





# **Does Multimodality Lead to Better Time Series Forecasting?**



2025/08/24

# I 引言

## 多模态时间序列预测(MMTS)

- 定义：结合时间序列数值（如气温）与辅助文本信息（如天气预报）以提升预测精度。

## 核心普遍的假设 vs. 现实的差距

- 普遍假设：更多的信息维度（文本+数值）理应带来更好的预测结果。
- 现实差距：这种增益是否普适？在何种条件下才能实现？目前缺乏系统性的答案。

## 核心研究问题：

- “多模态是否真的能提升预测性能？如果能，在何时、如何实现？”

## 本文贡献：

- 首次进行大规模、系统性的基准测试，从模型和数据两个维度深入剖析MMTS的有效性。



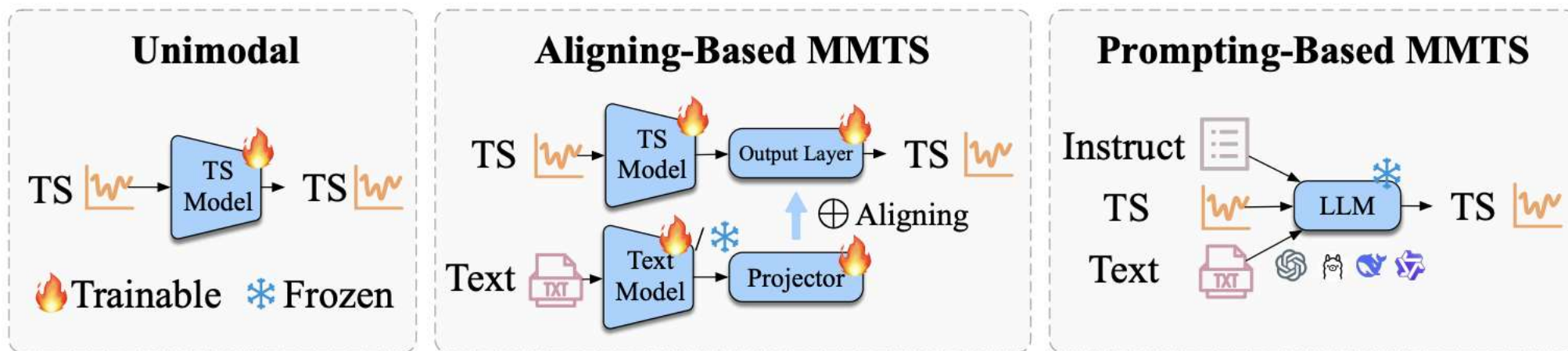
## II 相关工作

### 基于对齐的方法 (Aligning-Based Methods):

- 核心思想： 为不同模态设计专门的编码器，然后对齐它们的数学表征。
- 流程： 分别编码  $\rightarrow$  对齐/融合向量  $\rightarrow$  解码预测。
- 特点： 结构灵活，但设计复杂，对齐机制是关键挑战。

### 基于提示的方法 (Prompting-Based Methods):

- 核心思想： 利用大型语言模型 (LLM) 的预训练知识和推理能力。
- 流程： 将所有信息（数值+文本）格式化为自然语言提示  $\rightarrow$  输入LLM  $\rightarrow$  解析输出。
- 特点： 部署简单，但依赖LLM的数值处理能力。



# III MMTS 公式化与实验设置

## 任务目标

- 给定历史时间序列  $\mathbf{z}_{1:T} \in \mathbb{R}^{T \times d}$   $\mathbf{x} \in \mathcal{X}$  信息  $\mathbf{z}_{T+1:T+\tau} \in \mathbb{R}^{\tau \times d}$
- 形式化为:  $p(\mathbf{z}_{T+1:T+\tau} \mid \mathbf{z}_{1:T}, \mathbf{x}) = f(\mathbf{z}_{1:T}, \mathbf{x}), f: (\mathbb{R}^{T \times d}, \mathcal{X}) \rightarrow \mathcal{P}(\mathbb{R}^{\tau \times d})$

## 基于对齐方法的公式化:

- $p(\mathbf{z}_{T+1:T+\tau} \mid \mathbf{z}_{1:T}, \mathbf{x}) = g(\psi(\phi_z(\mathbf{z}_{1:T}), \phi_x(\mathbf{x})))$ .
- 关键组件: 时间序列编码器 ( $\phi_z$ ), 文本编码器 ( $\phi_x$ ), 融合模块 ( $\psi$ ), 预测器 ( $g$ ).

## 基于提示方法的公式化:

- $p(\mathbf{z}_{T+1:T+\tau} \mid \mathbf{z}_{1:T}, \mathbf{x}) = l(\pi(\mathbf{z}_{1:T}, \mathbf{x}))$ .
- 关键组件: 格式化器/提示模板 ( $\pi$ ), 大型语言模型 ( $l$ )

# III MMTS 公式化与实验设置

实验设置:

- 建模方法:
  - 对齐方法 (Aligning-Based Methods): 评估了多种时间序列模型 (如PatchTST、DLinear、FEDformer、Informer、iTransformer、Chronos) 与不同文本模型 (如BERT、GPT-2、T5、Qwen-1.5B、LLaMA-7B) 的组合。
  - 提示方法 (Prompting-Based Methods): 评估了多种LLMs (如LLaMA、Qwen、Mistral、GPT-4o-mini、Claude) 。
- 数据集: 14个真实世界的数据集, 覆盖7个不同的领域, 包括健康、环境、能源、经济等。这些数据集包括动态文本设置 (每时间步都有文本输入) 和静态文本设置 (整个时间序列关联一个文本描述)
- 评估指标: 使用均方误差 (MSE) 和平均绝对误差 (MAE) 评估点预测性能

Dataset Name	Type	Train Size	Val Size	Test Size	Context Length	Prediction Length
Agriculture [13]	Dynamic	318	45	94	24	6
Climate [13]	Dynamic	318	45	94	24	6
Economy [13]	Dynamic	267	38	79	24	6
Energy [13]	Dynamic	1000	138	284	24	12
Environment [13]	Dynamic	7700	1064	2173	24	48
Health [13]	Dynamic	937	129	266	24	12
Socialgood [13]	Dynamic	601	85	175	24	6
Traffic [13]	Dynamic	342	49	101	24	6
Fashion [25]	Static	3081	513	1983	1	11
Weather [12]	Dynamic	2475	412	361	7	7
Medical [12]	Dynamic	4575	762	706	7	7
PTF [28]	Static	7927	1321	2272	120	24
MSPG [28]	Static	7244	1207	2106	480	96
LEU [28]	Static	7968	1328	2320	240	48

## IV 实证结果分析



# IV - MMTS 模型是否始终有效？

## 实验结果

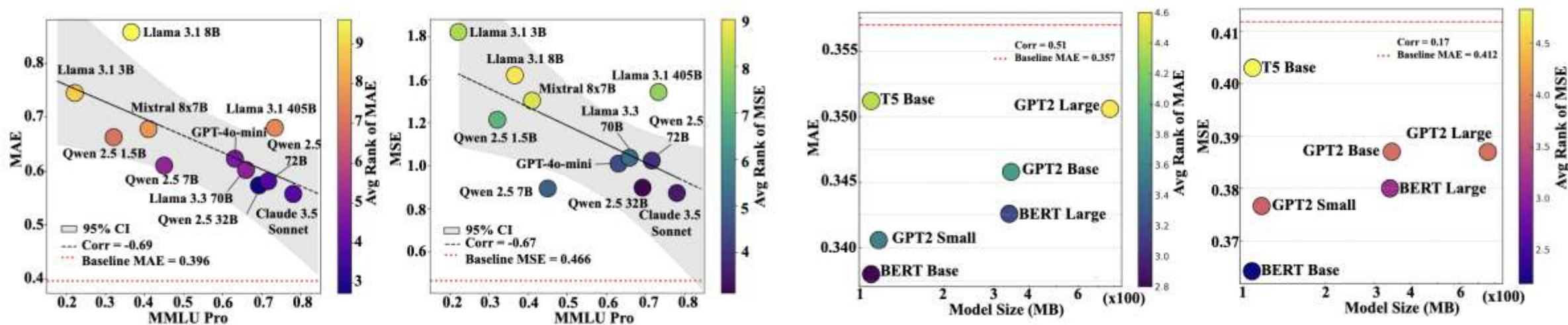
- 结论：多模态方法并不总是优于单模态方法，其效果高度依赖于模型架构、对齐策略和数据特性。
- 数据支撑：在14个数据集中，单模态模型在超过一半的情况下获得了最佳MSE或MAE。
- 统计显著性：对比最强的单模态模型（Chronos）和对齐驱动模型，性能差异在统计上不显著
- 对齐驱动模型普遍优于提示驱动模型，这可能源于当前LLM在精确数值推理上的局限性。

Dataset	Metric	Unimodal			Aligning				Prompting					
		PatchTST	DLinear	Chronos	PatchTST	DLinear	Chronos	Time-LLM	LLaMA-405B-Inst	Qwen-72B-Inst	Qwen-72B-Inst	Mistral 8x7B-Inst	GPT-4o (mini)	Claude 3.5
Agriculture [13]	MSE	0.205	0.664	0.161	0.219	0.652	0.167	0.249	0.152	<u>0.118</u>	<b>0.113</b>	0.153	0.123	0.137
	MAE	0.306	0.577	0.258	0.323	0.564	0.299	0.336	0.269	<u>0.229</u>	<b>0.227</b>	0.253	0.243	0.244
Climate [13]	MSE	1.498	<u>0.961</u>	1.200	1.521	<b>0.955</b>	1.077	1.127	1.913	1.468	1.265	1.786	1.519	1.733
	MAE	0.978	<u>0.786</u>	0.876	0.985	<b>0.783</b>	0.830	0.858	1.100	0.983	0.900	1.066	1.000	1.051
Economy [13]	MSE	<b>0.014</b>	0.161	0.061	0.019	0.137	0.153	<u>0.016</u>	0.053	0.036	0.053	0.057	0.039	0.040
	MAE	<b>0.093</b>	0.348	0.200	0.105	0.330	0.344	<u>0.103</u>	0.179	0.153	0.184	0.186	0.162	0.159
Energy [13]	MSE	0.130	<b>0.096</b>	0.112	0.130	<u>0.096</u>	0.138	0.102	0.131	0.125	0.139	0.144	0.137	0.131
	MAE	0.259	<b>0.221</b>	0.240	0.262	<u>0.221</u>	0.260	<b>0.221</b>	0.240	0.232	0.245	0.247	0.242	0.240
Environment [13]	MSE	0.558	0.475	<u>0.466</u>	0.557	0.474	<b>0.422</b>	0.490	2.895	1.112	1.256	1.468	0.641	0.710
	MAE	0.543	0.531	<b>0.486</b>	0.543	0.530	<u>0.488</u>	0.499	0.913	0.699	0.788	0.807	0.568	0.592
Health [13]	MSE	1.995	1.442	<b>1.169</b>	1.769	1.455	1.370	<u>1.271</u>	3.300	2.536	3.427	4.326	3.897	3.338
	MAE	0.792	0.809	<b>0.676</b>	0.781	0.818	0.833	<u>0.734</u>	1.096	0.947	1.045	1.100	1.218	1.218
Social Good [13]	MSE	0.924	1.045	0.914	0.924	1.028	0.957	1.019	1.078	<u>0.810</u>	<b>0.678</b>	0.866	0.908	0.788
	MAE	<u>0.404</u>	0.441	0.453	0.408	0.437	0.460	0.439	0.533	0.431	<b>0.378</b>	0.461	0.500	0.416
Traffic [13]	MSE	<b>0.144</b>	0.154	0.182	<u>0.147</u>	0.153	0.180	0.162	0.419	0.262	0.275	0.359	0.311	0.232
	MAE	0.202	<b>0.200</b>	0.259	0.207	<u>0.201</u>	0.244	0.237	0.477	0.360	0.362	0.458	0.408	0.278
Fashion [25]	MSE	0.525	<b>0.473</b>	0.485	0.524	<u>0.473</u>	0.482	0.513	0.607	0.566	0.539	0.642	0.552	0.598
	MAE	0.524	0.515	<b>0.468</b>	0.523	0.515	<u>0.472</u>	0.511	0.508	0.487	0.486	0.550	0.480	0.507
Weather [12]	MSE	0.180	<b>0.169</b>	0.184	0.182	<u>0.169</u>	0.248	0.172	0.289	0.241	0.253	0.295	0.227	0.257
	MAE	0.321	<b>0.314</b>	0.325	0.326	<u>0.314</u>	0.379	0.315	0.397	0.360	0.373	0.403	0.358	0.374
Medical [12]	MSE	0.539	0.694	0.596	<u>0.505</u>	0.694	<b>0.497</b>	0.527	1.083	0.712	0.688	0.765	0.642	0.718
	MAE	0.513	0.624	0.565	<u>0.506</u>	0.623	<b>0.503</b>	0.513	0.603	0.572	0.571	0.601	0.554	0.575
PTF [28]	MSE	0.099	0.134	0.110	0.098	0.120	<u>0.089</u>	<b>0.088</b>	0.886	0.519	0.618	1.305	0.738	0.287
	MAE	0.172	0.225	<u>0.167</u>	0.171	0.214	<b>0.162</b>	0.189	0.670	0.479	0.538	0.808	0.647	0.323
MSPG [28]	MSE	<u>0.328</u>	0.428	0.365	0.365	0.420	<b>0.300</b>	0.353	2.635	1.165	2.731	2.879	1.488	0.873
	MAE	0.230	0.234	<u>0.217</u>	0.237	0.233	<b>0.198</b>	0.244	0.535	0.416	0.717	0.610	0.539	0.425
LEU [28]	MSE	0.573	0.506	0.522	0.565	<u>0.504</u>	0.513	<b>0.490</b>	1.543	0.822	0.811	1.185	0.890	0.787
	MAE	0.412	0.389	<u>0.350</u>	0.414	0.388	<b>0.339</b>	0.371	0.603	0.452	0.465	0.568	0.498	0.482
Average	MSE	0.551	0.529	<b>0.466</b>	0.538	0.524	0.471	<u>0.470</u>	1.213	0.749	0.918	1.159	0.865	0.759
	MAE	0.411	0.444	<b>0.396</b>	0.414	0.441	0.415	<u>0.398</u>	0.580	0.486	0.520	0.580	0.530	0.492

# IV - 建模：文本模型容量是否影响性能？

## 实验结果

- 提示方法：较大的LLMs（如LLaMA-405B、Qwen-32B）在某些数据集上表现更好，但与最强的单模态模型相比仍有差距。
- 对齐方法：文本模型的容量对预测性能的影响较小。

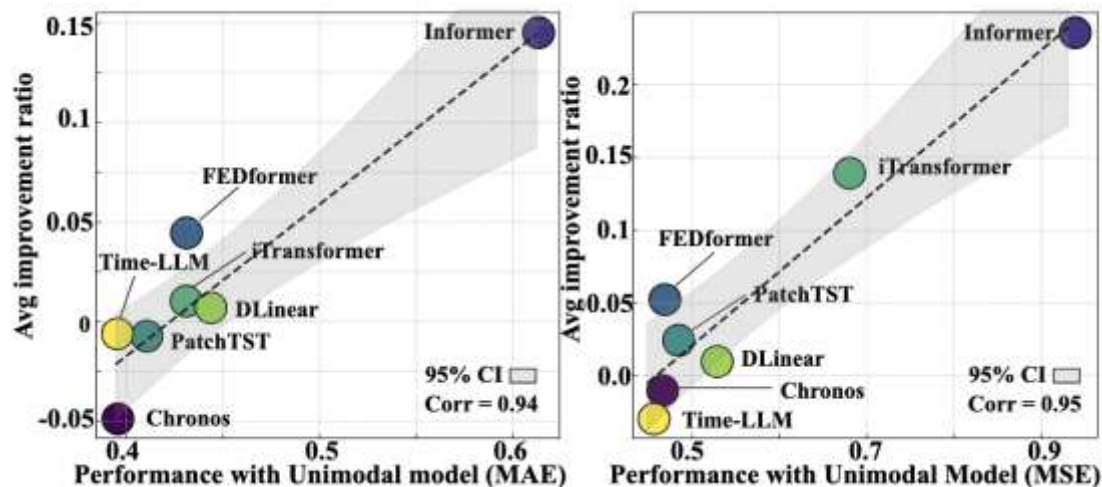




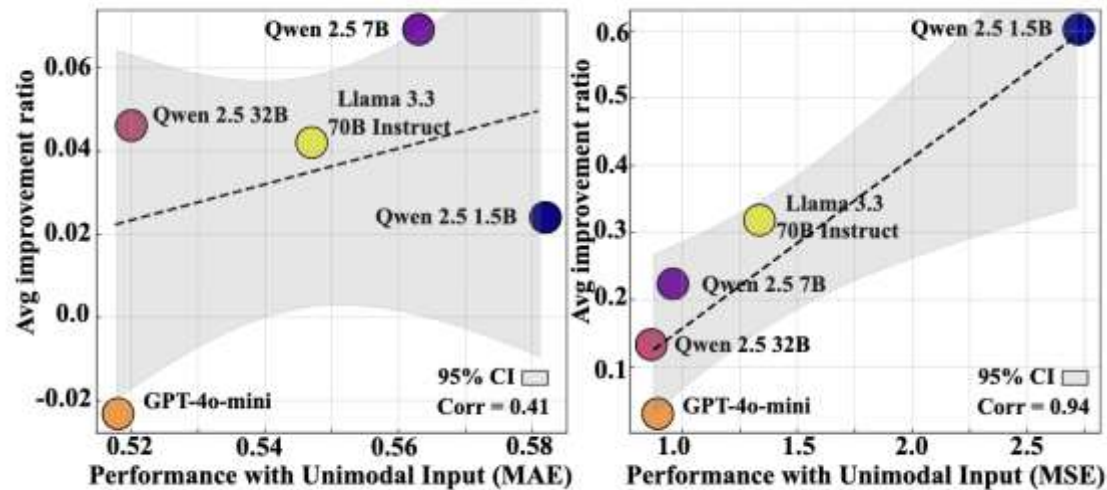
## IV - 建模：时间序列模型是否重要？

### 实验结果

- 对于基于对齐的MMTS方法，单模态预测能力与多模态改进之间存在强烈的负相关关系。
- 较弱的时间序列模型从文本信息中受益更多，而较强的单模态时间序列模型（如Chronos）在多模态设置中提升有限。
- 表明：当时间序列模型本身缺乏足够的容量来捕捉时间模式时，文本信息最有价值。



基于对齐



基于提示

## IV - 建模：对齐机制的设计有多重要？

### 实验结果

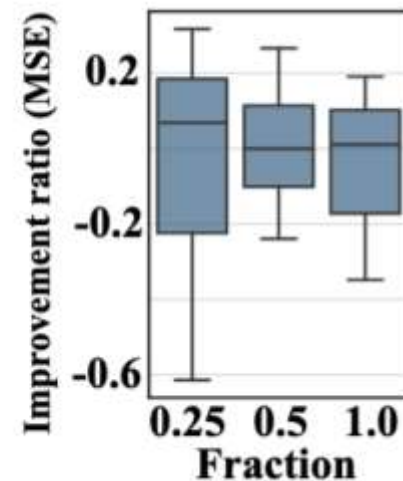
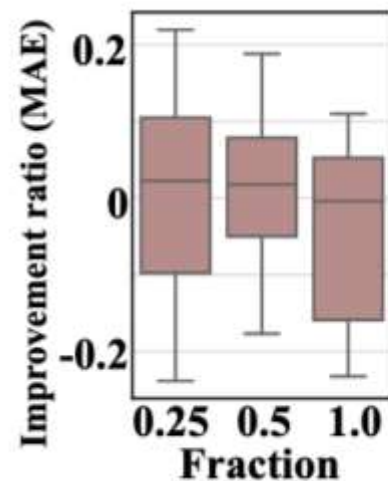
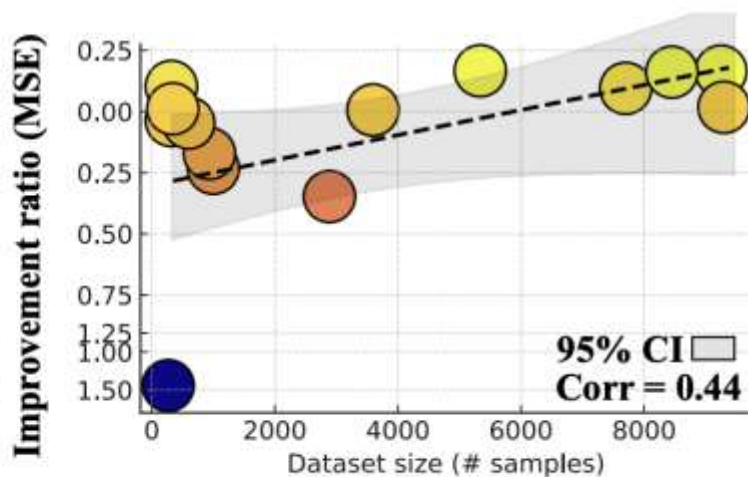
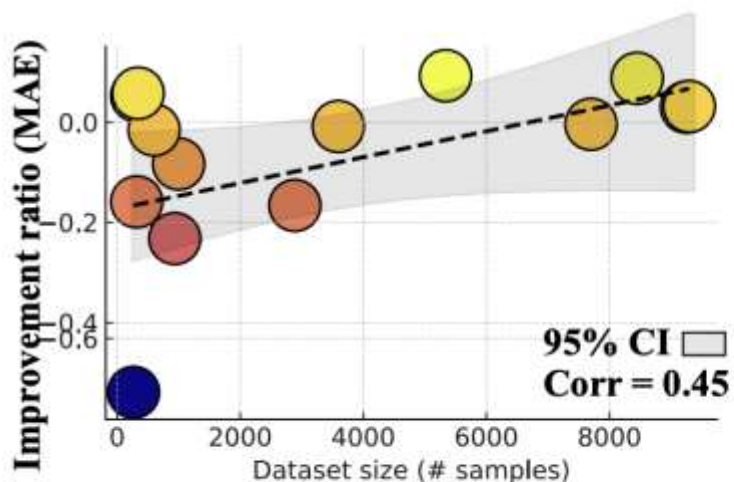
- 不同的对齐策略对性能有显著影响，最佳配置因数据集而异。
- 某些策略在平均性能上表现一致较好：通过加法聚合、平均池化、使用残差投影器、在时间序列解码器之后融合文本信息（晚期融合）以及应用高效微调。
- 启发：精心设计对齐机制以及需要灵活的架构以适应特定于数据集的特征。

Dataset	Metric	Uni	Aggregate		Pooling		Projector		Location			Fine-tuning			
			Add	Concat	Avg	CLS	Residual	MLP	Early	Mid	Late	Efficient	Fixed	Full	Two-stage
Environment	MSE	0.466	<b>0.422</b>	0.442	0.422	<b>0.391</b>	0.422	<b>0.410</b>	<b>0.411</b>	0.445	0.422	0.422	0.473	0.462	<b>0.420</b>
	MAE	0.486	<b>0.488</b>	0.504	0.488	<b>0.464</b>	<b>0.488</b>	0.489	<b>0.467</b>	0.496	0.488	0.488	0.488	0.500	<b>0.483</b>
Medical	MSE	0.596	<b>0.497</b>	0.585	<b>0.497</b>	0.570	<b>0.497</b>	0.523	0.703	0.559	<b>0.497</b>	0.497	0.538	0.617	<b>0.483</b>
	MAE	0.565	<b>0.503</b>	0.540	<b>0.503</b>	0.537	<b>0.503</b>	0.505	0.624	0.535	<b>0.503</b>	0.503	0.532	0.576	<b>0.488</b>
PTF	MSE	0.110	0.089	<b>0.088</b>	0.089	<b>0.084</b>	0.089	<b>0.088</b>	0.085	<b>0.083</b>	0.089	0.089	0.098	0.106	<b>0.085</b>
	MAE	0.167	0.162	<b>0.158</b>	0.162	<b>0.150</b>	0.162	<b>0.160</b>	0.163	0.166	<b>0.162</b>	0.162	0.157	0.162	<b>0.150</b>
MSPG	MSE	0.365	<b>0.300</b>	0.430	<b>0.300</b>	0.411	<b>0.300</b>	0.341	0.351	0.361	<b>0.300</b>	<b>0.300</b>	0.397	0.418	0.363
	MAE	0.217	<b>0.198</b>	0.230	<b>0.198</b>	0.229	<b>0.198</b>	0.204	0.205	0.209	<b>0.198</b>	<b>0.198</b>	0.229	0.223	0.217
LEU	MSE	0.522	<b>0.513</b>	0.551	<b>0.513</b>	0.546	<b>0.513</b>	0.519	0.515	<b>0.500</b>	0.513	<b>0.513</b>	0.526	0.531	0.568
	MAE	0.350	<b>0.339</b>	0.358	<b>0.339</b>	0.360	<b>0.339</b>	0.361	0.354	0.350	<b>0.339</b>	<b>0.339</b>	0.355	0.349	0.360
Average	MSE	0.412	<b>0.364</b>	0.419	<b>0.364</b>	0.400	<b>0.364</b>	0.376	0.413	0.390	<b>0.364</b>	<b>0.364</b>	0.406	0.427	0.384
	MAE	0.357	<b>0.338</b>	0.358	<b>0.338</b>	0.348	<b>0.338</b>	0.344	0.363	0.351	<b>0.338</b>	<b>0.338</b>	0.352	0.362	0.340

## IV - 数据：上下文长度和训练样本大小如何影响性能？

### 实验结果

- 在基准测试中，数据集大小与MMTS性能提升之间存在正相关关系。
- 较大的训练数据集有助于多模态模型的学习。
- MMTS的好处不仅取决于模型设计，还取决于是否有足够的数据来支持跨模态的学习。



# IV - 数据：文本和时间序列对齐质量如何影响性能？

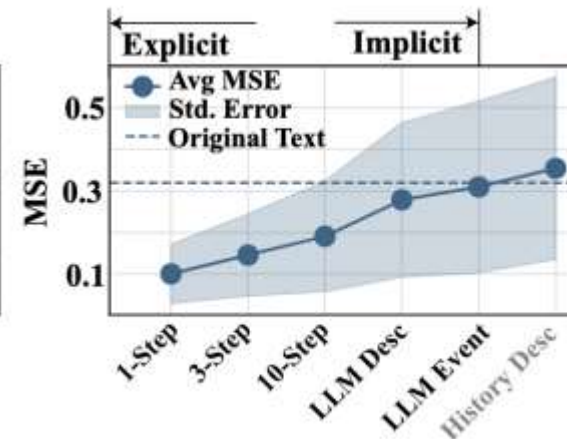
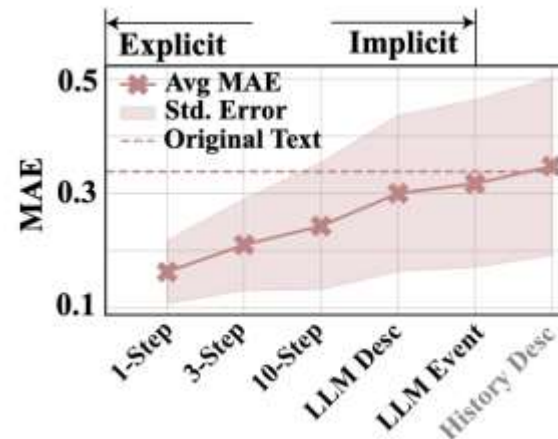
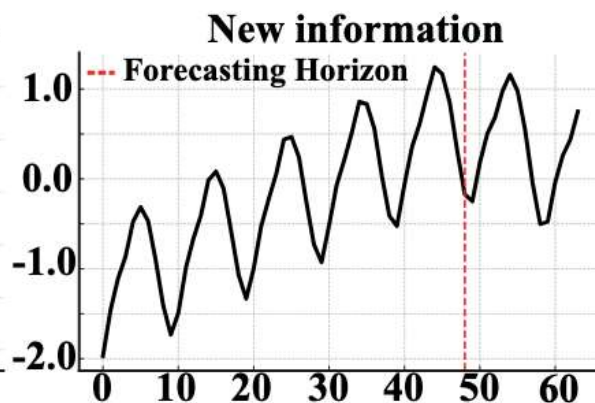
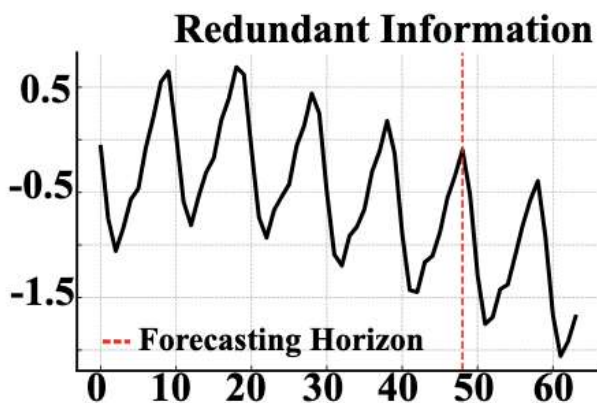
## 合成数据集实验

- 通过控制文本和时间序列之间的关系，研究文本信息是否提供独特的预测信号。
- 结论：只有当文本信息提供与时间序列不重叠的预测信号时，多模态方法才显著优于单模态方法。

## 真实数据集实验

- 通过生成不同类型的文本变体（如未来变化的精确描述、未来变化的平均值、LLM生成的描述等），研究文本信息的明确性对多模态预测性能的影响。
- 结论：文本信息越明确，多模态方法的性能提升越显著。

Model	Trend				Seasonality				Spike			
	Unique MSE	MAE	Redundant MSE	MAE	Unique MSE	MAE	Redundant MSE	MAE	Unique MSE	MAE	Redundant MSE	MAE
Unimodal	0.811	0.677	0.449	0.394	0.299	0.410	0.004	0.044	3.851	0.934	5.798	1.039
Multimodal	0.477	0.408	0.428	0.376	0.008	0.066	0.005	0.052	3.267	0.893	5.714	1.065







# V 结论与讨论



## 结论

- 多模态方法的有效性：多模态方法并不总是优于单模态方法，其效果高度依赖于模型架构、对齐策略和数据特性。
- 文本模型容量：较大的LLMs在提示方法中表现更好，但在对齐方法中影响较小。
- 时间序列模型容量：较弱的时间序列模型从文本信息中受益更多。
- 对齐策略：不同的对齐策略对性能有显著影响，最佳配置因数据集而异。
- 数据集大小：较大的训练数据集有助于多模态模型的学习。
- 文本和时间序列的对齐质量：只有当文本信息提供与时间序列不重叠的预测信号时，多模态方法才显著优于单模态方法。

## 未来方向：

- 探索更多模态（如图像、音频）
- 扩展测试数据集
- 解决将文本与时间序列结合可能会引发的隐私问题，例如在应用于医疗保健或金融等敏感领域时





谢谢！

