Order-Preserving Dimension Reduction for Multimodal Semantic Embedding

Chengyu Gong[1*†], **Gefei Shen**[2*†], **Luanzheng Guo**[3], **Nathan R. Tallent**[3], **Dongfang Zhao**[4‡]

[1]New York University
[2]Harvard University
[3]Pacific Northwest National Laboratory
[4]University of Washington Tacoma School of Engineering & Technology
cg4761@nyu.edu, gshen@g.harvard.edu, {lenny.guo, nathan.tallent}@pnnl.gov, dzhao@uw.edu

Abstract

Searching for the k-nearest neighbors in multimodal data retrieval is computationally expensive, particularly due to the inherent difficulty in comparing similarity measures across different modalities. Recent advances in multimodal machine learning address this issue by mapping data into a shared embedding space; however, the high dimensionality of these embeddings (hundreds to thousands of dimensions) presents a challenge for time-sensitive vision applications. This work proposes Order-Preserving Dimension Reduction (OPDR), aiming to reduce the dimensionality of embeddings while preserving the ranking of KNN in the lower-dimensional space. One notable component of OPDR is a new measure function to quantify KNN quality as a global metric, based on which we derive a closed-form map between target dimensionality and key contextual parameters. We have integrated OPDR with multiple state-of-the-art dimension-reduction techniques, distance functions, and embedding models; experiments on a variety of multimodal datasets demonstrate that OPDR effectively retains recall high accuracy while significantly reducing computational costs.

Introduction

Background and Motivation

Multimodal retrieval tasks, especially those involving heterogeneous modalities such as text, image (Zhou and Fan 2015), and audio, rely on diverse data sources where conventional indexing techniques such as DBSCAN or k-means become ineffective. Many real-world multimodal systems, such as those in material science (Rangel DaCosta et al. 2021), often consist of multiple images and associated textual descriptions, stored in databases that manage structured (tables) and unstructured (blobs) data separately. Traditional indexing approaches fail to capture the unified semantic similarity between different modalities, requiring an alternative representation that aligns heterogeneous data within a common space.

Multimodal machine learning (Li et al. 2022) provides a means to embed heterogeneous data into a shared vector space (Jia et al. 2021), enabling unified similarity search. However, embeddings from different modalities, such as text and images, are often concatenated, leading to an even higher-dimensional joint representation, which further exacerbates the curse of dimensionality (Aremu, Hyland-Wood, and McAree 2020). Given that embeddings generated by models such as BERT (Devlin et al. 2018) and ViT (Dosovitskiy et al. 2020) already contain hundreds to thousands of dimensions, concatenation leads to prohibitively high-dimensional representations. This dimensionality explosion results in increased storage requirements, computational inefficiency, and degraded nearest-neighbor search performance, particularly in large-scale scientific applications (Wu et al. 2020; Tshitoyan et al. 2019).

Dimensionality reduction offers a potential solution by projecting high-dimensional embeddings into a lower-dimensional space while preserving key relationships. Techniques such as PCA (Hotelling 1933) and ISOMAP (Tenenbaum, de Silva, and Langford 2000) aim to minimize the distortion in inter-point distances, facilitating efficient retrieval in reduced spaces. However, a fundamental challenge in applying dimensionality reduction to multimodal retrieval is the lack of a principled method to determine the optimal target dimensionality. Without an a prior understanding of how well nearest-neighbor structures are preserved post-reduction, practitioners must rely on heuristic or empirical tuning, limiting the robustness of these methods in real-world applications (Zhao et al. 2023). This uncertainty is particularly problematic in tasks such as cross-modal retrieval (Liang et al. 2024), Retrieval-Augmented Generation (Lewis et al. 2020), and scientific knowledge discovery (Gupta et al. 2022; Lu et al. 2023), where the integrity of neighborhood relationships impacts retrieval effectiveness.

These challenges underscore the necessity of understanding the trade-off between dimensionality reduction and nearest-neighbor preservation. A framework that enables quantitative estimation of neighbor preservation after dimensionality reduction would allow for more effective multimodal retrieval by mitigating the curse of dimensionality while maintaining computational efficiency.

[*]These authors contributed equally.
[†]Work done at University of Washington.
[‡]Corresponding author.

Proposed Work

To effectively reduce the dimensionality of high-dimensional embedding vectors while preserving the neighborhood structure, we introduce a mathematically rigorous measure function that quantifies the preservation of nearest neighbors across different metric spaces. This measure, formally defined over the power-set σ-algebra of the lower-dimensional space, quantifies the degree to which k-nearest neighbors are preserved. By enabling a well-defined aggregation over the entire space, this measure provides a concrete way to evaluate the fidelity of dimension-reduction transformations.

Building on this measure, we analyze the quantitative relationship between neighborhood preservation and target dimensionality. We construct a closed-form function that captures the intrinsic dependencies among key parameters, including the original and reduced dimensionality, the number of data points, and the expected neighborhood preservation accuracy. Our analysis reveals that the required lower-dimensional space is positively correlated with both the original dimensionality and dataset cardinality, but their influence varies. In particular, neighborhood preservation accuracy exerts an exponentially stronger effect than dataset size, indicating its dominant role in determining the optimal reduced dimensionality.

In addition to theoretical results, we demonstrate how the proposed measure and closed-form function integrate into practical vision applications. We detail their incorporation into dimensionality reduction pipelines using PCA (Hotelling 1933) and MDS (Kruskal and Wish 1978), evaluate their compatibility with state-of-the-art embedding models such as CLIP (Radford et al. 2021), ViT (Dosovitskiy et al. 2020), and BERT (Devlin et al. 2018), and examine their performance under different distance metrics, including Euclidean, cosine, and Manhattan. Extensive experiments on multimodal scientific datasets and image-text benchmarks validate the accuracy of the closed-form function in capturing the relationship between dimensionality reduction and neighborhood preservation. The results highlight the effectiveness of this framework in mitigating the curse of dimensionality while ensuring robust retrieval performance in vision-based applications.

Contributions This work makes the following key contributions:

- We introduce a mathematically measure function to quantify the preservation of k-nearest neighbors across different dimensional spaces, providing a formal foundation for evaluating dimensionality reduction methods.
- We derive a closed-form function that describes the relationship between neighborhood preservation and target dimensionality, offering a principled approach to determining the optimal reduced dimensionality.
- We empirically validate the proposed framework across multiple datasets and dimensionality reduction techniques, demonstrating that the closed-form function accurately predicts neighborhood preservation and effectively mitigates the curse of dimensionality.

Background and Related Work

Multimodal Data Analytics

Multimodal data analytics has attracted growing research interest across scientific domains. For example, the MELINDA dataset (Wu et al. 2020) focuses on classifying biomedical experiment methods by utilizing multiple data types. Similarly, a study on explainable AI (Jin, Li, and Hamarneh 2022) evaluates algorithms within the context of multimodal medical imaging tasks, emphasizing the importance of decision-making based on raw scientific data. In parallel, researchers have developed new large language models specifically designed for scientific data. For example, GIT-Mol (Liu et al. 2024) represents a notable effort, developing a multimodal large language model for molecular science that integrates graph, image, and text data. Additionally, MatSciBERT (Gupta et al. 2022) focuses on creating domain-specific language models tailored to materials science. While these models, along with others such as the unsupervised word embeddings for materials science literature (Tshitoyan et al. 2019), demonstrate strong potential in capturing domain-specific knowledge, they are often limited in scope, typically focusing solely on text-based data.

Multimodal data is often highly dimensional, making it crucial to find a trade-off between maintaining essential information and reducing the data's complexity. Given the high dimensionality of multimodal data, researchers have proposed methods to balance information preservation and complexity reduction. For instance, the study on Multiscale Feature Extraction and Fusion of Image and Text in VQA (Lu et al. 2023) introduces advanced techniques to fuse multimodal data, allowing for a more integrated analysis of different data types. The research on Semi-Supervised Multimodal Learning with Balanced Spectral Decomposition (Hu et al. 2020) focuses on optimizing information preservation through spectral decomposition, which enhances the ability to capture correlations within the data. Additionally, Unsupervised word embeddings in materials science (Tshitoyan et al. 2019) demonstrates the transformation of textual information into vectors, providing a compact and efficient representation of large-scale text data.

Contrastive Language-Image Pre-Training

With the introduction of "Attention is All You Need" (Vaswani et al. 2017), the development of large language models has surged, leading to significant advancements in the field. Following this breakthrough, models like BERT (Devlin et al. 2018) and ViT (Dosovitskiy et al. 2020) have focused on text analysis and image recognition respectively. Particularly noteworthy is the CLIP model (Radford et al. 2021), which integrates both text and image data, marking a significant step forward in multimodal data processing. Various improvements to the CLIP model have been proposed. For instance, TiMix (Jiang et al. 2024), address issues with noisy web-harvested text-image pairs by using mixed data samples, while CLIP-Event (Li et al. 2022) and SoftCLIP (Gao et al. 2024) further enhance the alignment between text and images.

In (Zhang et al. 2021), authors proposed to convert images

into binary codes (hash codes) that preserve image similarity, making it easier to compare images within large-scale datasets.

Some prior works proposed methods to perform dimension reduction over the embedding vectors. For example, in (Chen et al. 2022), authors proposed a new approach to reducing dimensionality following contrastive learning. Similarly, in (Huang et al. 2020), authors focused on relative positioning in data before and after LLM manipulation. However, none of the above works touched on the preservation of the set of k-nearest neighbors.

Dimension Reduction

After Multidimensional Scaling (MDS) (Torgerson 1952) was published, numerous methods were proposed that preserved pairwise distances in lower-dimensional spaces. For example, Colored Maximum Variance Unfolding (Song et al. 2007) preserved local distances while maximizing variance, and Neighborhood Preserving Embedding (He et al. 2005) retained local relative positions. Further innovations include Tensor Embedding Methods (Dai and Yeung 2006), which tackled the curse of dimensionality, and MultiMAP (Jain et al. 2023), which integrated multimodal data. Recent approaches like Similarity Order Preserving Discriminant Analysis (Hu, Feng, and Chen 2021) and Ordinal Data Clustering Algorithm with Automated Distance Learning (Zhang and Cheung 2020) specifically aimed to preserve data order post-reduction, highlighting the evolving focus on maintaining data integrity across various dimensions. However, none of the above works considered preserving the *set* of k-nearest neighbors during the dimension reduction.

As large language models (LLMs) become popular, scientists have also proposed methods that integrate LLMs with dimension-reduction techniques to enhance the analysis of complex data. For example, Transformer-based Dimensionality Reduction (Ran, Gao, and Fang 2022) introduced a method that decomposes autoencoders into modular components, leveraging the power of transformers to manage high-dimensional data efficiently. Similarly, in (George and Sumathy 2023), authors combined BERT with dimensionality reduction to enhance topic modeling by clustering textual data more effectively. However, the above works assumed that the users of LLMs were able to modify the model structure, which is unrealistic in many scientific applications.

Multimodal Embedding and Retrieval across Heterogeneous Data

Recent progress in multimodal representation learning has significantly advanced cross-modal retrieval tasks, which aim to identify semantically aligned items across heterogeneous modalities such as text, images, and audio. Traditional vector retrieval methods, including FAISS(Douze et al. 2024), ScaNN(Guo et al. 2020), and HNSW (Malkov and Yashunin 2018), enable efficient similarity search for high-dimensional embeddings. However, as embedding dimensionality increases, these methods face scalability challenges, leading to degraded retrieval efficiency and increased computational cost. While existing approaches focus on scaling similarity search or enhancing retrieval precision, few address the structural integrity of nearest-neighbor relations after dimensionality reduction. To address this limitation, our work introduces Order-Preserving Dimension Reduction (OPDR), which reduces embedding dimensionality while preserving k-nearest neighbor (kNN) structure, thereby improving retrieval efficiency in large-scale multimodal datasets.

Order-Preserving Dimension Reduction

This section first presents a new notion, i.e., Order-Preserving Measure (OPM), for the *set* of k-nearest neighbors that are preserved during a map between two metric spaces. Then, a closed-form function is constructed to quantify the relationship among the space dimensionality, the number of data points, and the OPM. Finally, a detailed implementation is presented about how to incorporate the new measure and function real-world multimodal retrieval systems.

Order-Preserving Measure

Informally, an OPM provides a metric of the number of nearest neighbors that do not change between two spaces. For example, if the 2-closest points (we assume that the spaces are metric spaces such that pair-wise distances are well defined) of $every$ point in a metric space (X, δ_X) are still the same 2-closest points (of each point) in a metric space (Y, δ_Y), then we say the map $f : X \to Y$ is order-preserving of 2, or OP_2. We will provide a more formal definition of OP_z, $z \in \mathbb{Z}^+ \cup \{0\}$ later; before that, we need to point out a common misunderstanding of this notion regarding inclusiveness, as follows.

The order-preserving notion defined above is *not* inclusive in the sense that in general, $OP_{k+1} \not\Rightarrow OP_k$, $1 \leq k \in \mathbb{Z}$. This can be understood as the *set* of top k nearest neighbors in the space Y do not necessarily respect the intrinsic order of elements in the *set* of the top k nearest neighbors in space X. The rationale behind this is that the result of a k-nearest neighbor (KNN) query in many cases will be the input of other analytical steps that are agnostic of the "internal" order of the set of points. This implies that, for example, an OP_2 map is not necessarily OP_1: A sorted list of points in space Y, $L_Y = (b, a, c)$ is clearly OP_2 if the original sorted list on space X is $L_X = (a, b, c)$ because $\{b, a\} = \{a, b\}$ (even if $(b, a) \neq (a, b)$ as ordered lists); However, L_Y is *not* OP_1 regarding L_X because $\{b\} \neq \{a\}$. For completeness, we will agree that OP_0 is trivially true for any pairs of lists.

Formally, a *measure* is a function, say μ, which maps a subset E_i of set X in a well-defined σ-algebra to a value in the extended real line $[0, \infty]$, such that (i) $\mu(\emptyset) = 0$ and (ii) $\mu\left(\bigcup_{i=1}^{n} E_i\right) = \sum_{i+1}^{n} \mu(E_i)$. Because we assume there exists a well-defined σ-algebra over (the set of) embedding vectors, say X, E_i's will be understood as *disjoint* subsets of X unless otherwise stated. In fact, we will explicitly construct the σ-algebra on the target space, Y. As a side note, a measure function μ can be also thought of as a *homomorphism* between the two monoids (i.e., groups without inverses) $(\mathcal{P}(X), \cup)$ and $(\mathbb{R}^+ \cup \{0\}, +)$ with kernel

$Ker_\mu = \{\emptyset\} \subseteq \mathcal{P}(X)$, where $\mathcal{P}(X)$ denotes the power set of X. However, our analysis in the following will not require any group-theoretical results.

The construction of the σ-algebra of embedding vectors is as follows. Let k denote the target number of preserved k-nearest neighbors. Let Y denote the target (i.e., low-dimensional) metric space (we omit the metric functions here) of the dimension-reduction map. In practice, the cardinality of Y is finite (i.e., there is an upper limit of number of embedding vectors in vision applications), which is of course countable. The σ-algebra $\mathcal{M}_Y$ of Y is simply the power set of all the mapped vectors in Y, i.e., $\mathcal{M}_Y = \mathcal{P}(Y)$, which is obviously a σ-algebra on Y because Y is countable and $\mathcal{P}(Y)$ is closed for any countable number of set unions.

After constructing its σ-algebra, we are ready to define the measure on $\mathcal{M}_Y$. Let

$E^Y_{k,i}$ denote the set of the k-nearest neighbors of y_i, where $y_i \in Y$. The same notation can be defined for space X, which denotes the original metric space where the original vectors of y_i reside,

i.e., $x_i = f^{-1}(y_i)$, where f denotes the dimension-reduction function. Let μ be a function defined as follows,

$$\mu(F_j) = \frac{\left|F_j \cap E^Y_{k,i} \cap E^X_{k,i}\right|}{k}, \tag{1}$$

where $F_j \in \mathcal{M}_Y$. It should be clear that $\mu \in [0,1] \subset [0,+\infty]$ because $\left|E^Y_{k,i}\right| \in [0,k]$, where $|\cdot|$ denotes the cardinality of a set.

We argue that the function defined in Eq. (1) is a measure on $\mathcal{M}_Y$. This means that we need to demonstrate two properties of μ (on $\mathcal{M}_Y$): (i) $\mu(\emptyset) = 0$ and (ii) $\mu(\bigcup_{i=1}^{\infty} F_j) = \sum_{i=1}^{\infty}(\mu(F_j))$, $\forall F_j \subseteq Y$ and F_j's are disjoint. Property (i) is trivially satisfied because when $F_j = \emptyset$, we have $\mu(F_j) = \frac{|\emptyset|}{k} = \frac{0}{k} = 0$. To prove property (ii), it suffices to show that if $F_j = F_1 \cup F_2$ and $F_1 \cap F_2 = \emptyset$, then $\mu(F_1 \cup F_2) = \mu(F_1) + \mu(F_2)$; this is because $\mathcal{M}_Y$ is finite and the above binary relationship can be extended finite times. To simplify the notation, let $E = E^Y_{k,i} \cap E^X_{k,i}$. It follows that $\mu(F_j) = \frac{|(F_1 \cup F_2) \cap E|}{k} = \frac{|(F_1 \cap E) \cup (F_2 \cap E)|}{k}$. Now, because $F_1 \cap F_2 = \emptyset$, we know $(F_1 \cap E) \cap (F_2 \cap E) = \emptyset$. This implies that there is no common element between $(F_1 \cap E)$ and $(F_2 \cap E)$, which means the sets are separate components and the cardinality of the union set is simply the addition of cardinality of each component: $|(F_1 \cap E) \cup (F_2 \cap E)| = |F_1 \cap E| + |F_2 \cap E|$. It follows that $\mu(F_1 \cup F_2) = \frac{|F_1 \cap E| + |F_2 \cap E|}{k} = \frac{|F_1 \cap E|}{k} + \frac{|F_2 \cap E|}{k} = \mu(F_1) + \mu(F_2)$, as desired.

The discussion about preserving the set of k-nearest neighbors between metric spaces around Eq. (1) is, in our humble opinion, a tip of the iceberg regarding the *neighborhood-preserving dimension-reduction maps*. The reasoning is that the k-nearest neighbors can be thought of the discrete representation of the neighborhood of the referred point, or vector. If we extend the idea to a continuous counterpart, i.e., given a point p and its neighborhood or an *open set* around it, say $\epsilon(p)$, the question becomes under what conditions the open set of the mapped vectors,

namely $f(p)$, corresponds to the open set of p in a dimension-reduction function f. In topology and analysis, such a function carrying the open sets in the forward direction is called an *open map*, which is unfortunately not the same (in fact, the inverse) condition under which the function is continuous. We leave this as an open question to the community.

Closed-Form Function

The previous section defines an additive measure μ on the σ-algebra of the projected space Y; This section investigates the relationship between μ and other parameters, such as the space dimensionality and the number of data points, or space cardinality. The idea is to construct a function for the above variables which we hypothesize to be critical for dimensionality reduction, largely based on our empirical observations and mathematical intuition. We will later verify the constructed function as a working hypothesis in the evaluation section.

We assume that we have an existing dimension-reduction method on hand and we are interested in quantifying the target dimensionality of the lower-dimensional space in which the set of k-nearest neighbors is an *invariant*, as defined in Eq. (1). It turns out that before attempting to solve the problem, we need to have a more "global" metric than $\mu(\cdot)$, which works as a local measure that is concerned with only the k-nearest neighbors of a single point.

We define the *global* metric for quantifying the overall closeness (or, similarity, accuracy) between two metric spaces regarding their k-nearest neighbors as follows. We first aggregate the measure for each point, then normalize the aggregate measure (AM) into $[0,1]$, and finally calculate the arithmetic mean of all normalized AMs (NAMs). We call the above average of NAMs as the *accuracy* of Y with respect to X on k-nearest neighbors. Formally, we define the accuracy A as

$$\mathrm{A}^X_k(Y) = \frac{1}{m} \cdot \sum_{i=1}^{m} \frac{\mu_i\left(Y \setminus \{y_i\}\right)}{k}, \tag{2}$$

where $m = |Y| = |X|$, $y_j \in Y$, and $\mu_i(\cdot)$ is the measure on Y over y_i as defined in Eq. (1). It is not hard to see that the accuracy A defined above falls within the range $[0,1]$, because $\mu_i\left(Y \setminus \{y_i\}\right)$ is bounded by k (inclusively).

The parameters in our hypothetical function include the accuracy A_k between two metric spaces X and Y, the number of data points $m = |X| = |Y|$, and the dimensionality of the spaces. In practice, the dimension of the domain space X is determined by the neural network model; therefore, our function would assume that X's dimension is a constant and only involves the dimension of Y, denoted by $dim(Y)$. Thus, we expect to construct a function g as follows,

$$dim(Y) = g\left(A_k, m\right),$$

such that the result $dim(Y)$ can be set as a parameter in a dimension-reduction function f. It follows that the real-world users can simply compose the functions g and f, i.e., $f \circ g$, to ensure that the set of k-nearest neighbors is an invariant between two spaces X and Y.

The construction of g is based on the following observations:

- Firstly, $dim(Y)$ is positively influenced by both the accuracy A_k and the cardinality m. That is, if the user expects to maintain a higher accuracy of the k-nearest neighbors in a lower-dimensional space Y, then the target dimension $dim(Y)$ should be also set higher. In the extreme case, if $Y = X$, then $A_k = 1.0$. Similarly, if there are a large quantity of data points, then the intuition is that we may need to keep a large number of dimensions in the target low-dimensional space Y.
- Secondly, the impact of the accuracy A_k should be higher than that of the cardinality m. This can be understood with the intuition that $dim(Y)$ is quite sensitive to A_k because A_k is a global metric of the entire space Y. On the other hand, a small change of the number of data points may or may not significantly change the distribution of the space.

Based on the above discussion, we postulate that the function g takes the following form

$$g(a, b) = \mathcal{O}(b \cdot 2^a), \tag{3}$$

where the exponential factor 2^a underscores the significantly higher impact of a. In other words, the closed-form function is in the form of

$$dim(Y) = \mathcal{O}(m \cdot 2^{A_k}).$$

If we simplify the asymptotic notation with constant coefficients, we can also rewrite the function as

$$A_k = c_0 \cdot \log \frac{dim(Y)}{m} + c_1, \tag{4}$$

where c_0 and c_1 are two constant values that can be estimated by various regression models. When $A_k = 1$, we say the low-dimensional space Y is OP_k to the original space X, which formalizes our previous notion of OP_K.

In the following sections, we will verify the effectiveness of the function defined in Eq. (4); but before that, we will describe how we implement the measure $\mu(\cdot)$ and the dimension-reduction function f guided by Eq. (4).

Integration into Multimodal Data Retrieval

We will first describe how the embedding vectors are generated from multimodal data, specifically the image-text pair and then present how we implement the hypothetical function and incorporate it into popular dimension-reduction methods.

To leverage transformer-based models, including CLIP (Contrastive Language-Image Pretraining), Vision Transformer (ViT), and BERT (Bidirectional Encoder Representations from Transformers), for converting multimodal data into embedding vectors, we conducted experiments on three distinct datasets: materials science data, Flickr30k, and OmniCorpus. Each dataset consists of multimodal data, incorporating text from HDF5 files, natural language text, and images in TIFF, PNG, or JPEG formats. The experimental results demonstrate that our approach maintains strong applicability across different domains, including both scientific data and natural images.

The generated embeddings retain the default dimensionality of each respective model: BERT and ViT both produce 768-dimensional embeddings. For CLIP, the text and image encoders each output 512-dimensional vectors. We construct unified multimodal representations by directly concatenating embeddings from different modalities—for example, combining CLIP text and image vectors into a single 1024-dimensional vector.

In the case of audio-text data from the ESC-50 dataset, we generate embeddings using BERT (768D) for textual labels and PANNs CNN14 (Kong et al. 2020) (2048D) for audio signals. These are concatenated into 2816-dimensional joint vectors. All embeddings, spanning image, text, and audio modalities, are stored for subsequent dimensionality reduction and retrieval analysis.

Following the embedding vector extraction, we focus on the dimensionality reduction phase. We evaluated several mainstream techniques including PCA, and MDS, among which PCA consistently outperformed others in terms of maintaining the integrity of location information in our datasets. Our investigations revealed a notable correlation between the effectiveness of PCA in preserving the spatial relationships and the ratio $\frac{n}{m}$, where $n = dim(Y)$ and $m = |Y|$, of the target dimension to the number of samples used in the PCA process. We adopted various regression models to elucidate the relationship between the accuracy of preserved relative location information and the ratio $\frac{n}{m}$. These models facilitate the prediction of an optimal embedding vector dimension required to achieve a predetermined accuracy level, given a known number of samples, m.

Evaluation

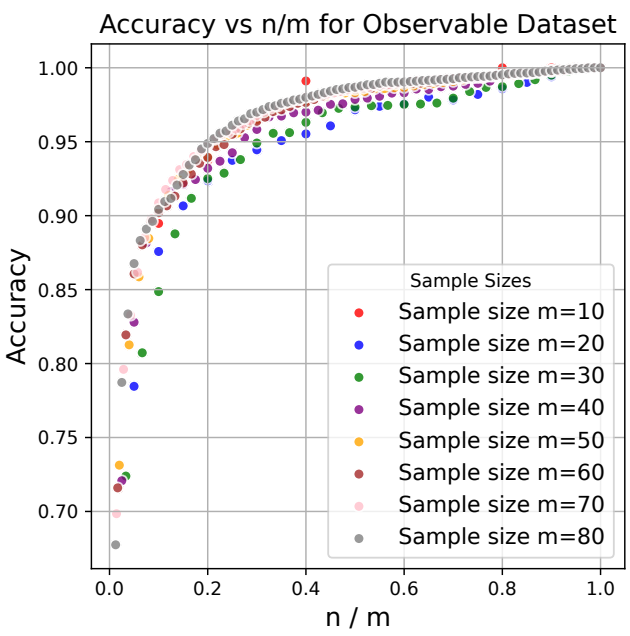
This section evaluates the effectiveness of the proposed method for KNN-preserved dimension reduction on multiple data sets. To ensure the robustness of our method across various scenarios, we evaluate the method with several Transformer-based models for extracting embedding vectors, two popular dimension-reduction techniques, and three distance metrics. All results suggest that the proposed method is highly effective.

More experimental results can be found in the full technical report: https://arxiv.org/abs/2408.10264

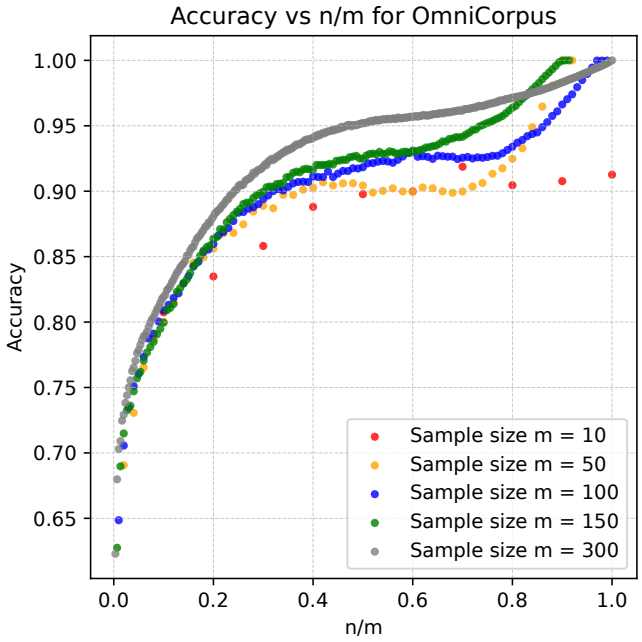
Experimental Setup

Data Sets The evaluation was conducted using seven datasets: four scientific datasets from the Materials Project (The Materials Project Accessed 2024), two natural image-text pair datasets (Flickr30k (Young et al. 2014) and OmniCorpus-037 CC (Li et al. 2024)), and one audio-text dataset (ESC-50 (Piczak 2015)).

Specifically, the observable, stable, metal, and magnetic datasets contain 33,990; 48,884; 72,252; and 81,723 data points, respectively. Flickr30k includes 31,014 image-text pairs, while OmniCorpus-037 CC provides 3,878,063 such pairs. The ESC-50 dataset comprises 2,000 environmental audio clips, each paired with a textual label describing the sound event (e.g., “dog barking”, “siren”, “rain”).



(a) Scientific Data (Observable Dataset)



(b) Large-scale Vision Data (OmniCorpus)

Figure 1: Performance Analysis.

Platform Our experiments were carried out on the NSF-sponsored computing platform CloudLab (Duplyakin et al. 2019), specifically the *d7525* nodes with the following specifications. The machine is equipped with two 16-core AMD EPYC 7302 CPUs at 3.00GHz, 128GB ECC Memory (8×16 GB 3200MT/s RDIMMs), and two 480 GB 6G SATA SSDs along with one 1.6 TB PCIe4 x4 NVMe SSD for storage. The network interface card (NIC) is a dual-port Mellanox ConnectX-6 DX 100Gb NIC, with one port supporting 200Gb connectivity for experimental use. Additionally, the machine is equipped with an NVIDIA 24GB Ampere A30 GPU.

We have installed the following libraries in conda environment: transformer, sklearn, torch, pandas, numpy, os, pickle, h5py, shutil, matplotlib, and pyprismatic, PIL.

OPDR on Various Data Sets

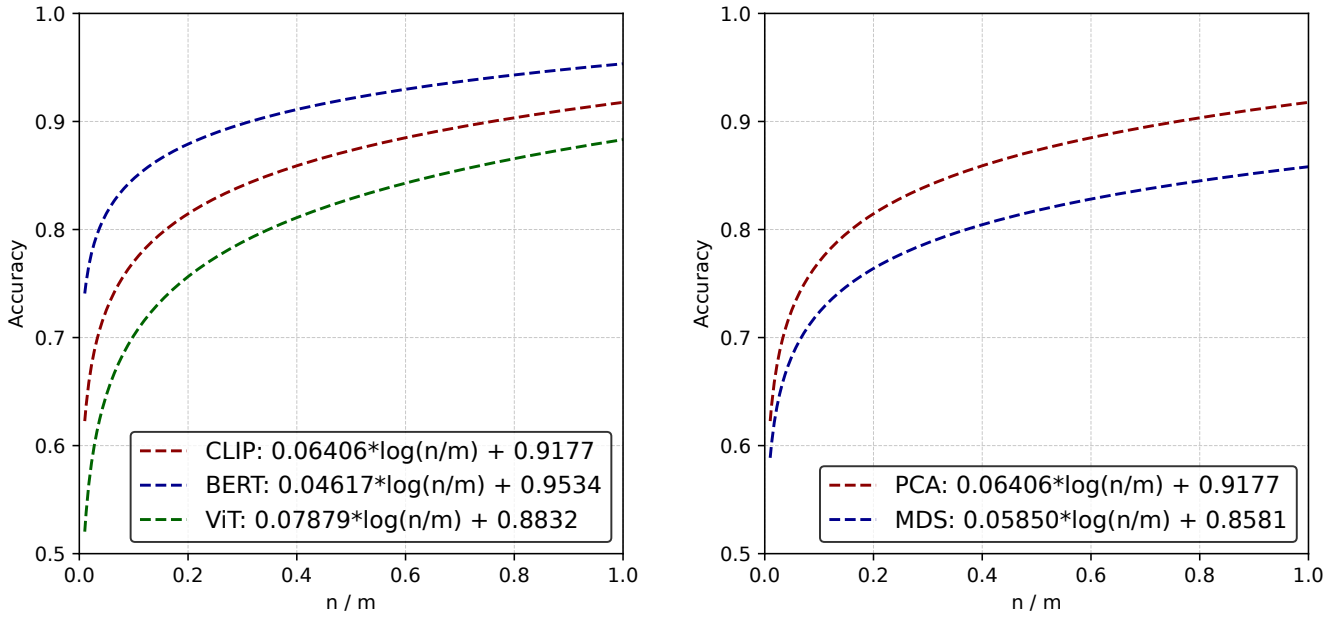
We first use the CLIP model to extract embedding vectors from six datasets, comprising four scientifically curated multimodal datasets (Observable, Unstable, Metal, Non-magnetic), two widely adopted vision-language benchmarks (Flickr30k and OmniCorpus), as well as the ESC-50 dataset (Piczak 2015) for audio-text representation. Subsequently, we computed the distances between vectors using the L2-norm and performed dimensionality reduction using PCA. To simplify notation, we let $n = \dim(Y)$ denote the dimension of the lower-dimensional space. For these four groups of material datasets, we further divided them into eight subsets each, with sample sizes $m \in \{10, 20, 30, 40, 50, 60, 70, 80\}$. For the remaining two multimodal datasets, we divided them into five subsets each, with sample sizes $m \in \{10, 50, 100, 150, 300\}$, to test whether the distribution patterns of data points remain consistent across different sample sizes.

Observations across all evaluated datasets indicate a strong positive correlation between accuracy and the ratio of n to m. As n approaches m, accuracy initially increases rapidly and then slows down, converging to a stable value. This trend is consistently observed across the four scientifically curated datasets as well as the multimodal benchmarks, despite differences in data modality and scale. Although the four different material-related datasets exhibit highly similar data distributions, minor differences may be attributed to the inherent distribution characteristics and randomness of the samples. Representative results illustrating these behaviors are shown in Figures 1a and Figures 1b.

Through comparative analysis discussed above, we observe that our proposed OPDR algorithm performs consistently across multiple data sets and exhibits the anticipated data patterns, validating our initial hypotheses. In the remainder of this section, we will focus on the Observable Material dataset, exploring the effects of using different Transformer-based models, dimensionality reduction techniques, and various distance metrics.

Influence of Embedding Models

Across all evaluated embedding models and datasets, we observe that the use of different neural network architectures does not substantially affect the overall accuracy trend from Figure 2a. In particular, the relationship between accuracy and the ratio n/m remains stable across models, indicating that the proposed closed-form function is applica-



(a) Embedding Models Fitting

(b) Dimension Reduction Fitting

Figure 2: Data Fitting Analysis on Flicker Dataset.

ble to a broad class of neural network embeddings. Consequently, in practical applications, the choice of embedding model should be guided primarily by dataset characteristics and task-specific requirements rather than concerns over disrupting the underlying data relationship.

Influence of Dimension-Reduction Methods

The datasets analyzed in this study exhibit extremely high dimensionality, often reaching hundreds of thousands of dimensions. Even after applying advanced embedding techniques, the resulting vectors typically retain several hundred to several thousand dimensions, presenting significant challenges for effective analysis and interpretation. To address these issues, we employed two principal dimensionality reduction methods, Multidimensional Scaling (MDS) and Principal Component Analysis (PCA), both widely recognized for their ability to efficiently reduce data complexity while preserving the structural integrity of data relationships. As many modern dimensionality reduction techniques are fundamentally derived from PCA and MDS, our study specifically focused on these methods to assess their practical effectiveness.

Across all evaluated datasets, we find that while the choice of dimensionality reduction method influences absolute accuracy values, it does not alter the overall relationship between accuracy and the ratio n/m. In general, PCA exhibits greater sensitivity to changes in n/m and converges to higher accuracy more rapidly than MDS. As illustrated by the representative results shown in Figure 2b, PCA achieves superior performance on the Flickr30k dataset, while both methods preserve the same global trend predicted by our hypothesis.

These observations indicate that although different dimensionality reduction techniques may affect specific numerical outcomes, the fundamental data structure remains stable. Therefore, in practical applications, selecting an appropriate dimensionality reduction method should be based on the desired balance between accuracy, interpretability, and computational efficiency, rather than concerns about altering the underlying data pattern.

Conclusion and Future Work

This paper proposes a new method, namely Order-Preserving Dimension Reduction (OPDR), to address the challenge of high-dimensional embeddings in multimodal data analytics. From a theoretical perspective, this work introduces concepts (both point-wise and space-level) to measure the closeness of two metric spaces in terms of their k-nearest neighbors and a closed-form function to characterize the relationship among dimensionality, cardinality, and similarity (i.e., the accuracy of the dimension-reduction map). The closed-form function is incorporated into a practical ecosystem by complementing other system components, such as dimension-reduction methods (PCA, MDS) and models (CLIP, ViT, Bert). Extensive evaluation demonstrates both the effectiveness and practical utility of OPDR.

Our future work will further explore the theoretical foundation of the closed-form function, particularly the relationship among metric spaces' dimensionality, cardinality, and the accuracy of the dimension-reduction map. Beyond multimodal retrieval, we aim to explore potential applications in medical imaging, where high-dimensional embeddings from MRI (Chamberland et al. 2019; Sridhar et al. 2022), CT scans, and vital records could benefit from this work.

Acknowledgments

This research is supported by the U.S. Department of Energy (DOE) through the Office of Advanced Scientific Computing Research's "Orchestration for Distributed & Data-Intensive Scientific Exploration" and the "Decentralized data mesh for autonomous materials synthesis" AT SCALE LDRD at Pacific Northwest National Laboratory. PNNL is operated by Battelle for the DOE under Contract DE-AC05-76RL01830. This work used TAMU ACES at Texas A&M University through allocation CHE240191 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603 and #2138296. Some results presented in this paper were partially obtained using the Chameleon testbed supported by the National Science Foundation.

References

Aremu, O. O.; Hyland-Wood, D.; and McAree, P. R. 2020. A machine learning approach to circumventing the curse of dimensionality in discontinuous time series machine data. *Reliability Engineering & System Safety*, 195: 106706.

Chamberland, M.; Raven, E. P.; Genc, S.; Duffy, K.; Descoteaux, M.; Parker, G. D.; Tax, C. M.; and Jones, D. K. 2019. Dimensionality reduction of diffusion MRI measures for improved tractometry of the human brain. *NeuroImage*, 200: 89–100.

Chen, S.; Gong, C.; Li, J.; Yang, J.; Niu, G.; and Sugiyama, M. 2022. Learning Contrastive Embedding in Low-Dimensional Space. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 6345–6357. Curran Associates, Inc.

Dai, G.; and Yeung, D.-Y. 2006. Tensor Embedding Methods. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 330–335. Boston, MA, USA: AAAI Press.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Douze, M.; Guzhva, A.; Deng, C.; Johnson, J.; Szilvasy, G.; Mazaré, P.-E.; Lomeli, M.; Hosseini, L.; and Jégou, H. 2024. The Faiss library.

Duplyakin, D.; Ricci, R.; Maricq, A.; Wong, G.; Duerig, J.; Eide, E.; Stoller, L.; Hibler, M.; Johnson, D.; Webb, K.; Akella, A.; Wang, K.; Ricart, G.; Landweber, L.; Elliott, C.; Zink, M.; Cecchet, E.; Kar, S.; and Mishra, P. 2019. The Design and Operation of CloudLab. In *Proceedings of the USENIX Annual Technical Conference (ATC)*, 1–14.

Gao, Y.; Liu, J.; Xu, Z.; Wu, T.; Zhang, E.; Li, K.; Yang, J.; Liu, W.; and Sun, X. 2024. SoftCLIP: Softer Cross-Modal Alignment Makes CLIP Stronger. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(3): 1860–1868.

George, L.; and Sumathy, P. 2023. An integrated clustering and BERT framework for improved topic modeling. *International Journal of Information Technology*, 15(4): 2187–2195.

Guo, R.; Sun, P.; Lindgren, E.; Geng, Q.; Simcha, D.; Chern, F.; and Kumar, S. 2020. Accelerating large-scale inference with anisotropic vector quantization. In *International Conference on Machine Learning*, 3887–3896. PMLR.

Gupta, T.; Zaki, M.; Krishnan, N. A.; and Mausam. 2022. MatSciBERT: A materials domain language model for text mining and information extraction. *npj Computational Materials*, 8(1): 102.

He, X.; Cai, D.; Yan, S.; and Zhang, H.-J. 2005. Neighborhood preserving embedding. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 2, 1208–1213. IEEE.

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

Hu, H.; Feng, D.-Z.; and Chen, Q.-Y. 2021. A novel dimensionality reduction method: Similarity order preserving discriminant analysis. *Signal Processing*, 182: 107933.

Hu, P.; Zhu, H.; Peng, X.; and Lin, J. 2020. Semi-supervised multi-modal learning with balanced spectral decomposition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 99–106.

Huang, Z.; Liang, D.; Xu, P.; and Xiang, B. 2020. Improve transformer models with better relative position embeddings. *arXiv preprint arXiv:2009.13658*.

Jain, M. S.; Polanski, K.; Conde, C. D.; Chen, X.; Park, J.; Mamanova, L.; Knights, A.; Botting, R. A.; Stephenson, E.; Haniffa, M.; Lamacraft, A.; Efremova, M.; and Teichmann, S. A. 2023. MultiMAP: Dimensionality Reduction and Integration of Multimodal Data. In *Proceedings of the ICML CompBio 2023*. ICML.

Jia, C.; Yang, Y.; Xia, Y.; Chen, Y.-T.; Parekh, Z.; Pham, H.; Le, Q. V.; Sung, Y.; Li, Z.; and Duerig, T. 2021. Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. arXiv:2102.05918.

Jiang, C.; Ye, W.; Xu, H.; Ye, Q.; Yan, M.; Zhang, J.; and Zhang, S. 2024. TiMix: Text-aware Image Mixing for Effective Vision-Language Pre-training. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Jin, W.; Li, X.; and Hamarneh, G. 2022. Evaluating Explainable AI on a Multi-Modal Medical Imaging Task: Can Existing Algorithms Fulfill Clinical Requirements? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, 11945–11953. Association for the Advancement of Artificial Intelligence (AAAI).

Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; and Plumbley, M. D. 2020. PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. arXiv:1912.10211.

Kruskal, J. B.; and Wish, M. 1978. *Multidimensional Scaling*. Sage Publications.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33: 9459–9474.

Li, M.; Xu, R.; Wang, S.; Zhou, L.; Lin, X.; Zhu, C.; Zeng, M.; Ji, H.; and Chang, S.-F. 2022. CLIP-Event: Connecting Text and Images with Event Structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Li, Q.; Chen, Z.; Wang, W.; Wang, W.; Ye, S.; Jin, Z.; Chen, G.; He, Y.; Gao, Z.; Cui, E.; et al. 2024. OmniCorpus: A Unified Multimodal Corpus of 10 Billion-Level Images Interleaved with Text. *arXiv preprint arXiv:2406.08418*.

Liang, X.; Yang, E.; Yang, Y.; and Deng, C. 2024. Multi-relational deep hashing for cross-modal search. *IEEE Transactions on Image Processing*.

Liu, P.; Ren, Y.; Tao, J.; and Ren, Z. 2024. GIT-Mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171: 108073.

Lu, S.; Ding, Y.; Liu, M.; Yin, Z.; Yin, L.; and Zheng, W. 2023. Multiscale feature extraction and fusion of image and text in VQA. *International Journal of Computational Intelligence Systems*, 16(1): 54.

Malkov, Y. A.; and Yashunin, D. A. 2018. Efficient and robust approximate nearest neighbor search using Hierarchical Navigable Small World graphs. arXiv:1603.09320.

Piczak, K. J. 2015. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, 1015–1018. New York, NY, USA: Association for Computing Machinery. ISBN 9781450334594.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Ran, R.; Gao, T.; and Fang, B. 2022. Transformer-based dimensionality reduction. arXiv:2210.08288.

Rangel DaCosta, L.; Brown, H. G.; Pelz, P. M.; Rakowski, A.; Barber, N.; O'Donovan, P.; McBean, P.; Jones, L.; Ciston, J.; Scott, M.; and Ophus, C. 2021. Prismatic 2.0 – Simulation software for scanning and high resolution transmission electron microscopy (STEM and HRTEM). *Micron*, 151: 103141.

Song, L.; Gretton, A.; Borgwardt, K.; and Smola, A. 2007. Colored Maximum Variance Unfolding. In Platt, J.; Koller, D.; Singer, Y.; and Roweis, S., eds., *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc.

Sridhar, C.; Pareek, P. K.; Kalidoss, R.; Jamal, S. S.; Shukla, P. K.; and Nuagah, S. J. 2022. Optimal medical image size reduction model creation using recurrent neural network and GenPSOWVQ. *Journal of Healthcare Engineering*, 2022(1): 2354866.

Tenenbaum, J. B.; de Silva, V.; and Langford, J. C. 2000. A global geometric framework for nonlinear dimensionality reduction. In *Science*, volume 290, 2319–2323.

The Materials Project. Accessed 2024. https://next-gen.materialsproject.org/materials.

Torgerson, W. S. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika*, 17(4): 401–419.

Tshitoyan, V.; Dagdelen, J.; Weston, L.; Dunn, A.; Rong, Z.; Kononova, O.; Persson, K. A.; Ceder, G.; and Jain, A. 2019. Unsupervised word embeddings capture latent knowledge from materials science literature. *Nature*, 571(7763): 95–98.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wu, T.-L.; Singh, S.; Paul, S.; Burns, G.; and Peng, N. 2020. MELINDA: A Multimodal Dataset for Biomedical Experiment Method Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Young, P.; Lai, A.; Hodosh, M.; and Hockenmaier, J. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics*, 2: 67–78.

Zhang, Y.; and Cheung, Y.-m. 2020. An ordinal data clustering algorithm with automated distance learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 6869–6876.

Zhang, Z.; Zhu, X.; Lu, G.; and Zhang, Y. 2021. Probability ordinal-preserving semantic hashing for large-scale image retrieval. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 15(3): 1–22.

Zhao, R.; Chen, H.; Wang, W.; Jiao, F.; Do, X. L.; Qin, C.; Ding, B.; Guo, X.; Li, M.; Li, X.; and Joty, S. 2023. Retrieving Multimodal Information for Augmented Generation: A Survey. arXiv:2303.10868.

Zhou, N.; and Fan, J. 2015. Automatic image–text alignment for large-scale web image indexing and retrieval. *Pattern Recognition*, 48(1): 205–219.