To Align or Not to Align: Strategic Multimodal Representation Alignment for Optimal Performance

Wanlong Fang[1,2], Tianle Zhang[2], Alvin Chan[2,3,4]*

[1]Artificial Intelligence-X (AI-X) @ NTU, Interdisciplinary Graduate Programme, Nanyang Technological University
[2]College of Computing and Data Science, Nanyang Technological University
[3]Lee Kong Chian School of Medicine, Nanyang Technological University
[4]Centre of AI in Medicine (C-AIM), Nanyang Technological University
{wanlong001, tianle006}@e.ntu.edu.sg, guoweialvin.chan@ntu.edu.sg

Abstract

Multimodal learning often relies on aligning representations across modalities to enable effective information integration—an approach traditionally assumed to be universally beneficial. However, prior research has primarily taken an observational approach, examining naturally occurring alignment in multimodal data and exploring its correlation with model performance, without systematically studying the direct effects of explicitly enforced alignment between representations of different modalities. In this work, we investigate how explicit alignment influences both model performance and representation alignment under different modality-specific information structures. Specifically, we introduce a controllable contrastive learning module that enables precise manipulation of alignment strength during training, allowing us to explore when explicit alignment improves or hinders performance. Our results on synthetic and real datasets under different data characteristics show that the impact of explicit alignment on the performance of unimodal models is related to the characteristics of the data: the optimal level of alignment depends on the amount of redundancy between the different modalities. We can find an optimal alignment strength that balances modality-specific signals and shared redundancy in the mixed information distributions. This work can help practitioners on when and how to enforce alignment for optimal unimodal encoder performance.

Introduction

A foundational principle in multimodal machine learning is that explicitly enforcing alignment between modalities in a shared semantic space is essential for effective knowledge fusion and improved performance (Wang et al. 2020; Zadeh et al. 2017; Jin et al. 2025; Zong, Mac Aodha, and Hospedales 2024; Wang and Shi 2023; Wang, Jian, and Xue 2023). This assumption has driven the development of state-of-the-art architectures, which often use contrastive learning objectives to maximize representation similarity (Yuan et al. 2021; Pielawski et al. 2020).

However, recent work shows that representation alignment can also emerge naturally from data itself without a direct alignment objective. The Platonic Representation Hypothesis (Huh et al. 2024a) posits that as unimodal models increase in scale and capability, their internal representations naturally converge toward a shared statistical model of the underlying reality. This emergent alignment is not coincidental; it correlates with the model's performance. While the Platonic Representation Hypothesis suggests a universal benefit, critical empirical findings from (Tjandrasuwita et al. 2025) demonstrate that the utility of alignment is far from universal. Their work reveals that the relationship between alignment and performance is highly conditional on the data's intrinsic information structure.

Prior research has primarily taken an observational approach, examining natural alignment in multimodal data and exploring its correlation with model performance. However, no studies have previously systematically varied the strength of alignment enforcement to assess its effects on model behavior. Gaining a clearer understanding of this area could offer valuable insights into how alignment influences both performance and representational structure, which are central considerations in the design of multimodal systems.

In this paper, we take a step toward addressing this gap by exploring the relationship between explicit alignment strength and its effects on model performance and representational similarity. We examine this relationship across different underlying information structures in the data, with the goal of better understanding how alignment may function under varying conditions. Our main research questions are as follows:

- ***RQ1***: Under what conditions does explicit alignment improve or hinder unimodal performance?
- ***RQ2***: How does this improvement generalize to real-world multimodal tasks?

To systematically investigate these questions, we introduce a controllable contrastive alignment loss weighted by a scalar factor λ, which is added to the task loss during unimodal encoder training. This allows us to precisely adjust alignment strength and evaluate its effects on both performance and representational similarity between modalities. In the synthetic dataset setting, we explicitly manipulate the degree of redundancy, enabling controlled testing under known information distributions. These experiments showed a direct link between representation alignment and improvement of

*Corresponding Author.

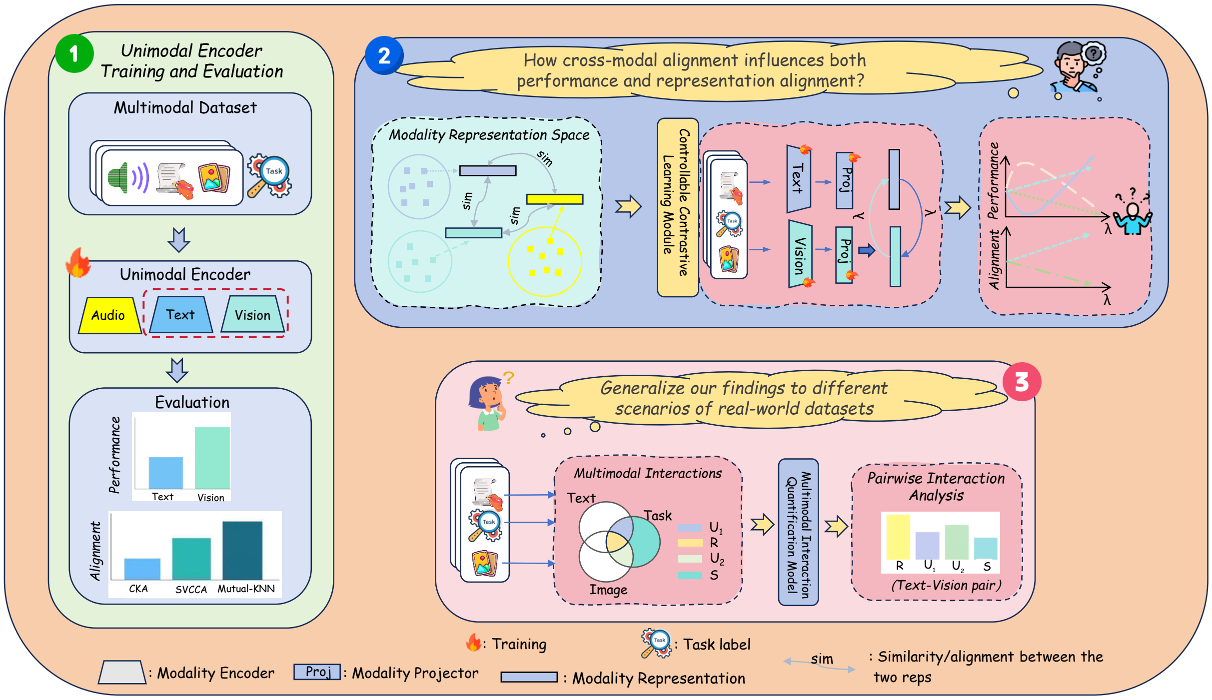

Figure 1: Overview of our experimental framework for analyzing the impact of explicit multimodal alignment. ① **Independent Training Baseline:** Each unimodal encoder is trained independently to establish a performance and representation baseline, capturing naturally emergent alignment. ② **Controlled Cross-modal Alignment:** We introduce a contrastive learning module to explicitly enforce alignment during training, enabling systematic analysis of how alignment strength affects unimodal performance and representational similarity. ③ **Real-world Generalization via PID:** We apply Partial Information Decomposition (PID) to quantify the redundancy–uniqueness–synergy structure across modality pairs, validating our findings under diverse real-world data conditions.

encoders when redundancy is present between the modalities. To validate these findings in practical applications, we first quantify the level of redundant, unique, and synergistic information across modality pairs in real-world datasets, using the Partial Information Decomposition (PID) framework (Williams and Beer 2010; Bertschinger et al. 2014). We then conduct experiments across all possible modality pairs within three diverse real-world benchmarks, offering a comprehensive analysis of alignment utility under varying data conditions. We also found that our approach to enforce alignment between modalities improve the performance of unimodal encoders, making it a practical tool for unimodal tasks.

Our findings provide compelling empirical evidence regarding the conditional utility of explicit alignment, making the following key contributions and impacts:

- ***A New Angle on Representation Alignment***: Unlike prior work, which often investigates implicit alignment, i.e. how representations naturally align within multimodal data. Specifically, we probe the direct link between forced representation alignment and its effects on both empirical performance and representation similarity.
- ***Findings on Improving Modality Encoders***: Our experimental results on synthetic datasets show that explicit cross-modal alignment can be a feasible approach to improve unimodal encoders. This improvement depends on the level of redundant information shared between the two modalities.
- ***Generalization to Real-world Datasets:*** By applying the PID framework, we validate our findings across real-world multimodal data under different modality-specific information structures, ensuring that insights from synthetic experiments are consistent and robust in practical, non-synthetic scenarios. This cross-validation strengthens the generalization of our findings and demonstrates the applicability of our analysis to real-world tasks.

Related Work

Multimodal Representation Alignment. Prior work has shown that alignment between modality-specific representations can emerge naturally without explicit supervision (Bonheme and Grzes 2022; Huh et al. 2024a), a phenomenon formalized by the *Platonic Representation Hypothesis* (Ziyin and Chuang 2025). Metrics such as CKA (Kornblith et al. 2019) and SVCCA (Raghu et al. 2017) have been used to study such emergent alignment, which is often correlated with performance gains. However, these studies are mostly observational and do not systematically intervene on alignment strength. Tjandrasuwita et al. (Tjandrasuwita et al. 2025) relate emergent alignment to information structure, but do not explicitly manipulate the alignment mechanism.

Cross-modal alignment via contrastive mearning Contrastive learning has become the standard approach for enforcing cross-modal alignment (Radford et al. 2021; Xu et al. 2021; Zolfaghari et al. 2021; Wang et al. 2021, 2025). While effective, most methods assume that stronger alignment is always beneficial (Cai et al. 2025; Jiang et al. 2023), potentially ignoring modality-specific structures such as redundancy and synergy (Dufumier et al. 2025). Our work departs from this assumption by introducing a controllable contrastive framework that explicitly varies alignment strength and measures its effect on unimodal performance under diverse information structures. *A full related work is provided in extended version.*

Background and Preliminaries

Multimodal Information Quantification To clarify the role of alignment in multimodal learning, we begin by analyzing how information is distributed across modalities. We consider two input modalities, X_1 and X_2, and a shared task label Y, where each sample (x_1, x_2, y) is drawn from a joint distribution $\mathcal{P}(X_1, X_2, Y)$. While our analysis focuses on two modalities for clarity, the methodology naturally extends to more modalities.

The total information that (X_1, X_2) provides about Y is captured by the multivariate mutual information $I(X_1, X_2; Y)$. However, conventional mutual information does not distinguish between different types of interactions among modalities. To address this, we adopt the Partial Information Decomposition (PID) framework (Williams and Beer 2010; Bertschinger et al. 2014), which decomposes $I(X_1, X_2; Y)$ into four interpretable components:

$$I(X_1, X_2; Y) = R + U_1 + U_2 + S, \tag{1}$$

where:

- R is the **redundancy** — information about Y that is shared by both X_1 and X_2;
- U_1, U_2 are the **unique information** contributed individually by X_1 and X_2, respectively;
- S is the **synergy** — information that only emerges when both modalities are considered jointly.

These components satisfy the following consistency constraints based on standard mutual information identities:

$$I(X_1; Y) = R + U_1, \quad I(X_2; Y) = R + U_2. \tag{2}$$

This decomposition provides a way to disentangle how each modality contributes to the task. Recent work (Liang et al. 2023; Yang, Wang, and Hu 2025) has proposed practical estimators for PID components in real-world datasets, which offers the foundation of our empirical analysis.

Method

The PID framework reveals the intrinsic, static information structure of a dataset—quantifying the redundant, unique, and synergistic components across modality-pairs. This decomposition enables us to generalize our findings across modality pairs with different information structures and provides actionable guidance on when and how strong alignment should be introduced. To operationalize this analysis, we introduce a controllable contrastive learning module that enables systematic manipulation of alignment strength. This allows us to directly examine the effect of enforced cross-modal alignment on both unimodal task performance and representation similarity.

Experimental Framework: A Systematic Approach to Probing Alignment Effects when multiple modalities (e.g., X_1, X_2) are used to predict the same target label Y, the representations learned by independently trained unimodal encoders (f_1, f_2) should inherently capture some shared, task-relevant information. For instance, a speaker's facial expressions, vocal tone, and text all convey cues about the same underlying sentiment, thus sharing redundant content while retaining unique signals. While prior work has relied on emergent or maximum alignment, few studies have systematically treated alignment strength as a controllable variable to investigate its impact on unimodal encoder learning. We propose the following procedure to address this gap (illustrated in Figure 1):

1. Unimodal encoders are trained independently to establish a baseline.
2. Experiment on how crossmodal alignment influence both performance and representation alignment.
3. Generalization to real-world dataset under different data characteristics scenarios.

Controllable Contrastive Learning Module To implement this controllable alignment in practice, we designed a module based on contrastive learning. We consider a batch of N paired samples $\left\{(x_i^A, x_i^B)\right\}_{i=1}^N$, where x_i^A and x_i^B are inputs from modality A (e.g., visual) and modality B (e.g., textual), respectively. Each input is processed by a corresponding unimodal encoder, $f_A(\cdot)$ or $f_B(\cdot)$, to extract modality-specific features. Following standard practice, these features are projected into a shared, normalized latent space—if dimensional alignment is needed—using separate multi-layer perceptron heads, $g_A(\cdot)$ and $g_B(\cdot)$. The resulting representations are $\mathbf{z}_i^A = g_A(f_A(x_i^A))$ and $\mathbf{z}_i^B = g_B(f_B(x_i^B))$.

Our symmetric contrastive loss, $\mathcal{L}_{\text{align}}$, is defined as the average of two InfoNCE (Oord, Li, and Vinyals 2018) losses, promoting bidirectional alignment. The loss from modality A to modality B, denoted as $\mathcal{L}_{A\to B}$, is formulated as:

$$\mathcal{L}_{A\to B} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\text{sim}(\mathbf{z}_i^A, \mathbf{z}_i^B)/\tau)}{\sum_{j=1}^{N} \exp(\text{sim}(\mathbf{z}_i^A, \mathbf{z}_j^B)/\tau)}. \tag{3}$$

where $\text{sim}(\mathbf{u}, \mathbf{v}) = \mathbf{u}^\top \mathbf{v}$ denotes the cosine similarity between two L2-normalized embedding vectors (i.e., their dot product), and τ is a fixed temperature hyperparameter that controls the sharpness of the softmax distribution. This loss encourages alignment between each positive pair $(\mathbf{z}_i^A, \mathbf{z}_i^B)$ while contrasting it against the $N - 1$ negative samples within the batch. The complete symmetric alignment loss, $\mathcal{L}_{\text{align}}$, is the average of the two directional losses:

$$\mathcal{L}_{\text{align}} = \frac{1}{2}(\mathcal{L}_{A\to B} + \mathcal{L}_{B\to A}) \quad (4)$$

Finally, this alignment loss is integrated as a regularization term into our model's total training objective:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda \cdot \mathcal{L}_{\text{align}} \quad (5)$$

In this formulation, $\mathcal{L}_{\text{task}}$ denotes the primary loss for the downstream task (e.g., cross-entropy for classification or L1 loss for regression), while λ is a scalar hyperparameter controlling the strength of the alignment regularization. Varying λ allows us to systematically explore how explicit alignment influences unimodal encoder performance under different information structures. When $\lambda = 0$, the model reduces to independently trained unimodal encoders, serving as the experimental baseline. As λ increases, the encoders are increasingly regularized to produce similar representations in the latent space. This formulation enables us to trace unimodal performance across the full spectrum—from no enforced alignment to strong alignment—and positions λ as the central experimental lever for addressing our primary research question (RQ1): *Under what conditions does explicit alignment improve or hinder unimodal performance?*

Experimental Setup

In this section, we describe the experimental setup for both synthetic and real-world datasets. (*Detailed dataset descriptions and training hyperparameters for encoders and modality-specific projectors can be found in the extended version.*)

Experimental Setup on Synthetic Dataset

To efficiently investigate our research question, we adopt a controlled experimental setup using synthetic datasets. This approach enables us to ***systematically manipulate*** the intrinsic properties of multimodal data, particularly the proportions of redundant and unique information.

Synthetic Data Generation We adopt the synthetic dataset proposed by (Tjandrasuwita et al. 2025), which is specifically designed to construct multimodal inputs (X_1, X_2) and an associated label Y with precisely controlled information composition. Following the PID framework, the task-relevant information for Y is decomposed into three conceptual components: a redundant component x_r shared across modalities, and two modality-specific unique components x_{u1} and x_{u2}. Each modality input is then formed by concatenating the shared and unique components:

$$x_1 = [x_r, x_{u1}], \quad x_2 = [x_r, x_{u2}].$$

An illustration of this generation process is provided in Figure 2. The degrees of redundancy (R) and uniqueness (U_i) are governed by the construction of the label Y, which is defined as a non-linear function over a selected subset $Q \subseteq \{x_r, x_{u1}, x_{u2}\}$. Specifically, R corresponds to the

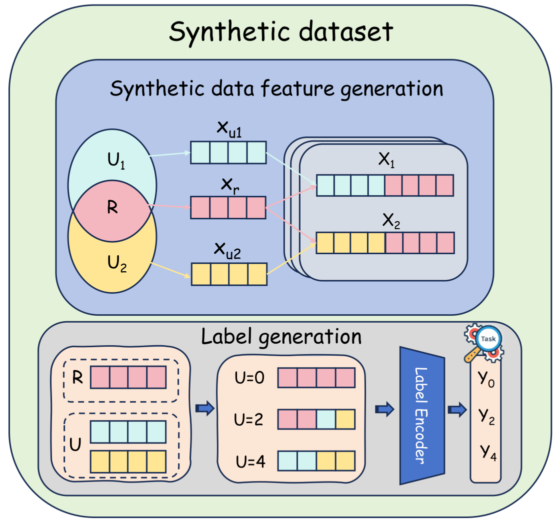

Figure 2: Synthetic Data Generation Process.

number of features in Q drawn from x_r, and U_i corresponds to the number of features drawn from x_{ui}. The total uniqueness is defined as $U = |Q| - R$. By fixing the total number of task-relevant features $|Q|$ and varying the allocation between redundant and unique components, we generate datasets that span a continuum of information structures—from fully redundant to fully modality-specific. This controlled design enables systematic investigation of how the alignment-performance relationship varies under different redundancy–uniqueness trade-offs.

Model Architecture and Training Protocol In our synthetic experiments, we use MLPs as unimodal encoders f_A and f_B, trained jointly with the controllable contrastive learning module. To ensure consistency and maintain a lightweight setup, each MLP has a fixed hidden dimension of 12—comprising 8 dimensions for shared features and 4 for unique features.

We evaluate model performance and compute alignment between the unimodal encoders f_A and f_B, each trained on X_1 and X_2, respectively, under varying levels of unique information and different alignment strengths $\lambda \in \{0, 0.2, 0.4, 0.6, 0.8, 1, 2\}$.

Experimental Setup on Real-world Datasets

To generalize our findings from synthetic dataset, we conduct analogous experiments on a selection of real-world multimodal datasets.

Dataset Selection and Characterization We select three representative datasets from **MultiBench** (Liang et al. 2021), each exhibiting distinct information characteristics quantified using the PID framework (***Table*** 1): **CMU-MOSEI** (Zadeh et al. 2018), **AV-MNIST** (Pérez-Rúa et al. 2019) and **MUSTARD** (Castro et al. 2019). These datasets cover a spectrum of redundant, unique, and synergistic information compositions, serving as a diverse testbed for evaluating the impact of alignment.

Model Architecture and Training Protocol For the affective computing datasets (CMU-MOSEI and MUStARD), we use transformer encoders trained on pre-extracted video, audio, and text features. For AV-MNIST, we adopt a vision transformer for digit images and a separate transformer for audio inputs, following the same architecture. Our training protocol mirrors that of the synthetic experiments. For datasets with more than two modalities (e.g., CMU-MOSEI, MUStARD), we conduct experiments on all pairwise combinations (e.g., Vision–Audio, Vision–Text, Audio–Text). For each modality pair, we learn dedicated projection heads and apply our controllable contrastive alignment module to regularize cross-modal representations. To study the effect of alignment strength, we vary the alignment weight λ over the range $\{0, 0.1, 0.25, 0.5, 0.6, 0.75, 1, 2, 4\}$ and evaluate unimodal task performance and representational alignment across modalities.

Results and Analysis

We introduce our evaluation protocols and alignment metrics, followed by results and analysis on synthetic and real-world datasets. For real-world data, we focus on representative modality pairs to illustrate alignment effects. *Due to space constraints, results for additional combinations are provided in the extended version.*

Evaluation Metrics & Alignment Computing We evaluate task performance using classification accuracy across all experiments, including both synthetic and real-world datasets. For MOSEI, we binarize the original regression labels into positive and negative sentiment classes to align with the classification setting. To quantify cross-modal alignment between unimodal encoders, we adopt three established metrics: **(i)** Centered Kernel Alignment (**CKA**) (Kornblith et al. 2019), **(ii)** Singular Vector Canonical Correlation Analysis (**SVCCA**) (Raghu et al. 2017), and **(iii)** Mutual K-Nearest Neighbors (**Mutual-KNN**) (Huh et al. 2024a). These metrics assess representational similarity across modalities, enabling us to investigate whether stronger explicit alignment leads to greater inter-modal consistency and improved downstream performance.

Performance and Alignment in Redundancy-Dominant Scenarios We first examine redundant information dominant scenarios, where redundancy is high and both unique and synergy are relatively low, as quantified by PID analysis (Table 1) and synthetic setup.

1. **Synthetic Dataset Results.** For high-redundancy synthetic datasets ($R = 6, 8$), we observe that unimodal encoder performance **monotonically improves and saturates** as alignment strength λ increases (Figure 3d, e). Alignment metrics also show a consistent upward trend (Figure 3i, j), confirming the effectiveness of the explicit alignment objective in this setting.
2. **Real-World Modalities with Minimal Unique Information.** Similar trends are observed in real-world scenarios where a modality contributes minimal unique information. In CMU-MOSEI (Vision in the Vision-Text pair), the Vision encoder exhibits negligible unique information ($U_1 = 0.001$) and moderate redundancy ($R = 0.123$). Its performance generally improves with increasing λ, peaking before a mild saturation (Figure 5e, pink bars), while alignment metrics steadily rise (Figure 5f). Likewise, in AV-MNIST (Audio in the Vision-Audio pair), the Audio encoder shows very low unique information ($U_2 = 0.03$). Its classification accuracy slightly increases as λ grows, reaching a peak (Figure 5a, blue bars), with alignment scores following a similar upward trend (Figure 5b).
3. **Analysis and Insights.** The consistent performance gains in both synthetic and real-world redundancy-dominant settings demonstrate that **explicit alignment is particularly effective when redundant information is existed and higher**. These findings empirically support the Platonic Convergence Hypothesis (Huh et al. 2024b), showing that maximizing inter-modal alignment is beneficial when an underlying shared structure exists. Notably, even within complex real-world datasets, individual modalities with minimal unique content can still benefit from alignment—provided redundancy exists. This directly answers our research question ***RQ2***, confirming that alignment benefits observed in controlled settings can transfer to practical, real-world scenarios. This highlights that alignment should not be applied solely based on coarse dataset-level labels, but rather guided by modality-level information decomposition. Strategic application of alignment, informed by fine-grained PID analysis, can thus maximize task performance by leveraging redundancy where present.

Dataset	Modality	R	U_1	U_2	S
CMU-MOSEI	Vision-Text	**0.123**	0.001	**0.163**	0.005
	Vision-Audio	**0.116**	0.010	0.001	0.012
	Audio-Text	**0.127**	0.001	**0.248**	0.002
AV-MNIST	Vision-Audio	0.170	**0.970**	0.030	0.040
MUSTARD	Vision-Text	**0.150**	0.020	0.010	**0.340**
	Vision-Audio	**0.140**	0.020	0.010	**0.200**
	Audio-Text	**0.160**	0.010	0.010	**0.370**

Table 1: PID statistics for selected MultiBench datasets. Bold indicates the dominant information type within each modality pair.

Performance and Alignment in Uniqueness-Dominant Scenarios We next examine scenarios where task-relevant information is primarily unique to individual modalities. This setting corresponds to low redundancy (R) and high uniqueness (U) as quantified by PID.

1. **Synthetic Dataset Results.** For low-redundancy synthetic datasets ($R = 0, 2$), we observe a clear performance decline as λ increases (Figure 5c, d). Both unimodal encoders show a **monotonic decrease** in accuracy, indicating that explicit alignment hinders their ability to preserve distinct, modality-specific information. Notably, all alignment metrics still increase with λ (Figure 3f, g), confirming that the alignment objective is tech-

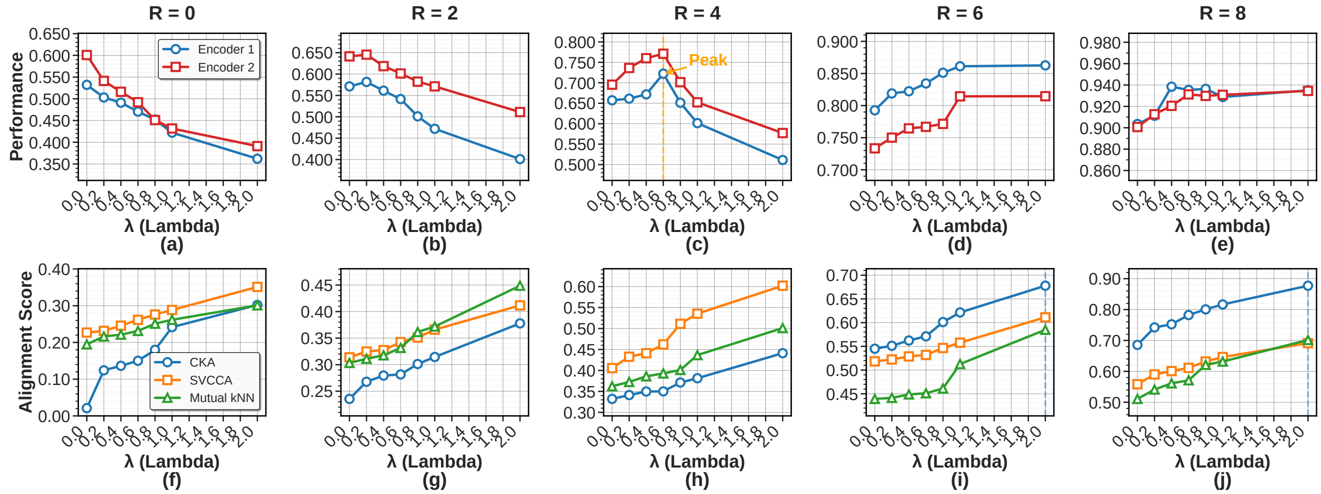

Figure 3: Experimental results on the synthetic dataset across varying redundancy levels (***R***). (a)–(e): Classification performance of Encoder 1 (blue) and Encoder 2 (red) as the alignment strength λ increases, under different redundancy levels ($R = 0, 2, 4, 6, 8$). (f)–(j): Cross-modal representation alignment scores (CKA, SVCCA, and Mutual-KNN) between the two encoders as a function of λ, evaluated at the same redundancy levels.

nically effective—even as it leads to degraded task performance.

2. **Real-World Case.** A similar pattern emerges in the AV-MNIST dataset, where the vision encoder in the vision-audio pair is characterized by extremely high unique information ($U_1 = 0.97$). Starting from a strong baseline, its performance slightly declines as λ increases, reaching a trough before a mild rebound at higher alignment strengths (Figure 5a, pink bars). Meanwhile, alignment metrics continue to rise (Figure 5b), further emphasizing the disconnect between alignment and performance in uniqueness-dominant conditions.

3. **Analysis and Insights.** These results provide a clear answer to ***RQ1***, when task-relevant information is predominantly modality-specific, explicit alignment can be detrimental. By enforcing representational similarity, alignment suppresses the very distinctions that enable strong unimodal performance. In such cases, improvements in alignment scores do not translate into better task outcomes—and can even be inversely correlated. The behavior of the AV-MNIST Vision encoder offers real-world validation of this effect. Despite the dataset's simplicity, the extreme dominance of unique visual information makes its performance particularly sensitive to forced alignment. These findings underscore the importance of tailoring alignment strategies to the underlying information structure of the data. In uniqueness-dominant settings, restraint in applying alignment is essential to preserve performance-critical modality-specific signals.

Performance and Alignment in Synergy-Dominant Scenarios We now investigate the impact of explicit alignment in scenarios dominated by synergistic information. Unlike redundancy or uniqueness, synergy reflects information that emerges only through the joint interpretation of modalities. This analysis focuses on the **MUSTARD dataset** (Vision-Audio pair), which our PID results (Table 1) identify as **synergy-dominant** ($S = 0.20$), with relatively low redundancy ($R = 0.14$) and negligible unique information ($U_1 = 0.02, U_2 = 0.01$).

1. **Observed Results.** As shown in Figure 5, increasing the alignment strength λ leads to a general improvement in unimodal encoder performance. Specifically, the Vision encoder accuracy rises from 0.54 to 0.62, and the Audio encoder from 0.58 to 0.66 (Figure 5c). In parallel, all alignment metrics increase monotonically with λ (Figure 5d), confirming that the explicit alignment objective is successfully promoting representational similarity.

2. **Analysis and Insights.** These improvements offer a nuanced insight into alignment behavior in synergy-driven tasks. Although synergy, by definition, reflects information unavailable to any single modality, the observed performance gains suggest that explicit alignment can still be beneficial. We attribute this to the model's ability to leverage a **small but non-negligible amount of redundancy** ($R = 0.14$) present between the Vision and Audio modalities. By aligning representations around this shared signal, the model improves the robustness of its unimodal encoders—despite synergy being the dominant factor. However, it is important to note that the performance ceiling remains relatively low. This is expected, as sarcasm detection in MUSTARD depends heavily on **synergistic information**—content that emerges only through the interaction or fusion of both modalities and cannot be captured through alignment alone or only one encoder. While explicit alignment can enhance whatever redundant signals exist, it cannot uncover the deeper cross-modal dependencies that characterize synergy. Therefore, the utility of alignment in such settings is fundamentally constrained. These findings clarify that while modest gains are possible in synergy-

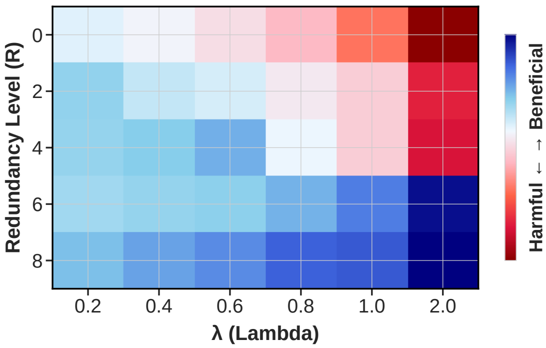

Figure 4: Effect of alignment strength λ on synthetic-dataset encoder1 performance under varying redundancy.

dominant tasks, alignment remains insufficient for fully modeling tasks where the most critical signals reside beyond unimodal spaces.

The Continuum of Alignment Utility: Towards Optimal Strategies We conclude by examining scenarios with mixed information distributions, where redundancy, uniqueness, and synergy coexist. These settings reveal that the utility of explicit alignment follows a **continuum**, often exhibiting an optimal alignment strength λ^*—a point beyond which further alignment becomes detrimental.

1. **Synthetic Dataset Results.** In synthetic datasets with moderate redundancy ($R = 4$), unimodal encoder performance follows a characteristic **concave (inverted U-shaped)** trend (Figure 3c). Performance initially improves with increasing λ, peaks around $\lambda = 0.4$–0.6, and then declines as over-alignment suppresses unique information. This trade-off is visually confirmed in Figure 4, where the transition from beneficial (blue) to harmful (red) impact is apparent at higher λ values. Meanwhile, alignment metrics continue to increase monotonically (Figure 3h), reinforcing the idea that stronger alignment does not necessarily imply better task performance.
2. **Real-World Case: CMU-MOSEI Text Encoder.** The Text encoder from the Vision-Text pair in CMU-MOSEI offers a real-world illustration of this trend. With moderate redundancy ($R = 0.123$) and substantial unique information ($U_2 = 0.163$), it exhibits a similar inverted U-shape: performance increases with λ, peaks around $\lambda = 0.75$, and slightly drops thereafter (Figure 5e, blue bars). Alignment metrics again increase steadily (Figure 5f), highlighting the divergence between alignment strength and actual task utility.
3. **Analysis and Insights.** These findings confirm that the effectiveness of explicit alignment is not binary, but depends on discovering an **optimal balance** between leveraging shared signals and preserving modality-specific information. While moderate alignment improves robustness by extracting redundancy, excessive alignment degrades performance by erasing valuable unique content. This validates our ***RQ1*** on the existence of an optimal λ^*, where alignment is most beneficial. The behavior of the CMU-MOSEI Text encoder exemplifies this balance: even as alignment improves, performance drops beyond λ^*, underscoring the importance of preserving distinct task-relevant signals. Overall, these results provide a principled guideline—**the optimal alignment strength should be tailored to the underlying information structure** of the multimodal data, offering practical insight for alignment-aware model design.

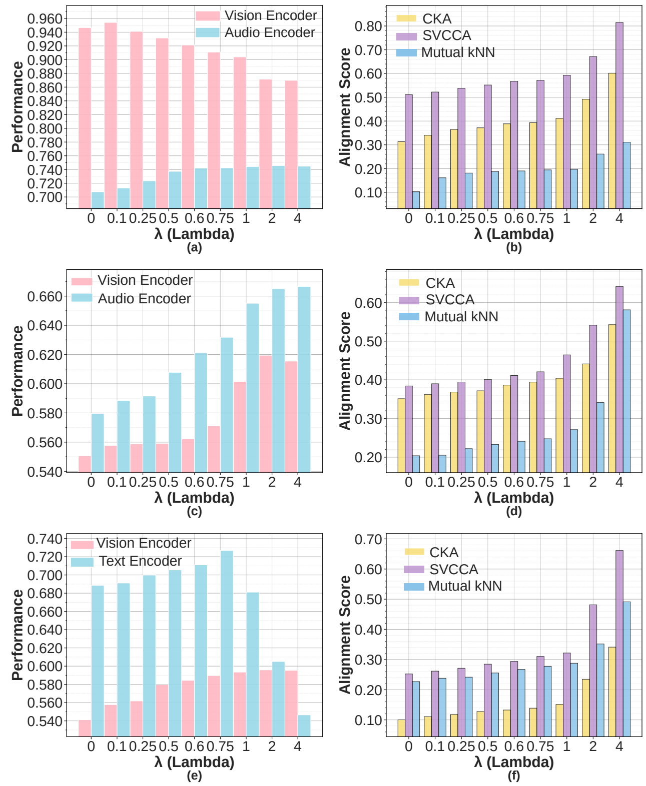

Figure 5: Performance and representation alignment as a function of alignment strength λ across different real-world datasets. (a)–(b): Results on AV-MNIST. (c)–(d): Results on MUStARD (V–A pair). (e)–(f): Results on CMU-MOSEI (V–T pair).

Conclusion

In this paper, we systematically examine when explicit cross-modal alignment helps or harms multimodal learning. We investigated the link between explicitly enforced alignment and its impact on both model performance and representation alignment under the different underlying information structure of data. To achieve this, we propose a controlled experimental framework that integrates a tunable alignment loss with PID to analyze the relationship between alignment strength and information structure. Our findings show that alignment is highly beneficial when modalities share redundant task-relevant information, but can be detrimental in uniqueness-dominant settings. We also find an optimal alignment strength in mixed-information regimes, suggesting that indiscriminate alignment may harm performance.

Acknowledgments

This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-001), NTU Start Up Grant and by the Ministry of Education, Singapore, under its Academic Research Fund Tier 1 (RG22/24). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. The computational work for this article was partially performed on resources of the National Supercomputing Centre (NSCC), Singapore (https://www.nscc.sg).

References

Bertschinger, N.; Rauh, J.; Olbrich, E.; Jost, J.; and Ay, N. 2014. Quantifying unique information. *Entropy*, 16(4): 2161–2183.

Bonheme, L.; and Grzes, M. 2022. How do Variational Autoencoders Learn? Insights from Representational Similarity. arXiv:2205.08399.

Cai, Y.; Liu, Y.; Gao, E.; Jiang, T.; Zhang, Z.; van den Hengel, A.; and Shi, J. Q. 2025. On the Value of Cross-Modal Misalignment in Multimodal Representation Learning. arXiv:2504.10143.

Castro, S.; Hazarika, D.; Pérez-Rosas, V.; Zimmermann, R.; Mihalcea, R.; and Poria, S. 2019. Towards multimodal sarcasm detection (an _obviously_ perfect paper). *arXiv preprint arXiv:1906.01815*.

Dufumier, B.; Castillo-Navarro, J.; Tuia, D.; and Thiran, J.-P. 2025. What to align in multimodal contrastive learning? arXiv:2409.07402.

Huh, M.; Cheung, B.; Wang, T.; and Isola, P. 2024a. The Platonic Representation Hypothesis. arXiv:2405.07987.

Huh, M.; Cheung, B.; Wang, T.; and Isola, P. 2024b. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*.

Jiang, C.; Ye, W.; Xu, H.; yan, M.; Zhang, S.; Zhang, J.; and Huang, F. 2023. Vision Language Pre-training by Contrastive Learning with Cross-Modal Similarity Regulation. arXiv:2305.04474.

Jin, Q.; Ge, E.; Xie, Y.; Luo, H.; Song, J.; Bi, Z.; Liang, C. X.; Guan, J.; Yeong, J.; and Hao, J. 2025. Multi-modal Representation Learning and Fusion. *arXiv preprint arXiv:2506.20494*.

Kornblith, S.; Norouzi, M.; Lee, H.; and Hinton, G. 2019. Similarity of Neural Network Representations Revisited. arXiv:1905.00414.

Liang, P. P.; Cheng, Y.; Fan, X.; Ling, C. K.; Nie, S.; Chen, R.; Deng, Z.; Allen, N.; Auerbach, R.; Mahmood, F.; et al. 2023. Quantifying & modeling multimodal interactions: An information decomposition framework. *Advances in Neural Information Processing Systems*, 36: 27351–27393.

Liang, P. P.; Lyu, Y.; Fan, X.; Wu, Z.; Cheng, Y.; Wu, J.; Chen, L.; Wu, P.; Lee, M. A.; Zhu, Y.; Salakhutdinov, R.; and Morency, L.-P. 2021. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. arXiv:2107.07502.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Pérez-Rúa, J.-M.; Vielzeuf, V.; Pateux, S.; Baccouche, M.; and Jurie, F. 2019. Mfas: Multimodal fusion architecture search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6966–6975.

Pielawski, N.; Wetzer, E.; Öfverstedt, J.; Lu, J.; Wählby, C.; Lindblad, J.; and Sladoje, N. 2020. CoMIR: Contrastive multimodal image representation for registration. *Advances in neural information processing systems*, 33: 18433–18444.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Raghu, M.; Gilmer, J.; Yosinski, J.; and Sohl-Dickstein, J. 2017. SVCCA: Singular Vector Canonical Correlation Analysis for Deep Learning Dynamics and Interpretability. arXiv:1706.05806.

Tjandrasuwita, M.; Ekbote, C.; Ziyin, L.; and Liang, P. P. 2025. Understanding the Emergence of Multimodal Representation Alignment. arXiv:2502.16282.

Wang, Y.; Huang, W.; Sun, F.; Xu, T.; Rong, Y.; and Huang, J. 2020. Deep multimodal fusion by channel exchanging. *Advances in neural information processing systems*, 33: 4835–4845.

Wang, Y.; Jian, X.; and Xue, B. 2023. Balance act: Mitigating hubness in cross-modal retrieval with query and gallery banks. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, 10542–10567.

Wang, Y.; and Shi, P. 2023. Video-text retrieval by supervised sparse multi-grained learning. In *Findings of the association for computational linguistics: EMNLP 2023*, 633–649.

Wang, Y.; Xue, B.; Cheng, Q.; Chen, Y.; and Zhang, L. 2021. Deep Unified Cross-Modality Hashing by Pairwise Data Alignment. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, 1129–1135.

Wang, Y.; Yuan, S.; Xue, B.; Jian, X.; Pang, W.; Wang, M.; and Yu, N. 2025. DREAM: Improving video-text retrieval through relevance-based augmentation using large foundation models. In *Proceedings of the 2025 conference of the nations of the americas chapter of the association for computational linguistics: Human language technologies (volume 1: Long papers)*, 3037–3056.

Williams, P. L.; and Beer, R. D. 2010. Nonnegative decomposition of multivariate information. *arXiv preprint arXiv:1004.2515*.

Xu, H.; Ghosh, G.; Huang, P.-Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; and Feichtenhofer, C. 2021. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*.

Yang, Z.; Wang, H.; and Hu, D. 2025. Efficient Quantification of Multimodal Interaction at Sample Level. *arXiv preprint arXiv:2506.17248*.

Yuan, X.; Lin, Z.; Kuen, J.; Zhang, J.; Wang, Y.; Maire, M.; Kale, A.; and Faieta, B. 2021. Multimodal contrastive training for visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6995–7004.

Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; and Morency, L.-P. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In Palmer, M.; Hwa, R.; and Riedel, S., eds., *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, 1103–1114. Copenhagen, Denmark: Association for Computational Linguistics.

Zadeh, A. B.; Liang, P. P.; Poria, S.; Cambria, E.; and Morency, L.-P. 2018. Multimodal language analysis in the wild: Cmu-mosei datasetp and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2236–2246.

Ziyin, L.; and Chuang, I. 2025. Proof of a perfect platonic representation hypothesis. arXiv:2507.01098.

Zolfaghari, M.; Zhu, Y.; Gehler, P.; and Brox, T. 2021. Crossclr: Cross-modal contrastive learning for multi-modal video representations. In *Proceedings of the IEEE/CVF international conference on computer vision*, 1450–1459.

Zong, Y.; Mac Aodha, O.; and Hospedales, T. 2024. Self-supervised multimodal learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.