Principled Multimodal Representation Learning

Xiaohao Liu   **Xiaobo Xia*** **See-Kiong Ng**   **Tat-Seng Chua**
National University of Singapore
xiaohao.liu@u.nus.edu   xbx@nus.edu.sg   seekiong@nus.edu.sg   dcscts@nus.edu.sg

ABSTRACT

Multimodal representation learning seeks to create a unified representation space by integrating diverse data modalities to improve multimodal understanding. Traditional methods often depend on pairwise contrastive learning, which relies on a predefined anchor modality, restricting alignment across all modalities. Recent advances have investigated the simultaneous alignment of multiple modalities, yet several challenges remain, such as limitations imposed by fixed anchor points and instability arising from optimizing the product of singular values. To address the challenges, in this paper, we propose Principled Multimodal Representation Learning (PMRL), a novel framework that achieves simultaneous alignment of multiple modalities without anchor dependency in a more stable manner. Specifically, grounded in the theoretical insight that full alignment corresponds to a rank-1 Gram matrix, PMRL optimizes the dominant singular value of the representation matrix to align modalities along a shared leading direction. We propose a softmax-based loss function that treats singular values as logits to prioritize the largest singular value. Besides, instance-wise contrastive regularization on the leading eigenvectors maintains inter-instance separability and prevents representation collapse. Extensive experiments across diverse tasks demonstrate PMRL's superiority compared to baseline methods. The source code will be publicly available.

1 Introduction

Humans perceive the world through a rich interplay of multimodal signals, integrating visual, auditory, textual, and tactile information to form cohesive representations of individual instances [1, 2, 3, 4, 5, 6, 7]. These modalities capture both shared and distinct concepts, completing one instance while enabling the differentiation from another. Inspired by this capability, multimodal representation learning (MRL) seeks to align diverse modalities within a unified space [8, 9, 10, 11, 12, 13, 14], where a representation from one modality can effectively retrieve or reconstruct corresponding representations from others.

Bimodal alignment, often achieved with contrastive learning [15, 16, 17], aligns one modality with another by comparing synthetic modality pairs [18, 19, 8, 20]. This paradigm demonstrates remarkable performance in tasks like image-text or audio-text understanding. Such success can also be replicated in MRL. Modality-binding methods, exemplified by ImageBind [10], designate one modal anchor as the centroid and adopt pairwise contrastive learning to align other modalities to it [11, 21, 22, 12, 23], as shown in Figure 1(left). Anchored alignment (*e.g.*, $m_1 \rightarrow m_3$ and $m_4 \rightarrow m_3$) is explicitly modeled, while the alignment among non-anchor modalities (*e.g.*, $m_1 \rightarrow m_4$) remains implicit. A branch of work proposes exploiting scaling data for pre-training [24, 25, 26, 11, 27, 28, 21], or introducing
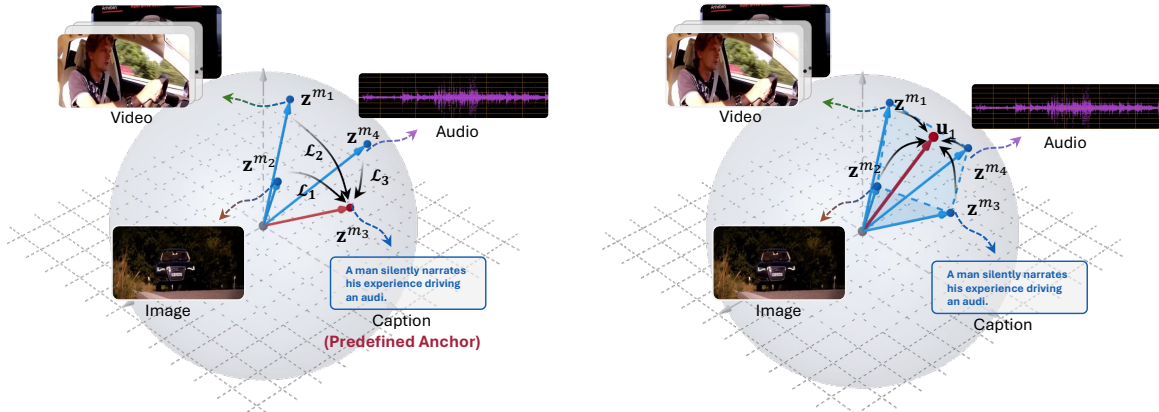
---

*Corresponding author.

Figure 1: **The illustration of multimodal representations within a hypersphere.** The left demonstrates pairwise contrastive learning to align multiple modalities with a predefined anchor (*i.e.*, caption), where modalities are sampled to multiple pairs. The right illustrates our method that aligns all modalities simultaneously with a leading direction.

auxiliary learning objectives, like language modeling loss [29, 30, 27] to improve MRL. Unfortunately, they remain reliant on pairwise contrastive learning, which keeps the alignment hinged with anchors.

A most recent work attempts to move beyond this holding paradigm by minimizing the volume of a parallelotope formed by multimodal representations [31]. It utilizes the determinant of the Gram matrix (numerically equal to the product of singular values) and interprets simultaneous alignment for all modalities in a geometric space. Unfortunately, it still depends on *predefined anchors* to construct negative instances (*i.e.*, replace the content of anchor modality to yield an unmatched multimodal instance). Moreover, its optimization on volume suffers from *instability*. Specifically, when the parallelotope collapses to a plane, optimization halts as the volume reaches zero, resulting in incomplete alignment. We also discuss this via singular value analysis (see Section 4.4). These limitations underscore the need for a more advanced MRL method with sound principles, motivating our framework development for multimodal alignment.

In this work, we initiate our research from the foundational goal of multimodal alignment, aiming to maximize the similarity between any modality pairs of a shared instance. This leads to a critical insight: establishing a fundamental connection between multimodal alignment and the rank of the Gram matrix, where full alignment is achieved when the rank equals one. This principle guides our development of a novel method for multimodal representation learning through rank-1 matrix approximation. To advance this, we propose strengthening the maximum singular value to encourage full alignment, drawing on the optimal low-rank approximation theory. The maximum singular value corresponds to a leading direction (*i.e.*, the dominant eigenvector), specialized for different instances. As this singular value increases, multimodal representations are aligned toward this direction adaptively, as depicted in Figure 1(right). This leading direction drifts with data itself, rather than privileging any one of the modalities. Motivated and implemented by the principle, we term our method as Principled Multimodal Representation Learning (PMRL) to highlight its theoretical grounding and pioneering design. PMRL removes anchor constraints, elevating any-to-one alignment to a straightforward *any-to-any* alignment for MRL. In addition, optimizing the maximum singular value relative to their sum also ensures *greater stability* than the previous volume-based method.

To this end, we propose a novel learning objective that directly aligns all modalities by optimizing singular values. Specifically, we employ a softmax-based loss that treats the singular values as logits and emphasizes the dominance of the maximum singular value. Besides, we incorporate instance-wise contrastive regularization over leading eigenvectors. These vectors serve as alignment centroids and are regularized to ensure inter-instance separability and prevent representation collapse. We verify the proposed PMRL with extensive experiments and demonstrate its superiority compared to baselines.

Before delving into details, we summarize our contributions as follows:

- We introduce Principled Multimodal Representation Learning (PMRL), a novel framework that encourages full alignment across multiple modalities simultaneously without relying on a predefined anchor modality. Our method is grounded in a theoretical connection between the singular values of multimodal representations and their full alignment under rank-1 approximation.

- To operationalize this insight, we reformulate the learning objective to strengthen the dominance of the maximum singular value, promoting full alignment across modalities, and incorporate instance-wise regularization to enhance inter-instance separability. By optimizing the maximum singular value, our method stabilizes the learning compared to directly reducing the products of singular values (*i.e.*, determinant-based volume).

- Extensive experiments on diverse tasks, including text-video retrieval, text-audio retrieval, and downstream classification, demonstrate PMRL's superior performance. Note that PMRL is capable of enhancing representations for broader fields, like medical applications (*e.g.*, autism diagnosis). Comprehensive analyses, including ablation studies, singular value analysis, regularization effects, noise robustness, and modality contribution, validate the efficacy and design rationale of PMRL, establishing its potential for advancing multimodal learning across varied applications.

2 Related Work

2.1 Multimodal Representation Learning

Multimodal representation learning begins with building connections between vision and language modalities (bi-modal representation learning) [9, 32, 33, 34]. Particularly, CLIP [9] learns deep semantic representations by matching vision concepts to linguistic inputs. This paradigm inspires a series of works to extend more modalities, *e.g.*, audio-to-text [35], point-to-text [20], and video-to-text [19, 36]. These methods utilize pairwise contrastive learning [16, 15, 37] to align two modality representations closer if they are from the same instance, while pushing away otherwise, thus building a joint embedding space. Building upon this bimodal paradigm, recent works introduce more modalities into a unified foundation model [38, 27, 31, 22, 23, 39, 13, 12, 10, 11, 21]. Subtitles [40, 41, 42] and audio [43, 8, 44, 45] are introduced and modeled together with vision and text. More training objectives, like next utterance prediction [41], masked prediction [40], and modality pair matching [29, 27], are adopted to further enhance the performance. Notably, VAST [27] pioneers the omni-modality foundation model, involving vision, audio, subtitle, and text. Alongside these innovations, ImageBind [10] builds upon CLIP and binds multiple modality representations, with vision modality being the anchor. By setting an anchor modality (*e.g.*, language [11], vision [10], and point cloud [22]), all the modalities will be aligned together through interactively contrastive learning. However, this privileging of one modality enforces a fixed representation to frame all modalities (*i.e.*, single-modal centrism), inherently ignoring the complexity and richness of multimodalities.

Differently, GRAM [31] proposes to align multimodal representations simultaneously, which spans a parallelotope, and minimizes its volume to achieve simultaneous alignment for all modalities. Nevertheless, its learning objective is implemented by contrastive learning, where the text modality serves as the anchor and is replaced to construct negative samples. A predefined anchor is still relied upon by GRAM. Additionally, the volume collapses when a singular value approaches zero, leading to unstable optimization. Our method goes beyond this by strengthening the maximum singular value, aligning all modalities to the leading direction automatically, and theoretically revealing its potential to achieve full alignment.

3

2.2 Principled Learning with SVD

Singular value decomposition (SVD) is a fundamental matrix factorization technique [46, 47, 48, 49, 13] with broad applications in machine learning [50, 51, 52]. SVD has been extensively utilized in domains such as image processing [53, 54, 55], model compression [56, 18, 57, 58], and parameter initialization [59]. Despite its versatility, the potential of SVD in multimodal learning remains underexplored, presenting opportunities for novel applications and theoretical advancements. Notably, contrastive loss minimization by gradient descent can be formulated as the SVD on a contrastive cross-covariance matrix, establishing the connection between SVD and multimodal contrastive learning [60]. In addition to the theoretical analysis, recent work [61] leverages SVD to construct the linear transformation from modality to representation, while being limited by a bimodal scenario. In this work, we further exploit the SVD analysis and establish a formal connection between singular values and full alignment in multimodal representation learning, which introduces a novel method that builds upon this theoretical insight.

3 Preliminary

Bimodal alignment. Given multimodal data $\mathcal{X} = \{\mathbf{x}_i | i \in \mathbb{Z}^+, i < N\}$, $\mathbf{x}_i$ is a multimodal instance containing k modalities, denoting $\mathbf{x}_i = \{\mathbf{x}_i^m | m \in \mathcal{M}\}$. Multimodal representation learning aims to learn the latent representation $\mathbf{z}_i^m \in \mathbb{R}^{d \times 1}$ from the corresponding multimodal data $\mathbf{x}_i^m$ through the encoder. The latent representations $\{\mathbf{z}_i^m | m \in \mathcal{M}\}$ from the common instance i are expected to be *similar*, in other words, being retrievable from each other. Cross-modal retrieval offers insights where two modalities (*e.g.*, m_1 and m_2) are aligned with pairwise contrastive learning, and the similarity is defined as the inner product between their representations:

$$\mathcal{L}^{m_1, m_2} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp(\mathbf{z}_i^{m_1 \top} \mathbf{z}_i^{m_2} / \tau)}{\sum_j^N \exp(\mathbf{z}_i^{m_1 \top} \mathbf{z}_j^{m_2} / \tau)}, \tag{1}$$

where τ is the temperature ratio and N denotes the number of data pairs. This bimodal alignment objective is also widely adopted for multimodal representation learning, including [11, 27, 10].

Multimodal alignment. The pairwise contrastive learning paradigm can also be extended to the multimodal scenario ($k > 2$). For example, the training on $\{m_1, m_2, m_3\}$ can be decomposed to the training sequences of $\mathcal{L}^{m_1, m_2}$, $\mathcal{L}^{m_1, m_3}$, and $\mathcal{L}^{m_2, m_3}$ [10, 27][2]. GRAM [31] proposes to project all modalities to form a parallelotope with a small volume, which can be defined by the determinant of the Gram matrix, *i.e.*, $\det(\mathbf{G}) = \det(\mathbf{Z}^\top \mathbf{Z})$. The performance improvement induced by aligning all modalities simultaneously motivates us to dive deeper into it.

SVD and eigenvalues. Given an arbitrary matrix $\mathbf{X} \in \mathbb{R}^{n \times n'}$, we have $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ via SVD. Here $\mathbf{U} \in \mathbb{R}^{n' \times n'}$ and $\mathbf{V} \in \mathbb{R}^{n \times n}$ are unitary matrices, satisfying $\mathbf{U}\mathbf{U}^\top = \mathbf{1}$ and $\mathbf{V}\mathbf{V}^\top = \mathbf{1}$. Besides, $\mathbf{\Sigma} \in \mathbb{R}^{n' \times n}$ is the matrix with non-negative entries on the diagonal and zeros off the diagonal. The diagonal ones (singular values) can be represented as $\sigma_1 \geq \sigma_2 \geq \cdots \geq 0$, square roots of the eigenvalues of $\mathbf{X}^\top \mathbf{X}$. The maximum eigenvalue $\lambda_1 = \sigma_1^2$ corresponds to the dominant singular direction.

4 Principled Multimodal Representation Learning

4.1 Principled Learning

Alignment goal. Given normalized representations $\{\mathbf{z}_i^m \mid m \in \mathcal{M}\}$ derived from a shared instance i, the alignment is typically quantified via pairwise inner products (*e.g.*, cosine similarity) [9, 10, 11, 27]. Therefore, the objective is to

---

[2]Here we do not highlight the asymmetric property of $\mathcal{L}^{m_1, m_2}$.

maximize such similarity: $\arg\max a_{\boldsymbol{\theta}}^{m_i,m_j} := (\mathbf{z}^{m_i})^\top \mathbf{z}^{m_j}, \forall\{m_i, m_j\} \subseteq \mathcal{M}^3$, where $\boldsymbol{\theta}$ denotes related parameters to be optimized. We can also express this in a matrix form as:

$$\underset{\boldsymbol{\theta}}{\operatorname{argmax}} \|\mathbf{G}\|_F = \sqrt{\sum_{i,j} |a^{m_i,m_j}|^2}, \quad \mathbf{G} = \begin{bmatrix} 1 & a^{m_1,m_2} & \cdots & a^{m_1,m_k} \\ a^{m_2,m_1} & 1 & \cdots & a^{m_2,m_k} \\ \vdots & \vdots & \ddots & \vdots \\ a^{m_k,m_1} & a^{m_k,m_2} & \cdots & 1 \end{bmatrix}, \tag{2}$$

where $\|\cdot\|_F$ is the Frobenius norm. The ideal case is that $\mathbf{z}^{m_1} = \mathbf{z}^{m_2} = \cdots = \mathbf{z}^{m_k}$, inducing maximum $\|\mathbf{G}\|_F$. Every entry in $\mathbf{G}$ equals 1. Notably, to avoid the extreme case where all the encoded representations are aligned with a common vector, there is typically an additional regularization, $a_{i,j} < a_{i,i}$, across different instances. We leave this discussion in Section 4.3.

Assumption 1. For a common instance, the alignment scores between pairs of modalities are nonnegative, *i.e.*, $a^{m_i,m_2} \geq 0, \forall\{m_1, m_2\} \subseteq \mathcal{M}$. This implies that the angle formed by any paired multimodal representations is not obtuse if they are sourced from the same instance [31].

In the following, we first draw the connection between multimodal full alignment and the rank of the Gram matrix (*cf.*, Lemma 1). Afterward, combined with the optimal rank-*r* approximation (*cf.*, Lemma 2), we derive our principled learning theory (*cf.*, Theorem 2) that motivates us to strengthen the maximum singular value to approach full alignment.

① **Alignment and** $\operatorname{rank}(\mathbf{G})$**.** Considering the maximum $\|\mathbf{G}\|_F$, the ultimate goal, it satisfies that every element in $\mathbf{G}$ equals 1, and meanwhile $\operatorname{rank}(\mathbf{G}) = 1$. We have the following equivalence lemma.

Lemma 1 (Full alignment $\Leftrightarrow$ Rank-1 Gram matrix)**.** *Let* $\mathbf{G} \in \mathbb{R}^{k \times k}$ *be a Gram matrix constructed from normalized modality representations* $\{\mathbf{z}^m\}_{m=1}^k$, *i.e.,* $\mathbf{G}_{i,j} = \langle \mathbf{z}^{m_i}, \mathbf{z}^{m_j} \rangle$ *with* $\|\mathbf{z}^{m_i}\| = 1$. *Then the following are equivalent: (1)* $\mathbf{G}_{i,j} = 1$ *for all* i, j, *and (2)* $\operatorname{rank}(\mathbf{G}) = 1$. *See proof in Appendix A.1.*

Remark. The proposed lemma establishes a fundamental connection between multimodal alignment and the rank of the Gram matrix. Therefore, we can transform the problem of multimodal alignment into achieving a rank-1 Gram matrix. A recent paper [31] investigates the connection between the determinant of the Gram matrix and geometric interpretation. However, it fails to achieve full alignment because its objective can be satisfied with a collapsed dimension. For a deeper understanding, we explore the connections between singular values and this work, and highlight the superiority of our method in Section 4.4. This equivalence offers a novel perspective that potentially inspires future research toward achieving full multimodal alignment.

② **Alignment and** σ_1**.** The goal is transformed to learn multimodal representations that yield the Gram matrix with rank 1. We propose a novel solution via SVD by maximizing the maximum singular value σ_1, supported by the following analysis (see Lemma 2 and Theorem 2).

Lemma 2 (Eckart-Young [62])**.** *The optimal rank-r approximation to* $\mathbf{X}$, *in a least-squares sense, is given by the rank-r SVD truncation* $\tilde{\mathbf{X}}$:

$$\underset{\tilde{\mathbf{X}}, s.t. \operatorname{rank}(\tilde{\mathbf{X}})=r}{\operatorname{argmin}} \|\mathbf{X} - \tilde{\mathbf{X}}\|_F = \tilde{\mathbf{U}}\tilde{\mathbf{\Sigma}}\tilde{\mathbf{V}}^\top. \tag{3}$$

Here $\tilde{\mathbf{U}}$ *and* $\tilde{\mathbf{V}}$ *denote the first r leading columns of* $\mathbf{U}$ *and* $\mathbf{V}$, *and* $\tilde{\mathbf{\Sigma}}$ *contains the leading* $r \times r$ *sub-block of* $\mathbf{\Sigma}$.

Remark. Lemma 2 reveals that the low-rank approximation can be optimally achieved via SVD, motivating a branch of work utilizing it for model compression [56, 18]. Despite the inspiration, in our context, the goal is to minimize $\|\mathbf{G} - \tilde{\mathbf{G}}\|_F$, where $\operatorname{rank}(\tilde{\mathbf{G}}) = 1$, by optimizing the learnable $\mathbf{Z}$. The optimal low-rank matrix ($\tilde{\mathbf{G}}$) is found, while $\mathbf{G} = \mathbf{Z}^\top \mathbf{Z}$ is under-resolved.

---

[3]We use superscript to denote modality indices and subscript for instance indices. Omitting the subscript indicates the same instance.
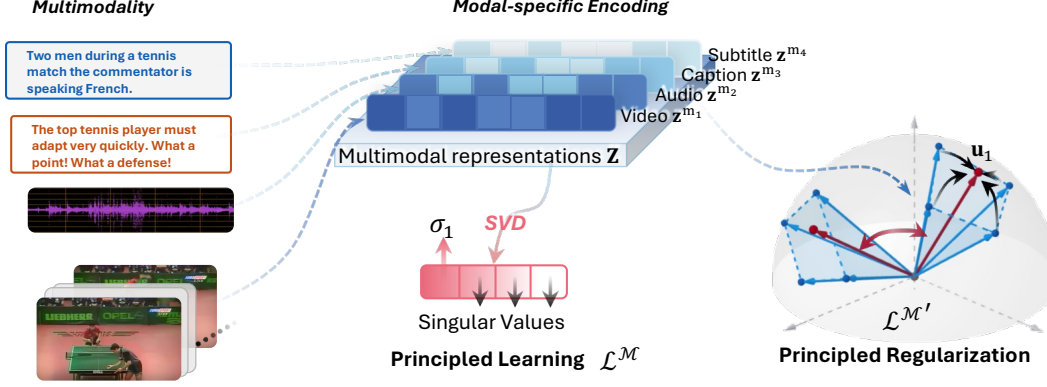
Figure 2: **The overall framework of PMRL.** Different modalities of the instance are encoded into multimodal representations $\mathbf{Z}$. PMRL utilizes SVD to obtain the maximum singular value σ_1 and maximizes it with the objective $\mathcal{L}^{\mathcal{M}}$. The leading directions (arrows in red) corresponding to σ_1 from different instances are regularized by $\mathcal{L}^{\mathcal{M}'}$.

Theorem 2 (Principled learning). *Let $\mathbf{Z} = [\mathbf{z}^{m_1}, \ldots, \mathbf{z}^{m_k}] \in \mathbb{R}^{d \times k}$ be a matrix of normalized modality representations from the same instance, i.e., $\|\mathbf{z}^{m_i}\| = 1$ for all i, and let σ_1 denote its maximum singular value. Then, we have (1) maximizing σ_1 maximizes the pairwise cosine similarities among $\{\mathbf{z}^m\}_{m=1}^k$, and (2) $\mathrm{rank}(\mathbf{G}) = 1$ is achieved if and only if $\sigma_1 = \sqrt{k}$. See proof in Appendix A.2.*

Remark. σ_1 reflects the strength of the leading direction of $\mathbf{u}_1$. By maximizing the σ_1, subject to the constraint $\sum_{i=1}^k \sigma_i^2 = k$, other singular values are minimized, finally aligning all representations $\mathbf{z}^m$ with the leading direction. Intuitively, this process adaptively identifies an optimal anchor for alignment at each training step, drawing all representations toward a common centroid. Figure 2 illustrates this concept for clarity.

4.2 Singular Value Maximization for Multimodal Alignment

Building upon the theoretical insights into multimodal alignment through Gram matrices and their spectral properties, we propose a novel learning objective that directly encourages alignment among heterogeneous modality representations. For each instance, we collect normalized embeddings from all available modalities and construct a compact representation matrix $\mathbf{Z}$. We apply SVD to extract the maximum singular value σ_1, which reflects the strength of the dominant alignment direction across modalities:

$$\mathrm{SVD}(\mathbf{Z}) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}, \quad \boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_k). \tag{4}$$

To enhance the prominence of the leading singular value during training, we introduce a softmax-based loss over the singular values:

$$\mathcal{L}^{\mathcal{M}} = -\frac{1}{N}\sum_{i=1}^N \log \frac{\exp[\sigma_1/\tau]}{\sum_j^k \exp[\sigma_j/\tau]}, \tag{5}$$

where τ is a temperature parameter. This formulation treats the singular values as logits and encourages σ_1 to stand out relative to the rest, thereby promoting strong alignment. The analysis on gradient propagation for improving alignment via singular value maximization is detailed in Appendix A.3. Below are the key insights that reveal its deeper significance. (1) Unlike contrastive objectives, which optimize local similarity at the bimodal-level $(\mathbf{z}^{m_1})^\top \mathbf{z}^{m_2}$, this loss operates at the level of global covariance structure of $\mathbf{Z}$. PMRL goes beyond conventional contrastive learning [9], shifting MRL from isolated pairs to the collective behavior of modalities. (2) Without predefined anchors, the model aligns modalities along a latent leading direction emerging from σ_1. By continuously amplifying the dominance through a differentiable competition among singular values, it fosters a self-discovering representation space.

6

4.3 Instance-wise Regularization

To prevent degenerate solutions where all embeddings collapse to a single point or become misaligned across instances, we incorporate *instance-wise contrastive regularization* that encourages separation between different instances:

$$\mathcal{L}^{\mathcal{M}'} = -\frac{1}{N} \sum_{i=1}^{N} \log \frac{\exp[(\mathbf{u}_1^{(i)})^\top \mathbf{u}_1^{(i)}/\tau]}{\sum_j^k \exp[(\mathbf{u}_1^{(i)})^\top \mathbf{u}_1^{(j)}/\tau]}, \tag{6}$$

where $\mathbf{u}_1$ corresponds to the maximum singular value σ_1, indicating the leading direction that all multimodal representations are aligned with. Furthermore, we employ the instance matching loss, which encourages the model to predict whether the multimodal data is matched or not as a binary question. The data obtained from all the encoders is concatenated and fed to a multimodal encoder. A two-layer multi-layer perceptron (MLP) serves as the predictor that returns $\hat{y}$, the matching probability. We follow [27, 31] with the hard negative mining strategy and employ the instance matching loss as follows:

$$\mathcal{L}_{\text{IM}} = \mathbb{E}_{(m_1, m_2, \ldots, m_k) \sim \mathcal{M}} [y \log \hat{y} + (1 - y) \log(1 - \hat{y})]. \tag{7}$$

The overall objective combines the alignment-driven singular value loss with auxiliary regularization terms:

$$\mathcal{L} = \mathcal{L}^{\mathcal{M}} + \lambda_1 \mathcal{L}^{\mathcal{M}'} + \lambda_2 \mathcal{L}_{\text{IM}}, \tag{8}$$

where λ_1 and λ_2 control the strength of regularizations. We set $\lambda_1 = 1$ and $\lambda_2 = 0.1$ by default.

4.4 Further Analysis

Connecting to the volume of the Gram matrix. Prior work [31] investigates minimizing the determinant of the Gram matrix, which can be interpreted geometrically, *i.e.*, the volume of the k-dimensional parallelotope formed by multimodal representations. Unfortunately, the theoretical connection between the volume and multimodal alignment is still unexplored. Here we highlight insights shared with GRAM [31] through singular value analysis while distinguishing our work through approaching full alignment. Specifically, the volume proposed by GRAM can be represented by the product of singular values as well:

$$\text{Vol}(\mathbf{G}) = \sqrt{\det \mathbf{G}} = \sqrt{\det \mathbf{Z}^\top \mathbf{Z}} = \prod_{i=1}^{k} \sigma_i. \tag{9}$$

Afterward, the volume achieves its minimum value of zero if and only if at least one of the singular values $\{\sigma_i\}_{i=1}^{k}$ is zero. In other words, by optimizing the minimum singular value σ_k to zero, the volume of the k-dimensional parallelotope reaches zero as well. In this case, the Gram matrix rank can remain larger than 1, preventing full alignment (see Lemma 1). We also illustrate the trend of singular values during training in Appendix C.3. Geometrically, this corresponds to the parallelotope *collapsing* to $k - 1$ dimensions. The collapsed volume is no longer optimized. In practice, GRAM is also achieved with a pre-defined anchor for contrastive learning. In comparison, we encourage the multimodal alignment in an *anchor-free* manner.

Robustness to noise. PMRL showcases robustness against noise in input data and labels, as supported by prior works [63, 64]. Noise, *e.g.*, Gaussian perturbations, disrupts the generation of multimodal representations, complicating alignment estimation. SVD effectively filters noisy data [65, 66] and recovers rank-1 matrices [63]. Additionally, noisy labels can destabilize learning processes [67, 68], but SVD-extracted low-rank matrices alleviate these negative effects [64, 69]. These findings highlight PMRL's robustness, approaching rank-1 matrices to promote full alignment.

5 Experiments

5.1 Experimental Setups

Datasets. VAST-150K [31], a downsized version of VAST-27M [27], is utilized for the multimodal training. This dataset involves four modalities, including vision, audio, subtitle, and text (*i.e.*, caption). For the downstream evaluation, we utilize MSR-VTT [70], DiDeMo [71], ActivityNet [72], and VATEX [73] for text-video retrieval, and AudioCaps [74] and Clotho [75] for text-audio retrieval. In addition, to demonstrate the broader potential of our method, we incorporate the ABIDE (Autism Brain Imaging Data Exchange) [76] dataset, a brain imaging dataset for autism classification, covering three modalities (*i.e.*, fMRI, sMRI, and text). PMRL is built upon VAST and employs a continual pre-training strategy to evaluate its effectiveness, following [31]. Therefore, we utilize VAST-150K to re-boost its zero-shot capabilities, and split downstream datasets for fine-tuning PMRL for specific tasks. All the downstream datasets involve over two modalities. See more details of the datasets in Appendix B.1.

Baselines and evaluation metrics. We select extensive baselines in our comparison, including Frozen [24], UMT [77], UMT-L [78], OmniVL [30], TVTSv2 [79], CLIP4Clip [36], ViCLIP [26], VideoCoCa [80], Norton [81], ImageBind [10], InternVideo-L [82], HiTeA [83], mPLUG-2 [84], VALOR-L [85], TEFAL [86], Bimodal T2M [87], T-MASS [88], vid-TLDR [89], VideoPrism-b [28], LanguageBind [11], AVFIC [25], VIP-ANT [90], VAST [27], and GRAM [31] (more details are shown in Appendix B.2). Wherein, GRAM serves as the state-of-the-art (SOTA) method. In this main comparison, we conduct the evaluation in the zero-shot setting. We also implement the fine-tuning setting in multimodal text-to-video retrieval, following [31]. We utilize *Recall* as the retrieval metric. Note that we implement VAST's evaluation algorithm, which uses a conventional cosine similarity-based method. For ABIDE, we use the training datasets to align new modalities (*e.g.*, fMRI and sMRI). Here we select AE-FCN [91], GCN [92], VanillaTF, BrainNetCNN [93], and BrainNetTF [94] as baselines for classification performance comparison. VAST [27] and GRAM [31] are also specialized for this task to ensure a fair comparison with our PMRL. *AUC* and *Accuracy* serve as the classification metric. The baselines and relevant metrics are detailed in Appendix B.2 and B.3.

Model architecture and hyperparameters. The PMRL model is built upon VAST [27] with the same architecture in our main comparison. Specifically, the vision, audio, and text encoders are implemented via EVAClip-ViT-G [95], BEATs [96], and BERT-B [97], respectively. We continue optimizing the parameters of VAST with our proposed objective. For autism evaluation, we implement our PMRL model (also VAST and GRAM) with BrainNetTF [94] for the fMRI modality, a multi-layer perceptron (MLP) module for the sMRI modality, and BERT for the textual modality. Built on generated representations, we add an MLP classifier. By default, we set the learning rate to 2×10^{-5}, the batch size to 64, and train the model for one epoch. We utilize AdamW [98] as the optimizer and a linear warmup scheduler. The experiments are conducted in a device equipped with $4 \times$NVIDIA H100-80GB GPUs. Detailed hyperparameter settings and model architecture design can be found in Appendix B.4 and Appendix B.5, respectively. We also provide the algorithm flow and pseudocode to facilitate reproducibility, as shown in Appendix B.6.

5.2 Main Results

In this subsection, we mainly explore the performance of PMRL via developed evaluations, like cross-modal retrievals, compared with existing state-of-the-art methods. We further showcase its broader impact for medical applications, manifested by the autism classification task.

Multimodal cross-modal retrieval. We evaluate the retrieval performance to indicate the alignments between modalities. It is well-established for several MRL methods, and typically focuses on text-video retrieval (*i.e.*, Table 1 for the zero-shot setting, while Table 2 for the fine-tuning setting), and text-to-audio retrieval (see Table 3). We follow the

---

[4]*Finetuning and evaluation with 12 frames.

Table 1: Multimodal text-to-video (T→V) and video-to-text (V→T) retrieval results (%) in the zero-shot setting, in terms of Recall@1 (R@1). Increment points are computed compared with VAST.

	MSR-VTT		DiDeMo		ActivityNet		VATEX	
	T→V	V→T	T→V	V→T	T→V	V→T	T→V	V→T
Fronzen [24]	18.7	-	21.1	-	-	-	-	-
UMT [77]	33.3	-	34.0	-	31.9	-	-	-
UMT-L [78]	40.7	37.1	48.6	49.9	41.9	39.4	-	-
OmniVL [30]	42.0	-	40.6	-	-	-	-	-
TVTSv2 [79]	38.2	-	34.6	-	-	-	-	-
ViCLIP [26]	42.4	41.3	18.4	27.9	15.1	24.0	-	-
VideoCoCa [80]	34.3	64.7	-	-	34.5	33.0	53.2	73.6
Norton [81]	10.7	-	-	-	-	-	-	-
ImageBind [10]	36.8	-	-	-	-	-	-	-
InternVideo-L [82]	40.7	39.6	31.5	33.5	30.7	31.4	49.5	69.5
HiTeA [83]	34.4	-	43.2	-	-	-	-	-
mPLUG-2 [84]	47.1	-	45.7	-	-	-	-	-
VideoPrism-b [28]	51.4	50.2	-	-	49.6	47.9	62.5	77.1
LanguageBind [11]	44.8	40.9	39.9	39.8	41.0	39.1	-	-
VAST [27]	50.5	48.8	46.4	45.3	51.7	48.8	75.9	74.8
GRAM [31]	51.5 (+1.0)	51.5 (+2.7)	49.8 (+3.4)	48.5 (+3.2)	54.5 (+2.8)	48.3 (-0.5)	77.5 (+1.6)	74.7 (-0.1)
PMRL (Ours)	**54.5 (+4.0)**	**52.4 (+3.6)**	**50.6 (+4.2)**	**48.4 (+3.1)**	**56.0 (+4.3)**	**49.6 (+0.8)**	**80.5 (+4.6)**	**75.2 (+0.4)**

Table 2: Multimodal text-to-video (T→V) and video-to-text (V→T) retrieval results (%) in the finetuning setting, in terms of Recall@1 (R@1). Increment points are computed compared with VAST[4].

	MSR-VTT		DiDeMo		ActivityNet		VATEX	
	T→V	V→T	T→V	V→T	T→V	V→T	T→V	V→T
UMT-L [78]	58.8*	58.6*	70.4*	65.7*	66.8*	64.4*	72.0*	86.0*
CLIP4Clip [36]	44.5	45.9	43.4	43.6	40.5	41.6	55.9	78.3
ViCLIP [26]	52.5	51.8	49.4	50.2	49.8	48.1	-	-
InternVideo-L [82]	55.2*	57.9*	57.9*	59.1*	62.2*	62.8*	71.1*	87.2*
HiTeA [83]	46.8	-	56.5	-	-	-	-	-
mPLUG-2 [84]	53.1	-	56.4	-	-	-	-	-
VALOR-L [85]	54.4	-	57.6	-	63.4	-	76.9	-
TEFAL [86]	52.0	-	-	-	-	-	61.0	-
Bimodal T2M [87]	36.8	-	-	-	-	-	-	-
T-MASS [88]	52.7	-	53.3	-	-	-	65.6	-
vid-TLDR [89]	58.5*	-	70.4*	-	65.2*	-	-	-
VAST [27]	64.4	64.3	68.4	65.4	68.1	65.4	83.1	81.3
GRAM [31]	60.0 (-4.4)	61.8 (-2.5)	68.7 (+0.3)	65.7 (+0.3)	67.6 (-0.5)	65.0 (-0.4)	82.5 (+0.6)	80.6 (-0.7)
PMRL (Ours)	**61.2 (-3.2)**	**60.7 (-3.6)**	**70.2 (+1.8)**	**66.4 (+1.0)**	**68.2 (+0.1)**	**66.4 (+1.0)**	**84.1 (+1.0)**	**83.4 (+2.1)**

conventional cosine-based similarity metric for retrieval evaluation [27][5]. From these results, we have the following observations. For text-video retrieval on four datasets, the PMRL model achieves substantial performance improvements, outperforming VAST by up to 5.3% in retrieval metrics. Furthermore, the results also showcase that PMRL surpasses GRAM in both settings. More results are detailed in Appendix C.1. For multimodal text-to-audio retrieval across two datasets, as shown in Table 3, PMRL brings up to 7.6% performance boost to VAST, and outperforms GRAM as well. Overall, the enhancement for the maximum singular value, the core objective of PMRL, brings performance boosts. The improvements indicate that a better multimodal representation can be learned from our proposed method. We can attribute it to our principled learning, exploring the fundamental goal of multimodal alignment and approaching it via resolving the algebraic problem. Specifically, we observe that GRAM performs worse compared to VAST in some cases. Its learning objective is specialized for volume as a measure of multimodal alignment, which is incompatible with the widely adopted cosine similarity. Despite the improvements brought by PMRL for most cases, we find that both GRAM and PMRL perform worse than VAST if we directly fine-tune the model on the MSR-VTT dataset. One possible reason is that MSR-VTT is cleaner and exhibits more manifest correlations between video and

---

[5]We adapt the baseline GRAM for cosine-based evaluation to ensure a fair comparison.

Table 3: Multimodal text-to-audio retrieval results (%) in the zero-shot setting, in terms of Recall@1 (R@1) and 10 (R@10) scores.

	AudioCaps		Clotho	
	R@1	R@10	R@1	R@10
AVFIC [25]	8.7	37.7	3.0	17.5
AVFIC [25]	10.6	45.2	-	-
VIP-ANT [90]	27.7	37.7	-	-
ImageBind [10]	9.3	42.3	6.0	28.4
LanguageBind [11]	19.7	67.6	16.7	52.0
VAST [27]	33.7	77.1	12.4	36.4
GRAM [31]	34.6 (+0.9)	77.4 (+0.3)	15.9 (+3.5)	43.6 (+7.2)
PMRL (Ours)	**36.1 (+2.4)**	**75.9 (-1.2)**	**16.8 (+4.4)**	**44.0 (+7.6)**

Table 4: Multimodal autism classification results (%) in terms of AUC and ACC.

	ABIDE	
	AUC	ACC
AE-FCN [91]	78.9	69.4
GCN [92]	60.0	56.8
VanillaTF	76.1	68.2
BrainNetCNN [93]	73.6	67.9
BrainNetTF [94]	78.7	70.6
VAST [27]	79.2	71.8
GRAM [31]	63.9	60.6
PMRL (Ours)	**80.5 (+1.8)**	**73.2 (+1.4)**

text modalities, which can be easily captured by vision-text specialized methods, like VAST. For datasets curated from wild (*e.g.*, DiDeMo), we can observe a relatively better performance of PMRL. We also provide more analysis, like ablation studies and any modality retrieval results in Section 5.3, which offers more insights about PMRL.

Multimodal autism classification. We demonstrate the broader impact of PMRL, especially focusing on multimodal autism classification. Table 4 provides the evaluation results on ABIDE concerning AUC and ACC metrics. We adopt multimodal representation learning objectives for the autism classification. Therefore, we introduce VAST, GRAM, and PMRL with a classification loss. Compared to previous methods, *e.g.*, BrainNetTF, more modalities benefit the performance improvements. Among these multimodal methods, PMRL outperforms others on both metrics (*e.g.,* 3.6% and 1.9% improvements *v.s.* VAST). Despite using modalities like fMRI and sMRI, PMRL shows its strong potential to enhance performance in more applications. We also observe a particularly low performance of GRAM when we conduct the training from scratch. Bolstered by the analysis on GRAM's volume collapse in Section 4.4, we can attribute it to its optimization leading to an incorrect direction to align multimodal representation, especially for the model with random initialization. We provide the singular value trends for both GRAM and PMRL in Appendix C.3 to illustrate PMRL's more stable and goal-oriented learning procedure.

5.3 Further Empirical Analysis

To elucidate PMRL, we perform a comprehensive analysis supported by further empirical results. We conduct an ablation study in PMRL's design and evaluate retrieval performance across any modalities. We also track changes in singular values during training and examine the efficacy of instance-wise regularization. We interpret modality contributions to alignment using eigenvectors. Finally, we evaluate PMRL's robustness to noise.

Ablation study. To evaluate the efficacy of our proposed PMRL, we conduct the ablation study on our core designs, including principled learning on singular values ($\mathcal{L}^{\mathcal{M}}$) and principled regularization
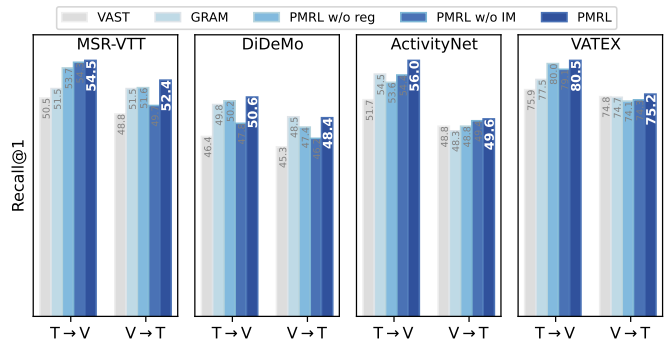

Figure 3: **The ablation study across 4 datasets in terms of Recall@1.** The instance-wise regularization loss (PMRL w/o reg) and instance matching loss (PMRL w/o IM) are canceled from PMRL and then compared with VAST and GRAM.

($\mathcal{L}^{\mathcal{M}'}$). We report the results on four datasets, as shown in Figure 3. Without regularization (*i.e.*, w/o reg or w/o IM), PMRL's performance declines across all scenarios. The integration of the proposed objectives yields great synergy to enhance multimodal representations.
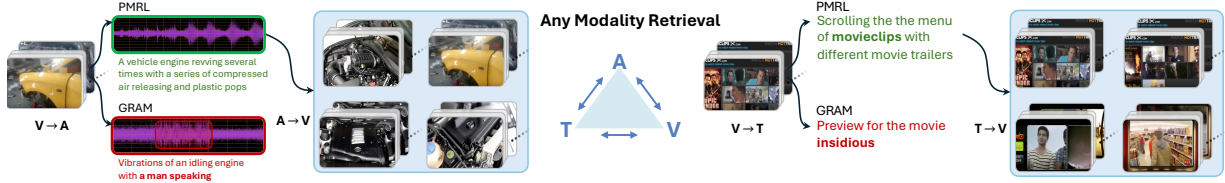
Figure 4: **The examples of any modality retrieval.** With a unified space, different modalities can retrieve others. PMRL is capable of retrieving from any modality pair with higher accuracy.
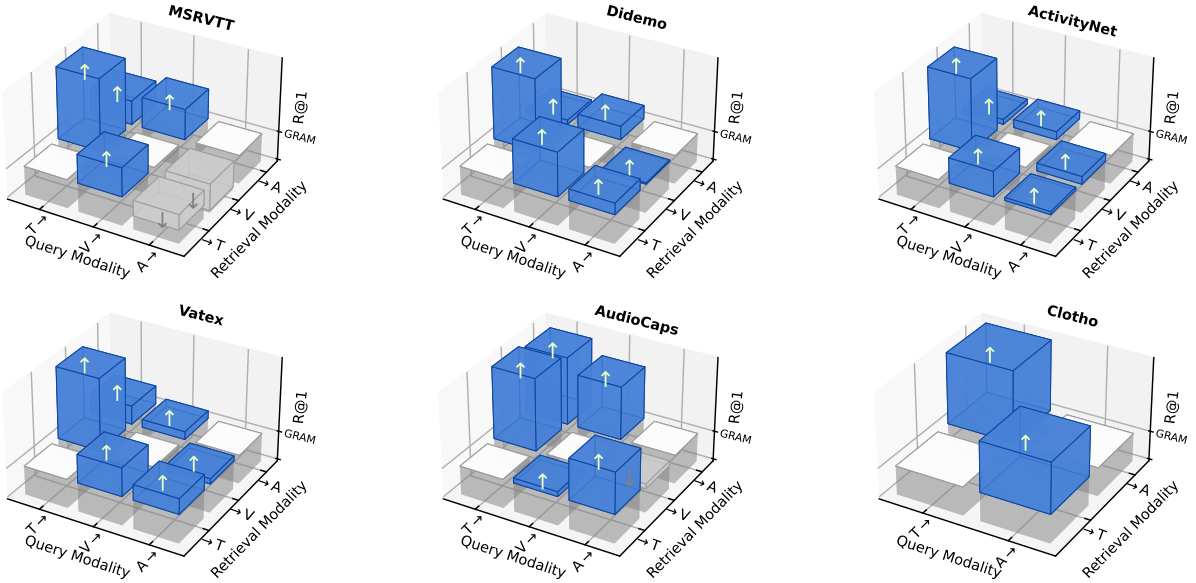

Figure 5: **Performance comparison for any modality retrieval across 6 benchmark datasets.** PMRL is compared with GRAM in terms of Recall@1. Blue regions highlight where PMRL outperforms GRAM, while gray regions indicate the opposite. Diagonal regions (colored in white) represent self-modal retrieval, which is not meaningful.

Any modality retrieval. PMRL is capable of encouraging full alignment without a predefined anchor, making it more stable for retrieval between any modalities, exemplified by Figure 4. We analyze the retrieval results among different modality pairs, *e.g.*, vision-audio, compared with GRAM, as illustrated in Figure 5. Compared to GRAM, we achieve higher performance for any modality retrieval (colored in blue) in most cases. The retrieval performance is not only greatly improved in text-related modalities, but also in other modalities. For instance, the performance on V→A retrieval boosts for all datasets, especially for AudioCaps. Due to the limited page, we provide more detailed results in terms of Recall@5 and Recall@10 in Appendix C.2.

Maximum singular value. We illustrate the changes of the maximum singular values along with the training procedure in Figure 6a. The result suggests an increasing trend, which can be attributed to our proposed principled learning objective. The singular value reaches a plateau afterward, indicating the convergence of the training.

Instance-wise regularization. We also investigate the effectiveness of principled regularization, intuited by keeping instances away, in terms of the leading eigenvector. To this end, we first measure the cosine similarity among leading eigenvectors (depicted in Figure 6b). Initially, the optimization is unstable, but continual regularization can still ensure its decrease, thereby enhancing the separability between instances. GRAM also implicitly introduces instance-wise regularization by comparing the volumes among instances. To isolate its impact, we modify GRAM to exclude this regularization, directly minimizing volume as $\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \text{Vol}(\mathbf{Z}_i)$. Results, shown in Table 5, reveal a performance drop for both PMRL models, underscoring the importance of instance-wise regularization. Note that GRAM exhibits more obvious degradation, indicating greater instability without regularization.
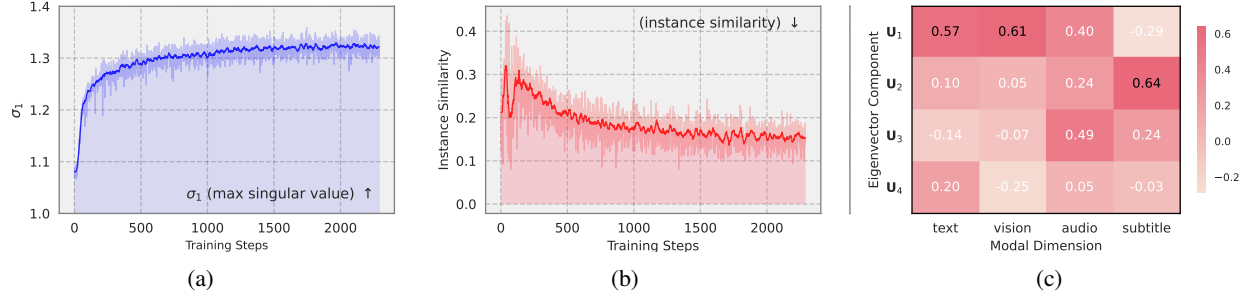
Figure 6: **Singular value analysis for PMRL.** Subfigure (a) illustrates the increase of the maximum singular value along with the training procedure induced by $\mathcal{L}^{\mathcal{M}}$. Subfigure (b) showcases the decrease of instance-wise similarity regularized by $\mathcal{L}^{\mathcal{M}'}$. Subfigure (c) depicts the contribution of each eigenvector to reconstruct the modal representation, which is interpreted by $\mathbf{V}$.

Table 5: The performance comparison without instance-wise regularization (w/o reg.) *w.r.t.* Recall@1 for T→V.

w/o reg.	MSR-VTT R@1	DiDeMo R@1	ActivityNet R@1	VATEX R@1
GRAM	50.9	40.2	20.1	58.7
PMRL	**53.7** (+2.8)	**50.2** (+10.0)	**53.6** (+33.5)	**80.0** (+21.3)

Table 6: The performance comparison with noise added to input and output (w/ noise) *w.r.t.* AUC and ACC.

w/ noise	VAST AUC	ACC	GRAM AUC	ACC	PMRL AUC	ACC
Input	71.5	64.4	61.0	57.4	**79.2** (+8.7)	**66.2** (+1.8)
Output	72.9	66.4	50.0	57.1	**77.2** (+4.3)	**70.3** (+3.9)

Modality contribution interpretation. PMRL also offers certain interpretability on modality contribution via SVD analysis, which is a core technique of our method. SVD decomposes the multimodal representation matrix $\mathbf{Z}$ into $\mathbf{U\Sigma V}$, where $\mathbf{U}$ represents transformed directions (eigenvectors), $\mathbf{\Sigma}$ contains singular values indicating the importance of each direction, and $\mathbf{V}$ shows how these directions are allocated to reconstruct different modalities. Therefore, the contribution of each modality to alignment can be roughly measured by $\mathbf{V}$ if we focus on the modality relevance to the first eigenvector (*i.e.*, $\mathbf{U}_1$). To visualize this, we average the absolute values of $\mathbf{V}$ in terms of instances to create a confusion matrix, as shown in Figure 6c, which highlights the relationships between modalities and the eigenvectors. Observed from this confusion matrix, text (m_1) and vision (m_2) are strongly tied to the leading eigenvector $\mathbf{U}_1$. The audio modality also shares a large proportion with $\mathbf{U}_1$, suggesting it shares overlap with text and vision, though to a lesser extent. In contrast, subtitle modality mostly corresponds to the second eigenvector $\mathbf{U}_2$. These findings indicate the well-aligned bimodality, *i.e.*, vision and text in semantics, also revealing the interpretability of PMRL for multimodal representation learning.

Robustness analysis. We show the robustness of PMRL to noise from inputs and outputs following the discussion in Section 4.4. We conduct the controlled experiments by adding Gaussian noise scaled by 0.4 to the normalized input features and randomly flipping class labels with a probability of 0.3. The results on ABIDE are reported in Table 6. Despite performance degradation due to noise across all methods, PMRL consistently outperforms others, reflecting the robustness in principle of maximizing the maximum singular value to encourage full alignment.

6 Conclusion and Discussion

In this paper, we propose to strengthen the dominance of the maximum singular value about multimodal representations and distinguish the corresponding leading eigenvectors from instances to encourage full multimodal alignment. The proposed method is grounded on the theoretical insight that connects the multimodal alignment and the rank of Gram matrices. Novel learning objectives are afterward introduced to maximize the maximum singular value and regularize instance-wise separability. A series of empirical results demonstrates the effectiveness of our PMRL framework and further showcases its rational design.

Our work provides new opportunities for multimodal representation learning by reframing the full alignment problem to resolving a rank-1 approximation. The proposed novel paradigm eliminates anchor constraints, empowering the model to self-discover the leading direction for alignment adaptively. PMRL provides certain interpretability for modality contributions to alignment, and demonstrates robustness to noise. Modalities, such as MRI in the medical application, can also be handled well by PMRL with enhanced multimodal representations. However, in this work, we do not focus on balancing the alignment and modality distinctness, and PMRL requires the concurrency of modalities to construct a unified representation space. Built upon our insight on approaching full alignment, future work can explore (1) trading off the perfect alignment and distinctness of multimodal representations according to our theoretical grounding, (2) scaling up training data and model parameters to develop more powerful multimodal representations, and (3) incorporating emerging modalities into PMRL.

Appendix

A Theoretical Analysis

A.1 Proof of Lemma 1

Recall. Let $\mathbf{G} \in \mathbb{R}^{k \times k}$ be a symmetric Gram matrix with diagonal entries equal to 1, *i.e.*, $\mathbf{G}_{i,i} = 1$, and off-diagonal entries defined as $\mathbf{G}_{i,j} = \langle \mathbf{z}^i, \mathbf{z}^j \rangle$, where each $\mathbf{z}^i \in \mathbb{R}^d$ satisfies $\|\mathbf{z}^i\| = 1$. Then the following are equivalent: (1) $\mathbf{G}_{i,j} = 1$ for all i, j, and (2) $\mathrm{rank}(\mathbf{G}) = 1$.

Proof. $(1) \Rightarrow (2)$: Suppose that $\mathbf{G}_{i,j} = 1$ for all i, j, *i.e.*,

$$\mathbf{G} = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ 1 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 1 \end{bmatrix} = \mathbf{1}\mathbf{1}^\top, \quad \mathbf{1} = [1, 1, \ldots, 1]^\top \in \mathbb{R}^k. \tag{10}$$

Then $\mathbf{G}$ is the outer product of a single vector with itself, so it has rank at most 1. Since $\mathbf{G} \neq \mathbf{0}$, we conclude $\mathrm{rank}(\mathbf{G}) = 1$. This proves the first direction.

$(2) \Rightarrow (1)$: Suppose that $\mathrm{rank}(\mathbf{G}) = 1$. Then $\mathbf{G}$ can be written as an outer product of two vectors:

$$\mathbf{G} = \mathbf{u}\mathbf{v}^\top = c\mathbf{v}\mathbf{v}^\top, \tag{11}$$

for some $\mathbf{u}, \mathbf{v} \in \mathbb{R}^k$. Since $\mathbf{G}$ is symmetric, we have $\mathbf{u} = c\mathbf{v}$ for some scalar c. Because $\mathbf{G}$ is a Gram matrix, it is also positive semidefinite. Therefore, $c > 0$, and we can normalize $\mathbf{v}$ such that:

$$\mathbf{G} = \mathbf{v}\mathbf{v}^\top. \tag{12}$$

Since $\mathbf{G}_{i,i} = \langle \mathbf{z}^i, \mathbf{z}^i \rangle = \|\mathbf{z}^i\|^2 = 1$, we have $\mathbf{v}_i^2 = 1$, which implies $\mathbf{v}_i = \pm 1$. However, recall that $\mathbf{G}_{i,j} = \langle \mathbf{z}^i, \mathbf{z}^j \rangle = \mathbf{v}_i \mathbf{v}_j \geq 0$. Therefore, $\mathbf{v}_i$ must all have the same sign (they are all $+1$). We then obtain:

$$\mathbf{v} = \mathbf{1}, \quad \Rightarrow \quad \mathbf{G} = \mathbf{1}\mathbf{1}^\top \quad \Rightarrow \quad \mathbf{G}_{i,j} = 1, \quad \forall\, i, j. \tag{13}$$

$\square$

This equivalence captures the ideal case in multimodal alignment, where all modality representations from the same instance are perfectly aligned.

A.2 Proof of Theorem 2

Recall. Let $\mathbf{Z} = [\mathbf{z}^{m_1}, \ldots, \mathbf{z}^{m_k}] \in \mathbb{R}^{d \times k}$ be a matrix of normalized modality representations from the same instance, *i.e.*, $\|\mathbf{z}^{m_i}\| = 1$ for all i, and let σ_1 denote its maximum singular value. Then, (1) maximizing σ_1 maximizes the pairwise cosine similarities among $\{\mathbf{z}^m\}_{m=1}^k$, and (2) $\mathrm{rank}(\mathbf{G}) = 1$ is achieved if and only if $\sigma_1 = \sqrt{k}$.

Proof. According to the Eckart-Young theorem [62] (see Lemma 2), the optimal rank-1 approximation $\tilde{\mathbf{Z}}$ of $\mathbf{Z}$ in the Frobenius norm is $\tilde{\mathbf{Z}} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^\top$, and the corresponding approximation error is:

$$\|\mathbf{Z} - \tilde{\mathbf{Z}}\|_F^2 = \sum_{i=2}^k \sigma_i^2. \tag{14}$$

Since $\|\mathbf{Z}\|_F^2 = \sum_{i=1}^k \sigma_i^2 = k$ (due to normalization $\|\mathbf{z}^i\| = 1$), we have:

$$\max \sigma_1 \iff \min \sum_{i=2}^k \sigma_i^2 \iff \min \|\mathbf{Z} - \tilde{\mathbf{Z}}\|_F^2. \tag{15}$$

Therefore, maximizing σ_1 minimizes the rank-1 approximation error. The perfect alignment can be achieved with $\sigma_1 = \sqrt{k}$.

Sufficiency: If $\sigma_1 = \sqrt{k}$, then $\sigma_2 = \cdots = \sigma_k = 0$, meaning $\mathbf{Z}$ is exactly rank-1:

$$\mathbf{Z} = \sqrt{k}\,\mathbf{u}_1\mathbf{v}_1^\top, \quad \text{with } \mathbf{v}_1 = \frac{1}{\sqrt{k}}\mathbf{1}_k. \tag{16}$$

This implies $\mathbf{z}^1 = \mathbf{z}^2 = \cdots = \mathbf{z}^k = \mathbf{u}_1$, achieving perfect alignment.

Necessity: Conversely, if $\mathbf{z}^1 = \mathbf{z}^2 = \cdots = \mathbf{z}^k = \mathbf{u}_1$, then:

$$\mathbf{Z} = \mathbf{u}_1\mathbf{1}_k^\top, \tag{17}$$

which has $\sigma_1 = \sqrt{k}$ and $\sigma_2 = \cdots = \sigma_k = 0$. $\qquad\square$

The Gram matrix $\mathbf{G} = \mathbf{Z}^\top\mathbf{Z}$ has eigenvalues $\sigma_1^2 \geq \sigma_2^2 \geq \cdots \geq \sigma_k^2$. When $\sigma_1 \to \sqrt{k}$, $\mathbf{G} \to \mathbf{1}_k\mathbf{1}_k^\top$, meaning $\mathbf{z}^{i\top}\mathbf{z}^j \to 1$ for all i, j. Therefore, maximizing σ_1 maximizes pairwise cosine similarities.

A.3 Gradient Analysis

In this section, we provide the gradient analysis for the proposed singular value-based contrastive loss:

$$\mathcal{L}^{\mathcal{M}} = -\frac{1}{N}\sum_{i=1}^N \log \frac{\exp(\sigma_1/\tau)}{\sum_{j=1}^k \exp(\sigma_j/\tau)},$$

where σ_1 denotes the maximum singular value of the normalized representation matrix $\mathbf{Z} \in \mathbb{R}^{d \times k}$, constructed from k modality-specific embeddings of the same instance.

Let us define the softmax-normalized weights over the singular values as:

$$p_j = \frac{\exp(\sigma_j/\tau)}{\sum_{j'=1}^k \exp(\sigma_{j'}/\tau)} \to \mathcal{L}^{\mathcal{M}} = -\frac{1}{N}\sum_{i=1}^N \log p_1^{(i)}, \tag{18}$$

where $p_1^{(i)}$ denotes the softmax weight corresponding to the maximum singular value $\sigma_1^{(i)}$ of the i-th instance.

Instance-level. For simplicity, we focus on one instance and drop the subscript i. The generalization to multiple instances follows directly. Using the chain rule and the earlier result $\frac{\partial \sigma_j}{\partial \mathbf{Z}} = \mathbf{u}_j\mathbf{v}_j^\top$, we compute:

$$\frac{\partial \mathcal{L}^{\mathcal{M}}}{\partial \mathbf{Z}} = \sum_{j=1}^k \frac{\partial \mathcal{L}^{\mathcal{M}}}{\partial \sigma_j} \cdot \frac{\partial \sigma_j}{\partial \mathbf{Z}} \tag{19}$$

$$= \sum_{j=1}^k \frac{\partial \mathcal{L}^{\mathcal{M}}}{\partial \sigma_j} \cdot \mathbf{u}_j\mathbf{v}_j^\top \qquad \left(\frac{\partial \sigma_i}{\partial \mathbf{Z}} = \mathbf{u}_i\mathbf{v}_i^\top\right) \tag{20}$$

$$= \frac{1}{\tau}\left[(p_1 - 1)\mathbf{u}_1\mathbf{v}_1^\top + \sum_{j=2}^k p_j\mathbf{u}_j\mathbf{v}_j^\top\right] \qquad \left(\frac{\partial \mathcal{L}^{\mathcal{M}}}{\partial \sigma_j} = \frac{1}{\tau}\begin{cases} p_1 - 1, & j = 1 \\ p_j, & j > 1 \end{cases}\right). \tag{21}$$

This expression reveals how the gradient shapes the learning dynamics:

- The term $(p_1 - 1)\mathbf{u}_1\mathbf{v}_1^\top$ pulls the leading direction $\mathbf{u}_1$ stronger, encouraging all columns of $\mathbf{Z}$ to align along $\mathbf{u}_1$.
- The terms $p_j\mathbf{u}_j\mathbf{v}_j^\top$ for $j > 1$ act to suppress other directions, pushing the representation space into a lower-dimensional subspace aligned with $\mathbf{u}_1$.

Modality-level. Let us denote the m-th column of $\mathbf{Z}$ as $\mathbf{z}^m$, representing the embedding of the m-th modality. Then, the gradient of the loss with respect to $\mathbf{z}^m$ can be extracted from the above expression:

$$\frac{\partial \mathcal{L}^\mathcal{M}}{\partial \mathbf{z}^m} = \frac{1}{\tau} \sum_{j=1}^{k} \frac{\partial \mathcal{L}^\mathcal{M}}{\partial \sigma_j} \cdot \mathbf{u}_j \mathbf{v}_{jm}, \tag{22}$$

where $\mathbf{v}_{jm}$ is the m-th entry of the right singular vector $\mathbf{v}_j$. This implies that each modality's representation is updated proportionally to its projection onto the dominant singular direction $\mathbf{u}_1$, weighted by the softmax probability p_j.

B Implementation Details

B.1 Training and Benchmark Datasets

We employ the training dataset **VAST-150K** [31], which is sampled from VAST-27M [27], following the training setting of GRAM [31]. VAST-27M is sampled from the large-scale HD_VILA_100M corpus [99], involving diverse categories of music, gaming, education, entertainment, animals, and more. Four modalities, *i.e.*, video, audio, caption, and subtitle, are collected for each example. More than that, we adopt several benchmark datasets as follows:

- **MSR-VTT** [70] is a large-scale video description dataset comprising approximately 10,000 short video clips (10-20 seconds each) sourced from YouTube, totaling around 200,000 video-text pairs. Each clip is annotated with 20 human-generated English captions, covering diverse scenarios such as sports, music, and daily activities. In our experiment, we extract the audio, which serves as one of three modalities. We adopt the standard split.

- **DiDeMo** [71] focuses on localized video descriptions, containing about 10,000 videos sourced from Flickr. Each video is annotated with four textual descriptions tied to specific temporal segments, emphasizing semantic diversity and temporal localization. These four short sentences are concatenated and arranged in temporal order. The official split is used.

- **ActivityNet** [72] is a large-scale video dataset tailored for human activity recognition, comprising approximately 20,000 YouTube videos totaling around 648 hours. It covers 200 activity classes (*e.g.*, cooking and sports) with temporally annotated segments and associated descriptions. Approximately 3,000 videos are unavailable online. Therefore, we remove them for our evaluation with the adopted official split.

- **VATEX** [73] is a multilingual video description dataset containing about 41,000 10-second video clips derived from the Kinetics-600 dataset, which covers 600 human activity categories. There are also some unavailable videos online. We adopt the split following [100, 27] and exclude these examples for evaluation.

- **AudioCaps** [74] is a large-scale audio description dataset, featuring approximately 51,000 10-second audio clips sourced from AudioSet. Each clip is paired with 1-5 human-annotated English captions describing diverse sound events (*e.g.*, natural, human, or mechanical sounds). We evaluate text-audio retrieval, following the same split protocol by [101].

- **Clotho** [75] contains 6,974 (in its expanded version) audio clips (15-30 seconds each) sourced from Freesound, each annotated with 5 detailed English captions. By emphasizing complex and diverse sound scenes, Clotho provides rich semantic descriptions for audio events. Its official split is adopted.

- **ABIDE** [76] is a neuroimaging dataset comprising brain imaging data (sMRI, fMRI, and *etc.*) from 871 subjects, including individuals with autism spectrum disorder (ASD) and healthy controls. Collected from multiple international sites, it includes functional connectivity data, structural imaging, and metadata (*e.g.*, age and gender). We utilize the metadata to construct the textual attribute as a modality. We follow the split protocol proposed by [94]. Note that we do not employ the cross-validation method for evaluation.

B.2 Baselines

We briefly introduce the used baselines in multimodal learning and autism classification.

- **Frozen** [24] is an end-to-end trainable model adapting ViT and Timesformer architectures with spatio-temporal attention, trained on both large-scale image and video captioning datasets using a curriculum learning approach.

- **UMT** [77] is the first to jointly optimize moment retrieval and highlight detection in videos by integrating multi-modal (visual-audio) learning, treating moment retrieval as keypoint detection with a query generator and decoder.

- **UMT-L** [78] enhances data efficiency by masking low-semantics video tokens and selectively aligning unmasked tokens with an image foundation model as an unmasked teacher, enabling faster convergence and multimodal compatibility.

- **OmniVL** [30] introduces a unified transformer-based foundation model that supports both image-language and video-language tasks through a single architecture, utilizing decoupled joint pretraining to enhance spatial and temporal vision-language modeling.

- **TVTSv2** [79] proposes a degradation-free pre-training strategy for video foundation models, preserving the text encoder's generalization by freezing shallow layers and tuning deep layers, while using a transcript sorting task with masking for scalable training.

- **CLIP4Clip** [36] adapts a CLIP image-language pre-training model for end-to-end video-text retrieval.

- **ViCLIP** [26] is a video-text representation learning model based on ViT-L, trained on a large-scale video-centric multimodal dataset with over 7 million videos and 234M clips, paired with 4.1B words of detailed descriptions.

- **VideoCoCa** [80] adapts a pretrained image-text contrastive captioner model for video-text tasks by leveraging its generative and contrastive attentional pooling layers for flattened frame embeddings.

- **Norton** [81] employs video-paragraph and clip-caption contrastive losses for video-language learning, which filters irrelevant clips and captions, realigns asynchronous pairs, and uses a soft-maximum operator to handle fine-grained frame-word misalignments.

- **ImageBind** [10] introduces a joint embedding method across six modalities with image-paired data, leveraging large-scale vision-language models to extend zero-shot capabilities to new modalities.

- **InternVideo-L** [82] presents a general video foundation model that combines generative masked video modeling and discriminative video-language contrastive learning to pretrain video representations.

- **HiTeA** [83] introduces a hierarchical temporal-aware video-language pre-training framework with cross-modal moment exploration to model detailed video moment representations and multi-modal temporal relation exploration to capture temporal dependencies across video-text pairs at varying time resolutions.

- **mPLUG-2** [84] introduces a modularized multi-modal pretraining framework with a multi-module composition network, sharing universal modules for modality collaboration while disentangling modality-specific modules to address entanglement.

- **VALOR-L** [85] proposes an end-to-end pretraining framework that jointly models vision, audio, and language using three separate encoders for modality-specific representations and a decoder for multimodal conditional text generation.

- **TEFAL** [86] introduces a text-conditioned feature alignment method for text-to-video retrieval, utilizing two independent cross-modal attention blocks to align text queries with audio and video representations separately.

- **Bimodal T2M** [87] proposes a hierarchical multimodal video retrieval model that enhances text-to-video retrieval by creating a shared embedding space using task-specific contrastive loss functions, designed to maximize mutual information between textual and cross-modal representations.

- **T-MASS** [88] introduces a stochastic text modeling approach for text-video retrieval, representing text as a flexible, resilient semantic "text mass" through a similarity-aware radius module and supporting text regularization.

- **vid-TLDR** [89] proposes a training-free token merging method for video Transformers, enhancing efficiency by merging background tokens using a saliency-aware strategy that leverages attention maps to focus on salient regions and drop irrelevant background tokens.

- **VideoPrism-b** [28] introduces a general-purpose video encoder pretrained on a diverse corpus of 36M video-caption pairs and 582M clips with noisy text, using a global-local distillation and token shuffling approach to enhance masked autoencoding.

- **LanguageBind** [11] proposes a multi-modal pretraining framework that extends video-language pretraining to multiple modalities by using a frozen language encoder from VL pretraining as the semantic bind.

- **AVFIC** [25] propose a multimodal transformer-based model trained on a new large-scale, weakly labeled audio-video captioning dataset with millions of paired clips and captions without additional manual effort.

- **VIP-ANT** [90] leverages shared image modality as a pivot in a tri-modal embedding space for audio-text alignment, eliminating the need for parallel audio-text data.

- **VAST** [27] trains a multimodal foundation model on the VAST-27M dataset, which is created by integrating vision and audio captions generated by separately trained captioners with subtitles using a large language model.

- **GRAM** [31] aligns multiple modalities in a higher-dimensional embedding space using a contrastive loss function that minimizes the Gramian volume of the k-dimensional parallelotope spanned by modality vectors.

For VAST, we utilize its pre-trained base model for zero-shot prediction. Due to it not releasing the fine-tuned versions, we fine-tune it for downstream tasks to evaluate its performance under the fine-tuning setting. For GRAM, we directly use their well-trained model weights for evaluation for two settings. Below, we introduce the baselines in autism classification.

- **AE-FCN** [91] integrates functional connectivity patterns from fMRI and volumetric correspondences of gray matter from sMRI, using a combination of unsupervised stacked autoencoders and supervised multilayer perceptrons.

- **GCN** [92] utilizes graph convolutional networks by representing populations as a sparse graph, where nodes incorporate imaging-based feature vectors and edges integrate phenotypic information as weights.

- **BrainNetCNN** [93] employs a convolutional neural network (CNN) to predict neurodevelopmental outcomes. It uses several convolutional filters (edge-to-edge, edge-to-node, node-to-graph) with the topological locality from structural brain networks.

19

- **DGM** [102] introduces a learnable function that predicts edge probabilities in graphs, enabling end-to-end training with convolutional graph neural network layers to infer graph structures directly from data.

- **BrainNetTF** [94] models brain networks as graphs with fixed-size, ordered nodes using connection profiles as node features for natural positional information and learns pairwise ROI connection strengths via efficient attention weights.

- **VanillaTF** is a simplified version of BrainNetTF, which consists of a two-layer Transformer and a concat-based readout.

VAST and GRAM are also utilized for comparison by equipping BrainNetTF, MLP, and BERT as modal encoders. PMRL follows the same model architecture for a fair comparison.

B.3 Evaluation Metrics

We evaluate the multimodal retrieval tasks with **Recall** as the metric, and evaluate autism classification (in binary) by using **AUC** (Area Under the Curve) and **ACC** (Accuracy) as metrics. Recall@K measures the proportion of relevant items successfully retrieved within the top K results. Let $\{q_1, q_2, \ldots, q_N\}$ denote the set of queries, and for each query q_i, let $\mathcal{R}_i^K \subseteq \mathcal{D}$ denote the set of top K retrieved items from the dataset $\mathcal{D}$. Let $\mathcal{S}_i \subseteq \mathcal{D}$ denote the set of true positive (relevant) items associated with query q_i. Recall at K is defined as: Recall@$K = \frac{1}{N} \sum_{i=1}^{N} \frac{|\mathcal{R}_i^K \cap \mathcal{S}_i|}{|\mathcal{S}_i|}$. In cases where each query corresponds to exactly one correct match, this simplifies to the ratio of queries for which the correct item appears among the top K retrieved results. Accuracy is defined as Accuracy $= \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(\hat{y}_i = y_i)$, where n is the total number of samples and $\mathbb{I}(\cdot)$ is the indicator function. Let $f(\mathrm{x}_i) \in [0, 1]$ denote the model's predicted probability for the sample x_i. The AUC estimates the probability that a randomly chosen positive sample is ranked higher than a randomly chosen negative one: AUC $= \frac{1}{n_+ n_-} \sum_{i:y_i=1} \sum_{j:y_j=0} \mathbb{I}(f(\mathrm{x}_i) > f(\mathrm{x}_j))$, where n_+ and n_- are the numbers of positive and negative samples, respectively.

B.4 Hyperparameter Settings

We utilize AdamW [98] as the optimizer, where the learning rate is set to 2×10^{-5}, $\beta_1 = 0.9$, and $\beta_2 = 0.98$. The linear schedule is employed for warmup, with a warmup ratio of 0.1. The weight decay is 0.01, and the gradient norm is limited to 2. All the representations are transformed into 512 dimensions. For our PMRL model, τ_1 is set to 0.05, τ_2 is set to 0.1; λ_1 is configured as 1.0 and λ_2 is 0.1. For autism classification, we employ 5-times trials and report the averaged performance. We set the learning rate to 1×10^{-4} and Adam [103] as the optimizer. The output representations are transformed into 128 dimensions. We adjust τ_2 to 0.4, and other settings are kept the same.

B.5 Model Architecture

We design the PMRL model architecture following the well-developed VAST model. Specifically, the vision encoder is set to use EVAClip-ViT-G [95], with 1.3B parameters. The input resolution for visual data is configured to 224×224 pixels. The text encoder is implemented with BERT, with the maximum caption length limited to 40 and the subtitle length to 70. The audio encoder is configured to use the BEATs model [96]. The audio input is processed into 64 mel-frequency bins, and the target input length is set to 1,024 frames.

For multimodal neuron imaging tasks, we implement the PMRL model by equipping it with an fMRI encoder as BrainNetTF [94] (built upon a graph transformer model), an sMRI encoder as a 2-layer MLP, and a text encoder as BERT as well. Resting-state fMRI data is preprocessed via a CPAC pipeline and a specified brain parcellation atlas (*i.e.*, CC200). For each subject, the mean time series of each brain region was extracted using the selected atlas. Subsequently, two types of functional connectivity matrices were computed: Pearson correlation and partial

---

Algorithm 1 PMRL: Principal Multimodal Representation Learning (Training)

Require: Inputs:
     Dataset $\mathcal{X} = \{(\mathbf{x}_i^{m_1}, \mathbf{x}_i^{m_2}, \dots, \mathbf{x}_i^{m_k})\}_{i=1}^N$ with $k = |\mathcal{M}|$ modalities per instance.
     Encoder networks $\{f^m(\cdot; \boldsymbol{\theta}^m)\}_{m=1}^k$, parameterized by $\boldsymbol{\theta}^m$.
     Temperature parameter $\tau > 0$, regularization weights λ_1, λ_2.
Ensure: Aligned multimodal representations.

1: **for** each training iteration **do**
2:     **Sample a batch of instances:** $\{\mathbf{x}_i^{m_1}, \mathbf{x}_i^{m_2}, \dots, \mathbf{x}_i^{m_k}\}_{i=1}^B$;
3:

    ① Modality-specific encoding:
$$\begin{cases} \text{Encode modality-specific embeddings: } \mathbf{z}^m = f^m(\mathbf{x}^m; \boldsymbol{\theta}^m), \quad \forall m \in \mathcal{M}; \\[2mm] \text{Normalize embeddings: } \mathbf{z}^m \leftarrow \dfrac{\mathbf{z}^m}{\|\mathbf{z}^m\|}, \quad \forall m \in \mathcal{M}; \\[2mm] \text{Stack normalized embeddings into matrix: } \mathbf{Z} = [\mathbf{z}^{m_1}, \mathbf{z}^{m_2}, \dots, \mathbf{z}^{m_k}] \in \mathbb{R}^{d \times k}; \end{cases}$$

4:     ② Perform SVD on $\mathbf{Z}$: $\text{SVD}(\mathbf{Z}) = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \quad \boldsymbol{\Sigma} = \text{diag}(\sigma_1, \dots, \sigma_k)$;
5:     Extract leading singular value σ_1 and corresponding left singular vector $\mathbf{u}_1$;

    ③ The core part of PMRL:
$$\begin{cases} \text{Compute the alignment loss } \mathcal{L}^{\mathcal{M}} \text{ via softmax over singular values (see Eq. (5));} \\[2mm] \text{Compute the contrastive regularization } \mathcal{L}^{\mathcal{M}'} \text{ using leading directions (see Eq. (6));} \end{cases}$$

6:     ④ Generate matched/mismatched pairs for instance matching loss $\mathcal{L}_{\text{IM}}$ (see Eq. (7));
7:     ⑤ Combine losses with weighting coefficients: $\mathcal{L} = \mathcal{L}^{\mathcal{M}} + \lambda_1 \mathcal{L}^{\mathcal{M}'} + \lambda_2 \mathcal{L}_{\text{IM}}$;
8:     ⑥ Update parameters $\boldsymbol{\theta}$ via gradient descent: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta \nabla_{\boldsymbol{\theta}} \mathcal{L}$;
9: **end for**
10: **Return:** Optimized encoder parameters $\boldsymbol{\theta}$ that encourage fully aligned multimodal representations.

---

correlation matrices, representing the pairwise relationships between brain regions. sMRI features are extracted from FreeSurfer-processed outputs. ComBat harmonization is applied to the sMRI features to mitigate site and batch effects, using site, age, sex, IQ, and diagnostic label as covariates. The resulting sMRI features are concatenated into a matrix. For the textual features, we combine age and gender attributes as "age: <attr_age>, gender: <attr_gender>" for each subject. The multimodal representations are averaged and fed to a 3-layer MLP that returns the predictions in binary for classification. We replace $\mathcal{L}_{\text{IM}}$ with the classification loss in implementing VAST, GRAM, and PMRL.

B.6 Algorithm Flow and Pseudo Code

To facilitate reproducibility, we provide the algorithm flow and pseudo code of PMRL. These materials demonstrate the straightforward implementation of integrating PMRL in just a few steps.

Integrating PMRL with four steps

```python
# 1. Singular Value Decomposition on Multimodal Representations >>>
U, S, _ = torch.linalg.svd(
    torch.stack([feat_t,feat_v,feat_a,feat_s], dim=-1)
    )

# 2. Principled learning via maximum singular values >>>
loss1 = F.cross_entropy(S/self.tau1, torch.zeros(S.shape[0]).to(S.device).long())
# Implemented by cross-entropy, and the singular value at the first position is
    the maximum one

# 3. Principled regularization via eigenvector corresponding to the maximum
    singular values >>>
U1 = U[:, :, 0]
loss2 = F.cross_entropy((U1 @ U1.T)/self.tau2, torch.arange(U1.shape[0]).to(U1.
    device).long())

......

# 4. Combine the loss >>>
loss = loss1 + self.lambda1 * loss2 + self.lambda2 * loss_IM
```

C   Additional Results

We provide the full results on multimodal text-video retrieval, especially in terms of Recall@1, Recall@5, and Recall@10 metrics as shown in Tables 7, 8, 9, and 10. Moreover, we illustrate more results of any modality retrieval on Recall@5 and Recall@10 in Figures 7 and 8. We also exhibit the trends of each singular value during training to reveal the collapse of GRAM compared to PMRL, as shown in Figure 9.

C.1   Multimodal Text-Video Retrieval

We report the available results in metrics of Recall@1, Recall@5, and Recall@10 for text-video retrieval. The performances under zero-shot and fine-tuning settings are shown in Tables 7, 8 and Tables 9, 10, respectively. Aligning with the results reported in the main content, our PMRL method also outperforms other methods in most cases.

C.2   Any Modality Retrieval

Figures 7 and 8 illustrate a performance comparison between PMRL and GRAM across six benchmark datasets (MSR-VTT, Didemo, ActivityNet, Vatex, AudioCaps, and Clotho) in terms of Recall@5 and Recall@10 for any-modality retrieval. Blue regions highlight areas where PMRL outperforms GRAM, indicating superior retrieval accuracy, while gray regions show where GRAM performs better. Diagonal regions, colored in white, represent self-modal retrieval, which is not meaningful for comparison and thus excluded from the analysis. The 3D bar charts visualize the performance differences across various modalities (denoted as A, T, V, *etc.*), with the height of the bars reflecting the recall scores, providing a clear visual representation of the relative strengths of PMRL and GRAM across different datasets and retrieval conditions. From the observations, it can be concluded that PMRL generally outperforms GRAM, with

Table 7: Multimodal text-to-video (T→V) and video-to-text (V→T) retrieval results on zero-shot setting (%) across MSR-VTT and DiDeMo.

| | MSR-VTT | | | | | | DiDeMo | | | | | |
| | T→V | | | V→T | | | T→V | | | V→T | | |
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
Fronzen [24]	18.7	39.5	51.6	-	-	-	21.1	46.0	59.2	-	-	-
UMT [77]	33.3	-	66.7	-	-	-	34.0	-	68.7	-	-	-
UMT-L [78]	40.7	63.4	71.8	37.1	-	-	48.6	72.9	79.0	49.9	-	-
OmniVL [30]	42.0	63.0	73.0	40.7	-	-	40.6	64.6	74.3	24.9	-	-
TVTSv2 [79]	38.2	62.4	73.2	-	-	-	34.6	61.9	71.5	-	-	-
ViCLIP [26]	42.4	-	-	41.3	-	-	18.4	-	-	27.9	-	-
VideoCoCa [80]	34.3	57.8	67.0	64.7	85.2	67.0	-	-	-	-	-	-
Norton [81]	10.7	24.1	31.6	-	-	-	-	-	-	-	-	-
ImageBind [10]	36.8	61.8	70.0	-	-	-	-	-	-	-	-	-
InternVideo-L [82]	40.7	-	-	39.6	-	-	31.5	-	-	33.5	-	-
HiTeA [83]	34.4	60.0	69.9	-	-	-	43.2	69.3	79.0	-	-	-
mPLUG-2 [84]	47.1	69.7	79.0	-	-	-	45.7	71.1	71.1	-	-	-
VideoPrism-b [28]	51.4	-	-	50.2	-	-	-	-	-	-	-	-
LanguageBind [11]	44.8	70.0	78.7	40.9	66.4	75.7	39.9	66.1	74.6	39.8	67.8	76.2
VAST [27]	50.5	69.0	74.3	48.8	69.9	75.6	46.4	67.5	73.5	45.3	68.7	75.4
GRAM [31]	51.5	71.5	77.9	51.5	73.5	79.5	49.8	71.0	76.3	48.5	70.1	75.5
PMRL (Ours)	54.5	73.2	80.4	52.4	73.8	79.8	50.6	72.7	77.4	48.4	70.8	78.3

Table 8: Multimodal text-to-video (T→V) and video-to-text (V→T) retrieval results on zero-shot setting (%) across ActivityNet and VATEX.

| | ActivityNet | | | | | | VATEX | | | | | |
| | T→V | | | V→T | | | T→V | | | V→T | | |
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UMT [78]	31.9	-	72.0	-	-	-	-	-	-	-	-	-
UMT-L [78]	41.9	-	-	39.4	-	-	-	-	-	-	-	-
ViCLIP [26]	15.1	-	-	24.0	-	-	-	-	-	-	-	-
VideoCoCa [80]	34.5	63.2	76.6	33.0	61.6	75.3	53.2	83.3	90.1	73.6	93.2	97.2
InternVideo-L [82]	30.7	-	-	31.4	-	-	49.5	-	-	69.5	-	-
VideoPrism-b [28]	49.6	-	-	47.9	-	-	62.5	-	-	77.1	-	-
LanguageBind [11]	41.0	68.4	80.0	39.1	69.8	81.1	-	-	-	-	-	-
VAST [27]	51.7	75.7	83.4	48.8	74.8	81.9	75.9	93.3	94.8	74.8	93.5	95.6
GRAM [31]	54.5	78.3	85.2	48.3	74.2	82.6	77.5	94.8	96.2	74.7	93.5	95.5
PMRL (Ours)	56.0	80.0	87.4	49.6	76.0	85.6	80.5	95.4	96.4	75.2	93.8	95.5

the strongest performance observed in text-relevant modality retrieval. Additionally, PMRL demonstrates significant improvement over GRAM in non-text-relevant modality retrieval as well, like V→T.

C.3 Eigenvalues Analysis

Figure 9 illustrates the trends of singular values during the training of a model from scratch, comparing two methods: GRAM and PMRL. The GRAM method primarily focuses on minimizing one specific singular value, as evidenced by the significant decline of the red line (σ_4) over the training steps, while the singular values of σ_2 and σ_3 remain relatively stable. In contrast, the PMRL method minimizes all singular values except the maximum one (σ_1) simultaneously, which is reflected in the gradual decrease of σ_2, σ_3, and σ_4, while σ_1 keeps increasing. This comparison highlights the different optimization strategies employed by GRAM and PMRL, with GRAM collapsing to minimize the minimum singular value and PMRL optimizing multiple values concurrently.
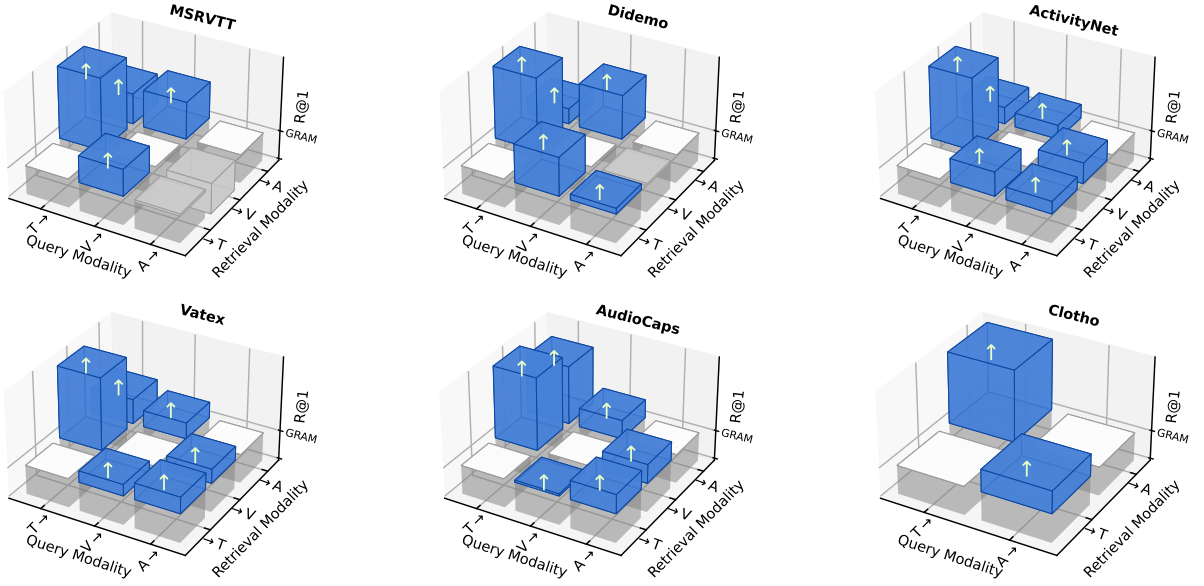
Figure 7: **Performance comparison of PMRL *v.s.* GRAM in terms of Recall@5 for any modality retrieval across 6 benchmark datasets.** Blue regions highlight where PMRL outperforms GRAM, while gray regions indicate the opposite. Diagonal regions (colored in white) represent self-modal retrieval, which is not meaningful for comparison.

Figure 8: **Performance comparison of PMRL *v.s.* GRAM in terms of Recall@10 for any modality retrieval across 6 benchmark datasets.** Blue regions highlight where PMRL outperforms GRAM, while gray regions indicate the opposite. Diagonal regions (colored in white) represent self-modal retrieval, which is not meaningful for comparison.
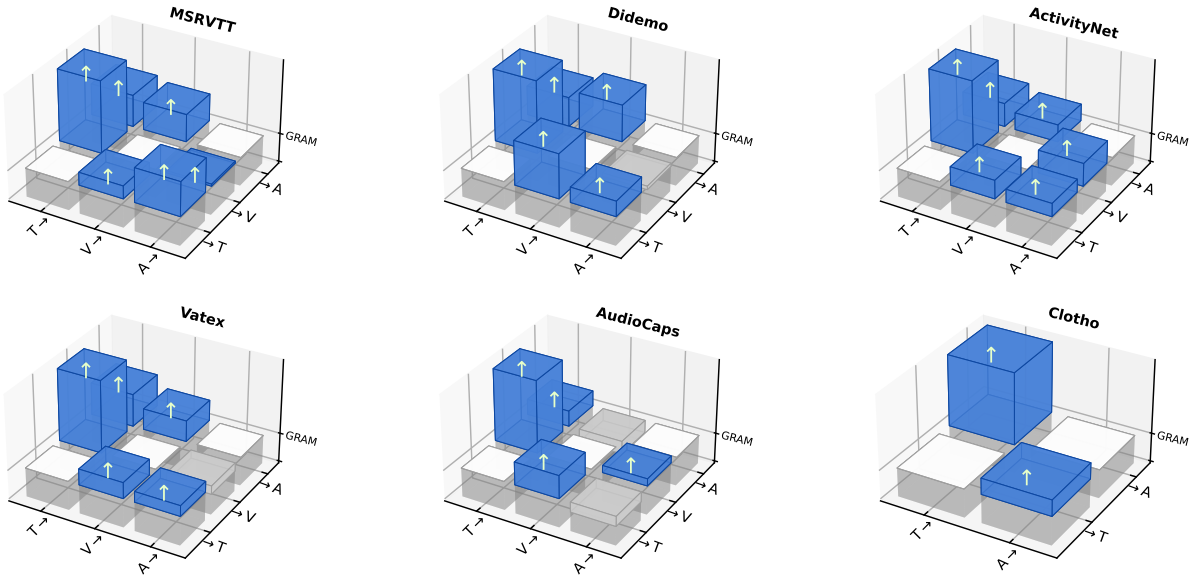
Table 9: Multimodal text-to-video (T→V) and video-to-text (V→T) retrieval results on finetuning setting (%) across MSR-VTT and DiDeMo.

| | MSR-VTT | | | | | | DiDeMo | | | | | |
| | T→V | | | V→T | | | T→V | | | V→T | | |
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UMT-L [78]	58.8*	81.0*	87.1*	58.6*	-	-	70.4*	90.1*	93.5*	65.7*	-	-
CLIP4Clip [36]	44.5	71.4	81.6	45.9	-	-	43.4	70.2	80.6	43.6	-	-
ViCLIP [26]	52.5	-	-	51.8	-	-	49.4	-	-	50.2	-	-
InternVideo-L [82]	55.2*	-	-	57.9*	-	-	57.9*	-	-	59.1*	-	-
HiTeA [83]	46.8	71.2	81.9	-	-	-	56.5	81.7	89.7	-	-	-
mPLUG-2 [84]	53.1	77.6	84.7	-	-	-	56.4	79.1	85.2	-	-	-
VALOR-L [85]	54.4	79.8	87.6	-	-	-	57.6	83.3	88.8	-	-	-
TEFAL [86]	52.0	76.6	86.1	-	-	-	-	-	-	-	-	-
Bimodal T2M [87]	36.8	-	-	-	-	-	-	-	-	-	-	-
T-MASS [88]	52.7	77.1	85.6	-	-	-	53.3	80.1	87.7	-	-	-
vid-TLDR [89]	58.5*	81.3*	86.9*	-	-	-	70.4*	90.5*	94.0*	-	-	-
VAST [27]	64.4	84.3	90.4	64.3	86.2	92.9	68.4	86.9	90.1	65.4	88.0	90.7
GRAM [31]	60.0	79.6	84.3	61.8	80.9	85.2	68.7	86.0	89.2	65.7	86.8	91.2
PMRL (Ours)	61.2	80.4	85.5	60.7	82.2	86.4	70.2	87.5	91.0	66.4	87.8	90.9

Table 10: Multimodal text-to-video (T→V) and video-to-text (V→T) retrieval results on finetuning setting (%) across ActivityNet and VATEX.

| | ActivityNet | | | | | | VATEX | | | | | |
| | T→V | | | V→T | | | T→V | | | V→T | | |
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
UMT-L [78]	66.8*	89.1*	94.9*	64.4*	-	-	72.0*	-	-	86.0*	-	-
CLIP4Clip [36]	40.5	72.4	-	41.6	-	-	55.9	89.2	95.0	78.3	-	-
ViCLIP [26]	49.8	-	-	48.1	-	-	-	-	-	-	-	-
InternVideo-L [82]	62.2*	-	-	62.8*	-	-	71.1*	-	-	87.2*	-	-
VALOR-L [85]	63.4	87.8	94.1	-	-	-	76.9	96.7	98.6	-	-	-
TEFAL [86]	-	-	-	-	-	-	61.0	90.4	95.3	-	-	-
T-MASS [88]	-	-	-	-	-	-	65.6	93.9	97.2	-	-	-
vid-TLDR [89]	65.2*	88.7*	94.5*	-	-	-	-	-	-	-	-	-
VAST [27]	68.1	89.5	95.7	65.4	88.7	94.9	83.1	98.1	99.2	81.3	98.4	99.6
GRAM [31]	67.6	89.4	95.4	65.0	88.4	94.5	82.5	98.0	98.9	80.6	98.0	99.2
PMRL (Ours)	68.2	89.1	94.6	66.4	88.4	94.1	84.1	97.3	98.3	83.2	97.8	98.4

D Reproducibility

We provide implementation details, involving illustrative algorithm descriptions and pseudo-code in Appendix B.6. The source code will be publicly released for reproducibility.

E Limitations

PMRL advances multimodal alignment by optimizing the maximum singular value of Gram matrices and ensuring instance-wise separability. However, the resource constraints, like the updated YouTube policies on video downloads, prevented us from collecting large-scale, high-quality multimodal datasets needed to fully enhance PMRL's capabilities. Therefore, PMRL has to employ continual training on pre-trained models. Despite the limitation, experimental results demonstrate the effectiveness of PMRL and the rationale of our core design.
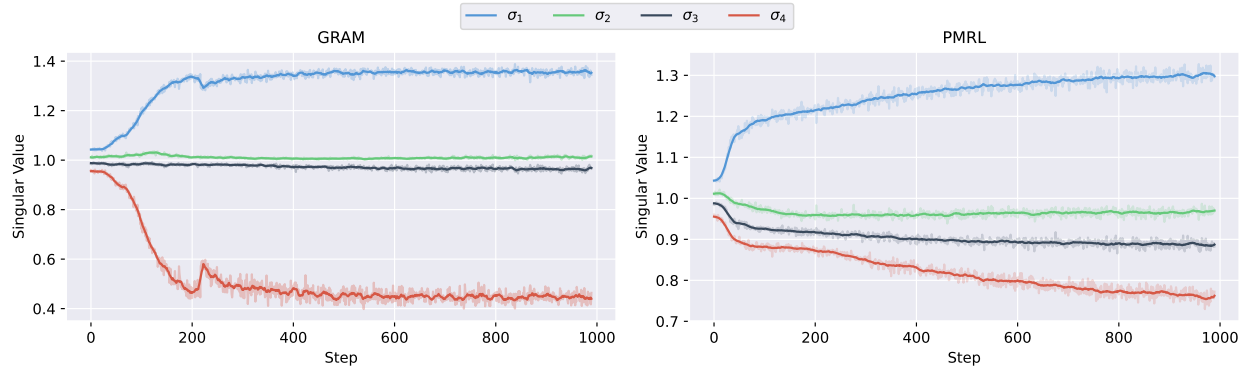
Figure 9: **The trends of singular values when training the model from scratch.** GRAM mainly focuses on minimizing one singular value, while PMRL minimizes all except the largest one simultaneously.

References

[1] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, Andrew Y Ng, et al. Multimodal deep learning. In ICML, volume 11, pages 689–696, 2011.

[2] Zhou Lu. A theory of multimodal learning. NeurIPS, 36:57244–57255, 2023.

[3] Yongshuo Zong, Oisin Mac Aodha, and Timothy Hospedales. Self-supervised multimodal learning: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[4] Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(10):12113–12132, 2023.

[5] Ye Zhu, Yu Wu, Nicu Sebe, and Yan Yan. Vision+ x: A survey on multimodal learning in the light of data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 46(12):9102–9122, 2024.

[6] Run Luo, Haonan Zhang, Longze Chen, Ting-En Lin, Xiong Liu, Yuchuan Wu, Min Yang, Minzheng Wang, Pengpeng Zeng, Lianli Gao, et al. Mmevol: Empowering multimodal large language models with evol-instruct. arXiv preprint arXiv:2409.05840, 2024.

[7] Run Luo, Xiaobo Xia, Lu Wang, Longze Chen, Renke Shan, Jing Luo, Min Yang, and Tat-Seng Chua. Next-omni: Towards any-to-any omnimodal foundation models with discrete flow matching. arXiv preprint arXiv:2510.13721, 2025.

[8] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In ICASSP, pages 4563–4567, 2022.

[9] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In ICML, pages 8748–8763, 2021.

[10] 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 CVPR, pages 15180–15190, 2023.

[11] Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. arXiv preprint arXiv:2310.01852, 2023.

[12] Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. arXiv preprint arXiv:2407.11895, 2024.

[13] Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Continual multimodal contrastive learning. arXiv preprint arXiv:2503.14963, 2025.

[14] Benoit Dufumier, Javiera Castillo-Navarro, Devis Tuia, and Jean-Philippe Thiran. What to align in multimodal contrastive learning? ICLR, 2025.

[15] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In CVPR, pages 9729–9738, 2020.

[16] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In ICML, pages 1597–1607, 2020.

[17] Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In ICML, pages 9929–9939. PMLR, 2020.

[18] Huanrui Yang, Minxue Tang, Wei Wen, Feng Yan, Daniel Hu, Ang Li, Hai Li, and Yiran Chen. Learning low-rank deep neural networks via singular vector orthogonality regularization and singular value sparsification. In CVPR, pages 678–679, 2020.

[19] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084, 2021.

[20] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In CVPR, pages 8552–8562, 2022.

[21] Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In CVPR, pages 26752–26762, 2024.

[22] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. arXiv preprint arXiv:2309.00615, 2023.

[23] Zehan Wang, Ziang Zhang, Xize Cheng, Rongjie Huang, Luping Liu, Zhenhui Ye, Haifeng Huang, Yang Zhao, Tao Jin, Peng Gao, et al. Freebind: Free lunch in unified multimodal space via knowledge fusion. arXiv preprint arXiv:2405.04883, 2024.

[24] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In CVPR, pages 1728–1738, 2021.

[25] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In ECCV, pages 407–426, 2022.

[26] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, et al. Internvid: A large-scale video-text dataset for multimodal understanding and generation. arXiv preprint arXiv:2307.06942, 2023.

[27] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. NeurIPS, 36:72842–72866, 2023.

[28] Long Zhao, Nitesh Bharadwaj Gundavarapu, Liangzhe Yuan, Hao Zhou, Shen Yan, Jennifer J Sun, Luke Friedman, Rui Qian, Tobias Weyand, Yue Zhao, et al. Videoprism: A foundational visual encoder for video understanding. In ICML, pages 60785–60811. PMLR, 2024.

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In ICML, pages 12888–12900. PMLR, 2022.

[30] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Luowei Zhou, Yucheng Zhao, Yujia Xie, Ce Liu, Yu-Gang Jiang, and Lu Yuan. Omnivl: One foundation model for image-language and video-language tasks. NeurIPS, 35:5696–5710, 2022.

[31] Giordano Cicchetti, Eleonora Grassucci, Luigi Sigillo, and Danilo Comminiello. Gramian multimodal representation learning and alignment. ICLR, 2025.

[32] Han Zhang, Jing Yu Koh, Jason Baldridge, Honglak Lee, and Yinfei Yang. Cross-modal contrastive learning for text-to-image generation. In CVPR, pages 833–842, 2021.

[33] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In ICML, pages 4904–4916, 2021.

[34] Yiwei Zhou, Xiaobo Xia, Zhiwei Lin, Bo Han, and Tongliang Liu. Few-shot adversarial prompt learning on vision-language models. In NeurIPS, pages 3122–3156, 2024.

[35] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap: learning audio concepts from natural language supervision. In ICASSP, pages 1–5, 2023.

[36] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing, 508:293–304, 2022.

[37] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. arXiv preprint arXiv:2003.04297, 2020.

[38] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. Valor: Vision-audio-language omni-perception pretraining model and dataset. arXiv preprint arXiv:2304.08345, 2023.

[39] Xiaohao Liu, Xiaobo Xia, Zhuo Huang, See-Kiong Ng, and Tat-Seng Chua. Towards modality generalization: A benchmark and prospective analysis. arXiv preprint arXiv:2412.18277, 2024.

[40] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353, 2020.

[41] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In CVPR, pages 16877–16887, 2021.

[42] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. 2021.

[43] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In ICASSP, pages 976–980, 2022.

[44] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. arXiv preprint arXiv:2006.09199, 2020.

[45] Ludan Ruan, Anwen Hu, Yuqing Song, Liang Zhang, Sipeng Zheng, and Qin Jin. Accommodating audio modality in clip for multimodal processing. In AAAI, volume 37, pages 9641–9649, 2023.

[46] Gene H Golub and Charles F Van Loan. Matrix computations. JHU press, 2013.

[47] Gene H Golub and Christian Reinsch. Singular value decomposition and least squares solutions. In Handbook for automatic computation: volume II: linear algebra, pages 134–151. Springer, 1971.

[48] Michael E Wall, Andreas Rechtsteiner, and Luis M Rocha. Singular value decomposition and principal component analysis. In A practical approach to microarray data analysis, pages 91–109. Springer, 2003.

[49] Hervé Abdi. Singular value decomposition (svd) and generalized singular value decomposition. Encyclopedia of measurement and statistics, 907(912):44, 2007.

[50] Alexander Mathiasen, Frederik Hvilshøj, Jakob Rødsgaard Jørgensen, Anshul Nasery, and Davide Mottin. What if neural networks had svds? NeurIPS, 33:18411–18420, 2020.

[51] Jake Levinson, Carlos Esteves, Kefan Chen, Noah Snavely, Angjoo Kanazawa, Afshin Rostamizadeh, and Ameesh Makadia. An analysis of svd for deep rotation estimation. NeurIPS, 33:22554–22565, 2020.

[52] Jiong Zhang, Qi Lei, and Inderjit Dhillon. Stabilizing gradients for deep neural networks via efficient svd parameterization. In ICML, pages 5806–5814. PMLR, 2018.

[53] Matthew Brand. Incremental singular value decomposition of uncertain data with missing values. In ECCV, pages 707–720. Springer, 2002.

[54] Yihong Gong and Xin Liu. Video summarization using singular value decomposition. In CVPR, pages 174–180, 2000.

[55] Zhenglin Zhou, Huaxia Li, Hong Liu, Nanyang Wang, Gang Yu, and Rongrong Ji. Star loss: Reducing semantic ambiguity in facial landmark detection. In CVPR, pages 15475–15484, 2023.

[56] Zhiteng Li, Mingyuan Xia, Jingyuan Zhang, Zheng Hui, Linghe Kong, Yulun Zhang, and Xiaokang Yang. Adasvd: Adaptive singular value decomposition for large language models. arXiv preprint arXiv:2502.01403, 2025.

[57] Xin Wang, Yu Zheng, Zhongwei Wan, and Mi Zhang. Svd-llm: Truncation-aware singular value decomposition for large language model compression. ICLR, 2025.

[58] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In ICCV, pages 7323–7334, 2023.

[59] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. In NeurIPS, pages 121038–121072, 2024.

[60] Ryumei Nakada, Halil Ibrahim Gulluk, Zhun Deng, Wenlong Ji, James Zou, and Linjun Zhang. Understanding multimodal contrastive learning and incorporating unpaired data. In AISTATS, pages 4348–4380. PMLR, 2023.

[61] Abhi Kamboj and Minh N Do. Leveraging perfect multimodal alignment and gaussian assumptions for cross-modal transfer. arXiv preprint arXiv:2503.15352, 2025.

[62] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. Psychometrika, 1(3):211–218, 1936.

[63] Daesung Kim and Hye Won Chung. Rank-1 matrix completion with gradient descent and small random initialization. NeurIPS, 36:10530–10566, 2023.

[64] Michael Dusenberry, Ghassen Jerfel, Yeming Wen, Yian Ma, Jasper Snoek, Katherine Heller, Balaji Lakshminarayanan, and Dustin Tran. Efficient and scalable bayesian neural nets with rank-1 factors. In ICML, pages 2782–2792. PMLR, 2020.

[65] Matan Gavish and David L Donoho. Optimal shrinkage of singular values. IEEE Transactions on Information Theory, 63(4):2137–2152, 2017.

[66] Brenden P Epps and Eric M Krivitzky. Singular value decomposition of noisy data: noise filtering. Experiments in Fluids, 60(8):126, 2019.

[67] Xiaobo Xia, Tongliang Liu, Bo Han, Mingming Gong, Jun Yu, Gang Niu, and Masashi Sugiyama. Sample selection with uncertainty of losses for learning with noisy labels. arXiv preprint arXiv:2106.00445, 2021.

[68] Xiaobo Xia, Tongliang Liu, Bo Han, Nannan Wang, Mingming Gong, Haifeng Liu, Gang Niu, Dacheng Tao, and Masashi Sugiyama. Part-dependent label noise: Towards instance-dependent label noise. In NeurIPS, pages 7597–7610, 2020.

[69] Sangamesh Kodge, Deepak Ravikumar, Gobinda Saha, and Kaushik Roy. Sap: Corrective machine unlearning with scaled activation projection for label noise robustness. In AAAI, volume 39, pages 17930–17937, 2025.

[70] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In ACL, pages 190–200, 2011.

[71] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In ICCV, pages 5803–5812, 2017.

[72] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In ICCV, pages 706–715, 2017.

[73] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In ICCV, pages 4581–4591, 2019.

[74] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In ACL, pages 119–132, 2019.

[75] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In ICASSP, pages 736–740. IEEE, 2020.

[76] Cameron Craddock, Yassine Benhajali, Carlton Chu, Francois Chouinard, Alan Evans, András Jakab, Budhachandra Singh Khundrakpam, John David Lewis, Qingyang Li, Michael Milham, et al. The neuro bureau preprocessing initiative: open sharing of preprocessed neuroimaging data and derivatives. Frontiers in Neuroinformatics, 7(27):5, 2013.

[77] Ye Liu, Siyuan Li, Yang Wu, Chang-Wen Chen, Ying Shan, and Xiaohu Qie. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In CVPR, pages 3042–3051, 2022.

[78] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In CVPR, pages 19948–19960, 2023.

[79] Ziyun Zeng, Yixiao Ge, Zhan Tong, Xihui Liu, Shu-Tao Xia, and Ying Shan. Tvtsv2: Learning out-of-the-box spatiotemporal visual representations at scale. arXiv preprint arXiv:2305.14173, 2023.

[80] Shen Yan, Tao Zhu, Zirui Wang, Yuan Cao, Mi Zhang, Soham Ghosh, Yonghui Wu, and Jiahui Yu. Videococa: Video-text modeling with zero-shot transfer from contrastive captioners. arXiv preprint arXiv:2212.04979, 2022.

[81] Yijie Lin, Jie Zhang, Zhenyu Huang, Jia Liu, Xi Peng, et al. Multi-granularity correspondence learning from long-term noisy videos. In ICLR, 2023.

[82] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191, 2022.

[83] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. Hitea: Hierarchical temporal-aware video-language pre-training. In ICCV, pages 15405–15416, 2023.

[84] Haiyang Xu, Qinghao Ye, Ming Yan, Yaya Shi, Jiabo Ye, Yuanhong Xu, Chenliang Li, Bin Bi, Qi Qian, Wei Wang, et al. mplug-2: A modularized multi-modal foundation model across text, image and video. In ICML, pages 38728–38748. PMLR, 2023.

[85] Jing Liu, Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, and Jinhui Tang. Valor: Vision-audio-language omni-perception pretraining model and dataset. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.

[86] Sarah Ibrahimi, Xiaohang Sun, Pichao Wang, Amanmeet Garg, Ashutosh Sanan, and Mohamed Omar. Audio-enhanced text-to-video retrieval using text-conditioned feature alignment. In ICCV, pages 12054–12064, 2023.

[87] Pranav Arora, Selen Pehlivan, and Jorma Laaksonen. Text-to-multimodal retrieval with bimodal input fusion in shared cross-modal transformer. In LREC-COLING, pages 15823–15834, 2024.

[88] Jiamian Wang, Guohao Sun, Pichao Wang, Dongfang Liu, Sohail Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. Text is mass: Modeling as stochastic embedding for text-video retrieval. In CVPR, pages 16551–16560, 2024.

[89] Joonmyung Choi, Sanghyeok Lee, Jaewon Chu, Minhyuk Choi, and Hyunwoo J Kim. vid-tldr: Training free token merging for light-weight video transformer. In CVPR, pages 18771–18781, 2024.

[90] Yanpeng Zhao, Jack Hessel, Youngjae Yu, Ximing Lu, Rowan Zellers, and Yejin Choi. Connecting the dots between audio and text without parallel data through visual knowledge transfer. arXiv preprint arXiv:2112.08995, 2021.

[91] Mladen Rakić, Mariano Cabezas, Kaisar Kushibar, Arnau Oliver, and Xavier Lladó. Improving the detection of autism spectrum disorder by combining structural and functional mri information. NeuroImage: Clinical, 25:102181, 2020.

[92] Sarah Parisot, Sofia Ira Ktena, Enzo Ferrante, Matthew Lee, Ricardo Guerrero, Ben Glocker, and Daniel Rueckert. Disease prediction using graph convolutional networks: application to autism spectrum disorder and alzheimer's disease. Medical Image Analysis, 48:117–130, 2018.

[93] Jeremy Kawahara, Colin J Brown, Steven P Miller, Brian G Booth, Vann Chau, Ruth E Grunau, Jill G Zwicker, and Ghassan Hamarneh. Brainnetcnn: Convolutional neural networks for brain networks; towards predicting neurodevelopment. NeuroImage, 146:1038–1049, 2017.

[94] Xuan Kan, Wei Dai, Hejie Cui, Zilong Zhang, Ying Guo, and Carl Yang. Brain network transformer. NeurIPS, 35:25586–25599, 2022.

[95] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389, 2023.

[96] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. Beats: Audio pre-training with acoustic tokenizers. In ICML, pages 5178–5193. PMLR, 2023.

[97] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In NAACL, pages 4171–4186, June 2019.

[98] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101, 2017.

[99] Hongwei Xue, Tiankai Hang, Yanhong Zeng, Yuchong Sun, Bei Liu, Huan Yang, Jianlong Fu, and Baining Guo. Advancing high-resolution video-language representation with large-scale video transcriptions. In CVPR, 2022.

[100] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. International Journal of Computer Vision, 123:94–120, 2017.

[101] Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. arXiv preprint arXiv:2105.02192, 2021.

[102] Anees Kazi, Luca Cosmo, Seyed-Ahmad Ahmadi, Nassir Navab, and Michael M Bronstein. Differentiable graph module (dgm) for graph convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 45(2):1606–1617, 2022.

[103] Diederik Kinga, Jimmy Ba Adam, et al. A method for stochastic optimization. In ICLR, 2015.