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

Siru Zhong[1], Weilin Ruan[1], Ming Jin[2], Huan Li[3], Qingsong Wen[4], Yuxuan Liang[1]
[1]The Hong Kong University of Science and Technology (Guangzhou), China
[2]Griffith University, Australia
[3]Zhejiang University, China
[4]Squirrel Ai Learning, USA.

ICML 2025

Introduction

Time series forecasting:
- Finance
- Climate
- Transportation

Method:
- RNNs
- Transformer

Issues:
- data-limited scenarios, particularly few-shot and zero-shot settings

→ augmenting time series forecasting with additional modalities

- Text-Augmented Models:
  - the modality gap (continuous time series ↔ discrete text)

  → information loss during representation alignment
  - pre-trained language knowledge is rare for capturing fine-grained temporal patterns (比如趋势拐点、短时波动)
- Vision-Augmented Models:
  - lack semantic interpretability

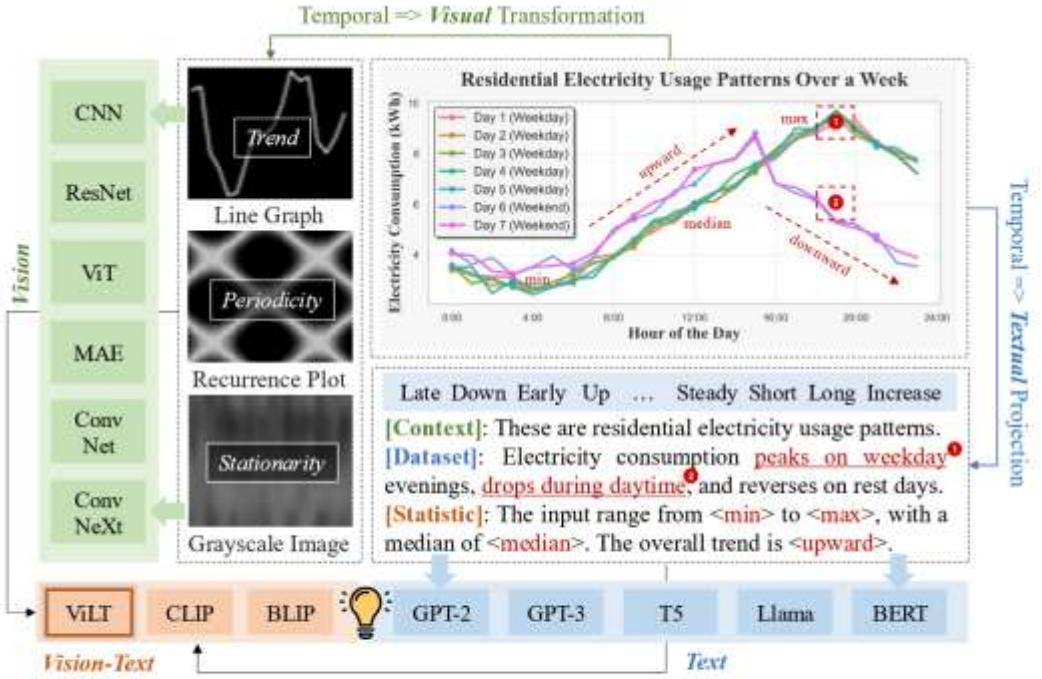

Figure 1: Our Time-VLM combines text (Right) and vision (Left) modalities to augment time series forecasting.

Introduction

Current approaches often focus on single modalities, failing to harness their combined strengths.

→ Time-VLM

- unifying temporal, visual, and textual information
- Temporal: Capture raw dynamics, trends, and periodicity
- Visual (Images): Reveal frequency and spatial patterns through visual representations
- Language (Text): Provide semantic context and domain knowledge

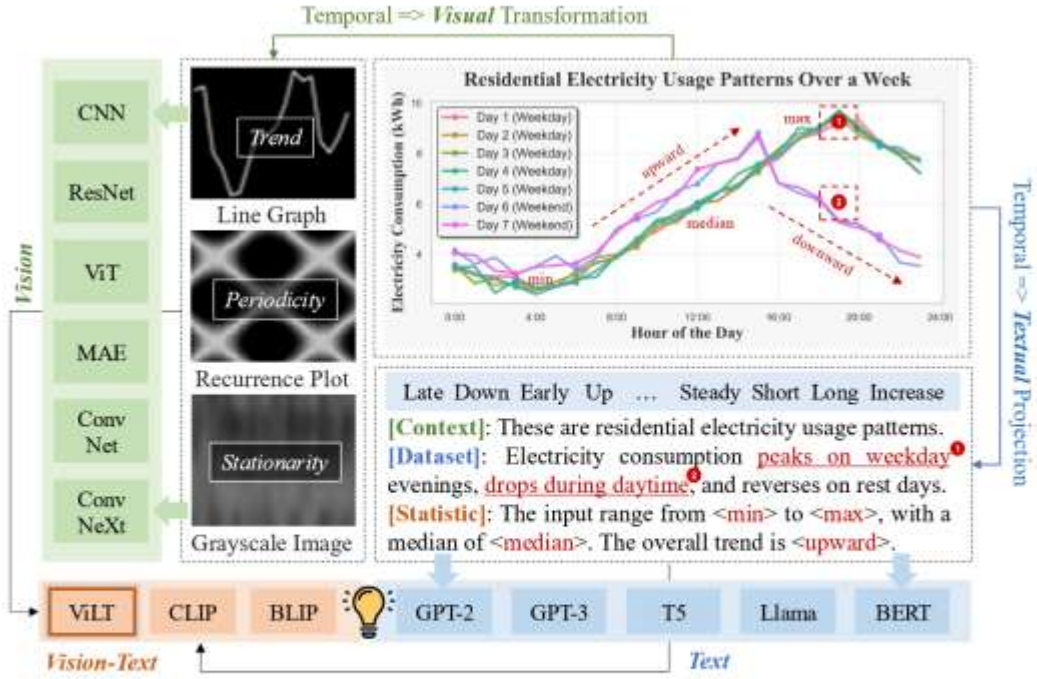

Figure 1: Our Time-VLM combines text (Right) and vision (Left) modalities to augment time series forecasting.

Method

- **Retrieval-Augmented Learner (RAL)**

- Patch Embedding
  embedded using positional encodings.

- Retrieval-Augmented Memory (理解当前时间序列信息的同时从记忆库中寻找相似信息)
  retrieve top-k similar patches from the memory bank M based on cosine similarity

$$\text{sim}(P, \mathcal{M}) = P \cdot \mathcal{M}^\top,$$

$$M_{\text{local}}^{(i)} = \text{MLP}(\text{topk}(E_p^{(i)})),$$

  The global memory is obtained by temporal averaging from current patch P

$$\text{Attn}(P) = \text{MultiHead}(Q, K, V),$$

$$M_{\text{global}} = \frac{1}{N_p} \sum_{i=1}^{N_p} \text{Attn}(P)_i.$$

  This fused representation supports dynamic retrieval and integration with other modalities

$$M_{\text{fused}} = M_{\text{local}} + M_{\text{global}},$$

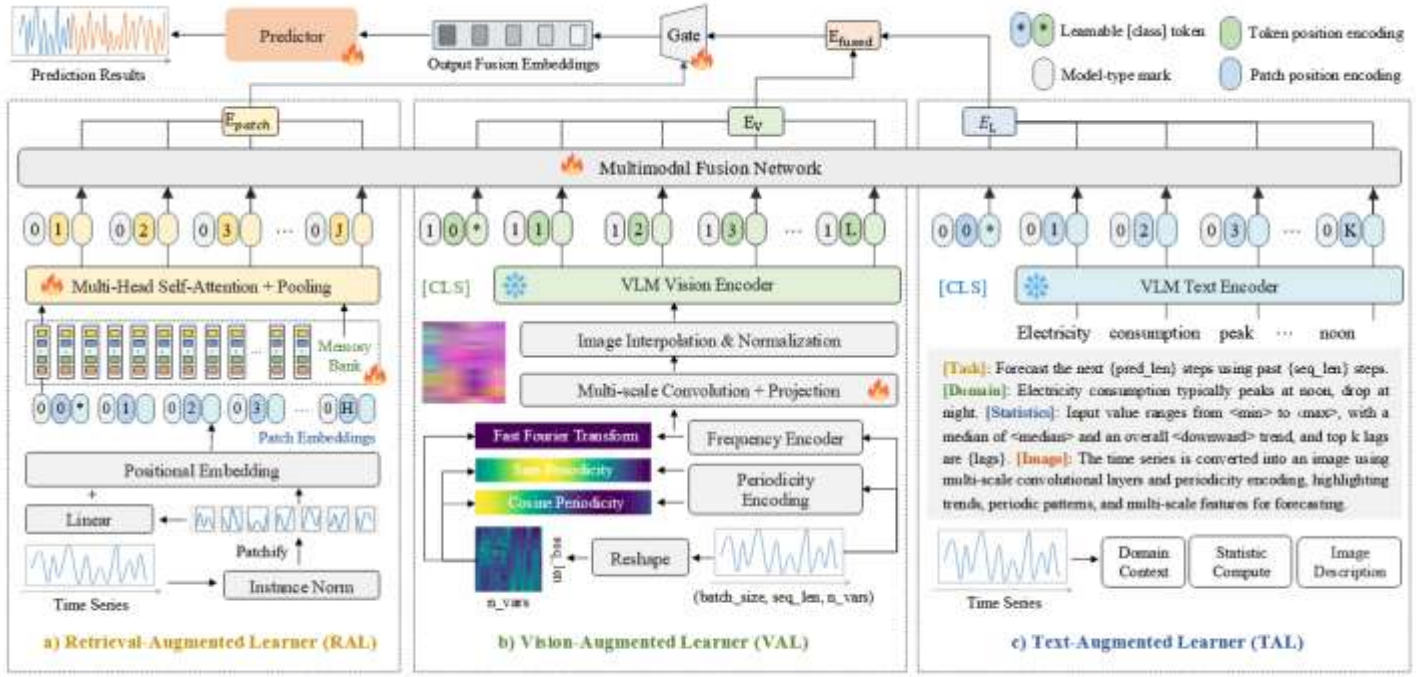

Figure 2: Overview of the Time-VLM framework.

Method

- **Vision-Augmented Learner**
  - Frequency Encoding: adding spectral information

$$\mathrm{FFT}(x_{\mathrm{enc}}) = \sum_{t=0}^{L-1} x_{\mathrm{enc}}(t) \cdot e^{-2\pi i k t / L},$$

  - Periodicity Encoding: adding time-domain information (P: periodicity hyperparameter)

$$\mathrm{encoding}(t) = \left[\sin\left(\frac{2\pi t}{P}\right), \cos\left(\frac{2\pi t}{P}\right) \right],$$

  - Multi-scale Convolution

  - Image Interpolation & Normalization

$$\mathbf{I}(x, y) = \sum_{i=1}^{2} \sum_{j=1}^{2} \mathbf{I}(x_i, y_j) \cdot w_{ij},$$

$$\mathbf{Inorm} = 255 \cdot \frac{\mathbf{I}_{\mathrm{raw}} - \mathrm{Min}(\mathbf{I}_{\mathrm{raw}})}{\mathrm{Max}(\mathbf{I}_{\mathrm{raw}}) - \mathrm{Min}(\mathbf{I}_{\mathrm{raw}}) + \epsilon},$$

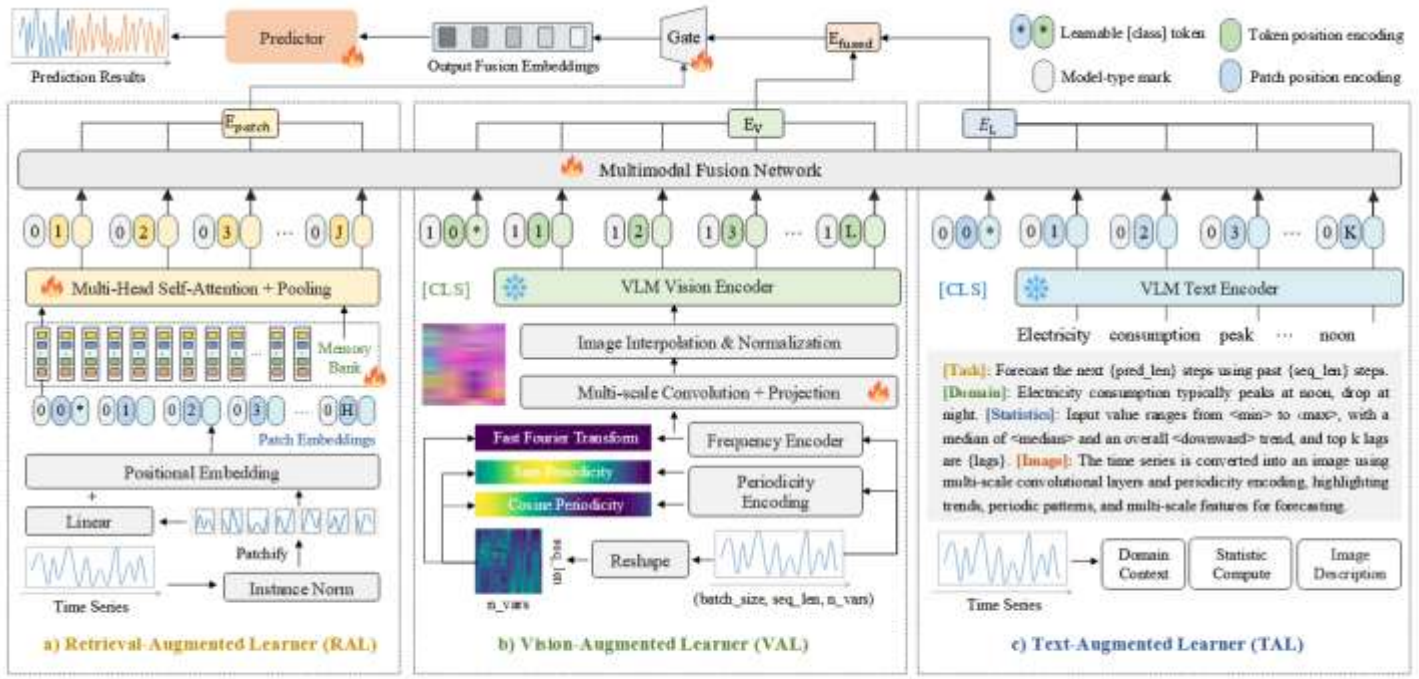

Figure 2: Overview of the Time-VLM framework.

Method

- **Vision-Augmented Learner**
  - Frequency-Domain Information
    - **细纹理** 表示高频成分（快速波动）
    - **大色块区域** 表示低频成分（慢趋势）
  - Multi-scale Periodic Encoding
    - **竖直条纹** 表示每日周期的重复
    - **水平图案** 表示每周模式（如周末变化）
  - Color Trends
    - 色彩亮度表示时间序列的数值大小
    - **深蓝** = 低值，**亮黄** = 高值
  - Abrupt Changes and Anomalies
    - 图像中**颜色突然变化**（例如从深色到亮色）意味着时间序列中出现了突变或异常。

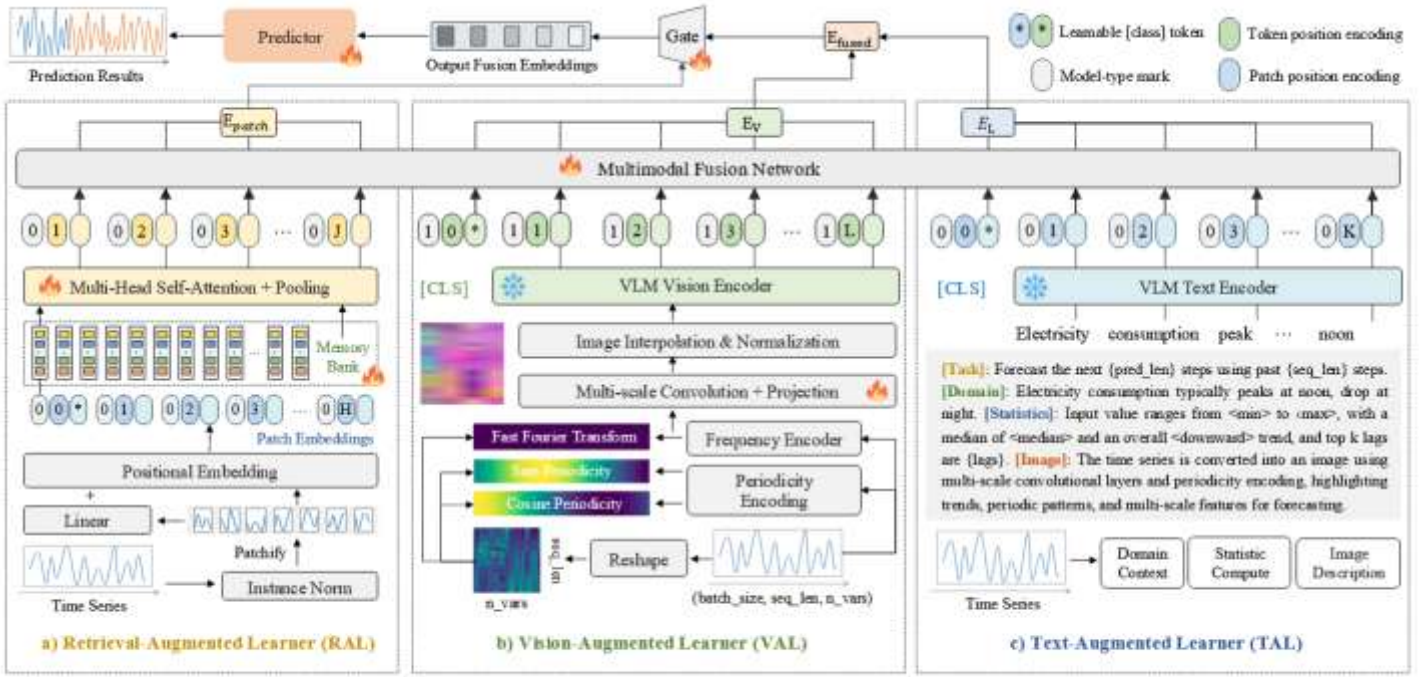

Figure 2: Overview of the Time-VLM framework.

Figure 6: Time series transformed images, capturing key temporal characteristics, including trends, stationarity, seasonality, sudden changes, and frequency-domain patterns.

Method

- **Text-Augmented Learner (TAL)**
  - generates natural language prompts describing the time series — task, domain context, and statistics like min, max, trend
  - tokenized and fed into a VLM's text encoder
    - **Statistical Properties:** Value range (min/max), central tendency (median), and overall trend direction.
    - **Contextual Information:** Periodic description, task-specific parameters (input window length and forecasting horizon), and domain-specific dataset characteristics.

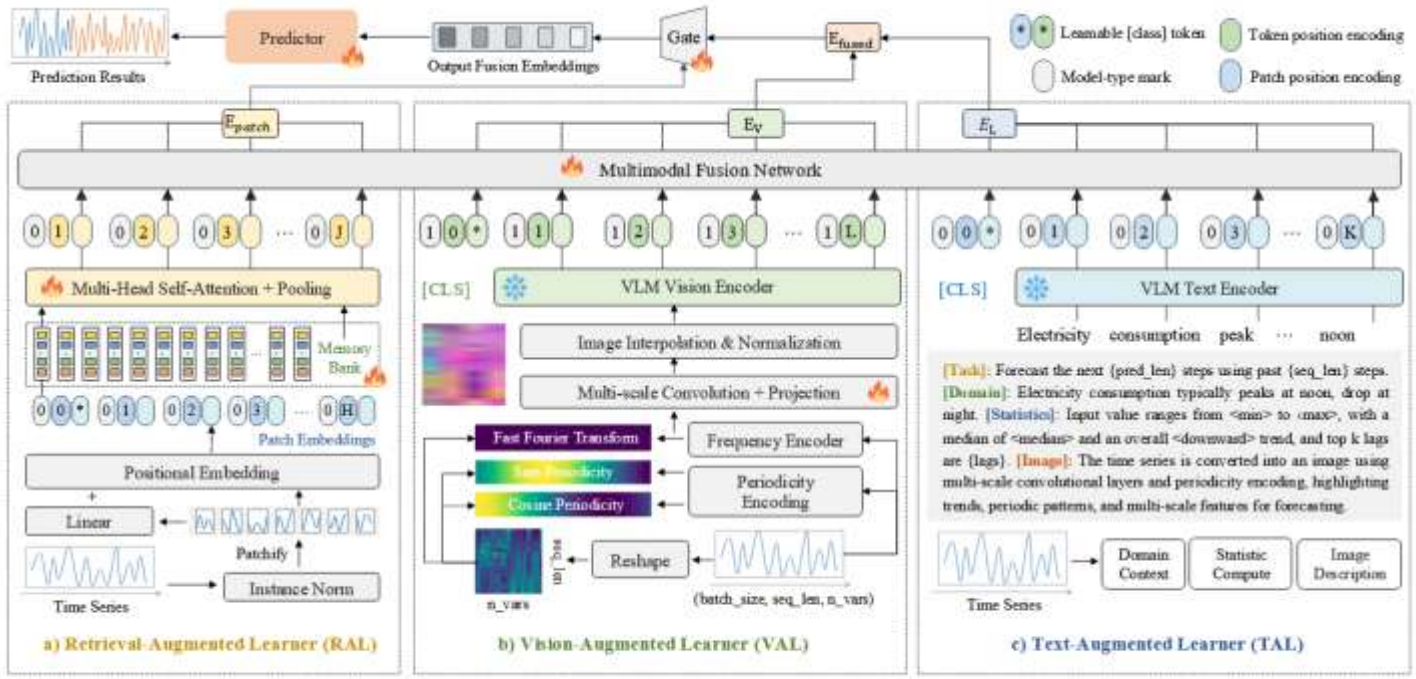

Figure 2: Overview of the Time-VLM framework.

Method

- **Multimodal Fusion with VLMs**
  - Multimodal Embeddings Extraction
    - Feature Alignment
  - Temporal Feature Fusion
    - Temporal memory embeddings as Q
    - multimodal embeddings as K and V

$$\text{CM-MHA}(Q, K, V) = \text{Cat}(\text{head}_1, \ldots, \text{head}_h)W^O, \quad (10)$$

$$\text{head}_i = \text{softmax}\left(\frac{QW_i^Q(KW_i^K)^\top}{\sqrt{d_k}}\right)VW_i^V. \quad (11)$$

$$\mathbf{F}_{\text{attn}} = \text{LayerNorm}(\mathbf{F}_{\text{tem}} + \text{CM-MHA}(Q, K, V)). \quad (12)$$

  - Gated Fusion

$$\mathbf{G} = \sigma(\mathbf{W}_g[\mathbf{F}_{\text{tem}}; \mathbf{F}_{\text{mm}}] + \mathbf{b}_g),$$
$$\mathbf{F}_{\text{fused}} = \mathbf{G} \odot \mathbf{F}_{\text{attn}} + (1 - \mathbf{G}) \odot \mathbf{F}_{\text{mm}},$$

  - Forecasting

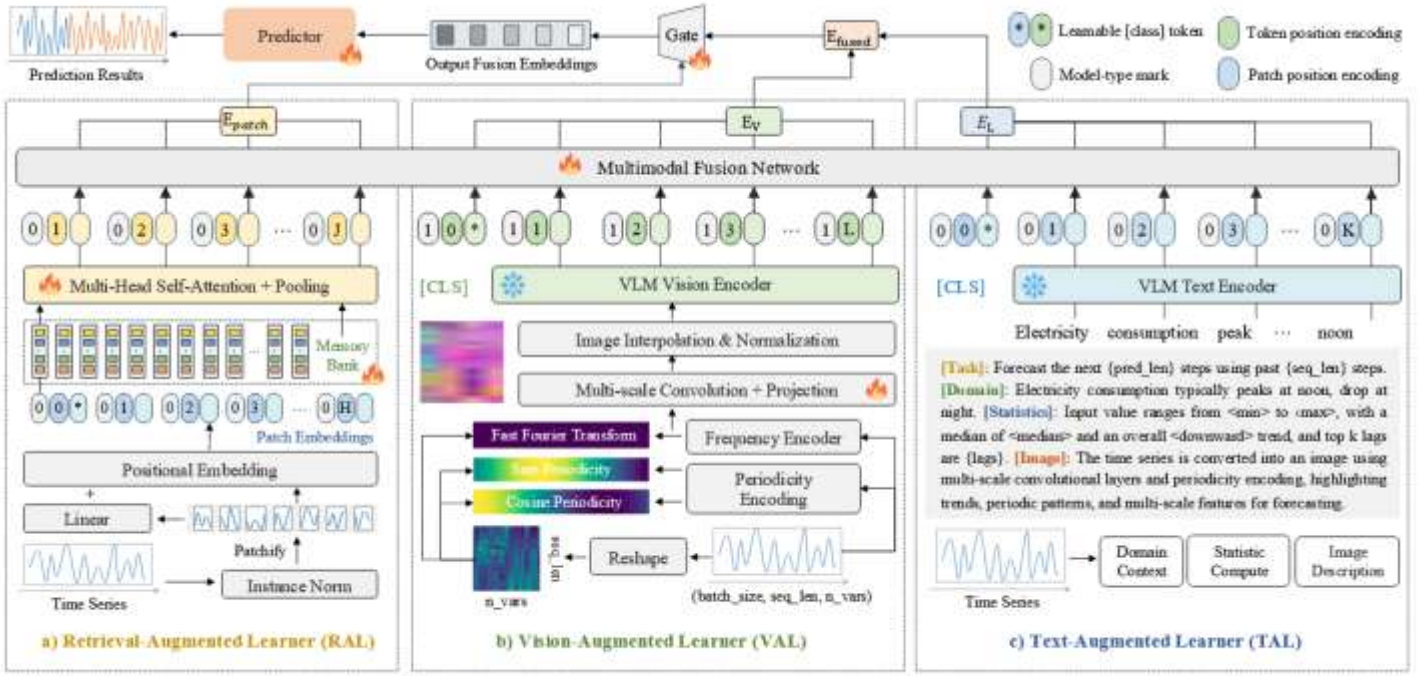

Figure 2: Overview of the Time-VLM framework.

Experiments

Dataset:

- Energy: ETTh1, ETTh2, ETTm1, ETTm2
- Weather
- Electricity: ECL
- Traffic

Traditional models: PatchTST, FEDformer

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 (Ours)		Time-LLM (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 3: Zero-shot learning results. Full results see Section B.2.

Methods	Time-VLM (Ours)		Time-LLM (2024)		LLMTime (2023)		GPT4TS (2023)		DLinear (2023)		PatchTST (2023)	
Metric	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
ETTh1 → *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 → *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 → *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 → *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 → *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 → *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 → *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 → *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

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 (Ours)		Time-LLM (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

Experiments

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)
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

$$\mathbf{G} = \sigma(\mathbf{W}_g[\mathbf{F}_{\text{tem}}; \mathbf{F}_{\text{mm}}] + \mathbf{b}_g),$$
$$\mathbf{F}_{\text{fused}} = \mathbf{G} \odot \mathbf{F}_{\text{attn}} + (1 - \mathbf{G}) \odot \mathbf{F}_{\text{mm}},$$

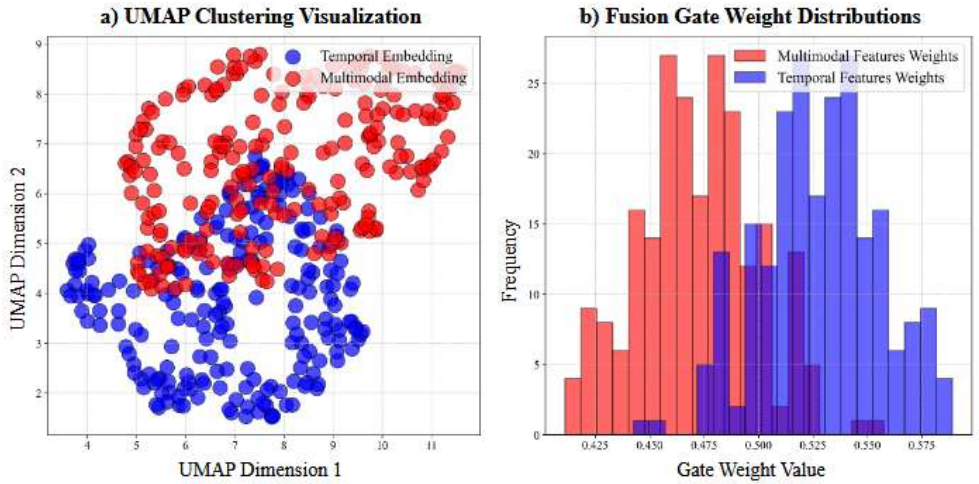

Figure 3: UMAP visualization (left) and gate weight distributions (right) of of multimodal and temporal embeddings.

左图：对齐到了统一语义空间
右图：根据不同样本自适应选择融合比例

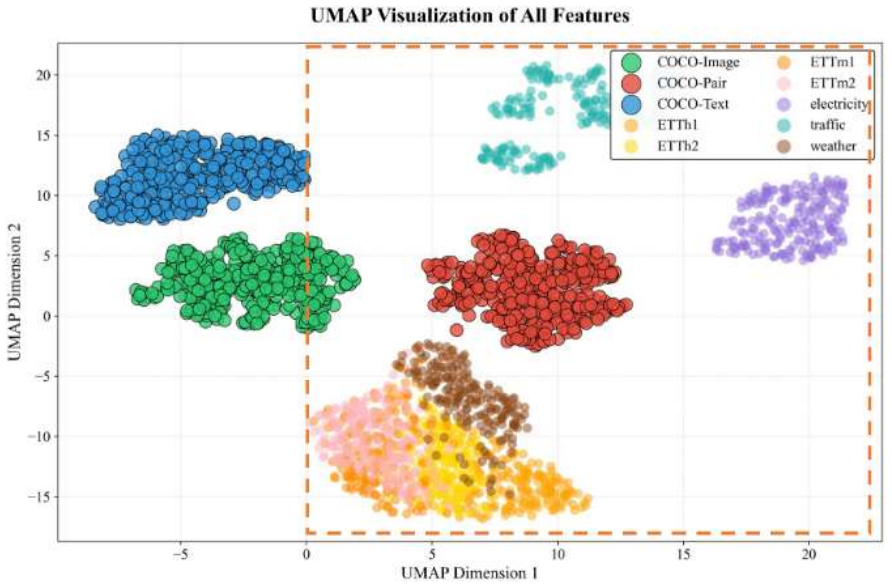

Figure 4: Interpretability visualization of Time-VLM: multimodal feature alignment via UMAP.

COCO-Image和COCO-Text跟时间序列相比较为孤立
表示语言模型难以直接对齐时间数据
COCO-Pair与时间序列模态**显著重叠**，靠近中心表明
VLM 学到的“图文跨模态知识”能迁移到时间序列中

↓

Time-VLM 的做法是：
把时间序列**变成图像 + 文本对**；
然后**嵌入 VLM 共享语义空间**；

Thanks