
AlignMamba-2: Enhancing Multimodal Fusion and Sentiment Analysis with Modality-Aware Mamba

Yan Li¹, Yifei Xing¹, Xiangyuan Lan¹, Xin Li¹, Haifeng Chen², Dongmei Jiang^{3,1*}

¹ Pengcheng Laboratory, Shenzhen, China

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

³ School of Computer Science, Northwestern Polytechnical University, Xi'an, China

In the era of large-scale pre-trained models, effectively adapting general knowledge to specific affective computing tasks remains a challenge, particularly regarding computational efficiency and multimodal heterogeneity. While Transformer-based methods have excelled at modeling inter-modal dependencies, their quadratic computational complexity limits their use with long-sequence data. Mamba-based models have emerged as a computationally efficient alternative; however, their inherent sequential scanning mechanism struggles to capture the global, non-sequential relationships that are crucial for effective cross-modal alignment. To address these limitations, we propose AlignMamba-2, an effective and efficient framework for multimodal fusion and sentiment analysis. Our approach introduces a dual alignment strategy that regularizes the model using both Optimal Transport distance and Maximum Mean Discrepancy, promoting geometric and statistical consistency between modalities without incurring any inference-time overhead. More importantly, we design a Modality-Aware Mamba layer, which employs a Mixture-of-Experts architecture with modality-specific and modality-shared experts to explicitly handle data heterogeneity during the fusion process. Extensive experiments on four challenging benchmarks, including dynamic time-series (on the CMU-MOSI and CMU-MOSEI datasets) and static image-related tasks (on the NYU-Depth V2 and MVSA-Single datasets), demonstrate that AlignMamba-2 establishes a new state-of-the-art in both effectiveness and efficiency across diverse pattern recognition tasks, ranging from dynamic time-series analysis to static image-text classification.

1. Introduction

In the field of pattern recognition, recognizing complex concepts often requires integrating information from heterogeneous sources such as audio, vision, and language (Li et al., 2024b; Xiao et al., 2025; Xu et al., 2025). A central problem in designing robust recognition systems lies in bridging the inherent *heterogeneity gap* among these modalities, where each is characterized by distinct statistical distributions and structural patterns (Xiao et al., 2024; Xu et al., 2026). Effectively aligning and fusing these disparate data streams to generate a cohesive and comprehensive representation remains a significant and open research problem.

In the era of large-scale pre-trained models, Transformer-based architectures have established themselves as the foundation of modern AI, powering state-of-the-art Large Language Models (LLMs) and Vision Language Models (VLMs). In the context of multimodal fusion, existing methods typically leverage these powerful backbones to model complex inter-modal dependencies. These methods can be broadly categorized into two paradigms: single-stream approaches (Kim et al., 2021; Li et al., 2019; Liu et al., 2023), which concatenate unimodal features and process them through a shared Transformer encoder, and multi-stream approaches (Huang and Xu, 2025; Tsai et al., 2019; Zheng et al., 2022), which employ dedicated encoders for each modality followed by cross-attention mechanisms for interaction. However, adapting these massive general-purpose models to affective

computing tasks faces a critical bottleneck. They are fundamentally constrained by the quadratic computational complexity of the self-attention mechanism (Li et al., 2025c; Vaswani et al., 2017). This limitation severely hinders their efficiency in finetuning and deployment, particularly for tasks involving long sequences or limited resources, which acts as a barrier to realizing the full potential of large-scale affective computing.

The recent introduction of State Space Models (SSMs) (Gu et al., 2021), particularly the Mamba architecture (Gu and Dao, 2023), offers a promising path for the next generation of efficient large-scale models. Mamba achieves linear computational complexity while maintaining strong performance in modeling long-range dependencies. This makes it an ideal candidate to address the efficiency challenges inherent in the current large-scale model era, effectively replacing the computationally expensive Transformer backbone. This breakthrough has sparked considerable interest in adapting Mamba for multimodal fusion and sentiment analysis tasks, with approaches ranging from direct feature concatenation (Qiao et al., 2024; Zhao et al., 2025) to multi-stream architectures (Dong et al., 2024; Gan et al., 2025; He et al., 2024). However, a direct application of Mamba to multimodal tasks reveals a critical limitation. As illustrated in Figure 1, Mamba’s core strength, its efficient sequential scanning mechanism, becomes a fundamental weakness when modeling cross-modal relationships. The sequential scan struggles to capture the global, non-sequential dependencies between a token being processed and all tokens from other modalities, especially those that have not yet been scanned (Li et al., 2025b). This issue can lead to incomplete cross-modal information exchange and suboptimal alignment, thereby compromising the quality of the final fused representation. For example, concurrent Mamba-based multimodal methods, such as VL-Mamba (Qiao et al., 2024) and Fusion-Mamba (Dong et al., 2024), predominantly rely on direct feature concatenation or simple multi-stream interaction. While effective for general sequence modeling, these approaches lack explicit mechanisms to align heterogeneous distributions before fusion and treat the Mamba backbone as a modality-agnostic processor. Consequently, they struggle to capture the fine-grained, non-sequential cross-modal dependencies that are critical for multimodal learning and sentiment analysis.

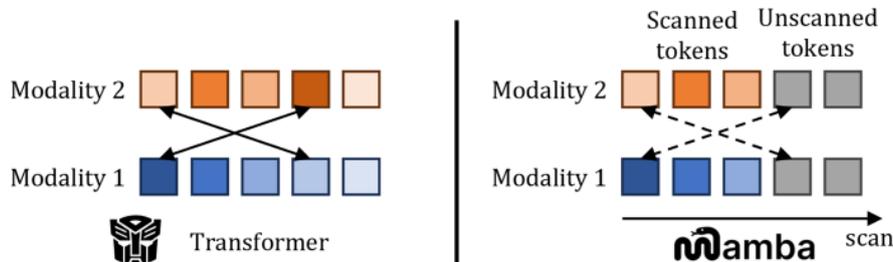


Figure 1 | The cross-attention mechanism in Transformer (left) and the scanning mechanism in Mamba (right).

To address these challenges, we propose a new framework, **AlignMamba-2**, which provides a holistic solution by enhancing Mamba for multimodal fusion at two crucial stages: principled alignment *before* fusion and modality-aware processing *during* fusion. First, we introduce a dual alignment strategy that serves as a powerful regularization term. This strategy employs two complementary distribution metrics: **Maximum Mean Discrepancy (MMD)** and **Optimal Transport (OT) distance**. MMD enforces consistency by matching the high-order statistical moments of the feature distributions, ensuring they share similar global properties. In parallel, OT distance evaluates the dissimilarity from a geometric perspective, minimizing the cost required to transform one distribution into another, thus encouraging a fine-grained alignment. Crucially, this dual alignment strategy directly complements the inherent characteristics of Mamba’s selective scan mechanism. Unlike Transformers, which can implicitly learn alignment via dense global attention maps, Mamba’s sequential nature limits its ability

to capture non-local, cross-modal correspondences spontaneously. By enforcing specific geometric (via OT) and statistical (via MMD) consistency prior to fusion, our strategy regularizes the latent space. This explicitly compensates for the lack of a global receptive field in the scanning process, ensuring that the computationally efficient selective scan operates on well-aligned representations, thereby maximizing the effectiveness of the linear-time fusion. Second, and more importantly, we introduce a novel **Modality-Aware Mamba layer**. By replacing standard projection layers with a Mixture-of-Experts (MoE) structure, composed of modality-specific and modality-shared experts, our model can process tokens differently based on their origin modality. This allows the fusion backbone to explicitly capture both unique unimodal properties and shared cross-modal patterns, leading to a more effective and nuanced fusion.

In summary, the main contributions of this work are fourfold:

- We present a critical analysis of existing Mamba-based multimodal methods, identifying the inherent limitations of the sequential scanning mechanism in capturing comprehensive cross-modal relationships and the modality-agnostic nature of the fusion process.
- We propose a dual alignment strategy using MMD and OT distance. This approach ensures robust cross-modal alignment from both statistical and geometric perspectives and, crucially, adds no extra computational cost during the inference phase.
- We introduce a novel Modality-Aware Mamba layer, which integrates a Mixture-of-Experts design to explicitly model both modality-specific and modality-invariant information within the fusion backbone, enabling a more sophisticated and effective fusion process.
- We conduct experiments on a diverse set of multimodal fusion and sentiment analysis benchmarks, including dynamic tasks (on CMU-MOSI and CMU-MOSEI datasets) and static tasks (on NYU-Depth V2 and MVSA-Single datasets), demonstrating the superior performance and broad applicability of AlignMamba-2.

This paper is an extended version of our preliminary work presented in (Li et al., 2025b). While the foundational idea of leveraging Mamba for aligned multimodal fusion is shared, this work introduces substantial advancements and a much broader scope of analysis. Specifically, the key extensions are threefold: **(1)** We replace the explicit OT matrix computation from our prior work with the OT distance-based loss. This architectural refinement not only simplifies the model but also eliminates all alignment-related computational overhead during inference, enhancing the model’s practicality for real-world deployment. **(2)** We move beyond simple pre-fusion alignment and introduce a novel Modality-Aware Mamba layer. This core architectural innovation endows the fusion backbone itself with the ability to process modality-specific and modality-invariant information, addressing a key limitation of modality-agnostic sequence models. **(3)** We significantly expand the empirical validation of our framework. In addition to the original multimodal sentiment analysis tasks, we now include comprehensive evaluations on the out-of-distribution setting and static multimodal tasks, namely RGB-D scene recognition and image-text classification, thereby demonstrating the versatility and generalizability of our approach across diverse data types and applications.

2. Related Work

In this section, we review prior work from four perspectives relevant to our proposed method: Transformer-based multimodal fusion, Mamba-based multimodal fusion, the Mixture-of-Experts paradigm, and techniques for multimodal representation alignment.

2.1. Transformer-based Multimodal Fusion

The Transformer architecture (Vaswani et al., 2017), with its powerful self-attention mechanism, has long dominated the field of multimodal representation learning. Existing approaches can be broadly classified into two main streams. **Single-stream** models, such as VisualBERT (Li et al., 2019) and ViLT (Kim et al., 2021), concatenate feature sequences from different modalities into a single sequence and process it through a unified Transformer encoder. This paradigm facilitates late fusion by allowing tokens from all modalities to interact within the same attention space (Zhang et al., 2023b). In contrast, **multi-stream** models like MulT (Tsai et al., 2019) and MM-PEAR-CoT (Li et al., 2025a) maintain separate backbones for each modality and employ cross-attention layers (Huang and Xu, 2025) to enable explicit, pairwise information exchange between them. These methods have demonstrated strong performance by capturing intricate cross-modal dependencies. However, their reliance on the self-attention mechanism, which has a computational complexity quadratic to the input sequence length, creates a fundamental efficiency bottleneck. This makes them prohibitively expensive for long-sequence tasks and resource-constrained environments.

In contrast, AlignMamba-2 directly addresses this fundamental efficiency bottleneck by adopting a linear-time Mamba backbone, while simultaneously introducing mechanisms to overcome Mamba’s inherent limitations in cross-modal modeling.

2.2. Mamba-based Multimodal Fusion

Inspired by the success of Mamba in natural language processing (Gu and Dao, 2023), a growing body of work has explored its application to multimodal tasks. These methods often adopt straightforward strategies, such as concatenating multimodal features for a single Mamba backbone (Qiao et al., 2024; Zhao et al., 2025) or using multi-stream designs (Dong et al., 2024; Gan et al., 2025; He et al., 2024). However, they face two significant challenges. First, as discussed, the inherent sequential scanning mechanism of Mamba makes it difficult to model the complex, non-sequential relationships between tokens across different modalities. Our previous work, AlignMamba-1 (Li et al., 2025b), partially addressed this by introducing an explicit pre-fusion alignment module. Second, existing approaches typically treat the Mamba block as a generic, modality-agnostic sequence processor. This overlooks the fact that different modalities possess unique intrinsic properties that may require specialized processing during the fusion stage (Bao et al., 2022).

Our work, AlignMamba-2, provides a more comprehensive solution. It not only employs a principled alignment strategy to address the scanning limitation but also redesigns the fusion core itself with a Modality-Aware Mamba layer to handle data heterogeneity, a capability absent in prior Mamba-based models.

2.3. Mixture-of-Experts

The Mixture-of-Experts paradigm is a powerful technique for increasing a model’s capacity without a proportional increase in computational cost. An MoE layer consists of multiple "expert" subnetworks and a "gating" network that sparsely activates a subset of these experts for each input token. This approach has been successfully employed to scale up large language models to trillions of parameters, such as in GShard (Lepikhin et al., 2021), Switch Transformers (Fedus et al., 2022), and DeepSeek-MoE (Dai et al., 2024a). In these applications, the primary goal is to expand model knowledge and capability while keeping inference costs manageable through learnable routing.

Our work introduces this powerful architecture into the Mamba framework for a novel purpose. We design a Modality-Aware Mamba layer that features a deterministic MoE structure composed of

both **modality-specific experts** and a **modality-shared expert**. This design enables the model to simultaneously learn modality-specific characteristics and general cross-modal patterns, leading to a more comprehensive and effective multimodal fusion representation.

2.4. Multimodal Representation Alignment

Bridging the heterogeneity gap between modalities necessitates effective representation alignment, which can be broadly categorized into implicit and explicit approaches. **Implicit alignment** methods learn a shared embedding space by optimizing a specific objective function, without computing an explicit token-to-token correspondence matrix (Dai et al., 2024b). This includes contrastive learning methods (Lin and Hu, 2023; Mai et al., 2022), which pull representations of corresponding (positive) multimodal pairs together (Li et al., 2023b). Another prominent direction is distribution matching. Techniques like Deep Canonical Correlation Analysis (DCCA) aim to maximize the correlation between the latent representations of different modalities (Sun et al., 2020), while Maximum Mean Discrepancy directly minimizes the distance between feature distributions in a Reproducing Kernel Hilbert Space (RKHS). These methods enforce global statistical consistency between modalities. **Explicit alignment** methods, on the other hand, focus on finding specific correspondences. A notable example is Optimal Transport (Villani, 2021), which computes a transport plan (an explicit matrix) that details the most efficient way to map tokens from one modality to another (Cao et al., 2022; Li et al., 2025b). While this approach provides strong, interpretable alignment, the computation of the transport matrix can introduce significant computational overhead, particularly for long sequences.

In contrast, we integrate a dual alignment strategy, using MMD and OT distance as regularization losses, directly with the Mamba architecture. This approach not only provides a robustly aligned foundation from both statistical and geometric perspectives but, crucially, does so entirely during the training phase. By formulating alignment as a training objective rather than a computational step at inference, our method ensures that Mamba can effectively process aligned features without incurring any extra computational cost once the model is deployed.

3. Methodology

In this section, we first provide a brief overview of the Mamba architecture as a preliminary. We then present the overall framework of our proposed AlignMamba-2. Subsequently, we detail its core components: the unimodal encoding and alignment strategy, and the novel Modality-Aware Mamba layer for fusion. Finally, we outline the training objective and provide a summary of the algorithm.

3.1. Preliminaries: The Mamba Architecture

The Mamba architecture (Gu and Dao, 2023) is built upon a structured State Space Model (SSM) (Gu et al., 2021), which maps a 1-D input sequence $u(t) \in \mathbb{R}$ to an output $y(t) \in \mathbb{R}$ through a latent state $h(t) \in \mathbb{R}^N$. The continuous-time formulation is given by:

$$\begin{aligned} h'(t) &= \mathbf{A}h(t) + \mathbf{B}u(t), \\ y(t) &= \mathbf{C}h(t), \end{aligned} \tag{1}$$

where $\mathbf{A} \in \mathbb{R}^{N \times N}$ is the state transition matrix, and $\mathbf{B}, \mathbf{C} \in \mathbb{R}^{N \times 1}$ are projection matrices. For use in deep learning models, these continuous parameters are discretized using a timestep Δ . A common discretization method is the zero-order hold (ZOH), which yields discrete parameters $\bar{\mathbf{A}}$ and $\bar{\mathbf{B}}$. Mamba introduces a key innovation: the *selective scan mechanism*, which makes the parameters \mathbf{B} , \mathbf{C} , and

Δ input-dependent functions of the input sequence $x \in \mathbb{R}^{T \times d}$. This allows the model to selectively focus on or ignore information at each timestep. The vanilla Mamba layer packages this SSM into a cohesive unit with input ($\text{Linear}_{\text{in}}$) and output ($\text{Linear}_{\text{out}}$) projections, a causal convolution, and gating mechanisms, making it a powerful and efficient replacement for the Transformer’s attention layer.

3.2. Overall Framework

As illustrated in Figure 2, our proposed AlignMamba-2 framework is designed to perform robust and efficient multimodal fusion. Using a trimodal scenario with audio, video, and language for illustration, the process begins by feeding raw data from each modality into its respective unimodal encoder to extract high-level feature sequences. These sequences are then processed by unimodal vanilla Mamba layers to model intra-modal contextual dependencies. Crucially, at the output of these unimodal Mamba layers, we apply our dual alignment strategy, which imposes both local geometric and global statistical constraints on the representations, prompting the alignment before fusion. Subsequently, the aligned feature sequences from all modalities are concatenated and fed into a stack of our Modality-Aware Mamba layers. These layers are specifically designed to handle heterogeneous data by jointly learning modality-specific and modality-invariant patterns, leading to a comprehensive fused representation. Finally, this representation is passed to a prediction head for the downstream task, such as sentiment analysis or scene classification.

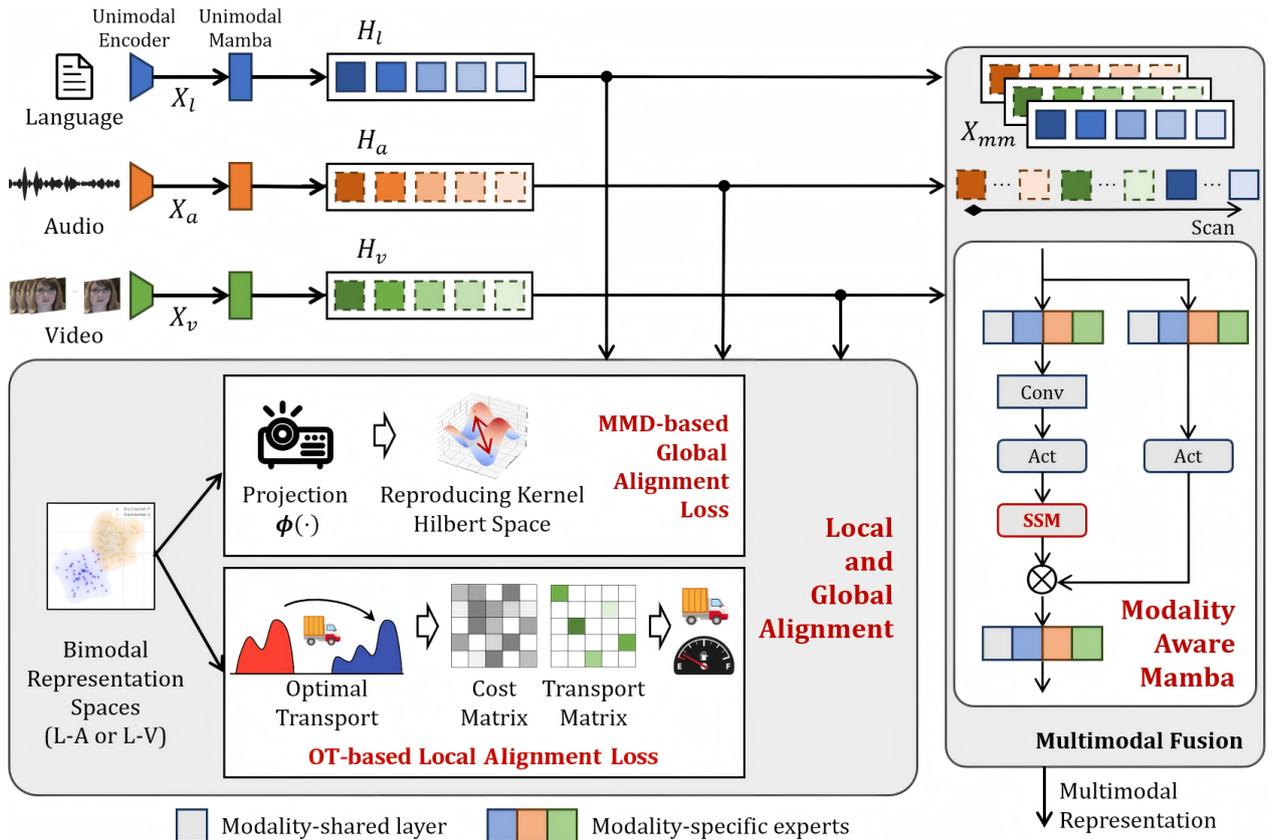


Figure 2 | The architecture of AlignMamba-2. The framework consists of three main stages: unimodal encoding, training-time dual alignment regularization (using OT and MMD), and a novel modality-aware Mamba fusion module based on a Mixture-of-Experts design.

3.3. Unimodal Encoding and Alignment

This stage aims to generate contextually-aware and well-aligned unimodal representations, laying a solid foundation for the subsequent fusion process.

Unimodal Encoding. Let the input feature sequences for audio, video, and language be denoted as $X_m \in \mathbb{R}^{T_m \times d_{in}}$ for $m \in \{A, V, L\}$. Each sequence is first processed by a shallow unimodal Mamba block to capture its internal temporal dependencies:

$$H_m = \text{Mamba}_m(X_m) \in \mathbb{R}^{T_m \times d}. \quad (2)$$

The resulting representations $\{H_A, H_V, H_L\}$ are then subjected to our dual alignment constraints. We use the language modality (H_L) as the anchor, aligning audio (H_A) and video (H_V) representations to it.

Local Alignment via OT Distance. To capture fine-grained, token-level correspondences, we use Optimal Transport (OT) distance (Villani, 2021) as a local alignment loss. Given two empirical distributions over token representations from modalities m and n , defined as $\mu = \frac{1}{T_m} \sum_{i=1}^{T_m} \delta_{h_{m,i}}$ and $\nu = \frac{1}{T_n} \sum_{j=1}^{T_n} \delta_{h_{n,j}}$, the OT distance is formulated as:

$$W(\mu, \nu) = \min_{\mathbf{P} \in \Pi(\mu, \nu)} \sum_{i=1}^{T_m} \sum_{j=1}^{T_n} P_{ij} C_{ij}, \quad (3)$$

where \mathbf{P} is the transport plan in the set of all valid plans $\Pi(\mu, \nu)$, and \mathbf{C} is the ground cost matrix. We define the cost C_{ij} as the cosine distance between tokens:

$$C_{ij} = 1 - \frac{h_{m,i} \cdot h_{n,j}}{\|h_{m,i}\| \|h_{n,j}\|}. \quad (4)$$

Solving Eq. 3 exactly is computationally expensive. Therefore, we use the Sinkhorn algorithm (Cuturi, 2013), which adds an entropic regularization term to make the problem efficiently solvable:

$$\mathcal{L}_{\text{OT}}(H_m, H_n) = \min_{\mathbf{P} \in \Pi(\mu, \nu)} (\langle \mathbf{P}, \mathbf{C} \rangle_F - \epsilon H(\mathbf{P})), \quad (5)$$

where $\epsilon > 0$ is the regularization strength and $H(\mathbf{P}) = -\sum_{i,j} P_{ij} \log(P_{ij})$ is the entropy of the plan. This loss penalizes misalignment by encouraging tokens with similar semantics (low cost) to be matched.

Global Alignment via MMD. To prompt consistency at a global, statistical level, we use the Maximum Mean Discrepancy (MMD) (Gretton et al., 2012). MMD measures the distance between two distributions as the squared norm of the difference between their mean embeddings in a Reproducing Kernel Hilbert Space (RKHS), \mathcal{H} . The empirical estimate of the squared MMD between representations H_m and H_n is:

$$\begin{aligned} \mathcal{L}_{\text{MMD}}(H_m, H_n) &= \left\| \frac{1}{T_m} \sum_{i=1}^{T_m} \phi(h_{m,i}) - \frac{1}{T_n} \sum_{j=1}^{T_n} \phi(h_{n,j}) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{T_m^2} \sum_{i,i'} k(h_{m,i}, h_{m,i'}) - \frac{2}{T_m T_n} \sum_{i,j} k(h_{m,i}, h_{n,j}) + \frac{1}{T_n^2} \sum_{j,j'} k(h_{n,j}, h_{n,j'}), \end{aligned} \quad (6)$$

where $\phi(\cdot)$ is the feature map to the RKHS and $k(\cdot, \cdot)$ is the associated kernel function. Here, we use the common Gaussian kernel:

$$k(x, y) = \exp(-\|x - y\|^2 / (2\sigma^2)). \quad (7)$$

This loss pulls the high-order statistical moments of the distributions closer, ensuring they are globally indistinguishable.

Our full alignment loss

$$\mathcal{L}_{\text{align}} = \lambda_{\text{OT}} \mathcal{L}_{\text{OT}} + \lambda_{\text{MMD}} \mathcal{L}_{\text{MMD}} \quad (8)$$

is applied to both the video-language and audio-language pairs. The two losses are complementary: OT enforces local geometric similarity, while MMD ensures global statistical consistency.

3.4. Modality-Aware Mamba for Fusion

A key limitation of vanilla Mamba in multimodal settings is its modality-agnostic nature. The shared projection and SSM layers process all tokens identically, regardless of their origin modality. To overcome this, we introduce the Modality-Aware Mamba layer, as depicted in Figure 3. Our motivation is to equip the Mamba block with the ability to learn both modality-specific features and shared cross-modal patterns simultaneously.

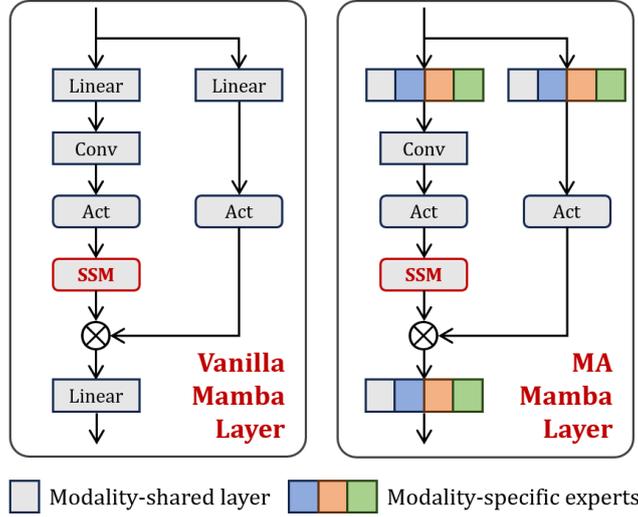


Figure 3 | Vanilla Mamba layer (left) and Modality-Aware Mamba layer (right).

We achieve this by replacing the standard linear input and output projection layers ($\text{Linear}_{\text{in}}$, $\text{Linear}_{\text{out}}$) with a Mixture-of-Experts (MoE) structure. For a concatenated input sequence $H_{\text{concat}} = [H_A; H_V; H_L]$, a token h_i belonging to modality m is routed deterministically based on its known origin.

Specifically, the MoE-based input projection consists of a set of modality-specific experts $\{E_{\text{in},m}\}_{m \in \{A,V,L\}}$ and one shared expert $E_{\text{in,shared}}$. All experts have an identical architecture. The projection for a token h_i from modality m is:

$$\text{MoE}_{\text{in}}(h_i) = E_{\text{in},m}(h_i) + E_{\text{in,shared}}(h_i). \quad (9)$$

Similarly, the output projection is also an MoE layer with experts $\{E_{\text{out},m}\}$ and $E_{\text{out,shared}}$:

$$\text{MoE}_{\text{out}}(h_i) = E_{\text{out},m}(h_i) + E_{\text{out,shared}}(h_i). \quad (10)$$

Crucially, the intermediate components, e.g., the 1D convolution, SiLU activation, and the selective SSM core, remain shared across all modalities. This design strikes a balance: the experts capture modality-specific transformations, while the shared core learns the universal dynamics of sequence modeling across all modalities. This enables a far more nuanced and effective fusion process than a standard Mamba block.

3.5. Training Objective and Algorithm

The entire AlignMamba-2 model is trained end-to-end. The final training objective is a sum of the task-specific loss and our alignment loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{align}}, \quad (11)$$

where $\mathcal{L}_{\text{task}}$ is the downstream loss (e.g., cross-entropy for classification, Mean Absolute Error for regression). The complete training procedure is summarized in Algorithm 1.

Algorithm 1 Training Process for AlignMamba-2

- 1: **Input:** Raw data for each modality $\{X_m^{\text{raw}}\}_{m \in \{A, V, L\}}$, labels Y .
 - 2: **Parameters:** Unimodal encoders $\{E_m\}$, unimodal Mamba $\{Mamba_m\}$, Modality-Aware Mamba $MAMamba$, prediction head C , weights $\lambda_{\text{OT}}, \lambda_{\text{MMD}}$.
 - 3:
 - 4: **for** each modality $m \in \{A, V, L\}$ **do**
 - 5: $X_m \leftarrow E_m(X_m^{\text{raw}})$ ▷ Extract unimodal features
 - 6: $H_m \leftarrow Mamba_m(X_m)$ ▷ Model intra-modal context
 - 7: **end for**
 - 8:
 - 9: $\mathcal{L}_{\text{OT}} \leftarrow \mathcal{L}_{\text{OT}}(H_A, H_L) + \mathcal{L}_{\text{OT}}(H_V, H_L)$ ▷ Compute total OT loss
 - 10: $\mathcal{L}_{\text{MMD}} \leftarrow \mathcal{L}_{\text{MMD}}(H_A, H_L) + \mathcal{L}_{\text{MMD}}(H_V, H_L)$ ▷ Compute total MMD loss
 - 11:
 - 12: $H_{\text{concat}} \leftarrow \text{Concatenate}(H_A, H_V, H_L)$ ▷ Concatenate unimodal features
 - 13: $Z \leftarrow MAMamba(H_{\text{concat}})$ ▷ Modality-aware fusion
 - 14: $\hat{Y} \leftarrow C(Z)$ ▷ Make prediction
 - 15:
 - 16: $\mathcal{L}_{\text{task}} \leftarrow \text{CrossEntropy} / \text{MAE}(Y, \hat{Y})$ ▷ Compute task-specific loss
 - 17: $\mathcal{L}_{\text{total}} \leftarrow \mathcal{L}_{\text{task}} + \lambda_{\text{OT}}\mathcal{L}_{\text{OT}} + \lambda_{\text{MMD}}\mathcal{L}_{\text{MMD}}$
 - 18: Perform backpropagation on $\mathcal{L}_{\text{total}}$.
-

4. Experiments

In this section, we conduct a comprehensive set of experiments to validate the effectiveness and efficiency of AlignMamba-2. We first introduce the datasets and implementation details (Section 4.1). We then present our main results on both dynamic multimodal tasks (Section 4.2) and static multimodal tasks (Section 4.3). Next, we provide an in-depth analysis of the model’s computational efficiency (Section 4.4), conduct ablation studies to dissect the contribution of each component (Section 4.5), and evaluate the model’s cross-dataset generalization capabilities (Section 4.6). Finally, we analyze the impact of key hyperparameters (Section 4.7).

4.1. Datasets and Implementation Details

Datasets To demonstrate the versatility and generalizability of our proposed AlignMamba-2, we evaluate it on four challenging multimodal fusion and sentiment analysis benchmarks spanning both dynamic (time-series) and static (image-based) settings.

- **CMU-MOSI (MOSI)** (Zadeh et al., 2016): A widely-used benchmark for multimodal sentiment analysis. It consists of 2,199 short video clips where aligned language, visual, and acoustic

streams are available. While the original annotations are on a scale from -3 to +3, we follow the standard evaluation protocol by framing the task as a binary classification (positive vs. negative).

- **CMU-MOSEI (MOSEI)** (Zadeh et al., 2018): A larger and more complex dataset for multimodal sentiment analysis, containing 22,856 video clips. MOSEI presents greater challenges due to its larger vocabulary, more diverse topics, and more complex emotional expressions. Similar to MOSI, we adhere to the standard binary classification setup (positive vs. negative).
- **NYU-Depth v2 (NYUDv2)** (Silberman et al., 2012): A popular indoor scene understanding dataset that provides tightly-coupled RGB and depth (D) modalities for each image. To ensure a fair comparison with prior works, we follow the evaluation setup from (Gao et al., 2024b) and frame the task as a 10-class scene recognition problem. This involves using 9 of the most common scene categories and grouping all remaining categories into a single "Others" class.
- **MVSA-Single (MVSA)** (Niu et al., 2016): A widely-used benchmark for image-text sentiment analysis. It contains over 4,800 image-text pairs sourced from Twitter, with each pair labeled as positive, negative, or neutral. A key characteristic of this dataset is the potential for semantic conflicts between the visual and textual modalities, posing a significant challenge for multimodal fusion methods.

Following previous work (Gao et al., 2024b; Li et al., 2025b), we adopt accuracy and F1 score as evaluation metrics on these four datasets.

Implementation Details To ensure a fair comparison, we adopt the experimental setup from previous methods (Gao et al., 2024b; Zadeh et al., 2016, 2018). Specifically, for the unimodal encoders on the MOSI and MOSEI datasets, we utilize the officially provided features extracted by FACET (for visual modality) and COVAREP (for audio modality) to maintain consistency with prior work. For the NYUDv2 and MVSA datasets, we employ a pre-trained ResNet as the encoder for both the RGB and depth modalities. For the text modality, we use the Hugging Face implementation of the Mamba model.

Our model is implemented using PyTorch. We use the Adam optimizer with an initial learning rate of $1e-3$ and employ a ReduceLROnPlateau learning rate scheduler. The gradient clipping parameter is set to 1.0. The number of layers for both the unimodal and modality-aware Mamba is selected from 1 to 3. To prevent overfitting, we apply a dropout rate of 0.2 and utilize an early stopping strategy.

Regarding the hyperparameter configurations for the dual alignment strategy, we specify the details as follows. For the MMD loss, we employ a Gaussian kernel, where the kernel bandwidth γ is set to the inverse of the input feature dimension d (i.e., $\gamma = 1/d$) to ensure scale invariance across different modalities. For the OT loss, we utilize the Sinkhorn algorithm implemented via the GeomLoss library. The entropic regularization strength ϵ (as denoted in Eq. 5) is controlled by setting the blur radius parameter to 0.05. Furthermore, instead of fixing a static number of iterations, we employ the library’s automatic epsilon-scaling mechanism with a tensorized backend. This adaptive strategy dynamically performs sufficient Sinkhorn iterations to guarantee numerical convergence and stability during training.

4.2. Results on MOSI and MOSEI

We present the results of AlignMamba-2 on the dynamic multimodal fusion benchmarks, CMU-MOSI and CMU-MOSEI, in Table 1. We compare our method against a wide range of state-of-the-art approaches, which can be broadly categorized into two groups: Transformer-based methods and the more recent Mamba-based methods.

Table 1 | Results on the MOSI and MOSEI datasets.

Method	Publication	MOSI		MOSEI	
		Acc	F1	Acc	F1
<i>Transformer-based Methods</i>					
MuT (Tsai et al., 2019)	ACL'19	84.1	83.9	82.5	82.3
ICCN (Sun et al., 2020)	AAAI'20	83.0	83.0	84.2	84.2
MISA (Hazarika et al., 2020)	MM'20	83.4	83.6	85.5	85.3
Self-MM (Yu et al., 2021)	AAAI'21	84.8	84.9	85.0	84.9
MMIM (Han et al., 2021a)	EMNLP'21	85.1	85.0	85.1	85.0
TokenFusion (Wang et al., 2022)	CVPR'22	84.8	84.7	85.0	84.9
HyCon (Mai et al., 2022)	TAFFC'22	85.2	85.1	85.4	85.6
ConFEDE (Yang et al., 2023)	ACL'23	85.5	85.5	85.8	85.8
MTMD (Lin and Hu, 2023)	TAFFC'23	86.0	86.0	86.1	85.9
MEA (Yang et al., 2024)	TCSVT'24	84.4	84.6	85.2	85.1
GeminiFusion (Jia et al., 2024)	ICML'24	85.7	85.8	86.1	85.8
EUAR (Gao et al., 2024a)	MM'24	86.3	86.3	86.6	<u>86.4</u>
DMD (Li et al., 2023a)	CVPR'24	85.8	85.8	86.0	86.1
MDKAT (Wang et al., 2025)	TCSVT'25	85.6	85.6	<u>86.5</u>	<u>86.4</u>
<i>Mamba-based Methods</i>					
CoupledMamba (Li et al., 2024a)	NeurIPS'24	-	-	85.7	85.6
Cobra (Zhao et al., 2025)	AAAI'25	82.3	82.1	81.8	81.4
Sigma (Wan et al., 2025)	WACV'25	86.3	86.3	86.1	86.2
AlignMamba-1 (Li et al., 2025b)	CVPR'25	<u>86.9</u>	<u>86.9</u>	86.6	86.5
AlignMamba-2	PR'26	87.0	87.0	<u>86.5</u>	86.5

Comparison with Transformer-based Methods The first group in the table comprises various Transformer-based fusion models, such as MuT (Tsai et al., 2019), MISA (Hazarika et al., 2020), and Self-MM (Yu et al., 2021), which leverage attention mechanisms to capture cross-modal interactions. More advanced methods like MMIM (Han et al., 2021a), ConFEDE (Yang et al., 2023), and MTMD (Lin and Hu, 2023) incorporate techniques like mutual information maximization or contrastive learning to enhance fusion quality. Compared to the Transformer-based MoE method (Gao et al., 2024a), our method achieves significant outperformance on the CMU-MOSI dataset and comparable results on the CMU-MOSEI dataset. This demonstrates the superiority of our new architecture, which replaces explicit alignment computation with a more efficient regularization scheme and introduces modality-awareness directly into the fusion core.

Comparison with Mamba-based Methods The second group of baselines includes recent methods that have also adopted Mamba for multimodal fusion, such as CoupledMamba (Li et al., 2024a), Cobra (Zhao et al., 2025), and Sigma (Wan et al., 2025). These models typically employ strategies like simple concatenation or co-scan mechanisms adapted for Mamba. While they benefit from Mamba’s computational efficiency, their performance on these challenging benchmarks is often limited, as seen with Cobra, or does not consistently outperform top-tier Transformer models. For instance, even strong Mamba-based contenders like Sigma achieve results (e.g., 86.3% Acc on MOSI) that fall short of our proposed method. This performance gap highlights a key insight: simply replacing Transformers with Mamba is insufficient. The architectural limitations of Mamba in handling cross-modal dependencies and multimodal fusion must be explicitly addressed.

Analysis of AlignMamba-2 AlignMamba-2 demonstrates state-of-the-art performance across both datasets. Notably, on the more challenging MOSEI dataset, it maintains a competitive F1 score of 86.5%, matching the performance of AlignMamba-1. The consistent high performance can be attributed to our unified framework. The OT and MMD alignment losses ensure that the features fed into the fusion module are already well-aligned from both geometric and statistical perspectives. More importantly, the novel Modality-Aware Mamba layer allows for a more nuanced fusion process by handling modality-specific characteristics and shared patterns. This combination proves to be highly effective for modeling the complex temporal dynamics inherent in multimodal fusion and sentiment analysis tasks.

4.3. Results on NYUDv2 and MVSA

To further assess the generalizability of our proposed framework beyond time-series data, we evaluate AlignMamba-2 on two static multimodal classification tasks: scene recognition on NYU-Depth V2 (RGB-D fusion) and sentiment classification on MVSA (Image-Text fusion). The results are summarized in Table 2. The results demonstrate that AlignMamba-2 consistently outperforms a variety of strong baselines across both datasets. On NYUDv2, it achieves an accuracy of 73.1%, surpassing recent state-of-the-art methods. Similarly, on the MVSA dataset, it reaches an accuracy of 82.7%, setting a new benchmark.

Table 2 | Results on the NYUDv2 and MVSA datasets.

Method	Publication	NYUDv2		MVSA	
		Acc	F1	Acc	F1
<i>General Fusion Methods</i>					
Late Fusion	-	69.1	68.3	76.9	75.7
TMC (Han et al., 2021b)	ICLR'21	71.1	69.8	76.1	74.6
TokenFusion (Wang et al., 2022)	CVPR'22	70.8	69.5	77.4	76.0
QMF (Zhang et al., 2023a)	ICML'23	70.1	68.7	78.1	77.2
GeminiFusion (Jia et al., 2024)	ICML'24	71.5	70.0	78.3	76.9
EAU (Gao et al., 2024b)	CVPR'24	72.1	70.6	79.2	78.4
EUAR (Gao et al., 2024a)	MM'24	71.7	70.7	79.6	78.0
MSFN (Zhang et al., 2025)	ToMM'25	-	-	79.0	<u>78.5</u>
<i>Mamba-based Methods</i>					
Cobra (Zhao et al., 2025)	AAAI'25	69.8	68.7	77.8	77.1
Sigma (Wan et al., 2025)	WACV'25	71.8	70.3	78.7	78.0
AlignMamba-1 (Li et al., 2025b)	CVPR'25	<u>72.4</u>	<u>70.9</u>	<u>81.1</u>	78.4
AlignMamba-2	PR'26	73.1	71.5	82.7	80.2

When comparing with Transformer-based fusion methods such as TokenFusion (Wang et al., 2022) and GeminiFusion (Jia et al., 2024), as well as uncertainty-aware models like EAU (Gao et al., 2024b), AlignMamba-2 shows a clear advantage. This indicates that our model’s effectiveness is not limited to temporal data but extends to tasks requiring the fusion of spatial (image) and semantic (depth, text) information. Compared to the Transformer-based MoE method (Gao et al., 2024a), our method also achieves significant outperformance on both datasets. The performance of Mamba-based counterparts like Cobra (Zhao et al., 2025) and Sigma (Wan et al., 2025), while superior to a simple Late Fusion baseline, does not reach the levels of top-performing methods. This again reinforces our core argument: a naive application of Mamba is insufficient for complex fusion tasks. Our previous

work, AlignMamba-1, achieves competitive results, but is surpassed by AlignMamba-2, highlighting the benefits of our refined alignment strategy and the novel modality-aware architecture.

In essence, the success on these static tasks validates that our dual alignment strategy and the Modality-Aware Mamba layer are general mechanisms for bridging the heterogeneity gap, regardless of whether the data has a temporal structure. The model effectively learns to align and fuse different data structures (patch-level features from images and depth maps, and token-level features from text), showcasing its robustness and wide applicability.

4.4. Computational Efficiency Analysis

To evaluate the computational efficiency of our proposed method, we conduct a comparative analysis of GPU memory consumption and inference latency. We benchmark AlignMamba-2 against our prior work, AlignMamba-1 (Li et al., 2025b), and a representative Transformer-based model, MulT (Tsai et al., 2019). To ensure a fair comparison, our analysis focuses exclusively on the fusion module, excluding the unimodal encoders. For all experiments, the batch size is set to 1, and we vary the input sequence length from 10k up to 100k. Each data point is the average of 10 independent runs to ensure stable and reliable measurements.

GPU Memory Usage Figure 4 illustrates the GPU memory usage of the three models. The Transformer-based MulT consumes approximately 10 GB of memory for a 10k sequence and encounters an out-of-memory error at 20k. This behavior starkly highlights the quadratic complexity of the self-attention mechanism, which leads to a dramatic increase in memory requirements as the sequence length grows.

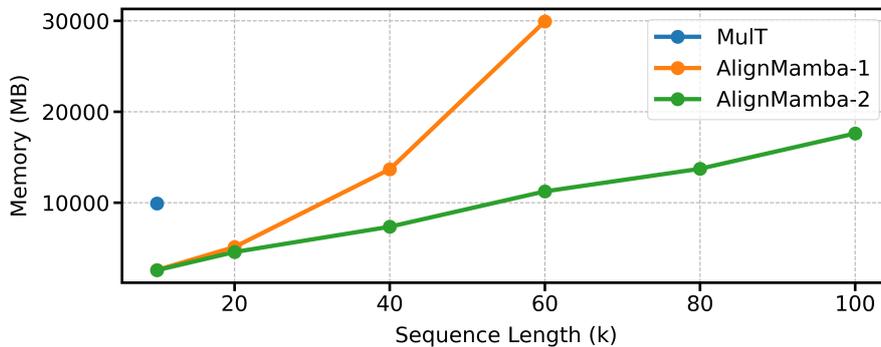


Figure 4 | GPU memory consumption (in MB) as a function of increasing sequence length. AlignMamba-2 demonstrates superior scalability compared to both Transformer-based and explicit-alignment-based Mamba models.

Both AlignMamba-1 and AlignMamba-2 exhibit comparable memory consumption at shorter sequence lengths (10k and 20k). However, as the length extends to 40k and 60k, the memory footprint of AlignMamba-1 increases significantly, culminating in an OOM error around 80k. This is because AlignMamba-1 explicitly computes the Optimal Transport (OT) matrix during inference, a process that incurs substantial resource costs for long sequences. In contrast, AlignMamba-2 leverages OT distance as a regularization loss *during training* and does not require this computation at inference time. Consequently, it maintains a linear and highly efficient memory profile, successfully processing sequences up to 100k and beyond.

Inference Time Figure 5 presents the comparison of inference latency. MulT requires nearly 100 milliseconds to process a 10k sequence, underscoring the severe latency bottleneck of Transformer-based fusion. AlignMamba-1 offers a substantial improvement, processing a 40k sequence in a similar

amount of time. AlignMamba-2, however, demonstrates even greater efficiency. It not only handles much longer sequences but also maintains a near-linear growth in processing time as the sequence length increases. This result shows AlignMamba-2 achieves remarkable gains in both speed and scalability, making it a highly practical solution for real-world, long-sequence multimodal applications.

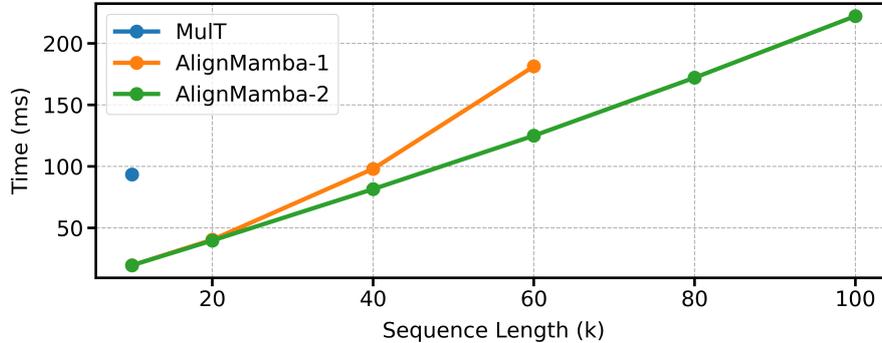


Figure 5 | Inference latency (in milliseconds) as a function of increasing sequence length. AlignMamba-2 maintains a linear increase in inference time, showcasing its efficiency for long-sequence processing.

4.5. Ablation Studies

To thoroughly investigate the individual contributions of the core components in AlignMamba-2, we conduct a series of ablation studies across all four datasets. We systematically remove or modify key parts of our model: the dual alignment loss, the Mixture-of-Experts mechanism in the fusion layers, and the deterministic routing strategy. The results are detailed in Table 3.

Table 3 | Ablation study of AlignMamba-2 on all four datasets. We report the performance (Accuracy and F1-Score) of the full model and its variants. "w/o" stands for "without", and "w/" stands for "with". Best results are in bold.

	MOSI		MOSEI		NYUDv2		MVSA	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1
AlignMamba-2	87.0	87.0	86.5	86.5	73.1	71.5	82.7	80.2
w/o dual alignment loss	84.5	84.4	84.1	84.0	71.0	69.5	79.2	78.1
w/o Mixture-of-Experts	85.8	85.7	85.3	85.2	72.0	70.4	80.5	78.2
w/o both	82.3	82.1	81.8	81.4	69.8	68.7	77.8	77.1
w/ Learnable Routing	86.6	86.4	86.1	86.2	72.8	71.2	82.2	79.5

Impact of the Dual Alignment Loss The first variant, "w/o dual alignment loss," removes both the OT and MMD regularization terms, meaning the unimodal features are directly concatenated and fed into the fusion module without any explicit alignment constraints. As shown in the table, this leads to the most significant performance degradation across all benchmarks. For instance, on CMU-MOSI, the F1 score drops by 2.6%, and on NYUDv2, it drops by 2.0%. This substantial decline underscores the critical importance of pre-fusion alignment. Without it, the Mamba-based fusion module struggles to bridge the large semantic and statistical gaps between heterogeneous modalities, validating our initial hypothesis that alignment is a prerequisite for effective fusion, especially within a sequential modeling framework like Mamba.

Impact of the Mixture-of-Experts Mechanism In the "w/o Mixture-of-Experts" setting, we replace our proposed Modality-Aware Mamba layers with standard, vanilla Mamba layers. In this configuration, the model still benefits from the dual alignment loss but loses the ability to process tokens in a modality-specific manner during fusion. The results show a consistent, smaller performance drop compared to removing the alignment loss. For example, the F1 score decreases by 1.3% on MOSI and 1.1% on NYUDv2. This demonstrates that while proper alignment is crucial, enabling the fusion backbone to explicitly model both modality-specific and modality-invariant information through the MoE architecture provides a significant additional benefit, leading to more nuanced and powerful representations. Removing both components ("w/o both"), which effectively reduces our model to a simple Mamba baseline, results in a drastic performance collapse, confirming that both alignment and awareness are indispensable.

Impact of Deterministic vs. Learnable Routing Finally, we explore the design of our MoE routing mechanism. The "Learnable Routing" variant replaces our deterministic, modality-based routing with a conventional learnable gating layer that dynamically decides which expert(s) to activate for each token. The results show that the performance on all datasets is slightly reduced, but still higher than that without the MoE architecture. This suggests that for multimodal fusion, the modality of a token is a powerful and reliable signal for routing. A learnable gate introduces additional complexity and potential optimization challenges, without providing a clear benefit over our simpler, more direct deterministic approach. For learnable routing without additional constraints, it is challenging to perform modality-specific modeling, which leads to multimodal representations that are insufficiently discriminative. This finding validates our design choice to leverage prior knowledge of token modality for efficient and effective expert selection.

4.6. Cross-Dataset Generalization Analysis

A robust multimodal fusion model should not only perform well within a specific data distribution but also demonstrate strong generalization capabilities to unseen, out-of-domain data. To evaluate this, we conduct a cross-dataset generalization experiment. Due to inconsistencies in the official visual feature dimensions between the CMU-MOSI and CMU-MOSEI datasets (Zadeh et al., 2016, 2018), we focus this analysis on a bimodal (audio-text) setup to ensure a fair and direct comparison. Specifically, we train our model and two representative baselines (Cobra (Zhao et al., 2025) and GeminiFusion (Jia et al., 2024)) on one dataset and test their performance on both the original (in-domain) and the other (out-of-domain) dataset in a zero-shot manner. The results of this rigorous evaluation are presented in Table 4.

Table 4 | Cross-dataset generalization performance on a bimodal (audio-text) setup. The table shows the in-domain and out-of-domain (zero-shot) performance for models trained on MOSI and MOSEI, respectively. Best results are in bold.

Training dataset	MOSI				MOSEI			
Test dataset	MOSI		MOSEI		MOSEI		MOSI	
Model	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Cobra	80.2	80.0	70.5	70.3	79.5	79.2	70.1	70.0
GeminiFusion	83.2	83.1	74.6	74.6	82.6	82.5	75.0	75.1
AlignMamba-1	83.8	83.8	75.6	76.8	83.2	83.1	76.4	76.3
AlignMamba-2	84.0	83.9	77.4	77.5	83.4	83.2	77.0	77.1

As expected, all models experience a significant performance drop when tested on an out-of-domain

dataset, which is attributable to the inherent domain shift in speaker characteristics, vocabulary, and emotional expression styles between MOSI and MOSEI. However, the key finding is the *relative* robustness of AlignMamba-2.

When trained on MOSI and tested on MOSEI, AlignMamba-2 achieves an F1 score of 77.5%, substantially outperforming GeminiFusion (74.6%) and Cobra (70.3%). A similar trend is observed in the reverse scenario (trained on MOSEI, tested on MOSI), where AlignMamba-2 again demonstrates superior performance. This indicates that our model learns more transferable and generalizable representations compared to the baselines.

Furthermore, while AlignMamba-2 achieves comparable in-domain performance to AlignMamba-1, it delivers superior results in out-of-domain settings, demonstrating enhanced cross-dataset generalization and robustness. We attribute this improvement to two key factors: First, unlike the explicit OT transport matrix computation in AlignMamba-1, the implicit OT loss regularization in AlignMamba-2 encourages the model to learn more robust, intrinsic representations. Second, the Modality-Aware MoE facilitates the disentanglement of modality-specific features, leading to improved transferability.

4.7. Hyperparameter Analysis

In this section, we investigate the impact of the key hyperparameters that govern our dual alignment strategy: the MMD loss weight, λ_{MMD} , and the OT loss weight, λ_{OT} . These weights control the strength of the alignment regularization relative to the main task objective. Figure 6 illustrates the model’s performance on the CMU-MOSI and CMU-MOSEI datasets as these two parameters are varied.

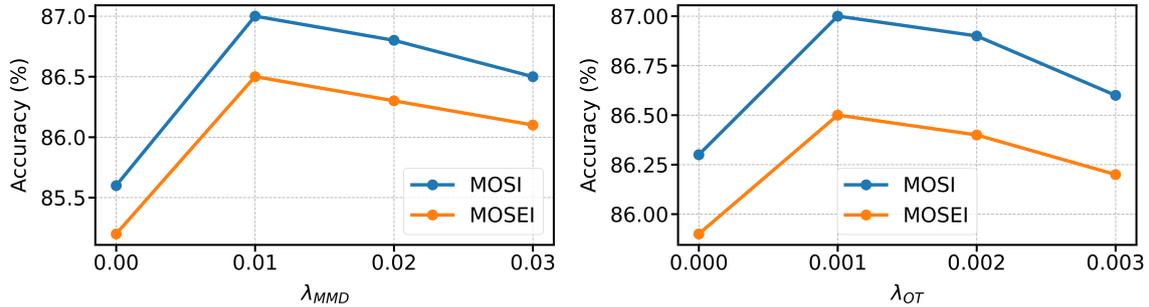


Figure 6 | Sensitivity analysis of model accuracy with respect to the MMD loss weight (λ_{MMD}) and the OT loss weight (λ_{OT}) on the MOSI and MOSEI datasets.

The left panel of the figure shows the sensitivity to λ_{MMD} while keeping λ_{OT} at its optimal value. When λ_{MMD} is set to zero, the model’s performance is sub-optimal, clearly indicating that the global statistical alignment provided by the MMD loss is beneficial. As the weight increases from zero, performance improves significantly. The model achieves its peak performance on both datasets at $\lambda_{MMD} = 0.01$, reaching a top accuracy of 87.0% on MOSI. Further increasing the weight beyond this point leads to a slight but consistent decline in performance. This trend suggests that while MMD alignment is crucial, overly strong regularization can start to interfere with the primary task objective.

A similar pattern is observed for the OT loss weight, λ_{OT} , as shown in the right panel, with λ_{MMD} held at its optimal setting. Performance without the OT loss is notably lower than the peak, confirming the positive contribution of the local geometric alignment. The model reaches the same optimal performance at a value of $\lambda_{OT} = 0.001$. As the weight increases past this point, we again observe a gentle degradation in accuracy, reinforcing the idea that balanced regularization is key to achieving the best results.

Overall, these results demonstrate the importance of both alignment losses for guiding the model

toward better fusion. They also show that while the model is sensitive to these hyperparameters, it is robust within a reasonable range around their optimal values. The fact that the optimal weights are relatively small suggests that the alignment objectives act as effective regularizers that steer the learning process without dominating the task-specific loss.

4.8. Case Study

To provide a more intuitive understanding of how AlignMamba-2 improves multimodal fusion, we conduct a qualitative case study on a representative sample from the CMU-MOSI dataset. As illustrated in Figure 7, we visualize the temporal evolution of sentiment in the video and text modalities, and compare the sentiment predictions of our model against two baselines: Cobra (Zhao et al., 2025) and our prior work, AlignMamba-1 (Li et al., 2025b).



Figure 7 | Qualitative analysis of a sample from the CMU-MOSI dataset.

The selected sample presents a classic case of inter-modal temporal misalignment. The visual modality (e.g., facial expressions) exhibits a neutral sentiment at the beginning before shifting to a sustained negative emotion for the remainder of the clip. In contrast, the textual modality remains neutral for the first half of the utterance and only introduces explicitly negative words towards the end. The ground-truth sentiment label for the entire sample is -0.8, indicating a weakly negative to neutral sentiment.

Cobra predicts a sentiment score of -2.1, indicating a strong negative emotion. Lacking an effective cross-modal alignment mechanism, it appears to over-emphasize the strong negative cues present at the end of each modality, failing to temper this with the initial neutral context. This results in a prediction that significantly deviates from the ground truth. AlignMamba-1, which incorporates an explicit OT-based alignment module, yields a more accurate prediction of -1.5. The alignment mechanism enables it to better correlate the temporal segments across modalities, leading to a more holistic understanding than Cobra. However, its prediction still leans more negative than the ground truth. In stark contrast, our proposed AlignMamba-2 delivers the most accurate prediction of -0.9. This superior result can be attributed to its two-fold architectural advancements. First, the dual alignment strategy (OT and MMD) ensures a more comprehensive alignment of the underlying feature distributions, allowing the model to correctly balance the influence of both the neutral and negative segments. Second, the Modality-Aware Mamba layer, with its specialized experts, can more effectively process the distinct temporal dynamics of visual and textual sentiment expression. By integrating these aligned and modality-aware insights, AlignMamba-2 achieves a nuanced fusion that closely mirrors the subtle, overall sentiment of the sample. This case study vividly demonstrates the practical benefits of our model’s design in handling complex, real-world multimodal interactions.

5. Conclusion

In this paper, we introduced AlignMamba-2, a novel framework for multimodal fusion and sentiment analysis. First, we proposed an efficient dual alignment strategy, using Optimal Transport distance and Maximum Mean Discrepancy as regularization to prompt comprehensive pre-fusion alignment without any inference cost. Second, we developed a novel Modality-Aware Mamba layer that leverages a Mixture-of-Experts design to explicitly model both modality-specific characteristics and shared cross-modal patterns. Through extensive experiments on a diverse set of dynamic and static multimodal benchmarks, we demonstrated that AlignMamba-2 consistently achieves state-of-the-art performance. Furthermore, our detailed efficiency analysis confirmed its significant advantages in memory consumption and inference speed for long-sequence tasks.

However, consistent with the "no free lunch" principle, the robust alignment capability comes with a trade-off. Specifically, calculating OT and MMD losses during the training phase introduces additional computational overhead compared to simple baseline methods. Our work not only presents a powerful solution for multimodal learning and sentiment analysis but also highlights a critical direction for future research: exploring linear-complexity alignment metrics or sparse mechanisms to achieve unified efficiency across both training and inference phases, ensuring discriminative multimodal representations with minimal computational cost.

6. Acknowledgement

This paper is supported by the China Postdoctoral Science Foundation (Grant No. 2025M781481) and the National Natural Science Foundation of China (Grant No. 62236006).

References

- H. Bao, W. Wang, L. Dong, Q. Liu, O. K. Mohammed, K. Aggarwal, S. Som, S. Piao, and F. Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in neural information processing systems*, 35:32897–32912, 2022.
- Z. Cao, Q. Xu, Z. Yang, Y. He, X. Cao, and Q. Huang. Otkge: Multi-modal knowledge graph embeddings via optimal transport. *Advances in neural information processing systems*, 35:39090–39102, 2022.
- M. Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.
- D. Dai, C. Deng, C. Zhao, R. Xu, H. Gao, D. Chen, J. Li, W. Zeng, X. Yu, Y. Wu, et al. Deepseekmoe: Towards ultimate expert specialization in mixture-of-experts language models. *arXiv preprint arXiv:2401.06066*, 2024a.
- Y. Dai, Y. Li, D. Chen, J. Li, and G. Lu. Multimodal decoupled distillation graph neural network for emotion recognition in conversation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(10):9910–9924, 2024b.
- W. Dong, H. Zhu, S. Lin, X. Luo, Y. Shen, X. Liu, J. Zhang, G. Guo, and B. Zhang. Fusion-mamba for cross-modality object detection. *arXiv preprint arXiv:2404.09146*, 2024.
- W. Fedus, B. Zoph, and N. Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

- C.-L. Gan, R.-S. Jia, H.-M. Sun, and Y.-C. Song. Multi-modal mamba framework for rgb-t crowd counting with linear complexity. *Pattern Recognition*, page 112522, 2025.
- Z. Gao, D. Hu, X. Jiang, H. Lu, H. T. Shen, and X. Xu. Enhanced experts with uncertainty-aware routing for multimodal sentiment analysis. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 9650–9659, 2024a.
- Z. Gao, X. Jiang, X. Xu, F. Shen, Y. Li, and H. T. Shen. Embracing unimodal aleatoric uncertainty for robust multimodal fusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26876–26885, 2024b.
- A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.
- A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- A. Gu, I. Johnson, K. Goel, K. Saab, T. Dao, A. Rudra, and C. Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. *Advances in neural information processing systems*, 34:572–585, 2021.
- W. Han, H. Chen, and S. Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *arXiv preprint arXiv:2109.00412*, 2021a.
- Z. Han, C. Zhang, H. Fu, and J. T. Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021b. URL <https://openreview.net/forum?id=00sR8BzCn15>.
- D. Hazarika, R. Zimmermann, and S. 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.
- X. He, K. Cao, K. Yan, R. Li, C. Xie, J. Zhang, and M. Zhou. Pan-mamba: Effective pan-sharpening with state space model. *arXiv preprint arXiv:2402.12192*, 2024.
- X. Huang and J. Xu. Multimodal spatiotemporal semisupervised transformer network for video-based group-level emotion recognition. *IEEE Transactions on Computational Social Systems*, 2025.
- D. Jia, J. Guo, K. Han, H. Wu, C. Zhang, C. Xu, and X. Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer. In *International Conference on Machine Learning*, pages 21753–21767. PMLR, 2024.
- W. Kim, B. Son, and I. Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR, 2021.
- D. Lepikhin, H. Lee, Y. Xu, D. Chen, O. Firat, Y. Huang, M. Krikun, N. Shazeer, and Z. Chen. Gshard: Scaling giant models with conditional computation and automatic sharding. In *International Conference on Learning Representations*, 2021.
- L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.
- W. Li, H. Zhou, J. Yu, Z. Song, and W. Yang. Coupled mamba: Enhanced multimodal fusion with coupled state space model. *Advances in Neural Information Processing Systems*, 37:59808–59832, 2024a.

- Y. Li, Y. Wang, and Z. Cui. Decoupled multimodal distilling for emotion recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6631–6640, 2023a.
- Y. Li, L. Zhang, X. Lan, and D. Jiang. Towards adaptable graph representation learning: An adaptive multi-graph contrastive transformer. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6063–6071, 2023b.
- Y. Li, W. Gan, K. Lu, D. Jiang, and R. Jain. Aves: An audio-visual emotion stream dataset for temporal emotion detection. *IEEE Transactions on Affective Computing*, 16(1):438–450, 2024b.
- Y. Li, X. Lan, H. Chen, K. Lu, and D. Jiang. Multimodal pear chain-of-thought reasoning for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(9):1–23, 2025a.
- Y. Li, Y. Xing, X. Lan, X. Li, H. Chen, and D. Jiang. Alignmamba: Enhancing multimodal mamba with local and global cross-modal alignment. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24774–24784, 2025b.
- Y. Li, Y. Zhao, X. Xia, and D. Jiang. Appearance-and relation-aware parallel graph attention fusion network for facial expression recognition. *IEEE Transactions on Affective Computing*, 2025c.
- R. Lin and H. Hu. Multi-task momentum distillation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 2023.
- H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023.
- S. Mai, Y. Zeng, S. Zheng, and H. Hu. Hybrid contrastive learning of tri-modal representation for multimodal sentiment analysis. *IEEE Transactions on Affective Computing*, 2022.
- T. Niu, S. Zhu, L. Pang, and A. El Saddik. Sentiment analysis on multi-view social data. In *International conference on multimedia modeling*, pages 15–27. Springer, 2016.
- Y. Qiao, Z. Yu, L. Guo, S. Chen, Z. Zhao, M. Sun, Q. Wu, and J. Liu. Vl-mamba: Exploring state space models for multimodal learning. *arXiv preprint arXiv:2403.13600*, 2024.
- N. Silberman, D. Hoiem, P. Kohli, and R. Fergus. Indoor segmentation and support inference from rgb-d images. In *European conference on computer vision*, pages 746–760. Springer, 2012.
- Z. Sun, P. Sarma, W. Sethares, and Y. 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*, volume 34, pages 8992–8999, 2020.
- Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, volume 2019, page 6558. NIH Public Access, 2019.
- A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- C. Villani. *Topics in optimal transportation*, volume 58. American Mathematical Soc., 2021.
- Z. Wan, P. Zhang, Y. Wang, S. Yong, S. Stepputtis, K. Sycara, and Y. Xie. Sigma: Siamese mamba network for multi-modal semantic segmentation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1734–1744. IEEE, 2025.

- J. Wang, C. Wang, L. Guo, S. Zhao, D. Wang, S. Zhang, X. Zhao, J. Yu, Y. Wang, Y. Yang, et al. Mdkat: Multimodal decoupling with knowledge aggregation and transfer for video emotion recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.
- Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang. Multimodal token fusion for vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12186–12195, 2022.
- L. Xiao, X. Yang, F. Peng, Y. Wang, and C. Xu. Oneref: Unified one-tower expression grounding and segmentation with mask referring modeling. *Advances in Neural Information Processing Systems*, 37: 139854–139885, 2024.
- L. Xiao, X. Yang, X. Lan, Y. Wang, and C. Xu. Towards visual grounding: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- M. Xu, X. Chen, B. Liu, Y.-R. Lin, Y.-H. Li, and J. Wang. A unified experience replay framework for spiking deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.
- M. Xu, Z. Wen, X. Chen, G. Zhao, J. Huang, and J. Wang. A generic competitive-cooperative actor-critic framework for deep reinforcement learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.
- D. Yang, M. Li, L. Qu, K. Yang, P. Zhai, S. Wang, and L. Zhang. Asynchronous multimodal video sequence fusion via learning modality-exclusive and-agnostic representations. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.
- J. Yang, Y. Yu, D. Niu, W. Guo, and Y. 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.
- W. Yu, H. Xu, Z. Yuan, and J. Wu. Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis. In *Proceedings of the AAI conference on artificial intelligence*, volume 35, pages 10790–10797, 2021.
- A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.
- A. B. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L.-P. 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.
- Q. Zhang, H. Wu, C. Zhang, Q. Hu, H. Fu, J. T. Zhou, and X. Peng. Provable dynamic fusion for low-quality multimodal data. In *International conference on machine learning*, pages 41753–41769. PMLR, 2023a.
- S. Zhang, J. Liu, Y. Jiao, Y. Zhang, L. Chen, and K. Li. A multimodal semantic fusion network with cross-modal alignment for multimodal sentiment analysis. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2025.
- Y. Zhang, M. Yin, H. Wang, and C. Hua. Cross-level multi-modal features learning with transformer for rgb-d object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 33 (12):7121–7130, 2023b.

- H. Zhao, M. Zhang, W. Zhao, P. Ding, S. Huang, and D. Wang. Cobra: Extending mamba to multi-modal large language model for efficient inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 10421–10429, 2025.
- J. Zheng, S. Zhang, Z. Wang, X. Wang, and Z. Zeng. Multi-channel weight-sharing autoencoder based on cascade multi-head attention for multimodal emotion recognition. *IEEE Transactions on Multimedia*, 25:2213–2225, 2022.