} + \sqrt{\frac{\nu}{m}} \right) \le C_{s,p}\sigma_\infty^2 \sqrt{\frac{\nu+1}{m}}.$$

Combining the two signal–noise terms with the noise–noise term gives, for $p = 2, 4, 6$ and uniformly over j, k,

$$\|\Delta_{jk}\|_{L^p} \le C_{s,p}\,\omega, \qquad \omega := \sigma_\infty \left(\frac{B_N}{m}\right)^{1/2} + \sigma_\infty^2 \sqrt{\frac{\nu+1}{m}},$$

so that $\omega = \eta_s$ after restoring the view index. Equivalently,

$$\mathbb{E}|\Delta_{jk}|^p \le C_{s,p}\,\omega^p, \qquad p = 2, 4, 6.$$

Taylor expansion. Taylor's theorem with integral remainder Firey [1960] gives

$$f(x_{jk} + \Delta_{jk}) - f(x_{jk}) = f'(x_{jk})\Delta_{jk} + \frac{1}{2}f''(x_{jk})\Delta_{jk}^2 + \mathcal{R}_{jk},$$

where

$$\mathcal{R}_{jk} = \frac{\Delta_{jk}^3}{2} \int_0^1 (1-r)^2 f^{(3)}(x_{jk} + r\Delta_{jk})\, dr.$$

Since $f^{(3)}$ is bounded by L_3,

$$|\mathcal{R}_{jk}| \le \frac{L_3}{6}|\Delta_{jk}|^3, \qquad \text{hence} \qquad \mathbb{E}|\mathcal{R}_{jk}|^2 \le CL_3^2\omega^6.$$

Define

$$A^{(1)}(j,k) := \frac{1}{N} f'(x_{jk})\Delta_{jk}, \qquad A^{(2)}(j,k) := \frac{1}{2N} f''(x_{jk})\Delta_{jk}^2, \qquad A^{(R)}(j,k) := \frac{1}{N}\mathcal{R}_{jk},$$

and the effective inner-product kernel matrix

$$K^\sharp(j,k) := \frac{1}{N} f\big(v_j^\top v_k + \mathbf{1}_{\{j=k\}}\nu\sigma_j^2\big) = \frac{1}{N} f(x_{jk}).$$

Then

$$K^{\mathrm{obs}} - K^\sharp = A^{(1)} + A^{(2)} + A^{(R)}.$$

Bias–fluctuation split. Define

$$\Upsilon' := \mathbb{E}[K^{\mathrm{obs}} - K^\sharp], \qquad \Xi := K^{\mathrm{obs}} - K^\sharp - \Upsilon'.$$

By construction, $\mathbb{E}\Xi = 0$. Since $\mathbb{E}\Delta_{jk} = 0$, we have $\mathbb{E}A^{(1)} = 0$, and hence

$$\Upsilon' = \mathbb{E}A^{(2)} + \mathbb{E}A^{(R)}.$$

For the second-order bias,

$$\|\mathbb{E}A^{(2)}\|_F^2 = \sum_{j,k} \left| \frac{f''(x_{jk})}{2N} \mathbb{E}\Delta_{jk}^2 \right|^2 \le \frac{L_2^2}{4N^2} \sum_{j,k} \big(\mathbb{E}\Delta_{jk}^2\big)^2 \le CL_2^2\omega^4,$$

so $\|\mathbb{E}A^{(2)}\|_F \le CL_2\omega^2$. For the Taylor-remainder bias,

$$\|\mathbb{E}A^{(R)}\|_F^2 \le \frac{1}{N^2} \sum_{j,k} \mathbb{E}|\mathcal{R}_{jk}|^2 \le CL_3^2\omega^6.$$

Combining the two estimates gives

$$\|\Upsilon'\|_F \le C_{s,b}\left(L_2\omega^2 + L_3\omega^3\right).$$

For the centered fluctuation,

$$\Xi = A^{(1)} + \left(A^{(2)} - \mathbb{E}A^{(2)}\right) + \left(A^{(R)} - \mathbb{E}A^{(R)}\right).$$

The three terms satisfy

$$\mathbb{E}\|A^{(1)}\|_F^2 = \frac{1}{N^2} \sum_{j,k} f'(x_{jk})^2 \mathbb{E}\Delta_{jk}^2 \le CL_1^2\omega^2,$$

$$\mathbb{E}\big\|A^{(2)} - \mathbb{E}A^{(2)}\big\|_F^2 \le \mathbb{E}\|A^{(2)}\|_F^2 \le \frac{L_2^2}{4N^2} \sum_{j,k} \mathbb{E}|\Delta_{jk}|^4 \le CL_2^2\omega^4,$$

$$\mathbb{E}\big\|A^{(R)} - \mathbb{E}A^{(R)}\big\|_F^2 \le \mathbb{E}\|A^{(R)}\|_F^2 \le CL_3^2\omega^6.$$

Since $\|A+B+C\|_F^2 \le 3(\|A\|_F^2 + \|B\|_F^2 + \|C\|_F^2)$ for any matrices A, B, C, we conclude that

$$\mathbb{E}\|\Xi\|_F^2 \le C_{s,v}\left(L_1^2\omega^2 + L_2^2\omega^4 + L_3^2\omega^6\right).$$

Diagonal correction. It remains to compare $K^\sharp$ with K^{sig}. The two matrices agree off the diagonal, so their difference is the deterministic diagonal matrix

$$D^{\mathrm{diag}}(j,k) := \mathbf{1}_{\{j=k\}} \frac{1}{N}\left[f\big(\|v_j\|_2^2 + \nu\sigma_j^2\big) - f(\|v_j\|_2^2)\right].$$

By the mean-value theorem,

$$|D^{\mathrm{diag}}(j,j)| = \frac{1}{N}\left|f(\|v_j\|_2^2 + \nu\sigma_j^2) - f(\|v_j\|_2^2)\right| \le \frac{L_1\nu\sigma_j^2}{N},$$

and therefore

$$\|D^{\mathrm{diag}}\|_F \le \frac{L_1\nu\left(\sum_{j=1}^N \sigma_j^4\right)^{1/2}}{N} \le \frac{L_1\nu\sigma_\infty^2}{\sqrt{N}}.$$

Conclusion. Restoring the view index, set

$$\Upsilon_s := \Upsilon' + D_s^{\mathrm{diag}}, \qquad \Xi_s := \Xi.$$

Then $K_s^{\mathrm{obs}} - K_s^{\mathrm{sig}} = \Upsilon_s + \Xi_s$ with $\mathbb{E}\Xi_s = 0$, and the bounds above give $\|\Upsilon_s\|_F \le b_s + d_s$ (taking $C_{s,b} \ge 1$) and $\mathbb{E}\|\Xi_s\|_F^2 \le v_s^2$. □

Lemma 4 (Linear-kernel specialization)**.** *Let $s \in \mathcal{S}_{\mathrm{lin}}$, so that*

$$k_s(u,u') = u^\top u',$$

and suppose that the inner-product noise regime of Assumption 2(1) holds. Assume, in addition, that the noise scale is homogeneous within the view, $\sigma_{s,j} \equiv \sigma_s$ for $j \in [N]$. Define

$$\chi_s^{\mathrm{lin}} := \frac{\nu_s\sigma_s^2}{N}, \qquad \Delta_s^{\mathrm{lin}} := K_s^{\mathrm{obs}} - \left(K_s^{\mathrm{sig}} + \chi_s^{\mathrm{lin}} I_N\right).$$

Then

$$\Delta_s^{\mathrm{lin}} = \Xi_s, \qquad \mathbb{E}\Xi_s = 0.$$

In particular, the deterministic finite-m_s bias vanishes: in the per-view decomposition $K_s^{\mathrm{obs}} - \alpha_s K_s^{\mathrm{sig}} - \chi_s I_N = \Upsilon_s + \Xi_s$ used in the proof of Theorem 1, with $\alpha_s = 1$ and $\chi_s = \chi_s^{\mathrm{lin}}$, one may take $\Upsilon_s = 0$, so $b_s = 0$. Moreover,

$$\mathbb{E}\,\|\Xi_s\|_F^2 \le v_s^2, \qquad v_s^2 := C_{s,v}\left[\frac{\sigma_s^2 B_{s,N}}{m_s} + \frac{\sigma_s^4(1+\nu_s)}{m_s}\right],$$

where $C_{s,v}$ depends only on (C_s, c_s, q_s) and a uniform upper bound on ν_s, and not on N or m_s.

Proof. We suppress the view index s and write

$$m = m_s, \qquad \sigma = \sigma_s, \qquad \nu = \nu_s, \qquad B_N = B_{s,N},$$

and $u_j = u_{s,j}$, $v_j = v_{s,j}$, $\varepsilon_j = \varepsilon_{s,j}$. As in the proof of Lemma 3, C denotes a constant that depends only on (C_s, c_s, q_s) and an upper bound on ν, and whose value may change from line to line. The observed entity vectors are $u_j = v_j + \sigma\varepsilon_j/\sqrt{m}$, and for the linear kernel

$$K^{\mathrm{obs}}(j,k) = \frac{1}{N}u_j^\top u_k, \qquad K^{\mathrm{sig}}(j,k) = \frac{1}{N}v_j^\top v_k.$$

Expanding the inner product gives

$$u_j^\top u_k = v_j^\top v_k + \frac{\sigma}{\sqrt{m}}\varepsilon_j^\top v_k + \frac{\sigma}{\sqrt{m}}\varepsilon_k^\top v_j + \frac{\sigma^2}{m}\varepsilon_j^\top\varepsilon_k.$$

Define

$$\begin{aligned}\Delta_{jk} &:= u_j^\top u_k - v_j^\top v_k - \mathbf{1}_{\{j=k\}}\nu\sigma^2 \\ &= \frac{\sigma}{\sqrt{m}}\varepsilon_j^\top v_k + \frac{\sigma}{\sqrt{m}}\varepsilon_k^\top v_j + \frac{\sigma^2}{m}\left(\varepsilon_j^\top\varepsilon_k - \mathbf{1}_{\{j=k\}}m\nu\right).\end{aligned}$$

We claim that $\mathbb{E}\Delta_{jk} = 0$. The two signal–noise terms have zero mean because the noise is centered. If $j \ne k$, independence and centering give $\mathbb{E}(\varepsilon_j^\top\varepsilon_k) = 0$; if $j = k$, the definition $\nu = \mathbb{E}\|\varepsilon_j\|_2^2/m$ gives

$$\mathbb{E}\left[\frac{\|\varepsilon_j\|_2^2}{m} - \nu\right] = 0.$$

The moment bounds established in the proof of Lemma 3 give, uniformly over $j, k \in [N]$,

$$\|\Delta_{jk}\|_{L^2} \le C\left[\sigma\left(\frac{B_N}{m}\right)^{1/2} + \sigma^2\sqrt{\frac{1+\nu}{m}}\right],$$

and hence

$$\mathbb{E}\,|\Delta_{jk}|^2 \le C\left[\frac{\sigma^2 B_N}{m} + \frac{\sigma^4(1+\nu)}{m}\right].$$

The residual matrix satisfies

$$\Delta^{\mathrm{lin}}(j,k) = K^{\mathrm{obs}}(j,k) - K^{\mathrm{sig}}(j,k) - \frac{\nu\sigma^2}{N}\mathbf{1}_{\{j=k\}} = \frac{1}{N}\Delta_{jk}.$$

Thus $\mathbb{E}\Delta^{\mathrm{lin}} = 0$, so we may set $\Xi := \Delta^{\mathrm{lin}}$ and $\Upsilon := 0$; in particular $b_s = 0$. Furthermore,

$$\mathbb{E}\,\|\Xi\|_F^2 = \frac{1}{N^2}\sum_{j,k=1}^{N}\mathbb{E}\,|\Delta_{jk}|^2 \le C_{s,v}\left[\frac{\sigma^2 B_N}{m} + \frac{\sigma^4(1+\nu)}{m}\right].$$

Restoring the view index gives $\mathbb{E}\|\Xi_s\|_F^2 \le v_s^2$. □

Lemma 5 (Gaussian distance-kernel bias–variance decomposition)**.** *Let $s \in \mathcal{S}_{\mathrm{gauss}}$, and suppose that the Gaussian distance-kernel regime of Assumption 2(2) holds. Define*

$$\Delta_s^{\mathrm{gauss}} := K_s^{\mathrm{obs}} - \left(\alpha_s K_s^{\mathrm{sig}} + \chi_s I_N\right), \qquad \alpha_s := \exp(-2\tau_s\sigma_s^2), \qquad \chi_s := \frac{1-\alpha_s}{N},$$

and

$$\eta_s^2 := \frac{\sigma_s^4\,\mathrm{tr}(\Psi_{s,m_s}^2)}{m_s^2} + \frac{\sigma_s^2\|\Psi_{s,m_s}\|_{\mathrm{op}}B_{s,N}}{m_s}.$$

Then there is a decomposition

$$\Delta_s^{\mathrm{gauss}} = \Upsilon_s + \Xi_s, \qquad \mathbb{E}[\Xi_s] = 0,$$

such that

$$\|\Upsilon_s\|_F \le b_s, \qquad \mathbb{E}\|\Xi_s\|_F^2 \le v_s^2,$$

where

$$b_s := C_{s,b}\left(\tau_s^2\eta_s^2 + \tau_s^3\eta_s^3\right), \qquad v_s^2 := C_{s,v}\left(\tau_s^2\eta_s^2 + \tau_s^4\eta_s^4 + \tau_s^6\eta_s^6\right),$$

and $C_{s,b}$ and $C_{s,v}$ are constants independent of N and m_s.

Proof. We suppress the view index s and write

$$m = m_s, \qquad \sigma = \sigma_s, \qquad \tau = \tau_s, \qquad \Psi = \Psi_{s,m_s}, \qquad B_N = B_{s,N},$$

and

$$\alpha = e^{-2\tau\sigma^2}, \qquad \chi = \frac{1-\alpha}{N}, \qquad \eta = \eta_s.$$

The signal vectors $\{v_j\}_{j=1}^N$ are treated as fixed, and all expectations below are taken with respect to the Gaussian noise only.

Step 1: Squared-distance perturbation. Set

$$\zeta_j := \frac{\sigma\varepsilon_j}{\sqrt{m}}, \qquad u_j = v_j + \zeta_j, \qquad \varepsilon_j \overset{\mathrm{iid}}{\sim} N(0,\Psi),$$

and, for $j \ne k$,

$$h_{jk} := v_j - v_k, \qquad \rho_{jk} := \|h_{jk}\|_2^2, \qquad a_{jk} := \zeta_j - \zeta_k.$$

Then $a_{jk} \sim N(0,\Gamma)$ with $\Gamma := 2\sigma^2\Psi/m$. Since $\mathrm{tr}(\Psi)/m = 1$, we have $\mathrm{tr}(\Gamma) = 2\sigma^2$, and consequently

$$\|u_j - u_k\|_2^2 = \|h_{jk} + a_{jk}\|_2^2 = \rho_{jk} + 2\sigma^2 + e_{jk},$$

where

$$e_{jk} := L_{jk} + Q_{jk}, \qquad L_{jk} := 2h_{jk}^\top a_{jk}, \qquad Q_{jk} := \|a_{jk}\|_2^2 - \mathrm{tr}(\Gamma).$$

For notational convenience we also set $h_{jj} := 0$ and $e_{jj} := 0$, so that the diagonal terms vanish in the sums over $j, k \in [N]$ below. Because a_{jk} is centered Gaussian, $\mathbb{E}[e_{jk}] = 0$. The cross term $\mathbb{E}[L_{jk}Q_{jk}]$ vanishes because L_{jk} is odd and Q_{jk} is even in a_{jk}. Therefore

$$\mathbb{E}[e_{jk}^2] = 4h_{jk}^\top\Gamma h_{jk} + 2\,\mathrm{tr}(\Gamma^2) = \frac{8\sigma^2}{m}h_{jk}^\top\Psi h_{jk} + \frac{8\sigma^4}{m^2}\,\mathrm{tr}(\Psi^2).$$

Step 2: Higher moments of e_{jk}. For $q = 1, 2, 3$, the elementary inequality $|a+b|^{2q} \le 2^{2q-1}(|a|^{2q} + |b|^{2q})$ gives

$$\mathbb{E}|e_{jk}|^{2q} \le C_q \left(\mathbb{E}|L_{jk}|^{2q} + \mathbb{E}|Q_{jk}|^{2q}\right).$$

Since L_{jk} is centered Gaussian with variance $4h_{jk}^\top \Gamma h_{jk}$,

$$\mathbb{E}|L_{jk}|^{2q} = \mathbb{E}|G|^{2q} \left(4h_{jk}^\top \Gamma h_{jk}\right)^q \le C_q \left(h_{jk}^\top \Gamma h_{jk}\right)^q, \qquad G \sim N(0,1).$$

For the quadratic term, diagonalize $\Gamma = U \operatorname{diag}(\lambda_1, \dots, \lambda_m) U^\top$. If $g \sim N(0, I_m)$, then $a_{jk} \overset{d}{=} \Gamma^{1/2} g$, and hence

$$Q_{jk} \overset{d}{=} \sum_{\ell=1}^m \lambda_\ell (g_\ell^2 - 1).$$

Thus Q_{jk} belongs to the second-order homogeneous Gaussian chaos, and Gaussian hypercontractivity [Nourdin et al., 2010, Proposition 2.6] gives $\|Q_{jk}\|_{L^p} \le (p-1)\|Q_{jk}\|_{L^2}$ for every $p \ge 2$. Taking $p = 2q$ and using $\mathbb{E}Q_{jk}^2 = 2\operatorname{tr}(\Gamma^2)$,

$$\mathbb{E}|Q_{jk}|^{2q} = \|Q_{jk}\|_{L^{2q}}^{2q} \le (2q-1)^{2q} \|Q_{jk}\|_{L^2}^{2q} \le C_q \left(\operatorname{tr}(\Gamma^2)\right)^q.$$

Combining the linear and quadratic bounds yields, for $q = 1, 2, 3$,

$$\mathbb{E}|e_{jk}|^{2q} \le C_q \left[\left(h_{jk}^\top \Gamma h_{jk}\right)^q + \left(\operatorname{tr}(\Gamma^2)\right)^q\right] \le C_q \left[h_{jk}^\top \Gamma h_{jk} + \operatorname{tr}(\Gamma^2)\right]^q.$$

We now use

$$h_{jk}^\top \Gamma h_{jk} \le \frac{2\sigma^2 \|\Psi\|_{\mathrm{op}}}{m} \|h_{jk}\|_2^2, \qquad \operatorname{tr}(\Gamma^2) = \frac{4\sigma^4}{m^2} \operatorname{tr}(\Psi^2),$$

together with the consequence of the signal-moment condition in Assumption 2(2),

$$\frac{1}{N^2} \sum_{j,k=1}^N \|h_{jk}\|_2^{2q} \le C_q B_N^q, \qquad q = 1, 2, 3,$$

to obtain

$$\frac{1}{N^2} \sum_{j,k=1}^N \mathbb{E}|e_{jk}|^{2q} \le C_q \eta^{2q}, \qquad q = 1, 2, 3, \qquad \eta^2 = \frac{\sigma^2 \|\Psi\|_{\mathrm{op}} B_N}{m} + \frac{\sigma^4 \operatorname{tr}(\Psi^2)}{m^2}. \tag{C.1}$$

Step 3: Taylor expansion of the kernel residual. For $j \ne k$, define

$$x_{jk} := \rho_{jk} + 2\sigma^2, \qquad y_{jk} := \|u_j - u_k\|_2^2 = x_{jk} + e_{jk},$$

both of which lie in $[0, \infty)$, and let $f(t) := e^{-\tau t}$ for $t \ge 0$. Since $\alpha K^{\mathrm{sig}}(j,k) = f(x_{jk})/N$, the off-diagonal residual is

$$\Delta^{\mathrm{gauss}}(j,k) = \frac{1}{N} \left[f(y_{jk}) - f(x_{jk})\right].$$

On the diagonal, the observed and target kernels agree exactly: $K^{\mathrm{obs}}(j,j) = 1/N$ and

$$\alpha K^{\mathrm{sig}}(j,j) + \chi = \frac{\alpha}{N} + \frac{1-\alpha}{N} = \frac{1}{N},$$

so $\Delta^{\mathrm{gauss}}(j,j) = 0$.

For $j \ne k$, Taylor's theorem with integral remainder [Firey, 1960] gives

$$f(y_{jk}) - f(x_{jk}) = f'(x_{jk}) e_{jk} + \frac{1}{2} f''(x_{jk}) e_{jk}^2 + \mathcal{R}_{jk},$$

where

$$\mathcal{R}_{jk} = \frac{e_{jk}^3}{2} \int_0^1 (1-r)^2 f^{(3)}(x_{jk} + r e_{jk})\, dr.$$

The segment between x_{jk} and y_{jk} lies in $[0,\infty)$, where $|f^{(3)}(t)| = \tau^3 e^{-\tau t} \le \tau^3$. Hence

$$|\mathcal{R}_{jk}| \le \frac{\tau^3}{6}|e_{jk}|^3, \qquad |\mathcal{R}_{jk}|^2 \le \frac{\tau^6}{36}|e_{jk}|^6,$$

and by (C.1) with $q = 3$,

$$\frac{1}{N^2}\sum_{j,k}\mathbb{E}|\mathcal{R}_{jk}|^2 \le C\tau^6\eta^6.$$

Define the random matrices

$$A^{(1)}(j,k) := \mathbf{1}_{\{j\neq k\}}\frac{f'(x_{jk})}{N}e_{jk}, \qquad A^{(2)}(j,k) := \mathbf{1}_{\{j\neq k\}}\frac{f''(x_{jk})}{2N}e_{jk}^2,$$

and

$$A^{(R)}(j,k) := \mathbf{1}_{\{j\neq k\}}\frac{\mathcal{R}_{jk}}{N},$$

so that $\Delta^{\text{gauss}} = A^{(1)} + A^{(2)} + A^{(R)}$.

Step 4: Bias. Define the bias and the centered fluctuation by

$$\Upsilon := \mathbb{E}[\Delta^{\text{gauss}}], \qquad \Xi := \Delta^{\text{gauss}} - \Upsilon.$$

Since $\mathbb{E}[e_{jk}] = 0$, we have $\mathbb{E}[A^{(1)}] = 0$, and hence $\Upsilon = \mathbb{E}[A^{(2)}] + \mathbb{E}[A^{(R)}]$. Because $|f''(x)| \le \tau^2$ for $x \ge 0$, Jensen's inequality gives

$$\|\mathbb{E}A^{(2)}\|_F^2 = \sum_{j\neq k}\left|\frac{f''(x_{jk})}{2N}\mathbb{E}[e_{jk}^2]\right|^2 \le \frac{\tau^4}{4N^2}\sum_{j\neq k}\mathbb{E}[e_{jk}^4] \le C\tau^4\eta^4.$$

Similarly,

$$\|\mathbb{E}A^{(R)}\|_F^2 \le \frac{1}{N^2}\sum_{j\neq k}\mathbb{E}|\mathcal{R}_{jk}|^2 \le C\tau^6\eta^6.$$

Therefore

$$\|\Upsilon\|_F \le \|\mathbb{E}A^{(2)}\|_F + \|\mathbb{E}A^{(R)}\|_F \le C\left(\tau^2\eta^2 + \tau^3\eta^3\right).$$

Step 5: Fluctuation. Write

$$\Xi = A^{(1)} + \left(A^{(2)} - \mathbb{E}A^{(2)}\right) + \left(A^{(R)} - \mathbb{E}A^{(R)}\right).$$

For the first-order term,

$$\mathbb{E}\|A^{(1)}\|_F^2 = \frac{1}{N^2}\sum_{j\neq k} f'(x_{jk})^2\mathbb{E}[e_{jk}^2] \le C\tau^2\eta^2.$$

For the centered second-order and remainder terms,

$$\mathbb{E}\left\|A^{(2)} - \mathbb{E}A^{(2)}\right\|_F^2 \le \mathbb{E}\|A^{(2)}\|_F^2 \le C\tau^4\eta^4, \qquad \mathbb{E}\left\|A^{(R)} - \mathbb{E}A^{(R)}\right\|_F^2 \le \mathbb{E}\|A^{(R)}\|_F^2 \le C\tau^6\eta^6.$$

Using $\|M_1 + M_2 + M_3\|_F^2 \le 3(\|M_1\|_F^2 + \|M_2\|_F^2 + \|M_3\|_F^2)$, we conclude that

$$\mathbb{E}\|\Xi\|_F^2 \le C\left(\tau^2\eta^2 + \tau^4\eta^4 + \tau^6\eta^6\right).$$

Restoring the view index s, the bounds $\|\Upsilon_s\|_F \le b_s$ and $\mathbb{E}\|\Xi_s\|_F^2 \le v_s^2$ hold with b_s and v_s^2 as defined in the statement. □

Lemma 6 (Mixed target eigengap)**.** *For $w \in \Delta_S$, define*

$$K_{\text{tar}}(w) := \sum_{s=1}^{S} w_s\alpha_s K_s^{\text{sig}} + \left(\sum_{s=1}^{S} w_s\chi_s\right) I_N.$$

Under Assumptions 3 and 4, the top-d eigenspace of $K_{\mathrm{tar}}(w)$ is $\mathrm{span}(Z_{\mathrm{sig}})$, and its eigengap is at least

$$\gamma(w) := \lambda_{\min}\left(\sum_{s=1}^{S} w_s \alpha_s \Lambda_s\right) - \left\|\sum_{s=1}^{S} w_s \alpha_s R_s^{\mathrm{sig}}\right\|_{\mathrm{op}}.$$

Moreover, with the per-view gaps $\gamma_s := \lambda_{\min}(\Lambda_s) - \|R_s^{\mathrm{sig}}\|_{\mathrm{op}}$,

$$\gamma(w) \geq \sum_{s=1}^{S} w_s \alpha_s \gamma_s.$$

Proof. By the shared-subspace decomposition in Assumption 3,

$$K_s^{\mathrm{sig}} = Z_{\mathrm{sig}} \Lambda_s Z_{\mathrm{sig}}^\top + R_s^{\mathrm{sig}}, \qquad R_s^{\mathrm{sig}} Z_{\mathrm{sig}} = 0, \qquad Z_{\mathrm{sig}}^\top R_s^{\mathrm{sig}} = 0.$$

Therefore

$$K_{\mathrm{tar}}(w) = Z_{\mathrm{sig}}\left(\sum_{s=1}^{S} w_s \alpha_s \Lambda_s\right) Z_{\mathrm{sig}}^\top + \sum_{s=1}^{S} w_s \alpha_s R_s^{\mathrm{sig}} + \left(\sum_{s=1}^{S} w_s \chi_s\right) I_N.$$

The last term is a scalar multiple of I_N; it shifts all eigenvalues equally and changes neither the eigenspaces nor the eigengaps, so we may ignore it. The remaining matrix is block diagonal with respect to the decomposition

$$\mathbb{R}^N = \mathrm{span}(Z_{\mathrm{sig}}) \oplus \mathrm{span}(Z_{\mathrm{sig}})^\perp.$$

On $\mathrm{span}(Z_{\mathrm{sig}})$ it acts as $\sum_s w_s \alpha_s \Lambda_s$ in the orthonormal basis Z_{sig}, so its d eigenvalues there are at least $\lambda_{\min}(\sum_s w_s \alpha_s \Lambda_s)$. On $\mathrm{span}(Z_{\mathrm{sig}})^\perp$ it acts as $\sum_s w_s \alpha_s R_s^{\mathrm{sig}}$, so its remaining eigenvalues are at most $\|\sum_s w_s \alpha_s R_s^{\mathrm{sig}}\|_{\mathrm{op}}$. Hence

$$\lambda_d\{K_{\mathrm{tar}}(w)\} - \lambda_{d+1}\{K_{\mathrm{tar}}(w)\} \geq \gamma(w).$$

Whenever $\gamma(w) > 0$, which holds by Assumption 4 and the bound below, the d largest eigenvalues are exactly those on $\mathrm{span}(Z_{\mathrm{sig}})$, so the top-$d$ eigenspace is $\mathrm{span}(Z_{\mathrm{sig}})$.

Finally, since $\lambda_{\min}$ is superadditive on symmetric matrices and the operator norm is subadditive,

$$\gamma(w) \geq \sum_{s=1}^{S} w_s \alpha_s \lambda_{\min}(\Lambda_s) - \sum_{s=1}^{S} w_s \alpha_s \|R_s^{\mathrm{sig}}\|_{\mathrm{op}} = \sum_{s=1}^{S} w_s \alpha_s \gamma_s.$$

□

Proof of Theorem 1. Let $K_{\mathrm{tar}}(w)$ be the target matrix of Lemma 6. By that lemma, the top-d eigenspace of $K_{\mathrm{tar}}(w)$ is $\mathrm{span}(Z_{\mathrm{sig}})$, with eigengap at least $\gamma(w)$.

For each view, let Υ_s, Ξ_s, $b_s := b_{s,N,m_s}$, $d_s := d_{s,N,m_s}$, and $v_s := v_{s,N,m_s}$ denote the bias matrix, the centered fluctuation matrix, and the corresponding error scales defined in Appendix B.3 and Lemmas 3, 4, and 5, according to the kernel type. Then

$$K_s^{\mathrm{obs}} - \alpha_s K_s^{\mathrm{sig}} - \chi_s I_N = \Upsilon_s + \Xi_s, \qquad \mathbb{E}\,\Xi_s = 0,$$

and

$$\|\Upsilon_s\|_F \leq b_s + \mathbf{1}_{\{s \in \mathcal{S}_{\mathrm{ip}}\}} d_s, \qquad \mathbb{E}\|\Xi_s\|_F^2 \leq v_s^2.$$

Since the centered matrices are independent across views and have mean zero, $\mathbb{E}\langle \Xi_s, \Xi_t \rangle_F = \langle \mathbb{E}\,\Xi_s, \mathbb{E}\,\Xi_t \rangle_F = 0$ for $s \neq t$. Consequently,

$$\mathbb{E}\left\|\sum_{s=1}^{S} w_s \Xi_s\right\|_F^2 = \sum_{s=1}^{S} w_s^2 \mathbb{E}\|\Xi_s\|_F^2 \leq \sum_{s=1}^{S} w_s^2 v_s^2.$$

Markov's inequality applied to the squared Frobenius norm gives

$$\mathbb{P}\left\{\left\|\sum_{s=1}^{S} w_s \Xi_s\right\|_F > \left(\frac{1}{\delta}\sum_{s=1}^{S} w_s^2 v_s^2\right)^{1/2}\right\} \leq \delta.$$

Hence, by the triangle inequality, with probability at least $1-\delta$,

$$\|K^{\mathrm{obs}}(w) - K_{\mathrm{tar}}(w)\|_F \le \sum_{s=1}^{S} w_s b_s + \sum_{s\in\mathcal{S}_{\mathrm{ip}}} w_s d_s + \left(\frac{1}{\delta}\sum_{s=1}^{S} w_s^2 v_s^2\right)^{1/2}.$$

The Frobenius-norm form of the Davis–Kahan $\sin\Theta$ theorem [Yu et al., 2015] gives

$$\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F \le \frac{C}{\gamma(w)}\|K^{\mathrm{obs}}(w) - K_{\mathrm{tar}}(w)\|_F,$$

so, with probability at least $1-\delta$,

$$\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F \le \frac{C}{\gamma(w)}\left[\sum_{s=1}^{S} w_s b_s + \sum_{s\in\mathcal{S}_{\mathrm{ip}}} w_s d_s + \left(\frac{1}{\delta}\sum_{s=1}^{S} w_s^2 v_s^2\right)^{1/2}\right].$$

Squaring and using $(a+b)^2 \le 2(a^2+b^2)$, we obtain, after enlarging C,

$$\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2 \le \frac{C}{\gamma(w)^2}\left[\left(\sum_{s=1}^{S} w_s b_s + \sum_{s\in\mathcal{S}_{\mathrm{ip}}} w_s d_s\right)^2 + \frac{1}{\delta}\sum_{s=1}^{S} w_s^2 v_s^2\right],$$

which is (4.4). □

C.2 Proof of Theorem 2 (end-to-end risk bound)

We now prove the end-to-end bound. The argument conditions on the unlabeled pretraining sample, so the empirical KPCA coordinate matrix $\widehat{Z}$ is fixed throughout the supervised analysis. The proof has three components: a fixed-embedding oracle inequality, a transfer bound from the empirical KPCA coordinates to the true latent coordinates, and a ReLU approximation bound for the smooth relation surfaces. Write

$$\mathcal{D}_\star := [-R_\star, R_\star]^d \times [-R_\star, R_\star]^d.$$

By Assumption 7, every latent coordinate satisfies $\|z_j^\star\|_2 \le R_\star - 1$, so $\mathcal{D}_\star$ contains all pairs of latent coordinates with a margin of 1. The transformed KPCA coordinates introduced below need not lie in $\mathcal{D}_\star$; the proof of Lemma 10 handles the entities whose coordinates are poorly recovered separately. Throughout this subsection, $\mathcal{F}$ denotes the class of relation-wise head collections $f = (f_1, \dots, f_K)$, where each f_r has the form (3.5) with depth at most D, at most W parameters, and clipping level B; for such f we write $f(\widehat{Z}; x) := f_r(\widehat{z}_h, \widehat{z}_t)$ for $x = (h, r, t)$, so that $\{x \mapsto f(\widehat{Z}; x) : f \in \mathcal{F}\} = \mathcal{F}_{\widehat{Z}}(D, W, B)$, conditionally on the pretraining sample.

Lemma 7 (Fixed-embedding oracle inequality)**.** *Condition on the pretraining sample and define the frozen class*

$$\mathcal{F}_{\widehat{Z}} := \{x \mapsto f(\widehat{Z}; x) : f \in \mathcal{F}\} = \mathcal{F}_{\widehat{Z}}(D, W, B).$$

Let p be any real number with $\mathrm{Pdim}(\mathcal{F}_{\widehat{Z}}) \le p \le n$. Under Assumption 6, with conditional probability at least $1 - 10(p/en)^p$,

$$\|\widehat{f}(\widehat{Z}; \cdot) - f^\star\|_P^2 \le 27 \inf_{f\in\mathcal{F}} \|f(\widehat{Z}; \cdot) - f^\star\|_P^2 + C_{\mathrm{or}}(\sigma_y^2 + B^2)\frac{p}{n}\log\frac{en}{p} + 6\delta_{\mathrm{opt}}, \tag{C.2}$$

where C_{or} is an absolute constant. Furthermore, for relation-wise ReLU heads with common depth D and parameter budget W, there is an absolute constant $C_{\mathrm{pd}} \ge 1$ such that, for every $\widehat{Z}$,

$$\mathrm{Pdim}(\mathcal{F}_{\widehat{Z}}) \le C_{\mathrm{pd}}\, KDW \log W.$$

Proof. Conditionally on the pretraining sample, $\widehat{Z}$ is deterministic, and the only randomness in the supervised stage comes from the labeled sample. The standard bounded least-squares oracle inequality for a fixed function class therefore applies to $\mathcal{F}_{\widehat{Z}}$: Theorem 1 of Liu et al. [2024], in the present notation, gives (C.2) for any δ_{opt}-approximate empirical risk minimizer. That theorem

is stated with $p = \mathrm{Pdim}(\mathcal{F}_{\widehat{Z}})$, but its proof uses the pseudo-dimension only through the uniform covering bound of Anthony and Bartlett [1999, Theorem 12.2],

$$\mathcal{N}_\infty(\epsilon, \mathcal{F}_{\widehat{Z}}, n) \le \sum_{i=0}^{\mathrm{Pdim}(\mathcal{F}_{\widehat{Z}})} \binom{n}{i} \Big(\frac{2B}{\epsilon}\Big)^i,$$

where $2B$ is the length of the range $[-B, B]$ of the clipped heads, applied with $\epsilon \le 2B$; every other occurrence of p is a free parameter of the argument. For $x := 2B/\epsilon \ge 1$ and any real p with $\mathrm{Pdim}(\mathcal{F}_{\widehat{Z}}) \le p \le n$, each index satisfies $i \le p$, so $x^i \le x^p$ and $(p/n)^i \ge (p/n)^p$; hence

$$\sum_{i=0}^{\mathrm{Pdim}(\mathcal{F}_{\widehat{Z}})} \binom{n}{i} x^i \le x^p \Big(\frac{n}{p}\Big)^p \sum_{i=0}^{n} \binom{n}{i} \Big(\frac{p}{n}\Big)^i = x^p \Big(\frac{n}{p}\Big)^p \Big(1 + \frac{p}{n}\Big)^n \le \Big(\frac{2enB}{\epsilon p}\Big)^p.$$

This is the covering bound used in that proof, now with p in place of the pseudo-dimension, so the conclusion holds for every such p. The pseudo-dimension bound follows from the standard estimate for ReLU networks, as in Lemma 3 of Liu et al. [2024]; with K heads of common architecture (D, W), the contributions add over relations and give $\mathrm{Pdim}(\mathcal{F}_{\widehat{Z}}) \le C_{\mathrm{pd}} KDW \log W$. □

Lemma 8 (Lipschitz property of the smooth oracle surfaces)**.** *Under Assumption 7, since $\beta \ge 1$, each relation surface $g_r^\star$ is Lipschitz on $\mathcal{D}_\star$. More precisely, there exists $L_\star \le C(d) S_\beta$ such that, for all $(a, b), (a', b') \in \mathcal{D}_\star$,*

$$|g_r^\star(a, b) - g_r^\star(a', b')| \le L_\star \big(\|a - a'\|_2 + \|b - b'\|_2\big).$$

Proof. The bound $\|g_r^\star\|_{W^{\beta,\infty}} \le S_\beta$ with $\beta \ge 1$ implies that all first-order weak derivatives are essentially bounded by S_β, so $\sup_z \|\nabla g_r^\star(z)\|_2 \le C(d) S_\beta$. Since $\mathcal{D}_\star$ is convex, the mean-value inequality gives

$$|g_r^\star(u) - g_r^\star(v)| \le \sup_z \|\nabla g_r^\star(z)\|_2 \, \|u - v\|_2 \le C(d) S_\beta \|u - v\|_2.$$

Writing $u = (a, b)$ and $v = (a', b')$ and using $\|(a - a', b - b')\|_2 \le \|a - a'\|_2 + \|b - b'\|_2$ gives the claim. □

Lemma 9 (KPCA coordinate alignment)**.** *Let*

$$Q \in \arg\min_{U \in O(d)} \|\widehat{Z} U - Z_{\mathrm{sig}}\|_F, \qquad \widetilde{Z} := \sqrt{N}\, \widehat{Z} Q A_\star^{-1},$$

and let $\widetilde{z}_j$ denote the jth row of $\widetilde{Z}$. Under Assumption 8, the entity-averaged coordinate error satisfies

$$\bar{\Delta}^2 := \frac{1}{N} \sum_{j=1}^{N} \|\widetilde{z}_j - z_j^\star\|_2^2 \le 2\kappa_A^2 \|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2.$$

Proof. By the standard Procrustes bound for principal angles,

$$\min_{U \in O(d)} \|\widehat{Z} U - Z_{\mathrm{sig}}\|_F \le \sqrt{2}\, \|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F.$$

Since $\sqrt{N} Z_{\mathrm{sig}} = Z^\star A_\star$,

$$\widetilde{Z} - Z^\star = \sqrt{N}\, \widehat{Z} Q A_\star^{-1} - \sqrt{N} Z_{\mathrm{sig}} A_\star^{-1} = \sqrt{N}\, (\widehat{Z} Q - Z_{\mathrm{sig}}) A_\star^{-1}.$$

Therefore

$$\bar{\Delta}^2 = \frac{1}{N} \|\widetilde{Z} - Z^\star\|_F^2 \le \|\widehat{Z} Q - Z_{\mathrm{sig}}\|_F^2 \, \|A_\star^{-1}\|_{\mathrm{op}}^2 \le 2\kappa_A^2 \|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2.$$

□

Lemma 10 (Transfer of the oracle bias to empirical KPCA coordinates)**.** *Suppose Assumptions 6–8 hold, and that there exists a relation-wise head collection $f^{\mathrm{app}} \in \mathcal{F}$, each of whose heads has at most $W/(2d)$ parameters, such that*

$$\sup_{r\in[K]} \sup_{(a,b)\in\mathcal{D}_\star} |f_r^{\mathrm{app}}(a,b) - g_r^\star(a,b)| \le \varepsilon_{\mathrm{app}}.$$

Then, with $\bar{L} := \max\{L_\star, 2B\}$ and $\bar{\Delta}$ as in Lemma 9,

$$\inf_{f\in\mathcal{F}} \|f(\widehat{Z};\cdot) - f^\star\|_P^2 \le 2\varepsilon_{\mathrm{app}}^2 + 8\bar{L}^2\bar{\Delta}^2, \tag{C.3}$$

and consequently

$$\inf_{f\in\mathcal{F}} \|f(\widehat{Z};\cdot) - f^\star\|_P^2 \le 2\varepsilon_{\mathrm{app}}^2 + 16\bar{L}^2\kappa_A^2 \|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2. \tag{C.4}$$

Proof. Let Q and $\widetilde{Z}$ be as in Lemma 9, and set

$$A := \sqrt{N}\, Q A_\star^{-1}, \qquad \text{so that} \qquad \widetilde{z}_j = A^\top \widehat{z}_j, \quad j \in [N].$$

For a relation-wise head f, define the linear reparameterization

$$f_r^{[A]}(a,b) := f_r(A^\top a, A^\top b).$$

For ReLU networks, precomposition with the fixed invertible map $(a,b) \mapsto (A^\top a, A^\top b)$ is absorbed into the first affine layer: the first-layer weight matrix is multiplied on the right by the $2d \times 2d$ matrix $\mathrm{blockdiag}(A^\top, A^\top)$, which leaves the depth and all other layers unchanged and gives each first-layer row at most $2d$ nonzero entries. Hence each reparameterized head has at most $2d$ times as many parameters as the original head, that is, at most W, and the output clipping is unaffected. Therefore

$$\widetilde{f}^{\mathrm{app}} := (f^{\mathrm{app}})^{[A]} \in \mathcal{F}.$$

Because the heads in $\mathcal{F}$ are clipped, $\widetilde{f}^{\mathrm{app}}$ takes values in $[-B, B]$.

Write $\Delta_j := \|\widetilde{z}_j - z_j^\star\|_2$ for $j \in [N]$, and fix a triple $x = (h, r, t)$. By construction,

$$\widetilde{f}^{\mathrm{app}}(\widehat{Z}; x) = \widetilde{f}_r^{\mathrm{app}}(\widehat{z}_h, \widehat{z}_t) = f_r^{\mathrm{app}}(\widetilde{z}_h, \widetilde{z}_t), \qquad f^\star(x) = g_r^\star(z_h^\star, z_t^\star).$$

We claim that

$$|\widetilde{f}^{\mathrm{app}}(\widehat{Z}; x) - f^\star(x)| \le \varepsilon_{\mathrm{app}} + \bar{L}(\Delta_h + \Delta_t). \tag{C.5}$$

Case 1: $\Delta_h \le 1$ and $\Delta_t \le 1$. By Assumption 7, $\|\widetilde{z}_h\|_2 \le \|z_h^\star\|_2 + \Delta_h \le R_\star$, and likewise for t; hence every coordinate of $\widetilde{z}_h$ and $\widetilde{z}_t$ lies in $[-R_\star, R_\star]$, and both $(\widetilde{z}_h, \widetilde{z}_t)$ and $(z_h^\star, z_t^\star)$ lie in $\mathcal{D}_\star$. By the approximation hypothesis and Lemma 8,

$$\begin{aligned} |\widetilde{f}^{\mathrm{app}}(\widehat{Z}; x) - f^\star(x)| &\le |f_r^{\mathrm{app}}(\widetilde{z}_h, \widetilde{z}_t) - g_r^\star(\widetilde{z}_h, \widetilde{z}_t)| + |g_r^\star(\widetilde{z}_h, \widetilde{z}_t) - g_r^\star(z_h^\star, z_t^\star)| \\ &\le \varepsilon_{\mathrm{app}} + L_\star(\Delta_h + \Delta_t). \end{aligned}$$

Case 2: $\Delta_h > 1$ or $\Delta_t > 1$. Both $\widetilde{f}^{\mathrm{app}}(\widehat{Z}; x)$ and $f^\star(x)$ lie in $[-B, B]$, by the clipping and by Assumption 6, respectively, so

$$|\widetilde{f}^{\mathrm{app}}(\widehat{Z}; x) - f^\star(x)| \le 2B < 2B(\Delta_h + \Delta_t).$$

In both cases (C.5) holds.

Squaring (C.5) and using $(x+y)^2 \le 2x^2 + 2y^2$ twice gives

$$|\widetilde{f}^{\mathrm{app}}(\widehat{Z}; x) - f^\star(x)|^2 \le 2\varepsilon_{\mathrm{app}}^2 + 4\bar{L}^2(\Delta_h^2 + \Delta_t^2).$$

By Assumption 6, the head and the tail of $X \sim P$ are each uniformly distributed on $[N]$, so $\mathbb{E}_{X\sim P}\Delta_h^2 = \mathbb{E}_{X\sim P}\Delta_t^2 = N^{-1}\sum_{j=1}^N \Delta_j^2 = \bar{\Delta}^2$. Taking the expectation over $X \sim P$ therefore gives

$$\|\widetilde{f}^{\mathrm{app}}(\widehat{Z};\cdot) - f^\star\|_P^2 \le 2\varepsilon_{\mathrm{app}}^2 + 8\bar{L}^2\bar{\Delta}^2.$$

Since $\widetilde{f}^{\mathrm{app}} \in \mathcal{F}$, this proves (C.3). Substituting Lemma 9 proves (C.4). □

Lemma 11 (Yarotsky approximation on the relation surfaces)**.** *There are constants C_{app} and W_0 depending only on $(d, \beta, R_\star, S_\beta, B)$ such that, for all $W \geq W_0$ and $D \gtrsim \log W$, there exists $f^{\mathrm{app}} \in \mathcal{F}$, each of whose heads has at most $W/(2d)$ parameters, satisfying*

$$\sup_{r\in[K]} \sup_{(a,b)\in\mathcal{D}_\star} |f_r^{\mathrm{app}}(a,b) - g_r^\star(a,b)| \leq C_{\mathrm{app}} \Big(\frac{W}{\log W}\Big)^{-\beta/(2d)}. \tag{C.6}$$

Proof. Each $g_r^\star$ is defined on a $2d$-dimensional cube and satisfies a uniform $W^{\beta,\infty}$ bound. After an affine rescaling of the input cube to $[0,1]^{2d}$ and a harmless output normalization, the ReLU approximation theorem of Yarotsky [2017] gives, for every $\varepsilon \in (0,1)$, a ReLU network of depth $O(\log(1/\varepsilon))$ with $O(\varepsilon^{-2d/\beta}\log(1/\varepsilon))$ weights whose uniform error is at most $C\varepsilon$. Applying this construction separately to the K relation surfaces under the common architecture budget gives a relation-wise collection in $\mathcal{F}$. Since the heads have bounded outputs, the clipping map $u \mapsto -B + (u+B)_+ - (u-B)_+$ can be appended at constant cost. Choosing

$$\varepsilon_W \asymp \Big(\frac{W}{\log W}\Big)^{-\beta/(2d)}$$

keeps the parameter count of each head at most $W/(2d)$ and the depth $O(\log W)$; the factor $2d$ only changes C_{app} and W_0, since d is fixed. This proves (C.6). □

Completion of the proof. *Step 1: conditional bound.* Condition on the pretraining sample. By Lemma 11, the hypothesis of Lemma 10 holds with

$$\varepsilon_{\mathrm{app}} = C_{\mathrm{app}} \Big(\frac{W}{\log W}\Big)^{-\beta/(2d)},$$

so (C.4) gives

$$\inf_{f\in\mathcal{F}} \|f(\widehat{Z};\cdot) - f^\star\|_P^2 \leq 2C_{\mathrm{app}}^2 \Big(\frac{W}{\log W}\Big)^{-\beta/d} + 16\bar{L}^2\kappa_A^2 \|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2.$$

Since $\bar{L}^2 \leq L_\star^2 + 4B^2 \leq 4(L_\star^2 + B^2)$, the last term is at most $64(L_\star^2+B^2)\kappa_A^2\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2$. Substituting this bound into Lemma 7 shows that, with conditional probability at least $1 - 10(p/en)^p$,

$$\begin{aligned}\|\widehat{f}(\widehat{Z};\cdot) - f^\star\|_P^2 \leq 54C_{\mathrm{app}}^2 \Big(\frac{W}{\log W}\Big)^{-\beta/d} + 1728(L_\star^2+B^2)\kappa_A^2\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2 \\ + C_{\mathrm{or}}(\sigma_y^2 + B^2)\frac{p}{n}\log\frac{en}{p} + 6\delta_{\mathrm{opt}}.\end{aligned} \tag{C.7}$$

We apply Lemma 7 with the deterministic choice

$$p := C_{\mathrm{pd}}\, KDW \log W,$$

which satisfies $\mathrm{Pdim}(\mathcal{F}_{\widehat{Z}}) \leq p \leq n$ for every realization of $\widehat{Z}$ whenever $C_{\mathrm{pd}}KDW\log W \leq n$. Since $C_{\mathrm{pd}} \geq 1$,

$$\frac{p}{n}\log\frac{en}{p} \leq C_{\mathrm{pd}}\,\frac{KDW\log W}{n}\log\frac{en}{KDW\log W}.$$

Step 2: combining the two stages. Let

$$\mathcal{D}_{\mathrm{pt}} := \{E_s : s \in [S]\}$$

denote the unlabeled pretraining data, equivalently $\{u_{s,j} : s \in [S],\, j \in [N]\}$. Define the pretraining event

$$\Omega_{\mathrm{pt}} := \Big\{\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F \leq a_{\mathrm{pt}}\Big\},$$

and let $\Omega_{\mathrm{sup}}(\widehat{Z})$ denote the event on which (C.7) holds, conditionally on $\mathcal{D}_{\mathrm{pt}}$. Suppose that

$$\mathbb{P}(\Omega_{\mathrm{pt}}) \geq 1 - \delta_{\mathrm{pt}}$$

and that, for every realization of $\mathcal{D}_{\mathrm{pt}}$,

$$\mathbb{P}\Big(\Omega_{\mathrm{sup}}(\widehat{Z}) \mid \mathcal{D}_{\mathrm{pt}}\Big) \geq 1 - \delta_{\mathrm{sup}}.$$

Then

$$\mathbb{P}\Big(\Omega_{\mathrm{pt}} \cap \Omega_{\mathrm{sup}}(\widehat{Z})\Big) = \mathbb{E}\Big[\mathbf{1}_{\Omega_{\mathrm{pt}}}\, \mathbb{P}\Big(\Omega_{\mathrm{sup}}(\widehat{Z}) \mid \mathcal{D}_{\mathrm{pt}}\Big)\Big] \geq (1-\delta_{\mathrm{sup}})\, \mathbb{P}(\Omega_{\mathrm{pt}})$$
$$\geq (1-\delta_{\mathrm{sup}})(1-\delta_{\mathrm{pt}}) \geq 1-\delta_{\mathrm{sup}}-\delta_{\mathrm{pt}}.$$

Step 3: conclusion. Let C_{pt} denote the constant C of Theorem 1. By that theorem, we may take

$$a_{\mathrm{pt}}^2 = C_{\mathrm{pt}}\, \mathcal{E}_{\mathrm{pt}}(w,\delta), \qquad \delta_{\mathrm{pt}} = \delta,$$

and, by Lemma 7 with $p = C_{\mathrm{pd}} KDW \log W$ as in Step 1, $\delta_{\mathrm{sup}} = 10(p/en)^p$, which does not depend on $\mathcal{D}_{\mathrm{pt}}$. On $\Omega_{\mathrm{pt}} \cap \Omega_{\mathrm{sup}}(\widehat{Z})$, inserting $\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_F^2 \leq C_{\mathrm{pt}}\mathcal{E}_{\mathrm{pt}}(w,\delta)$ and the pseudo-dimension bound of Step 1 into (C.7) yields (4.7) with probability at least $1 - 10(p/en)^p - \delta$ and constants

$$C_1 = 54\, C_{\mathrm{app}}^2, \qquad C_2 = C_{\mathrm{or}} C_{\mathrm{pd}}, \qquad C_3 = 1728\, C_{\mathrm{pt}}, \qquad C_4 = 6.$$

Thus C_1 depends only on $(d, \beta, R_\star, S_\beta, B)$, C_2 is absolute, C_3 depends only on the view-level noise constants of Theorem 1, and C_4 is absolute. This completes the proof of Theorem 2.

D Simulation Details

D.1 Pretraining recovery and view weighting

For pretraining, we use the linear kernel model:

$$u_{s,j} = v_{s,j} + \frac{1}{\sqrt{m_s}}\varepsilon_{s,j}, \qquad j = 1, \ldots, N,$$

where $u_{s,j} \in \mathbb{R}^{m_s}$ is the observed entity vector for entity j in view s, $v_{s,j}$ is the corresponding noiseless vector, and $\varepsilon_{s,j}$ is Gaussian noise across entities. We use

$$N = 60, \qquad d = 4, \qquad S = 4,$$

and a shared latent subspace $Z_{\mathrm{sig}} \in \mathbb{R}^{N\times d}$ satisfying $Z_{\mathrm{sig}}^\top Z_{\mathrm{sig}} = I_d$, the shared signal eigenspace of the theory. For each view,

$$V_s = Z_{\mathrm{sig}} T_s^\top, \qquad T_s^\top T_s = N\Lambda_s,$$

so that the noiseless linear kernel satisfies $\frac{1}{N} V_s V_s^\top = Z_{\mathrm{sig}} \Lambda_s Z_{\mathrm{sig}}^\top$. The view-specific signal spectra are

$$\mathrm{diag}(\Lambda_s) \in \{(10, 7, 4.5, 2),\ (7, 5, 3, 1.5),\ (4.8, 3.5, 2, 1),\ (8.5, 6, 3.5, 1.8)\}.$$

Each 4-tuple gives the $d = 4$ nonzero signal eigenvalues for one view; there are four tuples because $S = 4$.

We compare two noise regimes.
(i) In the isotropic regime,

$$\varepsilon_{s,j} \sim N(0, \sigma^2 I), \qquad \sigma = 0.25.$$

(ii) In the anisotropic regime,

$$\varepsilon_{s,j} \sim N(0, \Sigma_s), \qquad \Sigma_s = \sigma_s^2 I + Q_s D_s Q_s^\top,$$

where the anisotropic low-rank component is fixed for each view and is independent of Z_{sig}. The view-specific base noise levels are

$$(\sigma_s) = (0.15, 0.20, 0.25, 0.30),$$

with low-rank scales $(0.8, 1.2, 2.0, 2.8)$ and ranks $6, 6, 8, 8$, respectively.

For a given weight vector w, we form the aggregated empirical linear kernel

$$\widehat{K}(w) = \sum_{s=1}^{S} w_s \frac{1}{N} E_s E_s^\top.$$

The reported subspace recovery error is the operator-norm principal-angle error

$$\|\sin\Theta(\widehat{Z}, Z_{\mathrm{sig}})\|_{\mathrm{op}},$$

where $\widehat{Z}$ denotes the leading d-dimensional eigenspace of the aggregated empirical kernel $\widehat{K}(w)$. All reported means and standard deviations are computed over independent noise realizations, with Z_{sig}, the noiseless view matrices V_s, and the noise covariance matrices fixed throughout each experiment.

Experiment 1 varies the common entity dimension

$$m_s = m \in \{50, 100, 200, 400, 800\}$$

under uniform view weights. For each value of m and each noise regime, we compute the resulting subspace error over 20 runs. The isotropic-noise curve follows the predicted $m^{-1/2}$ rate almost exactly: the mean subspace error decreases from 0.0141 at $m = 50$ to 0.0034 at $m = 800$, with log–log slope -0.506 and $R^2 = 0.999$. The anisotropic-noise curve is higher for small and moderate m; thus, in the linear-kernel setting, anisotropy increases the finite-sample fluctuation constant instead of creating a persistent population eigenspace bias. In the long run, the decay of the rate remains of order $m^{-1/2}$, with constants depending on the spectral size and effective rank of the noise covariance (Theorem 1 and Appendix B.4). Consistently, the mean error decreases from 0.0318 to 0.0037 and the anisotropic curve approaches the isotropic curve as m grows.

Experiment 2 fixes heterogeneous entity dimensions $(m_1, m_2, m_3, m_4) = (800, 400, 200, 100)$ under the anisotropic noise setting and compares uniform weights, inverse-variance weights, and a simulation oracle. The inverse-variance weighting rule is a simulation-level implementation of the theory-guided inverse-variance weighting w^{IV}. It assigns

$$w_s \propto \frac{\widehat{\gamma}_s}{\widehat{v}_s^2 + 10^{-9}}.$$

We use the variance proxy

$$\widehat{v}_s^2 = \widehat{\sigma}_{\infty,s}^2 \frac{\widehat{B}_{s,N}}{m_s} + \frac{\widehat{\sigma}_{\infty,s}^4}{m_s},$$

where every quantity is computed from the observed matrix E_s only. Writing R_s for the residual of E_s after its rank-d SVD,

$$\widehat{\sigma}_{\infty,s}^2 = \frac{m_s \|R_s\|_{\mathrm{op}}^2}{(\sqrt{m_s} + \sqrt{N-d})^2}, \qquad \widehat{B}_{s,N} = N \max_{j,j'} |\widehat{K}_s^{\mathrm{sig}}(j, j')|,$$

where $\widehat{K}_s^{\mathrm{sig}}$ is the top-d part of $\frac{1}{N} E_s E_s^\top$ with its eigenvalues shifted down by $\|R_s\|_F^2/(N(N-d))$. The eigengap proxy $\widehat{\gamma}_s$ is the d-th eigengap of the observed kernel $\frac{1}{N} E_s E_s^\top$.

For reference, we also report a simulation oracle, a grid-search diagnostic over a finite candidate set of simplex weights (this is not guaranteed to be the exact optimizer over the continuous simplex). We use a dense candidate grid only to provide a near-oracle benchmark for comparing weighting rules; it is not a feasible procedure in real data because it uses the known Z_{sig}.

D.2 End-to-end supervised risk

Experiment 3 uses pretrained embeddings and learns bounded relation-wise ReLU networks on a nonlinear triple-score teacher with relation-specific bilinear interactions and Gaussian label noise $N(0, 0.3^2)$. We vary the labeled sample size n, pretraining sample size m, and the number of ReLU parameters per relation head W; the results are shown in Figure 2. The oracle baseline uses the true latent entity coordinates from the simulator while keeping the same supervised training procedure.

The pretraining model in Experiment 3 uses the same linear-kernel multi-view factor construction as in Experiments 1–2, with $N = 60$ entities, latent rank $d = 4$, and $S = 4$ views. The view-specific spectra Λ_s are the same as above. In the reported simulations, the pretraining samples are generated under the isotropic-noise regime with $\sigma = 2.5$, using uniform view weights. For each pretraining run, the embedding is obtained by forming the aggregated linear kernel from the noisy entity matrices and taking its leading d-dimensional eigenspace; all $N = 60$ rows, rescaled by $\sqrt{N}$, are then used as frozen entity representations for the downstream triple-score task.

We use $K_{\mathrm{rel}} = 3$ relation types. The true triple score is generated by a smooth one-hidden-layer tanh teacher with relation-specific bilinear terms, standardized to unit variance:

$$f^\star(h, r, t) = u^\top \tanh\left(A_{\mathrm{teach}} \begin{bmatrix} z_h \\ z_t \\ e_r \end{bmatrix} + b \right) + \langle \eta_r, z_h \odot z_t \rangle + c_r,$$

where e_r is the one-hot encoding of relation r, and $\eta_r \in \mathbb{R}^d$ and $c_r \in \mathbb{R}$ are relation-specific parameters. Training labels are noisy observations,

$$Y_i = f^\star(h_i, r_i, t_i) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, 0.3^2).$$

Test MSE is evaluated against the noiseless teacher score $f^\star(h, r, t)$ on all held-out triples, i.e., the triples among the $N^2 K_{\mathrm{rel}}$ that were not drawn into the labeled sample. Thus the reported risk measures estimation, approximation, and representation errors relative to the underlying true score function.

We compare two representations. The oracle representation uses the true latent entity coordinates $Z_{\mathrm{oracle}} = Z_{\mathrm{sig}}$ and trains the same relation-wise ReLU heads as all other methods, thereby removing pretraining error and isolating the supervised approximation and estimation components. The multi-view KPCA representation uses frozen coordinates estimated from the multi-view pretraining samples. The gap between multi-view KPCA and the oracle measures the downstream cost of imperfect representation recovery.

Each relation head is a one-hidden-layer ReLU network; a head with q hidden units has $W = q(2d + 2) + 1$ parameters. The labeled-sample sweep fixes $m = 40$ and $W = 81$ and varies

$$n \in \{200, 400, 800, 1600, 3200, 6400\}.$$

The pretraining-sample-size sweep fixes $n = 3000$ and $W = 161$ and varies

$$m \in \{5, 10, 20, 40, 80, 160, 400, 800\}.$$

The parameter sweep fixes $n = 900$ and $m = 40$ and varies

$$W \in \{11, 21, 41, 81, 121, 161, 201, 241\}.$$

For each configuration we average over 20 random seeds; each seed redraws the pretraining data, the labeled triples, the label noise, and the head initialization, and error bars show ± 1 standard deviation.

As the number of labeled triples increases, test MSE decreases consistently with the supervised-estimation term (Figure 2a). Multi-view KPCA remains at or above the oracle, with the gap representing the downstream cost of imperfect coordinate recovery; this gap shrinks as the pretraining sample grows (Figure 2b). The parameter sweep reveals a bias–variance trade-off in the relation heads: larger networks improve expressiveness but raise estimation variance when labeled data are limited (Figure 2c). All three trends match the risk decomposition in Theorem 2, where supervised sample complexity, approximation capacity, and pretraining error jointly determine the final risk.

E Real-Data Experimental Details

This section gives the full real-data protocol for WordNet and RENKI-KG: evaluation goals, dataset construction, side-information views and multi-view initialization, the encoder-induced kernels, split rules, negative sampling, relation-balanced objectives and metrics, and implementation details, followed by additional results.

E.1 Evaluation goals

The real-data experiments test the empirical implications of the two-stage framework. First, if entity-side information is aligned with the latent KG geometry, then pretrained entity coordinates should be much stronger than random coordinates for downstream relation-wise prediction. Second, when multiple side-information views have heterogeneous quality, multi-view aggregation can improve over a single view, and validation-selected weights can outperform uniform weighting. We use two graphs with textual side information: WordNet, a lexical-semantic graph whose synset definitions provide clean text, and RENKI-KG, a biomedical graph with heterogeneous entity types and type-specific relations, where encoders pretrained on different corpora may carry complementary information.

E.2 Datasets and preprocessing

WordNet. We preprocess WordNet and evaluate it as a synset-level knowledge graph. Starting from the WordNet pointer-relation data, we treat each retained synset as an entity and each observed pointer

relation between synsets as a positive KG triple. The preprocessing consists of: (i) canonicalizing synset identifiers so that all triples refer to synset-level nodes; (ii) retaining semantic and lexical pointer relations as relation types; (iii) removing triples whose head or tail synset lacks the textual information needed for side-information encoding; and (iv) constructing a text description for each retained synset from its synset type, representative lemma, and gloss/definition. Each synset carries one of five part-of-speech types (noun, verb, adjective, adjective satellite, adverb), which we use for type-matched negative sampling. After preprocessing, the graph contains 116,566 entities, 33 relation types, and 373,330 observed positive triples. Table 1 gives the split statistics and relation-family summary.

Table 1: WordNet statistics after preprocessing and a $60\%/20\%/20\%$ random edge split. Nodes are WordNet synsets and edges are semantic or lexical relations derived from WordNet pointers. The split sizes are identical across seeds; the assignment of triples is redrawn for each seed.

Quantity	Count	Description
Entities	116,566	Total number of retained synset nodes
Entity types	5	Noun, verb, adjective, adjective-satellite, and adverb synsets
Relation types	33	Number of retained base relation types
Observed positive triples	373,330	Total number of synset-level positive triples before splitting
Train positive triples	223,998	Positive triples assigned to training
Validation positive triples	74,666	Positive triples assigned to validation
Test positive triples	74,666	Positive triples assigned to testing
Node text	–	Synset type, representative lemma, and gloss/definition
Major relation families	–	Hypernymy/hyponymy, meronymy/holonymy, antonymy, derivational relations, domain relations, and other semantic or lexical WordNet pointer types
Most frequent relations	–	`hypernym`, `hyponym`, `lexical_derivationally_related_form`, `similar_to`, `member_holonym`, `member_meronym`, `part_holonym`, `part_meronym`, `instance_hypernym`, `instance_hyponym`

RENKI-KG. RENKI-KG is the medical KG constructed by Liu et al. [2024]. It links clinical concepts, diagnoses, and medications drawn from multiple curated medical ontologies and drug databases, and it covers both narrative and codified concepts used in electronic health records. Entities are UMLS clinical concepts, phecodes, and RxNorm medications, and each of the 9 relation types connects a fixed pair of entity types (e.g., concept–phecode, medication–phecode, phecode–phecode). The side information of each entity is its name string (e.g., "phenobarbital", "foreign body injury"). Table 2 summarizes the graph; we refer to Liu et al. [2024] for the full construction.

Table 2: RENKI-KG statistics, following Liu et al. [2024].

Quantity	Count
Entities	118,554
Clinical concepts (UMLS CUIs)	113,787
Diagnosis codes (phecodes)	1,846
Medications (RxNorm)	2,921
Relation types	9
Observed positive triples	698,368

E.3 Side-information views and multi-view initialization

Encoders. On WordNet, each synset description is encoded by four pretrained language models without task-specific fine-tuning: BERT-base (uncased), RoBERTa-base, ELECTRA-base (discriminator), and DeBERTa-v3-base. Each view is the mean-pooled last-layer hidden state (inputs truncated

to 32 tokens), with rows ℓ_2-normalized. On RENKI-KG, each entity name is encoded by seven BERT-family encoders pretrained or fine-tuned on different corpora: BERT, RoBERTa, NER-BERT, MentalBERT, SciBERT, BioBERT, and ClinicalBERT. Each view is the last-layer `[CLS]` hidden state, with rows ℓ_2-normalized. Each encoder induces one view as in Appendix E.4.

Multi-view KPCA. Let $\widetilde{E}_s \in \mathbb{R}^{N \times m'_s}$ be the column-centered encoder feature matrix of view s, with rows $(e_{s,j} - \bar{e}_s)^\top$ as in Appendix E.4. Given weights $w \in \Delta_S$, the shared coordinates are the top-d eigenvectors of $\sum_{s=1}^{S}(w_s/m'_s)\widetilde{E}_s\widetilde{E}_s^\top$; we use $d = 128$ on both graphs. A single-view representation is the same construction with one view, i.e., its top-d principal components. On RENKI-KG, the orthonormal coordinates are rescaled to the mean row norm of the raw BERT embedding, the input scale of the RENKI head network.

View weights. On WordNet, the weights are selected on validation AUROC from the 35 points of the simplex grid with step 0.25, averaging two seeds (101 and 103) that are disjoint from the evaluation seeds. The selected weights are $(1 : 2 : 1 : 0)$ over (BERT, RoBERTa, ELECTRA, DeBERTa-v3). On RENKI-KG, the seven views are aggregated with uniform weights, without weight search.

E.4 Encoder-induced kernels and coordinates

Each language model is regarded as a kernel and introduces a view-specific feature map, denoted as

$$\phi_s : \mathcal{U}_s \to \mathbb{R}^{m'_s}.$$

Let $u_{s,j} \in \mathcal{U}_s$ be the textual description of entity j. The encoder induces the entity kernel

$$k_s(u_{s,j}, u_{s,j'}) := \langle \phi_s(u_{s,j}), \phi_s(u_{s,j'}) \rangle_{\mathbb{R}^{m'_s}}.$$

Writing $e_{s,j} := \phi_s(u_{s,j})$, the corresponding kernel matrix is the centered Gram matrix of the explicit encoder features (centering, which Algorithm 1 omits, removes the common mean direction of the encoder outputs):

$$\widehat{K}_s(j, j') = \frac{1}{N}\langle e_{s,j} - \bar{e}_s, e_{s,j'} - \bar{e}_s \rangle, \qquad \bar{e}_s := \frac{1}{N}\sum_{j=1}^{N} e_{s,j}.$$

Accordingly, the encoder output is first interpreted as a feature map, and the kernel matrix is reconstructed from pairwise similarities between the mapped entity features. This places the language-model views directly within the multi-view entity-kernel framework developed above.

We aggregate the induced kernels as

$$\widehat{K}_{\mathrm{mv}}(w) := \sum_{s=1}^{S} w_s \widehat{K}_s, \qquad w_s \geq 0, \qquad \sum_{s=1}^{S} w_s = 1. \tag{E.1}$$

We use the leading $d = 128$ eigenvectors of the aggregated entity-kernel matrix as the pretrained entity coordinates:

$$\widehat{K}_{\mathrm{mv}}(w)\widehat{\psi}_\ell = \widehat{\lambda}_\ell \widehat{\psi}_\ell, \qquad \widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots, \qquad \widehat{Z} = [\widehat{\psi}_1, \ldots, \widehat{\psi}_{128}] \in \mathbb{R}^{N \times 128}. \tag{E.2}$$

E.5 Evaluation protocol

Edge splits. On WordNet, the observed positive triples are randomly permuted and split $60\%/20\%/20\%$ into train, validation, and test subsets. On RENKI-KG, each triple is independently assigned to train, validation, or test with probabilities $0.7/0.1/0.2$. In both cases the split is redrawn for each seed, and all reported results are overall test performance under these splits.

Negative sampling. Both graphs contain observed positive triples only. We construct a balanced binary link-prediction task by pairing each positive triple with one corrupted negative triple. For a positive triple (h, r, t), the negative triple replaces the head or the tail (chosen with equal probability) while keeping the relation fixed. The replacement entity is sampled uniformly from the entity type of the corrupted endpoint: the synset type on WordNet, and the type required by relation r on RENKI-KG. Type-matched corruption prevents the task from being solved through trivial type mismatch. Training negatives are resampled for every mini-batch and are not filtered against observed triples.

Detail	WordNet	RENKI-KG
Optimizer (PNKG)	AdamW	AdamW
Learning rate	2×10^{-3}	2×10^{-3}
Weight decay	0	0
Downstream KG batch size	1024	1024
Epochs	50	100
Coordinate dimension d	128	128
Relation head	$2d \to 256 \to 128 \to 1$, dropout 0.1	$2d \to 256 \to 1$
Encoder pooling	Mean, max. 32 tokens	[CLS], no truncation
LLM embedding batch size	32	1
Data-loader workers	2	– (data on GPU)
Train/validation/test split	Random edge, 60/20/20	Random edge, 70/10/20
Number of random seeds	10	10
GPU resources	NVIDIA GeForce RTX 5080 (16,303 MiB) and RTX 5070 (12,227 MiB)	
Driver/CUDA versions	NVIDIA driver 615.71.09; CUDA 13.4	
Approximate execution time	About 10 minutes per seed (four concurrent runs on two GPUs)	About 2 minutes per seed (one run per GPU)

Table 3: Experimental details for WordNet and RENKI-KG. KGE baselines use model-specific optimizers (Adam for TransE, DistMult, and RotatE; Adagrad for ComplEx) and otherwise share the settings above.

Relation-balanced objective and metrics. Both graphs have skewed relation frequencies, so a uniformly averaged loss can be dominated by frequent relation types. We therefore use relation-balanced weighting during training and evaluation. Let $\mathcal{D}_r$ denote the examples with relation r in the relevant split, and let $n_r = |\mathcal{D}_r|$. We assign relation weights

$$\omega_r = \frac{n_r^{-1}}{\sum_{r'} n_{r'}^{-1}}, \tag{E.3}$$

and optimize the weighted empirical objective

$$\widehat{L}(\theta) = \sum_{(x_i, y_i) \in \mathcal{D}_{\text{train}}} \omega_{r_i} \, \ell\{f_\theta(x_i), y_i\}, \tag{E.4}$$

where $x_i = (h_i, r_i, t_i)$ and ℓ is the squared error between the sigmoid score and the binary label. The same objective is used for PNKG and the KGE baselines. Relation weights computed on the validation and test splits are used to aggregate the corresponding metrics. We report relation-balanced AUROC, accuracy, and F1, with a positive prediction defined as a sigmoid score of at least 0.5. Results are averaged over 10 random seeds and shown as mean $\pm$ standard deviation.

Validation usage. The validation split is used only for model selection and is never included in training or final test evaluation. For each model, the reported checkpoint is the epoch with the best validation AUC: relation-balanced AUROC on WordNet, and, following Liu et al. [2024], the average of per-relation AUCs weighted by the number of validation triples on RENKI-KG. On WordNet, the validation set is also used to select the view weights (Appendix E.3). Once the checkpoint and view weights are selected, all reported results are evaluated on the held-out test split.

Models. In all text-derived PNKG variants, entity coordinates are frozen and only the relation-specific heads are trained. On WordNet, each relation head is an MLP $2d \to 256 \to 128 \to 1$ with ReLU activations and dropout 0.1 applied to the concatenated head and tail coordinates; the random variant trains its randomly initialized entity embeddings jointly with the head. On RENKI-KG, we use the RENKI head of Liu et al. [2024], an MLP $2d \to 256 \to 1$ with ReLU activation per relation; the random variant trains its randomly initialized entity embeddings jointly with the heads. The KGE baselines (TransE, DistMult, ComplEx, RotatE) use dimension 128 and are trained end to end with the same splits, negatives, objective, and checkpoint selection. Optimization and hardware details are given in Table 3.

E.6 Full results on WordNet and RENKI-KG

Tables 4 and 5 report all methods on WordNet and RENKI-KG, including each single view, uniform multi-view KPCA, and, on WordNet, multi-view KPCA with validation-selected weights. On both graphs the single views differ markedly in quality: test AUROC ranges from 0.8009 (DeBERTa-v3) to 0.9441 (BERT) on WordNet and from 0.9172 (NER-BERT) to 0.9607 (SciBERT) on RENKI-KG. Except for DeBERTa-v3 on WordNet, every text-derived view outperforms random coordinates. On WordNet, the other three views also outperform every graph-only baseline on all three metrics. On RENKI-KG, every view outperforms every graph-only baseline in accuracy and F1; in AUROC, SciBERT, BioBERT, and ClinicalBERT exceed every baseline, whereas BERT, RoBERTa, NER-BERT, and MentalBERT fall below RotatE. On WordNet, uniform weighting (0.9388 AUROC) falls slightly below the two strongest single views because it assigns weight to the weak DeBERTa-v3 view. The validation-selected weights $(1 : 2 : 1 : 0)$, which exclude that view, give the best result on all three metrics. On RENKI-KG, where the strongest views are close in quality, uniform multi-view KPCA alone is best on all three metrics.

Table 4: Full WordNet results (random edge split, relation-balanced test metrics). Text-derived `PNKG` variants use fixed entity coordinates; the random variant trains its entity embeddings. Mean $\pm$ standard deviation over 10 seeds. Best results are in bold.

	Method	AUROC	Accuracy	F1
KGE	TransE	0.8169 ± 0.0221	0.7616 ± 0.0086	0.6879 ± 0.0147
	DistMult	0.7274 ± 0.0134	0.6275 ± 0.0136	0.6742 ± 0.0128
	ComplEx	0.6985 ± 0.0142	0.6005 ± 0.0161	0.6709 ± 0.0099
	RotatE	0.5100 ± 0.0155	0.5622 ± 0.0152	0.6072 ± 0.0147
`PNKG`	Random	0.8460 ± 0.0065	0.7517 ± 0.0085	0.6900 ± 0.0178
	BERT	0.9441 ± 0.0064	0.8553 ± 0.0114	0.8451 ± 0.0137
	RoBERTa	0.9421 ± 0.0054	0.8538 ± 0.0112	0.8423 ± 0.0137
	ELECTRA	0.9195 ± 0.0067	0.8252 ± 0.0121	0.8126 ± 0.0136
	DeBERTa-v3	0.8009 ± 0.0074	0.7119 ± 0.0075	0.6859 ± 0.0134
	Multi-view (uniform)	0.9388 ± 0.0057	0.8455 ± 0.0103	0.8336 ± 0.0135
	Multi-view (selected)	$\mathbf{0.9469} \pm 0.0071$	$\mathbf{0.8597} \pm 0.0127$	$\mathbf{0.8496} \pm 0.0144$

Table 5: Full RENKI-KG results (random edge split, relation-balanced test metrics). Text-derived `PNKG` variants use fixed entity coordinates; the random variant trains its entity embeddings. Mean $\pm$ standard deviation over 10 seeds. Best results are in bold.

	Method	AUROC	Accuracy	F1
KGE	TransE	0.9352 ± 0.0016	0.8333 ± 0.0054	0.8102 ± 0.0077
	DistMult	0.8986 ± 0.0012	0.7800 ± 0.0026	0.8087 ± 0.0018
	ComplEx	0.7699 ± 0.0037	0.6944 ± 0.0026	0.7346 ± 0.0023
	RotatE	0.9382 ± 0.0018	0.7993 ± 0.0032	0.7568 ± 0.0050
`PNKG`	Random	0.8105 ± 0.0035	0.6314 ± 0.0030	0.4397 ± 0.0071
	BERT	0.9326 ± 0.0025	0.8492 ± 0.0052	0.8345 ± 0.0080
	RoBERTa	0.9364 ± 0.0020	0.8541 ± 0.0045	0.8395 ± 0.0068
	NER-BERT	0.9172 ± 0.0029	0.8364 ± 0.0039	0.8186 ± 0.0056
	MentalBERT	0.9225 ± 0.0026	0.8370 ± 0.0035	0.8185 ± 0.0053
	SciBERT	0.9607 ± 0.0016	0.8871 ± 0.0026	0.8794 ± 0.0033
	BioBERT	0.9556 ± 0.0017	0.8775 ± 0.0027	0.8681 ± 0.0039
	ClinicalBERT	0.9513 ± 0.0019	0.8730 ± 0.0048	0.8631 ± 0.0067
	Multi-view (uniform)	$\mathbf{0.9624} \pm 0.0007$	$\mathbf{0.8877} \pm 0.0046$	$\mathbf{0.8804} \pm 0.0062$

F Disclosure of AI Assistance

We used AI tools to assist with language editing, formatting, and code cleanup, and to check technical statements. These tools were not used to develop the theoretical results or their proofs, or to generate the original scientific content. The authors reviewed all AI-assisted edits and take full responsibility for the content of the paper.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state three contributions—multi-view KPCA subspace recovery theory, an end-to-end risk theorem with frozen pretrained coordinates, and experiments on WordNet and RENKI-KG—each of which is delivered in the corresponding sections of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 6 discusses limitations including frozen coordinates, shared-subspace misspecification, i.i.d. triple sampling, and unknown view-weighting quantities.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The full pretraining assumptions are stated in Appendix B, and supervised assumptions in Appendix B.2. Complete proofs of Theorems 1 and 2 are given in Appendix C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Appendix E specifies dataset preprocessing, splits, negative sampling, relation-balanced evaluation, seeds, and training settings. Simulation parameters and protocols are detailed in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Code is available at https://github.com/jifanzhang999/PNKG, as linked in the introduction. WordNet is publicly available, RENKI-KG is described in Liu et al. [2024] and accessible under the licenses of its source ontologies and drug databases, and the pretrained encoders are available via HuggingFace; the experimental protocols are described in Appendices D and E.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The experimental details are described in Section 5.2 and Appendices D and E, including data generation, splits, negative sampling, and relation balancing. Table 3 reports optimizer, learning rate, batch sizes, epochs, and real-data compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Real-data results are reported as mean $\pm$ standard deviation over 10 seeds, and Experiment 3 uses 20 seeds. Experiment 1 uses 20 independent noise realizations; the sources of variability and simulation protocols are described in Appendix D.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Table 3 in Appendix E reports GPU type, memory, and approximate per-seed runtimes for the real-data experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses WordNet, which is publicly available, RENKI-KG, which is built from public ontologies and drug databases and used under their licenses, and standard pretrained models; it involves no human subjects, no privacy-sensitive data collection, and poses no deception or dual-use concerns.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper does not include a broader impact statement. The work is foundational statistical theory for knowledge graph learning with no direct path to harmful applications. Potential positive societal impact includes improving biomedical knowledge discovery (e.g., drug indication and side-effect prediction via RENKI-KG), but this is not explicitly discussed.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release new models or datasets. The pretrained language models used as entity feature extractors (e.g., BERT, RoBERTa, BioBERT, ClinicalBERT) are standard publicly available models with no higher misuse risk than their existing releases.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All baseline methods (TransE, DistMult, ComplEx, RotatE), pretrained language models (BERT, RoBERTa, ELECTRA, and DeBERTa-v3 for WordNet; BERT, RoBERTa, NER-BERT, MentalBERT, SciBERT, BioBERT, and ClinicalBERT for RENKI-KG), and datasets (WordNet, RENKI-KG) are cited via their original papers in the related-work and experimental sections. WordNet is available under the Princeton WordNet License; RENKI-KG is introduced by Liu et al. and built from cited public ontologies and drug databases.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: The paper does not introduce new datasets or models.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper involves no crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects are involved in this research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Pretrained language models are used as off-the-shelf entity feature extractors; their role is described in Section 5.2 and Appendix E.4. Separately, Appendix F states that AI tools assisted with language editing, formatting, code cleanup, and checking technical statements, and were not used to develop the theoretical results or their proofs or to generate the original scientific content.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.