MulTiCast: A Multimodal Time Series Forecasting System

Sehyuk Park[*1, 2], **Caren Han**[*1, 2], **Hyunsuk Chung**[2]

[1]Pohang University of Science and Technology,
[2]The University of Melbourne
percy212@postech.ac.kr, caren.han@unimelb.edu.au, hyunsuk.chung.1@unimelb.edu.au

Abstract

Time-series forecasting plays an essential role in domains such as finance, healthcare, and energy. Yet most existing systems operate in a unimodal setting, overlooking complementary information available from visual and textual modalities. There are surprisingly few time-series forecasting demos, and multimodal, interpretable demos are rarer still. In practice, it is difficult for users to experiment with foundation models on their own time-series data due to strict input requirements, heavy setup burdens, and limited interpretability support. We present MulTiCast, an interactive **Mul**timodal **Ti**me series fore**Cast**ing system that enables users to combine numerical signals with visual and textual context to improve predictions. The system builds on pretrained models with lightweight adaptation, but its central contribution lies in the interactive demonstration platform. Through a web interface via Hugging Face Spaces, users can load datasets, toggle modality inclusion, and visualize forecasts together with the attention maps of each modality, providing insights into the reasoning path behind the predictions.

Demo — https://huggingface.co/spaces/adnlp/MulTiCast

Introduction

Accurate forecasting requires more than raw numerical sequences. Real-world scenarios often include auxiliary context, such as descriptive metadata and visual plots, that can enrich predictions and make outcomes easier to trust. Despite recent progress in Time Series Foundation Models (TSFMs) such as Chronos(Ansari et al. 2024), Timer(Liu et al. 2024), Moirai(Woo et al. 2024), and compact mixers(Ekambaram et al. 2024), the dominant paradigm remains unimodal. These models report strong benchmark scores but seldom expose interactive systems that non-experts can use.

There are surprisingly few time-series forecasting demos, and multimodal, interpretable demos are rarer still. In practice, people struggle to try foundation models on their own time series for several reasons. Data must be reshaped to strict input schemas (frequency, missing-value handling, scaling, window/horizon choices); environment setup is heavy (model checkpoints, tokenizer/feature extractors, dependencies); GPU/CPU latency can be high for interactive loops; and, crucially, interpretability tools are fragmented, making it hard to understand how models use signals across modalities. Most open demos prioritize static examples or offline leaderboards, not hands-on exploration of a user's custom slice of data with transparent reasoning aids.

Meanwhile, multimodal foundation models in vision and language, such as CLIP(Radford et al. 2021) and BLIP(Li et al. 2022), show how heterogeneous inputs can be aligned for downstream tasks. However, their integration into forecasting remains underexplored, and nearly no public systems let users see how visual and textual cues influence sequential predictions in real time. MulTiCast fills this gap by providing a usable, browser-based system that combines numerical, visual, and textual context and reveals its reasoning path via modality-specific attention maps. MulTiCast emphasizes accessibility, transparency, and interactivity by bringing multimodal and explainable forecasting to researchers and practitioners in a form they can actually use.

System Design and Implementation

MulTiCast, based on **UniCast**(Park, Han, and Hovy 2025), is designed as a web-based system that allows users to perform multimodal time-series forecasting without technical setup. The system integrates a numerical time-series forecasting backbone with vision and text encoders. It exposes these capabilities through a lightweight but fully interactive web interface hosted on Hugging Face Spaces. Its design focuses on lowering the barrier to entry while ensuring interpretability through attention-based visualizations.

A) Input Data Choice. At the entry point, users are provided with benchmark datasets such as NN5 and Australian Electricity, which are commonly used in time-series forecasting research. The interface also supports custom dataset uploads, allowing users to test MulTiCast on their own data. This dual option ensures that both first-time visitors and expert practitioners can engage meaningfully: newcomers can start with ready-to-use data, while researchers can examine forecasts on proprietary or domain-specific datasets.

B) Input Segment Selection. After choosing a dataset, the user selects a specific input window. The interface provides suggested sample segments for quick trials, but users can also define a custom segment by choosing the start and end points of interest. This feature mimics real forecasting workflows, where practitioners often need to focus on recent or

[*]Equal Contribution.

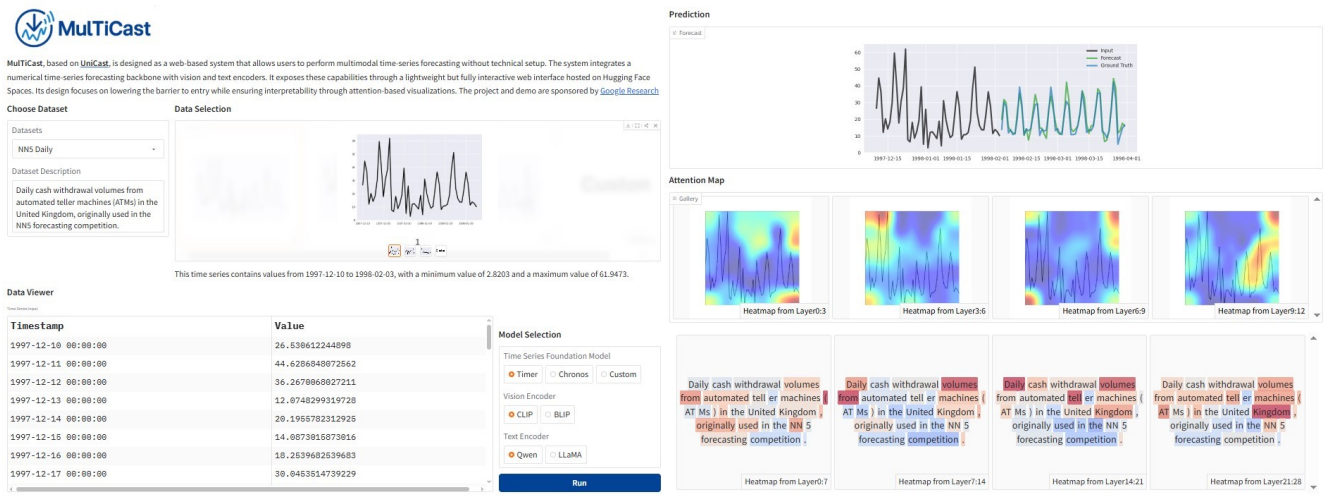

Figure 1: The MulTiCast demo interface. Users can select datasets, define input segments, and choose models. It generates forecasts compared with ground truth and visualizes modality-specific attention maps to highlight the reasoning process.

anomalous periods. Once selected, the chosen input segment becomes the context window that drives forecasting.

C) Model Configuration. The forecasting backend centers on UniCast, which aligns multiple modalities into a shared representational space before performing autoregressive or sequence-to-sequence forecasting. The demo interface allows toggling each modality on/off, enabling comparative evaluation between unimodal and multimodal configurations. This modularity is crucial for educational and diagnostic use, helping users see how each modality influences model behavior. MulTiCast also supports uploaded user-trained models[1], which are automatically wrapped by the same inference pipeline. This design future-proofs the system as practitioners increasingly fine-tune or adapt foundation models to their domains.

D) Forecasting and Visualization. Given the dataset, input window, and model, MulTiCast generates forecasts in real time. The results are plotted against the ground truth curve in the same chart, allowing immediate assessment of prediction accuracy. The visualization is responsive, enabling users to adjust input segments or modality settings and instantly view updated predictions. This dynamic feedback loop distinguishes MulTiCast from static demonstrations.

E) Attention-based Interpretability. A notable feature of MulTiCast is its attention-based interpretability layer, which reveals the internal reasoning pathway for each modality. Time-series attention highlights specific timesteps within the input sequence that the model considered most influential. Visual attention/saliency overlays heatmaps on plot patches to show what trends or anomalies influenced predictions. Textual attention displays token-level weights that indicate which metadata or descriptive phrases contributed to the forecast. These interpretability tools are computed directly from UniCast's intermediate attention tensors, ensuring that visualizations reflect actual model behavior rather than post-hoc approximations. The maps are presented in a unified panel next to the forecast, enabling users to correlate model decisions with domain knowledge.

User Scenario. Dr. Song, a data scientist at an energy company, often faces the challenge of predicting electricity demand under rapidly changing conditions. He opens the MulTiCast demo website and selects the Australian Electricity dataset. Curious about seasonal fluctuations, he customizes the input window to cover the most recent three months. He then chooses a trained forecasting model and toggles between numeric-only and multimodal settings. Within seconds, the forecasts appear against the ground-truth curve, revealing that the multimodal configuration captures seasonal spikes more accurately. Inspecting the attention maps, Dr. Song sees that the model focused on descriptors of the holiday season and visual patterns of sudden peaks, confirming that it leveraged contextual cues rather than simple extrapolation. In another scenario, Daniel, a medical graduate student, uploads a custom dataset of COVID-19 daily cases from his country. Selecting a two-month segment, he compares different models and observes how forecasts change. The attention maps highlight which timesteps and contextual signals most influenced predictions, helping him interpret the outcomes. Through these scenarios, MulTiCast demonstrates its value as an accessible and interpretable time-series forecasting platform that lowers the barrier for both practitioners and researchers.

Impact and Future Work

MulTiCast lowers technical barriers by enabling practitioners to explore foundation models in the browser, and also serves as a teaching tool to understand how multimodal signals shape forecasts. Looking ahead, we will enhance flexibility by streamlining support for user-trained models, so that the demo evolves from a static showcase into a platform for prototyping and comparative evaluation. Through these directions, it has the potential to bridge the gap between research advances and real-world adoption, fostering practical understanding of multimodal time-series forecasting.

[1] **Custom Training Code**: https://github.com/adlnlp/unicast

Acknowledgments

This research was supported by the Korea Planning & Evaluation Institute of Industrial Technology (KEIT) funded by the Ministry of Trade, Industry and Energy (No.RS-2025-25458052, Development of Core Technologies for Manufacturing Foundation Models) and the Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2025-02217259, Development of self-evolving AI bias detection-correctionexplain platform based on international multidisciplinary governance), (RS-2024-00395401, Development of VFX creation and combination using generative AI).

References

Ansari, A. F.; Stella, L.; Turkmen, C.; Zhang, X.; Mercado, P.; Shen, H.; Shchur, O.; Rangapuram, S. S.; Arango, S. P.; Kapoor, S.; et al. 2024. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Ekambaram, V.; Jati, A.; Dayama, P.; Mukherjee, S.; Nguyen, N.; Gifford, W. M.; Reddy, C.; and Kalagnanam, J. 2024. Tiny time mixers (ttms): Fast pre-trained models for enhanced zero/few-shot forecasting of multivariate time series. *Advances in Neural Information Processing Systems*, 37: 74147–74181.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, 12888–12900. PMLR.

Liu, Y.; Zhang, H.; Li, C.; Huang, X.; Wang, J.; and Long, M. 2024. Timer: generative pre-trained transformers are large time series models. *Proceedings of the 41st International Conference on Machine Learning*, 32369–32399.

Park, S.; Han, S. C.; and Hovy, E. 2025. UniCast: A Unified Multimodal Prompting Framework for Time Series Forecasting. *arXiv preprint arXiv:2508.11954*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PmLR.

Woo, G.; Liu, C.; Kumar, A.; Xiong, C.; Savarese, S.; and Sahoo, D. 2024. Unified Training of Universal Time Series Forecasting Transformers. *Proceedings of Machine Learning Research*, 235: 53140–53164.