Calibrated Multimodal Representation Learning with Missing Modalities

A PREPRINT

Xiaohao Liu[1] **Xiaobo Xia**[1] **Jiaheng Wei**[2] **Shuo Yang**[3] **Xiu Su**[4]
See-Kiong Ng[1] **Tat-Seng Chua**[1]
[1]National University of Singapore
[2]Hong Kong University of Science and Technology (Guang Zhou)
[3]Harbin Institute of Technology (Shenzhen) [4]Central South University

ABSTRACT

Multimodal representation learning harmonizes distinct modalities by aligning them into a unified latent space. Recent research generalizes traditional cross-modal alignment to produce enhanced multimodal synergy but requires all modalities to be present for a common instance, making it challenging to utilize prevalent datasets with missing modalities. We provide theoretical insights into this issue from an *anchor shift* perspective. Observed modalities are aligned with a local anchor that deviates from the optimal one when all modalities are present, resulting in an inevitable shift. To address this, we propose CalMRL for multimodal representation learning to calibrate incomplete alignments caused by missing modalities. Specifically, CalMRL leverages the priors and the inherent connections among modalities to model the imputation for the missing ones at the representation level. To resolve the optimization dilemma, we employ a bi-step learning method with the closed-form solution of the posterior distribution of shared latents. We validate its mitigation of anchor shift and convergence with theoretical guidance. By equipping the calibrated alignment with the existing advanced method, we offer new flexibility to absorb data with missing modalities, which is originally unattainable. Extensive experiments and comprehensive analyses demonstrate the superiority of CalMRL. Our code, model checkpoints, and evaluation raw data will be publicly available.

1 Introduction

The empirical world is perceived by humans through diverse modalities, *e.g.*, vision, audio, and text [1, 2, 3, 4, 5, 6, 7, 8, 9]. For example, eyes capture the shape or colors of one real instance, ears detect, and human language records. Different reflections on the instance shape its abstract form [10, 11], which connects heterogeneous modalities. The single sound you heard invokes your rough imagination in the brain. Multimodal representation learning [12, 13, 14, 15, 16, 17, 18], a key topic in multimodal learning, leverages aligned data to harmonize distinct unimodal encoders and bridges modalities, operating under this philosophy.

Recent research learns multiple modalities simultaneously, rather than grounding one to another [12, 19, 20, 21, 22, 23], fostering greater synergy among them [15, 24, 18]. To achieve multimodal alignment, they utilize geometric techniques to pull different unimodal representations together, ideally achieving a similar point (*i.e.*, anchor). Unfortunately, it requires training on datasets with all modalities present, *i.e.*, a complete set of modalities for one instance,

thus ensuring an unbiased alignment. This introduces an inevitable challenge: collecting such comprehensive datasets is costly and contrasts with the prevalence of incomplete modality data, like paired data (*e.g.*, vision-text [25, 26, 27] or audio-text [28, 29, 30]). The disparity exists between different data, hindering a collective effect. In this paper, we frame this as a *missing modality* problem that is underexplored in multimodal representation learning.

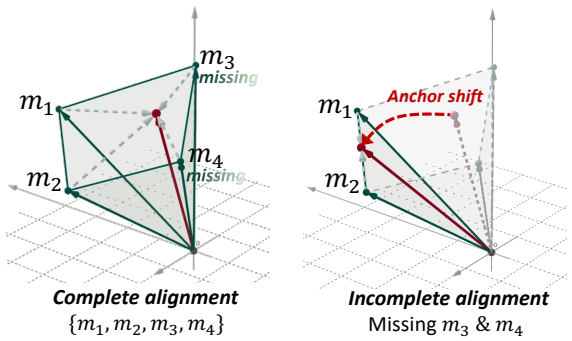

Figure 1: **Missing modalities result in distorted representation alignment.** Different modalities (in green) are aligned together with a virtual anchor (in red) implicitly with all modalities present. With missing modalities, observed ones are enforced to be aligned with a local anchor, deviating from the correct, *i.e.*, anchor shift.

Prior studies, like ImageBind [13] and LangugeBind [31], integrate several paired data (*i.e.*, incomplete modality data), using one existing modality (*i.e.*, vision or text) as the aligning anchor [32, 14, 33, 34]. To ensure the learning stability, the parameters of the corresponding encoder are fixed. They achieve aggregation of diverse modalities, yet bottlenecking the mutual improvements, and heavily depending on the performance of the fixed anchor. The issue of missing modalities is normally bypassed and set aside.

We delve into addressing this and analyzing with a natural perspective, anchor shifting. As illustrated in Figure 1, for complete alignment when all modalities are present, unimodal representations converge toward a virtual anchor within the space spanned by all modalities. In cases of incomplete alignment (*i.e.*, missing modalities m_3 and m_4), observed modalities are aligned with a local anchor, deviating from the optimal one associated with the instance, *i.e.*, resulting in *anchor shifting*. Reflecting on human perception, although we cannot touch or see a specific instance, we can still roughly connect it to other modalities via compensating for the missing ones, rather than being confined to solely what we can perceive. The compensation capability of humans inspires us to leverage the priors to impute missing modalities, thus calibrating the virtual anchor.

To this end, we introduce a novel multimodal representation learning framework, named **CalMRL**, which calibrates alignment in the presence of missing modalities. Specifically, we focus on the missing modality problem in the representation learning phase and theoretically analyze the inevitability of anchor shift. The main idea comes to light: construct oracle modalities if they are missing, thus compensating for the anchor shift. We leverage the priors of missing modalities and the inherent connections among modalities, and introduce a simple generative model, where modalities share common latents, and meanwhile, possess their own distinctness. To achieve this, we adopt a two-step learning method. We first derive a closed-form solution for the posterior distribution of the shared latents with fixed parameters. Then, using this posterior, we optimize the generative parameters. By iterating these two steps, we can progressively refine the parameters, updating only with the observed modalities. Missing modalities can be imputed by the shared latents, derived from the observations, and their priors. We further provide theoretical evidence to support the minor difference between the real and the imputed, and the convergence of our method. This method enables learning a calibrated alignment. We concatenate the observed and imputed modal representations and jointly optimize them with the objective of PMRL [18], aligning all modalities simultaneously. To validate the efficacy and rationale of CalMRL, we conduct extensive experiments and empirical analysis by comparing with state-of-the-art (SOTA) methods.

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

- We introduce Calibrated Multimodal Representation Learning (CalMRL) to address the overlooked missing modalities dilemma, which leads to anchor shift in our theoretical analysis.

- To calibrate the alignment, we propose using a generative model to impute oracle modalities, thus compensating for anchor shift. We refine the imputation precision by iterating the posterior inference and parameter optimization with theoretical grounding, followed by optimizing encoders by incorporating both observed and imputed modalities.
- We conduct extensive experiments to demonstrate the superiority of CalMRL and provide strong empirical evidence supporting our design rationale. Comprehensive empirical studies and discussions are also provided.

2 Related Work

Multimodal representation learning. To integrate multiple modalities, researchers align them within a unified latent space, where one unimodal representation can be retrieved from another when they correspond to a common instance [2, 35, 3, 16, 36, 37, 9]. For example, CLIP [12] pioneers grounding vision in language space through pairwise contrastive learning [38, 39, 40, 41], inspiring a series of methods for connecting bimodalities [22, 19, 42, 23, 20, 21]. Recent work goes beyond pairs, adapting the CLIP paradigm to incorporate more modalities [32, 43, 34, 33], exemplified by ImageBind [13] and LanguageBind [31]. Large-scale data collection [44, 45, 46, 26, 47, 48, 27] and architectural advancements [49, 50, 14, 51, 52, 53], like VAST [14], have furthered progress, though the learning objective remains constrained to pairwise alignment, *i.e.*, aligning one modality with a predefined anchor. GRAM [15] proposes to learn multimodal representations simultaneously in a geometric manner by minimizing the volume of a parallelotope spanned by the modality vectors. TRIANGLE [24] builds on this by minimizing the designed area for three modalities. PMRL [18] establishes principled foundations and proposes maximizing the largest eigenvalue for anchor-free alignment. However, to support its optimization, all modalities must be present to ensure unbiased estimation, unlike prevailing datasets that typically include only two modalities and thus have missing modalities. To endow this emerging paradigm with greater flexibility, we propose calibrating the incomplete alignment with missing modalities for multimodal representation learning.

Learning with missing modalities. This setting indicates the incomplete data with respect to modalities, necessitating the model to perform nearly as well as when all modalities are present [54, 55]. The prevailing way is modality imputation, which involves filling in the missing information by composing existing modalities and generating absent modalities [56]. Typically, this focuses on generating high-quality raw data for the missing modalities [56, 57] (*e.g.*, using an auxiliary adversarial loss to generate missing images [56]) or, more commonly, crafting the representations in the latent space based on the observed ones [58, 59, 60, 61, 62, 63, 64, 65] (*e.g.*, indexing from previous multimodal interactions [60, 65]). Some studies also employ advanced models for distillation to yield superior modality representations [66, 67, 68]. Another research line is to design a specialized fusion module to take advantage of available modalities [69, 70, 71]. For instance, SimMLM [70] introduces a gating network to weigh the contribution of experts corresponding to each modality, and [71] designs a modulation function to compensate for the missing modalities. However, the above methods are mainly specialized for downstream tasks. Learning better multimodal representations with incomplete modalities is not addressed in the fundamental perspective of multimodal alignment. This motivates this work, especially for the emergence of multimodalities in diverse applications.

3 Preliminaries

Aligning all modalities. Different modalities produce synergies, presenting inherent connections despite the heterogeneity. Such connections serve as the key assumption to align different modalities together to learn multimodal representations [2, 6]. Contrastive learning successfully optimizes the similarities of modality pairs [12, 13]. Recent work proposes manipulating the singular value of a GRAM matrix to align multimodalities simultaneously [18]. Specifically, all the uni-modal representations are concatenated together, *i.e.*, $\mathbf{z} = [\mathbf{z}^{m_1}, \cdots, \mathbf{z}^{|\mathcal{M}|}]$. σ_1 denotes the

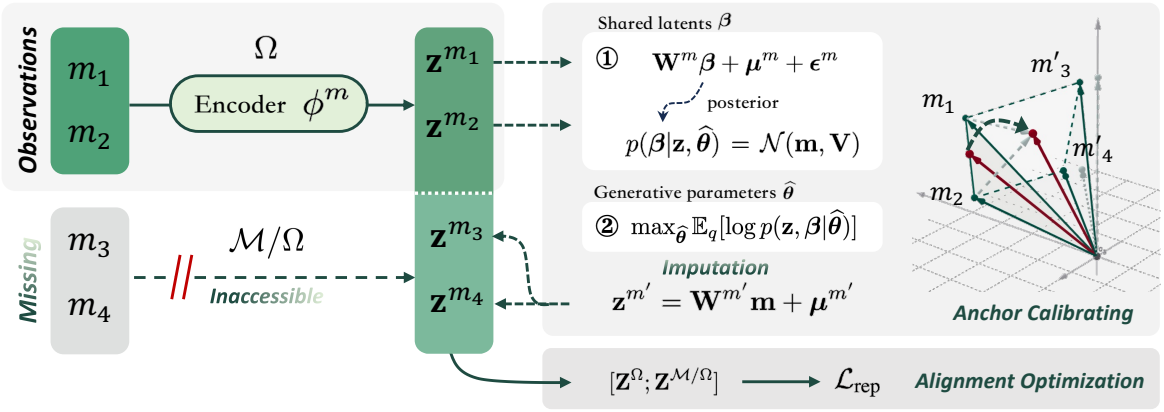

Figure 2: **The overall framework of CalMRL.** Observed unimodal content is first encoded to corresponding representations $\{\mathbf{z}^m\}_{m\in\Omega}$ with individual encoders ϕ^m in $\boldsymbol{\theta}$. Despite the missing modalities (*i.e.*, $\mathcal{M}/\Omega$), CalMRL calibrates multimodal alignment whereby missing modalities are imputed by generative parameters $\widehat{\boldsymbol{\theta}}$. Finally, $\mathcal{L}_{\text{rep}}$ optimizes the observed unimodal encoder to be aligned with the calibrated direction.

largest singular values of the GRAM matrix $\mathbf{z}\mathbf{z}^\top$. By maximizing the largest one σ_1 among others $\{\sigma_i\}_{>1}$, it excels pair-wise methods in alignment tasks, yet is limited by predefined modalities, hard to extend for arbitrary modalities, and can be enhanced with modality-missing datasets.

Modality missing. Given the observed multimodal features $\mathbf{X}^\Omega$, the latent/unified representation can be derived via $p(\widetilde{\mathbf{z}}|\mathbf{X}^\Omega) = p(\widetilde{\mathbf{z}}|\{\mathbf{x}^m\}_{m\in\Omega})$. Ω is the set of observed modalities and $\Omega \subseteq \mathcal{M}$, where $|\mathcal{M}| = k$. If $|\Omega| < k$, it incurs the phenomena of *modality missing*, which is quite common that some multimodal datasets only contain 2 modalities [25, 31], while some involve more [14]. $\widetilde{\mathbf{z}}$ indicates the shared information across modalities. This further informs that the independence among modalities conditioned by $\widetilde{\mathbf{z}}$, formally, $\mathbf{x}^m \perp \mathbf{x}^{m'}|\widetilde{\mathbf{z}}, \forall\{m, m'\} \subset \mathcal{M}$.

Probabilistic PCA. It introduces a generative model which works as $\mathbf{x} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\epsilon}$, where $\boldsymbol{\mu} \in \mathbb{R}^{d'}$ and $\mathbf{z} \in \mathbb{R}^d$ follows $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [72, 73]. $\mathbf{W} \in \mathbb{R}^{d'\times d}$ transforms the latents $\mathbf{z}$ into observed space and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$. Therefore, the conditional distribution of the observed variable is $p(\mathbf{x}|\mathbf{z}) = \mathcal{N}(\mathbf{W}\mathbf{z} + \boldsymbol{\mu}, \sigma^2\mathbf{I})$, and the marginal distribution is $p(\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{W}\mathbf{W}^\top + \sigma^2\mathbf{I})$. The closed-form solution is $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{M}\mathbf{x}_n$, $\sigma^2 = \frac{1}{d'-d}\sum_{i=d+1}^{d'}\lambda_i$, and $\mathbf{W} = \mathbf{U}_d(\Lambda - \sigma^2\mathbf{I})^{1/2}$. $\mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top = \text{SVD}(\mathbf{S})$ decomposes the data covariance $\mathbf{S}$ to eigenvectors $\mathbf{U}$ and eigenvalue matrix $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_{d'})$. We use similar modeling principles, specializing it for multimodal scenarios and emphasizing the connections and distinctiveness of modalities.

4 Calibrated Multimodal Representation Learning

Incomplete alignment. We consider a practical scenario where we cannot collect datasets with complete modalities. Formally, we represent the GRAM matrix as follows:

$$\mathbf{G}^\Omega = \begin{bmatrix} 1 & \langle \mathbf{z}^{m_1}, \mathbf{z}^{m_2}\rangle & \cdots & \langle \mathbf{z}^{m_1}, \mathbf{z}^{m_{k'}}\rangle \\ \langle \mathbf{z}^{m_2}, \mathbf{z}^{m_1}\rangle & 1 & \cdots & \langle \mathbf{z}^{m_2}, \mathbf{z}^{m_{k'}}\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \mathbf{z}^{m_{k'}}, \mathbf{z}^{m_1}\rangle & \langle \mathbf{z}^{m_{k'}}, \mathbf{z}^{m_2}\rangle & \cdots & 1 \end{bmatrix} = \phi(\mathbf{X}^\Omega)\phi(\mathbf{X}^\Omega)^\top, \tag{1}$$

where $\mathbf{G}^\Omega$ is one of the block matrices belonging to the complete one $\mathbf{G} \in \mathbb{R}^{k\times k}$. In practice, when training on a mix of data, it is hard to ensure that all the data instances have the same modalities involved. In other words, one contains

vision and text pairs, while another might contain audio and text data. They all miss at least one modality of data to implement the complete alignment.

Anchor shift $\boldsymbol{\Delta}(\mathbf{Z}, \mathbf{Z}^{\Omega})$. Multiple modalities are aligned together with a virtual centroid (*i.e.*, anchor). Intuitively, there is a deviation between anchors in incomplete modalities and the complete ones. To confirm this, we also provide the theoretical evidence for the anchor shift.

Theorem 1 (Anchor shift under incomplete modality alignment)**.** *Let $\mathbf{u}_1$ and $\mathbf{u}_1^{\Omega}$ be the leading left singular vectors of the full multimodal matrix $\mathbf{Z}$ and its observed submatrix $\mathbf{Z}^{\Omega}$, respectively. Define $\sigma_1 = \|\mathbf{Z}\|_2$, $\sigma_1^{\Omega} = \|\mathbf{Z}^{\Omega}\|_2$, and $\eta := \sqrt{\sum_{m\in\bar{\Omega}} \langle \mathbf{u}_1^{\Omega}, \mathbf{z}^m \rangle^2}$. Then the anchor shift satisfies*

$$\sqrt{2\left(1 - \frac{\sigma_1^{\Omega} + \eta^2}{\sigma_1}\right)} \;\leq\; \underbrace{\|\mathbf{u}_1 - \mathbf{u}_1^{\Omega}\|}_{\|\boldsymbol{\Delta}\|} \;\leq\; \frac{\sqrt{2}\,\|\mathbf{Z}^{\bar{\Omega}}\|_2}{\sigma_1 - \sigma_2}.$$

Remark. The lower bound is *strictly positive* whenever the missing modalities contribute non-zero alignment ($\eta > 0$) or the observed data fails to capture the full leading singular value ($\sigma_1^{\Omega} < \sigma_1$). This implies that *any modality loss inevitably perturbs the virtual anchor*, no matter how well the remaining modalities are aligned. The upper bound shows that this perturbation can be mitigated, but never eliminated, by a large spectral gap (*i.e., strong alignment*) and limited energy in the missing modalities.

Oracle modalities via imputation. We design the generation of unimodal representation to maintain intermodality connections following the conventional principles [2, 74]:

$$\mathbf{z}^m = \mathbf{W}^m \boldsymbol{\beta} + \boldsymbol{\mu}^m + \boldsymbol{\epsilon}^m, \boldsymbol{\epsilon}^m \sim \mathcal{N}(\mathbf{0}, (\sigma^m)^2 \mathbf{I}). \tag{2}$$

The representation for modality m is conditioned by shared latents $\boldsymbol{\beta}$ and its uniqueness $\boldsymbol{\mu}_m$. In this case, we can impute the missing modalities as oracles at the representation level to compensate for the anchor offset, leading to $\boldsymbol{\Delta}(\mathbf{Z}, [\mathbf{Z}^{\Omega}; \mathbf{Z}^{\bar{\Omega}}]) < \boldsymbol{\Delta}(\mathbf{Z}, \mathbf{Z}^{\Omega})$. To this end, even with the missing modalities, the observed modalities can be aligned in an approximated environment with complete modalities.

To optimize the parameters, we aim to maximize the marginal log-likelihood:

$$\arg\max_{\widehat{\boldsymbol{\theta}}} \sum_{i=1}^{N} \log p(\mathbf{z}) = \arg\max_{\widehat{\boldsymbol{\theta}}} \sum_{i=1}^{N} \log \int p(\mathbf{z} \mid \boldsymbol{\beta}) p(\boldsymbol{\beta})\, d\boldsymbol{\beta}, \tag{3}$$

where N denotes the batch size and $\widehat{\boldsymbol{\theta}} = \{\mathbf{W}^m, \boldsymbol{\mu}^m, \sigma^m\}_{m\in\mathcal{M}}$. Unfortunately, as different unimodal parameters are hinged with the common $\boldsymbol{\beta}$, we cannot resolve the parameters with a closed-form solution via naive Probabilistic PCA. For each example, we introduce an arbitrary distribution $q(\boldsymbol{\beta})$ and derive the evidence lower bound:

$$\log p(\mathbf{z}) = \log \mathbb{E}_q\Big[\frac{p(\mathbf{z}, \boldsymbol{\beta})}{q(\boldsymbol{\beta})}\Big] \geq \mathbb{E}_q \log\Big[\frac{p(\mathbf{z}, \boldsymbol{\beta})}{q(\boldsymbol{\beta})}\Big]. \tag{4}$$

By iteratively optimizing q and $\widehat{\boldsymbol{\theta}}$, we can elevate the likelihood.

Resolving $\widehat{\boldsymbol{\theta}}$. Accordingly, we obtain the generative parameters with a bi-step optimization.

① Fixing the parameters, we optimize the lower bound concerning q, *i.e.*, $\max_q \mathbb{E}_q[\log p(\mathbf{z}, \boldsymbol{\beta}|\widehat{\boldsymbol{\theta}})] - \mathbb{E}_q[\log q(\boldsymbol{\beta})]$. The equality of Eq. (4) holds if and only if $q(\boldsymbol{\beta}) = p(\boldsymbol{\beta}|\mathbf{z}, \widehat{\boldsymbol{\theta}})$. Besides, we can obtain the posterior $p(\boldsymbol{\beta}|\mathbf{z}, \widehat{\boldsymbol{\theta}}) = \mathcal{N}(\mathbf{m}, \mathbf{V})$ in a Gaussian distribution. In particular,

$$\mathbf{V} = \Big[\mathbf{I} + \sum_{m\in\Omega} \frac{1}{(\sigma^m)^2} \mathbf{W}^{m\top} \mathbf{W}^m\Big]^{-1}, \;\; \mathbf{m} \;\; = \mathbf{V} \sum_{m\in\Omega} \frac{1}{(\sigma^m)^2} \mathbf{W}^{m\top} (\mathbf{z}^m - \boldsymbol{\mu}^m), \tag{5}$$

and we can update *only* with the observed modalities in Ω.

② Given the posterior, we can maximize the lower bound concerning $\widehat{\boldsymbol{\theta}}$, *i.e.*, $\max_{\widehat{\boldsymbol{\theta}}} \mathbb{E}_q[\log p(\mathbf{z}, \boldsymbol{\beta}|\widehat{\boldsymbol{\theta}})]$. Here we derive the closed-form solution for each parameter as follows:

$$\begin{cases} \boldsymbol{\mu}^m = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N}(\mathbf{z}_i^m - \mathbf{W}^m\mathbf{m}_i), \\ \mathbf{W}^m = \left(\displaystyle\sum_{i=1}^{N}(\mathbf{z}_i^m - \boldsymbol{\mu}^m)\mathbf{m}_i^\top\right)\left(\displaystyle\sum_{i=1}^{N}\mathbb{E}[\boldsymbol{\beta}_i\boldsymbol{\beta}_i^\top]\right)^{-1}, \\ (\sigma^m)^2 = \dfrac{1}{Nd}\displaystyle\sum_{i=1}^{N}\left[\|\mathbf{z}_i^m - \boldsymbol{\mu}^m - \mathbf{W}^m\mathbf{m}_i\|^2 + \mathrm{Tr}[\mathbf{W}^{m\top}\mathbf{W}^m\mathbf{V}]\right]. \end{cases} \tag{6}$$

Alternatively, the parameters of the model can be optimized by minimizing the Gaussian negative log-likelihood loss. These generative parameters can be refined by the next iteration. For the generation of representation for missing modalities $\mathbf{z}^{m'}$, we have the following proposition.

Proposition 2 (Missing modality imputation)**.** *Given observed modalities Ω and learned generative parameters $\widehat{\boldsymbol{\theta}} = \{\mathbf{W}^m, \boldsymbol{\mu}^m, \sigma^m\}_{m\in\mathcal{M}}$, the representation of any missing modality $m' \in \mathcal{M} \setminus \Omega$ can be imputed in closed form as*

$$\widehat{\mathbf{z}}^{m'} = \mathbf{W}^{m'}\mathbf{m} + \boldsymbol{\mu}^{m'}, \quad \forall m' \in \mathcal{M}/\Omega. \tag{7}$$

We provide the detailed derivation in Appendix A.2.

Training objective. For the next phase, we optimize the encoders $\phi : \mathcal{X} \to \mathcal{Z}$ with parameters $\boldsymbol{\theta}$. We complete the multimodal representations and execute the maximum singular values following [18].

$$\mathbf{U}\boldsymbol{\Lambda}\mathbf{V} = \mathrm{SVD}([\mathbf{Z}^\Omega; \widehat{\mathbf{Z}}^{\mathcal{M}/\Omega}]), \quad \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_k). \tag{8}$$

Accordingly, the learning objective is designed to maximize the largest singular value to enforce full alignment, and the corresponding eigenvectors are used for instance-level regularization, ensuring uniformity, as follows:

$$\mathcal{L}_{\text{rep}} = -\frac{1}{N}\sum_{i=1}^{N}\left[\frac{\exp[\lambda_1/\tau]}{\sum_{j=1}^{k}\exp[\lambda_j/\tau]} + \frac{\exp[\mathbf{u}_1^{i\top}\mathbf{u}_1^i/\tau']}{\sum_{j=1}^{N}\exp[\mathbf{u}_1^{i\top}\mathbf{u}_1^j/\tau']}\right] + \alpha \cdot \mathbb{E}_{\{m\}^k\sim\mathcal{M}}[y\log\hat{y} + (1-y)\log(1-\hat{y})],$$

where N denotes the number of data points in a batch. The instance matching loss is applied only to observed modalities, and weighted by $\alpha = 0.1$ with a multimodal encoder and an MLP layer to predict whether the multimodal data is matched or not, returning the prediction $\hat{y}$ [14, 18, 24].

4.1 Analysis

To shed light on our method, we provide a more in-depth analysis of how it alleviates the anchor shift and whether it can converge.

Alleviating anchor shift. CalMRL introduces the imputation of missing modalities to mitigate the anchor shift in multimodal alignment. We derive the comparison results of the anchor offset before and after calibration as follows.

Corollary 3 (Less anchor shift with calibration)**.** *Let each imputation satisfies $\|\widehat{\mathbf{z}}^{m'} - \mathbf{z}^{m'}\|_2 \le \varepsilon, \forall m' \in \bar{\Omega}$. Then the calibration-induced anchor deviation obeys $\|\boldsymbol{\Delta}(\mathbf{Z}, [\mathbf{Z}^\Omega; \mathbf{Z}^{\bar{\Omega}}])\| < \|\boldsymbol{\Delta}(\mathbf{Z}, \mathbf{Z}^\Omega)\|$ if and only if*

$$\varepsilon < \frac{\sigma_1 - \sigma_2}{\sqrt{|\bar{\Omega}|}}\sqrt{1 - \frac{\sigma_1^\Omega + \eta^2}{\sigma_1}}. \tag{9}$$

Remark. This corollary reveals that the benefit of calibration for anchor shift, especially in the case of strong alignment among modalities ($\uparrow \sigma_1 - \sigma_2$) and relatively small contribution of missing modalities ($\downarrow \frac{\sigma_1^\Omega + \eta^2}{\sigma_1}$).

The convergence analysis. CalMRL utilizes the bi-step iteration to optimize the generative parameters, which raises concerns about its convergence. To answer this, we confirm the monotonicity of the log-likelihood as follows.

Corollary 4 (Monotonicity for CalMRL imputation)**.** *Let $\widehat{\boldsymbol{\theta}}^{(t)}$ be the parameters at iteration t applied to the observed-data log-likelihood $L(\widehat{\boldsymbol{\theta}}) = \sum_n \log p(\mathbf{z}_n^{\Omega} \mid \widehat{\boldsymbol{\theta}})$ under the generative model in Eq. (2). Given the solution of q and $\widehat{\boldsymbol{\theta}}$, $L(\widehat{\boldsymbol{\theta}}^{(t+1)}) \geq L(\widehat{\boldsymbol{\theta}}^{(t)})$.*

Remark. The monotonic increase in likelihood guarantees stable convergence of the generative model, generally to a local stationary point (see proof in Appendix A.4). In CalMRL, this subroutine operates within a larger alternating-optimization framework that updates the unimodal encoders, thereby providing a solid foundation for the overall convergence of the full training procedure.

Algorithm 1 CalMRL (Training)

Require: Multimodal dataset $\{\mathbf{x}_i\}_{i=1}^N$; encoder parameters $\boldsymbol{\theta}$; generative parameters $\widehat{\boldsymbol{\theta}}$; τ, τ'; α
Ensure: $\phi_{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$

Initialize parameters $\phi_{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\theta}}$
for each batch $\mathcal{B} = \{\mathbf{x}_i\}_{i=1}^N$ **do**
  Encode: $\mathbf{z}_i^m = \phi_{\boldsymbol{\theta}}^m(\mathbf{x}_i^m),\ \forall m \in \Omega_i$
  Oracle modalities via imputation
    Update: $\mathbf{m}_i, \mathbf{V}_i \leftarrow$ Eq. (5), $\widehat{\boldsymbol{\theta}} = \{\mathbf{W}^m, \boldsymbol{\mu}^m, \sigma^m\}_{\Omega} \leftarrow$ Eq. (6)
    Impute: $\mathbf{z}_i^{m'} \leftarrow \mathbf{W}^{m'}\mathbf{m}_i + \boldsymbol{\mu}^{m'}, m' \in \mathcal{M} \setminus \Omega_i$
  Update $\boldsymbol{\theta}$: $\boldsymbol{\theta} \leftarrow \mathcal{L}_{\text{rep}}$ (Eq. (9))
end for

4.2 Training Algorithm Flow

To implement and clarify our method, we provide the algorithm flow for training in Algorithm 1. Building upon our theoretical derivation, our method is straightforward, applying a few steps to update the parameters in batches of data. Observed unimodal content will be encoded and then used to determine the posterior, which also makes the generative parameters resolvable within this multimodal learning framework.

5 Experiments

In this section, we conduct experiments to address the following research question:

- **RQ1:** Does CalMRL outperform other multimodal representation methods under the missing-modality setting?
- **RQ2:** What is the contribution of CalMRL to different missing modalities? Does it satisfy the theoretical insight to calibrate for a better alignment with empirical results?
- **RQ3:** How is the training of CalMRL stable? Can it maintain the distribution of multimodal representations?

5.1 Experimental Setup

Datasets. We first adopt the VAST-150K with complete modalities for warming up parameters, following [14, 15]. Then we train the model on datasets with missing modality. Specifically, we select two datasets: MSRVTT [44] for vision-text pairs and AudioCaps [28] for audio-text pairs, covering two mainstream types of multimodal datasets (*i.e.*, V↔T, A↔T). To evaluate our method, we select 6 benchmarks, including MSR-VTT [75], DiDeMo [76], ActivityNet

Table 1: **Multimodal retrieval results (%) in terms of Recall@1** on video-text (T→V and V→T) and audio-text (T→A and A→T) datasets. "↑" indicates the model continually trained with missing modality datasets. The best result in each case is marked in bold, and the second-best result is underlined. Increment points are computed compared with VAST.

	MSR-VTT		DiDeMo		ActivityNet		VATEX		AudioCaps		Clotho		Avg.
	T→V	V→T	T→V	V→T	T→V	V→T	T→V	V→T	T→A	A→T	T→A	A→T	
ImageBind	36.8	-	-	-	-	-	-	-	9.3	-	6.0	-	-
InternVideo-L	40.7	39.6	31.5	33.5	30.7	31.4	49.5	69.5	-	-	-	-	-
LanguageBind	44.8	40.9	39.9	39.8	41.0	39.1	-	-	19.7	-	16.7	-	-
VAST	50.5	49.0	48.6	46.9	51.7	48.8	75.9	74.8	33.7	32.2	12.4	13.0	44.8
GRAM	52.1$^{+1.6}$	51.8$^{+2.8}$	53.1$^{+4.5}$	50.7$^{+3.8}$	54.5$^{+2.8}$	48.3$^{-0.5}$	77.5$^{+1.6}$	74.7$^{-0.1}$	34.6$^{+0.9}$	35.2$^{+3.0}$	15.9$^{+3.5}$	16.2$^{+3.2}$	47.1
TRIANGLE	54.3$^{+3.8}$	51.7$^{+2.7}$	53.4$^{+4.8}$	52.7$^{+5.8}$	55.4$^{+3.7}$	50.9$^{+2.1}$	79.9$^{+4.0}$	74.8$^{+0.0}$	37.2$^{+3.5}$	37.2$^{+5.0}$	15.3$^{+2.9}$	13.7$^{+0.7}$	48.0
PMRL	55.1$^{+4.6}$	53.5$^{+4.5}$	53.5$^{+4.9}$	51.3$^{+4.4}$	56.0$^{+4.3}$	49.6$^{+0.8}$	80.5$^{+4.6}$	75.2$^{+0.4}$	36.1$^{+2.4}$	33.9$^{+1.7}$	16.8$^{+4.4}$	16.1$^{+3.1}$	48.1
VAST↑	58.5$^{+8.0}$	60.2$^{+11.2}$	53.9$^{+5.3}$	53.1$^{+6.2}$	55.7$^{+4.0}$	53.9$^{+5.1}$	80.0$^{+4.1}$	77.9$^{+3.1}$	49.1$^{+15.4}$	**53.3**$^{+21.1}$	21.8$^{+9.4}$	21.8$^{+8.8}$	53.3
GRAM↑	59.7$^{+9.2}$	57.2$^{+8.2}$	54.8$^{+6.2}$	53.1$^{+6.2}$	56.2$^{+4.5}$	53.5$^{+4.7}$	80.5$^{+4.6}$	**79.2**$^{+4.4}$	49.1$^{+15.4}$	51.7$^{+19.5}$	20.6$^{+8.2}$	19.5$^{+6.5}$	52.9
TRIANGLE↑	57.6$^{+7.1}$	58.4$^{+9.4}$	51.7$^{+3.1}$	51.1$^{+4.2}$	54.2$^{+2.5}$	51.0$^{+2.2}$	77.9$^{+2.0}$	76.6$^{+1.8}$	48.3$^{+14.6}$	51.7$^{+19.5}$	19.9$^{+7.5}$	20.2$^{+7.2}$	51.6
PMRL↑	60.1$^{+9.6}$	59.2$^{+10.2}$	55.1$^{+6.5}$	53.3$^{+6.4}$	55.8$^{+4.1}$	**54.0**$^{+5.2}$	80.4$^{+4.5}$	78.7$^{+3.9}$	**50.4**$^{+16.7}$	52.0$^{+19.8}$	23.5$^{+11.1}$	**23.1**$^{+10.1}$	53.8
CalMRL↑	**61.1**$^{+10.6}$	**61.1**$^{+12.1}$	**55.4**$^{+6.8}$	**53.7**$^{+6.8}$	**57.1**$^{+5.4}$	53.6$^{+4.8}$	**81.3**$^{+5.4}$	**79.2**$^{+4.4}$	50.1$^{+16.4}$	51.0$^{+18.8}$	**23.8**$^{+11.4}$	22.4$^{+9.4}$	**54.2**

[77], and VATEX [78] for vision-text retrieval, and AudioCaps [28] and Clotho [29] for audio-text retrieval. Only test splits are used for evaluation to avoid any information leakage. We detail the statistics of datasets in Appendix B.1.

Baselines & evaluation metrics. We compare our method with existing well-trained models, including ImageBind [13], InternVideo [79], LanguageBind [31], VAST [14], GRAM [15], TRIANGLE [24], and PMRL [18]. We detail the implementation of baselines, especially their adaptation for missing modalities in Appendix B.3. Performance is evaluated using Recall@1. Results about Recall@$\{5, 10\}$ are detailed in Appendix C, with results reported for bidirectional retrieval (*e.g.*, T→V and V→T).

Implementation details. We implement our model upon VAST [14] which supports four modalities, *i.e.*, vision, caption, audio, and subtitle. The detailed model architecture is in Appendix B.2. VAST, GRAM, TRIANGLE, and PMRL are all continually trained on MST-VTT and AudioCaps training datasets, aligning with the missing modality training to ensure a fair comparison with CalMRL. We also report their base performance only trained on the full modality dataset (VAST) in our empirical results (*cf.*, Section 5.3).

5.2 Performance Comparison (RQ1)

We report the results on six datasets, covering two main multimodal retrieval tasks, as shown in Table 1. All the models are trained on a complete-modality dataset initially, and "↑" denotes the continual training on missing modality datasets. Accordingly, we draw the following conclusions:

① *Learning with complete-modality simultaneously achieves higher performance.* Initial models, *e.g.*, ImageBind, InternVideo-L, and LanguageBind, extend the pairwise contrastive learning paradigm. However, they generally perform worse than models trained on datasets where all modalities are available simultaneously. New learning objectives have been introduced specifically to optimize multiple modalities concurrently. Among these methods, PMRL demonstrates relatively better performance. The aligned modalities naturally reflect different aspects of a single common instance. This inherent alignment intuitively provides significant benefits for multimodal representation learning.

② *Missing modalities can boost the model yet incrementally.* When certain modalities are missing, virtually all models demonstrate performance improvements, particularly on in-domain datasets (*i.e.*, MSR-VTT and AudioCaps). Observable performance gains are also generalized to various out-of-domain datasets, such as DiDeMo and ActivityNet, though these gains are incremental. The performance boost on audio-text datasets (*i.e.*, AudioCaps and Clotho) is significantly greater compared to the boost observed on vision-text alignment tasks. This disparity indicates that

Table 2: **Multimodal retrieval results (%) for models trained on sole datasets.** "$\uparrow^{VT}$" and "$\uparrow^{AT}$" indicate the model continually trained with video-text and audio-text modality datasets, respectively. The best result in each case is marked in bold, and the second-best result is underlined. Increment points are computed compared with VAST.

	MSR-VTT		DiDeMo		ActivityNet		VATEX		AudioCaps		Clotho		Avg.
	T→V	V→T	T→V	V→T	T→V	V→T	T→V	V→T	T→A	A→T	T→A	A→T	
VAST$\uparrow^{AT}$	53.1$^{+2.6}$	50.9$^{+1.9}$	45.0$^{-3.6}$	46.0$^{-0.9}$	48.8$^{-2.9}$	48.5$^{-0.3}$	77.3$^{+1.4}$	75.9$^{+1.1}$	51.1$^{+17.4}$	52.1$^{+19.9}$	21.3$^{+8.9}$	21.1$^{+8.1}$	49.3
GRAM$\uparrow^{AT}$	49.0$^{-1.5}$	49.3$^{+0.3}$	48.5$^{-0.1}$	48.3$^{+1.4}$	49.2$^{-2.5}$	48.0$^{-0.8}$	58.1$^{-17.8}$	74.1$^{-0.7}$	53.0$^{+19.3}$	52.1$^{+19.9}$	22.6$^{+10.2}$	**22.5**$^{+9.5}$	47.9
TRIANGLE$\uparrow^{AT}$	55.8$^{+5.3}$	52.4$^{+3.4}$	50.1$^{+1.5}$	48.9$^{+2.0}$	50.2$^{-1.5}$	**50.0**$^{+1.2}$	**79.8**$^{+3.9}$	76.1$^{+1.3}$	47.0$^{+13.3}$	50.9$^{+18.7}$	16.6$^{+4.2}$	18.4$^{+5.4}$	49.7
PMRL$\uparrow^{AT}$	56.2$^{+5.7}$	52.7$^{+3.7}$	48.7$^{+0.1}$	49.4$^{+2.5}$	52.0$^{+0.3}$	49.3$^{+0.5}$	78.8$^{+2.9}$	**76.6**$^{+1.8}$	52.0$^{+18.3}$	54.0$^{+21.8}$	**22.7**$^{+10.3}$	21.8$^{+8.8}$	51.2
CalMRL$\uparrow^{AT}$	**56.4**$^{+5.9}$	**53.3**$^{+4.3}$	**50.5**$^{+1.9}$	**50.7**$^{+3.8}$	**53.8**$^{+2.1}$	49.8$^{+1.0}$	79.2$^{+3.3}$	**76.6**$^{+1.8}$	**53.1**$^{+19.4}$	**54.1**$^{+21.9}$	21.3$^{+8.9}$	21.6$^{+8.6}$	**51.7**
VAST$\uparrow^{VT}$	59.5$^{+9.0}$	**60.3**$^{+11.3}$	54.6$^{+6.0}$	**55.1**$^{+8.2}$	**57.6**$^{+5.9}$	**54.9**$^{+6.1}$	80.1$^{+4.2}$	78.3$^{+3.5}$	31.5$^{-2.2}$	33.4$^{+1.2}$	15.5$^{+3.1}$	15.5$^{+2.5}$	49.7
GRAM$\uparrow^{VT}$	60.1$^{+9.6}$	57.8$^{+8.8}$	56.5$^{+7.9}$	54.0$^{+7.1}$	56.2$^{+4.5}$	53.5$^{+4.7}$	78.9$^{+3.0}$	78.3$^{+3.5}$	**32.5**$^{-1.2}$	**36.8**$^{+4.6}$	16.1$^{+3.7}$	15.2$^{+2.2}$	49.7
TRIANGLE$\uparrow^{VT}$	57.8$^{+7.3}$	58.7$^{+9.7}$	53.1$^{+4.5}$	50.9$^{+4.0}$	56.3$^{+4.6}$	51.6$^{+2.8}$	78.4$^{+2.5}$	77.3$^{+2.5}$	31.2$^{-2.5}$	32.5$^{+0.3}$	16.3$^{+3.9}$	14.7$^{+1.7}$	48.2
PMRL$\uparrow^{VT}$	60.7$^{+10.2}$	60.0$^{+11.0}$	56.1$^{+7.5}$	53.5$^{+6.6}$	57.5$^{+5.8}$	53.2$^{+4.4}$	79.7$^{+3.8}$	78.1$^{+3.3}$	32.0$^{-1.7}$	34.8$^{+2.6}$	18.2$^{+5.8}$	**15.9**$^{+2.9}$	50.0
CalMRL$\uparrow^{VT}$	**61.1**$^{+10.6}$	**60.3**$^{+11.3}$	**57.5**$^{+8.9}$	54.4$^{+7.5}$	57.5$^{+5.8}$	52.2$^{+3.4}$	**81.5**$^{+5.6}$	**79.4**$^{+4.6}$	31.4$^{-2.3}$	34.8$^{+2.6}$	**18.9**$^{+6.5}$	15.8$^{+2.8}$	**50.4**

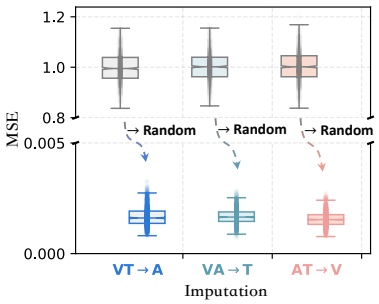

Figure 3: **MSEs between real and imputed representations.** "→" marks the direction of imputation; **Random** refers to representations drawn at random.

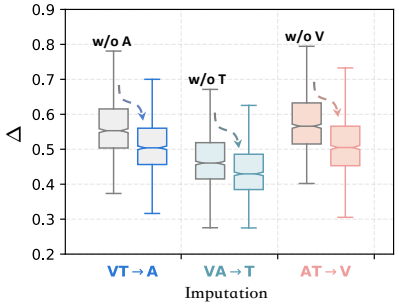

Figure 4: **Comparison of anchor shift (Δ) before and after calibration.** The left box with a gray border shows the anchor shift with missing modalities (w/o).

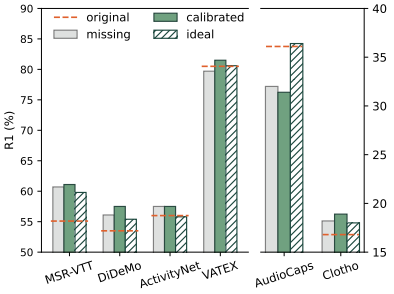

Figure 5: **The performance comparison across missing, calibrated, and full ("ideal") modalities.** All the models are trained on MSR-VTT.

vision-text alignment has benefited from extensive prior research and advancements, whereas audio-text alignment appears to have substantially greater room for improvement (corresponds to the visualization in Figure 6).

③ *CalMRL further improves without incorporating new information.* Among all the methods, CaLMRL demonstrates superior performance improvements over its backbone model VAST and surpasses the SOTA methods in most cases. Rather than requiring new datasets, CaLMRL adapts existing bimodal datasets to enhance the model using a simple generative approach, showcasing strong promise.

5.3 Further Analysis (R2 & R3)

Sole datasets with calibration (R2). To unravel the effect of different missing modality datasets, we conduct experiments on a single dataset with calibration. Table 2 showcases the performance evaluated on different variants of models. "$\uparrow^{AT}$" indicates the training with only audio-text dataset (*i.e.*, AudioCaps). The results for all given benchmarks are reported, and we have the following insights. ① *Training on the sole dataset can improve the relevant capabilities of modalities while harming others.* For instance, training on an audio-text dataset, we can observe a significant improvement on AudioCpas and Clotho (out-of-domain), and a performance drop on other video-text datasets, and vice versa. Such harmfulness can be found clearly under only the vision-text dataset training. ② *CalMRL heals to some extent and even sometimes achieves better performance.* In audio-text training, CalMRL resists performance degradation in most vision-text benchmarks, and even gains better results in AudioCaps. Combining the results in Table 1, CalMRL demonstrates the collective benefit of training on a mixture of datasets, resulting in improved per-

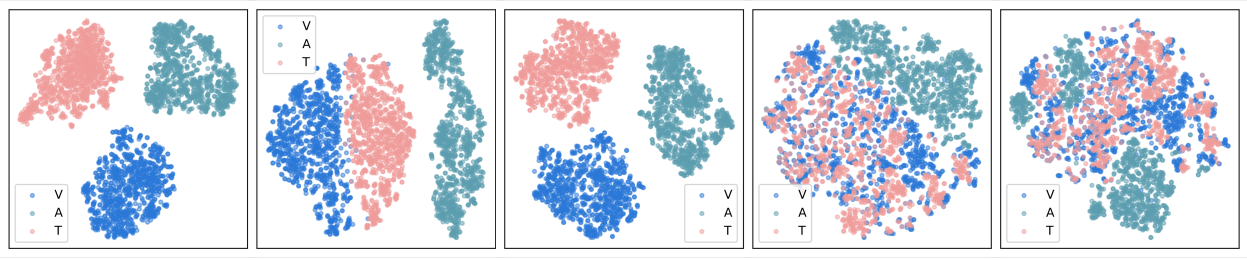

Figure 6: **t-SNE visualization on multimodal representations generated by different models.** Existing models, under missing modality training, present clearly separated clusters for each modality (distinct modal boundaries). Fortunately, CalMRL mitigates this issue.

formance. These results provide strong evidence to shed light on how CalMRL works: calibrating the alignment in oracle to resist the degradation for missing modalities to maintain an overall better performance.

Compensating for the anchor shift (R2). We measure the anchor shift before and after calibration to validate its effectiveness. Figures 3 and 4 illustrate that ① *our method can effectively reconstruct the original representations* and ② *the anchor shift is mitigated as expected*. This indicates that our method, even with a simple generative model, is capable of learning multimodal connections at the representation level. Without introducing any new information, the incomplete alignment can be approximated to the complete one via the mitigated anchor shift.

Approaching to the *"ideal"* anchor (R2). We conduct experiments where the model was trained on a dataset with synthetic complete modalities to mimic the complete-modality scenario. Figure 5 shows the comparison of the variant without calibration and the one with complete modalities on the MSR-VTT. Compared to the original performance, training on complete modalities can best maintain the alignment, achieving results around or above the red line. CalMRL outperforms the missing-modality variant in most cases and even the "ideal" (complete-modality training) in the VATEX dataset.

Training stability (R3). We plot the curves of the training loss $\mathcal{L}_{\text{rep}}$ for different models in Figure 7. VAST, GRAM, and TRIANGLE all utilize text-anchored contrastive learning, which leads to relatively larger losses, while CalMRL follows PMRL, optimizing with smaller ones. Overall, we can observe that the loss curve of CalMRL is more stable and exhibits a steadier decreasing trend compared to the others. Despite employing bi-step learning for missing modalities, CalMRL is capable of managing the training and achieving convergence.

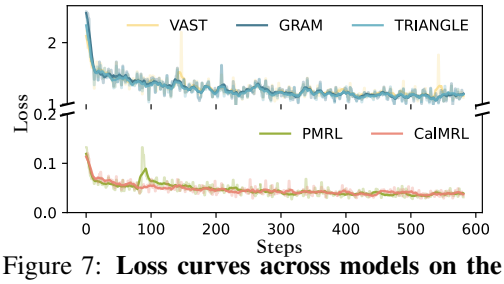

Figure 7: **Loss curves across models on the training phase.**

Case visualization (R3). To intuitively understand the inner workings of CalMRL, we visualize the representations as points within a 2D plane to analyze their distributions. As shown in Figure 6, VAST, GRAM, and TRIANGLE demonstrate distinct, well-separated regions for each modality. While PMRL successfully aligns the vision and text representations, it excludes the audio representation with a clear dividing line. In contrast, CalMRL alleviates this separation, pulling the multiple modalities closer together. We attribute this improvement to the alignment calibration, the core design philosophy of CalMRL, which effectively prevents the segregation of different modal representations.

6 Conclusion

In this paper, we presented CalMRL, a calibrated multimodal representation learning framework for handling missing modalities. We theoretically revealed the anchor shift phenomenon and proposed a generative imputation mechanism

to reconstruct oracle representations and calibrate alignment. Through a bi-step optimization with closed-form inference and iterative refinement, CalMRL achieves stable convergence and strong compatibility with existing alignment paradigms. Experiments across diverse benchmarks verify its superiority. Looking ahead, future explorations can extend our method to broader applications and focus on: calibrating the shift without imputation, incorporating more datasets to enhance multimodal models, and interpreting multimodal connections to inspire further advancement.

Appendix

A Theoretical Analysis

In this section, we detail the proof of Theorem 1 (*cf.*, Appendix A.1), the derivation of the closed-form solution of $\widehat{\boldsymbol{\theta}}$ (*cf.*, Appendix A.2), the anchor shift after calibration (*cf.*, Appendix A.3), and the convergence analysis (*cf.*, Appendix A.4).

A.1 Proof of Theorem 1

Recall. Let $\mathbf{u}_1$ and $\mathbf{u}_1^\Omega$ be the leading left singular vectors of the full multimodal matrix $\mathbf{Z}$ and its observed submatrix $\mathbf{Z}^\Omega$, respectively. Define $\sigma_1 = \|\mathbf{Z}\|_2$, $\sigma_1^\Omega = \|\mathbf{Z}^\Omega\|_2$, and $\eta := \sqrt{\sum_{m\in\bar{\Omega}}\langle \mathbf{u}_1^\Omega, \mathbf{z}^m\rangle^2}$. Then the anchor shift satisfies

$$\sqrt{2\left(1-\frac{\sigma_1^\Omega+\eta^2}{\sigma_1}\right)} \;\le\; \underbrace{\|\mathbf{u}_1-\mathbf{u}_1^\Omega\|}_{\|\Delta\|} \;\le\; \frac{\sqrt{2}\,\|\mathbf{Z}^{\bar{\Omega}}\|_2}{\sqrt{\lambda_1}-\sqrt{\lambda_2}}.$$

Proof. Let multimodal representation is expressed by the concatenation of different unimodal representations:

$$\mathbf{Z} = [\mathbf{z}^{m_1}, \mathbf{z}^{m_2}, \ldots, \mathbf{z}^{m_k}] \in \mathbb{R}^{d\times k}, \tag{10}$$
$$= \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^\top, \tag{11}$$
$$\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_d] \in \mathbb{R}^{d\times d}, \tag{12}$$
$$\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1, \ldots, \sigma_{\min\{d,k\}}), \tag{13}$$

where $\sigma_1 \ge \sigma_2 \ge \cdots \ge 0$. We define the virtual anchor as $\mathbf{u}_1$, which is the consensus direction in the spanned by multimodalities.

According to the Davis–Kahan Theorem [80], we have:

$$\sin\angle(\mathbf{u}_1, \mathbf{u}_1^\Omega) \le \frac{\|\mathbf{Z}^{\bar{\Omega}}\|_2}{\delta}, \tag{14}$$

where $\delta = \sigma_1 - \sigma_2$ represents the spectral gap and $\|\mathbf{Z}^{\bar{\Omega}}\|_2 = \sigma_{\max}(\mathbf{Z}^{\bar{\Omega}})$ denotes the spectral norm of the submatrix associated with the missing modalities.

$$\|\boldsymbol{\Delta}\| := \|\mathbf{u}_1 - \mathbf{u}_1^\Omega\| = [(\mathbf{u}_1-\mathbf{u}_1^\Omega)^\top(\mathbf{u}_1-\mathbf{u}_1^\Omega)]^{-\frac{1}{2}} \tag{15}$$
$$= [\|\mathbf{u}_1\|^2 + \|\mathbf{u}_1^\Omega\|^2 - 2\mathbf{u}_1^\top\mathbf{u}_1^\Omega]^{-\frac{1}{2}} \tag{16}$$
$$= [2-2\cos\theta]^{-\frac{1}{2}} \tag{17}$$
$$= [4\sin^2(\theta/2)]^{-\frac{1}{2}} \qquad (1-\cos\theta = 2\sin^2(\theta/2)) \tag{18}$$
$$= 2|\sin(\theta/2)| = 2\sin(\theta/2) \qquad (\theta \le \pi/2) \tag{19}$$
$$\le \sqrt{2}\sin(\theta) \tag{20}$$
$$\le \frac{\sqrt{2}\|\mathbf{Z}^{\bar{\Omega}}\|_2}{\sigma_1-\sigma_2} \qquad (\text{Eq. (14)}) \tag{21}$$
$$= \frac{\sqrt{2}\|\mathbf{Z}^{\bar{\Omega}}\|_2}{\sqrt{\lambda_1}-\sqrt{\lambda_2}} \qquad (\sigma_1(\mathbf{Z}) = \lambda_1(\mathbf{G})) \tag{22}$$
$$= \frac{\sqrt{2}\|\mathbf{Z}^{\bar{\Omega}}\|_2}{\mathcal{A}} \qquad (\mathcal{A} := \sqrt{\lambda_1}-\sqrt{\lambda_2}) \tag{23}$$

$$\leq \frac{\sqrt{2|\bar{\Omega}|}}{\mathcal{A}} \qquad (\|\mathbf{Z}^{\bar{\Omega}}\|_2 \leq \sqrt{|\bar{\Omega}|}) \tag{24}$$

This upper bound indicates that a better alignment can lead to better robustness for the case of missing modality.

Now we derive the lower bound as follows:

$$\begin{aligned}
\|\boldsymbol{\Delta}\| := \|\mathbf{u}_1 - \mathbf{u}_1^{\Omega}\| &= [2 - 2\cos\theta]^{1/2} && (25)\\
&\geq [2(1-\cos^2\theta)]^{1/2} && (26)\\
&= [2(1 - |\langle \mathbf{u}_1, \mathbf{u}_1^{\Omega}\rangle|^2)]^{1/2} && (27)\\
&\geq [2 - 2\frac{\|\mathbf{u}_1^{\Omega\top}\mathbf{Z}\|_2^2}{\sigma_1}]^{1/2} && (28)\\
&= [2 - 2\frac{\|\mathbf{u}_1^{\Omega\top}\mathbf{Z}^{\Omega}\|_2^2 + \|\mathbf{u}_1^{\Omega\top}\mathbf{Z}^{\bar{\Omega}}\|_2^2}{\sigma_1}]^{1/2} && (29)\\
&= [2 - 2\frac{(\sigma_1^{\Omega})^2 + \eta^2}{\sigma_1}]^{1/2} \qquad (\eta := \sqrt{\sum_{m\in\bar{\Omega}} \langle \mathbf{u}_1^{\Omega}, \mathbf{z}^m\rangle^2}) && (30)
\end{aligned}$$

This lower bound reveals that the anchor shift cannot be arbitrarily small. It is fundamentally limited by how much the missing modalities $\bar{\Omega}$ contribute to the leading singular direction. □

A.2 Resolving $\widehat{\theta}$

We consider the joint distribution of $\mathbf{z}$ and $\boldsymbol{\beta}$:

$$\begin{aligned}
\begin{bmatrix}\boldsymbol{\beta}_n \\ \mathbf{z}_n\end{bmatrix} &\sim \mathcal{N}\left(\begin{bmatrix}\mathbf{0} \\ \boldsymbol{\mu}\end{bmatrix}, \begin{bmatrix}\mathbf{I} & \mathbf{W}^\top \\ \mathbf{W} & \mathbf{W}\mathbf{W}^\top + \boldsymbol{\Sigma}\end{bmatrix}\right), && (31)\\
\boldsymbol{\mu} &= [\boldsymbol{\mu}^{1\top}, \ldots, \boldsymbol{\mu}^{|\Omega|\top}]^\top && (32)\\
\mathbf{W} &= [\mathbf{W}^{1\top}, \ldots, \mathbf{W}^{|\otimes|\top}]^\top && (33)\\
\boldsymbol{\Sigma} &= \mathrm{diag}((\sigma^1)^2\mathbf{I}, \ldots, (\sigma^{|\Omega|})^2\mathbf{I}) && (34)
\end{aligned}$$

The condition distribution is $p(\mathbf{z} \mid \mathbf{x}) = \mathcal{N}(\mathbf{z}; \mathbf{m}, \mathbf{V})$.

$$\begin{aligned}
\mathbf{V} &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}} - \boldsymbol{\Sigma}_{\boldsymbol{\beta}\mathbf{z}}\boldsymbol{\Sigma}_{\mathbf{z}}^{-1}\boldsymbol{\Sigma}_{\mathbf{z}\boldsymbol{\beta}} && (35)\\
&= \mathbf{I} - \mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top + \boldsymbol{\Sigma})^{-1}\mathbf{W} && (36)\\
&= \mathbf{I} - \mathbf{W}^\top[\boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1}\mathbf{W}(\mathbf{I}_K + \mathbf{W}^\top\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\boldsymbol{\Sigma}^{-1}]\mathbf{W} \qquad \text{(Woodbury matrix identity [81])}\\
&= \mathbf{I} - \mathbf{A} + \mathbf{A}(\mathbf{I}_K + \mathbf{A})^{-1}\mathbf{A} \quad (\mathbf{A} = \mathbf{W}^\top\boldsymbol{\Sigma}^{-1}\mathbf{W}) && (37)\\
&= \left(\mathbf{I} + \mathbf{W}^\top\boldsymbol{\Sigma}^{-1}\mathbf{W}\right)^{-1} \quad (\mathbf{A}(\mathbf{I}+\mathbf{A})^{-1} = \mathbf{I} - (\mathbf{I}+\mathbf{A})^{-1}) && (38)\\
&= \Big[\mathbf{I} + \sum_{m\in\Omega}\frac{1}{(\sigma^m)^2}\mathbf{W}^{m\top}\mathbf{W}^m\Big]^{-1} && (39)
\end{aligned}$$

$$\begin{aligned}
\mathbf{m} &= \mathbb{E}[\mathbf{z}_n \mid \mathbf{x}_n] && (40)\\
&= \mathbf{0} + \mathbf{W}^\top(\mathbf{W}\mathbf{W}^\top + \boldsymbol{\Sigma})^{-1}(\mathbf{z} - \boldsymbol{\mu}) && (41)\\
&= (\mathbf{I} + \mathbf{W}^\top\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}^\top\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu}) && (42)\\
&= \mathbf{V}\mathbf{W}^\top\boldsymbol{\Sigma}^{-1}(\mathbf{z} - \boldsymbol{\mu}) && (43)
\end{aligned}$$

$$= \mathbf{V} \sum_{m \in \Omega} \frac{1}{(\sigma^m)^2} \mathbf{W}^{m\top} (\mathbf{z}^m - \boldsymbol{\mu}^m) \tag{44}$$

Therefore, we have:

$$\begin{cases} \mathbf{V} = \Big[\mathbf{I} + \sum_{m \in \Omega} \frac{1}{(\sigma^m)^2} \mathbf{W}^{m\top} \mathbf{W}^m \Big]^{-1}, \\ \mathbf{m} = \mathbf{V} \sum_{m \in \Omega} \frac{1}{(\sigma^m)^2} \mathbf{W}^{m\top} (\mathbf{z}^m - \boldsymbol{\mu}^m), \end{cases} \tag{45}$$

The objective is updated as:

$$\mathbb{E}[\log p(\mathbf{z}, \boldsymbol{\beta} | \widehat{\boldsymbol{\theta}})] = \mathbb{E}\Big[\log p(\boldsymbol{\beta}) + \sum_{m \in \Omega} \log p(\mathbf{z}^m \mid \boldsymbol{\beta}, \widehat{\boldsymbol{\theta}}) \Big] \tag{46}$$

$$= \mathbb{E}\Big[-\frac{1}{2} \boldsymbol{\beta}^\top \boldsymbol{\beta} + \sum_{m \in \Omega} \Big[-\frac{d}{2} \log(\sigma^m)^2 - \frac{1}{2(\sigma^m)^2} \mathbb{E}[\|\mathbf{z}^m - \boldsymbol{\mu}^m - \mathbf{W}^m \boldsymbol{\beta}\|^2] \Big] + c \Big] \tag{47}$$

Let us compute the closed-form solution of $\boldsymbol{\mu}^m$ and $\mathbf{W}^m$ according to its related terms.

$$\mathbb{E}\|\mathbf{z}^m - \boldsymbol{\mu}^m - \mathbf{W}^m \boldsymbol{\beta}\|^2 = \|\mathbf{z}^m - \boldsymbol{\mu}^m - \mathbf{W}^m \mathbf{m}\|^2 + \mathrm{Tr}[\mathbf{W}^{m\top} \mathbf{W}^m \mathbf{V}]$$

$$\rightarrow \ \boldsymbol{\mu}^m = \frac{1}{N} \sum_n (\mathbf{z}_n^m - \mathbf{W}^m \mathbf{m}_n) \tag{48}$$

$$\rightarrow \ \mathbf{W}^m = \left(\sum_{n=1}^N (\mathbf{z}_n^m - \boldsymbol{\mu}^m) \mathbf{m}_n^\top \right) \left(\sum_{n=1}^N \mathbb{E}[\boldsymbol{\beta}_n \boldsymbol{\beta}_n^\top] \right)^{-1} \tag{49}$$

We can also calculate the solution for σ_m^2:

$$\sigma_m^2 = \frac{1}{Nd} \sum_{n=1}^N \big[\|\mathbf{z}_n^m - \boldsymbol{\mu}^m - \mathbf{W}^m \mathbf{m}_n\|^2 + \mathrm{Tr}[\mathbf{W}^{m\top} \mathbf{W}^m \mathbf{V}] \big] \tag{50}$$

Therefore, for missing modalities $m' \in \mathcal{M}/\Omega$ conditioned on the previous observations, we have:

$$\mathbf{z}^{m'} = \mathbf{W}^{m'} \bar{\boldsymbol{\beta}} + \boldsymbol{\mu}^{m'} + \bar{\boldsymbol{\epsilon}}^m \tag{51}$$

$$= \mathbf{W}^{m'} \mathbb{E}[\boldsymbol{\beta} | X] + \boldsymbol{\mu}^{m'} + \mathbb{E}[\boldsymbol{\epsilon}^m | X] \tag{52}$$

$$= \mathbf{W}^{m'} \mathbf{m} + \boldsymbol{\mu}^{m'} \tag{53}$$

A.3 Alleviating Anchor Shift

Let $\widehat{\mathbf{Z}}^{\bar{\Omega}} = [\mathbf{W}^{m'} \mathbf{m} + \mu^{m'} | m \in \mathcal{M}/\Omega]$.

$$\|\Delta_{\text{cal}}\| = \|\mathbf{u}_1 - \mathbf{u}_1^{\text{cal}}\| \tag{54}$$

$$\leq \frac{\sqrt{2} \|[\mathbf{Z}^\Omega, \widehat{\mathbf{Z}}^{\bar{\Omega}}] - \mathbf{Z}\|_2}{\sigma_1 - \sigma_2} \tag{55}$$

$$\leq \frac{\sqrt{2} \|[\mathbf{Z}^\Omega, \widehat{\mathbf{Z}}^{\bar{\Omega}}] - \mathbf{Z}\|_F}{\sigma_1 - \sigma_2} \tag{56}$$

$$\leq \frac{\sqrt{2|\bar{\Omega}|}\varepsilon}{\sigma_1 - \sigma_2} \qquad (\|\hat{\mathbf{z}}^{m'} - \hat{\mathbf{z}}^{m'}\|_2 \leq \varepsilon, \forall m' \in \bar{\Omega})$$

Here ε represents the imputation error at the representation level. From Theorem 1, we have:

$$\|\Delta\| \geq \sqrt{2(1 - \frac{\sigma_1^{\Omega} + \eta^2}{\sigma_1})}. \tag{57}$$

Therefore, we can derive the condition for $\|\Delta_{\text{cal}}\| \leq \|\Delta\|$ as follows:

$$\varepsilon < \frac{\sigma_1 - \sigma_2}{\sqrt{|\bar{\Omega}|}} \cdot \sqrt{1 - \frac{\sigma_1^{\Omega} + \eta^2}{\sigma_1}}. \tag{58}$$

This also provides a consistent conclusion to alleviate the anchor shift. The calibration is beneficial, especially in the case of strong alignment among modalities ($\uparrow \sigma_1 - \sigma_2$) and greater original shift ($\downarrow \frac{\sigma_1^{\Omega} + \eta^2}{\sigma_1}$).

A.4 Convergence Analysis.

Recall. Let $\widehat{\boldsymbol{\theta}}^{(t)}$ be the parameters at iteration t applied to the observed-data log-likelihood $L(\widehat{\boldsymbol{\theta}}) = \sum_n \log p(\mathbf{z}_n^{\Omega} \mid \widehat{\boldsymbol{\theta}})$ under the generative model in Eq. (2). Given the solution of q and $\widehat{\boldsymbol{\theta}}$, $L(\widehat{\boldsymbol{\theta}}^{(t+1)}) \geq L(\widehat{\boldsymbol{\theta}}^{(t)})$.

Proof. Let $L(\widehat{\boldsymbol{\theta}}^{(t+1)})$ be the log-likelihood for parameters at $t+1$ step.

$$L(\widehat{\boldsymbol{\theta}}^{(t+1)}) = \sum_{n=1}^{N} \log p(\mathbf{z}_n^{\Omega} | \widehat{\boldsymbol{\theta}}^{(t+1)}) \tag{59}$$

$$= \sum_{n=1}^{N} \log \int p(\mathbf{z}_n^{\Omega}, \boldsymbol{\beta}_n | \widehat{\boldsymbol{\theta}}^{(t+1)}) d\boldsymbol{\beta}_n \tag{60}$$

$$= \sum_{n=1}^{N} \log \mathbb{E}_{q_n}\Big[\frac{p(\mathbf{z}_n^{\Omega}, \boldsymbol{\beta}_n | \widehat{\boldsymbol{\theta}}^{(t+1)})}{q_n^{(t+1)}(\boldsymbol{\beta}_n)}\Big] \qquad (\mathbf{m}_n, \mathbf{V}_n \leftarrow \widehat{\boldsymbol{\theta}}^{(t)})$$

$$\geq \sum_{n=1}^{N} \mathbb{E}_{q_n}\Big[\log \frac{p(\mathbf{z}_n^{\Omega}, \boldsymbol{\beta}_n | \widehat{\boldsymbol{\theta}}^{(t+1)})}{q_n^{(t+1)}(\boldsymbol{\beta}_n)}\Big] \tag{61}$$

$$\geq \sum_{n=1}^{N} \mathbb{E}_{q_n}\Big[\log \frac{p(\mathbf{z}_n^{\Omega}, \boldsymbol{\beta}_n | \widehat{\boldsymbol{\theta}}^{(t)})}{q_n^{(t+1)}(\boldsymbol{\beta}_n)}\Big] := \mathcal{Q} \qquad (\widehat{\boldsymbol{\theta}}^{(t+1)} = \arg\max_{\widehat{\boldsymbol{\theta}}} \mathcal{Q}(q^{(t+1)}, \widehat{\boldsymbol{\theta}}))$$

$$= L(\widehat{\boldsymbol{\theta}}^{(t)}) - \underbrace{\text{KL}[q_n^{(t+1)}(\boldsymbol{\beta}_n) \| p(\boldsymbol{\beta}_n | \mathbf{z}_n^{\Omega}, \widehat{\boldsymbol{\theta}})]}_{0} \qquad (q(\boldsymbol{\beta}) = p(\boldsymbol{\beta} | \mathbf{z}, \widehat{\boldsymbol{\theta}}))$$

$$= L(\widehat{\boldsymbol{\theta}}^{(t)}) \tag{62}$$

Therefore, we have $L(\widehat{\boldsymbol{\theta}}^{*}) \geq \cdots \geq L(\widehat{\boldsymbol{\theta}}^{(t+1)}) \geq L(\widehat{\boldsymbol{\theta}}^{(t)}) \geq \cdots \geq L(\widehat{\boldsymbol{\theta}}^{(0)})$, indicating our method can converges to a stationary point of $L(\widehat{\boldsymbol{\theta}}^{*})$. □

Table 3: The statistics of datasets.

Benchmark	Modalities			#Train	#Test
	Video	Audio	Text		
MSR$\uparrow^{VT}$T	✓	-	✓	180,000	1,000
DiDeMo	✓	-	✓	-	1,003
ActivityNet	✓	-	✓	-	3,987
VATEX	✓	-	✓	-	1,225
AudioCaps	-	✓	✓	45,178	704
Clotho	-	✓	✓	-	1,045

B Implementation Details

B.1 Datasets

We employ the training dataset *VAST-150K* [15] at the warm-up stage to avoid a cold start. This dataset is a downsized version of VAST-27M [14] and used for previous works [15, 18]. VAST-27M involves four modalities, *i.e.*, video, audio, caption, and subtitle for each example. We also adopt the training splits of *MSR-VTT* [75] (around 180K, where 20 captions for each video) for vision-text and *AudioCaps* [28] (around 45K) for audio-text for the missing modality training. For benchmarking, we adopt several datasets with their test splits, including vision-text datasets *MSR-VTT* [75], *DiDeMo* [76], *ActivityNet* [77], and *VATEX* [78], and audio-text datasets *AudioCaps* [28] and *Clotho* [29]. All videos are sampled into 8 frames. For the testing splits, the complete modalities are provided [14]. The statistics of these benchmarks are shown in Table 3.

B.2 Model Architecture

We build our model based on VAST [14] to ensure a fair comparison rather than advancing the architecture design. Specifically, we construct the vision encoder with EVAClip-ViT-G [82], where the vision resolution is set to 224×224 pixels. BERT is utilized to implement the text encoder with a maximum caption length limited to 40. For subtitles, the maximum length is extended to 70. BEATs model [83] is adopted for audio encoding. Each audio is preprocessed into 63 mel-frequency bins, outputting 1024 frames.

B.3 Adaptating Baselines for Missing Modalities

For comparison with baselines under the missing-modality scenario, we adapt these methods with the following minor modifications. For the GRAM [15] and PMRL [18] methods, we can still calculate the GRAM matrix for the observed modalities. For instance, if given a V-T dataset, the GRAM matrix $\mathbf{G}$ should be 2-by-2. In this case, GRAM only needs to minimize one singular value, thus approaching PMRL. However, GRAM sets text as the fixed anchor modality by default, which may limit its optimization. For TRIANGLE [24], it is explicitly designed for datasets with three modalities since it optimizes the area of a triangle spanned by the three modalities. Therefore, for datasets with two modalities, we change it to maximize the cosine similarity between the two related modalities for stable training.

B.4 Warm-up Training

We employ the warm-up training for initialized generative parameters to avoid the cold-start issue. VAST-150K with complete modalities present is selected. To better mimic the missing modality setting, during the warming-up, we randomly choose one modality as the missing and the others as the observations. We only train the model in one epoch.

Table 4: Multimodal retrieval results (%) for models trained on full datasets and sole datasets with respect to Recall@5.

	MSR-VTT		DiDeMo		ActivityNet		VATEX		AudioCaps		Clotho	
	T→V	V→T	T→V	V→T	T→V	V→T	T→V	V→T	T→A	A→T	T→A	A→T
VAST ↑	76.8	81.8	74.6	75.6	80.2	80.8	95.7	95.8	81.5	84.7	49.0	45.1
GRAM ↑	78.7	81.1	75.3	75.7	80.0	80.3	96.7	96.8	82.4	83.9	47.2	43.6
Triangle ↑	78.4	79.9	73.9	74.9	79.3	79.4	94.5	95.3	80.4	84.1	42.8	45.0
PMRL ↑	79.6	78.8	74.9	76.5	80.9	81.3	95.6	96.2	81.7	83.5	49.0	47.6
CalMRL ↑	83.0	80.5	75.6	76.2	80.7	81.6	96.7	96.7	80.8	82.4	47.0	46.0
VAST$\uparrow^{AT}$	71.8	72.1	65.8	68.4	72.6	74.2	94.4	94.1	81.2	82.5	47.0	46.5
GRAM$\uparrow^{AT}$	67.1	73.0	70.3	70.2	73.2	74.0	66.7	92.8	84.7	85.1	49.0	46.6
Triangle$\uparrow^{AT}$	76.1	74.3	71.6	73.6	76.2	76.4	95.2	95.2	79.3	81.4	41.1	43.5
PMRL$\uparrow^{AT}$	77.1	73.8	70.4	72.3	77.0	76.1	94.9	94.8	84.5	85.2	49.9	48.0
CalMRL$\uparrow^{AT}$	76.9	74.2	71.7	72.8	78.1	76.7	95.3	94.5	82.8	83.1	45.8	46.5
VAST$\uparrow^{VT}$	77.0	82.1	76.8	76.8	82.0	81.8	96.0	96.7	64.3	66.2	32.8	34.0
GRAM$\uparrow^{VT}$	79.9	82.4	76.6	78.1	80.5	80.6	97.1	96.8	66.2	67.9	34.5	34.4
Triangle$\uparrow^{VT}$	78.8	81.2	75.6	75.8	81.0	80.0	95.3	95.6	61.5	64.2	35.2	34.2
PMRL$\uparrow^{VT}$	79.8	81.3	76.8	76.6	82.3	81.3	96.3	96.2	63.1	66.9	38.7	35.6
CalMRL $\uparrow^{VT}$	82.5	80.9	76.9	77.9	82.1	81.0	96.5	97.3	64.9	66.6	39.0	36.1

Table 5: Multimodal retrieval results (%) for models trained on full datasets and sole datasets with respect to Recall@10.

	MSR-VTT		DiDeMo		ActivityNet		VATEX		AudioCaps		Clotho	
	T→V	V→T	T→V	V→T	T→V	V→T	T→V	V→T	T→A	A→T	T→A	A→T
VAST ↑	81.4	85.7	79.2	80.2	86.8	88.5	97.3	97.1	89.9	92.9	60.4	59.5
GRAM ↑	86.0	86.9	80.2	81.0	86.9	87.9	98.2	98.4	90.8	94.2	60.4	57.9
Triangle ↑	83.8	85.1	79.1	79.6	86.8	87.4	95.8	97.1	90.1	93.6	57.0	59.1
PMRL ↑	84.5	85.2	79.6	81.6	88.0	89.0	97.0	97.3	90.3	92.8	58.9	59.7
CalMRL ↑	87.3	87.1	80.5	82.0	87.9	89.5	98.1	98.0	88.2	91.3	58.5	59.1
VAST$\uparrow^{AT}$	78.4	79.8	72.3	76.5	79.8	82.5	95.8	95.4	91.2	92.9	60.8	60.0
GRAM$\uparrow^{AT}$	72.4	79.1	74.8	77.2	80.0	82.5	67.1	94.9	93.3	94.3	61.5	60.6
Triangle$\uparrow^{AT}$	82.6	81.1	77.3	79.3	83.1	84.7	96.6	96.3	88.8	90.2	55.1	58.1
PMRL$\uparrow^{AT}$	82.7	80.7	75.9	78.9	84.1	85.2	96.1	96.3	92.0	92.8	61.1	61.8
CalMRL$\uparrow^{AT}$	83.3	81.1	77.2	78.7	85.3	85.7	96.2	96.6	90.5	91.5	59.0	59.0
VAST$\uparrow^{VT}$	82.9	86.2	80.9	82.7	88.6	89.2	97.4	97.6	75.9	79.3	43.9	44.6
GRAM$\uparrow^{VT}$	86.9	88.0	80.9	82.9	87.5	88.5	98.4	98.0	76.6	81.4	45.5	43.3
Triangle$\uparrow^{VT}$	84.3	86.8	80.2	80.5	87.9	88.1	97.5	97.5	73.6	78.7	47.0	45.5
PMRL$\uparrow^{VT}$	85.6	87.1	81.4	82.5	88.9	88.8	97.8	97.4	77.1	79.8	49.4	46.8
CalMRL $\uparrow^{VT}$	86.3	87.6	81.9	83.2	89.3	88.9	98.4	98.6	76.6	79.1	49.1	45.8

B.5 Hyperparameter Setting

We adopt the AdamW optimizer for training the parameters with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate is set to 1×10^{-5} and we apply the linear schedule with a warm-up ratio being 0.1. We set the temperature parameters $\tau = 0.05$ and $\tau' = 0.1$ and instance matching weight α to 0.1. All the unimodal representations are transformed with the number of dimensions as 512. The training batch size is set to 64. All the experiments are conducted on the device equipped with 2×NVIDIA H100-80GB GPUs.

C Additional Results

In Tables 4 and 5, we report the comprehensive results of multimodal retrieval using the metrics Recall@5 and Recall@10, respectively. A higher top K suggests higher absolute performance and a decreased performance gap. Overall, CalMRL achieves outperforming or comparable results compared to the state-of-the-art methods.

D Reproducibility

We provide implementation details, including illustrative algorithm descriptions and flows (*cf.*, Algorithm 1). The source code will be publicly released for reproducibility.

E Limitations

CalMRL enhances multimodal learning by flexibly handling datasets with missing modalities. Despite its effectiveness and design rationale, this core idea is model-architecture agnostic. Therefore, incorporating a more powerful backbone (such as recent advanced vision-text models) could significantly improve its capabilities. However, adopting a new backbone necessitates pre-training to adapt these models, which is computationally and data-collection intensive. To this end, we chose a recognized approach, the emerging complete-modality alignment frameworks (*e.g.*, GRAM and PMRL), to validate the effect of CalMRL.

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] 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.

[4] Qinglong Cao, Yuntian Chen, Lu Lu, Hao Sun, Zhengzhong Zeng, Xiaokang Yang, and Dongxiao Zhang. Generalized domain prompt learning for accessible scientific vision-language models. *Nexus*, 2(2), 2025.

[5] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. In *NeurIPS*, 2021.

[6] Paul Pu Liang, Yun Cheng, Xiang Fan, Chun Kai Ling, Suzanne Nie, Richard Chen, Zihao Deng, Nicholas Allen, Randy Auerbach, Faisal Mahmood, et al. Quantifying & modeling multimodal interactions: An information decomposition framework. In *NeurIPS*, pages 27351–27393, 2023.

[7] Yi Xin, Qi Qin, Siqi Luo, Kaiwen Zhu, Juncheng Yan, Yan Tai, Jiayi Lei, Yuewen Cao, Keqi Wang, Yibin Wang, et al. Lumina-dimoo: An omni diffusion large language model for multi-modal generation and understanding. *arXiv preprint arXiv:2510.06308*, 2025.

[8] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. Vit-lens: Towards omni-modal representations. In *CVPR*, pages 26647–26657, 2024.

[9] Run Luo, Lu Wang, Wanwei He, Longze Chen, Jiaming Li, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025.

[10] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *ICML*, 2024.

[11] Megan Tjandrasuwita, Chanakya Ekbote, Liu Ziyin, and Paul Pu Liang. Understanding the emergence of multimodal representation alignment. In *ICML*, 2025.

[12] 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.

[13] 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.

[14] 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. In *NeurIPS*, pages 72842–72866, 2023.

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

[16] Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Continual multimodal contrastive learning. In *NeurIPS*, 2025.

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

[18] Xiaohao Liu, Xiaobo Xia, See-Kiong Ng, and Tat-Seng Chua. Principled multimodal representation learning. *arXiv preprint arXiv:2507.17343*, 2025.

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

[20] 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.

[21] 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.

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

[23] 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.

[24] Giordano Cicchetti, Eleonora Grassucci, and Danilo Comminiello. A TRIANGLE enables multimodal alignment beyond cosine similarity. In *NeurIPS*, 2025.

[25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255, 2009.

[26] Yi Wang, Yinan He, Yizhuo Li, Kunchang Li, Jiashuo Yu, Xin Ma, Xinhao Li, Guo Chen, Xinyuan Chen, Yaohui Wang, Ping Luo, Ziwei Liu, Yali Wang, Limin Wang, and Yu Qiao. Internvid: A large-scale video-text dataset for multimodal understanding and generation. In *ICLR*. OpenReview.net, 2024.

[27] Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1m: A large-scale high-quality dataset for text-to-video generation. In *ICLR*. OpenReview.net, 2025.

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

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

[30] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, pages 1–5. IEEE, 2023.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

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

[36] Muhammad Abdullah Jamal and Omid Mohareri. Multi-modal contrastive masked autoencoders: A two-stage progressive pre-training approach for rgbd datasets. In *CVPR*, pages 17947–17957, 2025.

[37] Shawn Li, Huixian Gong, Hao Dong, Tiankai Yang, Zhengzhong Tu, and Yue Zhao. Dpu: Dynamic prototype updating for multimodal out-of-distribution detection. In *CVPR*, pages 10193–10202, June 2025.

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

[39] Mohammadreza Zolfaghari, Yi Zhu, Peter Gehler, and Thomas Brox. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *ICCV*, pages 1450–1459, 2021.

[40] 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.

[41] 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.

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

[43] 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.

[44] 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.

[45] 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.

[46] 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.

[47] Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations and structured captions. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang, editors, *NeurIPS*, 2024.

[48] Letian Fu, Gaurav Datta, Huang Huang, William Chung-Ho Panitch, Jaimyn Drake, Joseph Ortiz, Mustafa Mukadam, Mike Lambeta, Roberto Calandra, and Ken Goldberg. A touch, vision, and language dataset for multimodal alignment. In *ICML*. OpenReview.net, 2024.

[49] 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, 2022.

[50] 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.

[51] Siddharth Srivastava and Gaurav Sharma. Omnivec: Learning robust representations with cross modal sharing. In *WACV*, pages 1225–1237. IEEE, 2024.

[52] Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, and Mike Zheng Shou. VIT-LENS: towards omni-modal representations. In *CVPR*, pages 26637–26647. IEEE, 2024.

[53] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. Onellm: One framework to align all modalities with language. In *CVPR*, pages 26574–26585. IEEE, 2024.

[54] Renjie Wu, Hu Wang, Hsiang-Ting Chen, and Gustavo Carneiro. Deep multimodal learning with missing modality: A survey. *arXiv preprint arXiv:2409.07825*, 2024.

[55] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *CVPR*, pages 18177–18186, 2022.

[56] Lei Cai, Zhengyang Wang, Hongyang Gao, Dinggang Shen, and Shuiwang Ji. Deep adversarial learning for multi-modality missing data completion. In *KDD*, pages 1158–1166, 2018.

[57] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. In *NeurIPS*, pages 17117–17128, 2023.

[58] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In *ACL*, pages 2608–2618, 2021.

[59] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. Smil: Multimodal learning with severely missing modality. In *AAAI*, volume 35, pages 2302–2310, 2021.

[60] Chaohe Zhang, Xu Chu, Liantao Ma, Yinghao Zhu, Yasha Wang, Jiangtao Wang, and Junfeng Zhao. M3care: Learning with missing modalities in multimodal healthcare data. In *KDD*, pages 2418–2428, 2022.

[61] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *CVPR*, pages 15878–15887, 2023.

[62] Tao Jin, Xize Cheng, Linjun Li, Wang Lin, Ye Wang, and Zhou Zhao. Rethinking missing modality learning from a decoding perspective. In *MM*, pages 4431–4439, 2023.

[63] Zhenchao Tang, Guanxing Chen, Shouzhi Chen, Jianhua Yao, Linlin You, and Calvin Yu-Chian Chen. Modalnexus auto-encoder for multi-modality cellular data integration and imputation. *Nature Communications*, 15(1):9021, 2024.

[64] Mengxi Chen, Fei Zhang, Zihua Zhao, Jiangchao Yao, Ya Zhang, and Yanfeng Wang. Probabilistic conformal distillation for enhancing missing modality robustness. In *CVPR*, pages 36218–36242, 2024.

[65] Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. In *NeurIPS*, pages 98782–98805, 2024.

[66] Guanzhou Ke, Shengfeng He, Xiaoli Wang, Bo Wang, Guoqing Chao, Yuanyang Zhang, Yi Xie, and Hexing Su. Knowledge bridger: Towards training-free missing modality completion. In *CVPR*, pages 25864–25873, 2025.

[67] Mingrui Lao, Zheng Li, Yanming Guo, Xueyi Zhang, Siqi Cai, Zhaoyun Ding, and Haizhou Li. Boosting discriminability for robust multimodal entity linking with visual modality missing. In *SIGIR*, pages 989–999, 2025.

[68] Hu Wang, Congbo Ma, Jianpeng Zhang, Yuan Zhang, Jodie Avery, Louise Hull, and Gustavo Carneiro. Learnable cross-modal knowledge distillation for multi-modal learning with missing modality. In *MICCAI*, pages 216–226, 2023.

[69] Wenxin Xu, Hexin Jiang, and Xuefeng Liang. Leveraging knowledge of modality experts for incomplete multimodal learning. In *MM*, pages 438–446, 2024.

[70] Sijie Li, Chen Chen, and Jungong Han. Simmlm: A simple framework for multi-modal learning with missing modality. *arXiv preprint arXiv:2507.19264*, 2025.

[71] Md Kaykobad Reza, Ashley Prater-Bennette, and M Salman Asif. Robust multimodal learning with missing modalities via parameter-efficient adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[72] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 61(3):611–622, 1999.

[73] Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Factor analysis, probabilistic principal component analysis, variational inference, and variational autoencoder: Tutorial and survey. *arXiv preprint arXiv:2101.00734*, 2021.

[74] Sharut Gupta, Shobhita Sundaram, Chenyu Wang, Stefanie Jegelka, and Phillip Isola. Better together: Leveraging unpaired multimodal data for stronger unimodal models. *arXiv preprint arXiv:2510.08492*, 2025.

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

[76] 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.

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

[78] 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.

[79] 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.

[80] Ren-Cang Li. Matrix perturbation theory. *Handbook of linear algebra*, pages 15–21, 2006.

[81] Max A Woodbury. The stability of out-input matrices. *Chicago, IL*, 9:3–8, 1949.

[82] 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.

[83] 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.