

NeurIPS 2026

MMTA

**Benchmarking Multimodal Temporal Analysis
with Time Series, Text, and Vision**

Ziyang Zhang^{12*} · Shenyi Li^{2*} · Yilin Wang² · Ziyun Cui¹
Bowen Zhou² · Wen Wu² · Chao Zhang^{12†}

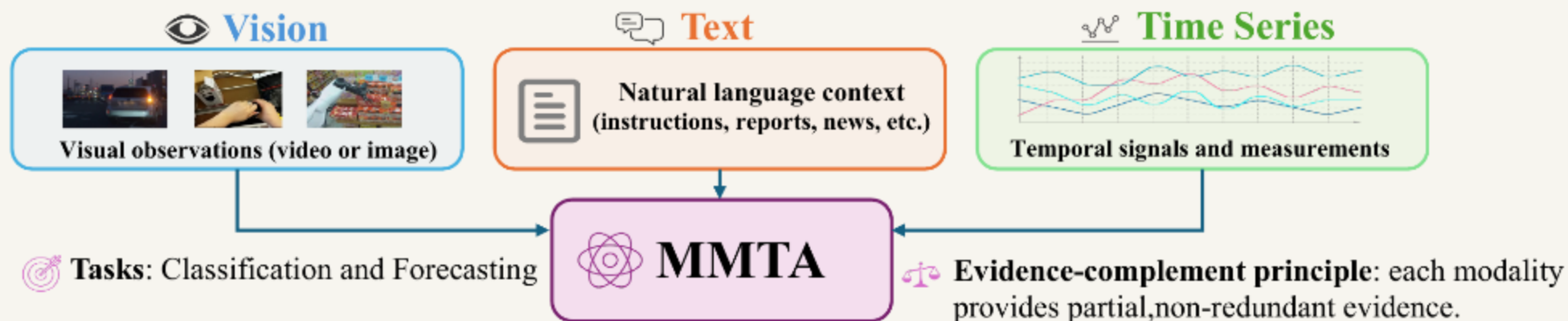
¹ Tsinghua University ² Shanghai Artificial Intelligence Laboratory

* Equal contribution † Corresponding author

windowso.github.io/MMTA/

**Real-world
evidence.
Native temporal
tokens.**

Complementary Evidence



Vision

Scenes, objects
and actions

Text

Instructions, reports
and event context

Time Series

Quantitative changes
and measurements

Joint evidence supports classification and forecasting

MMTA: nine domains, two input layouts

3,524

test samples

2,924

classification

600

forecasting

Two input layouts

Video + TS / Image + TS



Unified evaluation across real-world dynamic systems

TimeOmni-v: native temporal tokens

01 Encode

Chronos-2 encoding
16-step patch tokens

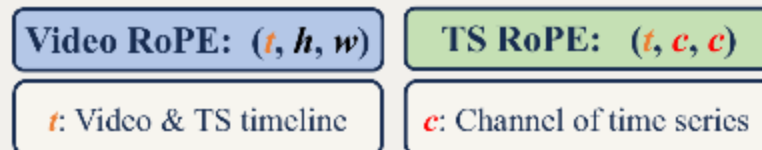
02 Align

TS-specific RoPE
Modality-aware layout

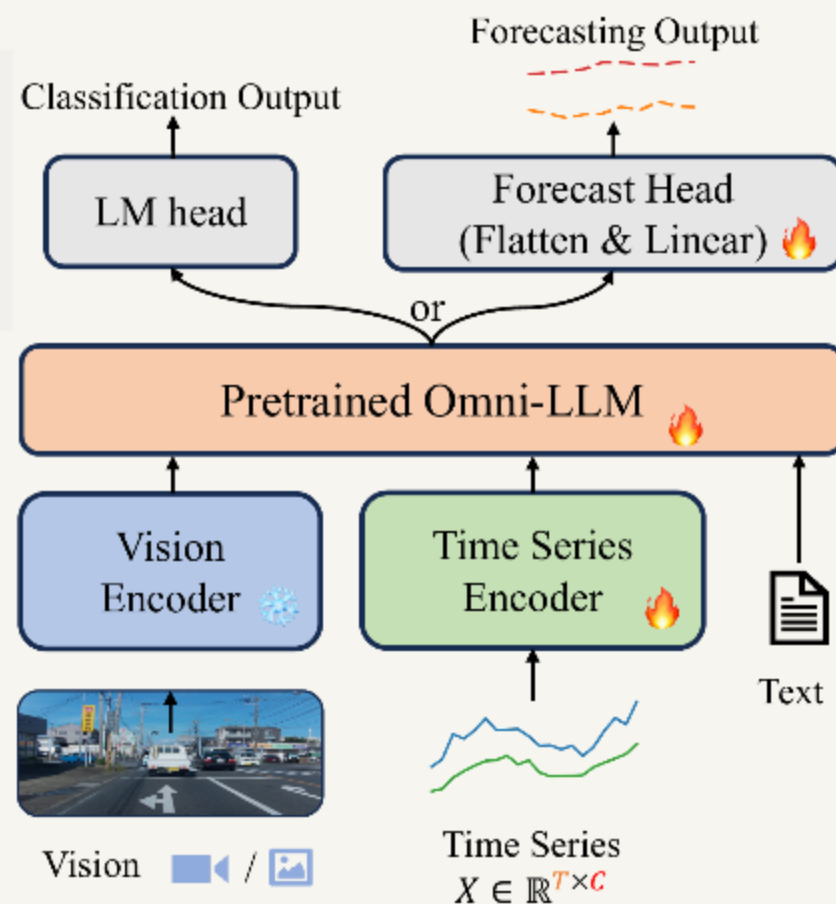
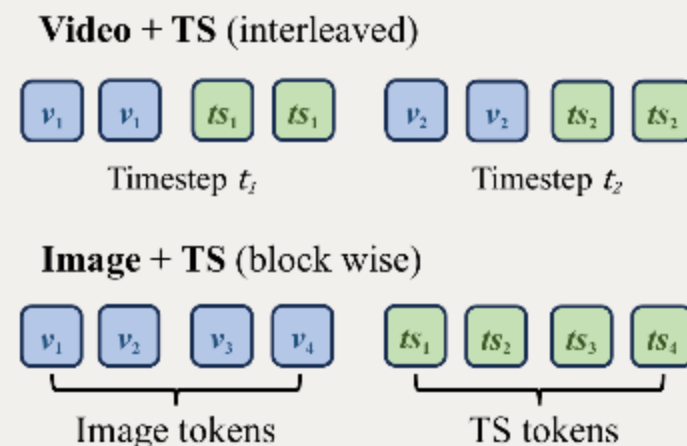
03 Predict

Classification
Numerical forecasting

A. TS-specific Multimodal RoPE



B. Modality-aware Token Layout



Compact tokens aligned with visual evidence

Stronger long-video classification

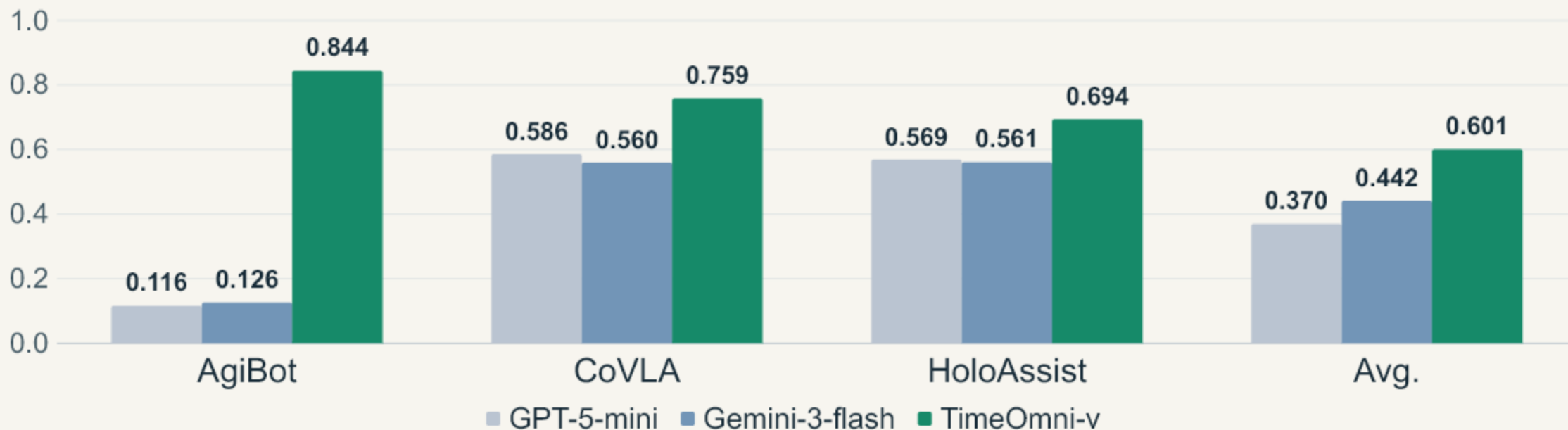
0.601

Avg-swF1

+36.0%

relative gain over 0.442

Highest average swF1
among evaluated models



8 reference MLLMs: **zero-shot**. TimeOmni-v: **per-domain fine-tuning**.
swF1 = Macro-F1 × SuccessRate. Avg. covers all classification domains.

Efficient inference on long time series

232.8×

Fewer TS tokens

23.0×

Lower latency

3.1×

Lower peak VRAM

AgiBot on a single H200

Matched dynamic-batching budgets

Inference cost	TS as text	TimeOmni-v
Median TS tokens	167,041	722
Latency / sample	27.70 s	1.20 s
Peak VRAM	108.6 GB	35.0 GB
Completed samples	113 / 131	131 / 131

Qwen2.5-Omni-7B with textualized TS vs TimeOmni-v

TimeOmni-v completes all 131 samples

MMTA and TimeOmni-v

Real-world evidence. Native temporal tokens.

MMTA

9 domains.

Classification and forecasting.

Stronger long-video classification

Efficient inference on long series.

Integration tested on two backbones.

TimeOmni-v

Compact **TS** encoding.

Temporal alignment.

Task-specific **output heads.**

Explore MMTA

windowso.github.io/MMTA/

github.com/windowso/TimeOmni-v