# On Achieving Perfect Multimodal Alignment for Cross-modal Transfer

Abhi Kamboj[1] Minh N. Do[1]
University of Illinois at Urbana-Champaign[1]
{akamboj2, minhdo}@illinois.edu

## Abstract

*Multimodal alignment aims to construct a joint latent vector space where two modalities representing the same concept map to the same vector. We formulate this as an inverse problem and show that, under certain conditions, perfect alignment can be achieved. When perfect alignment cannot be achieved, it can be approximated using the Singular Value Decomposition (SVD) of a multimodal data matrix. Experiments on synthetic multimodal Gaussian data verify the effectiveness of our perfect alignment method compared to the popular contrastive alignment method. We discuss how these findings can be applied to visual data and sensor data for unsupervised cross-modal transfer. We hope these findings inspire further exploration of the applications of perfect alignment for cross-modal learning.*

## 1. Introduction

Humans naturally perceive the same concept through multiple senses, a capability artificial intelligence (AI) aims to replicate with multimodal data. However, integrating diverse data types remains challenging due to differences in abundance, information richness, and annotation difficulty. For example, images and videos are plentiful and easy to label, while modalities like MRI, ECG, or IMU are scarce and harder to annotate. This diversity raises a fundamental question: How can we unify visual representations across such varied modalities to effectively transfer knowledge and improve AI performance? This challenge is especially important for modalities that are uncommon or more complex.

To interpret multimodal data, AI methods typically align the semantic meanings of different modalities within a shared latent space at the output of modality-specific encoders. For instance, models can associate an image with descriptive text [8] or with corresponding sounds, videos or other sensors [4]. While such alignment is often achieved through large-scale learned methods and specialized loss functions, these approaches remain fundamentally approximate and often lack theoretical rigor and interpretability.

Prior work explores contrastive alignment through geometric [9], probabilistic [1, 3], and information-theoretic [6, 7] perspectives. However, these analyses primarily reinterpret existing alignment frameworks rather than proposing new methodologies. In this work, we reframe multimodal alignment as a linear inverse problem, a class of problems well-studied in linear algebra and signal processing. This reframing enables the derivation of a representation space with perfect alignment between two modalities. We define perfect alignment as the existence of modality-specific encoders that map training instances from distinct modalities to identical latent representations. Notably, we demonstrate that empirical risk minimization and linear regression emerge as special cases of this framework, bridging classical machine learning paradigms to multimodal alignment.

We validate our perfect alignment approach on synthetic multimodal data, demonstrating that our method achieves strong alignment and competitive performance compared to contrastive learning-based methods. These results suggest that our framework has the potential to generalize to real multimodal data, motivating further exploration of perfect alignment for cross-modal tasks. Our work offers insights into the Platonic representation hypothesis [5], which posits that representations of the same semantic concepts from different modalities converge to a shared latent space. While our results neither fully confirm nor refute this hypothesis, they empirically demonstrate the existence of a perfect alignment space between two modalities, enabling tasks like zero-shot classification, cross-modal retrieval, and transfer learning. Our contributions are:

- **Theoretical Framework:** A novel inverse problem formulation for multimodal alignment, providing closed-form solutions for perfect alignment.
- **Empirical Validation:** Successful alignment on synthetic data, demonstrating the potential for tasks like zero-shot classification and cross-modal retrieval.

## 2. Methods

We propose a method to achieve **perfect alignment** between two modalities by solving an inverse problem to construct the aligned latent space. To formalize this, we first define key notation and assumptions.

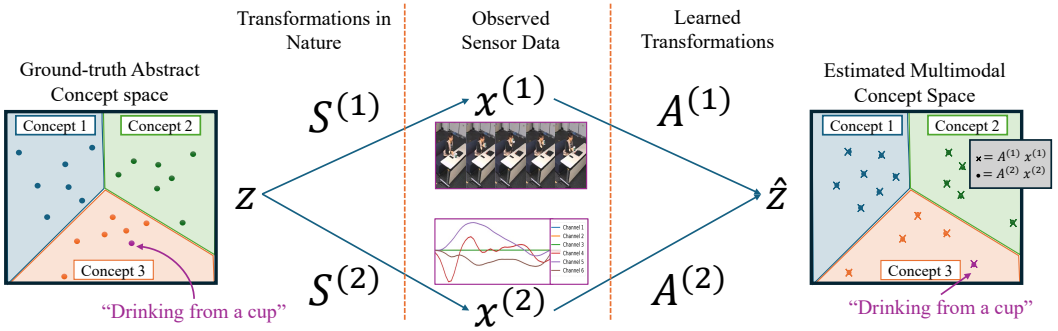


Figure 1. **Data generation model:** Latent concepts $\mathbf{z}_i$ are transformed through modality-specific matrices $\mathbf{S}^{(1)}, \mathbf{S}^{(2)}$ to generate observations in different modalities $\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)}$. Our goal is to recover alignment matrices $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}$ that invert these transformations. Ideally, the learned transformation preserves the latent structure such that a sample from a concept class in the original latent space remains in that class grouping in the estimated latent space

- Superscripts (e.g., $^{(1)}$, $^{(2)}$) denote modalities.
- Subscripts (e.g., $_i$, $_j$) index samples in a dataset.
- Lowercase bold letters (e.g., $\mathbf{x}$, $\mathbf{z}$) represent vectors.
- Uppercase bold letters (e.g., $\mathbf{A}$, $\mathbf{X}$) indicate matrices.
- Calligraphic letters (e.g., $\mathcal{X}$, $\mathcal{Z}$) denote vector spaces or sets.

**Definition 2.1** (Perfect Alignment)**.** Let $\mathcal{X}^{(1)}$ and $\mathcal{X}^{(2)}$ denote the input spaces of two modalities, with $\mathcal{Z}$ being their shared latent space. Given a dataset $\mathcal{D} = \{(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})\}_{i=1}^n$ of corresponding multimodal instances, *perfect alignment* is defined by the existence of encoder functions $f^{(1)} : \mathcal{X}^{(1)} \to \mathcal{Z}$ and $f^{(2)} : \mathcal{X}^{(2)} \to \mathcal{Z}$ satisfying:

$$\forall (\mathbf{x}^{(1)}, \mathbf{x}^{(2)}) \in \mathcal{D}, \quad f^{(1)}(\mathbf{x}^{(1)}) = f^{(2)}(\mathbf{x}^{(2)}) = \mathbf{z}, \quad (1)$$

where $\mathbf{z} \in \mathcal{Z}$ is the unified semantic representation of the shared concept underlying the pair $(\mathbf{x}^{(1)}, \mathbf{x}^{(2)})$.

Let $\mathcal{Z} \subseteq \mathbb{R}^k$ be a ground truth latent space representing semantic concepts. We assume each modality $m \in 1, 2$ is generated via linear transformations:

$$\mathbf{x}_i^{(m)} = \mathbf{S}^{(m)} \mathbf{z}_i, \quad (2)$$

where for sample $i$:
- $\mathbf{z}_i \in \mathcal{Z} \subseteq \mathbb{R}^k$ is the latent concept vector
- $\mathbf{S}^{(m)} \in \mathbb{R}^{d_m \times k}$ is the modality-specific generation matrix
- $\mathbf{x}_i^{(m)} \in \mathcal{X}^{(m)} \subseteq \mathbb{R}^{d_m}$ is the observed data in modality $m$

For aligned pairs $(\mathbf{x}_i^{(1)}, \mathbf{x}_i^{(2)})$ generated from the same $\mathbf{z}_i$, our goal is to recover projection matrices $\mathbf{A}^{(1)} \in \mathbb{R}^{k \times d_1}$ and $\mathbf{A}^{(2)} \in \mathbb{R}^{k \times d_2}$ such that:

$$\mathbf{A}^{(1)} \mathbf{x}_i^{(1)} = \mathbf{A}^{(2)} \mathbf{x}_i^{(2)} = \mathbf{z}_i \quad \forall i \in 1, \ldots, n. \quad (3)$$

This reduces to solving the system:

$$\mathbf{A}^{(1)} \mathbf{X}^{(1)} - \mathbf{A}^{(2)} \mathbf{X}^{(2)} = \mathbf{0}, \quad (4)$$

where $\mathbf{X}^{(m)} = [\mathbf{x}_1^{(m)} \cdots \mathbf{x}_n^{(m)}] \in \mathbb{R}^{d_m \times n}$ in which each column $\mathbf{x}_i^{(m)}$ represents the $i$-th data point in modality $m$.

When Eq. (4) holds, we achieve **perfect alignment** as defined in Theorem 2.1, where the encoder functions $f^{(1)}$ and $f^{(2)}$ correspond to the linear transformations $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, respectively. To recover these matrices we construct the combined matrices:

$$\mathbf{A} = \left[ \begin{array}{c:c} \mathbf{A}^{(1)} & \mathbf{A}^{(2)} \end{array} \right] \in \mathbb{R}^{k \times d}, \quad \mathbf{X} = \left[ \begin{array}{c} \mathbf{X}^{(1)} \\ \hdashline -\mathbf{X}^{(2)} \end{array} \right] \in \mathbb{R}^{d \times n}, \quad (5)$$

where $d = d_1 + d_2$.

This allows us to rewrite Eq. (4) as the linear inverse problem:

$$\mathbf{A}\mathbf{X} = \mathbf{0}, \quad (6)$$

where $\mathbf{0} \in \mathbb{R}^{k \times n}$ is the zero matrix. The goal is to find a non-trivial solution $\mathbf{A} \neq \mathbf{0}$ that satisfies this equation.

**Theorem 2.2** (Existence of Perfect Alignment)**.** *Given the inverse problem $\mathbf{A}\mathbf{X} = \mathbf{0}$ defined in Eq. (6), where $\mathbf{X} \in \mathbb{R}^{d \times n}$ is a given data matrix and $\mathbf{A} \in \mathbb{R}^{k \times d}$ is unknown, if $\mathbf{X}$ has a left null space $\mathcal{N}(\mathbf{X}^T)$ of dimension $\dim(\mathcal{N}(\mathbf{X}^T)) \geq k$, then there exists a closed-form solution for $\mathbf{A}$. Specifically, the rows of $\mathbf{A}$ can be formed by any $k$ linearly independent vectors spanning $\mathcal{N}(\mathbf{X}^T)$.*

*Proof.* The proof involves recognizing that any vector $\mathbf{a}$ in the left null space of $\mathbf{X}$ satisfies $\mathbf{a}^T\mathbf{X} = \mathbf{0}$. Therefore, if $\mathbf{X}$ has a null space of dimension at least $k$, we can select $k$ linearly independent vectors from this null space to form the rows of $\mathbf{A}$. This ensures that $\mathbf{A}\mathbf{X} = \mathbf{0}$ is satisfied. A full proof is given in Appendix Sec. 7.1. □

**Corollary 2.3** (Approximate Alignment)**.** *If $\mathbf{X} \in \mathbb{R}^{d \times n}$ has a left null space $\mathcal{N}(\mathbf{X}^T)$ with $\dim(\mathcal{N}(\mathbf{X}^T)) < k$, an approximation to $\mathbf{A}\mathbf{X} = \mathbf{0}$ can be obtained by selecting the $k$ basis vectors corresponding to the smallest singular values*

*of* $\mathbf{X}$. *This approximation minimizes the Frobenius norm* $\|\mathbf{AX}\|_F$.

*Proof.* This is a direct application of Eckhart-Young-Mirsky theorem. The full proof is shown in Sec. 7.1. □

**Method for Finding A.** To determine $\mathbf{A}$, compute the Singular Value Decomposition (SVD) of $\mathbf{X}$:

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

where $\mathbf{U} \in \mathbb{R}^{d\times d}$. Assuming $k \leq d$, Extract the last $k$ columns of $\mathbf{U}$, denoted $\mathbf{u}_{d-k+1}, \ldots, \mathbf{u}_d$, which correspond to the basis vectors of the left null space of $\mathbf{X}$ (if $\dim(\mathcal{N}(\mathbf{X}^T)) \geq k$) or its smallest singular values (otherwise). The solution for $\mathbf{A}$ is:

$$\mathbf{A}^* = \begin{bmatrix} \mathbf{u}_{d-k+1}^T \\ \vdots \\ \mathbf{u}_d^T \end{bmatrix}, \tag{7}$$

where $\mathbf{u}_j$ is the $j^{\text{th}}$ column of $\mathbf{U}$. This method achieves perfect alignment when $\dim(\mathcal{N}(\mathbf{X}^T)) \geq k$ and an optimal approximation in the Frobenius norm otherwise.

*Remark* 2.4 (Assumption on $k$ and $d$). The assumption $k \leq d$ is valid because the latent space dimension $k$ is typically smaller than the data dimension $d$ in representation learning. This reflects the common goal of compressing high-dimensional data into a lower-dimensional space while preserving essential semantic information.

*Remark* 2.5 (Achievability and Computational Cost of Perfect Alignment). Perfect alignment is often not achievable in practice because the number of data points $n$ is large, thus $\mathbf{X}$ becomes a wide matrix and the left null space of $\mathbf{X}$ has limited dimensionality. Furthermore, computing $\mathbf{A}$ via full SVD of $\mathbf{X} \in \mathbb{R}^{d\times n}$ has a time complexity of $O(d^2n + dn^2 + n^3)$. For large $d$ and $n$, this becomes prohibitive, motivating approximate methods like gradient descent. When only the $k$ smallest singular vectors are needed, truncated SVD reduces this to $O(dnk)$, making it feasible for moderate $k$.

*Remark* 2.6 (Comparison to Linear Regression). Standard linear regression minimizes $\arg\min_W \|\mathbf{Y} - \mathbf{WX}\|_F^2$, which is a special case of our alignment objective $\arg\min_A \|\mathbf{AX}\|_F^2$ when:

- $\mathbf{A}^{(1)} = \mathbf{I}_d$ (identity mapping for modality 1)
- $\mathbf{X}^{(1)} = \mathbf{Y}$ (one modality is the regression target)
- $\mathbf{A}^{(2)} = \mathbf{W}$ (learned regression weights for modality 2)
- $\mathbf{X}^{(2)} = \mathbf{X}$ (second modality is the data)

*Remark* 2.7 (Connection to Empirical Risk Minimization (ERM)). ERM seeks a model $f \in \mathcal{F}$ that minimizes the empirical risk:

$$R_{\text{emp}}(f) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), \mathbf{y}_i).$$

Our framework extends this to two hypothesis classes $\mathcal{F}^{(1)}, \mathcal{F}^{(2)}$, minimizing:

$$R_{\text{emp}}(f^{(1)}, f^{(2)}) = \frac{1}{n}\sum_{i=1}^{n} \ell\left(f^{(1)}(\mathbf{x}_i^{(1)}), f^{(2)}(\mathbf{x}_i^{(2)})\right),$$

where $f^{(1)}, f^{(2)}$ are linear transformations $\mathbf{A}^{(1)}, \mathbf{A}^{(2)}$, and $\ell$ is the alignment loss $\|\mathbf{AX}\|_F^2$.

## 2.1. Error Metrics for Perfect Alignment

Let $\hat{\mathbf{z}}_i^{(m)} = \mathbf{A}^{(m)}\mathbf{x}_i^{(m)}$ denote the latent vectors estimated from modality $m$. We define two error metrics to evaluate alignment quality:

1. **Cross-Modal Alignment Error (CMAE)**: Quantifies the discrepancy between latent representations from different modalities. CMAE is computed as:

$$\text{CMAE} = \frac{1}{n}\sum_{i=1}^{n} \left\|\hat{\mathbf{z}}_i^{(1)} - \hat{\mathbf{z}}_i^{(2)}\right\|_2. \tag{8}$$

   Note that for normalized latent vectors (i.e., $\|\hat{\mathbf{z}}_i^{(m)}\|_2 = 1$), minimizing CMAE is equivalent to maximizing their cosine similarity—the same objective as the InfoNCE loss used in contrastive learning [6]. *Unlike contrastive methods, our framework does not assume normalized latent vectors; thus, cosine similarity is not directly applicable as a metric.*
2. **Modality Latent Reconstruction Error (MLRE)**: Measures fidelity to the true latent space $\mathcal{Z}$, applicable only in synthetic experiments where $\mathbf{z}_i$ is known. For modality $m$, MLRE is:

$$\text{MLRE}^{(m)} = \frac{1}{n}\sum_{i=1}^{n} \left\|\mathbf{z}_i - \hat{\mathbf{z}}_i^{(m)}\right\|_2. \tag{9}$$

# 3. Experiments

**Data Generation.** We generate synthetic data from a ground-truth latent space $\mathcal{Z} \subseteq \mathbb{R}^2$, modeled as a mixture of two Gaussian distributions:

$$\mathcal{Z} \sim \pi_1\mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_1) + \pi_2\mathcal{N}(\boldsymbol{\mu}_2, \mathbf{\Sigma}_2), \tag{10}$$

where $\pi_1 = \pi_2 = 0.5$, $\boldsymbol{\mu}_1 = [0, 1]^T$, $\boldsymbol{\mu}_2 = [4, 5]^T$, and $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2 = \mathbf{I}_2$ (the 2D identity matrix).

The data generation process proceeds as follows:

1. Sample $n = 2000$ vectors: $\mathcal{D}_Z = \{\mathbf{z}_i\}_{i=1}^{2000}, \mathbf{z}_i \sim \mathcal{Z}$.
2. Project $\mathcal{D}_Z$ into two modalities using randomly generated matrices $\mathbf{S}^{(1)}, \mathbf{S}^{(2)} \in \mathbb{R}^{2\times 2}$, where each entry is uniformly sampled from $[-5, 5]$:

$$\mathcal{D}_{X^{(1)}} = \{\mathbf{x}_i^{(1)} = \mathbf{S}^{(1)}\mathbf{z}_i\}_{i=1}^{2000}, \tag{11}$$

$$\mathcal{D}_{X^{(2)}} = \{\mathbf{x}_i^{(2)} = \mathbf{S}^{(2)}\mathbf{z}_i\}_{i=1}^{2000}. \tag{12}$$

Fig. 1 illustrates this pipeline, showing the latent space clusters and their projections into the two modalities.

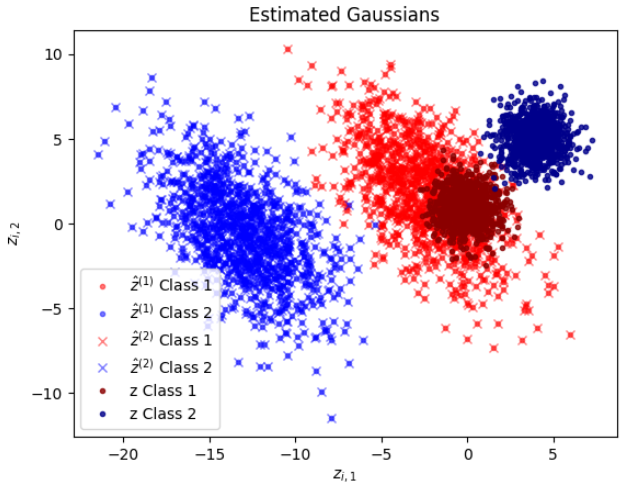


Figure 2. **Aligned Latent Space:** Recovered latent space $\hat{\mathcal{Z}}$ from synthetic data using our alignment method (see Sec. 3). Colors denote ground-truth cluster membership.

Table 1. **Alignment Errors:** Modality Latent Reconstruction Error (MLRE), Cross-Modal Alignment Error (CMAE) and Normalized CMAE (NCMAE) across alignment methods. **Perfect Alignment** achieves near-zero CMAE (even after normalization), while contrastive alignment exhibits higher errors. Normalized CMAE computes alignment error for L2-normalized latent vectors.

| **Alignment** | MLRE$^{(1)}$ | MLRE$^{(2)}$ | **CMAE** | **NCMAE** |
|---|---|---|---|---|
| Perfect (Ours) | 10.9 | 10.9 | $3.66e^{-15}$ | $3.14e^{-16}$ |
| Contrastive[8] | 8.73 | 4.34 | 5.44 | 0.0298 |

**Alignment and Reconstruction Errors.** Using the method in Sec. 2, we compute $\mathbf{A}^{(1)}$ and $\mathbf{A}^{(2)}$, then evaluate CMAE and MLRE as defined in Sec. 2.1. Tab. 1 shows:

- **Near-perfect alignment**: CMAE $\approx 3.66 \times 10^{-15}$ confirms latent representations from both modalities coincide almost exactly. This value is likely floating-point error.
- **High reconstruction error**: MLRE $\approx 10.9$ indicates the estimated latent space $\hat{\mathcal{Z}}$ differs from the ground-truth $\mathcal{Z}$.

Figure 2 visualizes $\hat{\mathcal{Z}}$, where aligned points $\hat{\mathbf{z}}_i^{(1)}$ and $\hat{\mathbf{z}}_i^{(2)}$ overlap perfectly but form clusters distinct from the original GMM. This arises because solutions to $\mathbf{AX} = \mathbf{0}$ are non-unique—any linear transformation of the basis in Eq. (7) yields valid solutions. While perfect alignment is achieved, perfect reconstruction requires identifying a specific transformation that maps $\hat{\mathcal{Z}}$ to $\mathcal{Z}$, a more constrained problem that requires more information about $\mathcal{Z}$

**Key Insight:** Despite high MLRE, the transformations preserve cluster structure (Gaussianity). This enables class separation in $\hat{\mathcal{Z}}$ and demonstrates that perfect alignment—not exact latent space recovery may suffice for cross-modal transfer tasks.

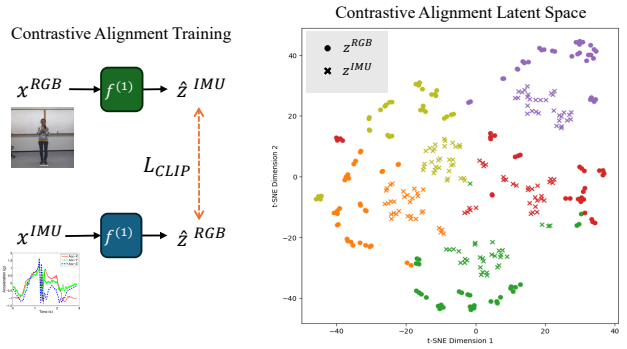


Figure 3. **Contrastive Alignment TSNE Plots:** TSNE visualization of latent representations for five activity classes (Bowling, Clap, Draw circle (clockwise), Jog, Basketball shoot) after contrastive alignment of video and IMU data on the UTD-MHAD [2] dataset. Notice how the latent representations of the RGB (o) and IMU (x) data seldom overlap, indicating imperfect alignment.

## 4. Limitations and Future Work

While our method demonstrates strong performance on synthetic data, several challenges remain for real-world applications. Scaling to real datasets, such as aligning visual and wearable sensor data for human activity recognition, is a primary goal. As illustrated in Fig. 3, t-SNE visualizations of latent representations generated through contrastive alignment of video and IMU data show relative class grouping but imperfect alignment between the modalities. We believe our perfect alignment approach can close this gap, enabling more precise cross-modal generation and transfer.

A key limitation is computational cost: perfect alignment in high-dimensional spaces is expensive, and our method assumes data is generated via linear transformations. In practice, real-world data is often highly non-linear, making direct application challenging.

To address these issues, we propose applying our perfect alignment method to the output space of pretrained modality-specific encoders. We are currently exploring two strategies: (1) using variational autoencoders (VAEs) for feature extraction, and (2) applying perfect alignment to the output features of pretrained CLIP encoders.

## 5. Conclusion

We introduced a method for perfect multimodal alignment by formulating the problem as an inverse projection onto a shared latent space. Experiments on synthetic data show near-zero alignment error. Importantly, the transformation preserves the relative structure of the data, potentially allowing for class identification in the estimated latent space. These results highlight the promise of our alignment technique for multimodal analysis and motivate further research into perfect alignment on complex, real-world datasets.

## 6. Acknowledgments

This material is based upon work supported by the National Science Foundation Graduate Research Fellowship.

## References

[1] Yongwei Che and Benjamin Eysenbach. The" law" of the unconscious contrastive learner: Probabilistic alignment of unpaired modalities. *arXiv preprint arXiv:2501.11326*, 2025. 1

[2] Chen Chen, Roozbeh Jafari, and Nasser Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *2015 IEEE International conference on image processing (ICIP)*, pages 168–172. IEEE, 2015. 4

[3] Zihao Chen, Chi-Heng Lin, Ran Liu, Jingyun Xiao, and Eva Dyer. Your contrastive learning problem is secretly a distribution alignment problem. *Advances in Neural Information Processing Systems*, 37:91597–91617, 2024. 1

[4] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, 2023. 1

[5] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, pages 20617–20642. PMLR, 2024. 1

[6] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1, 3

[7] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International conference on machine learning*, pages 5171–5180. PMLR, 2019. 1

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 4

[9] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International conference on machine learning*, pages 9929–9939. PMLR, 2020. 1

# 7. Appendix

## 7.1. Proofs of Theorem 2.2 and Corollary 2.3

**Theorem 2.2 Restated** (Existence of Perfect Alignment). *Given the inverse problem $AX = \mathbf{0}$ constructed in Eq. (6), where $X \in \mathbb{R}^{d\times n}$ is a given matrix and $A \in \mathbb{R}^{k\times d}$ is unknown, if $X$ has a left null space of at least $k$ dimensions, then there exists a closed-form solution for $A$. Specifically, the rows of $A$ can be formed by any $k$ vectors that constitute a basis for the left null space of $X$.*

*Proof.* Let $\mathcal{N}(X^T)$ denote the left null space of $X$, defined as:

$$\mathcal{N}(X^T) = \{\mathbf{y} \in \mathbb{R}^d \mid X^T\mathbf{y} = \mathbf{0}\}$$

By the rank-nullity theorem:

$$\dim(\mathcal{N}(X^T)) = d - \operatorname{rank}(X)$$

The theorem assumes $\dim(\mathcal{N}(X^T)) \geq k$. Therefore, there exist $k$ linearly independent vectors $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k\} \subseteq \mathcal{N}(X^T)$.

Construct $A \in \mathbb{R}^{k\times d}$ by setting these vectors as its rows:

$$A = \begin{bmatrix} \mathbf{v}_1^T \\ \mathbf{v}_2^T \\ \vdots \\ \mathbf{v}_k^T \end{bmatrix}$$

For each row $\mathbf{v}_i^T$ of $A$, we have:

$$\mathbf{v}_i^T X = \mathbf{0}^T \quad (\text{since } \mathbf{v}_i \in \mathcal{N}(X^T))$$

Therefore, the matrix product satisfies:

$$AX = \begin{bmatrix} \mathbf{v}_1^T X \\ \mathbf{v}_2^T X \\ \vdots \\ \mathbf{v}_k^T X \end{bmatrix} = \begin{bmatrix} \mathbf{0}^T \\ \mathbf{0}^T \\ \vdots \\ \mathbf{0}^T \end{bmatrix} = \mathbf{0}$$

The rows of $A$ are linearly independent by construction, as they form a basis for a $k$-dimensional subspace of $\mathcal{N}(X^T)$. This completes the proof that such an $A$ exists and satisfies $AX = \mathbf{0}$.

The closed-form solution arises from the fact that $\mathcal{N}(X^T)$ can be explicitly computed via:

- Singular Value Decomposition (SVD): If $X = U\Sigma V^T$, then $\mathcal{N}(X^T)$ is spanned by the last $d - \operatorname{rank}(X)$ columns of $U$.
- Reduced Row Echelon Form: For $X^T$, the null space basis vectors correspond to the free variables in $\operatorname{rref}(X^T)$.

Thus, any $k$ linearly independent vectors from these computed bases will satisfy the requirements for $A$. □

**Corollary 2.3 Restated** (Approximate Alignment in Frobenius Norm). *If $X \in \mathbb{R}^{d\times n}$ has a left null space with fewer than $k$ dimensions, an approximation of the solution to $AX = \mathbf{0}$ can be obtained by selecting the basis vectors corresponding to the $k$ smallest singular values of $X$. This approximation minimizes the Frobenius norm $\|AX\|_F$.*

*Proof.* Let $X = U\Sigma V^T$ be the SVD of $X$, where $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{d\times n}$ contains the singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r \geq 0$ with $r = \operatorname{rank}(X)$.

The left null space of $X$ is spanned by columns of $U$ corresponding to zero singular values. When $\dim(\mathcal{N}(X^T)) < k$, we instead use the $k$ columns of $U$ associated with the smallest singular values $\sigma_{d-k+1}, \ldots, \sigma_d$. Constructing $A$ as:

$$A = \begin{bmatrix} u_{d-k+1}^T \\ \vdots \\ u_d^T \end{bmatrix},$$

we compute:

$$AX = \begin{bmatrix} \sigma_{d-k+1} v_{d-k+1}^T \\ \vdots \\ \sigma_d v_d^T \end{bmatrix}.$$

The Frobenius norm $\|AX\|_F^2 = \sum_{i=d-k+1}^{d} \sigma_i^2$ is minimized because the Eckart-Young-Mirsky theorem ensures that truncating to the smallest $k$ singular values yields the optimal low-rank approximation in the Frobenius norm. Any other choice of vectors would include larger singular values, increasing the norm. □

## 7.2. Additional Experiments:

| Error Metric | Value |
|---|---|
| MLRE using $S^{(1)\dagger}$ | $2.98 \times 10^{-16}$ |
| MLRE using $S^{(2)\dagger}$ | $6.47 \times 10^{-16}$ |

Table 2. Sanity check: MLRE when using pseudo-inverse of ground-truth transformation matrices $S^{(m)\dagger}$. These near-zero errors validate our MLRE metric's ability to detect perfect reconstruction.

**Robustness Test** We vary 3 parameters to determine how it affects these errors, the number of data points $n$, the dimension of the data $d$, and the size of the latent dimension $k$. We further evaluate the method when the generated data $\hat{\mathbf{x}}^{(m)}$ has standard Gaussian noise added to it. The results are shown in Fig. 4.

## 7.3. Notation Reference:

We construct a table of our notation in Tab. 3

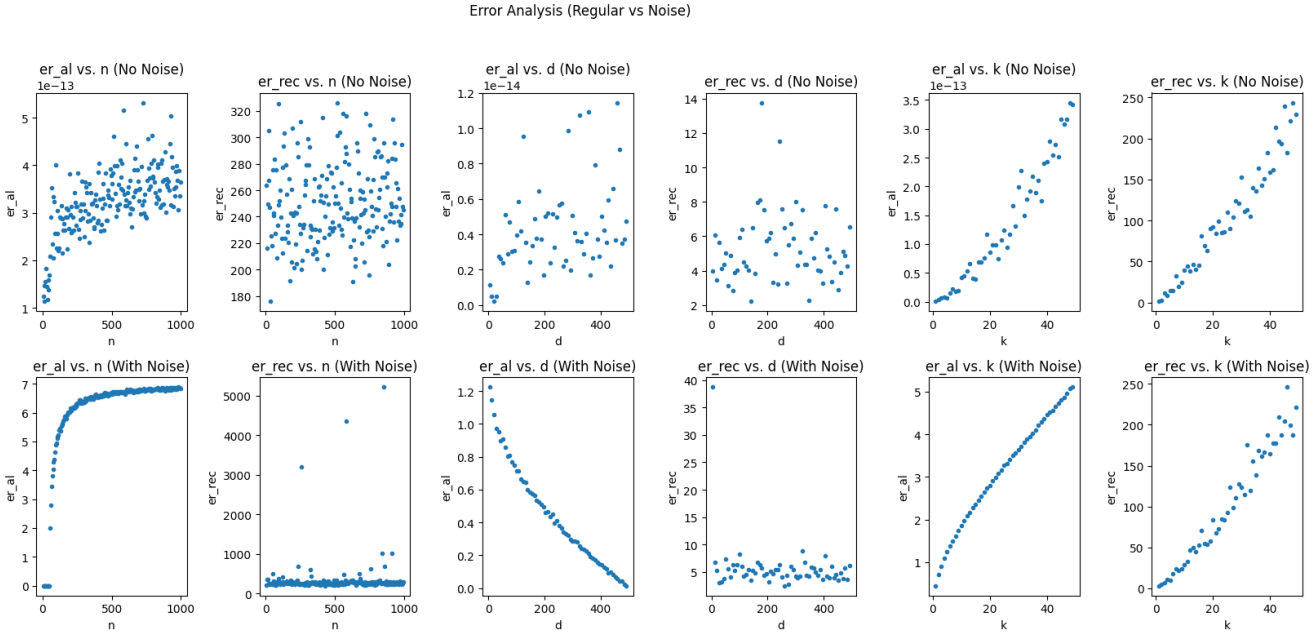


Figure 4. **Robustness Analysis:** CMAE/MLRE results using the proposed perfect alignment solver under varying parameters. (Columns 1-2) changes the number of data samples $n$, (Columns 3-4) varies the dimension of the data $d$, (Columns 5-6) alters the size of the latent dimension $k$. The first row shows the results without noise, the second row shows the results when the generated data has standard gaussian noise added to it..

Table 3. Notation Reference

| **Symbol** | **Description** | **Domain/Type** |
|---|---|---|
| $\mathcal{X}^{(m)}$ | Input space of modality $m$ | Vector space |
| $\mathcal{Z}$ | Shared latent space | Vector space |
| $\mathbf{x}_i^{(m)}$ | Data point $i$ from modality $m$ | $\mathbb{R}^{d_m}$ (column vector) |
| $\mathbf{S}^{(m)}$ | Ground-truth transformation matrix for modality $m$ | $\mathbb{R}^{d_m \times k}$ |
| $\mathbf{z}_i$ | Latent concept vector for data point $i$ | $\mathbb{R}^k$ |
| $f^{(m)}$ | Encoder function for modality $m$ | $\mathcal{X}^{(m)} \to \mathcal{Z}$ |
| $\mathbf{A}^{(m)}$ | Learned projection matrix for modality $m$ | $\mathbb{R}^{k \times d_m}$ |
| $\mathbf{X}^{(m)}$ | Data matrix for modality $m$ | $\mathbb{R}^{d_m \times n}$ |
| $\mathbf{A}$ | Combined projection matrix | $\mathbb{R}^{k \times d}$ $(d = \sum d_m)$ |
| $\mathbf{X}$ | Stacked data matrix | $\mathbb{R}^{d \times n}$ |
| $\mathbf{0}$ | Zero matrix in $\mathbf{AX} = \mathbf{0}$ | $\mathbb{R}^{k \times n}$ |
| $\mathbf{U}, \mathbf{\Sigma}, \mathbf{V}^T$ | SVD components of $\mathbf{X}$ | $\mathbf{U} \in \mathbb{R}^{d \times d}$, $\mathbf{\Sigma} \in \mathbb{R}^{d \times n}$, $\mathbf{V} \in \mathbb{R}^{n \times n}$ |
| $\pi_1, \pi_2$ | Mixture weights for GMM | $\pi_1 + \pi_2 = 1$ |
| $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ | GMM mean vectors | $\mathbb{R}^2$ |
| $\mathbf{\Sigma}_1, \mathbf{\Sigma}_2$ | GMM covariance matrices | $\mathbb{R}^{2 \times 2}$ |
| $n$ | Number of data points | $\mathbb{N}$ |
| $d_m$ | Dimension of modality $m$ | $\mathbb{N}$ |
| $d$ | Combined data dimension | $d = \sum d_m$ |
| $k$ | Latent space dimension | $\mathbb{N}$ |
| $---$ | Matrix concatenation operator | – |
| $\mathbf{I}_d$ | Identity matrix | $\mathbb{R}^{d \times d}$ |