TIMEVISTA: EXPLORING AND EXPLOITING VISION-LANGUAGE MODELS AS JUDGES FOR TIME SERIES FORECASTING

Anonymous authors
Paper under double-blind review

ABSTRACT

High-quality time series forecasting is pivotal for real-world decision-making. However, traditional numerical metrics often fail to reveal complex temporal patterns and are limited in their ability to align with human intuitive preferences. While the "LLM-as-a-Judge" paradigm has revolutionized text evaluation by providing flexible, human-aligned judgment, its application to time series remains largely unexplored. In this paper, we leverage Vision-Language Models (VLMs) as judges for time series forecasting, harnessing their ability to comprehend time series plots grounded in textual constraints. Specifically, we propose a novel framework that integrates Micro- and Macro-Level judgments informed by contextual information to evaluate time series forecasting. To this end, we introduce TimeVista, comprising 5,563 time series samples with evaluation rubrics, and Meta-TimeVista, comprising 1,025 forecast instances annotated by five experts. Meta-evaluation shows that VLM judges possess better alignment with human judgments than point-wise and structure-aware numerical baselines. Further analyses examine the effects of visual rendering and rubric design on VLM judgments. Building upon our benchmark, we comprehensively assess recent Time Series Foundation Models (TSFMs) under the VLM-as-a-Judge paradigm. Our results demonstrate that VLMs serve as robust and interpretable judges, providing a comprehensive, human-aligned standard for evaluating time series models.

1 INTRODUCTION

Time series forecasting supports informed decisions in applications ranging from finance (Tsay, 2005) to weather prediction (Wu et al., 2023). While forecasting models have evolved from statistical methods to Time Series Foundation Models (TSFMs) (Liang et al., 2024; Miller et al., 2024) and multimodal architectures (Wu et al., 2025; Ahamed et al., 2026), the evaluation paradigm still relies mainly on conventional point-wise metrics, notably mean squared error and mean absolute error.

While the practical utility of a forecast hinges on structural fidelity and domain context, point-wise metrics fail to convey such rich information. A forecast with low point-wise error may still fail to preserve key temporal patterns or meet domain-specific requirements. In the electrocardiogram (ECG) forecasting shown in Figure 1, preserving waveform morphology can be more important than matching exact peak times for downstream clinical interpretation. Yet, operating in a numerical vacuum, standard point-wise metrics penalize useful shifted peaks more heavily than useless smoothed waves, contradicting human intuition by prioritizing numerical precision over structural alignment.

Similar evaluation dilemmas arise in natural language generation, where n-gram-based metrics such as BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004) fail to fully reflect human preferences, motivating the "LLM-as-a-Judge" paradigm for more human-aligned feedback (Gu et al., 2026; Chen et al., 2024). However, extending this paradigm to time series evaluation poses several challenges for text-centric LLMs. First, discrete tokenization conflicts with the continuous numerical data (Jin et al., 2024), impairing magnitude comprehension. Second, extensive time series sequences expose LLMs to severe attention degradation, known as the "lost in the middle" effect (Liu et al., 2024). Third, representing time series as flat text obscures temporal patterns, such as trends and phases.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

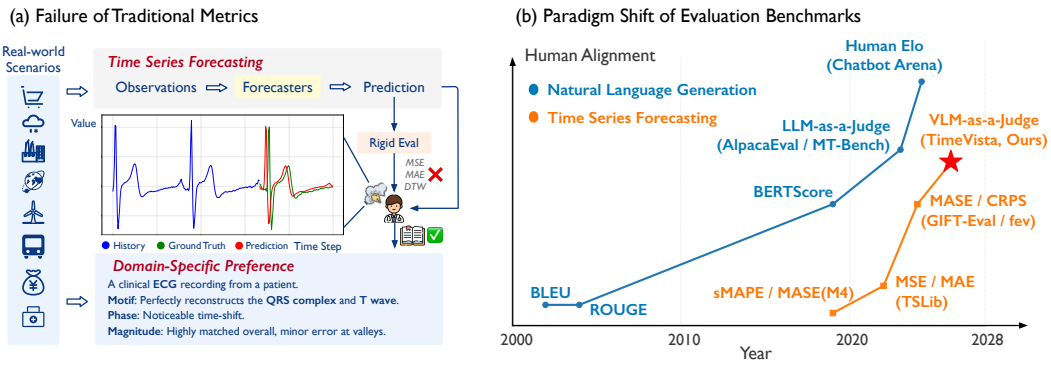


Figure 1: (a) Numerical metrics can conflict with domain-specific preferences in real-world scenarios. (b) The paradigm shift in evaluation benchmarks across natural language generation and time series forecasting is from rigid metrics toward model-based judging.

Inspired by how human experts assess forecast quality through visual inspection, we introduce the "VLM-as-a-Judge" paradigm for time series forecasting. Unlike general image or chart understanding, interpreting time series prediction requires joint comprehension of temporal dynamics and domain context. Therefore, we formulate the forecast evaluation as a multimodal task at two levels. At the Micro-Level, VLMs assess fine-grained temporal patterns, including trend, phase, motif, and magnitude, from the plotted prediction. At the Macro-Level, the evaluation is grounded in domain knowledge to judge whether the forecasts align with domain-specific constraints and preferences.

To this end, we introduce TimeVista, the first VLM-as-a-Judge benchmark for time series forecasting, comprising 5,563 time series paired with evaluation rubrics (Figure 2a). Meta-evaluation on Meta-TimeVista (Figure 2b) demonstrates strong alignment between VLM judgments and human annotations, while ablation studies reveal sensitivity to rubric design and visual input. We then exploit VLM judges to evaluate 13 forecasting models at the Micro-Level on GIFT-Eval (Aksu et al., 2024) and the Macro-Level on TimeVista. VLM-generated scores and explanations highlight pattern-specific performance and violations of domain-specific requirements, complementing MSE and MAE with explainable feedback. Our contributions are summarized as:

- We explore the use of VLMs as judges for time series forecasting and introduce TimeVista, the first VLM-as-a-Judge benchmark, featuring 5,563 time series paired with detailed evaluation rubrics.
- We construct Meta-TimeVista, containing 1,025 human-annotated forecast instances, to support comparisons of VLM judges and numerical baselines in terms of alignment with human judgment.
- We exploit VLM judgments to analyze 13 forecasting models on GIFT-Eval and TimeVista, identifying differences in structural strengths and adherence to domain-specific criteria.

2 Related Work

2.1 Time Series Forecasting Benchmarks

Forecasting benchmarks have expanded from dataset-specific evaluations such as ETT and Weather (Wang et al., 2026) to cross-domain, zero-shot assessment of TSFMs (Das et al., 2023; Woo et al., 2024; Ansari et al., 2025). GIFT-Eval (Aksu et al., 2024) evaluates generalization across datasets from diverse domains, fev-bench (Shchur et al., 2025) covers multivariate forecasting, and TIME (Qiao et al., 2026) introduces task-centric and pattern-level evaluation.

While these efforts substantially broaden data coverage and evaluation protocols, most benchmarks still focus on purely numerical, point-wise errors. To bridge this, as shown in Figure 1, TimeVista leverages the VLM-as-a-Judge paradigm to evaluate predictions from visual plots grounded in textual contexts, providing a multimodal and human-aligned complement to existing metrics.

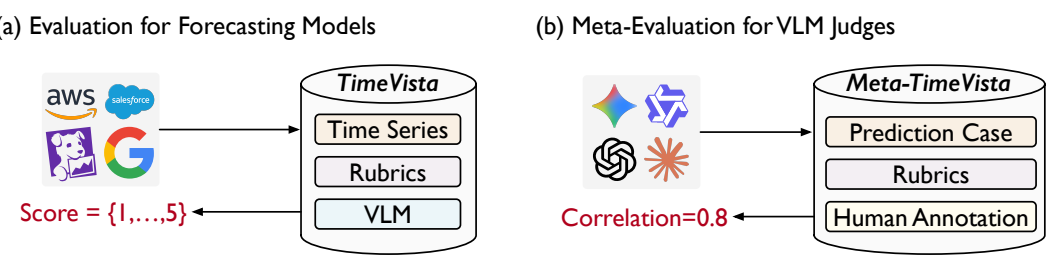


Figure 2: Illustration of the benchmarks proposed in this work. (a) Forecasting model evaluation on TimeVista using VLM judges with scoring rubrics. (b) Meta-evaluation on Meta-TimeVista to assess the alignment between VLM judges and human annotations.

2.2 Numerical Evaluations

Time series forecasts are commonly evaluated using numerical metrics, with MSE and MAE quantifying pointwise errors and CRPS assessing probabilistic forecasts (Ansari et al., 2024; Aksu et al., 2024). To account for temporal patterns, DTW compares sequences through nonlinear alignment (Sakoe & Chiba, 1978). Its extensions provide differentiable alignment through Soft-DTW (Cuturi & Blondel, 2017), emphasize local shape through derivative DTW (Keogh & Pazzani, 2001), and jointly penalize shape and timing distortions through DILATE (Le Guen & Thome, 2019). More targeted evaluation compares trends extracted by STL decomposition or HP filtering (RB, 1990; Hodrick & Prescott, 1997), measures timing deviations using cross-correlation lag (Knapp & Carter, 1976) or DTW warping offsets, and compares value distributions using Wasserstein distance (Gabriel & Marco, 2019).

However, these numerical metrics do not always reflect human priorities concerning certain patterns, acceptable distortions, and domain context. Section 4.3 therefore compares rubric-guided VLM judgments with both established numerical metrics and a deterministic implementation of the supplied rubrics, examining their alignment with human annotations.

2.3 LLM-as-a-Judge Evaluation

Evaluating open-ended generation in natural language processing has long suffered from lexical metrics such as BLEU and ROUGE (Papineni et al., 2002; Lin, 2004). Their reliance on rigid n-gram overlap often fails to capture semantic equivalence, coherence, and human preferences. The "LLM-as-a-Judge" paradigm (Gu et al., 2026) is proposed to address these limitations through rubric-guided assessments and explanations that align more closely with human judgments (Zheng et al., 2023; Chiang et al., 2024). It has been widely adopted for evaluating dialogue, instruction following, and reasoning (Bai et al., 2023; Fu et al., 2024). Recent work extends this approach to multimodal evaluation, using VLM judges to assess responses grounded in visual inputs (Lee et al., 2024).

Despite its success in NLP, this paradigm remains unexplored in time series analysis due to modality mismatch. Inspired by how human experts evaluate prediction quality through visual inspection, the proposed TimeVista leverages VLMs to bypass these textual constraints. By rendering forecasts as visual plots grounded in explicit scoring rubrics, we evaluate whether VLMs can jointly assess temporal dynamics and domain requirements in alignment with human annotations.

3 TimeVista

3.1 Problem Formulation

Given a historical look-back window $\mathbf{x} \in \mathbb{R}^L$, a forecasting model generates a prediction $\hat{\mathbf{y}} = \mathcal{F}_\theta(\mathbf{x}) \in \mathbb{R}^T$. Let $\mathbf{y} \in \mathbb{R}^T$ denote the corresponding ground truth over a forecast horizon of length T. For evaluation, the historical series, ground truth, and prediction are rendered together as a plot $\mathbf{V}$. For each evaluation dimension, textual context $\mathcal{C}$ specifies the evaluation criterion, its scoring rubric, and relevant task information. Conditioned on these inputs, the VLM judge returns:

$$(s, \mathcal{R}) \leftarrow \mathcal{P}_{\mathrm{VLM}}(\mathbf{V}, \mathcal{C}). \tag{1}$$

where $s \in \{1, \ldots, 5\}$ is a Likert score, and $\mathcal{R}$ is a textual justification for that rating.

3.2 Evaluation Protocol

The evaluation protocol comprises two complementary levels of assessment. Micro-Level evaluation assesses structural fidelity along four temporal patterns, whereas Macro-Level evaluation integrates operational constraints into a holistic judgment as illustrated in Figure 3(b).

3.2.1 Micro-Level

Micro-Level evaluation characterizes the discrepancy between a forecast and its ground truth along four complementary dimensions. These dimensions capture structural properties that are visually discernible in time series plots and are organized according to the conceptual representation:

$$y(t) = \mathcal{T}(t) + \mathcal{A}(t) \cdot \mathcal{M}\big(\mathcal{P}(t)\big). \tag{2}$$

The four dimensions are defined as follows:

- **Trend ($\mathcal{T}$):** The global low-frequency trajectory, capturing overall directional movement.
- **Magnitude ($\mathcal{A}$):** The vertical amplitude, representing the scale of temporal variations.
- **Motif ($\mathcal{M}$):** The local geometric shape, representing recurring patterns and sub-series.
- **Phase ($\mathcal{P}$):** The temporal alignment, capturing the alignment of peaks and troughs.

The scores distinguish structural strengths and weaknesses while providing complementary views of forecast quality. The rubrics in Appendix B provide operational definitions for each dimension.

3.2.2 Macro-Level

Macro-Level evaluation assesses forecast quality under a specified operational context and assigns a holistic *General* score. Real-world evaluation combines hard constraints, preferences for temporal patterns, and task-specific tolerances for forecasting errors. These priorities vary across different applications and are not represented by domain-agnostic numerical metrics, while textual rubrics make such priorities explicit, guiding VLM judges to assess overall forecast quality in context.

3.3 Dataset Construction

TimeVista comprises 5,563 time series windows with rubrics, each represented as:

$$\mathcal{S}_i = \big((\mathbf{x}_i, \mathbf{y}_i), \mathcal{C}_i\big), \tag{3}$$

Here $\mathbf{x}_i$ and $\mathbf{y}_i$ denote the historical lookback window and ground truth, and $\mathcal{C}_i$ specifies the rubric.

Time Series Collection We construct TimeVista using synthetic and real-world data to support complementary Micro- and Macro-Level evaluation. For Micro-Level evaluation, we generate 400 synthetic samples through a controllable generation process spanning diverse trends, motifs, phases, and magnitudes, enabling systematic assessment across the four evaluation dimensions. For Macro-Level evaluation, we collect and curate 5,163 real-world samples from established benchmarks across eight domains as presented in Figure 3. Synthesis Details are provided in Appendix A.2.

Rubric Construction Researchers with expertise in time series forecasting develop rubrics to specify priorities. Grounded in time series theory (Hamilton, 2020), the Micro-Level rubrics assess structural fidelity using visual cues such as slope deviations and peak-valley alignment. Macro-Level rubrics translate domain-specific requirements into scoring criteria. The rubrics appear in Appendix B, and Section 4.2 examines whether rubric-guided scores align with intuitive preferences.

Standardized Visualization For each forecast, we render the historical window, ground truth, and model prediction in a shared coordinate system. The default visualization uses fixed resolution, contrasting line styles and colors, labeled axes, aligned grids, and an explicit legend.

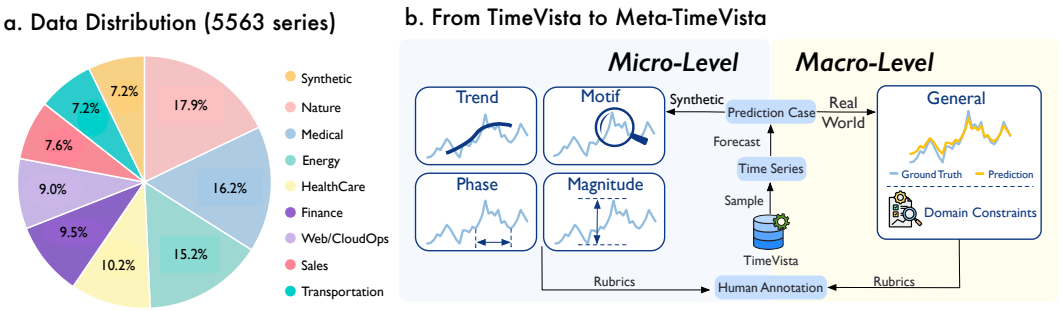


Figure 3: (a) TimeVista data distribution. (b) Meta-TimeVista construction.

4 Meta-Evaluation: Exploring VLMs as Judges

In this section, we conduct meta-evaluation to explore whether, and under what conditions, rubric-guided VLMs can provide reliable assessments of forecast quality that align with human preferences.

4.1 Experimental Setup

Meta-TimeVista We construct Meta-TimeVista to benchmark VLM judges against human annotations. Following the notation in Section 3.1, each meta-evaluation instance is represented as:

$$\begin{aligned} \mathcal{I}_{i,m} &= \big((\mathbf{x}_i, \mathbf{y}_i),\, \mathcal{C}_i,\, \hat{\mathbf{y}}_{i,m},\, s^{\mathrm{H}}_{i,m}\big), \\ \hat{\mathbf{y}}_{i,m} &= \mathcal{F}_{\theta_m}(\mathbf{x}_i). \end{aligned} \tag{4}$$

where $\hat{\mathbf{y}}_{i,m}$ is the zero-shot forecast generated by TSFM m for window i and $s^{\mathrm{H}}_{i,m}$ is the corresponding human-annotated score for the specified dimension. We sample time series from TimeVista and generate forecasts, yielding 480 synthetic and 545 real-world forecast instances. Five researchers with expertise in time series forecasting independently score each instance using the same five-point rubrics as the VLM judges. Their modal rating serves as the final human annotation.

Evaluation Protocol For meta-evaluation, the synthetic subset provides controlled structural variation for Micro-Level validation across four dimensions: Trend, Motif, Phase, and Magnitude. At the Macro-Level, real-world forecast instances are evaluated on the domain-aware General dimension. The 1,025 forecast instances thus yield 2,465 forecast-dimension evaluations. Within each dimension, we measure agreement with human annotations using Spearman's ρ and Kendall's τ.

Evaluated VLMs We evaluate six representative VLMs: GPT-5.4 (Achiam et al., 2023), Claude-Sonnet-4.6, Gemini-3.1-Pro (Team et al., 2023), LLaMA-4-Maverick (Touvron et al., 2023), Qwen-3.6-Plus (Yang et al., 2025), Seed-1.8 (Seed, 2026). The decoding temperature is set to 0 for all.

4.2 Validation of Human Annotations

Inter-Annotator Agreement The five annotators achieve a mean Krippendorff's α of 0.718 across the five evaluation dimensions, supporting the reliability of the aggregated human annotations.

Agreement with Intuitive Preferences To validate whether the rubrics align with human preferences, we sample pairs of two TSFM forecasts from Meta-TimeVista that share the same historical window and ground truth. The same five annotators independently select the better forecast or a tie for the specified dimension, without access to the rubrics or existing scores. We then compare their modal preference with the ordering implied by the rubric-guided human scores. Rubric-guided scores agree with intuitive human preferences in 88.29% of comparisons, supporting the rubrics' alignment with human preferences. Most disagreements arise from tied rubric scores rather than opposite forecast rankings. Detailed protocols and results are provided in Appendix C.2.

Table 1: Spearman (ρ) and Kendall-Tau (τ) correlations between VLMs and human annotations across Micro- and Macro-Level evaluation. VLMs exhibit strong alignment with human preferences.

Level	Micro-Level								Macro-Level	
Judge	Trend		Motif		Phase		Magnitude		General	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
Gemini-3.1-Pro	**0.859**	**0.771**	**0.870**	**0.796**	0.831	0.764	**0.907**	**0.849**	**0.865**	**0.820**
Qwen3.6-Plus	0.846	0.758	0.854	0.773	0.768	0.690	0.870	0.805	0.859	0.806
GPT-5.4	0.834	0.751	0.846	0.773	**0.854**	**0.781**	0.837	0.769	0.794	0.740
Claude-Sonnet-4.6	0.807	0.728	0.800	0.722	0.690	0.609	0.826	0.755	0.792	0.732
Seed-1.8	0.763	0.662	0.766	0.682	0.734	0.652	0.802	0.719	0.760	0.690
Llama-4-Maverick	0.750	0.668	0.729	0.645	0.704	0.629	0.763	0.680	0.674	0.608

4.3 Main Results

Alignment Performance Table 1 compares the six VLM judges in terms of their alignment with expert ratings across different evaluation dimensions. Gemini-3.1-Pro achieves the strongest overall alignment, while Qwen-3.6-Plus also delivers competitive performance. The results on domain-aware General further indicate that this alignment extends beyond individual temporal characteristics to overall forecast evaluation under specified domain requirements. Together, these results support the potential of rubric-guided VLMs to provide human-aligned assessments of forecast quality.

Comparison with Numerical Evaluators To validate the value of VLM judgments as a complement to numerical evaluation, we compare their alignment with human annotations against progressively more rubric-aware baselines. Structure-aware metrics quantify differences in temporal structure through established measures, such as distances between extracted trends, derivative-based shape comparisons, and temporal alignment errors. Rubric-based Rules translates the supplied rubrics into deterministic rules, mapping structural features and domain-constraint checks to five-point scores using predefined thresholds. This baseline tests whether the VLM's human-alignment advantage persists beyond explicitly encoding the same evaluation criteria.

Table 2: Human alignment of VLM and numerical evaluators. Error-based metrics (MSE, MAE, MASE, DTW, and structure-aware metrics) are negated so higher values indicate better forecasts.

Evaluator	Micro-Level								Macro-Level	
	Trend		Motif		Phase		Magnitude		General	
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	ρ	τ
VLM judge	**0.859**	**0.771**	**0.870**	**0.796**	**0.831**	**0.764**	**0.907**	**0.849**	**0.865**	**0.820**
MSE	0.406	0.313	0.303	0.227	0.268	0.201	0.371	0.282	-0.112	-0.086
MAE	0.416	0.321	0.312	0.234	0.278	0.208	0.383	0.292	-0.104	-0.079
MASE	0.679	0.529	0.563	0.428	0.548	0.417	0.636	0.497	0.308	0.235
DTW	0.442	0.340	0.330	0.247	0.290	0.217	0.406	0.309	-0.077	-0.059
Structure-aware metrics[†]	0.429	0.331	0.820	0.670	0.681	0.582	0.386	0.294	-0.066	-0.048
Rubric-based Rules	0.727	0.633	0.743	0.667	0.726	0.647	0.863	0.793	0.471	0.415

[†] Trend: HP error; Motif: DDTW; Phase: correlation lag; Magnitude: Wasserstein; General: Soft-DTW.

As shown in Table 2, the VLM judge achieves higher correlations with expert scores than the tested baselines across all five dimensions. Among the Micro-Level baselines in this table, structure-aware metrics perform best on Motif, while Rubric-based Rules performs best on Trend, Phase, and Magnitude, demonstrating the benefits of incorporating evaluation-specific criteria. On domain-aware General, Rubric-based Rules improves alignment to $\rho = 0.471$ but remains below the VLM judge, a gap that may reflect limitations in translating qualitative rubric criteria into fixed numerical features and thresholds. Historical-scale normalization improves the numerical baselines, while the VLM judge maintains its human-alignment advantage. Detailed results are provided in Appendix E.1.

4.4 Ablation Studies

We conduct ablation studies to assess the roles of input modality, scoring rubrics, and plot rendering in VLM-based evaluation. Unless otherwise specified, all analyses use Gemini-3.1-Pro.

Visual Modality To validate visual input's contribution, we compare visual-based evaluation with a numerical-text baseline. The judge and scoring rubrics are held fixed, while the same historical input $\mathbf{x}$, ground truth $\mathbf{y}$, and forecast $\hat{\mathbf{y}}$ are presented either as a plot $\mathbf{V}$ or as numerical text. The two formats are compared in terms of alignment with expert ratings and mean input tokens per judgment.

As shown in Figure 4(a), visual input achieves stronger alignment with expert ratings across all six forecast-length ranges, while keeping input token usage constant near 1,700. However, the input usage increases from approximately 4,300 to 20,700 for numerical text as forecast length grows. In the longest range, visual evaluation achieves $\rho = 0.834$, compared with 0.803 for numerical text, using approximately one-twelfth as many input tokens. Therefore, visual input provides a more token-efficient representation while maintaining higher alignment under the tested configuration.

Scoring Rubrics We examine sensitivity to rubric wording, presentation, and detail. As reported in Table 3, paraphrasing the original rubric leaves alignment largely unchanged. In contrast, rewriting the rubric as continuous prose reduces ρ to 0.809, which suggests a benefit of structured presentation. Independently, removing visual cues and thresholds while retaining the five score levels yields $\rho = 0.758$. These results suggest that some variation in the wording is compatible with strong alignment, while retaining detailed scoring criteria is important to maintaining alignment with expert ratings.

Table 3: Human alignment under rubric changes.

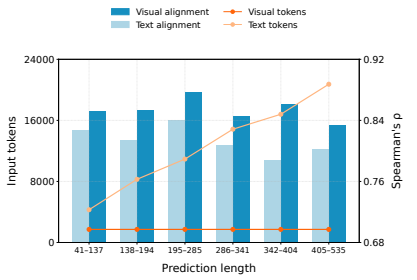
Setting	Configuration	ρ	τ
Synthetic *Magnitude*	Default	0.907	0.849
	Paraphrased	0.901	0.840
	Reformatted	0.809	0.733
	Simplified	0.758	0.693
Real-world *General*	Default	0.865	0.820
	Incorrect rules	-0.508	-0.434

We further test how judges respond to deliberately incorrect rules that contradict real-world domain constraints. Under these rules, alignment with the original human annotations drops from $\rho = 0.865$ to -0.508, indicating that supplied instructions can drive judgments away from domain-consistent assessments. This suggests that VLM judges may prioritize explicit scoring instructions over their prior domain knowledge, even when the supplied rules contradict real-world constraints.

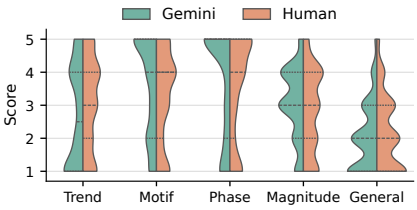
Table 4: Human alignment under rendering perturbations on synthetic *Magnitude*.

Default configuration: $\rho = 0.907$, $\tau = 0.849$							
Factor	**Configuration**	ρ	τ	**Factor**	**Configuration**	ρ	τ
Aspect ratio	6:5	0.896	0.836	Plot annotations	Axes removed	0.906	0.850
	18:5	0.903	0.846		Legend removed	0.895	0.836
	24:5	0.907	0.851		Legend text swapped	0.898	0.838
Vertical padding	50%	0.917	0.862	Resolution	50 DPI	0.902	0.843
	100%	0.912	0.855		300 DPI	0.896	0.835
Markers	None	0.908	0.851	Markers	Every 3 steps	0.906	0.846

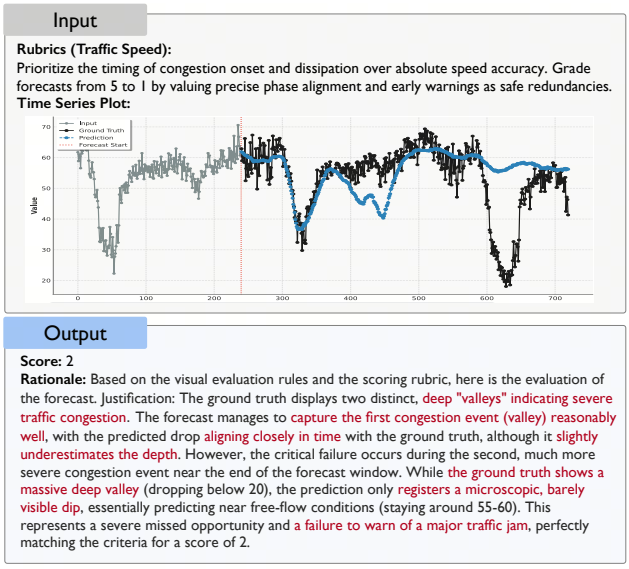
Plot Rendering We also explore the impact of stylistic changes on the judgments of the same forecast. As presented in Table 4, changes to aspect ratio, vertical padding, resolution, and marker placement yield correlations close to the default values on the synthetic *Magnitude* evaluation task, suggesting that judgments do not depend on a particular level of visual detail. Removing axis annotations or modifying legends also causes only small changes in alignment with expert ratings. Beyond these individual perturbations, randomizing colors and line styles across the five evaluation dimensions leaves 80.53% of scores unchanged as shown in Figure 7 in the Appendix. This complements the correlation results by showing stability at the individual-score level. Together, these findings suggest that VLM judges tolerate common changes in visual presentation, allowing flexibility in plot design without substantially altering evaluation outcomes under the tested conditions.



(a) Input modality comparison



(b) Score distributions



(c) Traffic speed judgment

Figure 4: Analysis of Gemini-3.1-Pro as a forecasting judge. (a) Comparison of alignment and efficiency between vision-based and text-based evaluation across different prediction lengths. (b) Distributions of Gemini scores and human ratings. (c) A traffic speed forecasting example.

4.5 Beyond Correlation

In this section, we focus beyond correlation to score distributions and the explanations accompanying individual judgments. Figure 4(b) compares the score distributions of Gemini-3.1-Pro and human annotators on different dimensions. Gemini assigns a score of 5 more frequently for Motif and Phase, whereas both predominantly assign low scores for General.

To demonstrate the role of domain-specific criteria in an individual assessment, we provide a traffic speed case in Figure 4(c). The judge assigns a score of 2, noting that the forecast captures the first congestion event but substantially underestimates the second. Under the domain rubric, capturing these events matters more than matching speed values during stable periods. By linking the identified error to this evaluation priority, the rationale provides an explicit justification for the score that can be checked against the plotted forecast. Additional examples are provided in Appendix E.6.

5 Evaluation: Exploiting VLMs as Judges

5.1 Experiment Setup

We use Gemini-3.1-Pro, which leads in most evaluation dimensions, to assess 13 forecasting models: four statistical baselines (SMA (Johnston et al., 1999), Seasonal Naive, AutoARIMA (Shumway & Stoffer, 2017), and AutoETS (Petropoulos et al., 2022)); eight TSFMs (Chronos (Ansari et al., 2024), Sundial (Liu et al., 2025b), Timer-S1 (Liu et al., 2026), Chronos2 (Ansari et al., 2025), TimesFM-2.5 (Das et al., 2023), TiRex (Auer et al., 2026), Flowstate (Graf et al., 2025), and Moirai2 (Liu et al., 2025a)); and the multimodal model Aurora (Wu et al., 2025). Model details are in Appendix D.1. We use these judgments for offline forecast analysis, examining structural fidelity on GIFT-Eval and domain-specific requirements on the real-world subset of TimeVista.

5.2 Micro-Level Evaluation on GIFT-Eval

We apply the four Micro-Level rubrics to assess existing time series models. The evaluation covers 846 randomly sampled time series from GIFT-Eval (Aksu et al., 2024), with scores averaged across series for each model and dimension. As shown in Figure 5, model rankings vary across evaluation dimensions. Chronos2 and Moirai2 receive the highest Trend scores, whereas Seasonal Naive

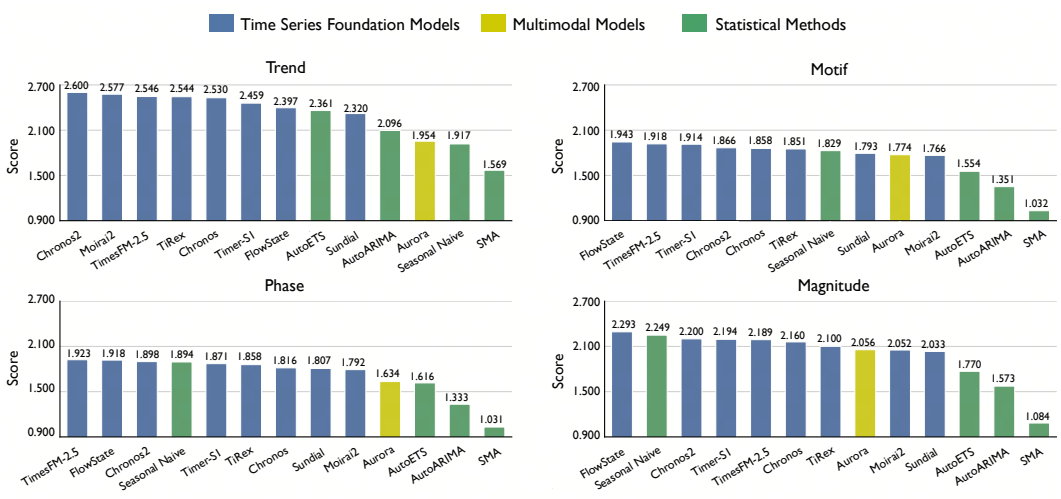


Figure 5: Micro-Level evaluation of 13 representative forecasting models on GIFT-Eval benchmark.

remains competitive on Phase and Magnitude. These differences expose complementary structural strengths that a single aggregate score can obscure, identifying phase and magnitude fidelity as targets for further model analysis. This application extends the Micro-Level rubrics to real-world forecasts; scores on this subset are VLM judgments without additional human annotations.

5.3 Macro-Level Evaluation on TimeVista

At the Macro-Level, focus shifts from individual temporal patterns to overall forecast quality under domain-specific requirements. The evaluation covers 945 time series randomly sampled from the real-world subset of TimeVista, with each model's General scores averaged across series. As shown in Figure 6, Chronos2 achieves the highest average score of 2.188 on the five-point scale. This indicates substantial room for improvement under these criteria. For example, negative solar-production forecasts violate an explicit domain requirement even when they remain numerically close to the ground truth.

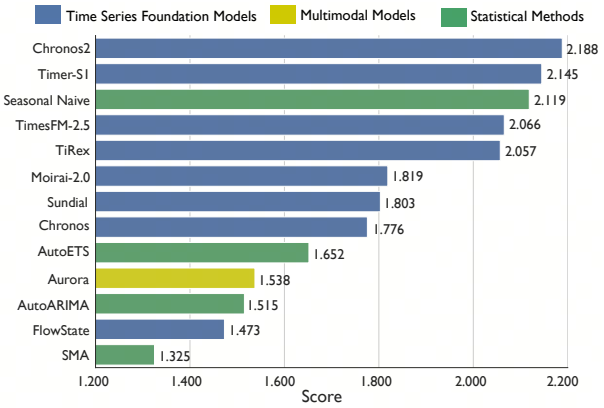


Figure 6: Macro-Level evaluation on TimeVista.

Seasonal Naive, a simple seasonal baseline, ranks third overall, demonstrating the value of preserving historical periodicity in practical deployment. Aurora scores below several unimodal models despite supporting textual context, illustrating that multimodal capability alone does not ensure adherence to the specified requirements. Together, the two evaluation levels provide complementary views of forecast behavior: Micro-Level scores distinguish structural strengths, while Macro-Level scores highlight adherence to domain-specific constraints and priorities.

6 Conclusion

We explore VLMs as judges for time series forecasting through TimeVista and Meta-TimeVista. Meta-evaluation shows strong human alignment under the specified rubrics, while sensitivity analyses distinguish tolerance to rendering changes from dependence on scoring criteria. Exploiting these judgments across 13 forecasting models reveals complementary structural strengths and differences in satisfying domain-specific requirements. Together, these findings support rubric-guided visual judgments as additional feedback alongside MSE and MAE, and provide a basis for further investigation of criterion design and judge generalization.

AI USE STATEMENT

This work evaluates vision-language models (VLMs) as judges for time series forecasting. Generative AI tools also assisted with language editing of the manuscript. The authors reviewed the AI-assisted materials and take responsibility for the final evaluation rubrics, benchmark data, experimental results, analyses, and manuscript.

ETHICS STATEMENT

This work includes a human annotation study involving five time series experts. All annotators provided informed consent and received compensation for their annotation work.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Md Atik Ahamed, Mihir Parmar, Palash Goyal, Chun-Liang Li, Qiang Cheng, Tomas Pfister, and Jinsung Yoon. Reasoning-aware training for time series forecasting. *arXiv preprint arXiv:2605.08625*, 2026.

Taha Aksu, Gerald Woo, Juncheng Liu, Xu Liu, Chenghao Liu, Silvio Savarese, Caiming Xiong, and Doyen Sahoo. Gift-eval: A benchmark for general time series forecasting model evaluation. *arXiv preprint arXiv:2410.10393*, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*, 2024.

Abdul Fatir Ansari, Oleksandr Shchur, Jaris Küken, Andreas Auer, Boran Han, Pedro Mercado, Syama Sundar Rangapuram, Huibin Shen, Lorenzo Stella, Xiyuan Zhang, et al. Chronos-2: From univariate to universal forecasting. *arXiv preprint arXiv:2510.15821*, 2025.

Andreas Auer, Patrick Podest, Daniel Klotz, Sebastian Böck, Günter Klambauer, and Sepp Hochreiter. Tirex: Zero-shot forecasting across long and short horizons with enhanced in-context learning. *Advances in Neural Information Processing Systems*, 38:57529–57580, 2026.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. Benchmarking foundation models with language-model-as-an-examiner. *Advances in Neural Information Processing Systems*, 36:78142–78167, 2023.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*, 2024.

Zhiyong Cui, Ruimin Ke, Ziyuan Pu, and Yinhai Wang. Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction. *arXiv preprint arXiv:1801.02143*, 2018.

Marco Cuturi and Mathieu Blondel. Soft-dtw: a differentiable loss function for time-series. In *International conference on machine learning*, pp. 894–903. PMLR, 2017.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. *arXiv preprint arXiv:2310.10688*, 2023.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4918–4927, 2024.

Jinlan Fu, See Kiong Ng, Zhengbao Jiang, and Pengfei Liu. Gptscore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6556–6576, 2024.

Peyré Gabriel and Cuturi Marco. Computational optimal transport with applications to data sciences. *Foundations and trends® in machine learning*, 11(5-6):355–607, 2019.

Lars Graf, Thomas Ortner, Stanisław Woźniak, and Angeliki Pantazi. Flowstate: Sampling-rate invariant time series foundation model with dynamic forecasting horizons. In *Recent Advances in Time Series Foundation Models Have We Reached the'BERT Moment'?*, 2025.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. A survey on llm-as-a-judge. *The Innovation*, 7(6), 2026.

James D Hamilton. *Time series analysis*. Princeton university press, 2020.

Robert J Hodrick and Edward C Prescott. Postwar us business cycles: an empirical investigation. *Journal of Money, credit, and Banking*, pp. 1–16, 1997.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. Time-llm: Time series forecasting by reprogramming large language models. In *International conference on learning representations*, volume 2024, pp. 23857–23880, 2024.

FR Johnston, John E Boyland, Maureen Meadows, and E Shale. Some properties of a simple moving average when applied to forecasting a time series. *Journal of the Operational Research Society*, 50(12):1267–1271, 1999.

Eamonn J Keogh and Michael J Pazzani. Derivative dynamic time warping. In *Proceedings of the 2001 SIAM international conference on data mining*, pp. 1–11. SIAM, 2001.

Charles Knapp and Glifford Carter. The generalized correlation method for estimation of time delay. *IEEE transactions on acoustics, speech, and signal processing*, 24(4):320–327, 1976.

Guokun Lai, Wei-Cheng Chang, Yiming Yang, and Hanxiao Liu. Modeling long-and short-term temporal patterns with deep neural networks. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pp. 95–104, 2018.

Vincent Le Guen and Nicolas Thome. Shape and time distortion loss for training deep time series forecasting models. *Advances in neural information processing systems*, 32, 2019.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11286–11315, 2024.

Yuxuan Liang, Haomin Wen, Yuqi Nie, Yushan Jiang, Ming Jin, Dongjin Song, Shirui Pan, and Qingsong Wen. Foundation models for time series analysis: A tutorial and survey. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 6555–6565, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Chenghao Liu, Taha Aksu, Juncheng Liu, Xu Liu, Hanshu Yan, Quang Pham, Silvio Savarese, Doyen Sahoo, Caiming Xiong, and Junnan Li. Moirai 2.0: When less is more for time series forecasting. *arXiv preprint arXiv:2511.11698*, 2025a.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the association for computational linguistics*, 12:157–173, 2024.

Yong Liu, Guo Qin, Zhiyuan Shi, Zhi Chen, Caiyin Yang, Xiangdong Huang, Jianmin Wang, and Mingsheng Long. Sundial: A family of highly capable time series foundation models. *arXiv preprint arXiv:2502.00816*, 2025b.

Yong Liu, Xingjian Su, Shiyu Wang, Haoran Zhang, Haixuan Liu, Yuxuan Wang, Zhou Ye, Yang Xiang, Jianmin Wang, and Mingsheng Long. Timer-s1: A billion-scale time series foundation model with serial scaling. *arXiv preprint arXiv:2603.04791*, 2026.

John A Miller, Mohammed Aldosari, Farah Saeed, Nasid Habib Barna, Subas Rana, I Budak Arpinar, and Ninghao Liu. A survey of deep learning and foundation models for time series forecasting. *arXiv preprint arXiv:2401.13912*, 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Fotios Petropoulos, Daniele Apiletti, Vassilios Assimakopoulos, Mohamed Zied Babai, Devon K Barrow, Souhaib Ben Taieb, Christoph Bergmeir, Ricardo J Bessa, Jakub Bijak, John E Boylan, et al. Forecasting: theory and practice. *International Journal of forecasting*, 38(3):705–871, 2022.

Zhongzheng Qiao, Sheng Pan, Anni Wang, Viktoriya Zhukova, Yong Liu, Xudong Jiang, Qingsong Wen, Mingsheng Long, Ming Jin, and Chenghao Liu. It's time: Towards the next generation of time series forecasting benchmarks. *arXiv preprint arXiv:2602.12147*, 2026.

CLEVELAND RB. Stl: A seasonal-trend decomposition procedure based on loess. *J Off Stat*, 6: 3–73, 1990.

Hiroaki Sakoe and Seibi Chiba. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing*, 26(1):43–49, 1978.

Bytedance Seed. Seed1. 8 model card: Towards generalized real-world agency. *arXiv preprint arXiv:2603.20633*, 2026.

Oleksandr Shchur, Abdul Fatir Ansari, Caner Turkmen, Lorenzo Stella, Nick Erickson, Pablo Guerron, Michael Bohlke-Schneider, and Yuyang Wang. fev-bench: A realistic benchmark for time series forecasting. *arXiv preprint arXiv:2509.26468*, 2025.

Robert H Shumway and David S Stoffer. Arima models. In *Time series analysis and its applications: with R examples*, pp. 75–163. Springer, 2017.

Zhijiang Tang, Jiaxin Qi, Yuhua Zheng, and Jianqiang Huang. A comprehensive benchmark for electrocardiogram time-series. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 6490–6499, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Ruey S Tsay. *Analysis of financial time series*. John wiley & sons, 2005.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Yong Liu, Chen Wang, Mingsheng Long, and Jianmin Wang. Deep time series models: A comprehensive survey and benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. Unified training of universal time series forecasting transformers. In *Forty-first International Conference on Machine Learning*, 2024.

Haixu Wu, Hang Zhou, Mingsheng Long, and Jianmin Wang. Interpretable weather forecasting for worldwide stations with a unified deep model. *Nature Machine Intelligence*, 5(6):602–611, 2023.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. *arXiv preprint arXiv:2509.22295*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

APPENDIX OVERVIEW

The appendix documents how TimeVista is constructed, how its judgments are validated, and which evaluation choices affect the results. Appendices A–B specify the data and scoring rubrics; Appendices C–D describe human validation and experimental implementation. Appendix E examines numerical scale, explicit rule evaluation, rubric design, rendering, and repeated judgments, with quantitative comparisons and illustrative cases.

A DATA CONSTRUCTION

TimeVista combines real-world sources with synthetic series to support domain-aware and structural evaluation. Table 5 summarizes source coverage, sequence lengths, and temporal frequencies. The following subsections describe source-specific curation, the controllable synthesis process, and the counting units used in each experiment. We will publicly release our dataset upon acceptance.

A.1 REAL-WORLD SOURCES AND CURATION

Table 5: Data sources and temporal characteristics of TimeVista.

Dataset	Domain	Input Length	Prediction Length	Frequency	# Series
CPSC2019	Medical	500	100	S	299
PTB	Medical	500	100	S	300
RDBH	Medical	500	100	T	300
FNSPID	Finance	42	7	D	500
CPHL	Nature	1818–17542	12, 336	15T, 30T, H	494
Supply Chain Customer	Sales	1642–1972	30	D	424
Australia Solar	Energy	32544–35040	24, 168	H	448
Azure2019	Web/CloudOps	594–8351	48, 288	5T	500
Crypto	Finance	2569–2809	30	D	29
Current Velocity	Nature	1536–31849	9, 432	10T, 15T, 20T, 5T, H	500
Covid Death	HealthCare	200–468	21	D	569
Loop Seattle	Transportation	7944–104256	48, 480	5T, H	400
Solar	Energy	8328–8712	48	H	400
Synthetic	Synthetic	107–1720	32, 601	-	400

Medical and healthcare. ECG windows are drawn from CPSC2019, PTB, and RDBH (Tang et al., 2025), covering different recording durations and signal configurations. Covid Death (Dong et al., 2020) provides daily cumulative death counts from the Johns Hopkins University repository, spanning January 22, 2020 to May 29, 2021 across countries and subnational regions.

Finance. FNSPID (Dong et al., 2024) provides historical stock prices alongside financial news. Crypto (Qiao et al., 2026) contains daily prices for Bitcoin, Ethereum, Litecoin, and Bitcoin Cash from December 20, 2017 to September 30, 2025.

Nature. CPHL (Qiao et al., 2026) measures chlorophyll concentration at four Australian marine stations. Current Velocity (Qiao et al., 2026) contains ocean-current and hydrographic measurements from the National Mooring Network Facility across five Australian states. Both sources have heterogeneous minute-level sampling frequencies. Their preprocessing includes hourly resampling, with excessively short Current Velocity sequences excluded.

Energy. Solar (Lai et al., 2018) records production at 137 photovoltaic plants in Alabama during 2006, originally sampled every ten minutes. Australia Solar (Qiao et al., 2026) provides hourly production records from three photovoltaic installations.

Transportation. Loop Seattle (Cui et al., 2018) contains measurements from roadway loop detectors on four Seattle-area freeways. We retain traffic speed observations from the DRIVE Net system.

Sales and cloud workloads. Supply Chain Customer (Qiao et al., 2026) contains five years of simulated delivery transactions, aggregated into daily customer-level order counts, volumes, and revenue. Azure2019 (Qiao et al., 2026) provides CPU utilization traces at five-minute intervals across delay-insensitive, interactive, and unclassified workloads.

A.2 Synthetic Series

Synthetic generation provides configurable coverage of temporal structures for Micro-Level evaluation. Building on *kernel-synth* (Ansari et al., 2025), we sample Gaussian processes that combine slowly varying trajectories with recurring local patterns. Smooth and linear components generate gradual rises, declines, and turning points; periodic components generate repeated peaks and troughs. Combining temporal scales produces waveforms beyond a single periodic pattern.

Component strengths control the balance between global trajectories and local fluctuations, periodicity controls event spacing, and amplitude variation controls fluctuation size. Sampling component combinations and sequence lengths therefore spans diverse trends, motifs, phases, and magnitudes within a common generation procedure.

A.3 Evaluation Subsets and Counting Units

We distinguish source windows, forecast instances, and forecast–dimension evaluations. Meta-TimeVista contains 480 synthetic and 545 real-world forecast instances. Evaluating each synthetic forecast on four Micro-Level dimensions and each real-world forecast on General yields $480 \times 4 + 545 = 2{,}465$ evaluations. Multiple model forecasts can share a source window.

The application study in Section 5.1 uses 846 source series from GIFT-Eval and 945 from the real-world portion of TimeVista. Each of the 13 forecasting models is evaluated on these series, with scores averaged across series within each dimension. This study compares forecasting models, whereas Meta-TimeVista compares evaluators against human annotations.

B Scoring Rubrics and Prompt Templates

The rubrics provide operational definitions of the evaluation dimensions through explicit five-point criteria. Each prompt combines a shared system instruction with criteria developed by researchers with expertise in time series forecasting. Micro-Level rubrics specify structural properties; Macro-Level rubrics specify domain requirements for General scoring. The templates below retain the scoring instructions and thresholds used in evaluation, making the criteria underlying both human and VLM judgments explicit.

B.1 Shared System Instruction

System Prompt

You are an expert evaluator in time series forecasting. You will be provided with a chart displaying historical data, ground truth, and model predictions. Your task is to evaluate the prediction quality based on a specific dimension and assign a score from 1 to 5. Please output your reasoning and the score strictly in the format: [[Score: X]], where X is an integer from 1 to 5.

B.2 MICRO-LEVEL RUBRICS

B.2.1 TREND

User Prompt: Trend

Evaluate the consistency of the [Macro Trend] (the overall baseline trajectory and major turning points).
Visual Evaluation Rubrics:
Imagine a smoothed "center of gravity" line for both curves. Compare these Macro paths.You may IGNORE minor local zig-zag noise or recurring periodic waves **IF AND ONLY IF** they are small in amplitude and ride on top of a clear, discernible Macro-trend line. **HOWEVER**, if the noise itself is of high amplitude and dominates the visual appearance of the prediction, obscuring the underlying trend, then you **MUST NOT** ignore it; instead, it should be heavily penalized according to the rubric's descriptions for scores 1-2. FOCUS ONLY on the overall direction (up/down/flat) and the timing of major Macro bends/turns.

Scoring Rubric:
5 (Perfect Fit): The overall visual baselines, movement directions, and major turning points align perfectly.
4 (Structurally Intact with Minor Flaws): Captures the primary direction and major turns. For monotonic trends, the slope angle has a minor deviation (the vertical gap between the two curves at the end of the chart is SMALLER than the total vertical height the ground truth climbed/dropped). For complex trends, a major turn happens slightly earlier or later.
3 (Noticeable Trend Degradation): Matches the trend in some parts but has severe visual flaws. Severe slope distortion (the vertical gap at the end is LARGER than the ground truth's total height change), or it visually misses a secondary Macro bend/turn.
2 (Significant Deviation or Excessive Noise): For 30%-50% of the horizontal axis, the prediction moves in the completely opposite direction (e.g., goes up while GT goes down) or flattens out horizontally (mean reversion). OR, hallucinates a mild slope when GT is almost visually flat. **OR, the prediction is visually dominated by high-frequency, seemingly random noise, making its underlying Macro trend difficult to discern or unreliable, even if some directional changes are occasionally matched.**
1 (Completely Contradictory or Degenerated to Pure Noise): For >50% of the horizontal axis, the curves move in opposite directions (e.g., forming an "X" shape) or the prediction globally flattens out. OR, hallucinates a strong, massive trend on a completely flat/stable ground truth. **OR, the prediction has completely degenerated into pure, seemingly random noise, making any underlying trend entirely unidentifiable or meaningless across the majority of the forecast horizon. It offers no reliable Macro directional information.**

B.2.2 MOTIF

User Prompt: Motif

Evaluate the consistency of [Motif] (the local geometric shape, waveform, or repeating structural unit).
Visual Evaluation Rubrics:
1. Treat the local pattern as a "Shape Template".
2. STRICTLY IGNORE horizontal shifts (phase/timing offsets) and global vertical stretching (absolute amplitude).
3. FOCUS ONLY on internal relative geometry: Check the peak-valley alternating rhythm, the relative height/width between adjacent peaks, and the sharpness/smoothness of the curves.
4. [Exception for Amplitude Collapse]: This rule overrides Rule #2 in extreme cases. If the forecast's amplitude is so severely compressed that its fluctuations become visually indistinguishable from random noise OR it fails to represent the essential structural differences between major peaks and minor valleys present in the ground truth, then the motif is considered broken. In such cases, rate as Severe Structural Mismatch (Score 2) or Complete Collapse (Score 1).
Scoring Rubric:
5 (Perfect Replication): The visual "shape templates" match exactly. Perfectly captures both the Macro skeleton (peak-valley order) and Micro-details (sharpness, local kinks).
4 (Structurally Intact with Minor Flaws): The core visual skeleton is correct (the sequence of peaks

User Prompt: Motif (continued)

and valleys matches perfectly). Allowed visual flaws: Slight smoothing (sharp peaks become rounded), minor top-clipping (flattened peaks), or slight distortions in relative proportions that do not change the recognizable shape.
3 (Noticeable Shape Degradation): The basic up-and-down fluctuation exists, but the internal template is broken. Visual cues: Missing secondary peaks/valleys, severe distortion of internal relative proportions (e.g., a dominant peak becomes tiny), or the shape visibly degrades into simpler waves for 30%-50% of the timeline.
2 (Severe Structural Mismatch): Visual cues: The predicted shape is completely wrong despite matching the duration; OR the rhythm/frequency is completely mismatched (e.g., predicting dense waves when GT is sparse); OR the shape degrades into generic noise for 50%-80% of the timeline.
1 (Complete Collapse): No structured local pattern exists. The prediction visually collapses into a featureless flat line, a simple straight trend, or pure random static.

B.2.3 PHASE

User Prompt: Phase

Evaluate the consistency of [Phase] (horizontal time alignment, frequency rhythm, and timing of events).
Visual Evaluation Rubrics:
1. Macro Rhythm First: Check the overall frequency (are the waves/events packed at the same density?). This is a prerequisite.
2. Micro Timing Second: Check the horizontal (left-right) alignment of specific peaks, valleys, and sudden changes.
3. STRICTLY IGNORE vertical height differences (amplitude) or overall vertical shifts. Focus purely on the X-axis alignment.
Scoring Rubric:
5 (Perfect Alignment): The Macro frequency matches perfectly. All key peaks, valleys, and sharp turns align exactly on the vertical axis with no left/right shift (no lag or lead).
4 (Accurate Rhythm with Minor Flaws): The overall frequency matches the ground truth. Most inflection points align precisely. Allowed visual flaws: May miss a few minor, non-key bumps, or exhibit very slight left/right timing shifts in less than 30% of the timeline.
3 (Noticeable Phase Shift or Degradation): The general frequency/rhythm is correct, but has noticeable visual flaws. Visual cues: (1) Systematic Shift: The predicted curve is visibly shifted left or right (overall lag/lead) but is NOT completely out-of-phase (i.e., peaks do not align with valleys); (2) Missing Details: Misses a significant number of minor bumps, though major peaks still align; (3) Local Degradation: Loses rhythm or is visibly misaligned in 30%-50% of the timeline.
2 (Severe Frequency Mismatch): Visual cues: The Macro rhythm is visibly wrong (e.g., predicted waves are much wider, narrower, or sparser than the ground truth). Even if a few peaks accidentally line up, it must be penalized if the overall wave density/frequency is incorrect.
1 (Complete Loss of Rhythm): Visual cues: For $>$50% of the timeline, the prediction completely loses its rhythmic or event-driven features, visually degrading into a flat line, a smooth monotonic trend, or pure random noise.

B.2.4 MAGNITUDE

User Prompt: Magnitude

Evaluate the consistency of [Magnitude] (the relative vertical stretch/amplitude and the absolute vertical height/bias).
Visual Evaluation Rubrics:
1. The "Lowest Score" Rule (CRITICAL): You MUST evaluate two sub-dimensions: (A) Relative Amplitude (the peak-to-valley vertical stretch) and (B) Absolute Height (the overall vertical shift/gap between the curves). Your final score MUST be the LOWER of these two evaluations.
2. Ignore Rare Outliers: Do not heavily penalize the model for missing 1 or 2 isolated, extreme spikes. Focus on the magnitude of the regular/bulk fluctuations.

User Prompt: Magnitude (continued)

Scoring Rubric:
5 (Perfect Fit): Both sub-dimensions are nearly perfect. (1) Amplitude: The peak-to-valley vertical stretch matches perfectly. (2) Height: No visible vertical shift (any gap is $<$10% of the average wave height).
4 (Good, Minor Flaws): The lowest of the two sub-dimensions is at level 4. (1) Amplitude: Visibly similar in scale, but slightly compressed or expanded (captures $>$2/3 of the true stretch). (2) Height: Visible but small vertical shift. The vertical fluctuation ranges (bounding boxes) of the two curves heavily OVERLAP.
3 (Severe Smoothing OR Severe Shift): The lowest of the two sub-dimensions is at level 3. (1) Amplitude (Over-smoothed): Severely compressed. The predicted waves look noticeably flatter (captures approximately 1/3 to 2/3 of the true stretch). (2) Height (Severe Shift): Massive vertical gap. The bottom valleys of one curve barely touch the middle/top of the other (very little overlap in their vertical ranges). (3) Local Degradation: Amplitude dies out or shifts drastically in 30%-50% of the timeline.
2 (Extreme Smoothing, but still some visible fluctuation): The lowest of the two sub-dimensions is at level 2. (1) Amplitude (Extreme Smoothing): The ground truth has massive waves, and the prediction shows some visible, albeit very small, fluctuations. The predicted amplitude is clearly present but captures significantly less than 40% of the true stretch. (2) Height (Extreme Disconnect): The two curves are completely separated vertically with ZERO OVERLAP in their Y-axis ranges. (3) Severe Degradation: Amplitude dies out to a flat line in 50%-80% of the timeline.
1 (Collapse or Hallucination): Visual cues: The forecast's fluctuations are so minor that they appear as 'texture' or 'noise' rather than distinct, individual waves. If you look at the forecast line and cannot clearly identify separate, meaningful up-and-down wave patterns that mirror the *intent* of the ground truth, it is a collapse. OR, Amplitude (Near Collapse): The prediction shows almost no vertical stretch, appearing as a nearly flat line, even if it follows the general trend.

B.3 Macro-Level Rubrics

The following rubrics cover the real-world sources in Table 5. Solar and Australia Solar share the same rubric, reproduced once below. Dataset-specific priorities are retained as written in the evaluation prompts.

B.3.1 Cumulative Deaths

User Prompt: Cumulative Deaths

Evaluate the prediction of [COVID-19 Cumulative Deaths].

Evaluation Criteria:
1. Monotonicity (Physical Constraint): Cumulative data should be non-decreasing. Differentiate between a macro-level failure (sustained downward trend) and a micro-level flaw (local zigzags between adjacent points).
2. Error Direction (Risk Assessment): Overestimation (prediction $>$ ground truth) acts as a safety buffer for medical resources. Underestimation (prediction $<$ ground truth) poses severe risks of resource shortages.
3. Error Magnitude (Fit Quality): Assess whether the gap between the prediction and ground truth is negligible, stable (bounded), or diverging.

Scoring Rubric:

5 (Excellent / Near-Perfect Fit):
Monotonicity: Strictly non-decreasing.
Deviation: Visually indistinguishable from the ground truth. Any deviations are marginal, non-compounding, and do not alter the overall trajectory.

4 (Good / Safe Overestimation):
Monotonicity: Strictly non-decreasing.
Deviation: Consistently overestimates (prediction $>$ ground truth), but the vertical gap remains stable and bounded. The predicted curve is structurally parallel to the ground truth, representing a safe and reasonable resource buffer.

User Prompt: Cumulative Deaths (continued)

3 (Fair / Sub-optimal Fit):
Monotonicity: Generally non-decreasing. (Strictly NO visible downward drops; flat segments are acceptable).
Deviation: Falls into one of two specific cases:
(a) Extreme Overestimation: The positive gap is excessively large or exponentially diverging, indicating unrealistic resource allocation.
(b) Mild Underestimation: The prediction is slightly below the ground truth but strictly preserves the correct trend shape without diverging further.

2 (Poor / Severe Underestimation or Noisy Output):
Falls into one of two cases:
(a) Severe Underestimation: Consistently and significantly underestimates (prediction $<$ ground truth), leading to high-risk resource shortages.
(b) Local Non-monotonicity: The overall macro trend is flat or upward, but the curve contains visible local downward fluctuations (zigzags/dips) between adjacent points, showing a lack of physical constraints.

1 (Failure / Macro Physical Violation):
Monotonicity: The predicted curve exhibits a sustained downward slope or a gradual negative drift across the forecast horizon (e.g., the end point is visibly lower than the start point). The model completely fails to learn the fundamental macro-level constraint that cumulative deaths cannot decrease.

B.3.2 ELECTROCARDIOGRAMS

User Prompt: Electrocardiograms

Evaluate the prediction of [ECG (Electrocardiogram) Signals].

Visual Evaluation Rules:
1. Count the Spikes (Rhythm): Count the most prominent, sharpest peaks (QRS complexes) in both the ground truth and the prediction. Are the total counts exactly the same?
2. Check the Sharpness (Morphology): Look at the predicted spikes. Are they sharp and needle-like (V-shaped/inverted V) or rounded and blunt (U-shaped/sine-wave)?
3. Check the Alignment and Baseline: Do the spikes align horizontally? Is the baseline stable or drifting?

Scoring Rubric:

5 (Perfect Diagnosis):
Visual cues: The prediction perfectly replicates the rhythm and the detailed P-QRS-T morphology. The peaks align perfectly on the time axis, and the sharpness of the spikes matches the ground truth exactly.

4 (Clinically Safe - Minor Flaws):
Visual cues: The total count of the main spikes matches exactly (correct heart rate). It accurately captures the rhythm. It may have minor baseline drift, slight vertical amplitude errors, or very small left/right phase shifts, but these do not affect the clinical heartbeat count.

3 (Diagnostic Ambiguity - Morphological Degradation):
Visual cues: The total count of the main spikes is EXACTLY the same as the ground truth, BUT the local shapes are visibly degraded. The sharp QRS spikes become noticeably wider or blunter, or the smaller secondary waves (P/T waves) are completely lost, creating a risk of misdiagnosis.

2 (High-Risk Misleading - Rhythm or Polarity Error):
Visual cues: Fatal clinical errors. The model predicts MORE or FEWER main spikes than the ground truth (rhythm/heart rate error). OR, the peaks are flipped upside down (polarity reversal). OR, the sharp spikes are extremely smoothed out into gentle, rolling sine waves.

1 (Loss of Biological Meaning):
Visual cues: The prediction completely loses the key fluctuations of the true signal. It degenerates into a flat horizontal line, a simple monotonic trend, pure random noise, or waves that have absolutely no structural match to the ground truth.

B.3.3 STOCK PRICES

User Prompt: Stock Prices

Evaluate the prediction of [daily closing price for a stock].

[Visual Evaluation Rules]
1. Trend over Amplitude: Generally focus on directional slopes. However, massive vertical gaps/level shifts between the last past value and the future ground truth MUST NOT be ignored.
2. Macro Trend Priority: The OVERALL macro direction (start to end) dictates the primary evaluation. A local dip or correction within a strong UPWARD macro trend is NOT a "fatal crash".
3. Downside Risk & Gap Downs: Missing a sudden crash OR a significant "gap down" (where ground truth drops heavily below the last known past value and stays low) is a fatal failure. A valid downward prediction must show a VISIBLY CLEAR downward slope or a corresponding lower level.

[Scoring Rubric]

5 (Perfect): Captures the macro trend AND all major turning points/local wiggles perfectly.

4 (Strong): Captures the macro trend AND the most prominent turning point, but misses minor wiggles or has slight timing delays.

3 (Direction Consistent): Overarching macro direction matches (both UP or both DOWN), but misses exact turning points.
Note: If ground truth crashes but prediction shows a CLEAR but gradual decline, score 3.

2 (Missed Trend / Flatline Failure): Ground truth shows a clear macro trend (goes UP, or shows a gradual/volatile DOWNWARD movement that is not a sudden crash), but prediction goes the opposite way, stays strictly FLAT, or shows only an insignificant/nearly flat slope. The model completely misses the macro movement and provides an uninformative prediction.

1 (Fatal Crash Miss): Ground truth crashes DOWN severely OR gaps down significantly below the last past value, but prediction goes UP, stays FLAT at a higher level, or shows only an insignificant decline (If the prediction has a visually obvious downward slope, even if gentle, it MUST NOT be scored 1). The model completely fails to warn of severe downside risk.

B.3.4 CRYPTOCURRENCY PRICES

User Prompt: Cryptocurrency Prices

Evaluate the prediction of [Cryptocurrency Daily Price].

[Visual Evaluation Rules]
1. Trend over Amplitude: Generally focus on directional slopes. However, massive vertical gaps/level shifts between the last past value and the future ground truth MUST NOT be ignored.
2. Macro Trend Priority: The OVERALL macro direction (start to end) dictates the primary evaluation. A local dip or correction within a strong UPWARD macro trend is NOT a "fatal crash".
3. Downside Risk & Gap Downs: Missing a sudden crash OR a significant "gap down" (where ground truth drops heavily below the last known past value and stays low) is a fatal failure. A valid downward prediction must show a VISIBLY CLEAR downward slope or a corresponding lower level.

[Scoring Rubric]

5 (Perfect): Captures the macro trend AND all major turning points/local wiggles perfectly.

4 (Strong): Captures the macro trend AND the most prominent turning point, but misses minor wiggles or has slight timing delays.

3 (Direction Consistent): Overarching macro direction matches (both UP or both DOWN), but misses exact turning points.
Note: If ground truth crashes but prediction shows a CLEAR but gradual decline, score 3.

2 (Missed Trend / Flatline Failure): Ground truth shows a clear macro trend (goes UP, or shows a gradual/volatile DOWNWARD movement that is not a sudden crash), but prediction goes the opposite way, stays strictly FLAT, or shows only an insignificant/nearly flat slope. The model completely misses the macro movement and provides an uninformative prediction.

1 (Fatal Crash Miss): Ground truth crashes DOWN severely OR gaps down significantly below the last past value, but prediction goes UP, stays FLAT at a higher level, or shows only an insignificant

User Prompt: Cryptocurrency Prices (continued)

decline (If the prediction has a visually obvious downward slope, even if gentle, it MUST NOT be scored 1). The model completely fails to warn of severe downside risk.

B.3.5 CHLOROPHYLL CONCENTRATION

User Prompt: Chlorophyll Concentration

Evaluate the prediction of [Ocean Chlorophyll Concentration (CPHL)].

Visual Evaluation Rules:
Before scoring, perform the following visual checks:
1. Find the Blooms: Locate any massive upward spikes in the ground truth, which represent algal blooms (red tides).
2. Check Amplitude and Timing: Compare the predicted spikes' height (Y-axis) and starting point (X-axis left/right shift) against the ground truth.
3. Check for Negatives: Scan the predicted curve for any values dropping below the zero line (concentration cannot be negative).

Scoring Rubric:

5 (Perfect Ecological Warning):
Visual cues: The baseline fluctuations and the massive spikes (algal blooms) align perfectly in timing, magnitude, and dissipation.

4 (Safe Conservative / Early Warning):
Visual cues: Absolutely NO negative values and NO missed spikes. The predicted spike is shifted to the LEFT (early warning, giving farmers time to react). OR, the predicted spike is visually TALLER than the ground truth (safe overestimation, triggering maximum defense measures).

3 (Sub-optimal / Friction Cost):
Visual cues: The prediction shows a massive spike while the ground truth is flat (false alarm, causing unnecessary panic/downtime). OR, the predicted spike is slightly SHORTER than the ground truth (mild underestimation). OR, the predicted spike is shifted slightly to the RIGHT (minor lag, 1-2 steps late).

2 (High-Risk Misleading / Severe Miss):
Visual cues: The prediction captures the correct general macro-trend of the ground truth, but fails to capture the critical local fluctuations/spikes. OR, the model attempts to predict the bloom but severely underestimates its magnitude (e.g., showing only a minor bump). OR, the prediction is severely shifted to the RIGHT (warning only after the bloom has peaked). OR, the predicted curve visibly drops BELOW the zero line (violating physical laws).

1 (Catastrophic Failure):
Visual cues: Complete visual disconnect in trend and volatility. The ground truth shows devastating spikes, but the prediction completely ignores them, appearing either as a completely flat line OR a smooth, irrelevant trend line (e.g., a slow, steady drift) that fails to capture any of the sudden bloom dynamics. OR, the ground truth is perfectly stable, but the prediction hallucinates continuous, wild, chaotic oscillations.

B.3.6 OCEAN CURRENT VELOCITY

User Prompt: Ocean Current Velocity

Evaluate the prediction of [Ocean Current Velocity].

Visual Evaluation Rules:
1. Check the Zero-Crossings (Direction): Look at where the curves cross the horizontal zero line. Do they cross at the exact same time? (Crossing the zero line means the current reverses direction).
2. Check the Amplitude (Over vs. Under): Look at the major extreme points (primary positive peaks and negative valleys). Are the predicted waves visually TALLER/DEEPER (safe overestimation) or FLATTER/SHALLOWER (dangerous underestimation) than the ground truth?

User Prompt: Ocean Current Velocity (continued)

3. Check the Phase (Alignment): Are the primary waves generally aligned, or is there a severe left/right shift causing them to be out-of-phase?
4. Ignore High-Frequency Noise (Smoothing Tolerance): Ground truth often contains local, high-frequency fluctuations (jaggedness). If the prediction smooths out these minor local fluctuations but accurately captures the macro-trend and major turning points, it should NOT be penalized.

Scoring Rubric:

5 (Perfect Navigation):
Visual cues: Excellent replication of the primary periodic oscillation and macro-trend. The zero-axis crossing points and major turning points align accurately. The amplitude of MAJOR peaks and valleys is well captured. NOTE: Smoothing out minor, high-frequency local fluctuations (jaggedness) is completely acceptable and MUST still be scored a 5.

4 (Conservative Operation - Safe Overestimation):
Visual cues: The zero-axis crossings align accurately. The predicted major peaks are visually TALLER and the valleys are DEEPER than the ground truth (conservatively overestimating the current speed, which keeps ships safely in port).

3 (Sub-optimal - Minor Shift, Mild Underestimation of MAJOR peaks, or Extreme Overestimation):
Visual cues: The zero-crossings are slightly shifted to the left or right (minor phase error). OR, the predicted MAJOR peaks/valleys are slightly shallower than the ground truth (mild underestimation of the macro-trend, not just local noise). OR, the predicted waves are massively larger than the ground truth.

2 (High-Risk Misleading - Severe Underestimation, Out-of-Phase, or Trend-Only):
Visual cues: The prediction still shows SOME macro-wave structure, but severely smooths out the primary waves. OR, the curves are severely out-of-phase. OR, the prediction loses the wave-like oscillations (e.g., becomes a simple monotonic sloped line) BUT successfully captures the overall macro-directional trend (the general upward or downward slope) of the ground truth, providing partial but incomplete information.

1 (Catastrophic Failure - Loss of BOTH Oscillation and Trend):
Visual cues: The prediction completely fails to capture BOTH the periodic wave characteristics AND the macro-directional trend. It degenerates into a completely flat horizontal line, OR a narrow band of meaningless high-frequency noise/jitter hovering around a constant mean, completely ignoring the massive macro-trends (large dips, peaks, or slopes) of the ground truth.

B.3.7 SOLAR ENERGY PRODUCTION

User Prompt: Solar Energy Production

Evaluate the prediction of [Solar Energy Generation].

Visual Evaluation Rules:
1. Check the Baseline (Nighttime & Negatives): Solar generation cannot be negative, and it must be strictly zero at night. Look for any visible dips BELOW the zero line, or any wiggles during the flat nighttime gaps.
2. Check the Vertical Position (Over vs. Under): During the daytime peaks, is the prediction visually BELOW (conservative/safe) or ABOVE (dangerous overestimation) the ground truth?
(Note: Tolerate mild errors. Microscopic deviations, such as being slightly lower, a tiny bit higher, or having negligible phase shifts, should be ignored and treated as a perfect match).

Scoring Rubric:

5 (Perfect Scheduling):
Visual cues: Excellent replication of the day-night rhythm. **Crucially, ignore mild errors: if the prediction's peak is only slightly lower or a tiny bit higher than the ground truth, or has negligible phase shifts, treat it as perfect.** The nighttime baseline is strictly flat and zero.

4 (Conservative Safety - Underestimation):
Visual cues: Strict zero at night. During the day, the predicted peaks sit CLEARLY and NOTICEABLY BELOW the ground truth (conservative underestimation). Do not penalize down to 4 if it is only "slightly" lower.

User Prompt: Solar Energy Production (continued)

3 (Sub-optimal - Mild Overestimation):
Visual cues: The predicted peaks sit NOTICEABLY ABOVE the ground truth during the day (but not massively). OR, there is a clear and obvious left/right phase shift. (Again, ignore very slight overestimations or tiny phase lags).
Peak Count Rule: In a multi-day chart, severe overestimation must occur in STRICTLY LESS THAN 30% of the daytime peaks.

2 (High-Risk Misleading - Severe Overestimation):
Visual cues: The prediction sits MASSIVELY ABOVE the ground truth (e.g., predicting a huge, tall peak when the true peak is very low). OR, there are visible, non-zero wiggles during the nighttime flat periods.
Peak Count Rule: In a multi-day chart, severe overestimation occurs in 30% OR MORE of the daytime peaks.

1 (Catastrophic Physics Violation):
Visual cues: The predicted curve drops visibly BELOW the zero line (showing clear negative values). OR, the rhythm is completely broken (e.g., predicting huge peaks during the true nighttime flat periods). OR, the prediction degrades into a continuous flat line or random noise.

B.3.8 TRAFFIC SPEED

User Prompt: Traffic Speed

Evaluate the prediction of [Traffic Speed].

Visual Evaluation Rules:
1. Identify the "Valleys": In traffic speed, normal flow is a high flat line, and congestion is a sudden downward drop forming a "deep valley".
2. Check Timing (Left vs. Right Shift): Focus on the exact moment the curve starts to drop. Is the predicted drop shifted LEFT (early warning - highly preferred) or RIGHT (lagging/after-the-fact - strictly penalized)?
3. Ignore Minor Vertical Gaps: Tolerate absolute vertical distance errors during the high flat (free-flow) periods. Focus primarily on the shape, timing, and existence of the valleys.

Scoring Rubric:

5 (Perfect Scheduling):
Visual cues: The start (drop), bottom, and recovery of the congestion "valley" align perfectly on the time axis. The shape and timing are a perfect match.

4 (Early Warning / Safe Redundancy):
Visual cues: The predicted valley starts to the LEFT of the ground truth (early warning). OR, the predicted valley is deeper than the ground truth (safely overestimating congestion). Minor vertical gaps during the high flat periods are completely acceptable.

3 (Reactive / Friction Cost):
Visual cues: The predicted drop starts slightly to the RIGHT of the ground truth (minor lag). OR, the prediction shows a medium-sized valley while the ground truth is a high flat line (false alarm). OR, the prediction captures the valley but it is visibly much shallower than the true deep valley (severe underestimation of congestion).

2 (Severe Misleading / Missed Opportunity):
Visual cues: The predicted drop is shifted severely to the RIGHT (massive lag, starting to drop only when the true congestion is already ending). OR, the ground truth shows a massive deep valley, but the prediction only shows a microscopic, barely visible dip (missing a major jam).

1 (Catastrophic Failure):
Visual cues: Complete visual disconnect. The ground truth shows a severe "deep valley" (massive congestion), but the prediction remains a high, stable flat line (predicting free flow). OR vice versa (hallucinating a massive jam on a completely clear road). OR missing ANY severe deep valley.

B.3.9 SUPPLY CHAIN DEMAND

User Prompt: Supply Chain Demand

Evaluate the prediction of [Supply Chain Order Volume and Revenue].

Visual Evaluation Rules:
1. Identify the Spikes: Supply chain data typically consists of a flat/stable baseline with sudden upward spikes (promotions or peak seasons). Locate these spikes.
2. Check the Amplitude (Over vs. Under): Are the predicted spikes visually TALLER (safe overstocking) or SHORTER/SMOOTHED (dangerous stockout risk) compared to the ground truth?
3. Check the Timing (Left vs. Right Shift): Are the predicted spikes shifted LEFT (early preparation - good) or RIGHT (lagging/after-the-fact - strictly penalized)?

Scoring Rubric:

5 (Perfect Supply Chain Scheduling):
Visual cues: The baseline trend and all sudden upward spikes align perfectly in both timing and magnitude. The wave shapes match exactly.

4 (Safe Stocking / Early Response):
Visual cues: The predicted spikes are visually TALLER than the ground truth (safe overestimation to prevent stockouts). OR, the predicted spikes are shifted slightly to the LEFT (early response). The baseline remains generally accurate.

3 (Sub-optimal / Friction Cost):
Visual cues: The predicted spikes are slightly SHORTER than the ground truth (mild underestimation, less than 35% lower). OR, the spikes are massively exaggerated (extreme overestimation leading to excess inventory). OR, the spikes are shifted slightly to the RIGHT (minor lag, 1-2 steps late).

2 (High-Risk Misleading / Severe Stockout):
Visual cues: The model severely smooths out the true spikes, predicting a much flatter curve during a massive demand burst (severe underestimation, 35% or more lower, causing fatal stockouts). OR, the predicted spikes are severely shifted to the RIGHT (massive lag, predicting a surge only after the true peak has already ended, leading to overstocking at the wrong time).

1 (Catastrophic Failure / Complete Disconnect):
Visual cues: The prediction completely ignores the pulse-like demand bursts, degenerating into a nearly flat line (even with minor trends or negligible fluctuations). OR, the macro trend moves in the completely opposite direction of the ground truth.

B.3.10 CLOUD CPU UTILIZATION

User Prompt: Cloud CPU Utilization

Evaluate the prediction of [Azure Cloud Server CPU Utilization].

Visual Evaluation Rules:
1. Focus on Amplitude (Over vs. Under): In auto-scaling, overestimating CPU (predicting ABOVE ground truth) safely provisions extra servers, while underestimating (predicting BELOW) causes server crashes. Amplitude is the most critical metric.
2. Check the Spikes (Severity over Frequency): Look for sudden, sharp computational peaks. Did the prediction capture their full height, or did it smooth them out? *Business Context Rule: Severity trumps frequency. Missing or severely underestimating even a few massive spikes is a critical failure, regardless of how well the majority of normal peaks are predicted.*
3. Check the Timing (Left vs. Right Shift): Are the predicted spikes shifted LEFT (early warning, allowing time for server cold starts - good) or RIGHT (lagging/after-the-fact - strictly penalized)?
4. CRITICAL OVERRIDE (The Flatline Rule): If the prediction degenerates into a completely flat horizontal line (flatline) or a featureless trend, it MUST receive a score of 1. This applies REGARDLESS of its amplitude. Even if a flatline is positioned massively above the ground truth, it represents a complete failure to capture time-series dynamics and must be strictly penalized with a 1.

Scoring Rubric:

5 (Perfect Elastic Scaling):
Visual cues: The baseline and all sudden spikes align perfectly in both timing and height. Any visual deviation is extremely minor (less than 10% error).

User Prompt: Cloud CPU Utilization (continued)

4 (High Availability / Safe Overestimation):
Visual cues: The prediction visually "wraps" the ground truth, sitting entirely or mostly ABOVE it (safe overestimation). OR, the predicted spikes are shifted slightly to the LEFT (early response). It successfully covers all true peaks.

3 (Sub-optimal / Cost Friction):
Visual cues: The prediction captures the dynamics/spikes but falls into one of these cases:
1. It is massively TALLER than the ground truth (extreme overestimation, wasting significant money), BUT it still shows dynamic fluctuations (NOT a flatline).
2. It is slightly SHORTER than the true peaks (mild underestimation, causing brief throttling). *Note: This applies ONLY if the underestimation is consistently mild across all peaks. If any massive spike is severely missed, it drops to a 2.*
3. It is shifted slightly to the RIGHT (minor lag).

2 (High-Risk Misleading / SLA Violation):
Visual cues: The model severely smooths out the true CPU spikes, predicting a much flatter curve during a massive traffic burst. *Crucially, even if the model predicts the majority of normal peaks accurately (good average geometric fit), severely underestimating just a few massive, critical bursts (e.g., cutting a peak in half) MUST result in a 2 due to the fatal risk of server crashes.* OR, the predicted spikes are severely shifted to the RIGHT (massive lag, scaling up only after the true peak has passed).

1 (Catastrophic Failure / Flatline):
Visual cues: The prediction completely ignores the dynamic nature of the data, degenerating into a completely flat horizontal line (flatline). *Note: ANY flatline prediction gets a 1, even if it is positioned high above the ground truth.* OR, it completely misses all massive bursts, predicting a low baseline while the ground truth spikes severely.

C Human Annotation and Validation

Human validation addresses two complementary questions: whether annotators consistently apply the rubrics, and whether the resulting score orderings agree with direct preferences. We assess these through inter-annotator agreement and pairwise comparison, respectively. Both tasks involve the same five researchers with expertise in time series forecasting.

C.1 Rubric-Guided Ratings and Inter-Annotator Agreement

Five researchers with expertise in time series forecasting independently rate each forecast on a five-point scale using the rubrics in Appendix B. The modal rating serves as the final human annotation.

The five annotators show agreement across all evaluation dimensions (Table 6). Krippendorff's α averages 0.718, with values from 0.666 on Motif to 0.839 on General. These results support the use of aggregated ratings as the human reference while making the variation in agreement across dimensions explicit.

Table 6: Agreement among the five annotators on rubric-guided ratings in Meta-TimeVista.

Metric	**Micro-Level**				**Macro-Level**
	Trend	Motif	Phase	Magnitude	General
Krippendorff's α	0.678	0.666	0.677	0.728	0.839
Kendall's τ	0.889	0.877	0.899	0.906	0.907
Spearman's ρ	0.907	0.895	0.917	0.920	0.916

C.2 Agreement with Intuitive Preferences

Pairwise evaluation tests whether rubric-guided scores preserve the preferences expressed in direct forecast comparisons. For each dimension, we sample forecast pairs sharing the same historical window and ground truth. The same five researchers independently choose the better forecast or

a tie without access to the rubrics or existing scores. We compare their modal preference with the ordering induced by the rubric-guided annotations.

Rubric-guided scores match intuitive preferences in 88.29% of the 1,495 valid comparisons (Table 7). The disagreement breakdown clarifies where the five-point scale loses resolution: most mismatches occur when equal rubric scores conceal a pairwise preference, whereas opposite rankings are rare (Table 8). Thus, the scores capture much of the preference ordering in these comparisons, while direct comparisons distinguish some forecasts within the same score level.

Table 7: Agreement between intuitive pairwise preferences and rubric-guided scores.

Dimension	N	**Agreement (%)**	**Macro-F1 (%)**	κ_w	**95% CI**
Trend	298	88.93	88.70	0.874	[84.91, 92.64]
Motif	300	90.33	90.18	0.889	[85.92, 94.31]
Phase	300	94.67	94.70	0.940	[91.39, 97.42]
Magnitude	299	85.62	85.15	0.822	[81.50, 89.54]
General	298	81.88	80.88	0.790	[76.64, 86.50]

N denotes valid comparisons with complete ratings and a unique modal preference; κ_w denotes linearly weighted Cohen's kappa. Confidence intervals for agreement are estimated from 10,000 bootstrap resamples of source series within each dimension.

Table 8: Pairwise annotator agreement and disagreement with rubric-induced orderings.

Evaluation setting	α	Unanimous (%)	Rubric tie	Human tie	Reversal	Total
Synthetic: Trend	0.856	82.00	21	11	1	33
Synthetic: Motif	0.868	83.67	20	8	1	29
Synthetic: Phase	0.918	89.33	5	11	0	16
Synthetic: Magnitude	0.857	82.33	32	5	6	43
Real-world: General	0.770	71.81	43	7	4	54

Nominal α uses comparisons with at least two ratings. Unanimity uses complete five-rater comparisons (300 per synthetic dimension and 298 for General); General has 299 comparisons with at least two ratings. Disagreement counts use complete comparisons with a unique mode. "Rubric tie" and "Human tie" denote a tie under only the rubric scores or only the modal pairwise judgment, respectively; "Reversal" denotes opposite non-tied orderings.

D Experimental Implementation

This section specifies the forecasting checkpoints, judge inputs, numerical metrics, and deterministic rule evaluator used in the comparisons. The implementation separates forecast generation from forecast evaluation and makes the information available to each evaluator explicit.

D.1 Forecasting Models

The forecasting comparison includes four statistical baselines, eight TSFMs, and the multimodal model introduced in Section 5.1. Pretrained models are evaluated without fine-tuning. For quantile-based models, the 0.5 quantile (median) provides the point forecast. Table 9 lists the foundation-model checkpoints and context limits; the model labeled Chronos uses the listed Chronos-Bolt checkpoint.

Table 9: Configurations of the forecasting models.

Model	Checkpoint	Context Length	Input Mode	Output Type
Timer-S1	`bytedance-research/Timer-S1`	11520	Unimodal	Quantile
TimesFM 2.5	`google/timesfm-2.5-200m-pytorch`	4096	Unimodal	Quantile
Chronos-2	`amazon/chronos-2`	8192	Unimodal	Quantile
Moirai 2.0	`Salesforce/moirai-2.0-R-small`	4000	Unimodal	Quantile
TiRex	`NX-AI/TiRex`	2048	Unimodal	Quantile
Sundial	`thuml/sundial-base-128m`	2880	Unimodal	Distribution
Chronos	`amazon/chronos-bolt-base`	2048	Unimodal	Quantile
FlowState	`ibm-research/flowstate`	4096	Unimodal	Quantile
Aurora	`DecisionIntelligence/Aurora`	2048	Multimodal	Distribution

D.2 VLM Judging and Input Representation

All six VLMs follow a common evaluation protocol: each judge receives a forecast plot together with the evaluation dimension, scoring rubric, and task information. We set the decoding temperature to zero and standardize plot presentation with contrasting colors, labeled axes, grids, and legends to support consistent comparisons across judges.

The modality comparison holds the judge and rubrics fixed and changes how the same historical window, ground truth, and forecast are represented: a plot or numerical text. Human alignment and mean input-token usage are measured across forecast-length groups on the synthetic data (Figure 4a). This comparison assesses the accuracy–token trade-off associated with input representation under a shared scoring task.

D.3 Numerical Metrics and Human Alignment

We compare numerical metrics and VLM judgments by their agreement with modal human ratings, measured using Spearman's ρ and Kendall's τ within each evaluation dimension. All numerical metrics compare the forecast $\hat{\mathbf{y}} \in \mathbb{R}^T$ with its ground truth $\mathbf{y} \in \mathbb{R}^T$. We negate errors, distances, and absolute temporal offsets so that higher values indicate better forecasts. Tables 1 and 2 use the same Gemini outputs and human annotations.

Pointwise and sequence-alignment metrics. MSE and MAE measure pointwise deviations:

$$\mathrm{MSE} = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{y}_t)^2, \qquad \mathrm{MAE} = \frac{1}{T}\sum_{t=1}^{T}|y_t - \hat{y}_t|. \tag{5}$$

For sequence alignment, DTW uses the normalized distance returned by the `symmetric2` step pattern, while Soft-DTW uses a smoothing parameter of $\gamma = 1$.

Dimension-specific metrics. For Trend, we compute MAE between components extracted separately from the forecast and ground truth using either HP filtering ($\lambda = 1600$) or STL decomposition. The STL period is 7 for $T \geq 14$ and $\max(2, \lfloor T/2 \rfloor)$ otherwise. For Motif, derivative DTW compares first differences using the same normalized DTW distance. Each difference sequence is standardized independently when its variance is nonzero.

For Phase, cross-correlation lag measures the absolute lag that maximizes full cross-correlation between standardized sequences. Standardization adds 10^{-8} to the standard deviation for numerical stability, and comparisons involving a constant sequence are excluded because the lag is undefined. We also evaluate DTW warping offset, the mean absolute index difference along the alignment path. For Magnitude, Wasserstein distance compares the empirical value distributions, capturing differences in value rather than temporal order.

Reporting and interpretation. Where multiple dimension-specific metrics are evaluated, Table 2 reports the one with the highest unrounded Spearman correlation. The alternative STL trend error yields $\rho = 0.425$ and $\tau = 0.328$, close to HP filtering. DTW warping offset yields $\rho = 0.681$ and $\tau = 0.536$, nearly matching cross-correlation lag in Spearman correlation but falling below its Kendall correlation ($\tau = 0.582$).

D.4 RUBRIC-BASED RULES

Rubric-based Rules tests whether a deterministic implementation of the supplied rubrics reproduces human alignment. It maps raw historical values, ground truth, and forecasts to five-point scores, using the same 480 synthetic and 545 real-world forecast instances and human targets as the VLM comparison.

Features and score mapping. For Micro-Level evaluation, Trend uses smoothed trajectories and directional agreement; Motif uses detrended shape and derivative-based distances; Phase uses peak/valley counts and matched event timing; Magnitude uses amplitude ratios, vertical offsets, and local degradation. Magnitude takes the lowest component score, following its rubric. For General, domain-specific rules combine these features with priorities such as non-negative solar production, non-decreasing cumulative deaths, and congestion-event detection. Ordered conditions map the features to score levels, with severe violations taking precedence.

Fixed configuration and implementation. Numerical score mappings operationalize the supplied criteria, with fixed heuristic thresholds for qualitative requirements. The peak-prominence fraction is 0.20 and the tolerance multiplier is 1 throughout; no human ratings or VLM scores are used to fit or select thresholds.

For the constraint-pass analysis in Appendix E.2, non-negativity and monotonicity are checked exactly. The visible-pass variant instead allows deviations within 1% of the reference span, matching the visibility tolerances used in the corresponding score rules.

E SUPPLEMENTARY ANALYSES AND EVALUATION CASES

The supplementary analyses examine whether the alignment results persist after numerical scale correction and explicit rule encoding, then assess sensitivity to rubrics, rendering, and repeated calls. Each comparison addresses a distinct evaluation choice. The final cases connect returned scores and explanations to visible forecasting errors and domain priorities.

E.1 HUMAN ALIGNMENT AFTER SCALE NORMALIZATION

Historical-scale normalization tests whether the VLM advantage in Table 2 can be explained by pooling forecasts with different numerical scales. We divide each forecast and its ground truth by the standard deviation of their shared historical window before recomputing the numerical metrics:

$$s_i = \mathrm{std}(\mathbf{x}_i), \qquad \widetilde{\mathbf{y}}_i = \mathbf{y}_i / s_i, \qquad \widetilde{\hat{\mathbf{y}}}_{i,m} = \hat{\mathbf{y}}_{i,m} / s_i. \tag{6}$$

The scale is estimated from historical values alone and shared by all forecasting models for a source window. Forecasts and ground truth are rescaled jointly, preserving amplitude and offset discrepancies. Every historical scale is nonzero, so all 480 synthetic and 545 real-world forecasts are retained. Human annotations and Gemini outputs remain fixed.

Table 10: Human alignment after historical-scale normalization of numerical metrics. Gemini scores and human annotations are unchanged from Table 2.

Dimension	VLM judge		MSE		MAE		DTW		Structure-aware metrics		
	ρ	τ	ρ	τ	ρ	τ	ρ	τ	Metric	ρ	τ
Trend	**0.859**	**0.771**	0.787	0.633	0.787	0.633	0.815	0.661	HP trend error	0.777	0.622
Motif	**0.870**	**0.796**	0.563	0.430	0.570	0.435	0.592	0.455	Derivative DTW	0.820	0.670
Phase	**0.831**	**0.764**	0.551	0.418	0.558	0.423	0.560	0.426	Cross-correlation lag	0.681	0.582
Magnitude	**0.907**	**0.849**	0.672	0.526	0.678	0.531	0.714	0.559	Wasserstein distance	0.651	0.507
General	**0.865**	**0.820**	0.508	0.400	0.497	0.388	0.515	0.401	Soft-DTW	0.306	0.234

ρ and τ denote Spearman's and Kendall's rank correlations. Numerical metrics are recomputed on the jointly rescaled sequences, with errors negated before correlation. Soft-DTW retains $\gamma = 1$ in normalized units.

Gemini retains higher correlations across all five dimensions after scale normalization, even as the numerical baselines improve substantially (Table 10). On General, MSE rises from $\rho = -0.112$ to 0.508 and DTW from -0.077 to 0.515. Thus, scale differences explain part of the original gap, while the remaining advantage persists beyond this correction; the gap narrows considerably on Trend. This comparison does not isolate domain knowledge as the source of the remaining advantage. Soft-DTW also changes its relative smoothing strength when sequences are rescaled with $\gamma = 1$, so its improvement combines scale and smoothing effects.

Alternative historical scales. We also evaluate non-seasonal MASE and RMSSE, which normalize forecast errors by historical one-step changes:

$$\mathrm{MASE} = \frac{\mathrm{MAE}}{\frac{1}{L-1}\sum_{t=2}^{L}|x_t - x_{t-1}|}, \qquad \mathrm{RMSSE} = \sqrt{\frac{\mathrm{MSE}}{\frac{1}{L-1}\sum_{t=2}^{L}(x_t - x_{t-1})^2}}. \tag{7}$$

Gemini also retains higher alignment across all dimensions under these alternative normalizations (Table 11). The numerical results vary with the historical scale, establishing its importance to baseline comparisons. Both denominators are positive for every instance, and no scale parameters are selected using human ratings.

Table 11: Human alignment of the VLM judge and numerical metrics normalized by historical one-step changes. The VLM judge is Gemini-3.1-Pro.

Dimension	VLM judge		MASE		RMSSE	
	ρ	τ	ρ	τ	ρ	τ
Trend	**0.859**	**0.771**	0.679	0.529	0.646	0.502
Motif	**0.870**	**0.796**	0.563	0.428	0.535	0.405
Phase	**0.831**	**0.764**	0.548	0.417	0.548	0.416
Magnitude	**0.907**	**0.849**	0.636	0.497	0.609	0.474
General	**0.865**	**0.820**	0.308	0.235	0.260	0.201

E.2 Supplementary Rule Comparisons

Gemini exceeds Rubric-based Rules on General by 0.394 in Spearman correlation (95% CI: $[0.292, 0.503]$), indicating an advantage even when rubric criteria are explicitly encoded (Table 2). Confidence intervals use 2,000 paired bootstrap resamples over 138 source-window groups, keeping forecasts with identical history–truth pairs together and recomputing ranks in each resample.

Among 98 forecasts from 42 source windows satisfying the selected non-negativity or monotonicity constraints, Gemini also achieves higher human alignment ($\rho = 0.923$ vs. 0.696; 95% CI for the difference: $[0.123, 0.376]$). This suggests that its advantage extends beyond detecting these hard-constraint violations.

E.3 Rubric Perturbation Protocol

The rubric ablations separate changes to wording, presentation, and scoring detail (Table 3). On synthetic Magnitude, paraphrasing preserves the criteria while changing their wording; reformatting converts the bullet points to continuous prose while preserving criteria, visual cues, and thresholds; simplification retains the five score levels while removing detailed cues and thresholds.

Alignment is more sensitive to presentation and scoring detail than to the tested wording change. Paraphrasing changes Spearman correlation by only 0.006, whereas simplification reduces it by 0.149. Reformatting reduces correlation from 0.907 to 0.809 despite preserving the scoring content. Together, these comparisons support structured, explicit criteria as useful components of rubric-guided evaluation in this setting.

On real-world General, deliberately incorrect rules reduce correlation with the original human annotations from $\rho = 0.865$ to -0.508. The negative correlation indicates an opposing tendency in

forecast ordering: changing domain requirements can redirect judgments, not merely weaken agreement. In combination with the Magnitude ablations, this result highlights rubric content as a consequential part of the evaluator. It characterizes sensitivity to supplied criteria without identifying the judge's internal reasoning process.

E.4 Plot Rendering and Prediction Length

Rendering protocol. Individual rendering perturbations test whether human alignment depends on a particular plot configuration (Table 4). The default aspect ratio is 12:5. Vertical padding extends each end of the value range by the stated fraction; sparse markers appear every three time steps. Removing axis annotations removes ticks, axis labels, and the title. The legend-swap condition exchanges ground-truth and prediction labels while retaining curve colors and styles.

Prediction-length subgroups. High alignment extends across the short, medium, and long forecast-horizon groups in Table 12. Spearman correlations are 0.888, 0.916, and 0.917, respectively, showing no observed decline for longer forecasts in this Magnitude evaluation. These are comparisons between different forecast groups, rather than a controlled test of changing the horizon of the same forecast.

Table 12: Human alignment for synthetic Magnitude by prediction length.

Subset	Prediction length	ρ	τ
Short	$T \leq 192$	0.888	0.829
Medium	$192 < T \leq 336$	0.916	0.857
Long	$T > 336$	0.917	0.862

Randomized colors and line styles. Score-level agreement complements the correlation results: randomizing colors and line styles preserves 1,985 of 2,465 scores (80.53%; Figure 7), and 390 of the 480 changed scores differ by only one point. The comparison covers all four Micro-Level dimensions for 480 synthetic forecasts and General for 545 real-world forecasts. Most judgments therefore remain unchanged or move to an adjacent score level across the two renderings.

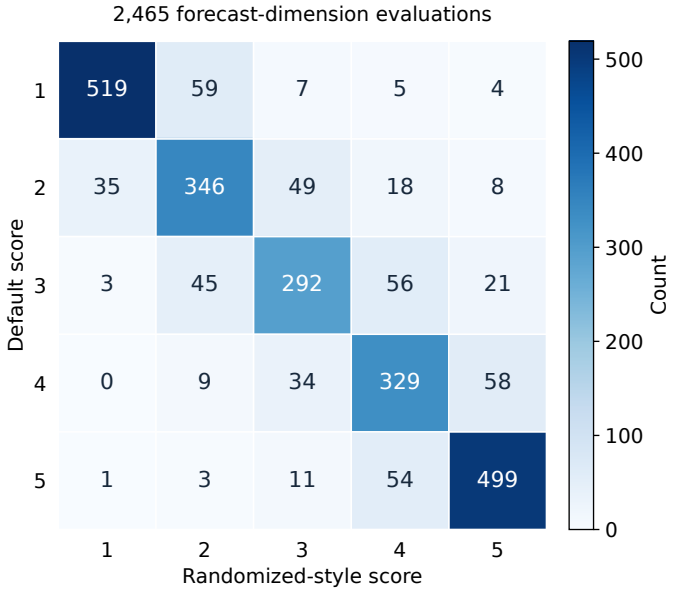


Figure 7: Score counts under default plot styles (rows) and randomized colors and line styles (columns). Diagonal entries indicate unchanged scores.

E.5 Consistency Across Repeated Runs

Across five repeated calls with fixed inputs, 72.54% of evaluations receive identical scores and 94.28% span at most one point (Table 13). For each of the 2,465 forecast–dimension evaluations, we compute the score range $\Delta s = \max_r s_r - \min_r s_r$; the mean within-evaluation sample standard deviation is 0.1652. The results show that most judgments remain within one or two adjacent score levels across all five calls, with larger fluctuations confined to a smaller subset.

Table 13: Score consistency across five independent runs.

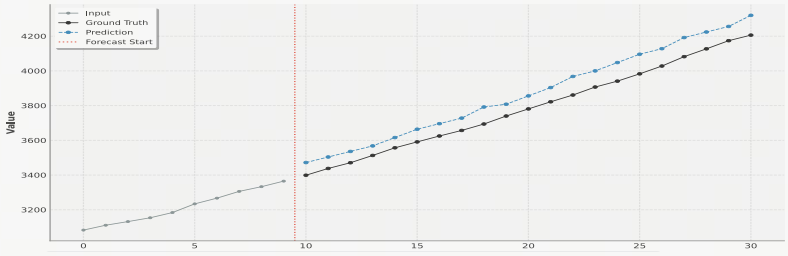
Score range Δs	Count	Percentage
0	1,788	72.54%
1	536	21.74%
2	112	4.54%
3	23	0.93%
4	6	0.24%
Total	2,465	100.00%

E.6 Qualitative Evaluation Cases

The cases in Figures 8–11 illustrate how textual justifications connect visible forecast errors to domain-specific scoring priorities. They span cumulative deaths, cryptocurrency, traffic speed, and solar production, with the assigned scores and decisive errors summarized in each caption. The traffic case complements Figure 4c by showing a less severe underestimation of congestion. These examples make individual judgments inspectable and illustrate the stated rationale.

Input

Rubrics (Covid Death):
Evaluate cumulative covid death predictions based on monotonicity, error direction, and error magnitude. Grade forecasts from 5 to 1 by prioritizing strict non-decreasing trends and safe overestimation as a resource buffer.
Time Series Plot:



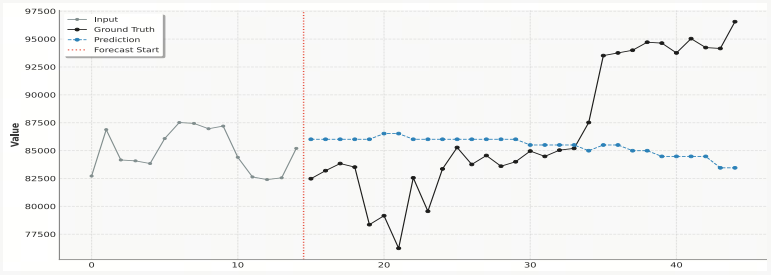
Output

Score: 4
Rationale: Based on the evaluation criteria, here is the assessment of the prediction: 1. Monotonicity: The predicted curve is strictly non-decreasing throughout the forecast horizon, correctly adhering to the physical constraint of cumulative data. There are no visible downward drops or local zigzags. 2. Error Direction: The forecast consistently overestimates the ground truth. In the context of medical resource planning, this serves as a safe buffer rather than a risky underestimation. 3. Error Magnitude: The vertical gap between the prediction and the ground truth remains relatively stable and bounded across the time steps. The predicted curve is structurally parallel to the actual data, representing a reasonable and safe overestimation without diverging into extreme or unrealistic values. Because the model maintains strict monotonicity and provides a safe, bounded overestimation that closely follows the structural trend of the ground truth, it aligns perfectly with the criteria for a "Good" fit.

Figure 8: Cumulative-death forecast: the judge assigns score 4 for a non-decreasing trajectory with bounded overestimation under the supplied rubric.

Input

Rubrics (Crypto):
Prioritize trend direction and early warning signs over absolute values. Grade predictions from 5 to 1 by rewarding precise alignment and proactive downturn alerts, tolerating false alarms or minor lag, and severely penalizing opposite trends, delayed reactions, flatline predictions in trending markets, and catastrophic failures to predict sharp market crashes.
Time Series Plot:



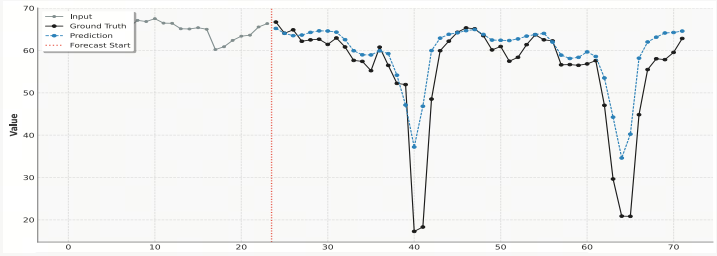
Output

Score: 2
Rationale: Based on the visual evaluation rules and the scoring rubric, here is the evaluation of the prediction: 1. Macro Trend Analysis: The ground truth initially experiences a dip but then exhibits a very strong and clear upward macro trend, rising from around 83,000 to over 95,000 by the end of the forecast horizon. 2. Forecast Analysis: The forecast stays almost entirely flat, with a very slight downward drift towards the end. It completely fails to anticipate the significant upward movement. 3. Rubric Application: The ground truth shows a clear upward macro trend, but the prediction stays nearly flat and uninformative, missing the macro movement entirely. This aligns perfectly with the criteria for a score of 2 (Missed Trend / Flatline Failure). [[Score: 2]]

Figure 9: Cryptocurrency forecast: the judge assigns score 2 for a nearly flat prediction that misses the rising trend.

Input

Rubrics (Traffic Speed):
Prioritize the timing of congestion onset and dissipation over absolute speed accuracy. Grade forecasts from 5 to 1 by valuing phase alignment and early warnings as safe redundancies.
Time Series Plot:



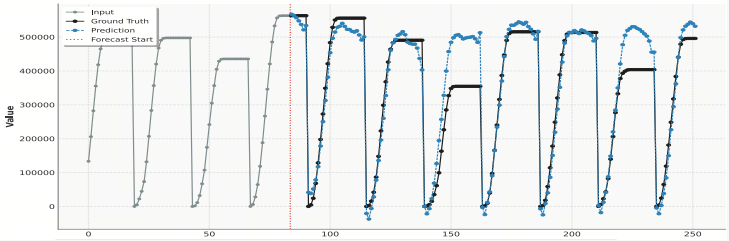
Output

Score: 3
Rationale: Based on the visual evaluation rules and the scoring rubric, here is the evaluation of the forecast: Justification: The ground truth displays two distinct, deep "valleys" indicating severe traffic congestion. The forecast successfully captures the existence and general timing of both congestion events. However, there are two main issues: 1. Severe Underestimation: Both predicted valleys are visibly much shallower than the true deep valleys. The ground truth drops to around17-20, while the forecast only drops to about 35, severely underestimating the intensity of the congestion. 2. Minor Lag: The start of the drop for the second valley in the forecast is shifted slightly to the RIGHT compared to the ground truth, indicating a minor lag (reactive rather than proactive). According to the rubric, a prediction that captures the valley but is visibly much shallower than the true deep valley, or shows a minor lag, falls under the "Reactive / Friction Cost" category. [[Score: 3]]

Figure 10: Traffic-speed forecast: the judge assigns score 3 for capturing congestion events while underestimating their severity and showing a timing lag.

Input

Rubrics (Solar Energy Production):
Evaluate solar power predictions based on diurnal rhythm alignment and grid safety. Grade forecasts from 5 to 1 by prioritizing perfect day-night cycles with safe, conservative underestimation, tolerating minor overestimation or phase shifts, and strictly penalizing severe overestimation, non-zero nighttime fluctuations, negative values, or complete rhythm collapse that threatens grid stability.
Time Series Plot:



Output

Score: 1
Rationale: Based on the visual evaluation rules and the provided chart, here is the assessment: 1. Baseline Check (Nighttime & Negatives): The most critical observation is that the forecast visibly drops below the zero line during the nighttime troughs. Solar energy generation cannot be negative, making this a clear violation of the physical constraints of the data. 2. Vertical Position: While the daytime peaks roughly follow the ground truth, the presence of negative values completely invalidates the prediction's reliability. According to the scoring rubric, a prediction that drops visibly below the zero line (showing clear negative values) constitutes a "Catastrophic Physics Violation.

Figure 11: Solar-production forecast: the judge assigns score 1 for negative predicted production despite partially matching the daytime pattern.