

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

利用局部和全局跨模态对齐增强多模态Mamba

汇报人：方文熙

2025年3月11日

问题背景：多模态表示融合（如音频、视频、文本）是跨模态信息整合和理解的关键技术，但由于不同模态之间的异质性（如统计特性和特征分布的差异），实现有效的跨模态对齐和融合仍然是一个重大挑战。

现有方法的局限性：基于Transformer的方法虽然有效，但其二次计算复杂度使其在处理长序列或大规模数据时效率低下。基于Mamba的方法虽然计算效率高，但由于其顺序扫描机制，难以全面建模跨模态关系。

解决方案：提出了**AlignMamba**，通过引入局部和全局跨模态对齐来增强多模态融合：

- **局部对齐：**基于最优传输（Optimal Transport, OT）显式学习不同模态之间的细粒度对应关系。
- **全局对齐：**基于最大均值差异（Maximum Mean Discrepancy, MMD）隐式地对齐不同模态的分布。

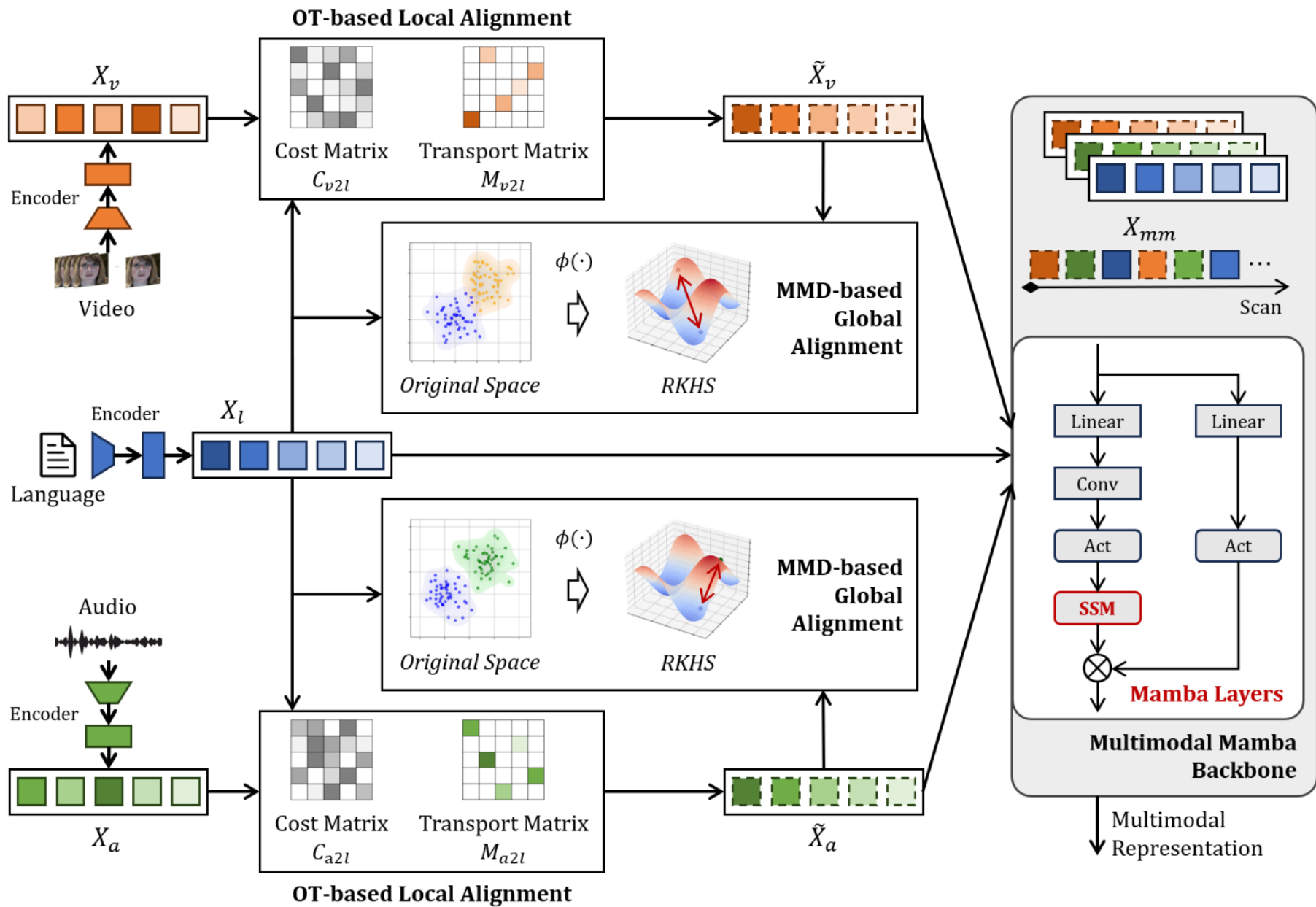
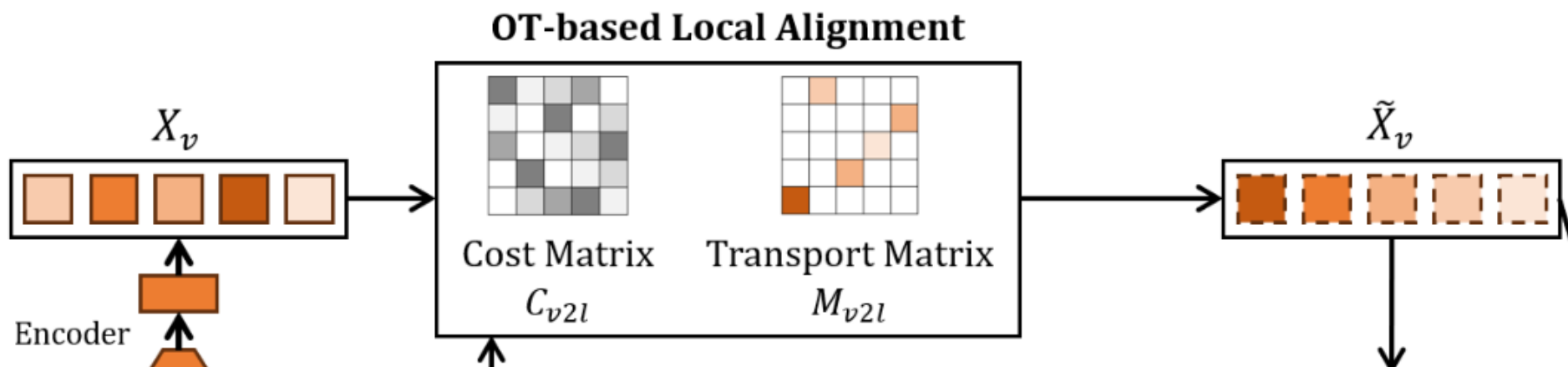


Figure 2. AlignMamba enhances multimodal Mamba by incorporating token-level alignment and distribution-level alignment, enabling more effective multimodal fusion.

• 局部对齐（基于OT）：

- 将特征序列视为离散分布，通过学习传输矩阵来对齐不同模态的特征。
- 使用余弦距离作为成本矩阵，最小化特征传输成本。
- 采用松弛的OT公式以降低计算复杂度。



$$X_a \in \mathbb{R}^{T_a \times d}, X_v \in \mathbb{R}^{T_v \times d}, X_l \in \mathbb{R}^{T_l \times d}$$

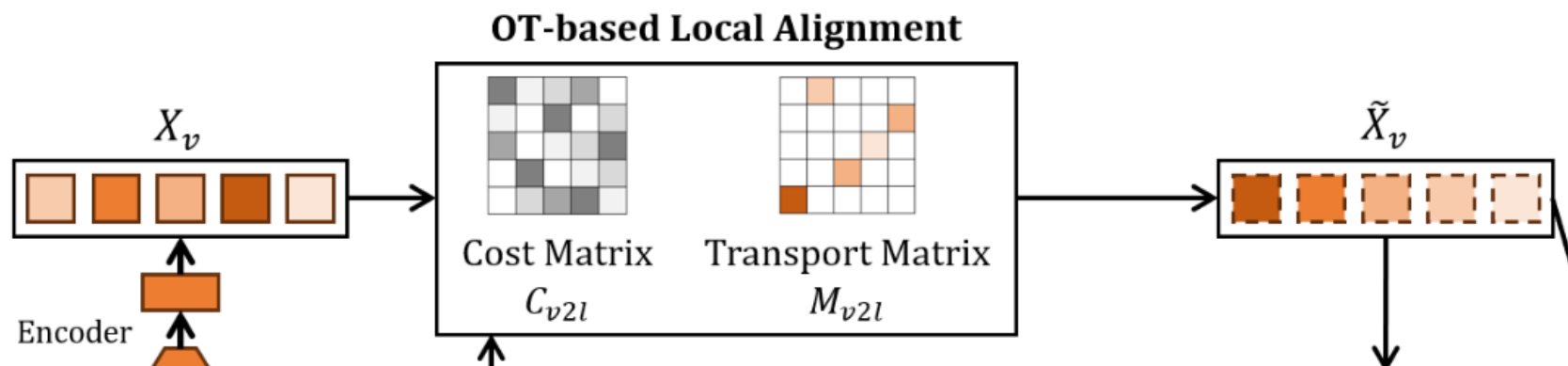
audio

video

language

T_a, T_v, T_l 表示不同模态的序列长度

d : 特征维度



在这个模块，我们需要做局部特征融合，局部对齐（基于OT），以音频和视频空间的信号，跟语言空间信号的序列长度对齐为例，
通过以下最优化问题：

通过此最优化矩阵，求出M
矩阵（传输矩阵），M (i, j)
表示视频模态时刻i（第i行）
对语言模态时刻j（第j行）
的权重

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

其中C为开销矩阵，对应元素
相乘，和为0，即总开销最小

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

行归一，多对一
列归一，一对多

权重必须大于0

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

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

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

开销矩阵，初始化为 1 - 余弦相似度

解决以上优化问题太复杂了，可用更宽松的OT，减少一个约束条件，允许一对多

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

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

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

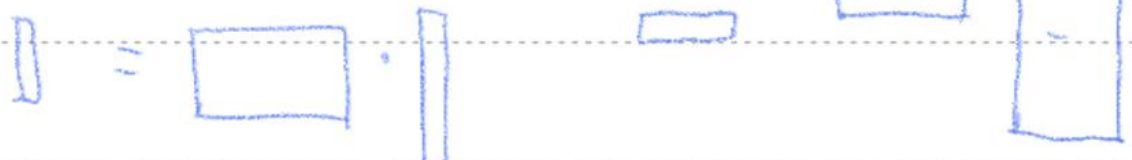
M的最优解，每行只有一个非0

简化版的OT，非常好解，显然，最优解为：
$$M_{v2l}(i, j) = \begin{cases} \frac{1}{T_v}, & j = \arg \min_j C_{v2l}(i, j), \\ 0, & j \neq \arg \min_j C_{v2l}(i, j). \end{cases}$$

M (i, j) 表示，video空间，i时刻，对，language空间，j时刻的影响权重，
此最优解表示，对video空间i时刻，将余弦相似度最高的，即最相似的，权重设置为非0，其余时刻均设置为0

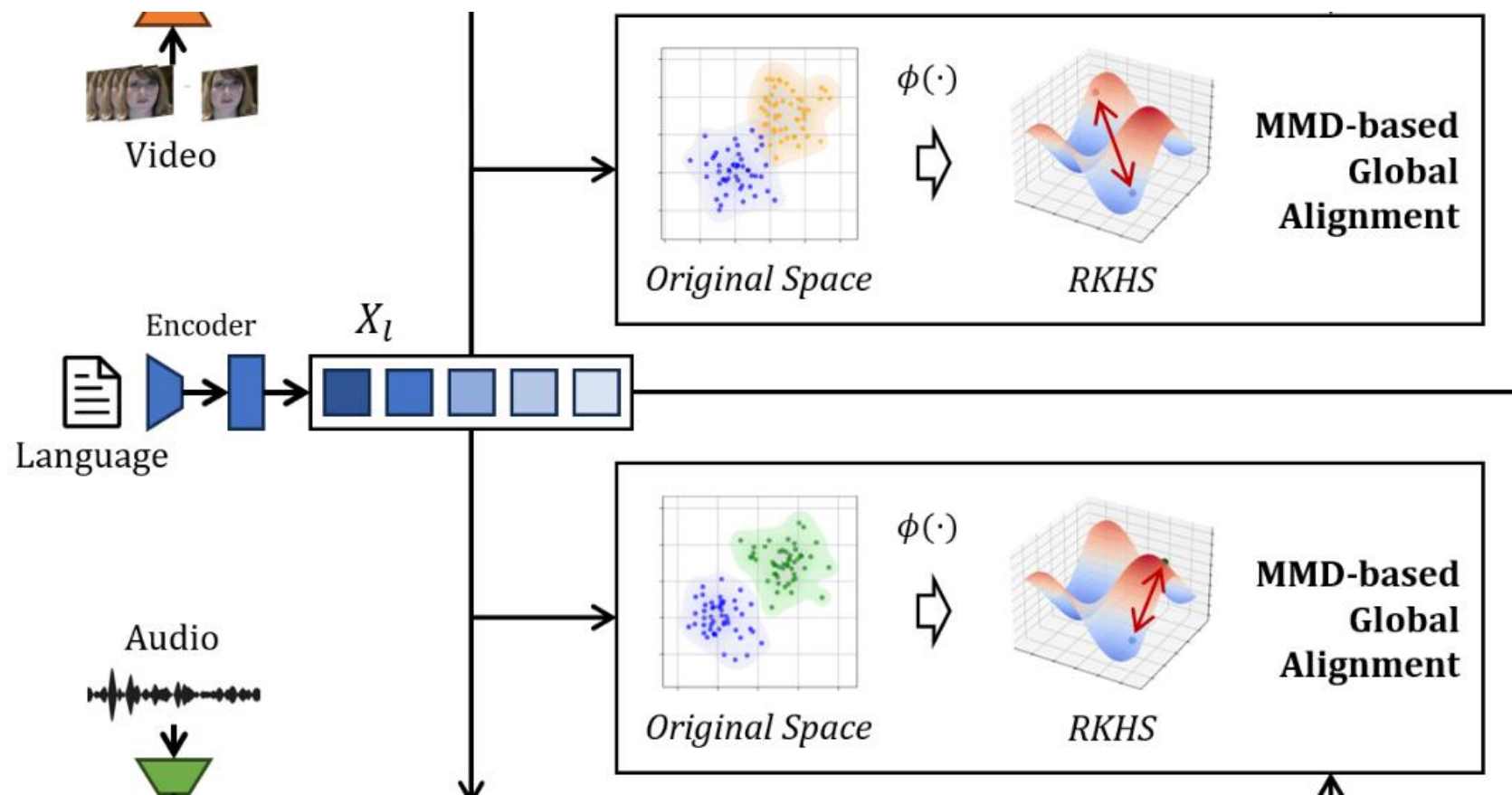
同理，可求出音频空间对语言空间的传输矩阵，根据传输矩阵，将音频空间和视频空间都映射到语言空间中

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

$$\hat{X}_v = M_{v2l}^T X_v \Leftrightarrow \hat{X}_v^T = X_v^T \cdot M_{v2l}$$


•全局对齐（基于MMD）：

- 在高维再生核希尔伯特空间（RKHS）中测量不同模态之间的统计差异。
- 使用高斯核计算MMD距离，确保模态分布的一致性。



在 Hilbert 空间中定义距离, MMD^2

104041

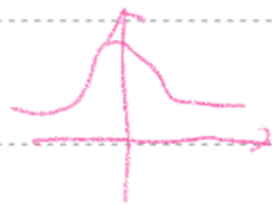
$$MMD^2(X, Y) \triangleq \left\| \frac{1}{I} \sum_{i=1}^I \phi(x_i) - \frac{1}{I} \sum_{j=1}^I \phi(y_j) \right\|_H^2$$

$$x \xrightarrow{\phi} \sum \phi(x_i)$$

利用高斯核计算

$$MMD^2(X, Y) \triangleq \frac{1}{I^2} \left(\sum_{i=1}^I \sum_{i'=1}^I k(x_i, x_{i'}) + \sum_{j=1}^I \sum_{j'=1}^I k(y_j, y_{j'}) - 2 \sum_{i=1}^I \sum_{j=1}^I k(x_i, y_j) \right)$$

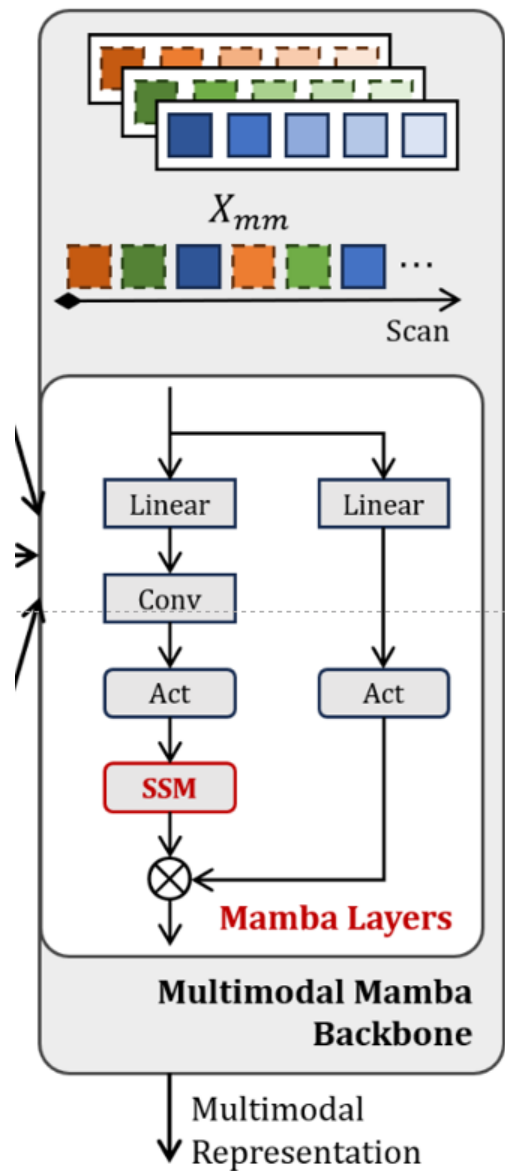
$$k(x, y) \triangleq \exp\left(-\frac{\|x-y\|_2^2}{2\sigma^2}\right)$$



x, y 距离越小, $k(x, y)$ 值越大

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

$MMD^2(X, Y)$ 希望同模态 不同时刻 距离大
不同模态 距离小



Input
(sequence)



$\mathbf{x}(t)$

State Space Model
(SSM)

Output
(sequence)



$\mathbf{y}(t)$

$$\mathbf{h}'(t) = \mathbf{A}\mathbf{h}(t) + \mathbf{B}\mathbf{x}(t)$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{h}(t) + \mathbf{D}\mathbf{x}(t)$$

$$h_t = \bar{\mathbf{A}}h_{t-1} + \bar{\mathbf{B}}x_t$$

$$y_t = \mathbf{C}h_t$$

SSM

$$h_t = \overline{A}h_{t-1} + \overline{B}x_t$$

$$y_t = Ch_t$$

Mamba-based Fusion and Optimization 基于Mamba的融合优化

将 $\tilde{X}_a, \tilde{X}_v, X_l$ 拼成一个新的序列

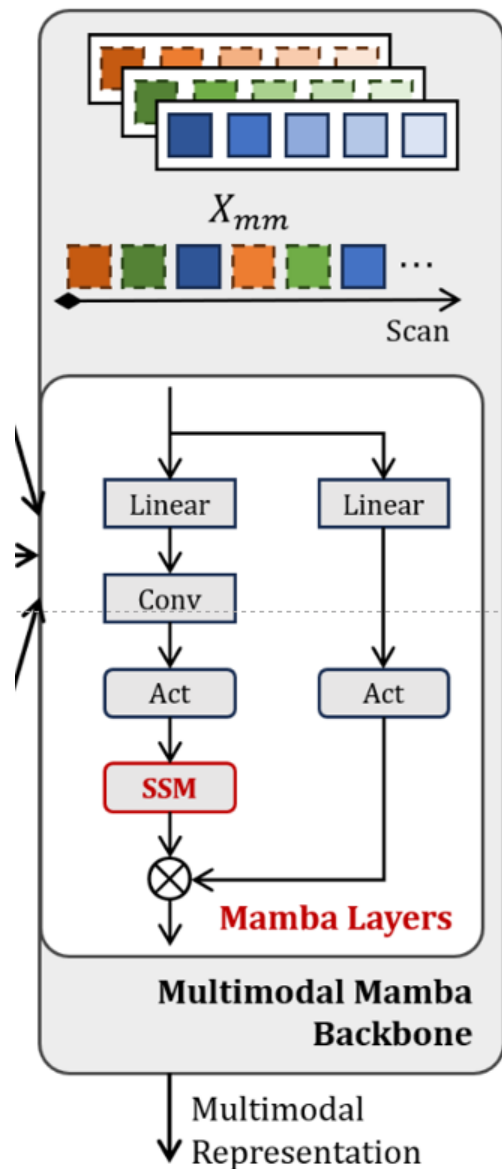
$$X_{mm} = [\tilde{X}_a^1, \tilde{X}_v^1, X_l^1, \tilde{X}_a^2, \tilde{X}_v^2, X_l^2, \dots, \tilde{X}_a^T, \tilde{X}_v^T, X_l^T]$$

上标为时间索引

时间优先级扫描

$$\mathcal{L} = \mathcal{L}_{align} + \mathcal{L}_{task}$$

下游任务的损失



实验部分

- **数据集：**CMU-MOSI和CMU-MOSEI，包含带有视觉、声学 and 文本模态的视频片段，并标注了情感分数。
- **任务：**完整多模态融合（所有模态可用）和不完整多模态融合（推理时某些模态缺失）。
- **结果：**
 - **完整融合：**AlignMamba在CMU-MOSI数据集上实现了最先进的性能，分类准确率提高了0.9%。
 - **不完整融合：**AlignMamba在模态缺失的情况下表现出更强的鲁棒性，优于现有方法。
- **效率：**与基于Transformer的方法相比，AlignMamba显著减少了GPU内存使用（减少20.3%）和推理时间（减少83.3%）。

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 / F₁ (%). Δ : performance drop from 10% to 70% missing rate (lower is better).

	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 (%).

OT 局部对齐模块 和 MMD 全局对齐损失 都对模型性能有显著贡献，其中 OT 局部对齐模块的影响更大。

单纯的 Mamba 架构（单流或多流）无法有效实现多模态融合，凸显了跨模态对齐的必要性。

语言模态 对模型性能的影响最大，其次是视频模态和音频模态。

时间和空间的开销

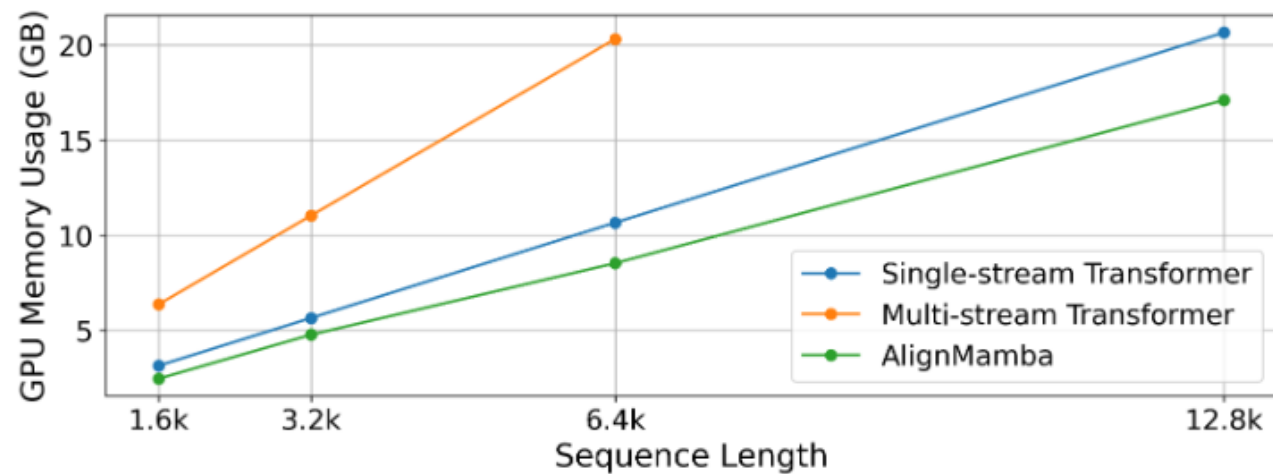


Figure 3. GPU memory usage comparison with varying lengths.

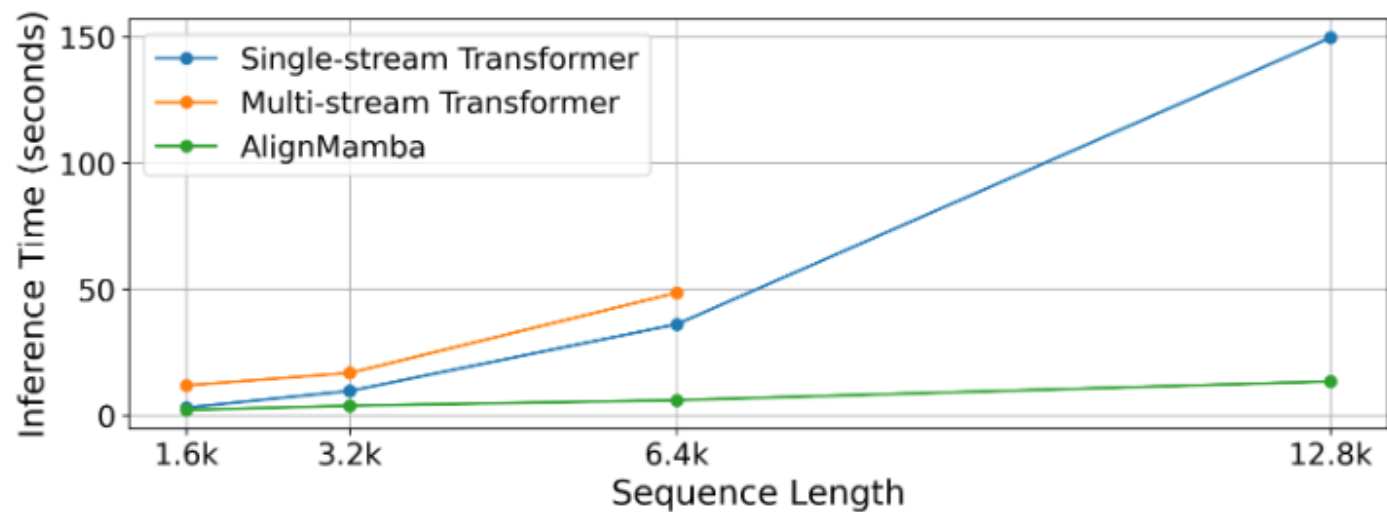


Figure 4. Inference time comparison with varying lengths.

END