AlignMamba: Enhancing Multimodal Mamba with Local and Global Cross-modal Alignment

Yan Li¹, Yifei Xing¹, Xiangyuan Lan¹, Xin Li¹, Haifeng Chen², Dongmei Jiang^{1*} ¹Pengcheng Laboratory, Shenzhen, China

²Shaanxi University of Science and Technology, Xi'an, China

liyan4ai@gmail.com, {xingyf, lanxy, lix07, jiangdm}@pcl.ac.cn, chenhaifeng@sust.edu.cn

Abstract

Cross-modal alignment is crucial for multimodal representation fusion due to the inherent heterogeneity between modalities. While Transformer-based methods have shown promising results in modeling inter-modal relationships, their quadratic computational complexity limits their applicability to long-sequence or large-scale data. Although recent Mamba-based approaches achieve linear complexity, their sequential scanning mechanism poses fundamental challenges in comprehensively modeling cross-modal relationships. To address this limitation, we propose Align-Mamba, an efficient and effective method for multimodal fusion. Specifically, grounded in Optimal Transport, we introduce a local cross-modal alignment module that explicitly learns token-level correspondences between different modalities. Moreover, we propose a global cross-modal alignment loss based on Maximum Mean Discrepancy to implicitly enforce the consistency between different modal distributions. Finally, the unimodal representations after local and global alignment are passed to the Mamba backbone for further cross-modal interaction and multimodal fusion. Extensive experiments on complete and incomplete multimodal fusion tasks demonstrate the effectiveness and efficiency of the proposed method. For instance, on the CMU-MOSI dataset, AlignMamba improves classification accuracy by 0.9%, reduces GPU memory usage by 20.3%, and decreases inference time by 83.3%.

1. Introduction

In recent years, multimodal representation fusion has emerged as a critical technology for integrating and understanding information across different modalities (e.g., audio, video, language). This capability is fundamental to a wide range of applications such as visual-language understanding [41] and audio-visual analysis [13, 40]. However, due to the inherent heterogeneity between modalities - each with its distinct statistical properties and feature distributions - achieving effective cross-modal alignment and fusion remains a significant challenge.

Traditional approaches to this challenge have primarily relied on Transformer-based [31] architectures, which can be broadly categorized into two paradigms. Single-stream methods (e.g., VisualBERT [15], ViLT [11], LLaVA [21]) concatenate features from different modalities into a unified sequence and process them through a shared Transformer layer. In contrast, multi-stream approaches (e.g., LXMERT [29], ViLBERT [23], MulT [30], CMA [44]) employ separate encoders for each modality with cross-modal Transformers to facilitate information exchange. While these methods have demonstrated promising results in capturing dynamic cross-modal interactions, they suffer from a fundamental limitation: the quadratic computational complexity of attention mechanisms makes them inefficient for processing long-sequence or large-scale data common in real-world multimodal applications.

Recent advances in sequence modeling have introduced the Mamba [3] architecture, based on State Space Models (SSMs) [4, 5], which achieves linear computational complexity while maintaining strong modeling capabilities. By incorporating selection mechanisms and hardware-aware parallel algorithms into SSMs, Mamba effectively captures long-range dependencies without the computational burden of attention mechanisms. This breakthrough has sparked considerable interest in adapting Mamba for multimodal fusion tasks, with approaches ranging from direct feature concatenation (e.g., VL-Mamba [26], Cobra [42], RoboMamba [22]) to multi-stream architectures (e.g., Pan-Mamba [8], Fusion-Mamba [2], MambaDFuse [17]). However, our analysis reveals a critical limitation. As shown in Fig. 1, Mamba's sequential scanning mechanism, while computationally efficient, struggles to capture comprehensive cross-modal relationships, particularly with unscanned tokens. This inherent limitation leads to suboptimal alignment between modalities and consequently affects the quality of learned multimodal fusion representations.

1



Figure 1. Transformer leverages attention mechanisms to model relationships across different modalities (top left), whereas Mamba struggles to achieve this due to its sequential scanning mechanism (top right). In contrast, the proposed AlignMamba utilizes both local (OT-based) and global (MMD-based) cross-modal alignment information to achieve efficient and effective multimodal fusion (bottom).

To address these issues, we propose AlignMamba, which integrates local and global cross-modal alignment information into Mamba for efficient and effective multimodal fusion. Specifically, we introduce a local alignment module based on Optimal Transport (OT), which learns a transport plan to align features across different modalities by minimizing the cost of feature transportation. While local alignment captures token-level cross-modal relationships, it does not account for distributional differences between modalities. Therefore, we also propose a global alignment loss based on Maximum Mean Discrepancy (MMD). Leveraging the theoretical advantages of reproducing kernel Hilbert space, MMD maps the feature distributions of different modalities into a high-dimensional space and achieves implicit alignment by minimizing their distributional differences. After local and global cross-modal alignment, all unimodal features are combined and fed into the Mamba backbone for further multimodal fusion. This dual alignment strategy ensures that Mamba can exploit both local and global relationships between modalities, thereby learning more comprehensive multimodal representations.

In summary, the contributions of this paper are threefold:

- We observe the limitation of directly applying Mamba to multimodal fusion tasks, which ignores more comprehensive cross-modal alignment information, and propose the AlignMamba framework to achieve efficient and effective multimodal fusion.
- We introduce an OT-based local alignment module for explicit learning of token-level correspondences, complemented by an MMD-based global alignment loss for implicit distribution alignment. These two types of alignment information complement each other, achieving comprehensive cross-modal alignment.

 Extensive experiments on both complete and incomplete multimodal fusion tasks demonstrate that AlignMamba achieves state-of-the-art results in terms of both effectiveness and efficiency.

2. Related work

2.1. Transformer-based Multimodal Fusion

Transformer [31], with its powerful modeling capabilities, has become the cornerstone architecture in modern neural networks. Existing multimodal fusion methods mainly rely on Transformers to model relationships between different modalities and learn multimodal fusion representations. These approaches can be categorized into two main types: multi-stream and single-stream methods.

Multi-stream methods employ cross-modal Transformers to model interactions between any two modalities. For vision-language pre-training tasks, models like ViL-BERT [23] and LXMERT [29] utilize two co-attention Transformer layers to model bidirectional relationships between visual and textual modalities. For audio-visualtextual trimodal fusion tasks, MulT [30] leverages crossmodal Transformers to model pairwise modal interactions, and then concatenate all bimodal fusion representations to obtain trimodal fusion representations. Similarly, CMA [30], based on cross-modal attention mechanisms, was proposed to fuse features from three modalities. More recently, BLIP-2 [14] introduced Q-Former, a lightweight querying Transformer architecture, to align vision-language modalities and learn multimodal fusion representations.

Single-stream methods adopt a more straightforward strategy by concatenating features from different modalities and feeding them into a Transformer encoder for cross-modal interaction and multimodal fusion. For instance, in vision-language pre-training tasks, VisualBERT [15] extracts features from key regions using object detectors and concatenates these region feature sequences with text to-ken embeddings before feeding them into a Transformer. In contrast, ViLT [11] replaces region feature sequences with image patch embedding sequences, discarding the object detection backbone and improving efficiency. Recent multimodal pre-training models, such as LLaVA [21], have adopted similar approaches to model cross-modal correspondences and learn multimodal fusion representations for downstream tasks.

Existing methods achieve cross-modal interaction and fusion through cross-attention or self-attention mechanisms, learning comprehensive and effective multimodal fusion representations. However, the quadratic time complexity of Transformers limits their efficiency when processing large-scale or long-sequence data. This limitation necessitates the development of novel multimodal fusion methods that balance effectiveness and efficiency.

2.2. Mamba-based Multimodal Fusion

As a novel architectural paradigm, Mamba [3] incorporates selection mechanisms and hardware-aware parallel algorithms into SSMs [4, 5], achieving efficient and effective sequence modeling in the language domain. Inspired by its success, recent studies have explored adapting Mamba for multimodal fusion tasks. For instance, Pan-mamba [8] and Fusion-mamba [2] incorporate features from other modalities as inputs to unimodal Mamba to enable cross-modal interaction and fusion. Similarly, MambaDFuse [17] and MTMamba [19] utilize multimodal representations as inputs to unimodal Mamba for cross-modal interaction and fusion. In contrast, some approaches adopt a simpler strategy: VL-Mamba [26] and Cobra [42], for example, concatenate visual and textual representation sequences before feeding them into Mamba for sequence modeling and multimodal fusion.

While these Mamba-based approaches demonstrate significant computational advantages compared to Transformer-based multimodal fusion methods, they face inherent limitations due to Mamba's sequential scanning mechanism. This mechanism makes it challenging to effectively learn cross-modal correspondences, particularly with unscanned tokens. The resulting loss in cross-modal alignment information may constrain the effectiveness of learned multimodal fusion representations. Therefore, how to effectively leverage cross-modal relationships within the Mamba framework to learn more comprehensive multimodal fusion representations remains an open research challenge.

3. Method

3.1. Overview

Fig. 2 presents the framework of our proposed Align-Mamba. Using audio-visual-language trimodal data as a case study, the framework first processes raw signals from each modality through modality-specific encoders to generate corresponding unimodal embedding sequences X_a , X_v , and X_l . The framework then employs two complementary alignment mechanisms: an OT-based local alignment module that captures token-level correspondences, and an MMD-based global alignment loss that ensures distribution-level consistency. These mechanisms yield aligned embedding sequences \tilde{X}_a and \tilde{X}_v (illustrated here by aligning audio and visual modalities to the language modality as the anchor). The aligned unimodal embeddings, which now incorporate cross-modal correspondence information, are subsequently processed by the Mamba backbone for multimodal fusion. The following sections provide a detailed description of each component.

3.2. OT-based Local Cross-modal Alignment

Optimal Transport provides a principled framework for comparing and aligning probability distributions by finding the optimal way to transform one distribution into another while minimizing the transportation cost [32]. In our multimodal alignment context, OT offers a natural way to establish token-level correspondences between different modalities by treating feature sequences as discrete distributions.

Given the unimodal feature sequences $X_a \in \mathbb{R}^{T_a \times d}$, $X_v \in \mathbb{R}^{T_v \times d}$, and $X_l \in \mathbb{R}^{T_l \times d}$ from audio, video, and language modalities respectively, where T_a, T_v , and T_l denote the sequence lengths of different modalities and d is the feature dimension, we aim to learn the transport matrix M that capture fine-grained correspondences between different modalities. Take video-to-language alignment as an example, the classical optimal transport problem can be formulated as follows:

$$\min_{T_{v2l}} \sum_{i=1}^{T_v} \sum_{j=1}^{T_l} M_{v2l}(i,j) C_{v2l}(i,j).$$
(1)

The optimization is constrained by:

$$\begin{cases} \sum_{j=1}^{T_l} M_{v2l}(i,j) = \frac{1}{T_v}, & \forall i \in [1,T_v] \\ \sum_{i=1}^{T_v} M_{v2l}(i,j) = \frac{1}{T_l}, & \forall j \in [1,T_l] \\ M_{v2l}(i,j) \ge 0, & \forall i,j \end{cases}$$
(2)

where $C_{v2l} \in \mathbb{R}^{T_v \times T_l}$ is the cost matrix. Given that the cosine distance emphasizes angular relationships between feature vectors while providing numerical stability through its bounded range, we use cosine distance as the cost matrix:

$$C_{v2l}(i,j) = 1 - \frac{X_v^i \cdot X_l^j}{||X_v^i||_2 ||X_l^j||_2}.$$
(3)

However, solving this OT problem is extremely computationally expensive. Following [12], we adopt a relaxed version by removing the incoming sum constraint:

$$\begin{cases} \sum_{j=1}^{T_l} M_{v2l}(i,j) = \frac{1}{T_v}, & \forall i \in [1, T_v] \\ M_{v2l}(i,j) \ge 0, & \forall i,j \end{cases}$$
(4)

This relaxed formulation allows each textual feature to be matched with multiple video features without constraining the total incoming flow, significantly reducing the computational complexity while maintaining the ability to capture meaningful cross-modal correspondences. The corresponding solution is defined as:

$$M_{v2l}(i,j) = \begin{cases} \frac{1}{T_v}, & j = \arg\min_j C_{v2l}(i,j), \\ 0, & j \neq \arg\min_j C_{v2l}(i,j). \end{cases}$$
(5)

Similarly, we compute the transport matrix M_{a2l} for audioto-language alignment. Finally, the aligned video and audio



Figure 2. AlignMamba enhances multimodal Mamba by incorporating token-level alignment and distribution-level alignment, enabling more effective multimodal fusion.

features can then be obtained through:

$$\begin{cases} \tilde{X}_v = M_{v2l}^\top X_v \in \mathbb{R}^{T_l \times d}, \\ \tilde{X}_a = M_{a2l}^\top X_a \in \mathbb{R}^{T_l \times d}. \end{cases}$$
(6)

This relaxed OT-based alignment process provides an efficient way to capture fine-grained cross-modal correspondences while maintaining computational tractability. The resulting transport matrices provide interpretable alignment information between different modalities. However, while this token-level alignment effectively captures local correspondences, ensuring global distribution-level consistency across modalities requires additional consideration, which we address through our MMD-based global alignment mechanism in the following section.

3.3. MMD-based Global Cross-modal Alignment

To ensure distribution-level consistency across modalities, we employ Maximum Mean Discrepancy as the global alignment metric. MMD measures the statistical discrepancy between different modalities in a high-dimensional Reproducing Kernel Hilbert Space (RKHS) by comparing all orders of their statistics. For two feature sequences X and Y, the squared MMD distance is defined as:

$$\text{MMD}^{2}(X,Y) = \left\| \frac{1}{T} \sum_{i=1}^{T} \phi(x_{i}) - \frac{1}{T} \sum_{j=1}^{T} \phi(y_{j}) \right\|_{\mathcal{H}}^{2}, \quad (7)$$

where $\phi(\cdot)$ is a feature mapping to a RKHS \mathcal{H} . Using the kernel trick, this can be computed as:

$$MMD^{2}(X,Y) = \frac{1}{T^{2}} \sum_{i=1}^{T} \sum_{i'=1}^{T} k(x_{i}, x_{i'}) + \frac{1}{T^{2}} \sum_{j=1}^{T} \sum_{j'=1}^{T} k(y_{j}, y_{j'}) - \frac{2}{T^{2}} \sum_{i=1}^{T} \sum_{j=1}^{T} k(x_{i}, y_{j}),$$
(8)

where $k(\cdot, \cdot)$ is a positive definite kernel function. In our implementation, we adopt the Gaussian kernel:

$$k(x,y) = \exp(-\frac{\|x-y\|_2^2}{2\sigma^2}),$$
(9)

where σ is the kernel bandwidth parameter.

For the aligned audio features X_a , the aligned video features \tilde{X}_v , and the language features X_l , the global alignment loss is defined as the sum of MMD distances between each pair of modalities:

$$\mathcal{L}_{\text{align}} = \text{MMD}^2(\tilde{X}_v, X_l) + \text{MMD}^2(\tilde{X}_a, X_l).$$
(10)

By minimizing this loss during training, we encourage the feature distributions of different modalities to be aligned in the RKHS. While OT establishes token-level correspondences, MMD ensures the consistency of overall feature distributions, providing complementary alignment signals at different granularities. This dual-alignment strategy facilitates more effective multimodal fusion in subsequent processing stages.

3.4. Mamba-based Fusion and Optimization

Mamba-based Multimodal Fusion. Following the local and global alignment processes, we employ Mamba to facilitate efficient multimodal fusion while maintaining its inherent linear computational complexity. Unlike traditional Transformer-based methods that process all tokens simultaneously through self-attention mechanisms, our approach implements a time-priority scanning strategy that preserves Mamba's sequential nature while enabling effective crossmodal interactions. Given the aligned audio features \tilde{X}_a , the aligned video features \tilde{X}_v , and the language features X_l , we construct a unified multimodal feature sequence X_{mm} by interleaving features from different modalities at each timestep:

$$X_{mm} = [\tilde{X}_{a}^{1}, \tilde{X}_{v}^{1}, X_{l}^{1}, \tilde{X}_{a}^{2}, \tilde{X}_{v}^{2}, X_{l}^{2}, ..., \tilde{X}_{a}^{T}, \tilde{X}_{v}^{T}, X_{l}^{T}],$$
(11)

where the superscript denotes the temporal index. This temporal-priority organization ensures that features from different modalities at the same timestep are processed sequentially, allowing the selective scan mechanism of Mamba to effectively capture both intra- and inter-modal dependencies. The fused representations are obtained by processing the constructed sequence through multiple Mamba layers.

Training Objective. The framework is optimized end-toend using a composite loss function that combines the taskspecific objective with the alignment constraints:

$$\mathcal{L} = \mathcal{L}_{task} + \lambda \mathcal{L}_{align}, \tag{12}$$

where \mathcal{L}_{task} is determined by the downstream task (e.g., cross-entropy loss for classification or mean squared error for regression), \mathcal{L}_{align} is the MMD-based alignment loss, and λ is a hyperparameter that balances the two objectives. During training, minimizing \mathcal{L}_{task} drives the model to learn task-relevant multimodal representations, while \mathcal{L}_{align} ensures consistent feature distributions across modalities.

4. Experiment

We evaluate our proposed method on two distinct multimodal fusion scenarios: complete multimodal fusion and incomplete multimodal fusion. In the complete fusion setting, all modalities are available during both training and inference, which tests the model's ability to effectively integrate complementary information across modalities. The incomplete fusion scenario, where certain modalities may be missing during inference, presents a more challenging yet practical setting that evaluates the model's robustness and adaptability to partial observations. Through extensive experiments on these two scenarios, we demonstrate the effectiveness of our approach in both ideal conditions and more challenging practical situations.

4.1. Datasets and Evaluation Metrics

We conduct experiments on two multimodal representation fusion benchmarks: CMU-MOSI [39] and CMU-MOSEI [40]. Both datasets consist of video segments collected from online platforms, containing visual (facial expressions), acoustic (voice), and textual (transcribed speech) modalities. Compared to CMU-MOSI, CMU-MOSEI exhibits greater diversity in terms of speakers, topics, and recording conditions. Each segment in both datasets is annotated with a sentiment score ranging from -3 (highly negative) to +3 (highly positive). These scores are binarized into positive and negative sentiments for classification. To evaluate the effectiveness of our method, we adopt the following metrics based on previous works [18, 34]: binary accuracy and binary F_1 score.

4.2. Comparison with SoTA methods

4.2.1. Results on Complete Multimodal Fusion Tasks

Table 2 presents a comprehensive comparison between our approach and various state-of-the-art methods on the complete multimodal representation fusion task, which can be categorized into three main groups: (1) LSTM methods, including ICCN [28], MISA [7], and MMIM [6]; (2) Cross-modal Transformer methods: MuIT [30], Self-MM [38], and DMD [16]; (3) Contrastive learning methods: HyCon [24], Confede [36], and MTMD [20].

On one hand, AlignMamba additionally incorporates token-level alignment to enhance multimodal fusion compared to contrastive learning approaches. On the other hand, AlignMamba's advantage over cross-modal Transformer methods lies in its consideration of distributional alignment relationships. Consequently, AlignMamba attains the best performance on all metrics in both datasets. For example, on the CMU-MOSI dataset, AlignMamba achieves a binary classification accuracy of 86.9%, representing a 0.9% improvement over previous methods. These results can be ascribed to AlignMamba's capacity to con-

Dataset	Missing	DCCA [1]	DCCAE [33]	MCTN [25]	MMIN [43]	GCNet [18]	IMDer [34]	AlignMamba
MOSI	10%	72.1 / 72.2	74.5 / 74.7	78.4 / 78.5	81.8 / 81.8	82.3 / 82.3	84.9 / 84.8	85.7 / 85.6
	20%	69.3 / 69.1	71.8/71.9	75.6 / 75.7	79.0 / 79.1	79.4 / 79.5	83.5 / 83.4	84.3 / 84.1
	30%	65.4 / 65.2	67.0 / 66.7	71.3 / 71.2	76.1 / 76.2	77.2 / 77.2	81.2 / 81.0	82.2 / 82.2
	40%	62.8 / 62.0	63.6 / 62.8	68.0 / 67.6	71.7 / 71.6	74.3 / 74.4	78.6 / 78.5	80.0 / 79.6
	50%	60.9 / 59.9	62.0/61.3	65.4 / 64.8	67.2 / 66.5	70.0 / 69.8	76.2 / 75.9	77.6 / 77.3
	60%	58.6 / 57.3	59.6 / 58.5	63.8 / 62.5	64.9 / 64.0	67.7 / 66.7	74.7 / 74.0	75.8 / 75.1
	70%	57.4 / 56.0	58.1 / 57.4	61.2 / 59.0	62.8 / 61.0	65.7 / 65.4	71.9 / 71.2	73.8 / 73.2
	Avg.	63.8 / 63.1	65.2 / 64.8	69.1 / 68.5	71.9 / 71.5	73.8/73.6	78.7 / 78.4	79.9 / 79.6
	Δ	14.7 / 16.2	16.4 / 17.3	17.2 / 19.5	19.0 / 20.8	16.6 / 16.9	13.0 / 13.6	11.9 / 12.4
MOSEI	10%	77.4/77.3	78.4 / 78.3	81.8 / 81.6	81.9 / 81.3	82.3 / 82.1	84.8 / 84.6	85.4 / 85.4
	20%	73.8 / 74.0	75.5 / 75.4	79.0 / 78.7	79.8 / 78.8	80.3 / 79.9	82.7 / 82.4	83.6 / 83.3
	30%	71.1 / 71.2	72.3 / 72.2	76.9 / 76.2	77.2 / 75.5	77.5 / 76.8	81.3 / 80.7	82.5 / 81.0
	40%	69.5 / 69.4	70.3 / 70.0	74.3 / 74.1	75.2 / 72.6	76.0 / 74.9	79.3 / 78.1	81.7 / 80.5
	50%	67.5 / 65.4	69.2 / 66.4	73.6 / 72.6	73.9 / 70.7	74.9 / 73.2	79.0 / 77.4	80.1 / 78.7
	60%	66.2 / 63.1	67.6/63.2	73.2 / 71.1	73.2 / 70.3	74.1 / 72.1	78.0 / 75.5	79.4 / 78.2
	70%	65.6/61.0	66.6 / 62.6	72.7 / 70.5	73.1 / 69.5	73.2 / 70.4	77.3 / 74.6	78.8 / 76.9
	Avg.	70.2 / 68.8	71.4 / 69.7	75.9 / 75.0	76.3 / 74.1	76.9 / 75.6	80.3 / 79.0	81.6 / 80.6
	Δ	11.8 / 16.3	11.8 / 15.7	9.1 / 11.1	8.8 / 11.8	9.1 / 11.7	7.5 / 10.0	6.6 / 8.5

Table 1. Performance comparison on CMU-MOSI and CMU-MOSEI datasets. Results are reported as Accuracy / F_1 (%). Δ : performance drop from 10% to 70% missing rate (lower is better).

Method	CMU-MOSI	CMU-MOSEI	
ICCN [28]	83.0 / 83.0	84.2 / 84.2	
MISA [7]	83.4 / 83.6	85.5 / 85.3	
MulT [30]	84.1 / 83.9	82.5 / 82.3	
MAG-BERT [27]	84.3 / 84.6	84.8 / 84.7	
CM-BERT [37]	84.5 / 84.5	83.6 / 83.6	
ULGM [9]	84.5 / 84.5	85.0 / 85.1	
FDMER [35]	84.6 / 84.7	86.1 / 85.8	
Self-MM [38]	84.8 / 84.9	85.0 / 84.9	
MMIM [6]	85.1 / 85.0	85.1 / 85.0	
HyCon [24]	85.2 / 85.1	85.4 / 85.6	
Confede [36]	85.5 / 85.5	85.8 / 85.8	
AOBERT [10]	85.6 / 86.4	86.2 / 85.9	
DMD [16]	85.8 / 85.8	86.0 / 86.1	
MTMD [20]	86.0 / 86.0	86.1 / 85.9	
AlignMamba	86.9 / 86.9	86.6 / 86.5	

Table 2. Performance comparison on CMU-MOSI and CMU-MOSEI datasets. Results are reported as Accuracy / F_1 (%).

duct extensive cross-modal alignment by leveraging its local alignment module and global alignment loss, thereby adeptly exploiting cross-modal correlations across different granularities and enabling the learning of more effective multimodal fusion representations.

4.2.2. Results on Incomplete Multimodal Fusion Tasks

Table 1 presents experimental results on incomplete multimodal representation fusion tasks. We compare Align-Mamba with various state-of-the-art methods, which can be categorized into two main groups: (1) modality recovery approaches, including MCTN [25], MMIN [43], GCNet [18], and IMDer [34], which attempt to reconstruct missing modalities from available ones; and (2) non-recovery approaches, such as DCCA [1] and DCCAE [33], which directly learn from available modalities.

The results demonstrate that AlignMamba consistently outperforms existing methods across different missing rates, achieving an average accuracy of 79.9% on the CMU-MOSI dataset, a 1.2% improvement over previous methods. More importantly, AlignMamba demonstrates stronger robustness to increasing modality missing rates. For instance, on the CMU-MOSI dataset, while MMIN and IMDer experience significant performance degradation with accuracy drops of 19.0% and 13.0% respectively, AlignMamba shows better resilience with only an 11.9% decrease in binary classification accuracy.

In conclusion, these improvements on both complete and incomplete multimodal fusion tasks can be attributed to the proposed dual alignment strategy: the local token-level alignment and global distribution-level alignment mechanisms work together to capture comprehensive cross-modal correspondences. This dual alignment strategy, combined with Mamba's efficient sequence modeling capabilities, not only enables learning more comprehensive and accurate multimodal fusion representations in complete multimodal scenarios, but also improves the robustness of learned representations in incomplete multimodal settings.

4.3. Efficiency Analysis

We conduct comprehensive efficiency analysis for Align-Mamba and compare them against both single-stream and multi-stream Transformer methods. Our evaluation metrics consist of GPU memory usage, inference time, and computational complexity. For a fair comparison, we specifically focus on the cross-modal interaction and fusion components, excluding the computational costs of unimodal encoders. All experiments are performed under identical conditions.

4.3.1. GPU Memory Usage

First, we report the GPU memory usage of each method with respect to varying input sequence lengths in Fig. 3. We exclude multi-stream Transformers on the 12.8k-token setting as they encounter an out-of-memory error. Align-Mamba consistently achieves the best trade-off between sequence length and memory usage across all settings, surpassing other Transformer-based approaches by a non-trivial margin. For instance, when processing 6.4k tokens, AlignMamba requires only 8.53 GB of memory, achieving 20.3% and 58.0% memory reduction compared to single-stream (10.7 GB) and multi-stream (20.3 GB) Transformers, respectively. This significant advantage in memory consumption is particularly valuable for processing longer sequences and deploying models on resource-constrained devices.



Figure 3. GPU memory usage comparison with varying lengths.

4.3.2. Inference Time

Next, we report the inference time of each method with respect to varying input sequence lengths in Fig. 4. To ensure fairness, we aggregate the running time of 50 inference passes for each model. AlignMamba again demonstrates consistent and substantial speed advantages over Transformer-based approaches across all settings. For instance, when processing 6.4k tokens, AlignMamba takes only 6.05 seconds, achieving 83.3% and 87.6% reduction in inference time compared to single-stream (36.13s) and multi-stream (48.61s) Transformers, respectively.



Figure 4. Inference time comparison with varying lengths.

4.3.3. Computational Complexity

Finally, we analyze the FLOPs required by each method to quantify their computational efficiency. Without loss of generality, we fix the input sequence length to 1024 for each model. AlignMamba demonstrates superior efficiency with only 46.7G FLOPs, compared to 101.6G FLOPs for single-stream Transformer and 203.2G FLOPs for multi-stream Transformer. This represents a reduction of more than 54% compared to single-stream and 77% compared to multi-stream approaches, highlighting AlignMamba's computational advantages in cross-modal alignment and multi-modal fusion tasks. This also justifies the lower memory consumption and swift inference speed as presented in the previous sections.

4.4. Ablation study

We conduct comprehensive ablation studies from three aspects to evaluate our proposed method, as shown in Table 3.

Component analysis. First, we evaluate the effectiveness of the OT-based local alignment module and MMD-based global alignment loss. Removing either component led to performance degradation across both datasets. For instance, on the CMU-MOSI dataset, accuracy dropped by 2.3% and 1.1% respectively. Notably, the OT-based alignment module demonstrated superior performance compared to the MMD-based alignment loss, likely because OT-based alignment provides explicit alignment plans while MMD-based alignment only imposes implicit alignment constraints.

Mamba-based fusion. Furthermore, we ablate Align-Mamba with regular single-stream [42] and multi-stream Mamba-based fusion methods [2] to show the effectiveness of our method in terms of multimodal fusion. Results demonstrate reduced performances in these two Mambabased methods, suggesting their lack of explicit consideration of inter-modal correspondences, which makes it difficult to learn comprehensive cross-modal relationships. This shows that naive Mamba architecture alone does not suffice in effective multimodal fusion and highlights both the lim-



Figure 5. The learned optimal transport plan. We only show the transport plan between video and language modalities for brevity.

itations of Mamba's original scanning mechanism and the necessity of our proposed cross-modal alignment.

Modality ablations. Lastly, we conduct modality ablation experiments by removing one modality at a time. When the text modality is removed, we only align the audio modality with the video modality. This results in significant performance degradation, likely due to the strong correlation between language and emotions. In contrast, removing the audio modality results in a smaller performance drop, possibly due to the sizable presence of irrelevant information in the audio modality such as background noise, reducing its impact on the overall performance.

	CMU-MOSI	CMU-MOSEI					
AlignMamba	86.9 / 86.9	86.6 / 86.5					
Alignment							
w/o Local	84.6 / 84.4	84.1 / 84.0					
w/o Global	85.8 / 85.7	85.7 / 85.5					
Fusion							
Single-stream	82.3 / 82.1	81.8 / 81.4					
Multi-stream	83.7 / 83.5	83.5 / 83.2					
Modality							
w/o Audio	84.4 / 84.6	83.9 / 83.5					
w/o Video	83.7 / 83.8	83.3 / 82.8					
w/o Language	65.3 / 63.4	64.6 / 62.2					

Table 3. Ablation studies on CMU-MOSI and CMU-MOSEI datasets. Results are reported as Accuracy / F_1 (%).

4.5. Further Analysis

4.5.1. Cross-modal Alignment

To quantitatively assess our dual alignment strategy, we measure the A-distance between modality pairs in Table 4. The A-distance $\in [0, 2]$ is a common metric for domain discrepancy, with higher values indicating greater modality differences. A_{al} and A_{vl} represents the audio-language and video-language distances respectively. The results reveal significant and consistent reductions in inter-

modal distances through our dual alignment strategy in both CMU-MOSI and CMU-MOSEI. These improvements demonstrate the effectiveness of our strategy in bridging the modality gap by learning meaningful cross-modal correlations, leading to more robust multimodal fusion representations.

	CMU-MOSI		CMU-MOSE	
	$ \mathcal{A}_{al} $	\mathcal{A}_{vl}	$\mid \mathcal{A}_{al}$	\mathcal{A}_{vl}
w/ Dual-align	1.59	1.49	1.61	1.53
w/o Dual-align	1.68	1.57	1.72	1.65

4.5.2. Optimal Transport Plan

Here, we qualitatively present the learned optimal transport plan. Figure 5 illustrates an example from the CMU-MOSI dataset. Notice that modalities exhibit temporal misalignment: the sentiment correspondence between different modalities may appear at different timesteps, which poses a challenge for multimodal representation fusion. For instance, the visual modality exhibits negative expressions at the beginning, while the textual modality introduces negative words toward the end. The original Mamba model struggles to explicitly learn these correspondences due to its sequential scanning mechanisms. In contrast, our proposed method leverages optimal transport to explicitly transform and align features in different temporal stages across modalities, reducing the modality gap and improving the effectiveness of multimodal fusion.

5. Conclusion

In this paper, we proposed AlignMamba, an efficient and effective method for multimodal representation fusion. By integrating OT-based local alignment and MMD-based global alignment, our method captures comprehensive cross-modal relationships while maintaining lower computational complexity. Extensive experiments on both complete and incomplete multimodal fusion tasks demonstrate that AlignMamba achieves state-of-the-art performance with significantly reduced computational costs.

References

- Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *International conference on machine learning*, pages 1247–1255. PMLR, 2013. 6
- [2] Wenhao Dong, Haodong Zhu, Shaohui Lin, Xiaoyan Luo, Yunhang Shen, Xuhui Liu, Juan Zhang, Guodong Guo, and Baochang Zhang. Fusion-mamba for cross-modality object detection. arXiv preprint arXiv:2404.09146, 2024. 1, 3, 7
- [3] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1, 3
- [4] Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021. 1, 3
- [5] Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. *Advances in Neural Information Processing Systems*, 35:35971–35983, 2022. 1, 3
- [6] Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*, 2021. 5, 6
- [7] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the* 28th ACM international conference on multimedia, pages 1122–1131, 2020. 5, 6
- [8] Xuanhua He, Ke Cao, Keyu Yan, Rui Li, Chengjun Xie, Jie Zhang, and Man Zhou. Pan-mamba: Effective pan-sharpening with state space model. arXiv preprint arXiv:2402.12192, 2024. 1, 3
- [9] Yewon Hwang and Jong-Hwan Kim. Self-supervised unimodal label generation strategy using recalibrated modality representations for multimodal sentiment analysis. In *Findings of the Association for Computational Linguistics: EACL* 2023, pages 35–46, 2023. 6
- [10] Kyeonghun Kim and Sanghyun Park. Aobert: Allmodalities-in-one bert for multimodal sentiment analysis. *Information Fusion*, 92:37–45, 2023. 6
- [11] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Visionand-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021. 1, 2
- [12] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957– 966. PMLR, 2015. 3
- [13] Jun-Tae Lee, Mihir Jain, Hyoungwoo Park, and Sungrack Yun. Cross-attentional audio-visual fusion for weaklysupervised action localization. In *International conference on learning representations*, 2020. 1

- [14] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730– 19742. PMLR, 2023. 2
- [15] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. arXiv preprint arXiv:1908.03557, 2019. 1, 2
- [16] Yong Li, Yuanzhi Wang, and Zhen Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of* the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6631–6640, 2023. 5, 6
- [17] Zhe Li, Haiwei Pan, Kejia Zhang, Yuhua Wang, and Fengming Yu. Mambadfuse: A mamba-based dual-phase model for multi-modality image fusion. arXiv preprint arXiv:2404.08406, 2024. 1, 3
- [18] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023. 5, 6
- [19] Baijiong Lin, Weisen Jiang, Pengguang Chen, Yu Zhang, Shu Liu, and Ying-Cong Chen. Mtmamba: Enhancing multitask dense scene understanding by mamba-based decoders. arXiv preprint arXiv:2407.02228, 2024. 3
- [20] Ronghao Lin and Haifeng Hu. Multi-task momentum distillation for multimodal sentiment analysis. *IEEE Transactions* on Affective Computing, 2023. 5, 6
- [21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. arXiv preprint arXiv:2304.08485, 2023. 1, 2
- [22] Jiaming Liu, Mengzhen Liu, Zhenyu Wang, Lily Lee, Kaichen Zhou, Pengju An, Senqiao Yang, Renrui Zhang, Yandong Guo, and Shanghang Zhang. Robomamba: Multimodal state space model for efficient robot reasoning and manipulation. arXiv preprint arXiv:2406.04339, 2024. 1
- [23] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. Advances in neural information processing systems, 32, 2019. 1, 2
- [24] Sijie Mai, Ying Zeng, Shuangjia Zheng, and Haifeng Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 2022. 5, 6
- [25] Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6892–6899, 2019. 6
- [26] Yanyuan Qiao, Zheng Yu, Longteng Guo, Sihan Chen, Zijia Zhao, Mingzhen Sun, Qi Wu, and Jing Liu. Vl-mamba: Exploring state space models for multimodal learning. arXiv preprint arXiv:2403.13600, 2024. 1, 3
- [27] Wasifur Rahman, Md Kamrul Hasan, Sangwu Lee, Amir Zadeh, Chengfeng Mao, Louis-Philippe Morency, and Ehsan Hoque. Integrating multimodal information in large pre-

trained transformers. In *Proceedings of the conference*. *Association for Computational Linguistics*. *Meeting*, page 2359. NIH Public Access, 2020. 6

- [28] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 8992–8999, 2020. 5, 6
- [29] Hao Tan and Mohit Bansal. Lxmert: Learning crossmodality encoder representations from transformers. arXiv preprint arXiv:1908.07490, 2019. 1, 2
- [30] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, page 6558. NIH Public Access, 2019. 1, 2, 5, 6
- [31] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 1, 2
- [32] Cédric Villani. *Topics in optimal transportation*. American Mathematical Soc., 2021. 3
- [33] Weiran Wang, Raman Arora, Karen Livescu, and Jeff Bilmes. On deep multi-view representation learning. In *International conference on machine learning*, pages 1083– 1092. PMLR, 2015. 6
- [34] Yuanzhi Wang, Yong Li, and Zhen Cui. Incomplete multimodality-diffused emotion recognition. Advances in Neural Information Processing Systems, 36, 2024. 5, 6
- [35] Dingkang Yang, Shuai Huang, Haopeng Kuang, Yangtao Du, and Lihua Zhang. Disentangled representation learning for multimodal emotion recognition. In *Proceedings of the 30th* ACM International Conference on Multimedia, pages 1642– 1651, 2022. 6
- [36] Jiuding Yang, Yakun Yu, Di Niu, Weidong Guo, and Yu Xu. Confede: Contrastive feature decomposition for multimodal sentiment analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7617–7630, 2023. 5, 6
- [37] Kaicheng Yang, Hua Xu, and Kai Gao. Cm-bert: Crossmodal bert for text-audio sentiment analysis. In Proceedings of the 28th ACM international conference on multimedia, pages 521–528, 2020. 6
- [38] Wenmeng Yu, Hua Xu, Ziqi Yuan, and Jiele Wu. Learning modality-specific representations with self-supervised multitask learning for multimodal sentiment analysis. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10790–10797, 2021. 5, 6
- [39] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. arXiv preprint arXiv:1606.06259, 2016. 5
- [40] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th*

Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2236–2246, 2018. 1, 5

- [41] Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. Multimodal chain-ofthought reasoning in language models. arXiv preprint arXiv:2302.00923, 2023. 1
- [42] Han Zhao, Min Zhang, Wei Zhao, Pengxiang Ding, Siteng Huang, and Donglin Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. arXiv preprint arXiv:2403.14520, 2024. 1, 3, 7
- [43] Jinming Zhao, Ruichen Li, and Qin Jin. Missing modality imagination network for emotion recognition with uncertain missing modalities. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pages 2608– 2618, 2021. 6
- [44] Jiahao Zheng, Sen Zhang, Zilu Wang, Xiaoping Wang, and Zhigang Zeng. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE Transactions on Multimedia*, 25: 2213–2225, 2022. 1