FLAME: Flow Enhanced Legendre Memory Models for General Time Series Forecasting

Xingjian Wu*, Zhengyu Li*, Hanyin Cheng*, Xiangfei Qiu*, Jilin Hu, Chenjuan Guo, Bin Yang†

East China Normal University

* Equal contribution † Corresponding author

byang@dase.ecnu.edu.cn

NeurIPS 2026

Introduction

Time series foundation models support zero-shot forecasting, but probabilistic models often rely on large backbones and expensive generation. Lightweight alternatives mainly target point forecasting. FLAME combines Legendre Memory with a normalizing-flow head to support both deterministic and probabilistic forecasting.

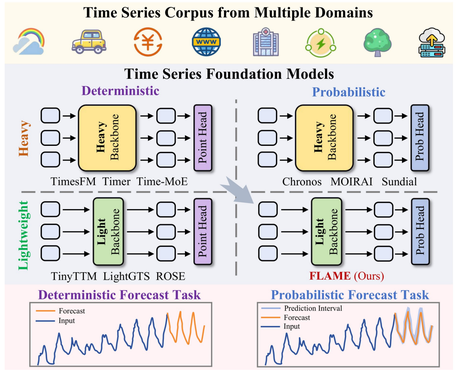


Figure 1: FLAME combines a lightweight backbone with a generative forecasting head. The model is pretrained on time series from multiple domains and supports point forecasts as well as predictive distributions. Existing foundation models are shown for comparison.

Legendre Memory encodes temporal context in a compact state. Translated Legendre memory (LegT) summarizes period-aligned local environments; scaled Legendre memory (LegS) supports long-range decoding. The flow-based head models dependencies within each forecast patch.

- FLAME Small, Base and Large contain approximately 2M, 6M and 10M parameters, respectively.
- Local perception supplements fixed-length patch tokens with variable-length, period-aware context.
- A masked autoregressive flow provides conditional forecast distributions and an exact likelihood.

Ablation study

Table 3: Weather CRPS, averaged over horizons (lower is better).

Variant	CRPS ↓
FLAME Large	**0.074**
w/o Local-Perception	0.104
Self-Attention Decoder	0.106
Diffusion Head	0.077

Paper:

Code:

Methodology

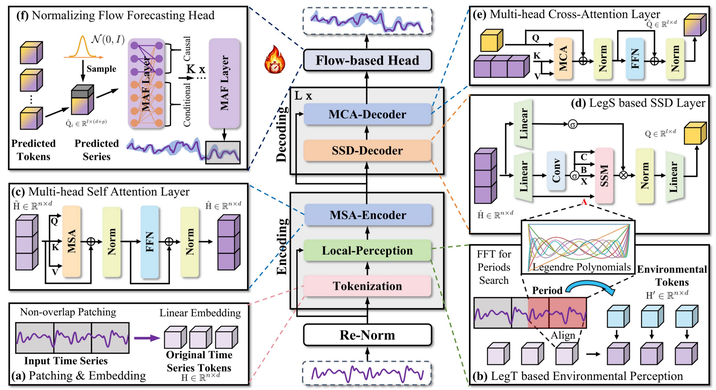


Figure 2: Overall framework of FLAME. Input variables are processed independently. The encoder combines patch embeddings and local context; the decoder predicts future tokens, which condition the flow-based head.

Local perception. A period-aligned environment provides context for each fixed-length patch. LegT compresses this variable-length context into a fixed-size vector and fuses it with the patch embedding before self-attention.

Long-range decoding. A LegS-based structured state-space duality (SSD) decoder summarizes the history and predicts future tokens. Cross-attention then retrieves relevant representations from the encoder.

$$m_t = A_t m_{t-1} + B_t x_t, \quad y_t = C_t^{\mathrm{T}} m_t$$

Flow-based forecasting. A conditional masked autoregressive flow models future patches and their within-patch ordering. The inverse flow maps Gaussian noise to forecast samples, conditioned on each decoded token.

Evaluation uses 100 samples for probabilistic forecasting and their average for point forecasts. FLAME assumes regularly sampled time series.

Forecasting example

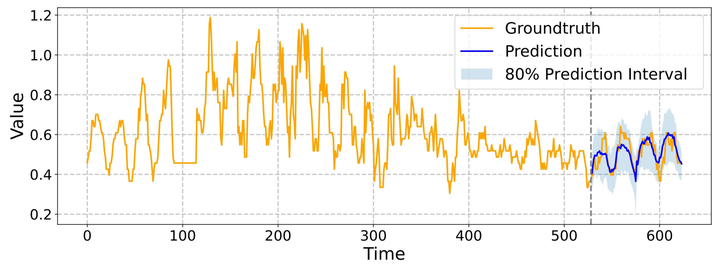


Figure 3: ETTh1 example from the paper (Fig. 7b); shaded area: 80% prediction interval.

Experiments

Table 1: Zero-shot deterministic forecasting. Average MSE over horizons; all dataset groups from the paper are included.

Dataset	FLAME-L	Sundial	LightGTS	MOIRAI
ETT (Avg)	0.344	**0.335**	0.339	0.382
Weather	0.227	0.234	**0.219**	0.260
Electricity	0.217	**0.169**	0.233	0.188
Traffic	**0.546**	—	0.610	—
Solar	**0.168**	0.221	0.219	0.714
PEMS08	**0.293**	—	0.812	—
Wind	**1.042**	1.186	1.292	1.299
NYSE	0.529	0.880	0.620	**0.492**

FLAME-L denotes FLAME Large. Bold red: best among displayed models. —: excluded because of overlap with pretraining data.

Table 2: Zero-shot probabilistic forecasting. Average CRPS over horizons; all dataset groups from the paper are included.

Dataset	FLAME-L	Sundial	Chronos	MOIRAI
ETT (Avg)	0.236	**0.231**	0.290	0.366
Weather	**0.074**	0.087	0.142	0.179
Electricity	**0.081**	**0.081**	—	0.247
Traffic	**0.247**	—	0.269	—
Exchange	**0.042**	0.045	0.044	0.045
ILI	0.154	**0.148**	0.170	0.159

FLAME Large achieves 15 first-place CRPS results. Its average CRPS is 34.3% lower than MOIRAI Large and 18.1% lower than Chronos Large (paper, Section 4.2).

Efficiency analyses

Table 4: ETTm1, prediction horizon 720, batch size 1. Parameters, peak GPU memory and inference time are reported.

Model	Params	GPU (MB)	Time (s)
FLAME Large	**10.6M**	**129**	**0.037**
Sundial Base	128M	603	0.1
Chronos Large	700M	10,269	34.33

FLAME Large requires 0.238G multiply–accumulate operations.

Short-term forecasting

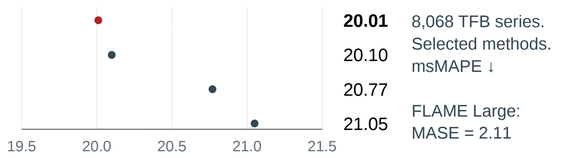


Figure 4: Short-term msMAPE comparison; values from the paper, Section 4.1.