清華大學
Tsinghua University

终期汇报：

基于多模态融合的时间序列文本问答大模型

汇报人：李沐晟　　指导教师：谷源涛 教授



清華大學
Tsinghua University

内 容
contents

1 背景介绍

2 模型方法

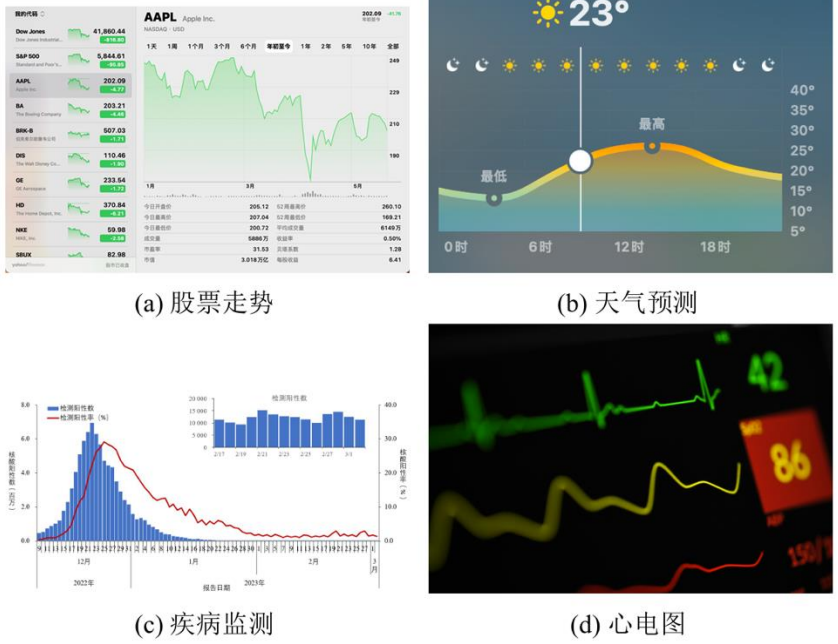
3 数据合成

4 实验结果

5 总结展望

背景介绍

时间序列无处不在



(a) 股票走势　(b) 天气预测

(c) 疾病监测　(d) 心电图

图 1.1　典型的时间序列应用场景

时间序列分析已成为众多领域进行数据建模和决策支持的核心手段。

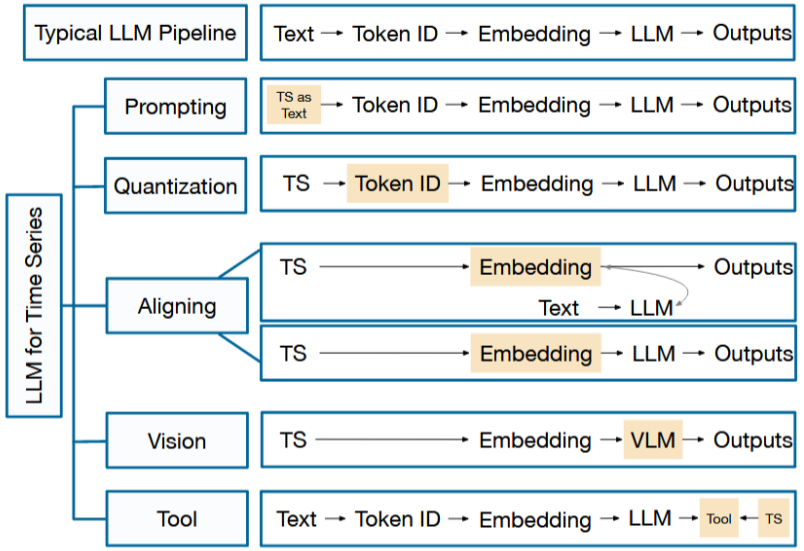
时间序列分析方法变革

20世纪初	经典统计学 与信号处理领域	自回归、移动平均 傅立叶变换、小波变换
20世纪末	基于统计学习的 模型方法	隐马尔可夫模型 状态空间模型
21世纪以来	机器学习与 深度学习	RNN、LSTM CNN、Transformer等
近几年来	大语言模型与 多模态融合技术	TS-MLLM ……

时间序列多模态大语言模型：仍处在初步探索阶段

背景介绍

时间序列多模态大语言模型(TS-MLLM)：实现方法



- **Prompt(提示)**：纯文本输入
- **Quantize(量化)**：量化为有限token
- **Align(对齐)**：与LLM嵌入特征对齐
- **Vision(视觉)**：绘图使用VLM
- **Tool(工具)**：调用外部工具

当前主要挑战

- 模型框架**简单直接迁移，缺乏可解释性**研究
- 模型能力局限，**通用性不足，泛化能力不佳**
- 从头构建数据集，数据**不互通**，缺乏**横向比较**

本工作目标

- 探索基于嵌入空间对齐的模型结构的**有效性与可解释性**
- 探究模型中对于时间序列建模分析能力的**重要影响因素**
- **统一测试**比较本领域的模型性能，对比差异，寻找**瓶颈**

Zhang Xiyuan, et al. Large Language Models for Time Series: A Survey.

模型方法

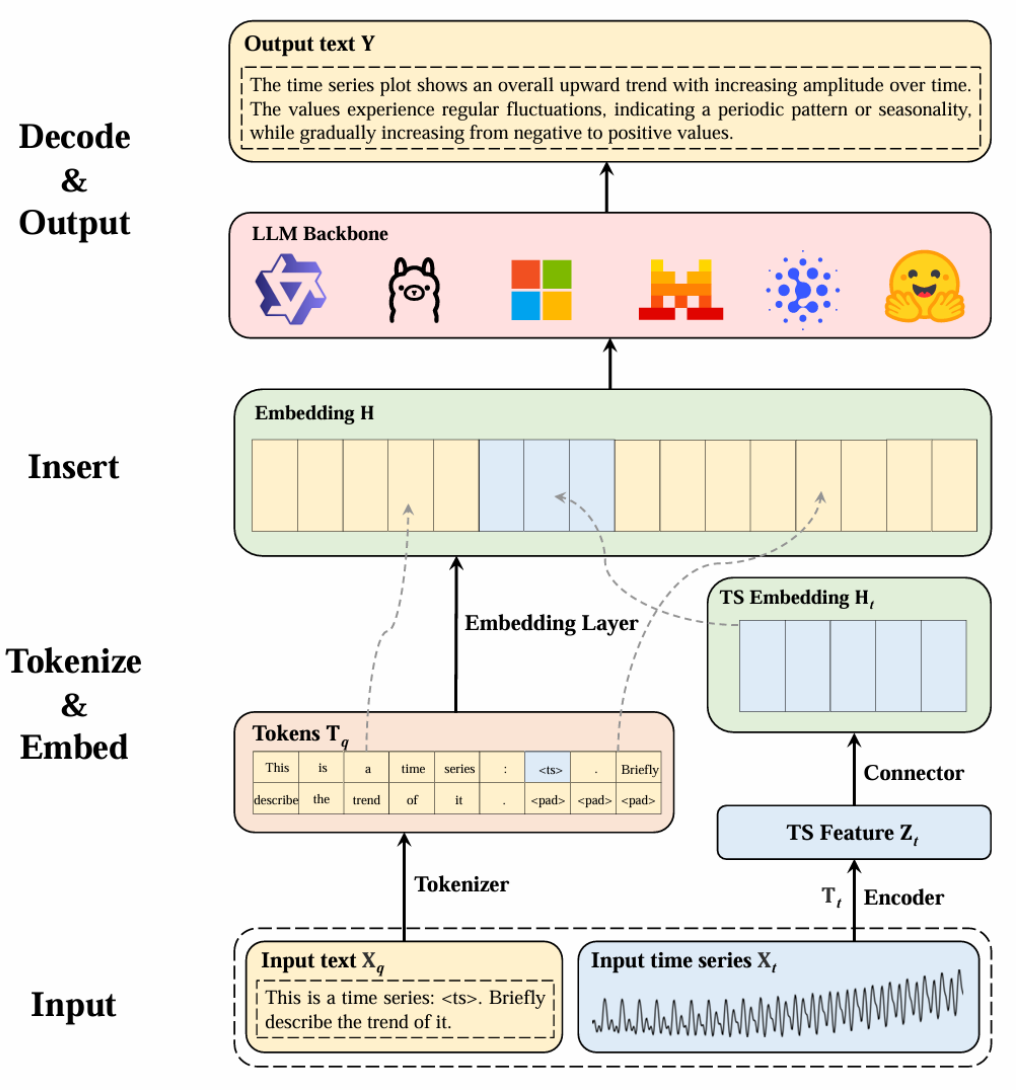


图 2.1　本工作模型结构概览图

主要模块

- 时间序列编码器
- LLM骨干网络

推理过程

- **数据输入**：文本输入+时间序列输入
- **分词嵌入**：文本按照LLM网络方法分词嵌入，时间序列输入编码器切片后投影到LLM嵌入空间
- **插入拼接**：将时间序列特征插入到文本嵌入特征对应位置处，在输入文本中用<ts>标记
- **解码输出**：把整体嵌入特征送入注意力层解码输出

训练策略

- ***冷启动***：使用CLIP方法对比学习实现模态初步对齐
- **预训练**：冻结LLM骨干网络，微调时间序列编码器
- **微调**：LLM骨干网络与时间序列编码器全参数微调

数据合成

时间序列数据合成

基于高斯过程(GP)采样的时间序列合成

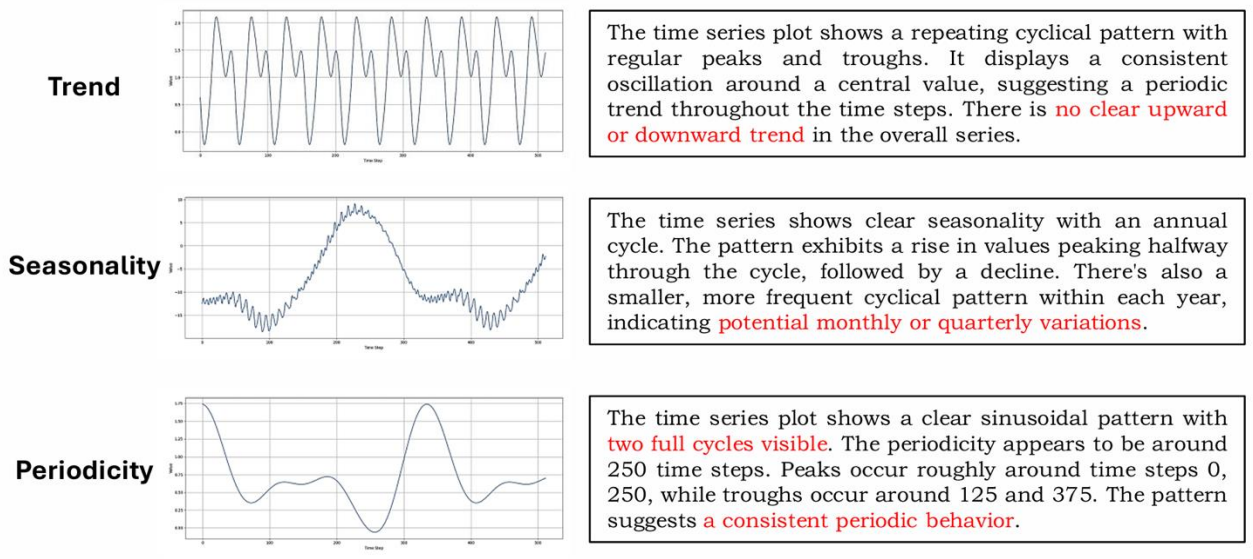
参考Chronos[1]工作的方法，从高斯核仓库中随机选取并组合高斯核，从得到的高斯过程中采样时间序列。

基于时间-趋势(ST)分解的时间序列合成

参考TADNet[2]工作的方法，随机采样超参数，分别合成季节项、趋势项、残差项，最终合成丰富的时间序列数据。

文本标注生成

将时间序列绘制为图像，使用通用视觉大模型ChatGPT-4o生成时间序列基础属性描述与简单问答文本，形成文本-时间序列对。



经人工抽样筛查
合成数据质量高，符合使用要求

[1] Ansari A F, et al. Chronos: Learning the Language of Time Series.
[2] Zhang Z, et al. Unravel Anomalies: An End-to-end Seasonal-Trend Decomposition Approach for Time Series Anomaly Detection.

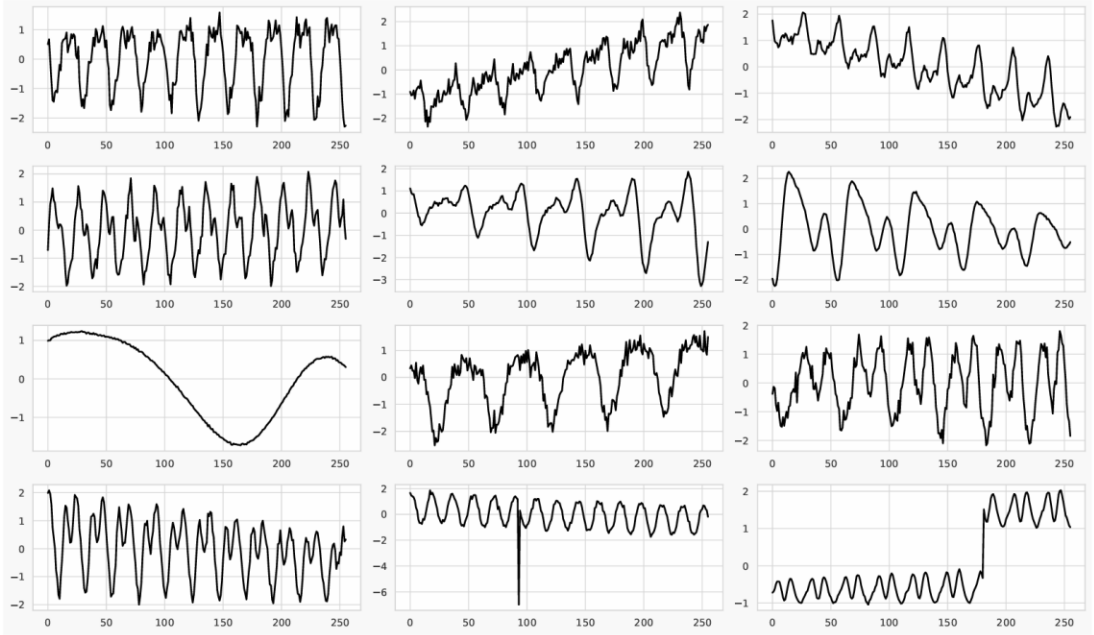
数据合成

数据增强

操作	描述
Jittering	在原始序列上添加小幅高斯噪声，使数据略微扰动
Scaling	将整段时间序列幅度按某个缩放因子放大或缩小
Local Scaling	将时间序列的局部幅度按某个缩放因子放大或缩小
Multi-piece Scaling	将序列分为多个片段，并分别对每段进行不同的缩放
Permutation	将时间序列划分为多个片段，并打乱它们的顺序
Reverse	将整个时间序列反转
Local Reverse	随机选取时间序列的某个片段进行反转
Random Erasing	随机选取时间序列的某个片段置为 0
Time Warp	使用样条插值对时间轴进行非线性扭曲
Upsample	通过重复点数来扩展序列长度
Downsample	通过等距保留部分点实现降采样
Amplitude Spectrum Perturbation	在频域中扰动幅度谱
Phase Spectrum Perturbation	在频域中扰动相位谱

- **时间序列**：添加噪声、缩放、反转、抹除等
- **文本**：多样化文本表述，多样化文本任务

其他来源时间序列



TSQA 数据集样例图，竖直方向依次表示趋势、波动性、季节性、异常值四类任务，水平方向代表同类任务中的三种不同情况。

图 3.4　TSQA 数据集样例

- **TSQA**：基础分析选择题，取自ChatTime[1]论文
- **MMTS**：复杂推理选择题，本组其他成员合成

Wang C, et al. ChatTime: A Unified Multimodal Time Series Foundation Model Bridging Numerical and Textual Data.

实验结果

TS-MLLM 模型的有效性分析

表 4.1　不同模型在 TSQA 测试数据上的表现

Model	Initialized	Pretrained	Finetuned	Qwen(text)	GPT-4o(text)	GPT-4o(full)
Trend	0.34	0.84	**1.00**	0.34	0.16	0.69
Volatility	0.37	0.54	**0.98**	0.37	0.27	0.54
Seasonality	0.38	0.41	0.54	0.39	0.38	**0.71**
Outliers	0.36	0.51	0.58	0.36	0.14	**0.98**
Average	0.3625	0.5750	**0.8200**	0.3650	0.2375	0.7675

表 4.2　不同模型在 MMTS-Benchmark 测试数据上的表现

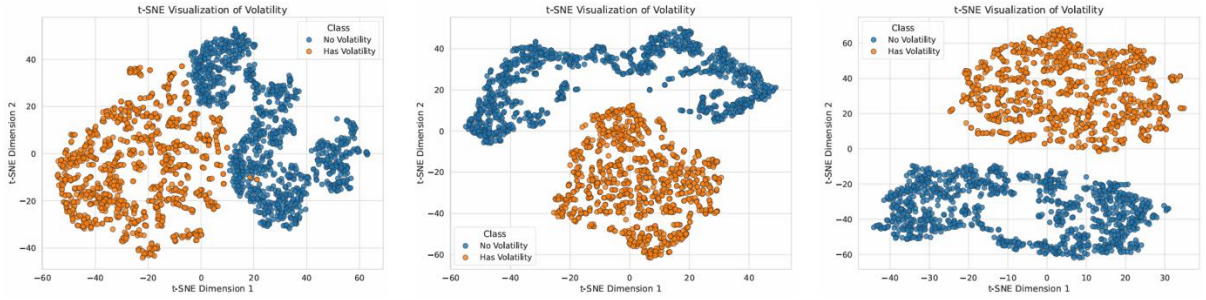
Model Name	Initialized	Pretrained	Finetuned	Qwen(text)	GPT-4o(text)	GPT-4o(full)
Accuracy	0.3333	0.4493	-	0.3551	0.3696	**0.5662**

- 初始化模型表现接近随机猜测，**预训练**对齐后性能**有所提升**，**微调**后性能**进一步提升**
- 纯文本输入表现接近随机猜测，与未训练模型性能相当，表明**题干未泄漏关键信息**
- 微调后的模型能在部分任务上超过图片输入的GPT-4o，在复杂推理任务上模型性能有所提升，但GPT-4o能力更强。

当前模型结构能够有效实现模态对齐与交互
“预训练+微调”训练策略能让模型逐步学习时间序列建模分析能力

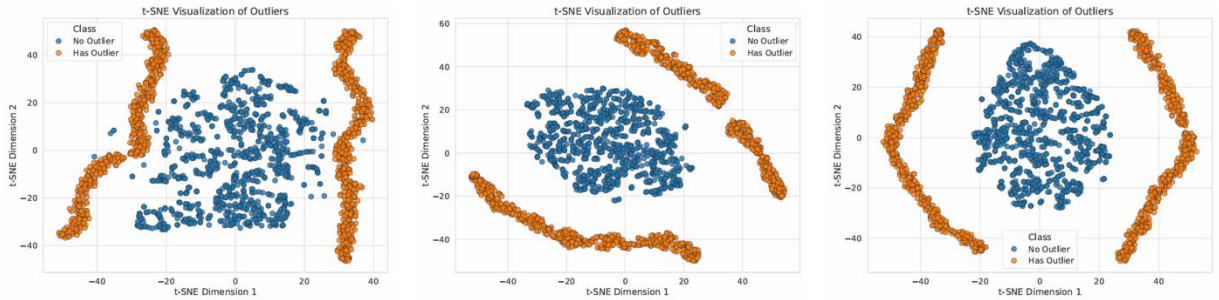
实验结果

TS-MLLM 模型的可解释性探究：聚类分析



(a) 初始化　(b) 预训练　(c) 微调

图 4.2　不同阶段模型编码时间序列特征二维 t-SNE 聚类效果图：波动性



(a) 初始化　(b) 预训练　(c) 微调

图 4.4　不同阶段模型编码时间序列特征二维 t-SNE 聚类效果图：异常值

表 4.4　不同阶段模型编码时间序列特征聚类评价指标：波动性

Stage	Initialized	Pretrained	Finetuned
FDR	941	**2650**	1070
SC	0.4161	0.4070	**0.5734**

表 4.6　不同阶段模型编码时间序列特征聚类评价指标：异常值

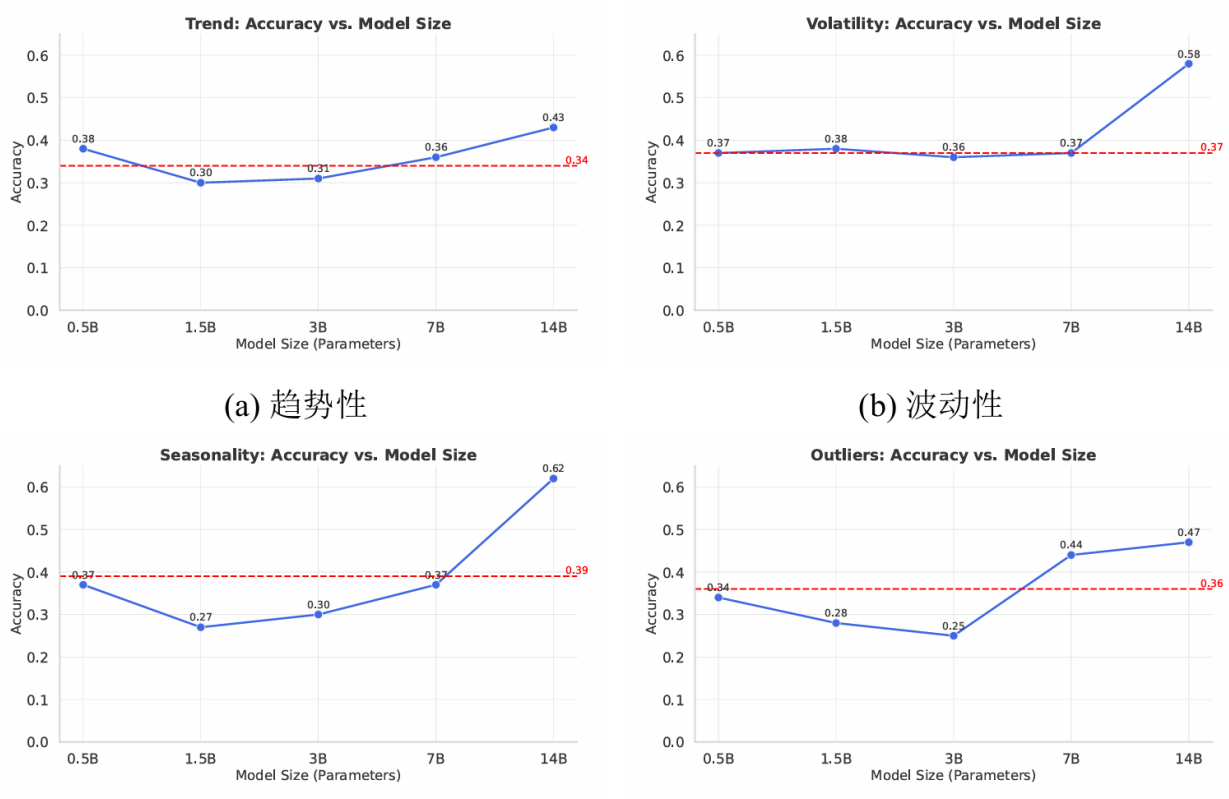
Stage	Initialized	Pretrained	Finetuned
FDR	866	**2414**	973
SC	0.1243	**0.2051**	0.1774

- 初始化编码器也能一定程度区分时间序列特征
- 从图像上看，训练有助于分离簇团，优化输出空间
- 从指标上看，训练有助于提升聚类效果

预训练与微调能优化编码器输出空间，有利于提取时间序列特征与促进模态耦合

实验结果

影响时间序列分析建模能力的关键因素分析



(a) 趋势性　(b) 波动性

(c) 季节性　(d) 异常值

图 4.6　使用不同规模参数量骨干网络的模型在 TSQA 数据集上的正确率



图 4.5　使用不同规模参数量骨干网络的模型的训练过程损失曲线

- 整体上**更大**的LLM骨干网络对应**更好**的表现
- **考虑计算资源与模型性能的平衡**

实验结果

影响时间序列分析建模能力的关键因素分析

表 4.7　不同层数 MLP 编码器对应模型在 TSQA 与 MMTS 测试集上的表现

MLP Layers	1	2	3	5	7
Params	118,272	12,996,912	25,815,552	51,512,832	77,210,112
Trend	0.34	0.34	**0.57**	0.52	0.35
Volatility	0.29	0.41	0.29	**0.42**	0.37
Seasonality	0.37	**0.51**	**0.51**	0.44	0.33
Outliers	0.37	**0.52**	0.47	0.43	0.40
TSQA Average	0.3425	0.4450	**0.4600**	0.4525	0.3625
MMTS	0.1957	0.3406	0.3913	0.4058	0.3986

表 4.9　编码器切片长度对模型性能的影响

Window Length	8	16	32	64
Accuracy	**0.4493**	0.4203	0.3768	0.3986

- 随着编码器参数规模扩大，模型表现**先升后降**，不同任务的**最优点不同**
- 更小的切片长度对应更大的**输入分辨率**，使模型性能提升

LLM骨干网络参数量、时间序列编码器深度与切片长度、训练超参数*等对模型能力有重大影响，为实验提供经验与数据指导

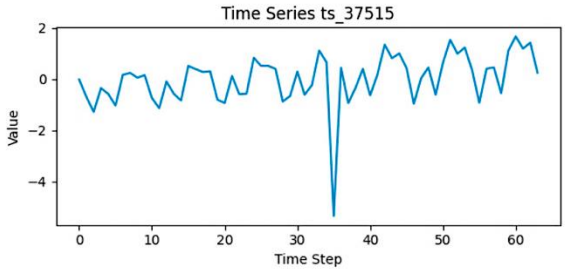
此外，还探究了批大小、学习率、轮次等超参数的影响，详见论文正文

实验结果

多模态时间序列大模型统一测试评估

Model	Ours(Init)	Ours(Pretrain)	Ours(Finetune)	ChatTime	ChatTS
Trend	0.34	0.84	**1.00**	0.98	0.92
Volatility	0.35	0.54	**0.98**	0.94	0.37
Seasonality	0.39	0.41	0.54	**0.74**	0.49
Outliers	0.31	0.51	0.58	**0.95**	0.91
Average	0.3475	0.5750	0.8200	**0.9025**	0.6475
MMTS	0.3333	0.4493	-	-	**0.4638**

- 当前主流模型均在**基础**时间序列分析任务上有不错的性能，但存在**不平衡**问题
- 本项目微调模型与ChatTime模型因微调后**丧失通用语言能力**而无法作答MMTS
- 由于不确定ChatTime工作对数据集的划分，测试时可能存在**一定的数据泄露**问题



The time series appears to have a significant outlier around the 30th data point, where the value sharply decreases from approximately -2.5 to -4. This drop is more pronounced compared to other fluctuations in the series.

已初步具备时间序列分析能力，但因训练数据不平衡不充分，导致整体泛化性能不佳，并存在定位不准的幻觉问题。

Xie Z, et al. ChatTS: Aligning Time Series with LLMs via Synthetic Data for Enhanced Understanding and Reasoning.
Wang C, et al. ChatTime: A Unified Multimodal Time Series Foundation Model Bridging Numerical and Textual Data.

总结展望

清華大學 Tsinghua University

本工作主要贡献

1. 验证了本工作使用的模型架构的有效性，并根据不同阶段的编码器输出聚类探究其可解释性
2. 针对影响模型性能的关键因素进行系统性的对比实验，为训练实践提供了启发与指导
3. 对本领域模型进行统一测试与案例研究，揭示了训练数据不平衡不充分的问题，并发现不同模型均存在泛化性能较差和定位不准的幻觉问题

未来进一步研究方向

- 更高效的**编码器结构设计**——针对时间序列分析任务特性设计专用编码器
- 更平衡的**训练数据集构建**——构建数据分布更丰富更平衡的训练集，提升泛化性能
- 更精准的**数据输入形式探索**——尝试同时输入时间序列与时间轴，消除定位不准问题
- 更广泛的**模型训练策略优化**——尝试正则化方法、强化学习等训练策略提升模型性能

清華大學
Tsinghua University

感谢垂听，敬请批评指正！



附录

数据标注结果

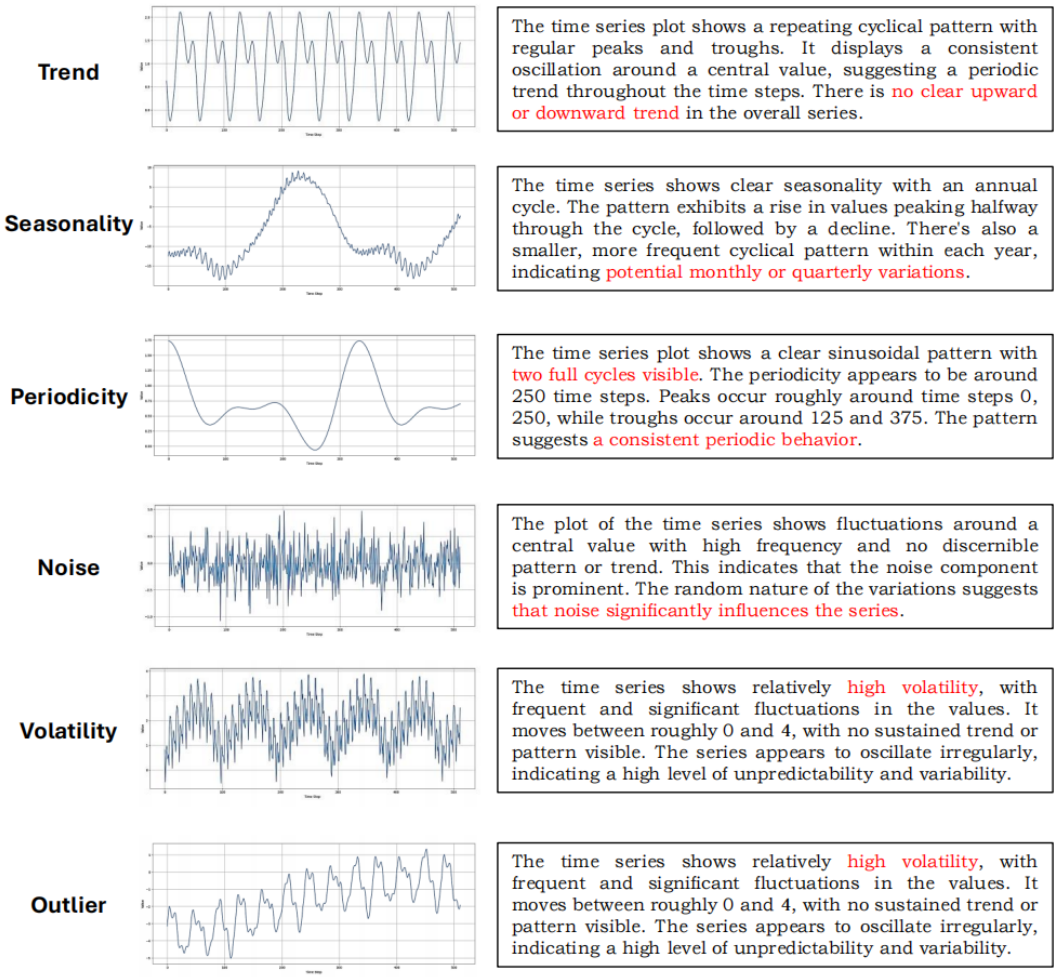


图 B.9　使用 MLLM 标注生成的时间序列-文本数据样例

数据标注提示词（部分）

- **System Prompt：** You are a helpful AI assistant, you should help the user answer the given question correctly and concretely according to the given plot of a time series.
- **Trend：** This is a plot of a time series:<ts>. Briefly describe the trend of the given time series plot.
- Seasonality：This is a plot of a time series:<ts>. Briefly describe the seasonality of the given time series.
- **Periodicity：** This is a plot of a time series:<ts>. Briefly describe the periodicity of the given time series plot.
- **Noise：** This is a plot of a time series:<ts>. Briefly evaluate whether the noise component in this time series is prominent.
- **Volatility：** This is a plot of a time series:<ts>. Briefly describe the volatility of the this time series.
- **Outlier：** This is a plot of a time series:<ts>. Briefly describe the outlier(not necessarily present) of the time series according to the plot.
- **Global Extreme Values：** This is a plot of a time series:<ts>. Briefly describe the the extreme values of the time series, including its minimum, maximum, and overall range.

附录

本工作用到的数据概况

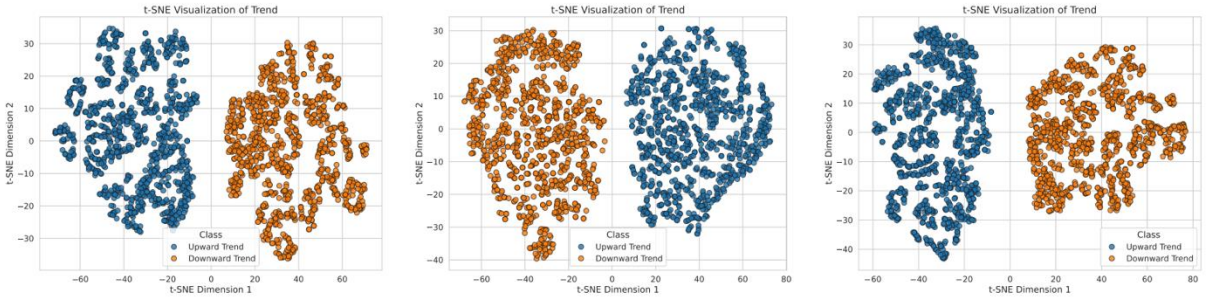
表 3.1　本工作数据集统计情况

数据集名称	时间序列数量	样本数量
GP-DESC	1000	9000
STL-DESC	1500	13500
Aug-DESC-QA	2500	5000
TSQA[16]	48000	48000
MMTS-Benchmark	1200	800

- TSQA 数据集：该合成数据集来自 ChatTime[16]，包含 48000 条简单时间序列问答数据，包含趋势（Trend）、波动性（Volatility）、季节性（Seasonality）、异常值（Outliers）四个方面的选择题。其中时间序列的长度有 64、128、256、512，整体分布均匀，其数据示例如图 3.4所示。在本工作中，将该数据集按照 8:2 的比例划分为训练集和测试集，用于训练与测试模型的基础时间序列分析建模能力。
- MMTS-Benchmark 测试集：GuLab 时间序列小组 Benchmark 工作的数据集，由同组其他成员合成构建，本工作中仅使用该数据集进行测试。数据集包含 800 个真实时间序列数据复杂推理问题，涵盖交通、电力、等领域，包含单变量与多变量选择题，用于测试模型的时间序列复杂问题推理分析能力。在本文中，将其简称为 MMTS 测试集。

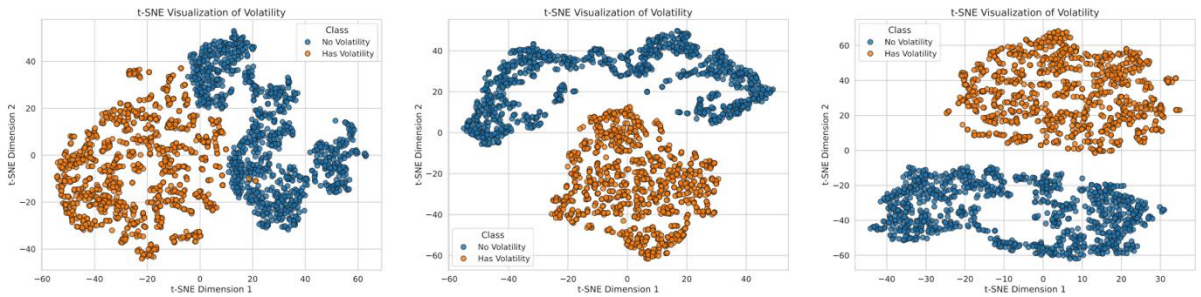
附录

聚类研究



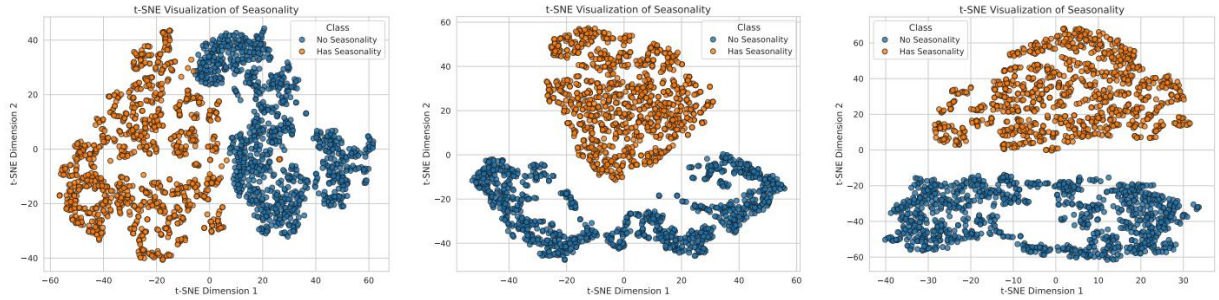
(a) 初始化　(b) 预训练　(c) 微调

图 4.1　不同阶段模型编码时间序列特征二维 t-SNE 聚类效果图：趋势



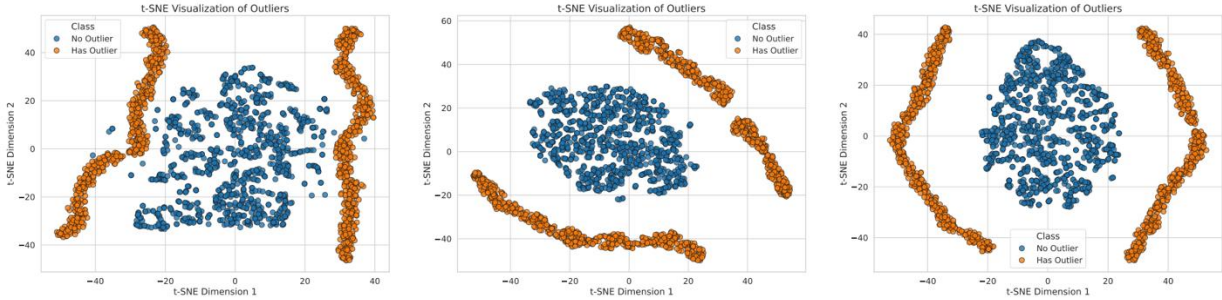
(a) 初始化　(b) 预训练　(c) 微调

图 4.2　不同阶段模型编码时间序列特征二维 t-SNE 聚类效果图：波动性



(a) 初始化　(b) 预训练　(c) 微调

图 4.3　不同阶段模型编码时间序列特征二维 t-SNE 聚类效果图：季节性

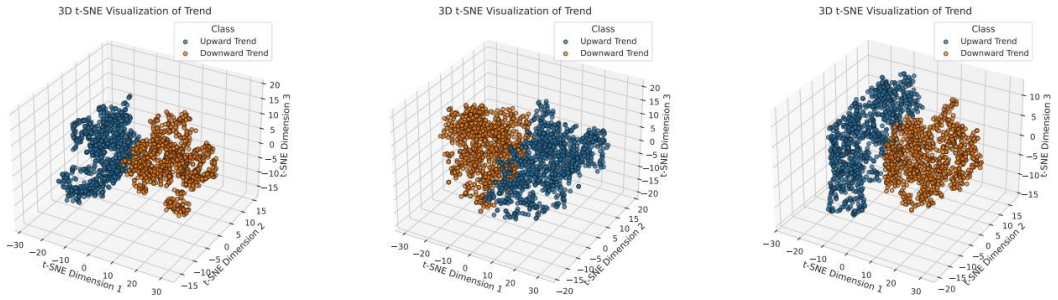


(a) 初始化　(b) 预训练　(c) 微调

图 4.4　不同阶段模型编码时间序列特征二维 t-SNE 聚类效果图：异常值

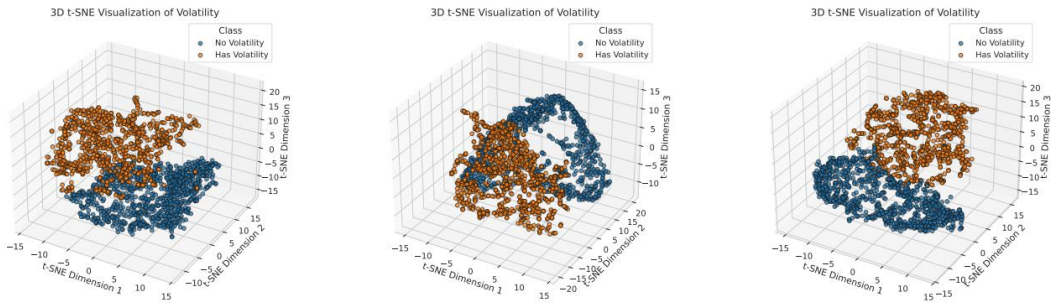
附录

聚类研究



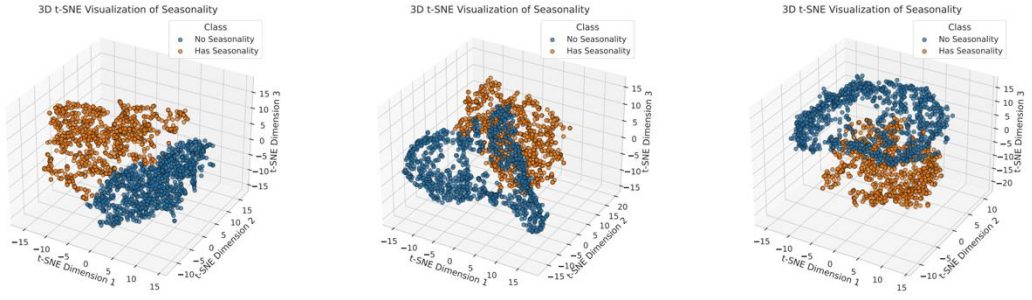
(a) 初始化　(b) 预训练　(c) 微调

图 B.1　不同阶段模型编码时间序列特征三维 t-SNE 聚类效果图：趋势



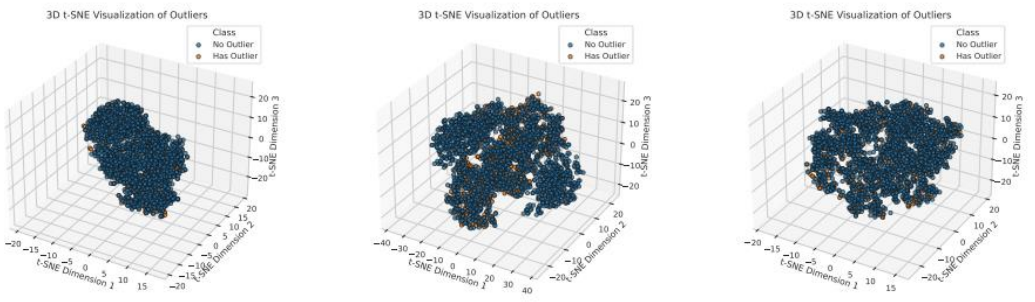
(a) 初始化　(b) 预训练　(c) 微调

图 B.2　不同阶段模型编码时间序列特征三维 t-SNE 聚类效果图：波动性



(a) 初始化　(b) 预训练　(c) 微调

图 B.3　不同阶段模型编码时间序列特征三维 t-SNE 聚类效果图：季节性



(a) 初始化　(b) 预训练　(c) 微调

图 B.4　不同阶段模型编码时间序列特征三维 t-SNE 聚类效果图：异常值

附录

聚类研究

Fisher判别比（FDR）

$$\mathrm{FDR} = \frac{\|\mu_1 - \mu_2\|^2}{\sigma_1^2 + \sigma_2^2}$$

值越大代表分类效果越好

轮廓系数（SC）

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\ b(i)\}}$$

取值范围为 $[-1, +1]$，趋近+1代表聚类效果好。

表 4.3　不同阶段模型编码时间序列特征聚类评价指标：趋势

Stage	Initialized	Pretrained	Finetuned
FDR	17338	25506	**39161**
SC	0.5909	**0.6037**	0.5849

表 4.4　不同阶段模型编码时间序列特征聚类评价指标：波动性

Stage	Initialized	Pretrained	Finetuned
FDR	941	**2650**	1070
SC	0.4161	0.4070	**0.5734**

表 4.5　不同阶段模型编码时间序列特征聚类评价指标：季节性

Stage	Initialized	Pretrained	Finetuned
FDR	933	**2591**	1055
SC	0.4211	0.4108	**0.5718**

表 4.6　不同阶段模型编码时间序列特征聚类评价指标：异常值

Stage	Initialized	Pretrained	Finetuned
FDR	866	**2414**	973
SC	0.1243	**0.2051**	0.1774

附录

LLM推理公式表示

$$\mathbf{T}_q = [\mathbf{X}_{q,1}, \mathbf{X}_{q,2}, \cdots, \mathbf{X}_{q,k-1}, \mathbf{X}_{q,k}, \mathbf{X}_{q,k+1}, \cdots, \mathbf{X}_{q,m}] \tag{2.1}$$

$$\mathbf{H}_q = \mathbf{Embed}(\mathbf{T}_q) = [\mathbf{H}_{q,1}, \mathbf{H}_{q,2}, \cdots, \mathbf{H}_{q,k-1}, \mathbf{H}_{q,k}, \mathbf{H}_{q,k+1}, \cdots, \mathbf{H}_{q,m}] \tag{2.2}$$

$$\mathbf{T}_t = [\mathbf{X}_{t,1}, \mathbf{X}_{t,2}, \cdots, \mathbf{X}_{t,n}] \tag{2.3}$$

$$\mathbf{Z}_t = \mathbf{Encoder}(\mathbf{T}_t) = [\mathbf{Z}_{t,1}, \mathbf{Z}_{t,2}, \cdots, \mathbf{Z}_{t,n}] \tag{2.4}$$

$$\mathbf{H}_t = \mathbf{W} \cdot \mathbf{Z}_t \tag{2.5}$$

$$\mathbf{H} = \mathbf{Insert}(\mathbf{H}_t, \mathbf{H}_q) \tag{2.6}$$

$$= [\mathbf{H}_{q,1}, \mathbf{H}_{q,2}, \cdots, \mathbf{H}_{q,k-1}, \mathbf{H}_t, \mathbf{H}_{q,k+1}, \cdots, \mathbf{H}_{q,m}] \tag{2.7}$$

$$\mathbf{Y} = \mathbf{Decoder}(\mathbf{Transformer}(\mathbf{H})) \tag{2.8}$$

对比学习冷启动示意图

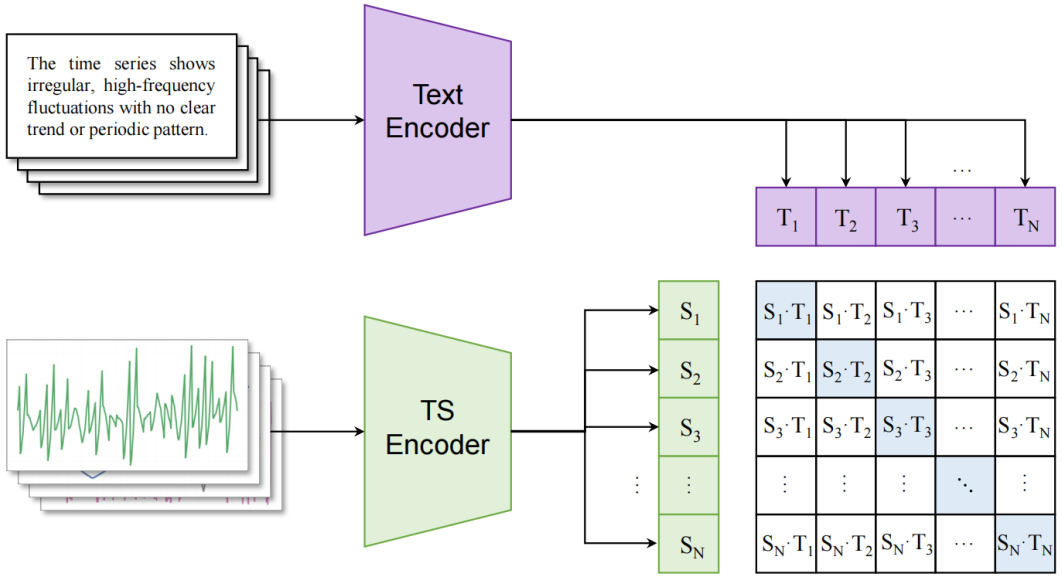


图 2.3　对比学习示意图

附录

LLM骨干网络对复杂推理能力的影响

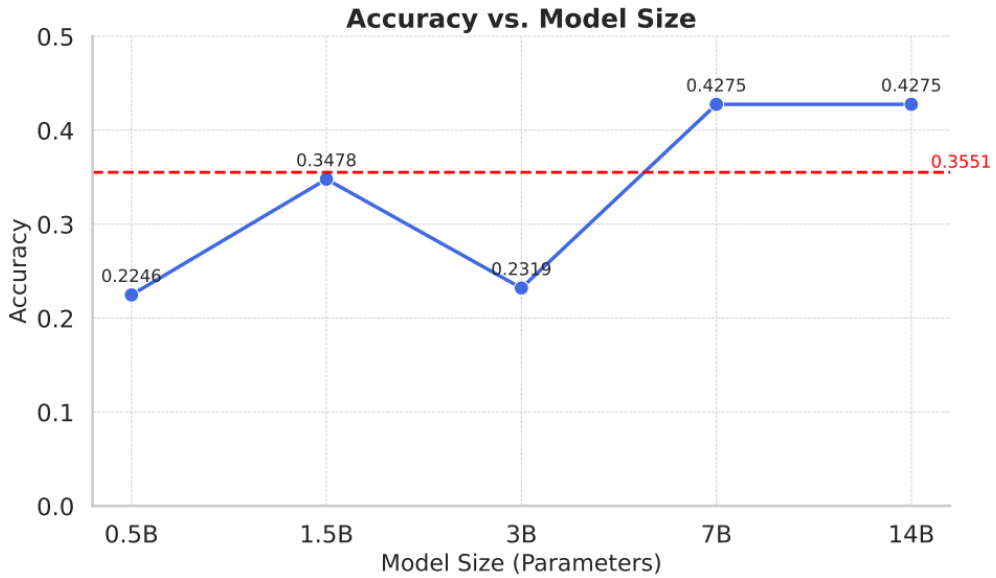


图 4.7　使用不同规模参数量骨干网络的模型在 MMTS 数据集上的正确率

PatchTST层数对性能的影响

表 4.8　不同层数 PatchTST 编码器对应模型在 TSQA 与 MMTS 测试集上的表现

PatchTST Layers	1	2	3	5
Params	1,127,555	1,260,036	1,392,517	1,657,479
Trend	0.37	0.34	0.33	0.34
Volatility	0.30	0.30	0.30	0.30
Seasonality	0.37	0.34	0.33	0.34
Outliers	0.40	0.40	0.40	0.40
TSQA Average	0.3600	0.3450	0.3400	0.3450
MMTS	0.3969	0.3913	0.4275	0.3696

附录

不同切片长度训练损失曲线

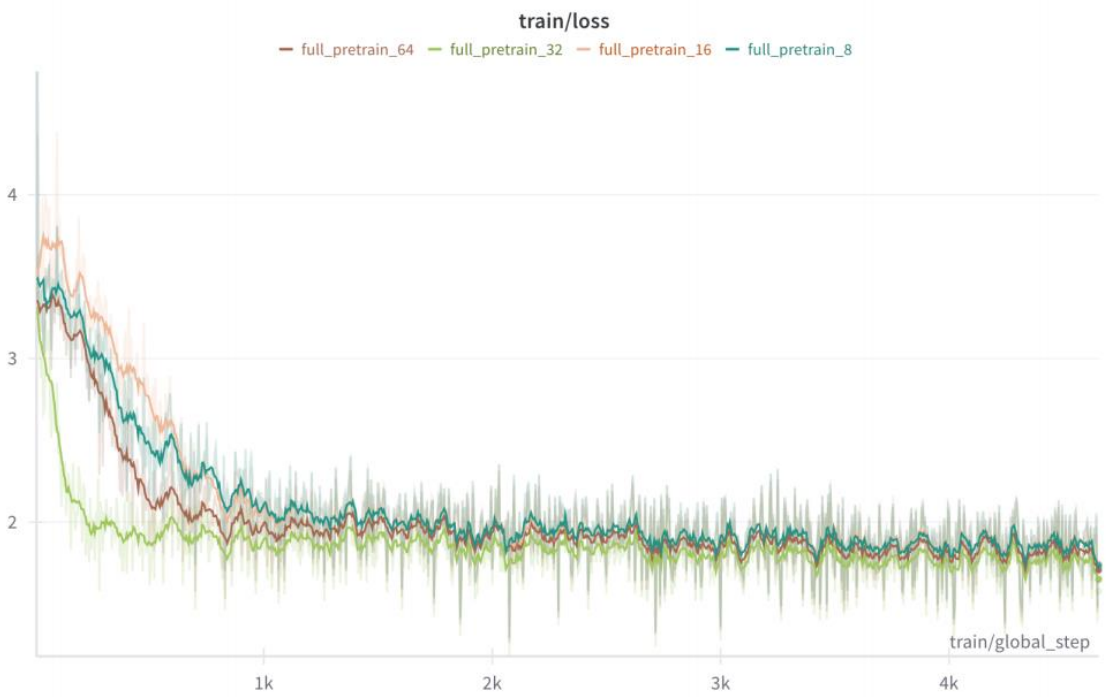


图 4.8　不同编码器切片长度对应的训练过程损失曲线

训练数据规模的影响

表 4.10　不同训练数据下的模型在 TSQA 测试集上的表现

Train Dataset	2k	4k	8k	Full	Full-Aug	Qwen(text)	Init
Trend	0.32	0.30	0.31	0.52	**0.84**	0.34	0.34
Volatility	0.30	0.25	0.32	0.65	**0.85**	0.37	0.35
Seasonality	0.38	0.30	0.39	**0.52**	0.41	0.39	0.39
Outliers	0.33	0.45	0.37	**0.67**	0.51	0.36	0.31
Average	0.3325	0.3250	0.4130	0.4275	**0.4493**	0.3551	0.3333

表 4.11　不同训练数据下的模型在 MMTS 测试集上的表现

Train Dataset	2k	4k	8k	Full	Full-Aug	Qwen(text)
Trend	0.2319	0.3913	0.4130	0.4275	**0.4493**	0.3551

附录

学习率的影响

表 4.12　不同学习率下训练的模型在 TSQA 与 MMTS 测试集上的表现

Learning Rate	1e-6	3e-6	1e-5	3e-5	1e-4	3e-4
Trend	0.30	0.45	0.58	0.54	0.65	**0.74**
Volatility	0.40	0.30	0.61	**0.71**	0.55	0.33
Seasonality	0.43	0.31	0.56	0.63	0.54	**0.69**
Outliers	0.41	0.40	0.56	0.55	0.47	**0.57**
TSQA Average	0.3850	0.3900	0.5775	0.6075	0.4700	**0.5825**
MMTS	**0.4493**	0.3551	0.3913	0.4203	0.4275	0.3913

批大小的影响

表 4.13　不同批大小训练的模型在 TSQA 与 MMTS 测试集上的表现

Batch Size	4	8	16	32	64	128
Trend	**0.84**	0.71	0.52	0.34	0.60	0.62
Volatility	**0.75**	0.70	0.65	0.33	0.38	0.33
Seasonality	**0.63**	0.62	0.52	0.39	0.46	0.39
Outliers	**0.75**	0.43	0.67	0.41	0.44	0.40
TSQA Average	**0.7425**	0.6140	0.5900	0.3675	0.4700	0.4350
MMTS	0.4348	0.4058	0.4275	0.4275	0.4420	**0.4565**

附录

训练轮次的影响

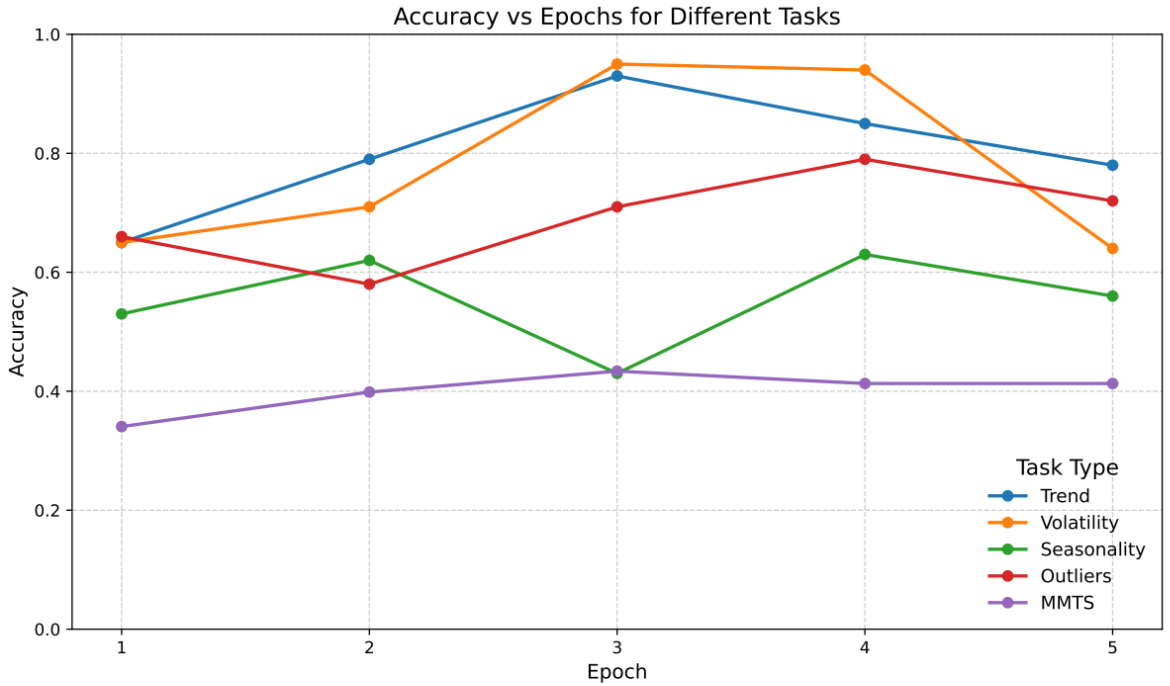


图 4.9　不同训练轮次下的模型在不同任务上的正确率折线图

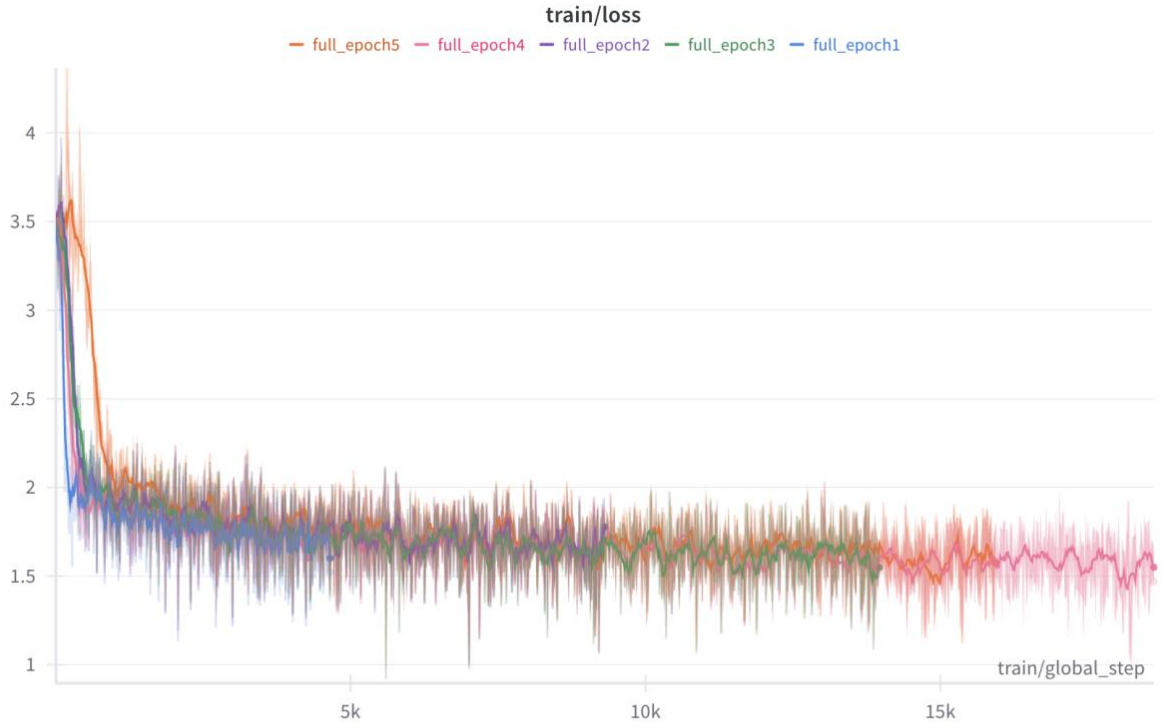
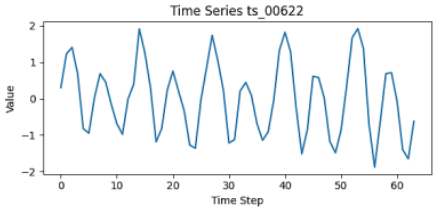


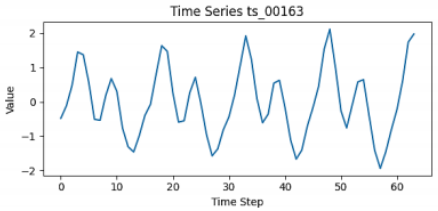
图 4.10　不同训练轮次下的模型训练过程损失曲线

附录

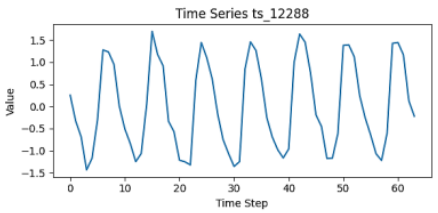
更多案例研究



The time series plot shows a repeating pattern with peaks and troughs, suggesting periodic behavior. The amplitude of the oscillations appears to increase over time, indicating a possible trend component combined with cyclical variations.

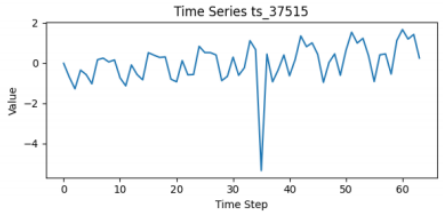


The time series shows a pattern of fluctuations with peaks and troughs occurring at regular intervals. It appears to have a cyclical component, suggesting periodic behavior. The amplitude of the oscillations varies slightly over time, indicating some variability in the cycle's strength.

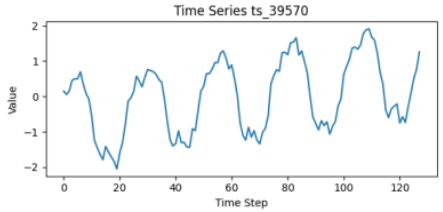


The time series exhibits a clear cyclical pattern with regular peaks and troughs, indicating periodic behavior. It starts at a high point, decreases to a low point, then rises again in a smooth, symmetrical manner. This suggests a sinusoidal or wave-like pattern, typical of phenomena like seasonal effects or oscillations.

图 B.7　多模态模型针对时间序列的整体描述样例



The time series appears to have a significant outlier around the 30th data point, where the value sharply decreases from approximately -2.5 to -4. This drop is more pronounced compared to other fluctuations in the series.



The time series plot shows a clear pattern with regular peaks and troughs, suggesting a periodic or seasonal component. There do not appear to be any significant outliers, as the data points consistently follow the established pattern without any abrupt deviations from it.

图 B.8　多模态模型针对时间序列的异常值描述样例

大量的案例研究表明本工作的模型在时间序列分析理解上已初步具备能力，然而同样也暴露出了众多问题。模型在对时间序列进行通用描述时总偏向于描述其周期性、季节性性质，揭示了训练数据存在不平衡的问题。此外，模型在数值理解与索引定位上存在较大的幻觉，可能是由于训练数据中缺乏准确的界定与引导。此外，模型在一些时间序列的分析描述上会给出完全错误的结果。