THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Time-VLM: Exploring Multimodal Vision-Language Models for Augmented Time Series Forecasting

Siru Zhong, Yuxuan Liang* et al.

ICML 2025

June 2025

Background

- **Time series forecasting (TSF)** plays a pivotal role across numerous domains.
- Traditional deep learning models (LSTM, RNN, Transformer) are widely used but limited in generalizing across diverse time series domains.
- Recent advancements have explored augmenting models with text or vision modalities to improve accuracy.
  - **Text**: provide contextual understanding
  - **Vision**: captures intricate temporal patterns

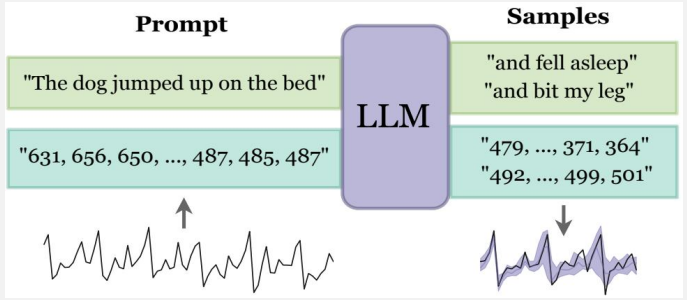
Related Work — Text-Augmented Models

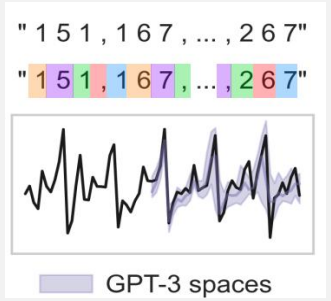
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Textual data, such as task specific knowledge, provides valuable context for TSF.
- With the development of NLP, researchers have tried to transfer the good few-shot / zero-shot capabilities of LLMs to TSF.
- Recent methods can be divided into two paradigms.

No-tuning based (LLMTime)

directly encode numerical data as text, use pre-trained LLMs without fine-tuning or training.

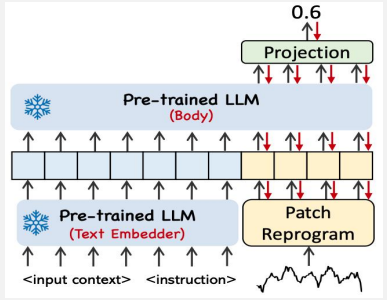


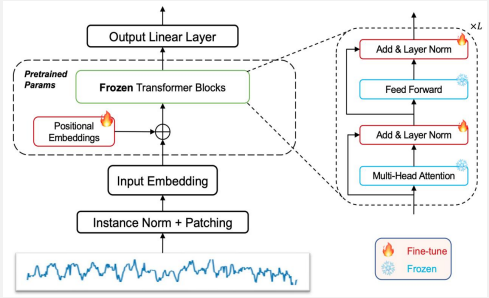


Gruver, Nate, et al. “Large language models are zero-shot time series forecasters.” NeurIPS (2024).

Tuning based (Time-LLM, GPT4TS)

align time series with LLM embeddings, need fine-tuning adaptation layer or LLM itself.



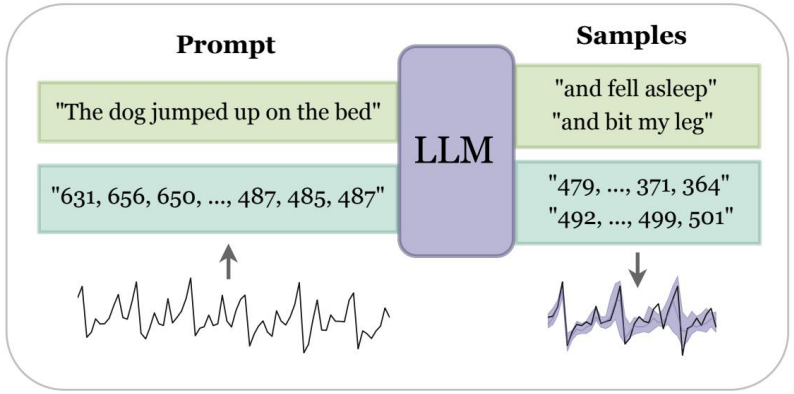


Jin, Ming, et al. "Time-llm: Time series forecasting by reprogramming large language models." ICLR (2024).
Zhou, Tian, et al. "One fits all: Power general time series analysis by pretrained lm." NeurIPS (2023)

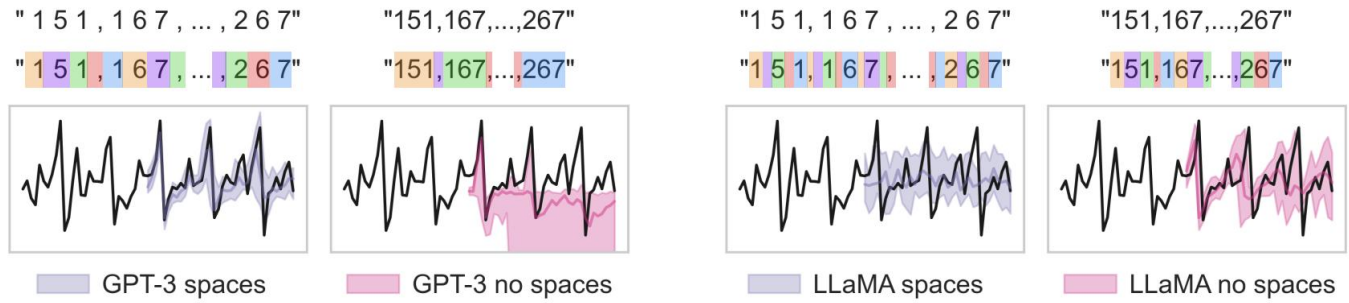
Related Work — Text-Augmented Models

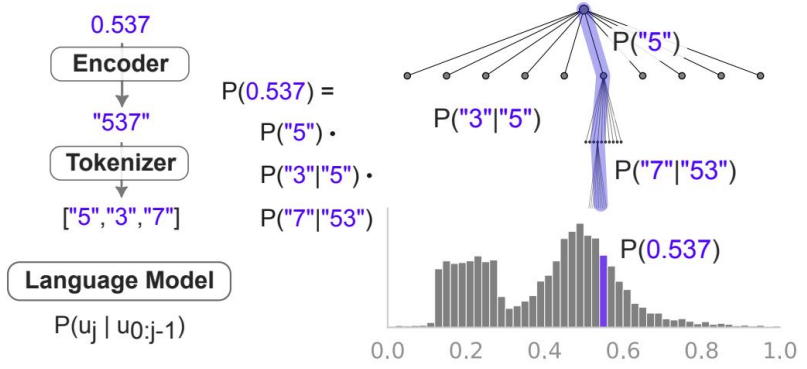
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

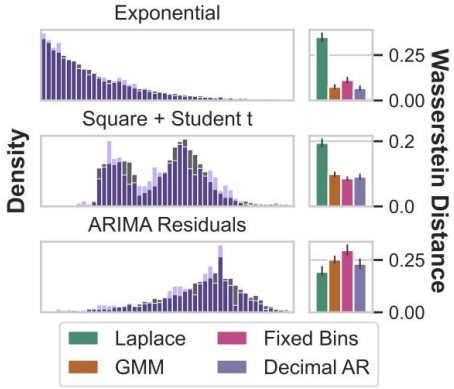
- **No-tuning based method**: methods like LLMTime directly encode numerical data as text, use pre-trained LLMs without fine-tuning or training.



1. LLM's restricted context windows
2. LLM's weakness in arithmetic and recursive operations





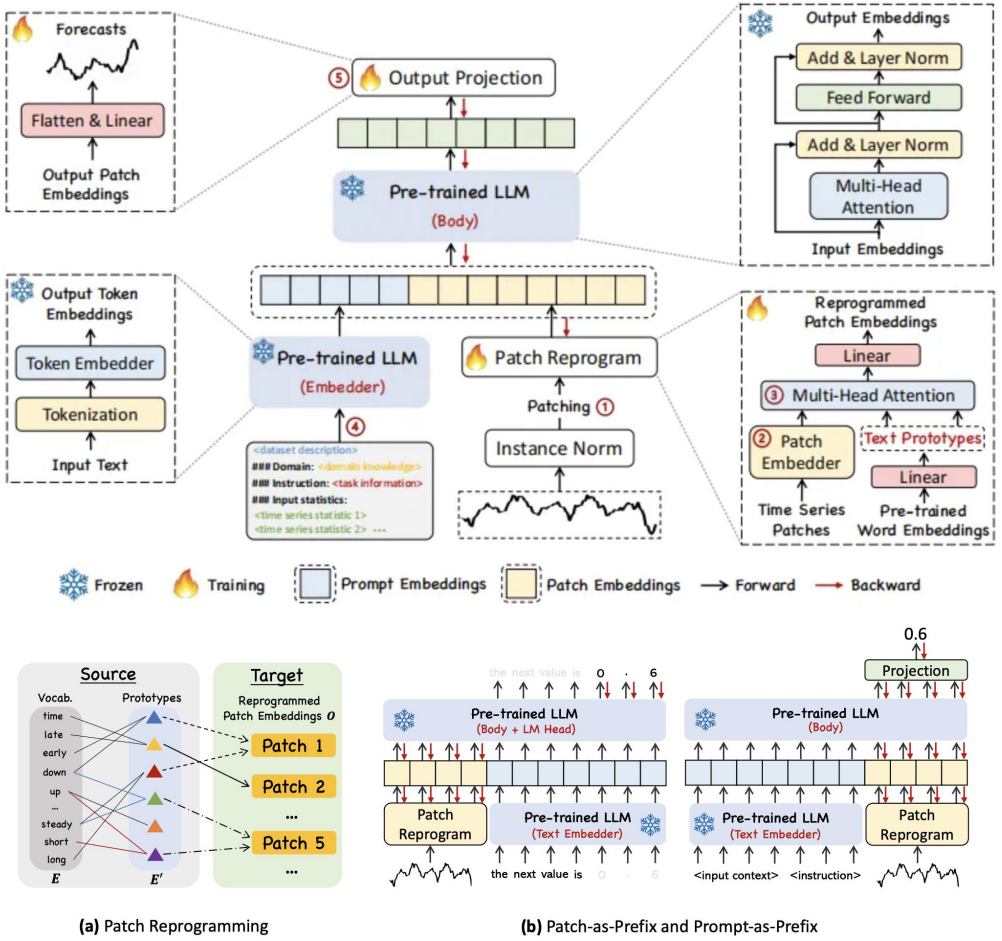


Gruver, Nate, et al. "Large language models are zero-shot time series forecasters." NeurIPS (2024).

Related Work — Text-Augmented Models

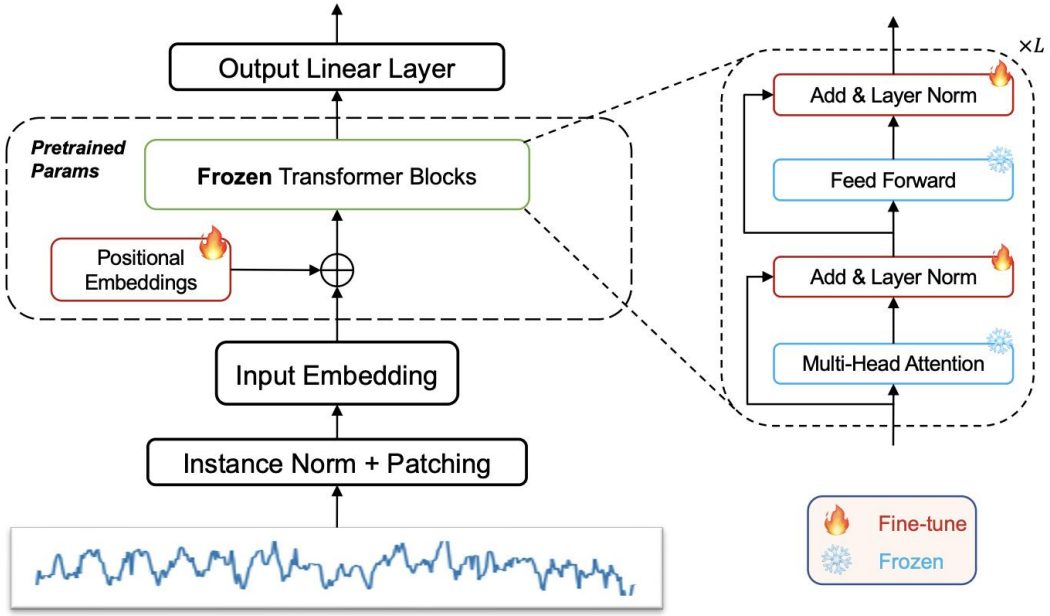
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **Tuning based method**: methods like Time-LLM and GPT4TS align time series with LLM embeddings, need fine-tuning adaptation layer or LLM itself.



(a) Patch Reprogramming

(b) Patch-as-Prefix and Prompt-as-Prefix



1. Modality gap between time series and text
2. Pre-trained word-embeddings are rare for TSF

Jin, Ming, et al. "Time-llm: Time series forecasting by reprogramming large language models." ICLR (2024).
Zhou, Tian, et al. "One fits all: Power general time series analysis by pretrained lm." NeurIPS (2023)

Related Work — Vision-Augmented Models

- Visual data is more similar to time series and can well reflect its characteristics.
- With the development of CV, researchers have tried to transform time series into visual representations, enables vision models to identify and exploit underlying patterns.

	Characteristics	Origin	Information
Time series	continuous	physical systems	high redundancy
Image	continuous	physical systems	high redundancy
Text	discrete	human cognitive construct	semantically dense

Related Work — Vision-Augmented Models

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Early approaches explore converting time series into different types of images, and then using a visual model for feature extraction and prediction.

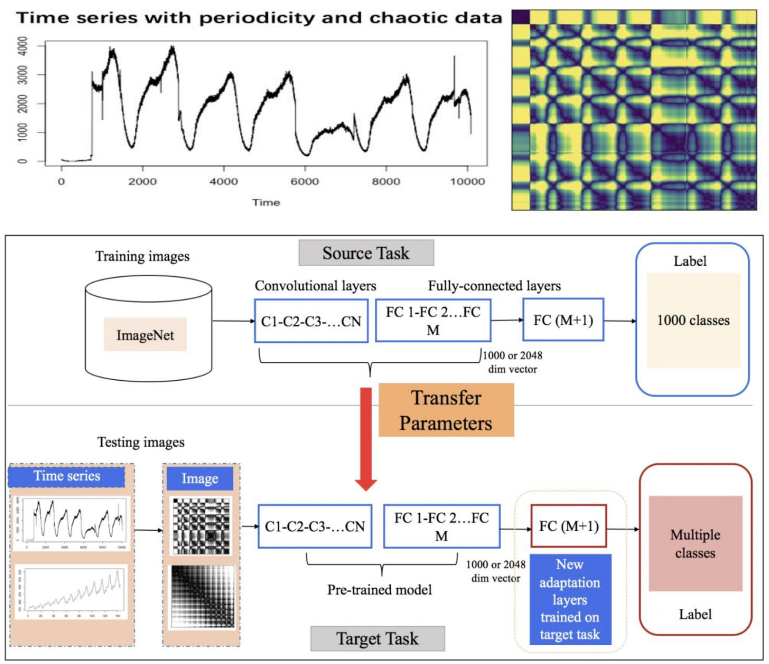


Image: Recurrence plot
Vision model: Convolutional Neural Network
Loss function: OWA (MASE + sMAPE)

Forecasting with time series imaging, Expert Systems with Applications 2020

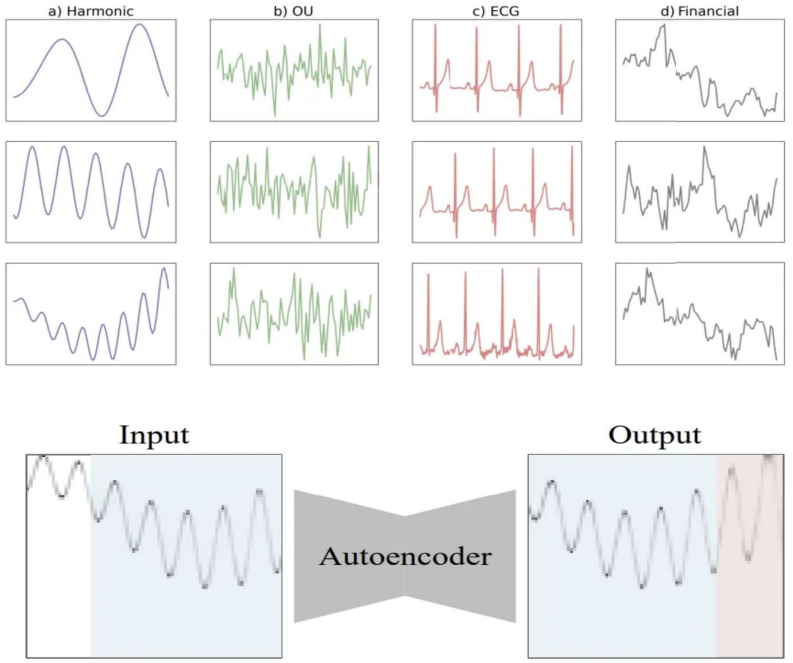


Image: X-time points, Y-normalized values.
Vision model: Convolutional Autoencoder
Loss function: Jensen-Shannon divergence

Visual Time Series Forecasting: An Image-driven Approach, ICAIF 2021

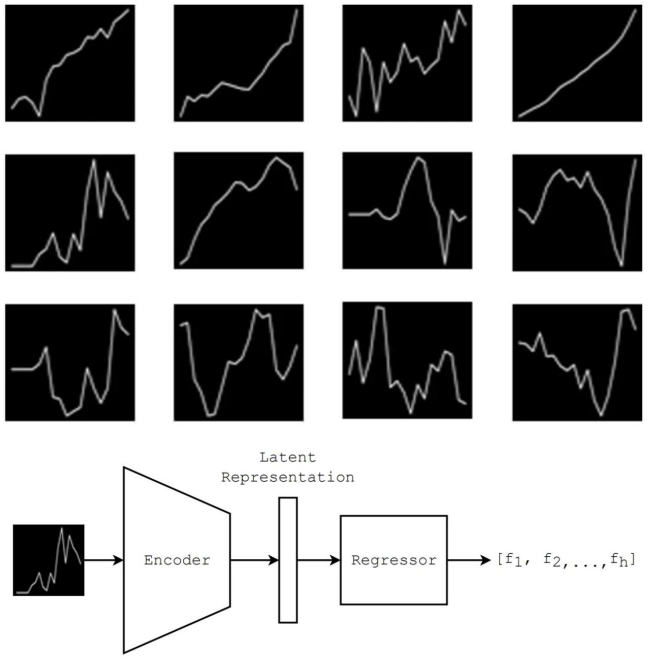


Image: Line graph
Vision model: Convolutional Neural Network
Loss function: MSE

Image-based time series forecasting: A deep convolutional neural network approach, Neural Networks 2023

Related Work — Vision-Augmented Models

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **TimeNet** advances the field with multi-scale frequency-based time image transformations.

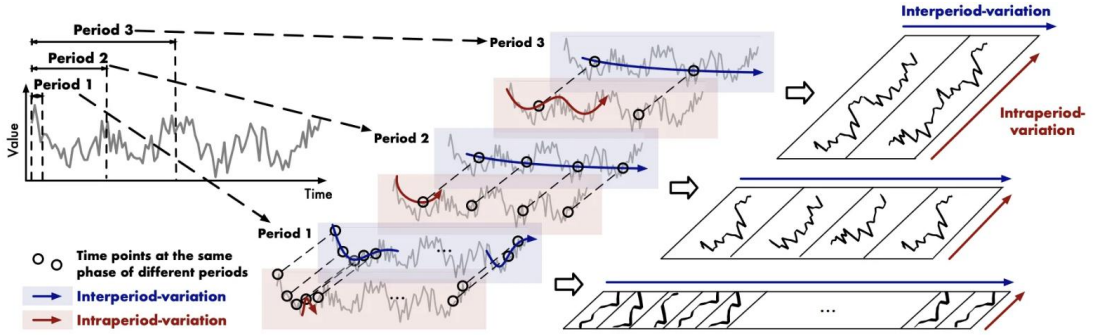


Figure 1: Multi-periodicity and temporal 2D-variation of time series. Each period involves the intraperiod-variation and interperiod-variation. We transform the original 1D time series into a set of 2D tensors based on multiple periods, which can unify the intraperiod- and interperiod-variations.

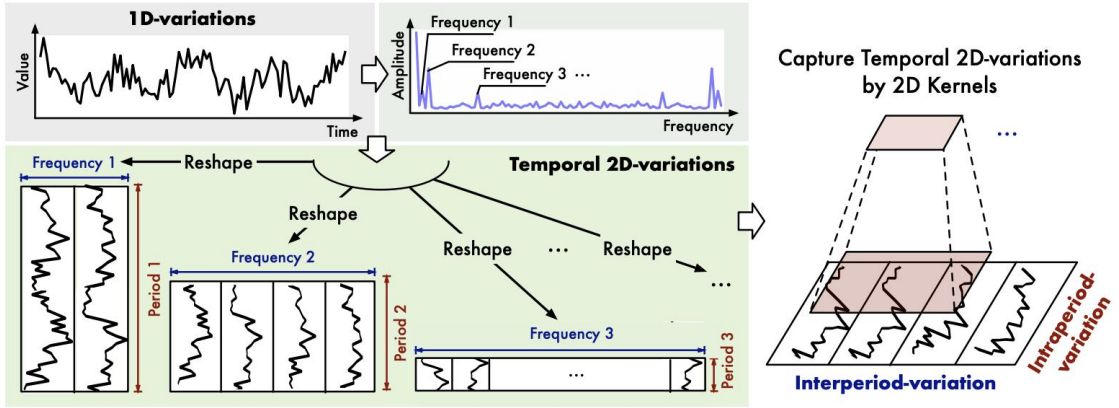


Figure 2: A univariate example to illustrate 2D structure in time series. By discovering the periodicity, we can transform the original 1D time series into structured 2D tensors, which can be processed by 2D kernels conveniently. By conducting the same reshape operation to all variates of time series, we can extend the above process to multivariate time series.

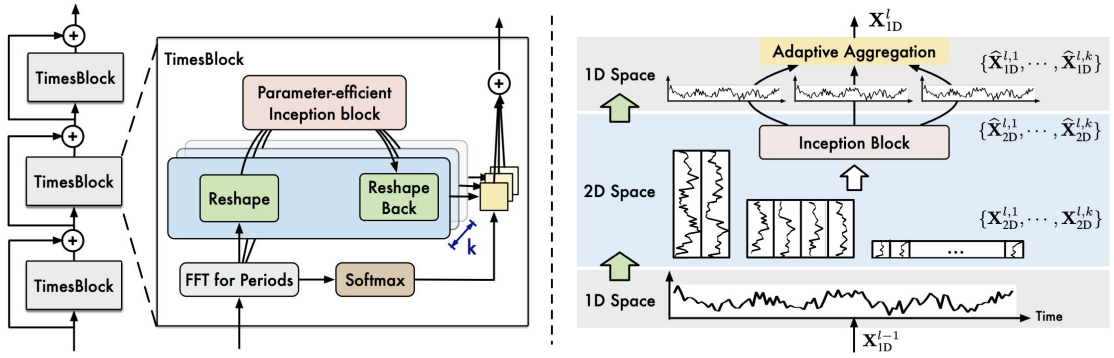


Figure 3: Overall architecture of TimesNet. TimesNet is stacked by TimesBlocks in a residual way. TimesBlocks can capture various temporal 2D-variations from k different reshaped tensors by a parameter-efficient inception block in 2D space and fuse them based on normalized amplitude values.

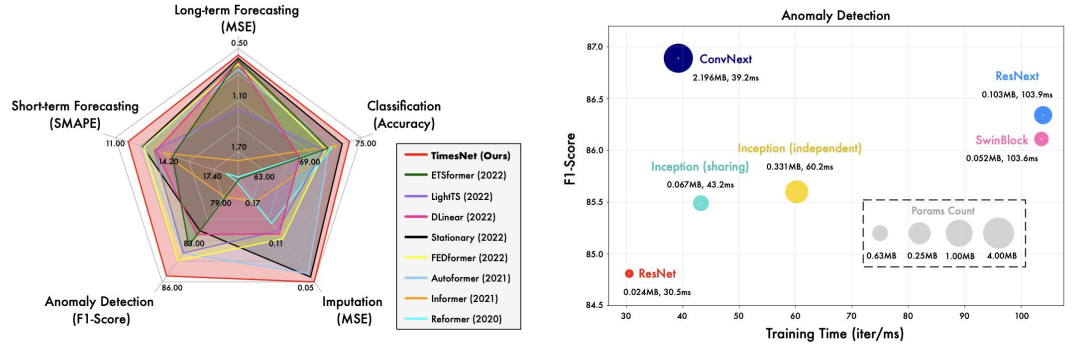


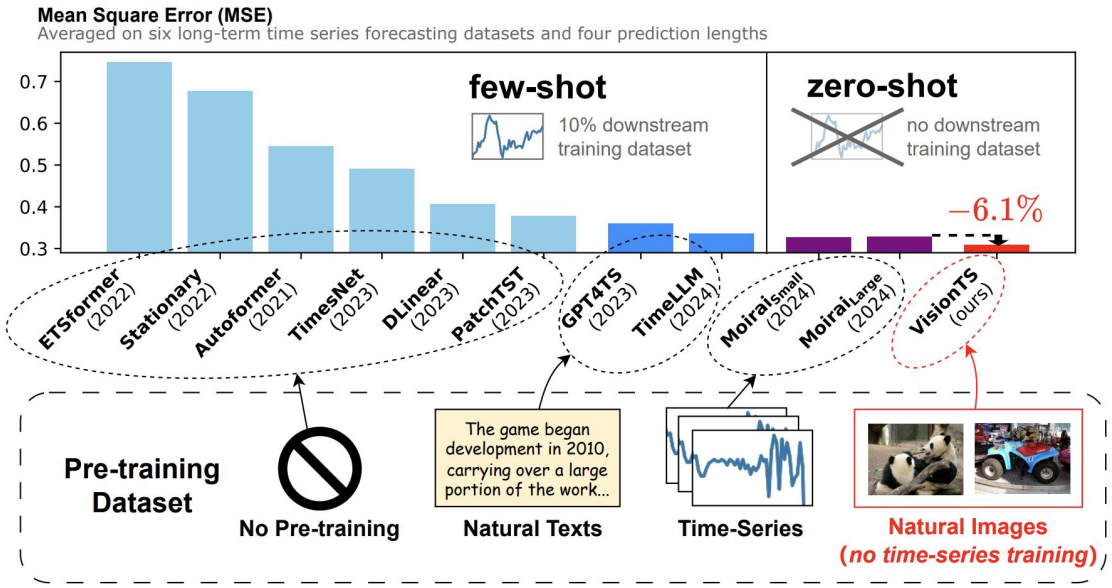
Figure 4: Model performance comparison (left) and generality in different vision backbones (right).

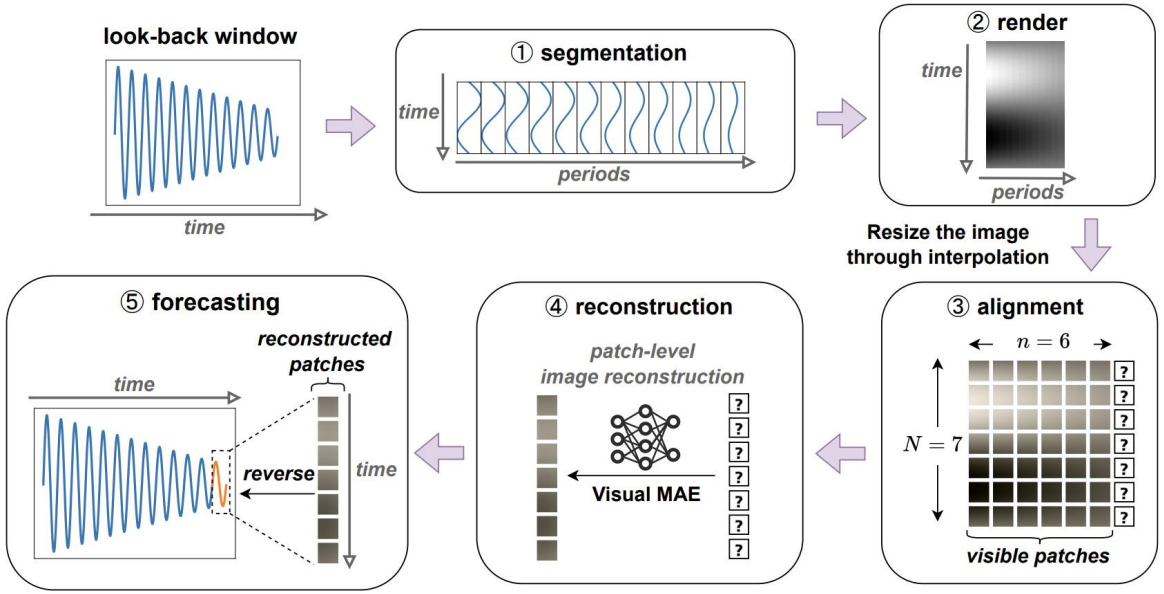
Wu, Haixu, et al. "Timesnet: Temporal 2d-variation modeling for general time series analysis." ICLR (2022).

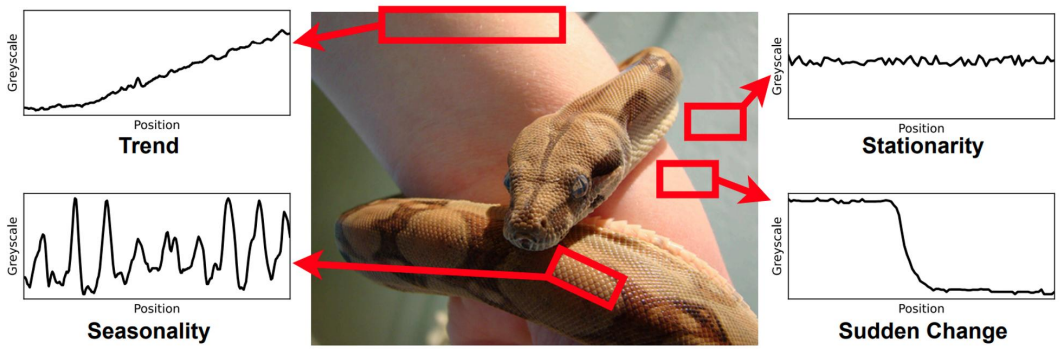
Related Work — Vision-Augmented Models

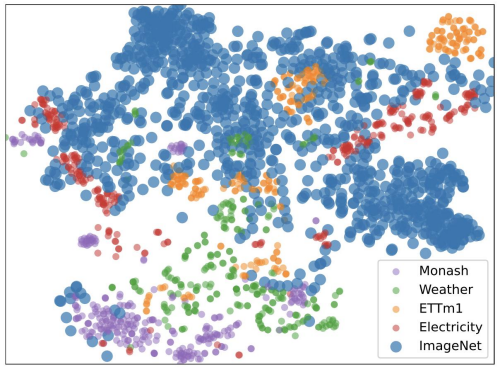
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- **VisionTS** pioneers pre-trained visual encoders with grayscale time series images.









Chen, Mouxiang, et al. "VisionTS: Visual Masked Autoencoders Are Free-Lunch Zero-Shot Time Series Forecasters." arXiv (2024).

Motivation

- **Time-VLM** (Time-Vision-Language Model)
- Can we bridge temporal, visual, and textual modalities for enhanced TSF?

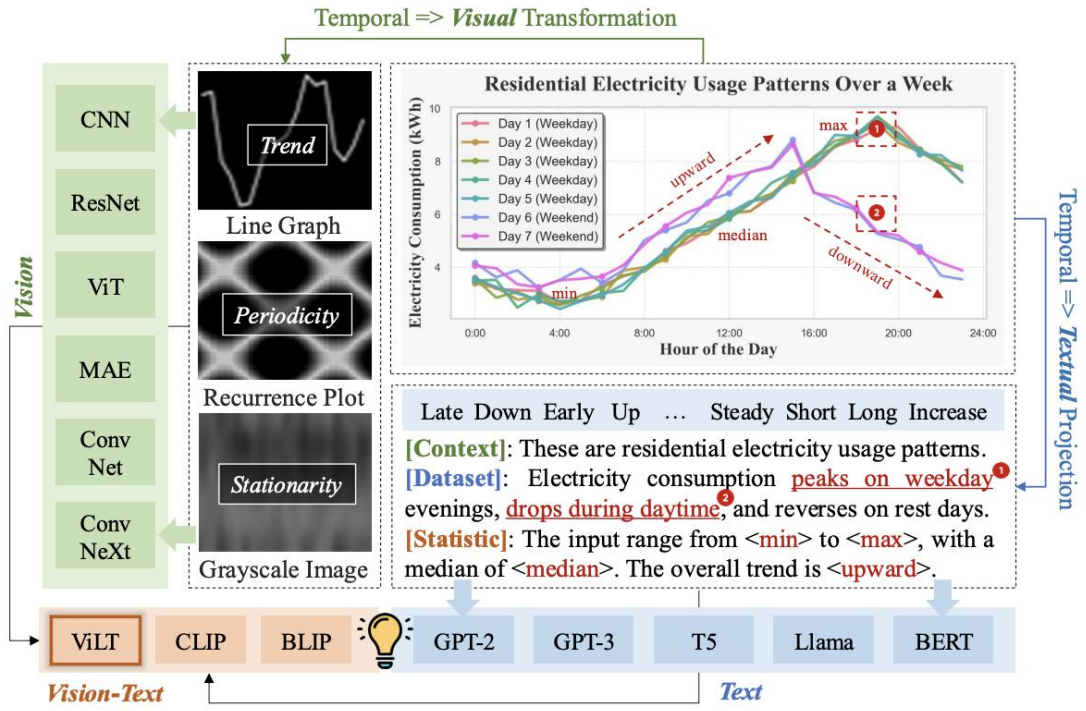


Figure 1: Our Time-VLM combines text (Right) and vision (Left) modalities to augment time series forecasting.

Methodology

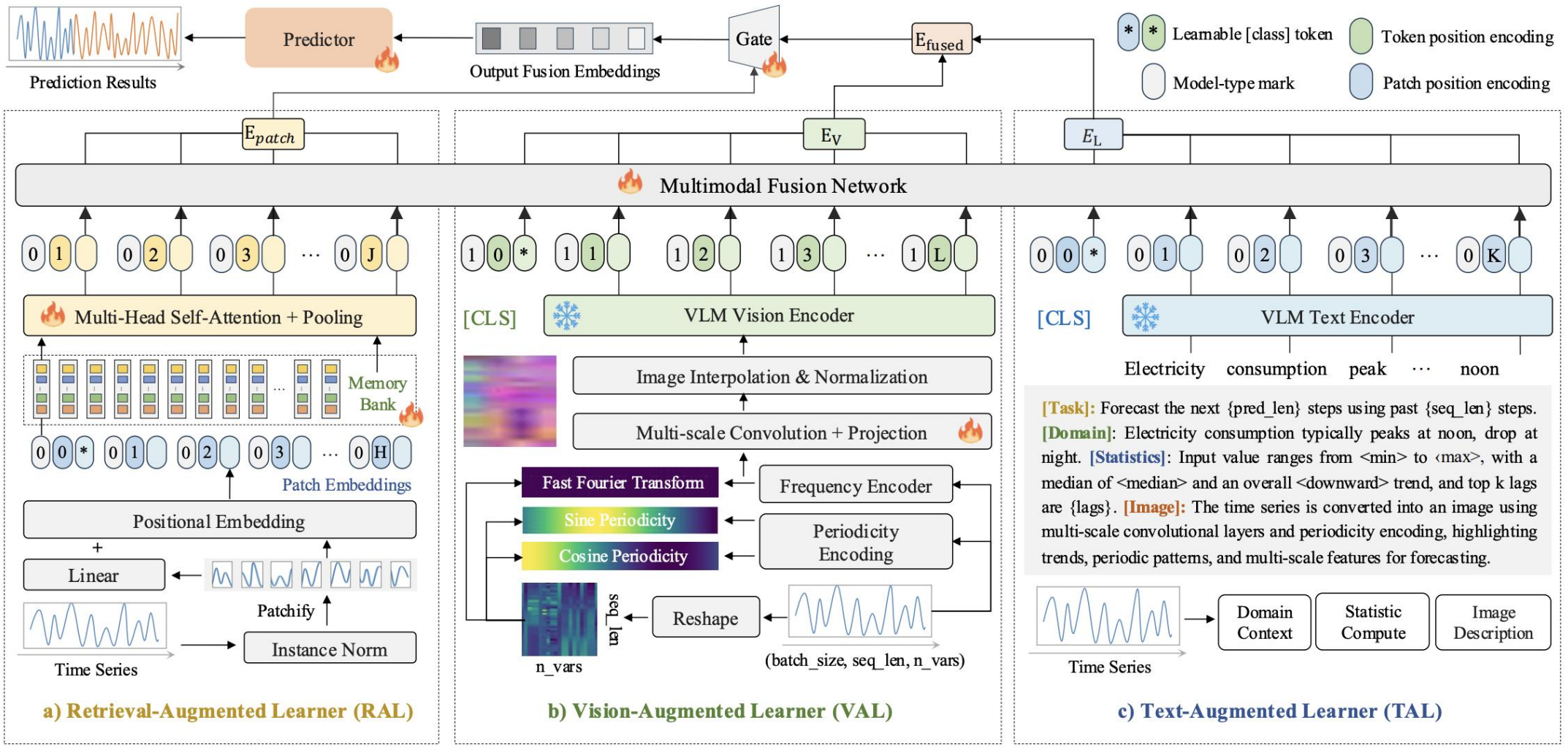


Figure 2: Overview of the Time-VLM framework.

Experiments

- Dataset Details

Table 8: Summary of benchmark datasets. Each dataset includes multiple time series (Dim.) with varying sequence lengths, split into training, validation, and testing sets. Data are collected at different frequencies across various domains.

Tasks	Dataset	Dim.	Series Length	Dataset Size	Frequency	Domain	Periodicity
Long-term Forecasting	ETTm1	7	{96, 192, 336, 720}	(34465, 11521, 11521)	15 min	Temperature	96
	ETTm2	7	{96, 192, 336, 720}	(34465, 11521, 11521)	15 min	Temperature	96
	ETTh1	7	{96, 192, 336, 720}	(8545, 2881, 2881)	1 hour	Temperature	24
	ETTh2	7	{96, 192, 336, 720}	(8545, 2881, 2881)	1 hour	Temperature	24
	Electricity	321	{96, 192, 336, 720}	(18317, 2633, 5261)	1 hour	Electricity	24
	Traffic	862	{96, 192, 336, 720}	(12185, 1757, 3509)	1 hour	Transportation	24
	Weather	21	{96, 192, 336, 720}	(36792, 5271, 10540)	10 min	Weather	144
Short-term Forecasting	M4 - Yearly	1	6	(23000, 0, 23000)	Yearly	Demographic	1
	M4 - Quarterly	1	8	(24000, 0, 24000)	Quarterly	Finance	4
	M4 - Monthly	1	18	(48000, 0, 48000)	Monthly	Industry	3
	M4 - Weekly	1	13	(359, 0, 359)	Weekly	Macro	4
	M4 - Daily	1	14	(4227, 0, 4227)	Daily	Micro	1
	M4 - Hourly	1	48	(414, 0, 414)	Hourly	Other	24

Experiments

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Optimization Settings

Table 9: Default Model Architecture Parameters

Parameter	Default Value	Description
image_size	64	Size of generated image representation
d_model	128	Dimension of hidden embeddings
d_fusion	256	Dimension of gated fusion module
num_workers	32	Number of data loader workers
e_layers	2	Number of encoder layers
d_layers	1	Number of decoder layers
dropout	0.1	Dropout rate
vlm_fused_len	156	Token length of VLM multimodal embedding
vlm_hidden_dim	768	Hidden dimension of VLM

Table 10: Default Training Parameters

Parameter	Default Value	Description
batch_size	32	Training batch size
learning_rate	0.001	Initial learning rate
training_epochs	10	Number of training epochs
patience	3	Early stopping patience
loss	MSE	Mean square error
seq_len	512	Input sequence length
c_out	7 (ETTh1/h2/m1/m2) 21 (Weather) 321 (Electricity) 862 (Traffic)	Output dimension (dataset-specific)
pred_len	96/192/336/720	Prediction length
periodicity	24 (ETTh1/h2/Electricity/Traffic) 96 (ETTm1/m2) 144 (Weather)	Dataset periodicity (dataset-specific)
norm_const	0.4	Normalization coefficient
patch_len	16	Patch length
padding	8	Padding length
stride	8	Stride length
num_queries	8	Number of learnable queries for temporal memory
n_heads	4	Number of attention heads

- Evaluation Metrics

$$\text{MSE} = \frac{1}{H}\sum_{h=1}^{T}(\mathbf{Y}_h - \hat{\mathbf{Y}}_h)^2, \qquad \text{MAE} = \frac{1}{H}\sum_{h=1}^{H}|\mathbf{Y}_h - \hat{\mathbf{Y}}_h|,$$

$$\text{SMAPE} = \frac{200}{H}\sum_{h=1}^{H}\frac{|\mathbf{Y}_h - \hat{\mathbf{Y}}_h|}{|\mathbf{Y}_h| + |\hat{\mathbf{Y}}_h|}, \qquad \text{MAPE} = \frac{100}{H}\sum_{h=1}^{H}\frac{|\mathbf{Y}_h - \hat{\mathbf{Y}}_h|}{|\mathbf{Y}_h|},$$

$$\text{MASE} = \frac{1}{H}\sum_{h=1}^{H}\frac{|\mathbf{Y}_h - \hat{\mathbf{Y}}_h|}{\frac{1}{H-s}\sum_{j=s+1}^{H}|\mathbf{Y}_j - \mathbf{Y}_{j-s}|}, \qquad \text{OWA} = \frac{1}{2}\left[\frac{\text{SMAPE}}{\text{SMAPE}_{\text{Naïve2}}} + \frac{\text{MASE}}{\text{MASE}_{\text{Naïve2}}}\right],$$

Experiments

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Few-shot learning results

Table 1: Few-shot learning on 5% training data. Results are averaged over forecasting horizons $H \in \{96, 192, 336, 720\}$. Lower values indicate better performance. Full results see Section B.1. **Red**: best, Blue: second best.

Methods	Time-VLM$_{143M}$ (**Ours**)		Time-LLM$_{3405M}$ (2024)		GPT4TS (2023)		DLinear (2023)		PatchTST (2023)		TimesNet (2023a)		FEDformer (2022)		Autoformer (2021)		Stationary (2022b)		ETSformer (2022)		LightTS (2022)		Informer (2021)		Reformer (2020)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh1$	**0.442**	**0.453**	0.627	0.543	0.681	0.560	0.750	0.611	0.694	0.569	0.925	0.647	0.658	0.562	0.722	0.598	0.943	0.646	1.189	0.839	1.451	0.903	1.225	0.817	1.241	0.835
$ETTh2$	**0.354**	**0.402**	0.382	0.418	0.400	0.433	0.694	0.577	0.827	0.615	0.439	0.448	0.463	0.454	0.441	0.457	0.470	0.489	0.809	0.681	3.206	1.268	3.922	1.653	3.527	1.472
$ETTm1$	**0.364**	**0.385**	0.425	0.434	0.472	0.450	0.400	0.417	0.526	0.476	0.717	0.561	0.730	0.592	0.796	0.620	0.857	0.598	1.125	0.782	1.123	0.765	1.163	0.791	1.264	0.826
$ETTm2$	**0.262**	**0.323**	0.274	**0.323**	0.308	0.346	0.399	0.426	0.314	0.352	0.344	0.372	0.381	0.404	0.388	0.433	0.341	0.372	0.534	0.547	1.415	0.871	3.658	1.489	3.581	1.487
$Weather$	**0.240**	**0.280**	0.260	0.309	0.263	0.301	0.263	0.308	0.269	0.303	0.298	0.318	0.309	0.353	0.310	0.353	0.327	0.328	0.333	0.371	0.305	0.345	0.584	0.527	0.447	0.453
ECL	0.218	0.315	0.179	**0.268**	**0.178**	0.273	0.176	0.275	0.181	0.277	0.402	0.453	0.266	0.353	0.346	0.404	0.627	0.603	0.800	0.685	0.878	0.725	1.281	0.929	1.289	0.904
$Traffic$	0.558	0.410	0.423	0.298	0.434	0.305	0.450	0.317	**0.418**	**0.296**	0.867	0.493	0.676	0.423	0.833	0.502	1.526	0.839	1.859	0.927	1.557	0.795	1.591	0.832	1.618	0.851

Table 2: Few-shot learning on 10% training data. We use the same protocol in Table 1. Full results see Section B.1.

Methods	Time-VLM$_{143M}$ (**Ours**)		Time-LLM$_{3405M}$ (2024)		GPT4TS (2023)		DLinear (2023)		PatchTST (2023)		TimesNet (2023a)		FEDformer (2022)		Autoformer (2021)		Stationary (2022b)		ETSformer (2022)		LightTS (2022)		Informer (2021)		Reformer (2020)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh1$	**0.431**	**0.442**	0.556	0.522	0.590	0.525	0.691	0.600	0.633	0.542	0.869	0.628	0.639	0.561	0.702	0.596	0.915	0.639	1.180	0.834	1.375	0.877	1.199	0.809	1.249	0.833
$ETTh2$	**0.361**	0.405	0.370	**0.394**	0.397	0.421	0.605	0.538	0.415	0.431	0.479	0.465	0.466	0.475	0.488	0.499	0.462	0.455	0.894	0.713	2.655	1.160	3.872	1.513	3.485	1.486
$ETTm1$	**0.360**	**0.382**	0.404	0.427	0.464	0.441	0.411	0.429	0.501	0.466	0.677	0.537	0.722	0.605	0.802	0.628	0.797	0.578	0.980	0.714	0.971	0.705	1.192	0.821	1.426	0.856
$ETTm2$	**0.263**	**0.323**	0.277	**0.323**	0.293	0.335	0.316	0.368	0.296	0.343	0.320	0.353	0.463	0.488	1.342	0.930	0.332	0.366	0.447	0.487	0.987	0.756	3.370	1.440	3.978	1.587
$Weather$	**0.233**	0.274	0.234	**0.273**	0.238	0.275	0.241	0.283	0.242	0.279	0.279	0.301	0.284	0.324	0.300	0.342	0.318	0.323	0.318	0.360	0.289	0.322	0.597	0.495	0.546	0.469
ECL	0.198	0.291	**0.175**	0.270	0.176	**0.269**	0.180	0.280	0.180	0.273	0.323	0.392	0.346	0.427	0.431	0.478	0.444	0.480	0.660	0.617	0.441	0.489	1.195	0.891	0.965	0.768
$Traffic$	0.484	0.357	**0.429**	0.306	0.440	0.310	0.447	0.313	0.430	**0.305**	0.951	0.535	0.663	0.425	0.749	0.446	1.453	0.815	1.914	0.936	1.248	0.684	1.534	0.811	1.551	0.821

- Zero-shot learning results

Table 3: Zero-shot learning results. Full results see Section B.2.

Methods	Time-VLM $_{143M}$ (**Ours**)		Time-LLM$_{3405M}$ (2024)		LLMTime (2023)		GPT4TS (2023)		DLinear (2023)		PatchTST (2023)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
$ETTh1 \rightarrow ETTh2$	**0.338**	**0.385**	0.353	0.387	0.992	0.708	0.406	0.422	0.493	0.488	0.380	0.405
$ETTh1 \rightarrow ETTm2$	0.293	0.350	**0.273**	**0.340**	1.867	0.869	0.325	0.363	0.415	0.452	0.314	0.360
$ETTh2 \rightarrow ETTh1$	0.496	0.480	**0.479**	**0.474**	1.961	0.981	0.757	0.578	0.703	0.574	0.565	0.513
$ETTh2 \rightarrow ETTm2$	0.297	0.353	**0.272**	**0.341**	1.867	0.869	0.335	0.370	0.328	0.386	0.325	0.365
$ETTm1 \rightarrow ETTh2$	**0.354**	**0.397**	0.381	0.412	0.992	0.708	0.433	0.439	0.464	0.475	0.439	0.438
$ETTm1 \rightarrow ETTm2$	**0.264**	**0.319**	0.268	0.320	1.867	0.869	0.313	0.348	0.335	0.389	0.296	0.334
$ETTm2 \rightarrow ETTh2$	0.359	**0.399**	**0.354**	0.400	0.992	0.708	0.435	0.443	0.455	0.471	0.409	0.425
$ETTm2 \rightarrow ETTm1$	0.432	**0.426**	**0.414**	0.438	1.933	0.984	0.769	0.567	0.649	0.537	0.568	0.492

Experiments

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Full-shot learning results

Table 4: Short-term time series forecasting results on M4. The forecasting horizons are in [6, 48] and the three rows provided are weighted averaged from all datasets under different sampling intervals. Full results see Section B.3.

Methods		Time-VLM$_{143M}$ (**Ours**)	Time-LLM$_{3405M}$ (2024)	GPT4TS (2023)	TimesNet (2023a)	PatchTST (2023)	N-HiTS (2023)	N-BEATS (2020)	ETSformer (2022)	LightTS (2022)	DLinear (2023)	FEDformer (2022)	Stationary (2022b)	Autoformer (2021)	Informer (2021)	Reformer (2020)
Average	SMAPE	**11.894**	11.983	12.690	12.880	12.059	12.035	12.250	14.718	13.525	13.639	13.160	12.780	12.909	14.086	18.200
	MASE	**1.592**	1.595	1.808	1.836	1.623	1.625	1.698	2.408	2.111	2.095	1.775	1.756	1.771	2.718	4.223
	OWA	**0.855**	0.859	0.940	0.955	0.869	0.869	0.896	1.172	1.051	1.051	0.949	0.930	0.939	1.230	1.775

Table 5: Long-term forecasting results. We use the same protocol in Table 1. Full results see in Section B.4.

Methods		Time-VLM$_{143M}$ (**Ours**)		Time-LLM$_{3405M}$ (2024)		GPT4TS (2023)		DLinear (2023)		PatchTST (2023)		TimesNet (2023a)		FEDformer (2022)		Autoformer (2021)		Stationary (2022b)		ETSformer (2022)		LightTS (2022)		Informer (2021)		Reformer (2020)	
Metric		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1		**0.405**	**0.420**	0.408	0.423	0.465	0.455	0.422	0.437	0.413	0.430	0.458	0.450	0.440	0.460	0.496	0.487	0.570	0.537	0.542	0.510	0.491	0.479	1.040	0.795	1.029	0.805
ETTh2		0.341	0.391	0.334	0.383	0.381	0.412	0.431	0.446	**0.330**	**0.379**	0.414	0.427	0.437	0.449	0.450	0.459	0.526	0.516	0.439	0.452	0.602	0.543	4.431	1.729	6.736	2.191
ETTm1		0.347	0.377	**0.329**	**0.372**	0.388	0.403	0.357	0.378	0.351	0.380	0.400	0.406	0.448	0.452	0.588	0.517	0.481	0.456	0.429	0.425	0.435	0.437	0.961	0.734	0.799	0.671
ETTm2		**0.248**	**0.311**	0.251	0.313	0.284	0.339	0.267	0.333	0.255	0.315	0.291	0.333	0.305	0.349	0.327	0.371	0.306	0.347	0.293	0.342	0.409	0.436	1.410	0.810	1.479	0.915
Weather		**0.224**	0.263	0.225	**0.257**	0.237	0.270	0.248	0.300	0.225	0.264	0.259	0.287	0.309	0.360	0.338	0.382	0.288	0.314	0.271	0.334	0.261	0.312	0.634	0.548	0.803	0.656
Electricity		0.172	0.273	**0.158**	**0.252**	0.167	0.263	0.166	0.263	0.161	**0.252**	0.192	0.295	0.214	0.327	0.227	0.338	0.193	0.296	0.208	0.323	0.229	0.329	0.311	0.397	0.338	0.422
Traffic		0.419	0.303	**0.388**	0.264	0.414	0.294	0.433	0.295	0.390	**0.263**	0.620	0.336	0.610	0.376	0.628	0.379	0.624	0.340	0.621	0.396	0.622	0.392	0.764	0.416	0.741	0.422

Experiments

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- ## Ablation Studies

Table 6: Ablation study on multimodal components.

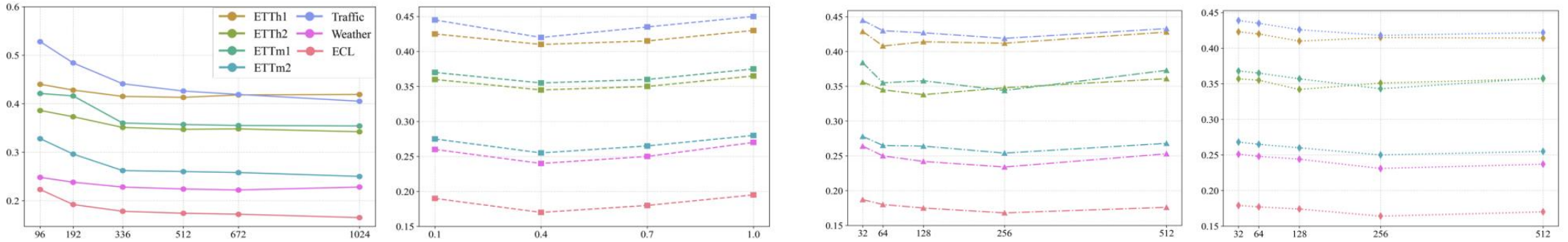
Horizon	Full		w/o RAL		w/o VAL		w/o TAL	
	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
96	**0.160**	**0.213**	0.273	0.324	0.213	0.266	0.165	0.218
192	**0.203**	**0.252**	0.297	0.338	0.237	0.298	0.208	0.257
336	**0.253**	**0.291**	0.325	0.354	0.255	0.302	0.258	0.295
720	**0.317**	**0.340**	0.369	0.383	0.309	0.357	0.322	0.345
Avg	**0.233**	**0.274**	0.316	0.350	0.254	0.306	0.238	0.279
%Deg	–	–	35.6% ↑	27.7% ↑	9.0% ↑	11.7% ↑	2.1% ↑	1.8% ↑

- ## Computation Studies

Table 7: Computational efficiency comparison between Time-VLM and Time-LLM across datasets. "-" denotes memory exceeds 49GB, infeasible on a single GPU. Results are averaged over multiple prediction steps under consistent conditions.

Method	Metric	ETTh1	ETTh2	ETTm1	ETTm2	Weather	ECL	Traffic
Time-VLM	Param. (M)	**143.6**	**143.6**	**143.6**	**143.6**	**143.6**	**143.6**	**143.6**
	Mem. (MiB)	**2630**	**2630**	**2640**	**2640**	**1968**	**10818**	**24916**
	Speed (s/iter)	**0.481**	**0.438**	**0.277**	**0.210**	**0.296**	**0.206**	**0.323**
Time-LLM	Param. (M)	3404.6	3404.6	3404.6	3404.6	-	-	-
	Mem. (MiB)	37723	37723	37849	37849	-	-	-
	Speed (s/iter)	0.607	0.553	0.349	0.265	-	-	-

- ## Hyperparameter Studies



a) Analysis on Input Length b) Analysis on Norm Constant c) Analysis on Model Dimension d) Analysis on Gate Network Dimension

Experiments

- Multimodal and Few/Zero-shot Analysis

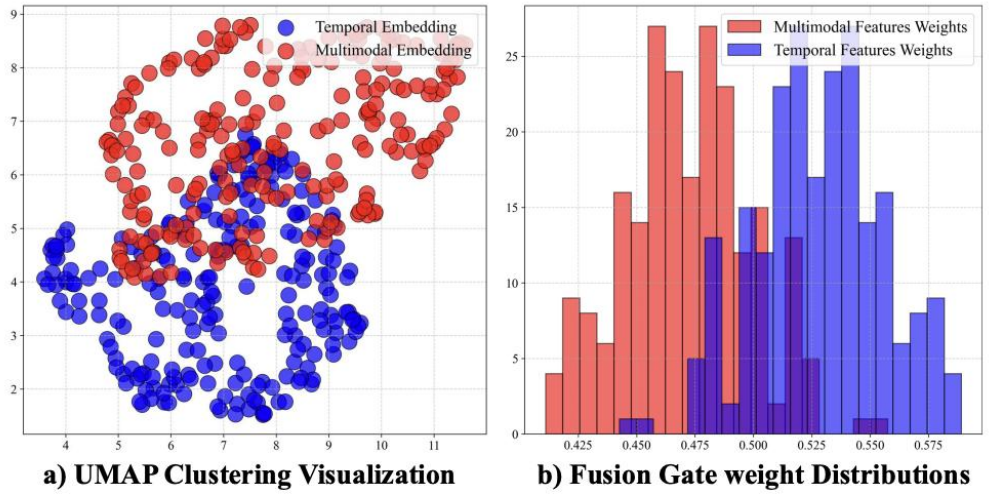


Figure 3: 2D UMAP visualization (Left) and Gate weight distributions (Right) of multimodal and temporal memory embeddings, highlighting their complementary behavior.

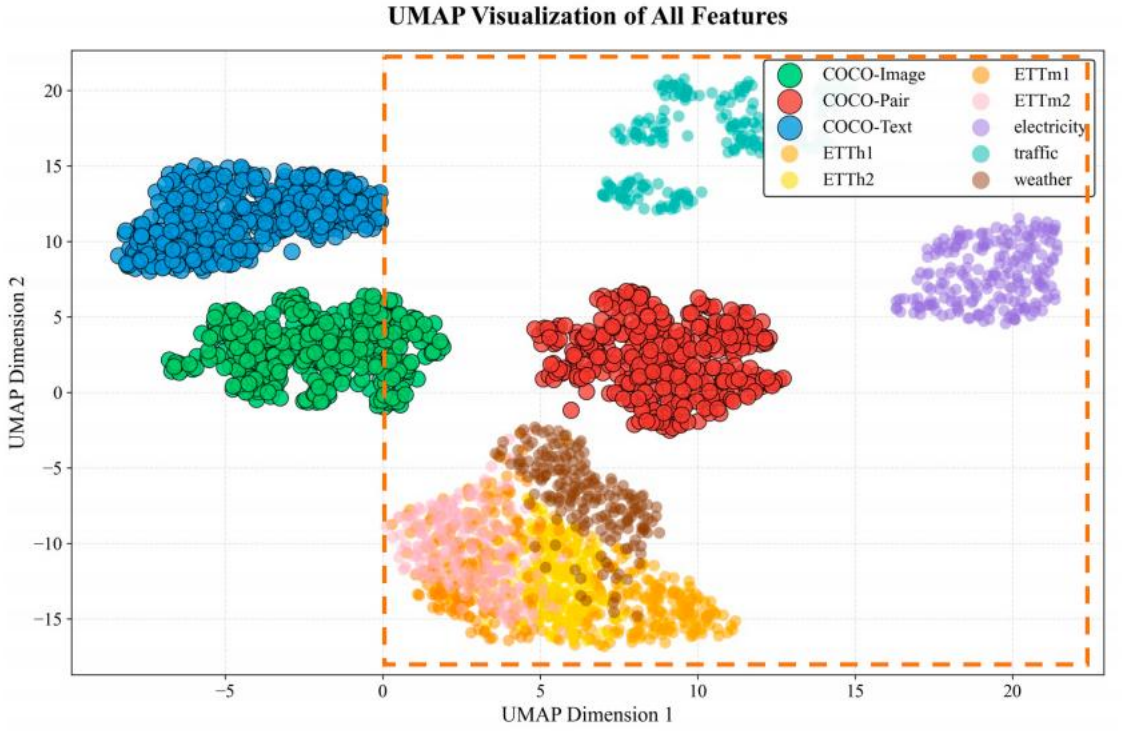


Figure 4: Interpretability visualization of Time-VLM: multimodal feature alignment via UMAP.

Experiments

- Visualization of Generated Time Series Images

Figure 5: Time series transformed images, capturing key temporal characteristics, including trends, stationarity, seasonality, sudden changes, and frequency-domain patterns.

Experiments

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Visualization of prediction results

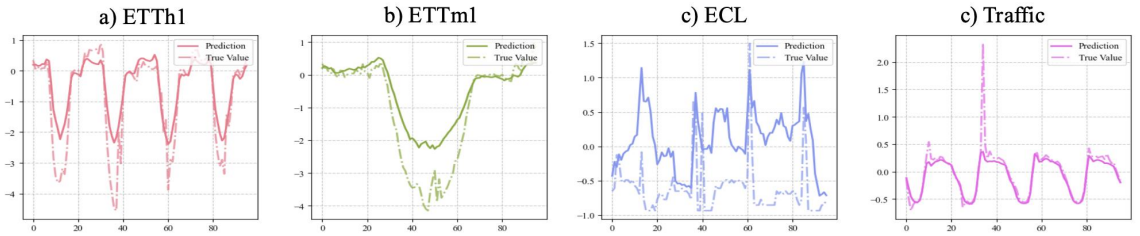


Figure 6: Prediction results visualization for ETTh1, ETTm1, ECL, and Traffic datasets at 96 prediction lengths. True values (solid line) and predicted values (dashed line) are shown for each dataset and horizon.

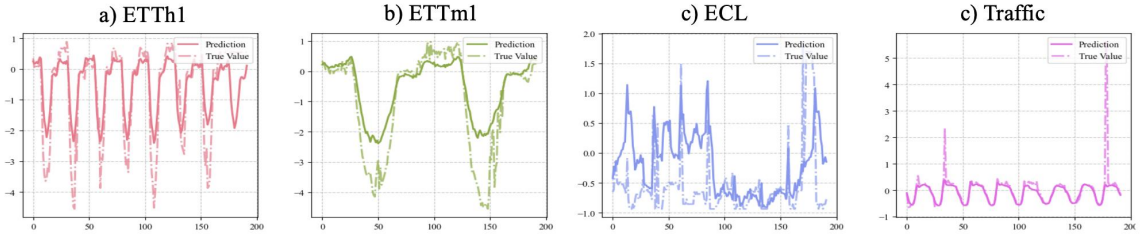


Figure 7: Prediction results visualization for ETTh1, ETTm1, ECL, and Traffic datasets at 192 prediction lengths. True values (solid line) and predicted values (dashed line) are shown for each dataset and horizon.

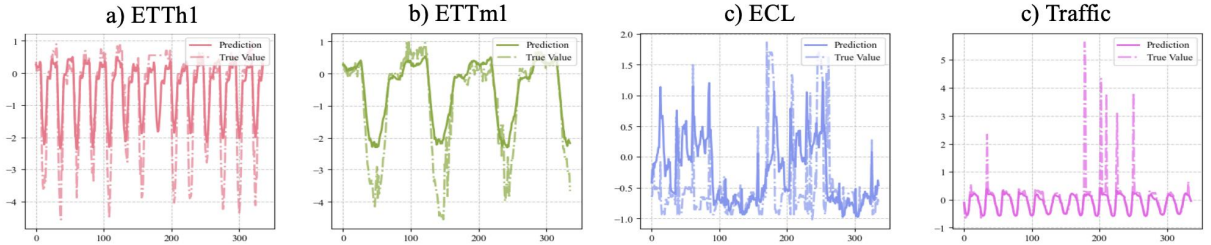


Figure 8: Prediction results visualization for ETTh1, ETTm1, ECL, and Traffic datasets at 336 prediction lengths. True values (solid line) and predicted values (dashed line) are shown for each dataset and horizon.

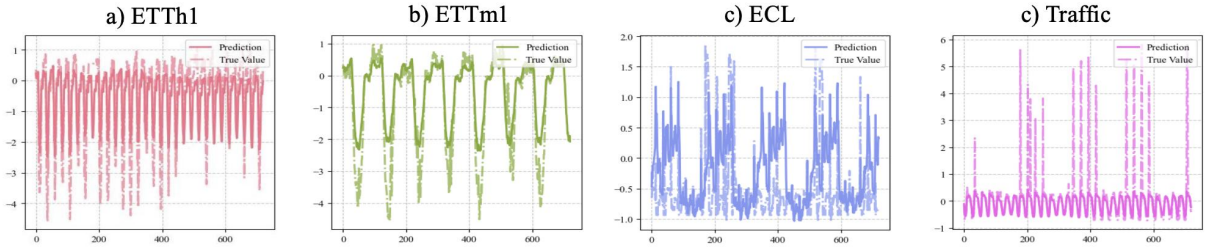


Figure 9: Prediction results visualization for ETTh1, ETTm1, ECL, and Traffic datasets at 720 prediction lengths. True values (solid line) and predicted values (dashed line) are shown for each dataset and horizon.

Conclusion

- Time-VLM is a novel framework leveraging pretrained VLMs to unify temporal, visual, and textual modalities for time series forecasting.
- It is self-enhanced, operating solely on original time series data without external information. By integrating the RAL, VAL, and TAL, Time-VLM bridges modality gaps, enabling rich cross-modal interactions.

Future Work

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Future Work:
  - High quality multi-modal time series benchmark => foundation model for multi-tasks.
  - Time series imaging — solving the problem of information density misalignment when different input step lengths are mapped to fixed resolution images.
  - Visual distillation — visual encoder may be redundant compared to time series.
  - Enhanced text encoder with time series semantic understanding capabilities.
  - Interpretability, modal contribution and analysis.

- VLM variants analysis.

Table 8: Comparison of different VLM variants on ETTh2 in terms of performance and computational efficiency.

VLM Type	Params (M)	Mem. (MiB)	Speed (s/iter)	MSE (avg)	MAE (avg)
ViLT	**128.9**	1346	0.36	**0.336**	**0.388**
CLIP	168.4	**1174**	**0.12**	0.339	0.391
BLIP-2	3763.1	25200	0.98	0.342	0.393
Custom	213.2	1474	0.17	0.348	0.397

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

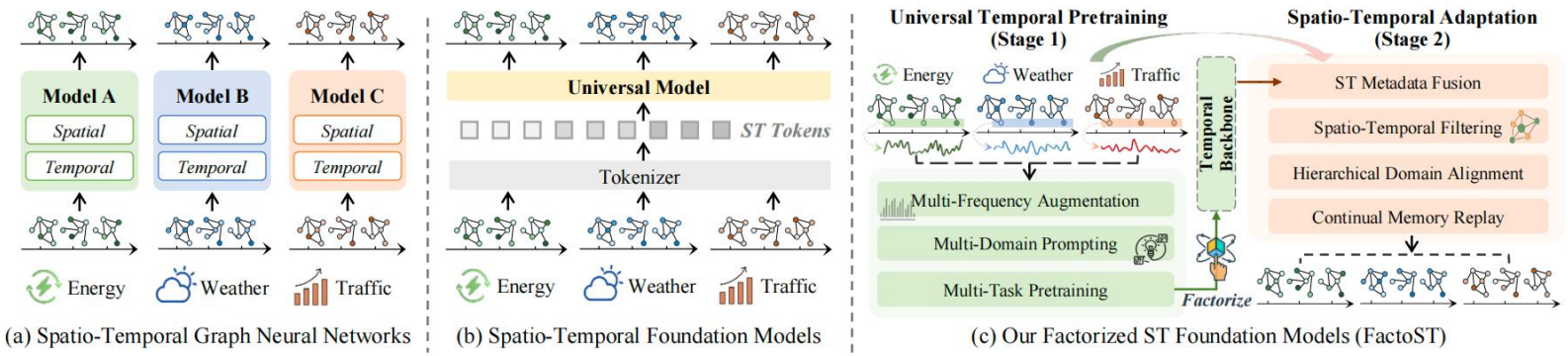
Thanks for listening!

June 2025

New work

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

- Motivation
  - 时间可以泛化，空间不可以泛化
  - 当前 STFM 预训练资源消耗大，
    且泛化性，性能在长期预测场景都不好



(a) Spatio-Temporal Graph Neural Networks (b) Spatio-Temporal Foundation Models (c) Our Factorized ST Foundation Models (FactoST)

Learning to Factorize Spatio-Temporal Foundation Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Spatio-Temporal Foundation Models (STFMs) promise zero/few-shot generalization across various datasets, yet joint spatio-temporal pretraining is computationally prohibitive and struggles with domain-specific spatial correlations. To this end, we introduce FactoST, a factorized STFM that decouples universal temporal pretraining from spatio-temporal adaptation. The first stage pretrains a space-agnostic backbone with multi-frequency reconstruction and domain-aware prompting, capturing cross-domain temporal regularities at low computational cost. The second stage freezes or further fine-tunes the backbone and attaches an adapter that fuses spatial metadata, sparsifies interactions, and aligns domains with continual memory replay. Extensive forecasting experiments reveal that, in few-shot setting, FactoST reduces MAE by up to 38.5% versus UniST, uses 46.2% fewer parameters, and achieves 68% faster inference than OpenCity, while remaining competitive with expert models. We believe this factorized view offers a practical and scalable path toward truly universal STFMs. The code will be released upon notification.

New work

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

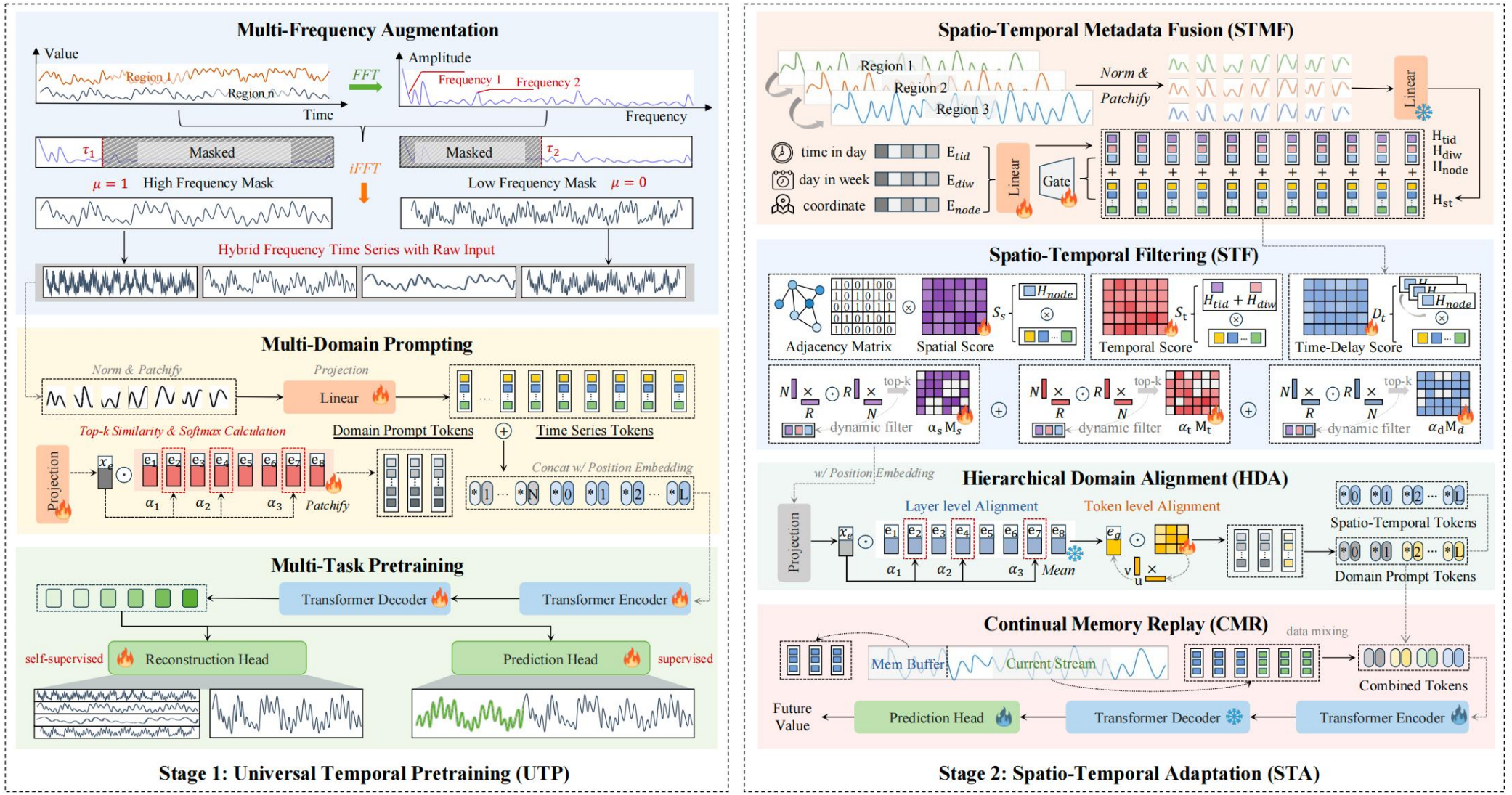


Figure 2: Overview of FactoST.

New work

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Table 2: Few-shot short-term forecasting (12→12) results on 10% training data across multiple ST datasets. Lower values indicate better performance. **Red**: the best. Blue: the second best.

Method Type		Foundation Model					Expert Model					
		Spatio Temporal			*Time Series*		*Spatio Temporal*			*Time Series*		
Method		FactoST	OpenCity	UniST	TimesFM	Moirai	STAEformer	STID	D2STGNN	PatchTST	DLinear	Informer
PEMS-03	MAE	**17.54**	17.90	40.39	21.99	21.40	30.79	22.93	18.55	21.97	21.94	23.24
	RMSE	**28.10**	28.80	53.44	35.31	32.38	47.67	34.10	29.21	35.59	35.30	37.98
PEMS-04	MAE	**23.93**	24.78	42.76	27.84	33.73	48.23	26.72	24.86	28.11	28.37	29.81
	RMSE	**37.44**	40.41	59.07	43.15	54.09	68.46	40.31	38.43	44.13	44.57	45.59
PEMS-07	MAE	26.48	44.43	40.77	32.61	35.69	33.50	31.46	**25.51**	31.19	31.89	37.55
	RMSE	41.92	65.47	54.86	50.20	51.36	51.43	46.72	**39.81**	48.91	49.65	62.55
PEMS-08	MAE	**18.94**	32.16	35.70	22.06	38.01	36.15	23.17	19.55	22.42	23.10	31.69
	RMSE	**29.59**	48.47	46.74	33.87	53.05	51.05	34.09	30.51	35.64	36.35	51.53
PEMS-Bay	MAE	**1.96**	2.77	5.14	2.25	2.26	2.01	2.00	1.99	2.15	2.21	2.96
	RMSE	**4.51**	6.08	8.28	5.49	5.49	4.62	4.57	4.72	5.23	5.20	6.23
METR-LA	MAE	4.77	4.18	8.79	5.56	4.95	4.61	**4.00**	**4.00**	4.34	4.57	4.93
	RMSE	9.88	8.33	14.34	12.87	12.75	8.91	8.20	**8.03**	9.75	9.82	9.20
ETTh2	MAE	0.272	0.513	0.425	0.284	**0.135**	1.208	0.756	0.916	0.721	1.885	2.125
	RMSE	0.424	0.710	0.545	0.410	**0.307**	1.673	1.224	1.433	1.211	2.946	2.898
Electricity	MAE	**0.374**	0.412	0.565	0.529	0.837	0.858	0.575	0.686	0.840	1.282	1.598
	RMSE	**0.545**	1.740	3.276	0.801	1.036	8.289	1.085	4.535	5.097	8.837	15.649
Weather	MAE	**0.087**	0.414	0.239	0.138	0.184	0.575	0.330	0.587	0.296	0.383	0.958
	RMSE	**0.276**	0.660	0.381	0.323	0.432	1.085	0.920	1.269	1.074	1.046	1.783

Table 3: Few-shot long-term forecasting (96 → 96) results on 10% training data across multiple spatio-temporal datasets. Lower values indicate better performance. **Red**: the best, Blue: the second

Method Type		Foundation Model					Expert Model					
		Spatio Temporal			*Time Series*		*Spatio Temporal*			*Time Series*		
Method		FactoST	OpenCity	UniST	TimesFM	Moirai	STAEformer	STID	D2STGNN	PatchTST	DLinear	Informer
PEMS-03	MAE	**28.57**	34.21	67.70	38.47	51.40	77.42	45.45	OOM	61.22	76.41	46.27
	RMSE	**46.78**	54.82	94.00	59.77	79.47	115.67	65.35	OOM	100.33	113.63	69.41
PEMS-04	MAE	**42.04**	67.24	85.14	64.43	81.30	64.12	78.13	OOM	70.71	85.61	54.26
	RMSE	**64.89**	112.20	112.11	93.44	116.26	91.95	111.12	OOM	104.00	125.44	83.42
PEMS-07	MAE	**45.60**	50.70	101.20	157.10	134.46	61.45	71.32	OOM	80.09	106.68	52.82
	RMSE	**72.47**	78.36	134.98	208.36	200.30	91.06	106.95	OOM	118.54	147.43	81.78
PEMS-08	MAE	**35.69**	49.47	73.81	89.93	68.73	68.45	75.87	OOM	57.31	76.77	44.25
	RMSE	**56.15**	82.07	96.45	125.27	97.89	96.14	103.15	OOM	87.33	109.15	68.43
PEMS-Bay	MAE	**2.96**	7.40	5.17	5.18	5.78	3.28	3.10	OOM	4.32	4.62	3.27
	RMSE	**6.21**	12.38	8.27	9.97	10.97	6.65	6.63	OOM	9.22	9.52	6.81
METR-LA	MAE	6.93	9.71	13.16	14.23	12.17	6.15	**5.94**	OOM	7.20	7.65	6.38
	RMSE	13.07	13.62	19.96	22.56	22.39	**11.56**	11.91	OOM	14.17	13.42	12.29
ETTh2	MAE	0.358	0.751	0.488	0.365	**0.325**	1.295	1.066	OOM	0.943	1.069	2.960
	RMSE	0.561	1.040	0.622	0.541	**0.465**	1.918	1.751	OOM	1.609	1.781	3.783
Electricity	MAE	**0.265**	0.303	0.494	0.305	0.312	0.733	0.440	OOM	0.442	0.558	1.693
	RMSE	**0.409**	1.240	2.512	0.465	0.484	6.562	2.755	OOM	2.747	3.262	16.536
Weather	MAE	**0.226**	0.653	0.348	0.270	0.262	1.171	0.740	OOM	0.708	0.731	2.249
	RMSE	**0.426**	3.730	0.491	0.484	0.465	1.804	1.446	OOM	1.409	1.424	3.403

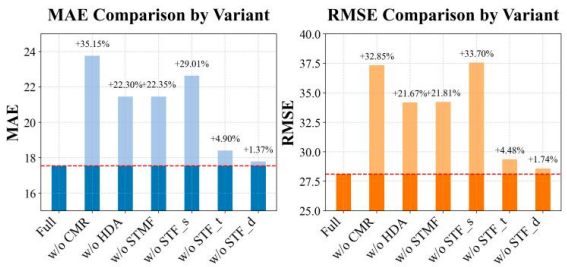


Figure 3: Ablation studies of various components.

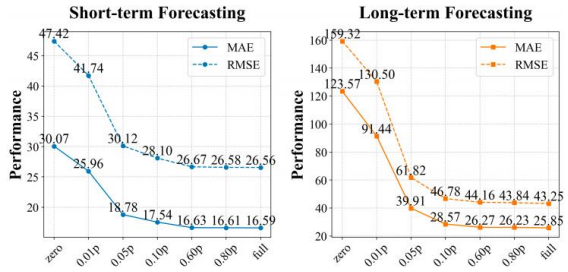


Figure 4: Data scaling analysis.

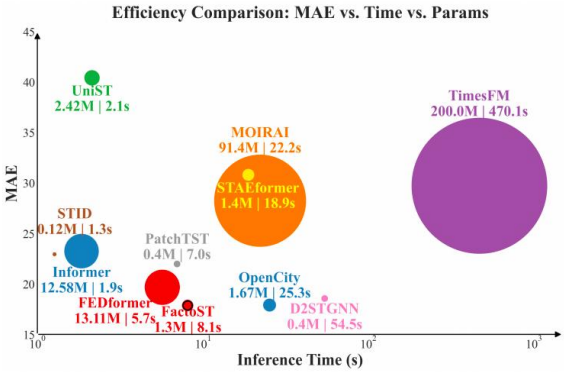


Figure 5: Efficiency comparison with baselines.

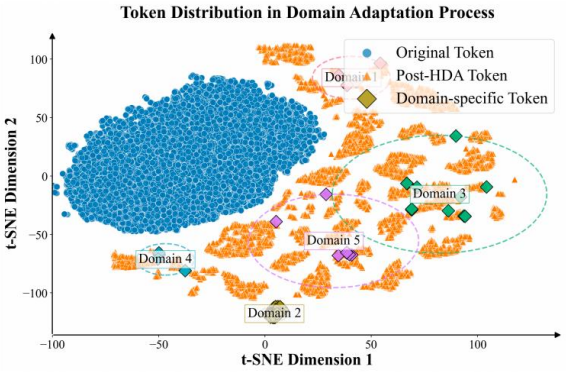
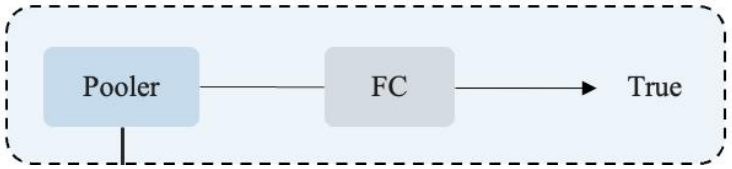


Figure 6: Domain adaptation visualization.

Methodology — Loss Design

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Image Text Matching

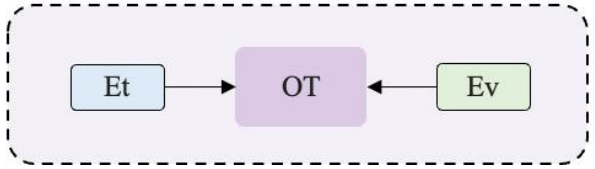


Adapt the model to align the embedding of paired images and text, learning the global consistency between them:

- Randomly replace some images not to match the text during training
- Using a single-layer linear classifier to predict whether an image-text pair matches
- **Binary Cross-Entropy Loss**

$$\mathcal{L}_{\text{ITM}} = -\frac{1}{N}\sum_{i=1}^{N}\left(y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right)$$

Word Patch Alignment

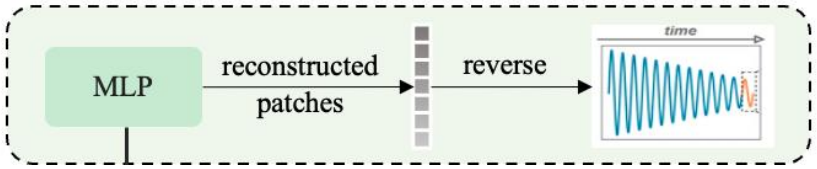


Aligns words in text with visual patches for fine-grained understanding.

- Separates text features Et and visual features Ev
- Uses Optimal Transport with the IPOT algorithm to align features by calculating a transport matrix P that minimizes the **Wasserstein distance** between Et and Ev.

$$\mathcal{L}_{\text{WPA}} = \sum_{i,j} P_{ij} \cdot C_{ij}$$

Masked Visual Modeling



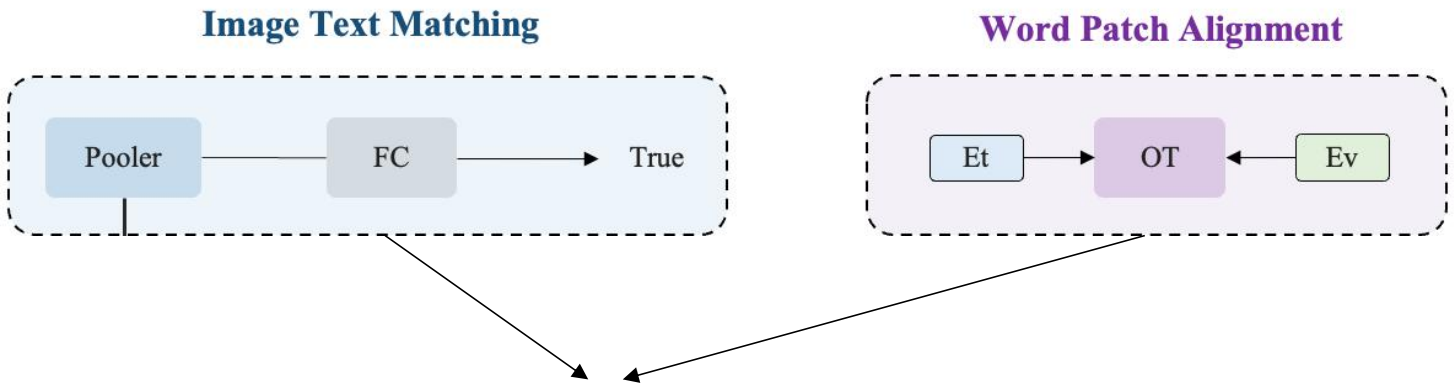
Predicts future time steps by masking parts of the visualized time series, training the model to reconstruct missing data.

- Masks future time points, with visible patches representing known past steps.
- Adapts to different horizons, e.g., predicting 96, 192, 336, 720 steps ahead.
- Uses **MSE** and **MAE** for robust forecasting across datasets.

$$\mathcal{L}_{\text{MVM}}^{\text{MSE}} = \frac{1}{T}\sum_{i=1}^{T}(y_{t+i} - \hat{y}_{t+i})^2 \qquad \mathcal{L}_{\text{MVM}}^{\text{MAE}} = \frac{1}{T}\sum_{i=1}^{T}|y_{t+i} - \hat{y}_{t+i}|$$

$$\mathcal{L}_{\text{Total}} = \lambda_{\text{ITM}}\mathcal{L}_{\text{ITM}} + \lambda_{\text{WPA}}\mathcal{L}_{\text{WPA}} + \lambda_{\text{MVM}}(\mathcal{L}_{\text{MSE}} + \mathcal{L}_{\text{MAE}})$$

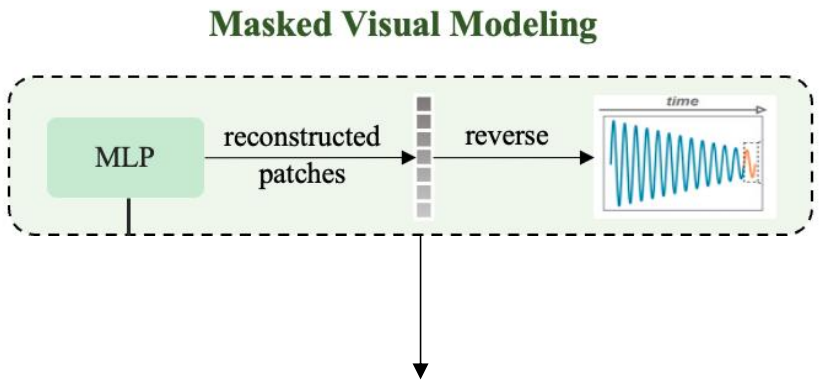
Methodology — Training Strategy



Stage1: Self-supervised learning

Model adapts to low-information images and text

- Image-Text Matching (BCE)
- Word Patch Alignment (OT)
- Masked Image Prediction (MAE)



Stage2: Supervised prediction fine-tuning

Achieve better predictive capabilities

- Masked Visual Modeling (MAE/MSE)