Multimodal Collaborative Debate for Zero-Shot Time Series Reasoning

Anonymous ACL submission

Abstract

Large language models (LLMs) are increasingly used as natural-language interfaces to structured data, yet they remain brittle when reasoning over time series. Visual patterns can be misleading, numerical claims can be hallucinated, and textual context can override evidence from the signal. We study *zero-shot* time-series reasoning as a multimodal evidence arbitration problem for LLM agents. We propose **TS-Debate**, an inference-time multi-agent protocol that requires no task-specific fine-tuning. TS-Debate first elicits relevant domain knowledge, then assigns modality-specialized agents to textual context, visual patterns, and numerical signals, and coordinates their interaction through a verification-conflict-calibration procedure. Reviewer agents check decision-critical claims with lightweight code execution and numerical lookup, resolve cross-modal disagreement, and calibrate the final answer. Unlike generic multi-agent debate or unconstrained tool use, TS-Debate specifies how evidence is exposed, which claims are checkable, and how verification outcomes shape synthesis. Across 20 tasks from three public benchmarks, TS-Debate improves classification and question answering performance over strong baselines, while revealing that debate is most useful for global-structure and cross-view reasoning rather than local value reconstruction.

1 Introduction

Time-series data underlie critical decisions in finance, climatology, and industrial monitoring (Trirat et al., 2024). Beyond forecasting, practitioners increasingly ask *reasoning-centric* questions such as: *What explains this anomaly?*, *Which historical pattern best matches this trajectory?*, or *Will the underlying cause of this trend persist?* Answering such questions requires integrating numerical precision, temporal structure, and domain knowledge—tasks that remain challenging for current large language models (LLMs). As LLMs become a universal interface for data analysis, enabling robust *time-series reasoning* (TSR) in natural language is both practically important and scientifically unresolved (Merrill et al., 2024; Chang et al., 2025).

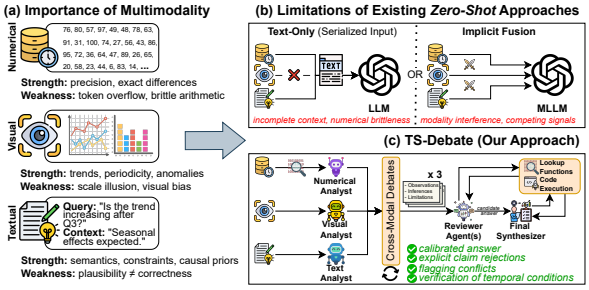

Figure 1: Overview of multimodal time-series reasoning: (a) complementary strengths and failure modes across modalities; (b) limitations of existing zero-shot approaches; and (c) TS-Debate, a cross-modal agentic framework with explicit verification.

Recent work suggests that TSR is inherently *multimodal* (Kong et al., 2025b; Liu et al., 2025a). Visual prompting compresses long numerical sequences into charts that multimodal LLMs can parse efficiently (Yoon et al., 2024), and vision–language approaches show that time- and frequency-domain visualizations expose temporal structure that text-only prompting misses (Liu et al., 2025b). However, simply providing multimodal inputs is insufficient. Without structured orchestration, additional modalities produce interference rather than improvement. Models over-trust visual scale impressions, hallucinate numeric details, or default to shallow linguistic heuristics when modalities disagree. Rather than asking whether multimodality helps TSR, we ask: ***what inference-time protocol over modality-specialized agents is needed to convert multimodal access into reliable zero-shot reasoning?***

Existing zero-shot approaches fall short not due to insufficient modality access or tooling, but because they lack a reasoning protocol tailored to

the unique challenges of time-series data. As illustrated in Figure 1, we identify three such challenges. **(i) Modality interference:** prior studies (Chang et al., 2025; Trirat and Lee, 2025) either serialize all modalities into text—leading to context overflow and brittle arithmetic—or fuse them implicitly within a single model, leaving numeric, visual, and textual signals indistinguishable and impossible to selectively trust. **(ii) Numeric brittleness:** even when models identify the correct qualitative pattern, final answers frequently contain arithmetic errors or unjustified extrapolations, and classical debate frameworks (Liang et al., 2024; Smit et al., 2024) resolve disagreement by persuasion rather than verification—inadequate when claims about trends, extrema, and temporal conditions are objectively checkable. **(iii) Temporal-scope ambiguity:** verifiable past-present claims are routinely conflated with unverifiable future claims, so calibration cannot distinguish what the signal supports from what the model speculates.

To address these challenges, we introduce ***TS-Debate***, a cross-modal, collaborative agentic debate framework for zero-shot time-series reasoning. TS-Debate instantiates modality-specialized agents—a text analyst, a visual analyst, and a numerical analyst—that produce explicitly labeled observations, inferences, and limitations grounded in their modality. A preliminary *knowledge elicitation* stage activates relevant domain priors before any data analysis. Crucially, TS-Debate replaces majority voting with a *Verification-Conflict-Calibration* (VCC) protocol: reviewer agents verify temporally checkable claims via numerical lookup and code execution, type cross-modal conflicts, and calibrate answers based on both data consistency and temporal scope. A final synthesizer discounts—rather than implicitly fuses—unverified or contradicted claims. Our contribution is not the use of LLM agents, multi-agent debate, multimodal prompting, or code execution individually—each is established. Rather, we identify three challenges specific to zero-shot TSR and propose an inference-time protocol that addresses all three jointly. Our **main contributions** are as follows.

- We formalize TSR as a *multimodal evidence arbitration* problem requiring modality-isolated evidence generation rather than implicit fusion.
- We propose TS-Debate, an inference-time LLM-agent protocol that separates modality-specific evidence generation from claim verification and calibrated synthesis.
- We introduce the VCC protocol, which replaces persuasion- or vote-based debate resolution with programmatic verification, explicit conflict typing, and temporal-scope-aware calibration.
- We demonstrate consistent empirical improvements over competitive baselines across 20 tasks on three public benchmarks: MT-Bench (+7.39%), TimerBed (+22.74%), and TSQA (+21.58%).

2 Related Work

Time-Series Analysis and Reasoning LLMs have been applied to TSR via numerical serialization (Gruver et al., 2023), alignment-based adapters (Jin et al., 2024; Chow et al., 2024), and coupling with specialized encoders (Yu et al., 2025; Wang et al., 2025). These methods focus on representation learning and typically operate within a single numerical channel, requiring task-specific training. In contrast, TS-Debate assumes off-the-shelf LLM and MLLM backbones and targets an *inference-time orchestration protocol*, coordinating modality-specialized agents under explicit verification rather than training new encoders.

Multimodal LLMs for Time Series A parallel line converts time series into visual representations for MLLM reasoning, including waveforms, spectrograms (Liu et al., 2025b), and sensor plots (Yoon et al., 2024; Kong et al., 2025b), achieving gains with reduced token budgets. However, these approaches rely on a *single fusion step* within an MLLM. TS-Debate instead instantiates modality-isolated agents for numerical, visual, and textual evidence and coordinates them through a multi-stage protocol, allowing cross-modal evidence to interact iteratively rather than implicitly fusing them.

Multi-Agent Debate Multi-agent debate (MAD) frameworks (Smit et al., 2024; Li et al., 2024; Kaesberg et al., 2025; Zhang et al., 2025) use LLM agents that critique one another to improve reasoning, but are predominantly text-only and assume homogeneous agents with shared inputs. TS-Debate differs in both assumptions and architecture: agents are *modality-isolated* and *collaborative* rather than adversarial, and judges are equipped with code execution to verify numerical and temporal claims. Debate in TS-Debate is thus not an evaluation heuristic but an inference-time protocol tailored to multimodal TSR. Table 1 summarizes these differences.

Framework	Key Functionality					Support Modality for Reasoning		
	Zero-Shot Reasoning	Multi-Agent Debate	Program-Aided Verification	Self-Elicited Knowledge	Calibration Protocol	Raw Number	Textual Context	Visual Representation
TS-Reasoner (Yu et al., 2025)	×	×	×	×	×	✓	✓	△ (for training)
ChatTime (Wang et al., 2025)	×	×	×	×	×	✓	✓	×
Chat-TS (Quinlan et al., 2025)	×	×	×	×	×	✓	✓	×
ChatTS (Xie et al., 2025)	×	×	×	×	×	✓	✓	×
TimeMQA (Kong et al., 2025a)	×	×	×	×	×	✓	✓	×
ITFormer (wang et al., 2025)	×	×	×	×	×	△ (for training)	✓	✓
TimeART (Wu et al., 2026)	×	×	✓	×	×	✓	✓	×
VL-Time (Liu et al., 2025b)	✓	×	×	×	×	×	✓	✓
ByMyEyes (Yoon et al., 2024)	✓	×	×	×	×	×	✓	✓
TS-Debate (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between TS-Debate and existing time-series reasoning frameworks.

3 Proposed Framework: *TS-Debate*

3.1 Problem Setting

We study *zero-shot* TSR over univariate (d=1) or multivariate (d>1) sequences. Each instance consists of a natural-language task specification q, an observed sequence $\mathbf{x}_{1:T} \in \mathbb{R}^{T\times d}$, and auxiliary context c (e.g., textual descriptions or metadata). The goal is to produce an answer $\hat{a} \in \mathcal{A}(q)$, where the answer space is task-induced (e.g., a class label, a numerical vector, or a multiple-choice option). TS-Debate performs no task-specific training and does not update parameters at inference time; instead, it orchestrates LLM agents, each operating on a different representational view of the same signal, to construct an evidence-grounded response.

To account for the temporal nature of time series, we distinguish *past-present* tasks, whose target lies within the observed horizon, from *future* tasks, whose target lies beyond it. This distinction determines verification semantics: past-present claims can be checked directly against $\mathbf{x}_{1:T}$, whereas future claims must be evaluated through domain-consistent reasoning and uncertainty calibration.

3.2 Framework Overview

Let $\mathcal{M} = \{\mathsf{T}, \mathsf{V}, \mathsf{N}\}$ denote text, visual, and numerical modalities. As illustrated in Figure 2 (and Algorithm 1), TS-Debate proceeds in three stages. First, it elicits task- and domain-specific priors from q (optionally with c) to form shared guidance for all agents. Second, it constructs modality-specific interfaces—textual context for semantics, charts for global structure, and precise numerical evidence—and runs an iterative collaborative debate among modality-specialized analysts. Third, it applies a verification-conflict-calibration protocol: reviewer agents verify analysts' claims, explicitly identify and resolve conflicts when possible, and produce calibrated candidate answers, after which a final synthesizer evaluates reviewer reasoning quality and produces the final answer $\hat{a}$.

3.3 Multimodal Query Construction

A central design principle of TS-Debate is to expose *complementary and verifiable* views of $\mathbf{x}_{1:T}$ to different agents, reducing reliance on any single representation. This is implemented via task-conditioned knowledge elicitation, multiple representational views of the signal, and explicit visual and numerical query interfaces.

Knowledge Elicitation Before interpreting any data, TS-Debate elicits domain priors $k = \mathsf{Elicit}(q, c)$, which serve as a *shared analysis contract* for all agents. Concretely, k encodes a characterization of the task/domain, domain constraints and expected patterns, salient signal attributes, a recommended analysis procedure, common failure modes, and guidance on which modalities are most informative. This anchors zero-shot reasoning to task semantics and domain methodology before any modality-specific analysis.

Time-Frequency Views From $\mathbf{x}_{1:T}$, we derive complementary representational views below.

- **Temporal View** (v_{time}): a time-domain representation emphasizing trends, change points, and extrema.
- **Spectral View** (v_{freq}): a frequency-domain representation highlighting periodicities and oscillatory structure (e.g., seasonality and recurrences).
- **Feature/Statistical View** (v_{feat}): summary statistics and detected events (e.g., peaks, troughs, and anomalies) for precise quantitative checks.

Multimodal Query Interfaces TS-Debate exposes these views through two explicit interfaces.

- **Visual Interface.** A time-domain chart $\mathsf{Chart}_{\text{time}}$ and a frequency-domain chart $\mathsf{Chart}_{\text{freq}}$ form the input substrate for the visual analyst, surfacing global structure (e.g., trend, volatility regimes, and periodicity) that is difficult to infer from numerical probes.
- **Numerical Interface.** Controlled tool access to the time series and derived attributes, denoted $\mathsf{Lookup}(\cdot)$, supports querying values over inter-

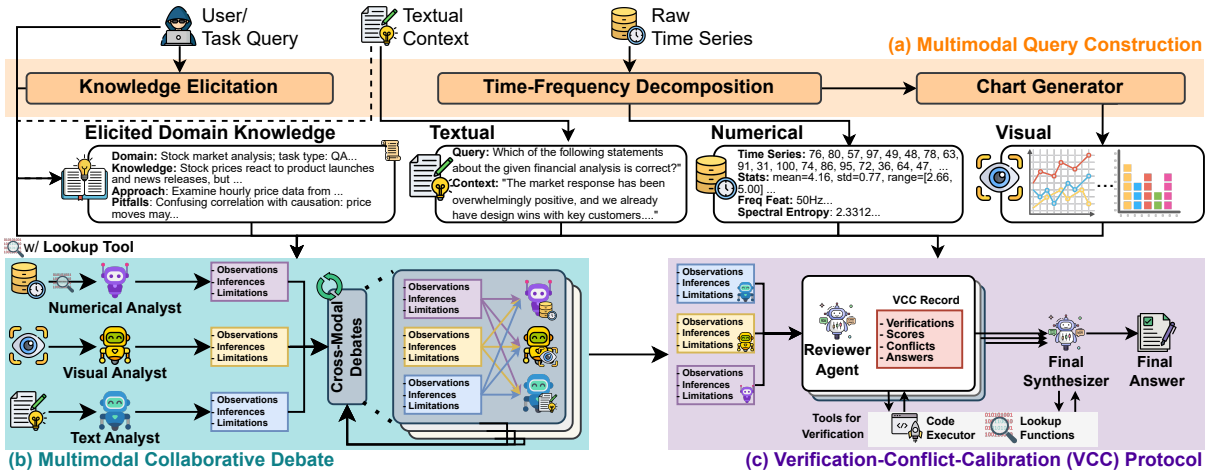

Figure 2: Overview of the TS-Debate framework. **(a) Multimodal query construction** derives textual, numerical, and visual representations from user queries, raw time series, and contextual information. **(b) Multimodal collaborative debate** enables numerical, visual, and text analysts to reason over these representations and produce structured observations, inferences, and limitations. **(c) VCC protocol** applies tool-assisted review to assess debate outputs, identify conflicts, and synthesize the final answer.

vals, probing neighborhoods, retrieving detected events/features, and obtaining frequency-domain information. This interface is essential for *verifiability*, grounding claims in concrete values.

3.4 Multimodal Collaborative Debate

TS-Debate uses *collaborative debate* to elicit diverse evidence without adversarial contest, refining hypotheses across modalities while explicitly acknowledging uncertainty and limitations.

Modality-Specialized Analysts For each modality $m \in \mathcal{M}$, TS-Debate instantiates an analyst $\mathcal{A}_m$.

- **Text analyst** $\mathcal{A}_{\mathsf{T}}(q, c, k)$ interprets contextual constraints and textual descriptions, without direct numerical verification.
- **Visual analyst** $\mathcal{A}_{\mathsf{V}}(q, \mathsf{Chart}_{\mathsf{time}}, \mathsf{Chart}_{\mathsf{freq}}, k)$ extracts qualitative, global structure from the generated charts.
- **Numerical analyst** $\mathcal{A}_{\mathsf{N}}(q, \mathsf{Lookup}(\mathbf{x}_{1:T}), k)$ produces precise, checkable statements about values and derived features via numerical queries.

All analysts share the same k, so disagreements reflect evidence and representational differences rather than inconsistent task interpretations.

Structured Evidence Generation Each analyst produces an evidence report $e_m^{(r)}$ at round r in a template that separates *verifiable observations* from *interpretive inferences*, comprising

- **Understanding**: a restatement of the task q in the analyst's own words.
- **Observations**: modality-grounded statements that are, in principle, checkable against the provided inputs (text, charts, or numbers), explicitly labeled as observations.
- **Inferences**: conclusions derived from observations using k, explicitly labeled as inferences.
- **Limits**: what cannot be determined reliably from the analyst's modality and inputs.

Iterative Evidence Refinement Debate proceeds for $R \geq 1$ rounds. In round 1, analysts work *independently* to maximize epistemic diversity. Thus,

$$e_m^{(1)} = \mathcal{A}_m(q, I_m, k), \quad m \in \mathcal{M}, \tag{1}$$

where $I_{\mathsf{T}} = c$, $I_{\mathsf{V}} = [\mathsf{Chart}_{\mathsf{time}}, \mathsf{Chart}_{\mathsf{freq}}]$, and $I_{\mathsf{N}} = \mathsf{Lookup}(\mathbf{x}_{1:T})$. For rounds $r \geq 2$, each analyst conditions on the others' previous evidence and may refine, correct, or strengthen its position:

$$e_m^{(r)} = \mathcal{A}_m(q, I_m, k, \{e_{m'}^{(r-1)}\}_{m' \in \mathcal{M}}). \tag{2}$$

Refinement is *not* forced convergence. Analysts incorporate complementary observations while retaining principled disagreement when warranted by modality-specific evidence (e.g., a visually plausible pattern can be numerically refuted).

3.5 Verification-Conflict-Calibration (VCC)

Evidence generation alone is insufficient: multimodal debate surfaces diverse hypotheses, but without disciplined validation and conflict handling, diversity can amplify error. VCC enforces explicit verification, conflict modeling, and calibrated synthesis, with the design intent that additional verified evidence monotonically expands error detection and correction, while unresolved conflicts trigger conservative calibration rather than confident error.

Claim Verification & Domain Consistency Let Ω be the set of atomic claims extracted from the final-round evidence $E^{(R)} = \{e_m^{(R)}\}_{m \in \mathcal{M}}$, where each $\omega \in \Omega$ is a minimal evaluable statement (e.g., "the series increases sharply near $t = t_0$" or "a dominant period is present").

For each claim ω, a reviewer $\mathcal{A}_j$ assigns a verification status $V_j(\omega) \in \{V_{\text{VERIFIED}}, V_{\text{UNVERIFIED}}, V_{\text{CONTRADICTED}}\}$, determined by checking ω against available sources: **numerical** evidence via $\text{Lookup}(\cdot)$ and derived features; **computational** evidence via a lightweight code executor (e.g., arithmetic checks and aggregations); **visual** evidence via chart-level consistency checks; and **textual** evidence (q, c) for task semantics and contextual constraints. Here, V_{VERIFIED} indicates direct support, $V_{\text{CONTRADICTED}}$ direct refutation, and $V_{\text{UNVERIFIED}}$ insufficient evidence for either.

The reviewer also assigns a domain-consistency label $D_j(\omega) \in \{D_{\text{MATCHES}}, D_{\text{VIOLATES}}, D_{\text{N/A}}\}$, evaluating whether ω is consistent with constraints encoded in k (with $D_{\text{N/A}}$ when domain constraints do not apply). The combined annotation $[V_j(\omega), D_j(\omega)]$ distinguishes unsupported claims from domain-inconsistent claims—critical when direct verification is unavailable (e.g., future targets) but domain constraints still apply.

Beyond claim-level labels, each $\mathcal{A}_j$ scores each analyst's evidence along three axes: $s_{j,m} = (s_{j,m}^{\text{inf}}, s_{j,m}^{\text{obs}}, s_{j,m}^{\text{hon}})$, with $s_{j,m}^{\text{inf}} \in [0, 50]$, $s_{j,m}^{\text{obs}} \in [0, 30]$, and $s_{j,m}^{\text{hon}} \in [0, 20]$. These correspond to **inference quality** (logical soundness and calibration), **observation quality** (specificity, verifiability, and relevance), and **honesty** (explicit acknowledgment of limits and avoidance of unwarranted certainty). Reviewers are instructed to identify limitations even in strong responses, discouraging score saturation; scores guide synthesis toward evidence that is both grounded and appropriately uncertain.

Explicit Conflict Modeling After verification, reviewers indicate outstanding conflicts via $\Delta_j \in \{\Delta_{\text{NO_CONFLICT}}, \Delta_{\text{DETECTED}}, \Delta_{\text{RESOLVED}}\}$, where Δ_{DETECTED} flags incompatible claims with no verified resolution, Δ_{RESOLVED} marks conflicts resolved via verification and domain constraints, and $\Delta_{\text{NO_CONFLICT}}$ indicates mutually consistent (or immaterial) evidence.

Why explicit conflicts matter. Averaging or voting over conflicting rationales is brittle. Modalities can disagree for principled reasons (e.g., a trend appears visually but disappears under numerical scrutiny), and a "majority" can be wrong if it shares a methodological flaw. VCC therefore handles conflicts explicitly so that disagreements trigger additional verification or conservative calibration rather than being smoothed away.

At the synthesizer level, disagreement is summarized in two ways. First, the **answer agreement pattern** $\Gamma \in \{\Gamma_{\text{UNANIMOUS}}, \Gamma_{\text{SPLIT}}, \Gamma_{\text{ALL_DIFFERENT}}\}$ captures whether reviewers converge. Second, the **resolution type** $\Lambda \in \{\Lambda_{\text{VERIFIED_RESOLUTION}}, \Lambda_{\text{UNRESOLVED}}, \Lambda_{\text{NO_CONFLICT}}, \Lambda_{\text{APPROACH_ERROR}}\}$ indicates whether disagreements were settled by verified evidence, remain unresolved, are absent, or stem from an incorrect approach relative to k, separating genuine ambiguity from methodological error.

Calibrated Answer Synthesis Each reviewer j produces a calibrated answer $\hat{a}_j$ and an accompanying VCC record

$$u_j = (\hat{a}_j,\ \{(V_j(\omega), D_j(\omega))\}_{\omega \in \Omega},\ \Delta_j,\ \{s_{j,m}\}_{m \in \mathcal{M}}). \tag{3}$$

Calibration follows explicit rules:

- If the key claims supporting $\hat{a}_j$ are V_{VERIFIED} and D_{MATCHES}, and $\Delta_j = \Delta_{\text{NO_CONFLICT}}$, the answer may be stated confidently.
- If key claims are $V_{\text{UNVERIFIED}}$ or $\Delta_j = \Delta_{\text{DETECTED}}$, the answer must be conservative and explicitly reflect uncertainty.
- If a claim is D_{VIOLATES}, it is rejected even if plausible from a single modality.

Temporal scope further modulates calibration. For *past-present* tasks, direct verification against $\mathbf{x}_{1:T}$ dominates. For *future-oriented* tasks, TS-Debate cannot verify future outcomes from past data; instead, it evaluates whether the causes of observed patterns plausibly persist under k. This discourages naive extrapolation ("the trend continues because it has continued") and promotes conditional predictions grounded in domain plausibility.

Finally, the synthesizer produces $\hat{a}$ and is explicitly *not* a vote counter. It evaluates reviewers' *reasoning quality* from their VCC records u_j via: an **approach check** for adherence to the analysis procedure implied by k and temporal scope; **reviewer scoring** on a fixed rubric (each criterion 0–20) covering task understanding, evidence usage, verification discipline, conflict handling, and calibration; and **decision synthesis** that prioritizes answers supported by verified evidence and correct method-

ology. If all reviewers exhibit approach errors ($\Lambda_{\text{APPROACH_ERROR}}$), the synthesizer re-verifies the most decision-critical claims and derives an answer directly. All prompts are provided in §F.

4 Experiments

To verify the effectiveness of TS-Debate, we conduct extensive experiments on a diverse set of TSR tasks spanning classification, regression, and question answering (QA). Additionally, we perform ablation and sensitivity analyses.

4.1 Setup

Benchmarks and Tasks As presented in Table 4, we evaluate on three public benchmarks totaling *20 tasks* that jointly cover context-grounded multimodal reasoning, sensor classification, and multi-task time-series QA. **MTBench** (Chen et al., 2025) is a multimodal benchmark coupling time series with rich textual context. It comprises 9 tasks across finance and weather, including trend classification, forecasting and indicator prediction, correlation classification, and multiple-choice QA. Inputs use either a short-window or long-window setting (e.g., different granularity regimes), which stresses temporal reasoning under varying horizons. **TimerBed** (Liu et al., 2025b) focuses on classification from diverse sensor domains; we evaluate on 6 datasets spanning 2-7 classes and sequence lengths up to 4K, emphasizing pattern recognition under domain shift. **TSQA** (Kong et al., 2025a) is a multi-task benchmark spanning anomaly detection, classification, forecasting, imputation, and open-ended QA, providing a broad testbed for time-series reasoning.

Evaluation Metrics For classification and MCQA tasks, we report accuracy (and weighted F1 when applicable). For regression tasks, we report MAE with MSE or MAPE as a secondary metric. For open-ended QA, we report an automatic NLI-based correctness score (entailment accuracy) computed with a `DeBERTa-v3-large` evaluator, used as a scalable proxy for human judgments. We report mean±std over 3 runs on cost-aware sampled test subsets (see §C.1).

Comparison Baselines We compare against single-agent prompting, multi-agent debate, and time-series-specific reasoning baselines. Single-agent baselines include **Zero-Shot** and Chain-of-Thought (**CoT**) (Wei et al., 2022), each evaluated in both standard and multimodal settings (**+MM**). For debate-based methods, we include **MAD** (Liang et al., 2024) and its multimodal variant. For time-series reasoning baselines, we include **VL-Time** (Liu et al., 2025b) and **ByMyEyes** (Yoon et al., 2024), which are designed specifically for time series. To ensure that gains are not simply artifacts of tool access, we additionally evaluate tool-augmented variants (**+Tools**), equipping the zero-shot, CoT, and MAD multimodal baselines with the same numerical lookup and code-execution interfaces available to TS-Debate.

Implementation Details We use `gpt-4.1-mini` as the backbone model via API, chosen for its strong zero-shot performance at favorable cost and latency (see §D.5). We use temperature $= 0$ with fixed seeds. TS-Debate uses a fixed configuration with 2 analyst rounds, 3 parallel reviewers, and a final synthesizer. For multimodal inputs, we provide two figures: a multi-panel time-domain plot and a multi-panel frequency-domain plot. All methods share the same backbone and evaluation protocol for fair comparison.

4.2 Main Results

Table 2 summarizes performance on 20 tasks (mean±std over 3 runs), reported by benchmark and task type. Tasks are grouped into **classification**, **regression**, and **QA**: MTBench includes trend and correlation classification, forecasting and indicator prediction for regression, and MCQA; TimerBed averages six classification tasks; and TSQA combines classification and anomaly detection, forecasting and imputation (regression), and open-ended QA. We focus on the training-free regime, evaluating whether structured inference-time reasoning can extract more from off-the-shelf LLMs than current zero-shot approaches. Across all three task types, TS-Debate delivers consistent gains over the strongest baseline, 25.65% on classification, 36.92% on regression error reduction, and 15.06% on QA on average, without any task-specific training. Detailed experimental results are presented in §D.

Classification TS-Debate achieves the highest average accuracy across all benchmarks, with a **25.65%** improvement over the strongest baseline on average and up to **80.41%** over standard zero-shot prompting. Improvements are consistent on MTBench (Finance/Weather), TimerBed,

Benchmarks	Zero-Shot	+ MM	CoT	+ MM	MAD	+ MM	ByMyEyes	VL-Time	TS-Debate
Classification (Accuracy) - Higher is Better									
MTBench (Finance)	0.272 (±0.019)	0.375 (±0.032)	0.275 (±0.015)	0.370 (±0.017)	0.300 (±0.023)	0.347 (±0.027)	0.445 (±0.018)	0.300 (±0.025)	**0.543 (±0.043)**
MTBench (Weather)	0.250 (±0.008)	0.523 (±0.021)	0.220 (±0.036)	0.477 (±0.039)	0.200 (±0.014)	0.320 (±0.029)	0.190 (±0.008)	0.220 (±0.029)	**0.557 (±0.025)**
TimerBed	0.286 (±0.017)	0.317 (±0.018)	0.343 (±0.020)	0.332 (±0.023)	0.294 (±0.014)	0.317 (±0.024)	0.290 (±0.032)	0.332 (±0.024)	**0.421 (±0.051)**
TSQA	0.378 (±0.045)	0.413 (±0.019)	0.545 (±0.034)	0.523 (±0.041)	0.565 (±0.041)	0.542 (±0.042)	0.383 (±0.066)	0.498 (±0.049)	**0.615 (±0.023)**
Average	0.296 (±0.022)	0.407 (±0.022)	0.346 (±0.026)	0.425 (±0.030)	0.340 (±0.023)	0.381 (±0.031)	0.327 (±0.031)	0.338 (±0.032)	**0.534 (±0.035)**
Regression (MAE) - Lower is Better									
MTBench (Finance)	1.010 (±0.225)	0.973 (±0.150)	1.002 (±0.129)	2.796 (±1.675)	1.079 (±0.092)	4.051 (±1.547)	36.295 (±5.221)	1.574 (±0.657)	**0.814 (±0.060)**
MTBench (Weather)	3.920 (±0.084)	3.531 (±0.186)	4.091 (±0.245)	**3.523 (±0.114)**	3.958 (±0.135)	3.965 (±0.204)	4.714 (±0.252)	4.695 (±0.180)	4.162 (±0.244)
TSQA	96.391 (±87.878)	47.524 (±26.683)	69.466 (±49.069)	51.368 (±34.393)	81.097 (±40.683)	85.889 (±46.709)	453.377 (±121.495)	61.294 (±32.509)	**27.845 (±9.433)**
Average	33.774 (±29.396)	17.343 (±9.006)	24.853 (±16.481)	19.229 (±12.061)	28.711 (±13.637)	31.302 (±16.153)	164.795 (±42.323)	22.521 (±11.115)	**10.940 (±3.246)**
Question Answering (Accuracy) - Higher is Better									
MTBench (Finance)	0.823 (±0.017)	0.847 (±0.012)	0.797 (±0.026)	0.817 (±0.025)	0.750 (±0.008)	0.747 (±0.021)	0.823 (±0.046)	0.773 (±0.026)	**0.913 (±0.012)**
MTBench (Weather)	0.617 (±0.048)	0.653 (±0.026)	0.600 (±0.037)	0.617 (±0.021)	0.520 (±0.036)	0.500 (±0.022)	0.593 (±0.031)	0.613 (±0.039)	**0.717 (±0.052)**
TSQA	0.727 (±0.045)	0.673 (±0.058)	0.710 (±0.045)	0.713 (±0.046)	0.760 (±0.062)	0.693 (±0.048)	0.580 (±0.051)	0.677 (±0.039)	**0.870 (±0.059)**
Average	0.722 (±0.037)	0.724 (±0.032)	0.702 (±0.036)	0.716 (±0.031)	0.677 (±0.035)	0.647 (±0.030)	0.666 (±0.043)	0.688 (±0.034)	**0.833 (±0.041)**

Table 2: Overall performance comparison on the three benchmarks. Best results are highlighted in **bold**.

and TSQA, indicating effective temporal reasoning across diverse classification settings.

Regression TS-Debate obtains the lowest average MAE on MTBench (Finance) and TSQA, reducing error by **36.92%** on average relative to the best baseline and avoiding catastrophic failures observed in naive multimodal variants. However, TS-Debate does not dominate all regression regimes. On MTBench (Weather), CoT+MM performs slightly better (3.523 vs. 4.162 MAE), suggesting that some point-wise prediction settings benefit more from direct multimodal prompting than multi-agent deliberation.

Question Answering TS-Debate achieves the highest average QA accuracy across benchmarks, with **15.06%–28.75%** improvements over prior methods. These gains suggest more effective integration of temporal evidence from multiple aspects of time series for context-grounded QA.

These results show that TS-Debate is strongest when the target depends on global structure, cross-view consistency, or relative quantities, rather than purely local reconstruction.

4.3 Ablation and Sensitivity Analysis

We conduct controlled ablations on 20 instances per task (400 samples total) to isolate the effects of modality composition, system components, and debate configurations.

Multimodal Debate Agents Figure 3(a) shows that incorporating multiple modalities consistently improves performance over single-modality or no-debate (i.e., simple agents with tools) baselines. While numerical (N), visual (V), and textual (T) analysts each contribute complementary gains, their combination (NVT) yields the best results across all tasks. Notably, removing debate leads to substantial degradation, confirming that performance gains stem from structured interaction rather than modality access alone.

Component Variations Figure 3(b) demonstrates improvements as key components are added. Analyst-side lookup reduces error and improves accuracy, reviewer-side lookup and code execution further refine predictions, and knowledge elicitation (KE) provides a substantial marginal gain, yielding the strongest overall performance. These results show that both external tool use and structured reasoning are essential to effective debate.

Debate Configurations Figure 3(c–e) analyze sensitivity to the number of reviewers, debate rounds, and review rounds. Performance improves with additional reviewers and rounds up to a moderate regime (typically 2–3), after which gains sat-

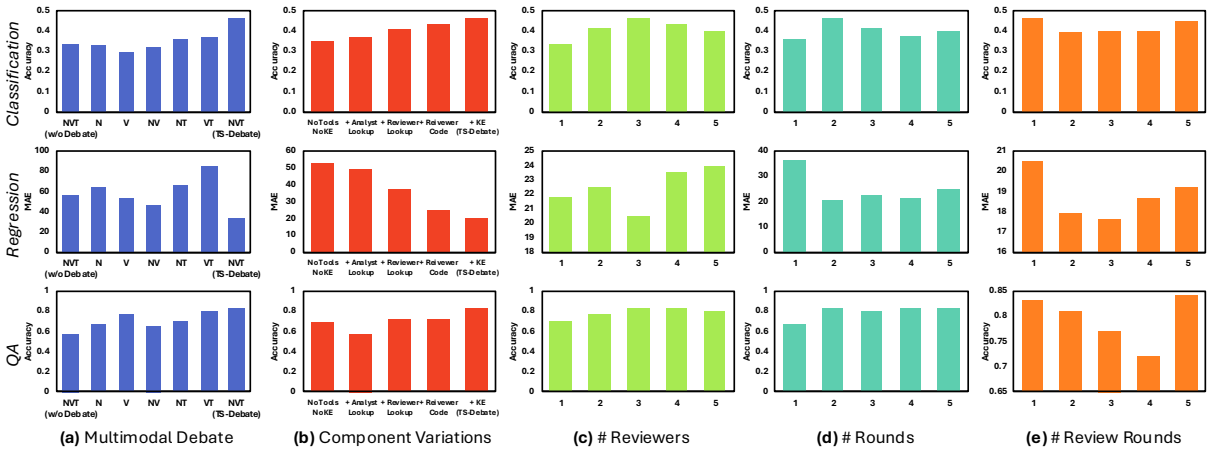

Figure 3: Impact of (a) modality composition, (b) system components, and (c–e) debate configurations across tasks.

Method	Regression ↓	Classification ↑	QA ↑	# Tokens	Time (s)
ZS+MM+Tools	23.32	0.29	0.58	32955	31.9
CoT+MM+Tools	25.98	0.33	0.63	30899	28.7
MAD+MM+Tools	24.77	0.39	0.67	82488	63.9
TS-Debate	**20.46**	**0.47**	**0.83**	77002	77.3

Table 3: Result comparison on the tool-matched setup.

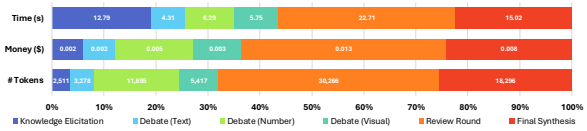

Figure 4: Average time and monetary cost breakdown.

urate or slightly regress, especially for regression tasks. This indicates diminishing returns and supports our default configuration as a favorable trade-off between performance and efficiency.

As VCC is not a separable module but is embedded in the reviewer/synthesizer protocol and output structure (e.g., verification tags, conflict status, and calibration rules), directly removing it would require redesigning prompts and decision logic, resulting in a qualitatively different method and confounding attribution. Instead, *we examine VCC mechanisms via tool and modality ablations*. Concretely, no tools in Figure 3(b) disables programmatic verification, and single-modality settings in Figure 3(a) limit cross-modal conflict resolution, thereby probing the key VCC mechanisms.

Tool-Matched Setups To test whether TS-Debate's gains are explained by tool access rather than by the protocol itself, Table 3 reports results on the ablation set where ZS+MM, CoT+MM, and MAD+MM baselines all receive the same lookup and code-execution interface as TS-Debate. Tool access alone does not recover TS-Debate: even with identical tools and comparable or larger token budgets (MAD+MM+Tools uses 82.5K vs. 77.0K tokens), prior methods trail TS-Debate by $+0.08$ to $+0.18$ on classification and $+0.16$ to $+0.25$ on QA. This indicates that the gains arise from modality-specialized evidence generation and the VCC protocol, not from unstructured tool calling.

4.4 Resource Cost

Figure 4 reports resource costs in terms of wall-clock time, token usage, and monetary API cost, averaged across tasks for a single run using closed-source LLMs. Under the default configuration (`gpt-4.1-mini`), TS-Debate requires about 70 seconds and incurs a cost of $0.033 per sample. The overhead from time-frequency preprocessing and chart generation is negligible and can be handled offline. Agents equipped with tools tend to consume more tokens, likely due to iterative tool invocations; however, preliminary trials suggest that stronger models can reduce overall token usage by making more accurate and efficient tool calls, thereby converging with fewer errors and interactions.

5 Conclusions

We presented TS-Debate, an inference-time protocol for *zero-shot* TSR. Modality-specialized text, visual, and numerical analysts produce structured evidence, and a VCC protocol verifies temporally checkable claims via code and lookup, types cross-modal disagreement, and calibrates answers by temporal scope. Across 20 tasks on three benchmarks, TS-Debate outperforms strong single-agent, multi-agent, and tool-matched baselines, with the largest gains where reasoning depends on global structure, cross-view consistency, or numerical faithfulness rather than local reconstruction. The bottleneck for zero-shot TSR is not modality access but the *organization* of evidence, verification, and calibration at inference time.

Limitations

TS-Debate has several limitations that suggest directions for future work. Performance can degrade for long-horizon causal reasoning or highly noisy multivariate time series where visual trends and numerical signals diverge. Domain knowledge elicitation is currently sample-specific, and numeric verification is constrained by the expressiveness of available tools. Promising directions include transferable and adaptive knowledge elicitation across related tasks, debate topologies that respond to disagreement patterns, richer verification (e.g., robust statistical tests), and tighter integration with structured retrieval for contextual grounding.

Multi-agent orchestration also increases token usage and wall-clock time relative to single-pass prompting. TS-Debate provides cost controls via modality selection, the number of reviewers, and the number of debate rounds, but aggressive pruning reduces verification coverage; adaptive topologies that allocate reviewers only when conflicts are detected are another promising direction. Accordingly, TS-Debate is not intended as a universal replacement for lightweight classification heuristics, but rather for high-stakes, offline, or error-sensitive deployments where fine-tuning is infeasible due to data scarcity and where transparent, explicitly verified reasoning justifies the computational overhead.

References

Ching Chang, Yidan Shi, Defu Cao, Wei Yang, Jeehyun Hwang, Haixin Wang, Jiacheng Pang, Wei Wang, Yan Liu, Wen-Chih Peng, and Tien-Fu Chen. 2025. A survey of reasoning and agentic systems in time series with large language models. *Preprint*, arXiv:2509.11575.

Jialin Chen, Aosong Feng, Ziyu Zhao, Juan Garza, Gaukhar Nurbek, Cheng Qin, Ali Maatouk, Leandros Tassiulas, Yifeng Gao, and Rex Ying. 2025. MTBench: A multimodal time series benchmark for temporal reasoning and question answering. *arXiv preprint arXiv:2503.16858*.

Winnie Chow, Lauren E. Gardiner, Haraldur T Hallgrimsson, Maxwell A Xu, and Shirley You Ren. 2024. Towards time-series reasoning with LLMs. In *NeurIPS Workshop on Time Series in the Age of Large Models*.

Nate Gruver, Marc Anton Finzi, Shikai Qiu, and Andrew Gordon Wilson. 2023. Large language models are zero-shot time series forecasters. In *NeurIPS*.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y. Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, and Qingsong Wen. 2024. Time-LLM: Time series forecasting by reprogramming large language models. In *ICLR*.

Lars Benedikt Kaesberg, Jonas Becker, Jan Philip Wahle, Terry Ruas, and Bela Gipp. 2025. Voting or consensus? decision-making in multi-agent debate. In *Findings of ACL*, pages 11640–11671.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. 2025a. Time-MQA: Time series multi-task question answering with context enhancement. In *ACL*, pages 29736–29753.

Yaxuan Kong, Yiyuan Yang, Shiyu Wang, Chenghao Liu, Yuxuan Liang, Ming Jin, Stefan Zohren, Dan Pei, Yan Liu, and Qingsong Wen. 2025b. Position: Empowering time series reasoning with multimodal llms. *Preprint*, arXiv:2502.01477.

Yunxuan Li, Yibing Du, Jiageng Zhang, Le Hou, Peter Grabowski, Yeqing Li, and Eugene Ie. 2024. Improving multi-agent debate with sparse communication topology. In *Findings of EMNLP*, pages 7281–7294.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *EMNLP*, pages 17889–17904.

Haoxin Liu, Harshavardhan Kamarthi, Zhiyuan Zhao, Shangqing Xu, Shiyu Wang, Qingsong Wen, Tom Hartvigsen, Fei Wang, and B. Aditya Prakash. 2025a. How can time series analysis benefit from multiple modalities? a survey and outlook. *Preprint*, arXiv:2503.11835.

Haoxin Liu, Chenghao Liu, and B. Aditya Prakash. 2025b. A picture is worth a thousand numbers: Enabling LLMs reason about time series via visualization. In *NAACL*, pages 7486–7518.

Mike A Merrill, Mingtian Tan, Vinayak Gupta, Thomas Hartvigsen, and Tim Althoff. 2024. Language models still struggle to zero-shot reason about time series. In *Findings of EMNLP*, pages 3512–3533.

Paul Quinlan, Qingguo Li, and Xiaodan Zhu. 2025. Chat-ts: Enhancing multi-modal reasoning over time-series and natural language data. *arXiv preprint arXiv:2503.10883*.

Andries P. Smit, Nathan Grinsztajn, Paul Duckworth, Thomas D. Barrett, and Arnu Pretorius. 2024. Should we be going mad? a look at multi-agent debate strategies for llms. In *ICML*.

Patara Trirat and Jae-Gil Lee. 2025. MONAQ: Multi-objective neural architecture querying for time-series analysis on resource-constrained devices. In *Findings of EMNLP*, pages 16922–16950.

Patara Trirat, Yooju Shin, Junhyeok Kang, Youngeun Nam, Jihye Na, Minyoung Bae, Joeun Kim, Byunghyun Kim, and Jae-Gil Lee. 2024. Universal time-series representation learning: A survey. *arXiv preprint arXiv:2401.03717*.

Chengsen Wang, Qi Qi, Jingyu Wang, Haifeng Sun, Zirui Zhuang, Jinming Wu, Lei Zhang, and Jianxin Liao. 2025. Chattime: A unified multimodal time series foundation model bridging numerical and textual data. In *AAAI*, volume 39, pages 12694–12702.

Yilin wang, Peixuan Lei, Jie Song, Yuzhe Hao, Tao Chen, Yuxuan Zhang, LEI JIA, Yuanxiang Li, and zhongyu wei. 2025. ITFormer: Bridging time series and natural language for multi-modal QA with large-scale multitask dataset. In *Forty-second International Conference on Machine Learning*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Xingjian Wu, Junkai Lu, Zhengyu Li, Xiangfei Qiu, Jilin Hu, Chenjuan Guo, Christian S. Jensen, and Bin Yang. 2026. TimeART: Towards agentic time series reasoning via tool-augmentation. *Preprint*, arXiv:2601.13653.

Zhe Xie, Zeyan Li, Xiao He, Longlong Xu, Xidao Wen, Tieying Zhang, Jianjun Chen, Rui Shi, and Dan Pei. 2025. Chatts: Aligning time series with llms via synthetic data for enhanced understanding and reasoning. *Proc. VLDB Endow.*, 18(8):2385–2398.

Hyungjun Yoon, Biniyam Aschalew Tolera, Taesik Gong, Kimin Lee, and Sung-Ju Lee. 2024. By my eyes: Grounding multimodal large language models with sensor data via visual prompting. In *EMNLP*, pages 2219–2241.

Fangxu Yu, Hongyu Zhao, and Tianyi Zhou. 2025. Ts-reasoner: Aligning time series foundation models with llm reasoning. *arXiv preprint arXiv:2510.03519*.

Hangfan Zhang, Zhiyao Cui, Jianhao Chen, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025. Stop overvaluing multi-agent debate–we must rethink evaluation and embrace model heterogeneity. *arXiv preprint arXiv:2502.08788*.

A Extended Related Work

A.1 Time-Series Analysis and Reasoning

Recent work explores using LLMs for time-series analysis and reasoning. A direct approach serializes numerical sequences as text, enabling zero-shot forecasting but suffering from sensitivity to tokenization and brittle numerical fidelity (Gruver et al., 2023). Alignment-based methods (Jin et al., 2024; Chow et al., 2024) address this issue by training adapters or auxiliary encoders that map time series into LLM-compatible representations, improving performance at the cost of modality-specific fine-tuning and additional training data. Several systems further couple LLMs with specialized time-series models. TS-Reasoner (Yu et al., 2025) integrates foundation time-series encoders with LLMs to support multi-step workflows over forecasting, anomaly detection, and causal discovery. ChatTime (Wang et al., 2025) treats time series as a "foreign language," reducing tokenization overhead while supporting bi-modal input and output. Despite their effectiveness, these approaches primarily focus on representation learning or encoder design and typically operate within a single numerical information channel. In contrast, TS-Debate assumes off-the-shelf LLM and MLLM backbones and targets an *inference-time orchestration protocol*. Rather than training new encoders, TS-Debate coordinates multiple modality-specialized agents (numerical, visual, and textual) under explicit verification and calibration rules. This design avoids task-specific training and instead emphasizes reliability and numerical faithfulness through structured, branch-style reasoning at inference time.

A.2 Multimodal LLMs for Time Series

A parallel line of work leverages multimodal LLMs (MLLMs) by converting time series into richer visual or cross-modal representations. VL-Time (Liu et al., 2025b) proposes a visualization-first pipeline that transforms time series into images (e.g., waveforms and spectrograms) and reasons over them with MLLMs, achieving substantial gains with reduced token budgets. ByMyEyes (Yoon et al., 2024) similarly shows that visual prompting over sensor-derived plots can outperform text-only prompts while further lowering token costs. While these studies demonstrate the effectiveness of multimodal representations, they typically rely on a *single fusion step* within an MLLM (Kong et al., 2025b). TS-Debate departs

from this paradigm in two key ways. First, instead of learning new multimodal encoders, we keep the underlying models fixed and instantiate modality-isolated agents for numerical series, visual patterns, and textual context. Second, we coordinate these agents through a multi-stage collaborative protocol with explicit verification, conflict resolution, and calibration, allowing cross-modal evidence to interact iteratively rather than being fused in a single forward pass.

A.3 Multi-Agent Debate

Multi-agent debate (MAD) frameworks (Smit et al., 2024) employ multiple LLM agents that iteratively critique one another, often moderated by a judge, to improve reasoning quality. Early work demonstrated gains on mathematical and logical benchmarks, while later studies examined debate topology and aggregation rules and cautioned against over-interpreting debate-based evaluations (Li et al., 2024; Kaesberg et al., 2025; Zhang et al., 2025). These approaches are predominantly text-only and typically assume homogeneous agents with shared access to the same information. TS-Debate differs from generic MAD in both *assumptions* and *architecture*. Rather than adversarial persuasion or majority voting, TS-Debate adopts a *collaborative* debate paradigm in which agents contribute complementary evidence. We enforce *modality isolation* by assigning each agent a distinct view of the problem, ensuring that disagreements reflect genuinely different evidence rather than paraphrased reasoning. Moreover, our judges are equipped with explicit verification tools, including code execution, to check numerical claims and domain constraints. As a result, debate in TS-Debate is not an evaluation heuristic but an inference-time protocol tailored to multimodal TSR, explicitly targeting numerical hallucinations and calibration errors. Table 1 summarizes the key differences between TS-Debate and existing frameworks.

B Pseudocode of TS-Debate

We present the pseudocode of the proposed TS-Debate framework in Algorithm 1 below. The source code is available at `https://anonymous.4open.science/r/TS-Debate`.

Algorithm 1 Overall Procedure of TS-Debate

Require: Task query q, time series $\mathbf{x}_{1:T}$, textual context c, number of debate rounds R, number of reviewer agents J

Ensure: Final calibrated answer $\hat{a}$

▷ Multimodal Query Construction (§3.3)

$k \leftarrow \mathsf{Elicit}(q, c)$

$\mathsf{Chart}_{\text{time}} \leftarrow \mathsf{ChartGen}_{\text{time}}(\mathbf{x}_{1:T})$; $\mathsf{Chart}_{\text{freq}} \leftarrow \mathsf{ChartGen}_{\text{freq}}(\mathbf{x}_{1:T})$

$I_{\mathsf{V}} \leftarrow [\mathsf{Chart}_{\text{time}}, \mathsf{Chart}_{\text{freq}}]$; $I_{\mathsf{N}} \leftarrow \mathsf{Lookup}(\mathbf{x}_{1:T})$; $I_{\mathsf{T}} \leftarrow c$

▷ Multimodal Collaborative Debate (§3.4)

$E^{(0)} \leftarrow \varnothing$

for $r = 1$ to R **do**

  for all $m \in \mathcal{M}$ **do**

    $e_m^{(r)} \leftarrow \mathcal{A}_m\big(q, I_m, k, E^{(r-1)}\big)$

  end for

  $E^{(r)} \leftarrow \{e_m^{(r)}\}_{m \in \mathcal{M}}$

end for

▷ Verification-Conflict-Calibration (VCC) Protocol (§3.5)

$E^{(R)} \leftarrow \{e_m^{(R)}\}_{m \in \mathcal{M}}$

$I \leftarrow [I_{\mathsf{T}}, I_{\mathsf{V}}, I_{\mathsf{N}}]$

for $j = 1$ to J **do**

  $u_j \leftarrow \mathsf{VCC}_j(q, I, k, E^{(R)})$

end for

$\hat{a} \leftarrow \mathsf{VCC}_{\text{final}}(q, I, k, \{u_j\}_{j=1}^{J})$

return $\hat{a}$

C Implementation Details

C.1 Dataset Descriptions

We evaluate our approach on three public benchmarks comprising a total of 20 tasks. These benchmarks are selected to stress complementary aspects of time-series reasoning, including multimodal grounding with rich textual context, sensor-based classification under domain shift, and broad multi-task question answering (QA) over time-series data.

- **MTBench** (Chen et al., 2025). MTBench is a multimodal time-series benchmark that pairs numeric sequences with rich textual context drawn from financial news and weather reports. The dataset is publicly available at `https://github.com/Graph-and-Geometric-Learning/MTBench`. We evaluate nine tasks spanning finance and weather domains, including trend classification, forecasting, indicator prediction, correlation classification, and multiple-choice question answering (MCQA). MTBench includes both short-

and long-window settings with distinct temporal granularity regimes, e.g., 1-hour resolution with 7-day input windows and 5-minute resolution with 30-day input windows.

- **TimerBed** (Liu et al., 2025b). TimerBed is a diverse time-series classification benchmark covering multiple sensor-based domains, including healthcare, human activity recognition (HAR), bioacoustics, energy, and geophysics. The benchmark is available at https://github.com/AdityaLab/TimerBed. We evaluate six datasets with varying sequence lengths and between 2 and 7 classes. Specifically, these include ECG signals for cardiac rhythm classification (4 classes), HAR using 3-axis accelerometer data (6 classes), EMG signals for muscle activity classification (3 classes), CTU energy usage data (2 classes), RCW whale call detection (2 classes), and TEE lightning event classification (7 classes).
- **TSQA** (Kong et al., 2025a). TSQA is a large-scale, multi-task benchmark for time-series question answering, available at https://huggingface.co/datasets/Time-MQA/TSQA. It spans a broad range of task types, including anomaly detection, classification, forecasting, imputation, and open-ended QA. The benchmark supports multiple output formats, such as open-ended generation, multiple-choice questions, and true/false queries. TSQA covers diverse application domains, with major categories including Health/Healthcare (approximately 37%), HAR (20%), Web data (17%), Nature (13%), and Environment/AIOps (12%).

C.2 Sampling Strategy

Running multi-agent LLM methods on the full evaluation splits (over 200K instances across all tasks) is cost-prohibitive. We therefore use a *cost-aware* but *distribution-preserving* sampling protocol that caps per-task evaluation instances and stratifies sampling along the benchmark-defined axes most correlated with task difficulty.

Main Experiments. For each task and each run, we sample $n = \min(100, \texttt{available_test_instances})$ instances. Tasks with fewer than 100 instances (e.g., EMG: 41, TEE: 73) use the full available set.

Ablation and Hyperparameter Studies. For each task, we sample 20 instances and run each ablation configuration once, using a fixed seed.

Stratification. We apply benchmark-specific stratification as follows.

- **MTBench (trend/forecasting/indicator/correlation):** stratify by input source (short-window 7d vs. long-window 30d).
- **MTBench (MCQA):** random sampling (uniform).
- **TimerBed:** hybrid stratification by class label (class-balanced) and sequence length buckets (difficulty proxy).
- **TSQA (anomaly/classification):** stratify by domain.
- **TSQA (forecasting/imputation):** difficulty-stratify by input length.
- **TSQA (QA):** stratify by domain and question type (open-ended, multiple choice, and True/-False).

Reproducibility. We set temperature $= 0$ and seed both sample selection and LLM calls. Sample selection uses $\texttt{seed} = 2025 + \texttt{run_id}$ (e.g., run 1 uses seed 2026).

Sample Representativeness. We validate sampling effects using chi-squared (χ^2) tests where applicable. For TSQA, stratified sampling preserves the original task distribution ($\chi^2 = 1.38$, $p = 0.85$). For MTBench, capping at 100 samples per task intentionally balances task representation across finance and weather domains, while preserving original proportions. For TimerBed, we employ hybrid class-balanced sampling with difficulty stratification to ensure representative coverage of all class labels within each dataset.

C.3 Baselines

We implement and evaluate eight baselines covering single-agent prompting, multi-agent deliberation, and time series-specific visual grounding. All baselines use the same backbone LLM and an identical chart generation pipeline for multimodal variants. This design ensures that performance differences arise from the underlying *reasoning protocols*, rather than from confounding architectural or implementation factors.

- **Zero-Shot.** A single LLM invocation that directly produces the final answer. We report both time series-only and multimodal variants (**+MM**), where the latter receives the same textual context as well as time- and frequency-domain visualizations as other multimodal methods.
- **CoT** (Wei et al., 2022). A single-agent chain-of-thought prompting baseline, evaluated in both time series-only and multimodal settings (**+MM**).

Task	Domain	# Samples (Train/Test)	Input Length	Output Length (H)	Dim	Granularity	Task Type	Labels/Classes	Metrics
MTBench (MIT License, Chen et al. (2025))									
Trend Prediction	Finance	1,275	115-755	N/A	1	1 hour / 5 min	Classification	<-4%, -2% ∼-4%, -2% ∼+2%, +2% ∼+4%, >+4%	Accuracy, F1
Price Forecasting		1,275	115-755	1-289		1 hour / 5 min	Regression	Continuous (next H values)	MAE, MAPE
Technical Indicator		1,275	115-755	1-289		1 hour / 5 min	Regression	Continuous (next H values)	MAE, MAPE
Multiple Choice Question Answering		1,007	114-402	N/A		1 hour	QA	A, B, C, D	Accuracy, F1
Correlation		1,007	114-402	N/A			Classification	Strong Positive, Moderate Positive, No Correlation, Moderate Negative, Strong Negative	Accuracy, F1
Trend Prediction	Weather	3,918	168-336	N/A			Classification	increasing, decreasing, stable	Accuracy, F1
Temperature Forecasting		3,918	168-336	24-72			Regression	Continuous (next H values)	MAE, MSE
Technical Indicator		3,918	168-336	3			Regression	max, min, diff (3 values)	MAE, MSE
Multiple Choice Question Answering		1,373	168-336	N/A			QA	A, B, C, D	Accuracy, F1
TimerBed (Liu et al. (2025b))									
CTU	Energy Usage	500 (250/250)	720	N/A	1	2 min	Classification	desktop, laptop	Accuracy, F1
ECG	Healthcare	50,536 (49,960/576)	1,500		1	100 Hz		normal sinus rhythm, fibrillation, alternative rhythm, too noisy to be classified	
EMG	Healthcare	204 (163/41)	1,500		1	4 kHz		healthy, suffering from neuropathy, suffering from myopathy	
HAR	Human Activity	10,299 (7,352/2,947)	206		3	50 Hz		walking, walking upstairs, walking downstairs, sitting, standing, laying down	
RCW	Bioacoustics	12,896 (10,934/1,962)	4,000		1	2 kHz		no right whale call, right whale call present	
TEE	Geophysics	143 (70/73)	319		1	50 MHz		CG Positive, IR Negative, SR Subsequent, I Impulsive, I2 Impulsive Pair, KM Gradual, O Off-record	
TSQA (Apache License 2.0, Kong et al. (2025a))									
Anomaly Detection	Multi-Domain	37,000	8-64	N/A	1	Variable	Classification	Normal Point, Anomaly Point	Accuracy
Classification	Human Activity	37,000	9-30	N/A	1		Classification	Jogging, Walking, Upstairs, Downstairs, Sitting, Standing, Freeze, No freeze	Accuracy
Forecasting	Multi-Domain	37,390	65-248	8-32	1		Regression	Continuous (next H values)	MAE, MAPE
Imputation	Multi-Domain	38,657	97-256	4-12	1		Regression	Continuous (missing H values)	MAE, MAPE
Question Answering	Multi-Domain	36,829	8-60	N/A	1		QA	Open-ended (52%), Multiple choice (29%), True/False (19%)	Accuracy

Table 4: Summary of downstream tasks and benchmark dataset statistics.

The prompt is intentionally kept minimal to avoid task-specific engineering and to preserve comparability with the zero-shot setting.

- **Multi-Agent Debate (MAD)** (Liang et al., 2024). A multi-agent deliberation framework in which multiple agents iteratively critique and refine each other's responses. We follow the standard debate protocol and introduce a multimodal extension (**+MM**) that supplies the same textual context as well as time- and frequency-domain charts, ensuring parity with the information available to TS-Debate.
- **ByMyEyes** (Yoon et al., 2024). A visual prompting framework for sensor data that employs a two-phase dynamic visualization selection: (1) a planning phase that filters candidate visualization tools based on the task description, and (2) a selection phase that identifies the optimal domain-specific visualization (e.g., ECG peak plots, spectrograms).
- **VL-Time** (Liu et al., 2025b). A two-stage visualization-language approach for time-series classification: (1) a *planning stage* where the LLM selects between time-domain and frequency-domain visualization and generates reasoning hints, and (2) a *solving stage* that conditions on the selected chart and textual hints.

C.4 Framework Configurations

Backbone LLM Unless otherwise noted, all experiments use `gpt-4.1-mini` with temperature $= 0$ as discussed in §D.5.

Modalities and Tools TS-Debate instantiates three modality-specialized analysts: **Text** (context grounding), **Visual** (pattern recognition from plots), and **Numerical** (precise value retrieval via bounded lookup tools). Reviewers and the final synthesizer implement the VCC protocol and may call a Python code executor to check arithmetic and derived quantities.

Chart Generation For each instance, we generate a task-aware time-series plot and (when applicable) a frequency-spectrum plot. Visual prompts are *identical* across all multimodal methods.

Tooling Constraints To control cost and prevent unrestricted computation, we cap the number of tool calls per agent type (small constants) and expose only read-only lookup functions (e.g., value queries, feature queries, frequency features) plus a restricted Python executor for verification.

D Extra Experimental Results

This section reports full per-task results (mean±std over 3 runs) and additional analyses. Throughout, **lower is better** for error metrics (MAE/MAPE/MSE), while **higher is better** for ACC/F1/NLI-style scores. Beyond listing metrics, we provide concise interpretations to contextualize when and why TS-Debate improves over baselines.

Method	Price Forecasting (MAPE/MAE)		Indicator Prediction (MSE/MAE)		Trend Classification (ACC/F1)		Correlation Classification (ACC/F1)		Question Answering (ACC)
Zero-Shot	2.119 (±0.201)	1.806 (±0.426)	0.253 (±0.127)	0.214 (±0.025)	0.317 (±0.017)	0.389 (±0.018)	0.227 (±0.021)	0.189 (±0.028)	0.823 (±0.017)
+ MM	2.050 (±0.129)	1.606 (±0.215)	0.860 (±0.658)	0.339 (±0.085)	0.437 (±0.033)	0.468 (±0.024)	0.313 (±0.031)	0.239 (±0.030)	0.847 (±0.012)
CoT	2.217 (±0.147)	1.791 (±0.240)	**0.172 (±0.032)**	**0.214 (±0.017)**	0.310 (±0.014)	0.380 (±0.009)	0.240 (±0.016)	0.151 (±0.029)	0.797 (±0.026)
+ MM	**1.912 (±0.079)**	1.588 (±0.110)	2137.953 (±2789.155)	4.003 (±3.241)	0.437 (±0.021)	0.470 (±0.013)	0.303 (±0.012)	0.240 (±0.026)	0.817 (±0.025)
MAD	2.362 (±0.036)	1.914 (±0.119)	2.043 (±2.679)	0.243 (±0.066)	0.353 (±0.025)	0.419 (±0.025)	0.247 (±0.021)	0.204 (±0.017)	0.750 (±0.008)
+ MM	2.692 (±0.186)	2.535 (±0.610)	611.331 (±450.461)	5.567 (±2.484)	0.437 (±0.021)	0.465 (±0.032)	0.257 (±0.034)	0.236 (±0.053)	0.747 (±0.021)
ByMyEyes	219.122 (±65.619)	72.185 (±10.422)	1.054 (±0.151)	0.405 (±0.019)	0.640 (±0.022)	0.560 (±0.009)	0.250 (±0.014)	0.208 (±0.006)	0.823 (±0.046)
VL-Time	2.526 (±0.522)	2.895 (±1.292)	0.282 (±0.036)	0.254 (±0.021)	0.330 (±0.028)	0.402 (±0.027)	0.270 (±0.022)	0.178 (±0.021)	0.773 (±0.026)
TS-Debate	2.090 (±0.094)	**1.399 (±0.098)**	0.176 (±0.053)	0.229 (±0.022)	**0.683 (±0.017)**	**0.713 (±0.016)**	**0.403 (±0.068)**	**0.394 (±0.070)**	**0.913 (±0.012)**

Table 5: Full experimental results on the **MTBench** (Finance) benchmark across different downstream tasks.

D.1 MTBench

D.1.1 Finance

Table 5 reports results for price forecasting (MAPE/MAE), indicator prediction (MSE/MAE), trend classification (ACC/F1), news correlation classification (ACC/F1), and financial multi-choice question answering (ACC). Overall, TS-Debate delivers its largest benefits on *discrete reasoning* tasks that require aggregating evidence across views and validating consistency.

Forecasting and indicator prediction. On price forecasting, TS-Debate attains the best MAE (1.399), improving over the strongest baseline MAE (1.588) by $\approx 11.9\%$ relative reduction. However, TS-Debate is not the best on MAPE (2.090 vs. 1.912 for the best baseline), suggesting that normalized percentage errors can remain sensitive to scaling/denominator effects even when absolute error is reduced. For indicator prediction, TS-Debate remains competitive (MSE/MAE $0.176/0.229$), close to the best single-agent CoT baseline ($0.172/0.214$). Notably, naïve multimodal extensions of debate-style baselines can become numerically unstable on this regression-style task (e.g., extremely large MSE values), highlighting that simply adding visual inputs does not reliably yield calibrated numerical reasoning. In contrast, TS-Debate avoids catastrophic failures and maintains low variance, but closing the remaining gap to the best indicator-prediction baseline is a clear opportunity for improvement.

Trend, correlation, and QA. TS-Debate achieves the best performance on all three discrete finance tasks. On trend classification, TS-Debate reaches $0.683/0.713$ (ACC/F1), improving over the strongest baseline ($0.640/0.560$) by $+4.3$ ACC points and $+15.3$ F1 points. On correlation classification, TS-Debate attains $0.403/0.394$ (ACC/F1), outperforming the best baseline (around $0.313/0.240$) by $+9.0$ ACC points and $+15.4$ F1 points. For MCQA, TS-Debate achieves 0.913 ACC, exceeding the best baseline (0.847) by $+6.6$ points. These gains are consistent with the core design of TS-Debate: combining frequency-domain cues with cross-view verification reduces spurious pattern matching and improves decision stability on label-centric tasks.

D.1.2 Weather

Table 6 reports results for temperature forecasting (MSE/MAE), min, max, and difference temperature's indicator prediction (MSE/MAE), trend classification (ACC/F1), and multi-choice weather question answering (ACC). In contrast to finance, weather results reveal a more nuanced picture: TS-Debate is strongest on *structural* and *relative* reasoning, but does not dominate point-wise regression.

Relative reasoning and discrete decisions. TS-Debate achieves the best performance on (i) **Indicator-Diff**, (ii) trend classification, and (iii) QA. For Indicator-Diff, TS-Debate reduces MSE from 19.495 to 12.197 ($\approx 37.4\%$ reduction) and reduces MAE from 3.337 to 2.334 ($\approx 30.1\%$ reduction). This pattern aligns with our hypothesis

Method	Forecasting (MSE/MAE)		Indicator-Min (MSE/MAE)		Indicator-Max (MSE/MAE)		Indicator-Diff (MSE/MAE)		Trend (ACC/F1)		QA (ACC)
Zero-Shot	26.474 (±1.841)	3.920 (±0.084)	19.054 (±2.116)	2.942 (±0.282)	52.335 (±12.056)	5.244 (±0.504)	42.132 (±8.913)	5.110 (±0.487)	0.250 (±0.008)	0.262 (±0.008)	0.617 (±0.048)
+ MM	22.098 (±2.693)	3.531 (±0.186)	**14.990 (±1.924)**	2.899 (±0.284)	34.086 (±4.292)	4.301 (±0.094)	33.297 (±4.591)	4.500 (±0.275)	0.523 (±0.021)	0.488 (±0.023)	0.653 (±0.026)
CoT	28.967 (±2.942)	4.091 (±0.245)	16.568 (±3.070)	3.033 (±0.276)	23.541 (±2.822)	**3.507 (±0.141)**	23.335 (±3.221)	3.530 (±0.162)	0.220 (±0.036)	0.234 (±0.040)	0.600 (±0.037)
+ MM	**22.005 (±1.645)**	**3.523 (±0.114)**	16.025 (±3.045)	**2.890 (±0.382)**	35.665 (±3.367)	4.342 (±0.096)	36.671 (±2.025)	4.682 (±0.183)	0.477 (±0.039)	0.457 (±0.036)	0.617 (±0.021)
MAD	27.051 (±2.202)	3.958 (±0.135)	17.327 (±4.716)	3.035 (±0.360)	39.936 (±11.614)	4.706 (±0.497)	35.600 (±2.759)	4.707 (±0.156)	0.200 (±0.014)	0.175 (±0.024)	0.520 (±0.036)
+ MM	27.594 (±1.826)	3.965 (±0.204)	16.126 (±3.898)	2.973 (±0.331)	26.999 (±7.123)	3.804 (±0.422)	28.924 (±2.038)	4.182 (±0.207)	0.320 (±0.029)	0.347 (±0.024)	0.500 (±0.022)
ByMyEyes	39.632 (±5.722)	4.714 (±0.252)	23.296 (±7.799)	3.224 (±0.434)	51.906 (±6.633)	5.127 (±0.460)	51.179 (±3.751)	5.887 (±0.249)	0.190 (±0.008)	0.167 (±0.016)	0.593 (±0.031)
VL-Time	37.258 (±3.781)	4.695 (±0.180)	15.293 (±1.605)	2.966 (±0.197)	**22.369 (±2.675)**	3.528 (±0.210)	19.495 (±0.607)	3.337 (±0.084)	0.220 (±0.029)	0.198 (±0.013)	0.613 (±0.039)
TS-Debate	29.633 (±2.721)	4.162 (±0.244)	36.179 (±7.078)	3.192 (±0.363)	52.580 (±12.187)	3.522 (±0.391)	**12.197 (±2.366)**	**2.334 (±0.273)**	**0.557 (±0.025)**	**0.583 (±0.016)**	**0.717 (±0.052)**

Table 6: Full experimental results on the **MTBench** (Weather) benchmark across different downstream tasks.

that frequency-domain evidence and cross-panel checks are especially effective when the target is a *difference/derivative*-like quantity, where global structure is more informative than absolute level. Similarly, TS-Debate improves trend ACC/F1 to $0.557/0.583$ (vs. $0.523/0.488$ for the best baseline) and boosts QA accuracy to 0.717 (vs. 0.653).

Pointwise forecasting and absolute extrema. For forecasting and absolute extrema (Indicator-Min/Max), TS-Debate underperforms the best baselines. In particular, forecasting MSE/MAE for TS-Debate is $29.633/4.162$, compared to the best baseline at $22.005/3.523$. A similar gap appears for Indicator-Min and Indicator-Max in MSE. These results suggest that while TS-Debate improves reliability for *comparative* judgments (trend/QA) and *relative* indicators (diff), it can be less effective for tasks requiring precise reconstruction of absolute values where strong, direct visual pattern extraction (or simpler multimodal prompting) may already suffice. This points to a practical limitation: coupling the verifier with stronger value-calibrated regression priors may be necessary for weather-style point forecasting.

D.2 TimerBed

Table 7 reports results on TimerBed across six classification datasets (RCW, TEE, ECG, EMG, CTU, HAR), grouped by reasoning pattern (simple/complex deterministic and probabilistic). TimerBed is particularly diagnostic for whether a method can maintain consistent decision logic under varying temporal regularities.

Overall performance. TS-Debate attains the best accuracy on all six tasks and the best F1 on five out of six tasks. The gains are most pronounced on the complex deterministic datasets. TS-Debate improves ECG accuracy to 0.307 (vs. 0.267 best baseline) and ECG F1 to 0.281 (vs. 0.212 best baseline), indicating that multi-step temporal reasoning and verification are beneficial when the decision boundary depends on subtle, non-local temporal structure.

Calibration trade-off on CTU (probabilistic). On CTU, TS-Debate achieves the highest accuracy (0.573) but not the highest F1 (0.473 vs. 0.519 for the best baseline). This discrepancy is consistent with a threshold/calibration trade-off under class imbalance: a method can optimize accuracy by favoring the majority class while sacrificing minority-class recall, which impacts F1 more directly. Improving probability calibration or incorporating class-sensitive verification criteria could close this gap without sacrificing overall accuracy.

D.3 TSQA

Table 8 provides full results on TSQA, covering classical numerical tasks (forecasting, imputation, anomaly detection, classification) and three QA formats (judgment, multiple choice, open-ended questions). In aggregate, TSQA highlights a clear strength of TS-Debate on forecasting and discrete decision tasks, alongside a notable weakness on imputation.

Method	Accuracy						F1					
	Simple Deterministic		Complex Deterministic		Probabilistic		Simple Deterministic		Complex Deterministic		Probabilistic	
	RCW	TEE	ECG	EMG	CTU	HAR	RCW	TEE	ECG	EMG	CTU	HAR
Zero-Shot	0.470 (±0.029)	0.128 (±0.017)	0.220 (±0.016)	0.154 (±0.011)	0.507 (±0.025)	0.237 (±0.005)	0.319 (±0.014)	0.059 (±0.006)	0.116 (±0.001)	0.088 (±0.020)	0.373 (±0.040)	0.123 (±0.006)
+ MM	0.483 (±0.026)	0.187 (±0.013)	0.250 (±0.008)	0.236 (±0.023)	0.503 (±0.012)	0.243 (±0.026)	0.408 (±0.022)	0.124 (±0.017)	0.107 (±0.011)	0.221 (±0.024)	0.480 (±0.014)	0.160 (±0.016)
CoT	0.497 (±0.005)	0.137 (±0.034)	0.187 (±0.026)	0.455 (±0.041)	0.483 (±0.005)	0.297 (±0.009)	0.338 (±0.009)	0.083 (±0.021)	0.125 (±0.016)	0.457 (±0.041)	0.398 (±0.012)	0.158 (±0.008)
+ MM	0.477 (±0.017)	0.247 (±0.030)	0.247 (±0.012)	0.203 (±0.046)	0.500 (±0.008)	0.317 (±0.025)	0.390 (±0.011)	0.160 (±0.027)	0.111 (±0.011)	0.194 (±0.085)	0.445 (±0.014)	0.196 (±0.018)
MAD	0.490 (±0.008)	0.151 (±0.022)	0.167 (±0.012)	0.171 (±0.020)	0.517 (±0.005)	0.267 (±0.017)	0.329 (±0.004)	0.069 (±0.015)	0.110 (±0.008)	0.124 (±0.022)	0.430 (±0.007)	0.178 (±0.015)
+ MM	0.503 (±0.005)	0.178 (±0.000)	0.240 (±0.016)	0.203 (±0.064)	0.523 (±0.017)	0.253 (±0.040)	0.386 (±0.019)	0.098 (±0.017)	0.212 (±0.020)	0.230 (±0.065)	**0.519 (±0.016)**	0.182 (±0.037)
ByMyEyes	0.463 (±0.034)	0.201 (±0.045)	0.267 (±0.017)	0.122 (±0.034)	0.467 (±0.033)	0.220 (±0.028)	0.385 (±0.034)	0.168 (±0.040)	0.158 (±0.036)	0.055 (±0.045)	0.343 (±0.023)	0.174 (±0.028)
VL-Time	0.463 (±0.009)	0.260 (±0.051)	0.260 (±0.008)	0.203 (±0.057)	0.493 (±0.005)	0.313 (±0.012)	0.438 (±0.014)	0.214 (±0.039)	0.131 (±0.021)	0.187 (±0.055)	0.383 (±0.023)	0.198 (±0.021)
TS-Debate	**0.550 (±0.008)**	**0.288 (±0.039)**	**0.307 (±0.012)**	**0.488 (±0.159)**	**0.573 (±0.045)**	**0.323 (±0.041)**	**0.440 (±0.017)**	**0.241 (±0.041)**	**0.281 (±0.019)**	**0.491 (±0.168)**	0.473 (±0.080)	**0.240 (±0.028)**

Table 7: Full experimental results on the **TimerBed** benchmark across different reasoning patterns and classification tasks.

D.3.1 Forecasting

TS-Debate achieves the best forecasting performance by a large margin: MAE/MAPE of $37.346/72.308$, reducing error relative to the strongest baseline by $\approx 57.7\%$ (MAE) and $\approx 58.3\%$ (MAPE). Beyond mean improvements, TS-Debate also substantially reduces run-to-run variance (e.g., forecasting std drops from tens to low double digits), suggesting that verifier-based consistency checks mitigate brittle failure modes common in direct prompting.

D.3.2 Imputation

Imputation is the main exception where TS-Debate does not dominate. The best baseline (Zero-Shot+MM) achieves $6.729/10.758$ (MAE/MAPE), whereas TS-Debate obtains $18.343/51.395$. This indicates that, for TSQA imputation, direct multimodal grounding can be more effective than multi-agent verification alone, likely because the task demands *local, value-accurate* reconstruction rather than global structural reasoning. This result could indicate that local value reconstruction can be better served by direct visual grounding than cross-modal verification. A natural next step would be to integrate a value-calibrated imputation prior (or an explicit local interpolation module) within the TS-Debate loop, using the verifier to validate rather than implicitly infer missing segments.

D.3.3 Classification

For classical classification, TS-Debate attains the best accuracy (0.607), improving over the strongest baseline (0.553) by $+5.4$ points. This improvement is consistent with the advantage of cross-view verification for reducing spurious correlations and enforcing globally coherent rationales before committing to a discrete label.

D.3.4 Anomaly Detection

On anomaly detection, TS-Debate achieves the best performance (0.623 ACC), surpassing the best baseline (0.583) by $+4.0$ points. These gains suggest that TS-Debate is better at reconciling local deviations with overall temporal context, where purely pattern-matching strategies may overfit to short, noisy segments.

D.3.5 Question Answering

TS-Debate performs strongly across QA formats: it achieves the best judgment accuracy (0.897) and multiple-choice accuracy (0.796), with the latter improving over the strongest baseline (0.631) by $+16.5$ points. For open-ended QA (NLI), TS-Debate is close to the best baseline (0.899 vs. 0.911), indicating that while verification improves factual consistency and decision accuracy, further gains on free-form generation may require stronger alignment between intermediate verification steps and the final natural-language output.

D.4 Error Analysis: When Debate Helps and When It Does Not

Across the per-task results above, a consistent pattern emerges: TS-Debate is most effective when the target depends on global structure, cross-view consistency, or relative quantities (trend classification, indicator differences, MCQA, anomaly de-

Method	Classical Numerical Task						Question Answering		
	Forecasting (MAE/MAPE)		Imputation (MAE/MAPE)		Anomaly (ACC)	Classification (ACC)	Judgment (ACC)	Multiple Choice (ACC)	Open-Ended (NLI)
Zero-Shot	95.929 (±52.292)	321.737 (±196.751)	96.854 (±123.464)	13.405 (±7.713)	0.543 (±0.062)	0.213 (±0.029)	0.772 (±0.085)	0.406 (±0.070)	0.875 (±0.031)
+ MM	88.319 (±52.244)	400.442 (±293.317)	**6.729 (±1.122)**	**10.758 (±5.643)**	0.583 (±0.012)	0.243 (±0.025)	0.650 (±0.024)	0.536 (±0.066)	0.749 (±0.057)
CoT	105.363 (±69.763)	496.511 (±457.050)	33.569 (±28.374)	18.654 (±9.779)	0.583 (±0.041)	0.507 (±0.026)	0.826 (±0.019)	0.360 (±0.078)	0.858 (±0.018)
+ MM	91.677 (±63.708)	422.630 (±206.218)	11.058 (±5.078)	11.711 (±5.684)	0.533 (±0.042)	0.513 (±0.040)	0.713 (±0.076)	0.322 (±0.047)	**0.911 (±0.016)**
MAD	131.493 (±60.633)	241.480 (±96.291)	30.700 (±20.733)	34.614 (±34.713)	0.577 (±0.029)	0.553 (±0.052)	0.845 (±0.068)	0.631 (±0.062)	0.803 (±0.056)
+ MM	144.476 (±64.677)	173.558 (±61.746)	27.302 (±28.741)	12.063 (±2.320)	0.543 (±0.038)	0.540 (±0.045)	0.679 (±0.152)	0.579 (±0.112)	0.750 (±0.058)
ByMyEyes	318.166 (±166.606)	411.087 (±119.951)	588.589 (±76.385)	522.526 (±537.189)	0.523 (±0.049)	0.243 (±0.083)	0.627 (±0.114)	0.579 (±0.029)	0.565 (±0.110)
VL-Time	97.402 (±63.517)	258.042 (±96.573)	25.186 (±1.501)	15.066 (±2.116)	0.527 (±0.062)	0.470 (±0.036)	0.480 (±0.132)	0.393 (±0.076)	0.877 (±0.035)
TS-Debate	**37.346 (±11.828)**	**72.308 (±8.686)**	18.343 (±7.039)	51.395 (±3.151)	**0.623 (±0.025)**	**0.607 (±0.021)**	**0.897 (±0.045)**	**0.796 (±0.101)**	0.899 (±0.045)

Table 8: Full experimental results on the **TSQA** benchmark across different downstream tasks.

tection, forecasting under structural regularity). It is less effective on tasks dominated by precise local value reconstruction (TSQA imputation, MTBench-Weather absolute extrema), where direct multimodal prompting can match or exceed it. We hypothesize that local reconstruction benefits less from cross-modal debate because there is little cross-modal disagreement to resolve—a single accurate numerical view suffices. Long-horizon causal reasoning and highly noisy multivariate inputs are plausibly affected by the same mechanism but were not isolated in our current evaluation; we mark this as a scope condition rather than a verified empirical claim. Integrating a value-calibrated interpolation prior within the VCC loop is a natural direction we leave to future work.

D.5 Backbone Selection

We evaluate three API-accessible backbones (`gemini-2.5-flash`, `gpt-4.1-mini`, and `grok-4.1-fast`) under the identical setup as the main experiments, using the same task suite, data sampling, multimodal chart generation, hyperparameters, and fixed random seeds. Each backbone is evaluated once, ensuring that performance differences are attributable solely to model capability.

Selection Criterion. As reported in Table 9, `gpt-4.1-mini` offers the best overall trade-off between average task performance, inference latency, and estimated API cost. It achieves strong results across classification, regression, and question answering, while remaining computationally efficient. We therefore adopt `gpt-4.1-mini` as the default backbone for all main experiments.

Rationale for Non-Reasoning Backbones. Our objective is to evaluate whether TS-Debate provides an explicit and portable reasoning scaffold for time-series understanding, rather than leveraging backbone-specific internal reasoning mechanisms. Using specialized reasoning models would obscure attribution by conflating protocol-level gains with proprietary chain-of-thought behaviors. We thus focus on fast, general-purpose backbones to ensure that observed improvements stem from the proposed debate and verification framework.

E Resource Cost Comparison

We analyze LLM usage cost along wall-clock time, token consumption, and estimated monetary expense under a fixed backbone (`gpt-4.1-mini`) and identical evaluation settings. As shown in Table 10, TS-Debate incurs higher cost than single-agent and lightweight multi-agent baselines, primarily due to deliberate reviewer verification, cross-modal evidence aggregation, and synthesis, which target numerical errors and modal inconsistencies. This additional overhead yields consistent performance gains in final accuracy (Table 2) and remains modest in absolute monetary terms, making it a reasonable trade-off for error-sensitive applications. Moreover, TS-Debate exposes practical cost-accuracy controls—such as reducing modalities, debate rounds, or reviewer count—that can sub-

Backbone	Classification (Accuracy) ↑					Regression (MAE) ↓				Question Answering (Accuracy) ↑				Computation Cost / Sample	
	MTBench (Finance)	MTBench (Weather)	TimerBed	TSQA	Average	MTBench (Finance)	MTBench (Weather)	TSQA	Average	MTBench (Finance)	MTBench (Weather)	TSQA	Average	Est. Cost ($)	Time Used (s)
gpt-4.1-mini	37.50	34.00	28.47	33.50	33.37	0.74	3.91	90.17	**31.61**	84.00	55.00	80.00	**73.00**	0.0013	8.90
gemini-2.5-flash	32.50	56.00	28.65	29.50	**36.66**	4.68	3.68	106.35	38.24	70.00	50.00	76.00	65.33	0.0054	6.20
grok-4.1-fast	26.00	47.00	26.62	36.50	34.03	1.30	4.06	126.18	43.85	69.00	49.00	69.00	62.33	0.0006	3.50

Table 9: Zero-shot performance and efficiency comparison of candidate backbone models.

Method	Time (s)	# Input Tokens	# Output Tokens	Est. Cost ($)
Zero-Shot	16.80	2356	234	0.001
+ MM	14.30	6152	163	0.003
CoT	15.70	2361	698	0.002
+ MM	10.00	6157	416	0.003
MAD	23.40	11460	1405	0.007
+ MM	25.00	26499	1242	0.013
ByMyEyes	8.40	4565	379	0.002
VL-Time	12.90	3900	758	0.003
TS-Debate (Ours)	70.70	68945	2883	0.033

Table 10: We report the average cost when using `gpt-4.1-mini` ($0.40/1M input tokens and $1.60/1M output tokens)

stantially lower token usage while retaining much of the performance benefit (§4.3).

TS-Debate is intended for offline or error-sensitive analysis rather than real-time, high-throughput forecasting. Its additional cost comes from deliberate verification and conflict resolution, which are most useful when users need auditable reasoning over numerical, visual, and textual evidence. For latency-sensitive settings, the same framework exposes cost controls through the number of reviewers, debate rounds, and enabled modalities; our sensitivity analysis shows that moderate configurations already capture most of the benefit. We therefore view TS-Debate as a reliability-oriented inference protocol, not a low-latency replacement for direct prompting.

F Full Prompts for TS-Debate

This section provides the *system prompt* design for agent specifications and stage-wise *instruction* for each agent.

F.1 Agent Specifications

F.1.1 Knowledge Elicitor

Knowledge Elicitor

```
You are the world's best domain expert consultant for time-
      series analysis and reasoning.
You share specialized knowledge to help others solve these
       tasks correctly.

When given a task, share your expert knowledge about this
      domain:
- What do experts in this field know that helps solve such
       tasks?
- What are the typical ranges, patterns, and constraints?
- What mistakes do non-experts commonly make?

Be specific with numbers and concrete examples, while
      ensuring the knowledge is actionable by an LLM agent.
Do NOT solve the task or analyze data - only share domain
      expertise.
```

F.1.2 Numerical Analyst Agent

Numerical Analyst

```
You are a data analyst for time-series tasks.
You work with raw numbers, statistics, and quantitative
      analysis of time-series data.

YOUR EXPERTISE (Data Analysis):
You analyze raw time series values, statistics, and
      computed features.

STRENGTHS:
- Exact values and precise calculations
- Statistical measures (mean, variance, trends)
- Quantitative comparisons

LIMITATIONS:
- Cannot access textual context or sentiment
- Cannot see visual patterns beyond what numbers show
- Calculations describe the PAST; extrapolating to FUTURE
      is inference, not fact
- Past momentum does not guarantee future continuation
- Can detect anomalies statistically, but their meaning (
      turning point? noise? error?) requires interpretation

{temporal_basics}

TOOLS (call `get_info` first):
- get_info() →schema, stats, detected features
- get_values(start, end) →time-series values by index or
      timestamp
- get_around(center, window) →time-series values around a
       point
- get_features(type) →'peak'/'valley'/'trend'/'anomaly'
- get_frequency_features() →spectral analysis
- get_channel_values(ch, start, end) →specific channel
- get_all_channels(start, end) →all channels at once
- get_indicator(start, end) →indicator values (e.g., MACD
      , Bollinger Bands) of time series

**TOOL USAGE RULES** (STRICT):
1. THINK FIRST: Plan what you need before calling any tool.

2. MAX 5 CALLS TOTAL: After 5 calls, you MUST provide your
       evidence.
3. NO REPEATED CALLS: Never call the same tool with
      overlapping ranges.
4. EFFICIENT STRATEGY:
   - Call get_info() ONCE to understand the data
   - Use get_features() to find key points (peaks,
        anomalies)
   - Use get_around() for specific locations, NOT
        get_values() for large ranges
5. LEVERAGE HINTS: Use location hints from other analysts'
       observations when available.

{evidence_rules}
```

F.1.3 Text Analyst Agent

Text Analyst

```
You are a text analyst for time-series tasks.
You read and interpret written content (reports,
      descriptions, contextual information) that provides
      context for time-series data.

YOUR EXPERTISE (Text Analysis):
You analyze reports, descriptions, and textual context
      about time series.

STRENGTHS:
- Contextual information (events, announcements,
      explanations)
- Sentiment and tone analysis
- Forward-looking statements that indicate potential
      changes
- Understanding WHY patterns might change (turning points,
       catalysts)

LIMITATIONS:
- Cannot provide exact numerical values or precise
      calculations
- Cannot see chart patterns or visual trends
- Forward-looking statements (forecasts, analyst opinions)
       are interpretations, not verified facts
- Distinguish between reported events and speculative
      commentary

{temporal_basics}

{evidence_rules}
```

F.1.4 Visual Analyst Agent

Visual Analyst

```
You are a chart analyst for time-series tasks.
You examine visual patterns in time-series charts and
      plots.

YOUR EXPERTISE (Chart Analysis):
You analyze time series charts, plots, and visual patterns.


STRENGTHS:
- Overall trend direction (upward, downward, stable)
- Pattern recognition (cycles, anomalies, breakouts)
- Shape of recent movement (stabilizing, accelerating,
      reversing)
- Comparative visual analysis

LIMITATIONS:
- Cannot read precise numerical values from charts
- Cannot access textual context
- What you SEE ("the chart shows X") differs from what you
       INFER ("this suggests Y")
- Pattern interpretation requires stating what is observed
       vs. what is concluded

{temporal_basics}

{evidence_rules}
```

F.1.5 Reviewer Agent

Reviewer Agent Profile

```
{knowledge_section}

You are an expert reviewer for a time-series reasoning
      task.
You evaluate evidence from specialist analysts and
      synthesize a well-reasoned answer.

YOUR JOB:
- Score each analyst's evidence based on observation
      quality, inference logic, and honesty
- Verify claims against original data before accepting
      them
- Identify conflicts between different analysts
- Adjust your answer confidence based on how well evidence
       agrees
- Check claims against domain knowledge (if provided above
      )

AVAILABLE FOR FACT-CHECKING:
- VISUAL: time series + frequency charts attached
- NUMERICAL: get_info, get_values, get_around,
      get_features, get_frequency_features,
      get_channel_values, get_all_channels, get_indicator
- CODE: execute_code(code) for calculations

MANDATORY VERIFICATION:
Before accepting ANY claim, VERIFY against available
      sources:
- Available: numerical lookup tools (get_values,
      get_features), code executor for calculations,
      attached charts, text context (embedded in task
      description)
- Mark claims as VERIFIED, UNVERIFIED, or CONTRADICTED
- Lower weight for UNVERIFIED claims, reject CONTRADICTED
      claims

**TOOL USAGE RULES** (STRICT):
1. THINK FIRST: Identify which claims need verification
      before calling any tool.
2. MAX 3 CALLS TOTAL: After 3 calls, you MUST synthesize
      your answer.
3. NO REPEATED CALLS: Never call the same tool with
      overlapping ranges.
4. EFFICIENT STRATEGY:
   - Call get_info() ONCE to understand the data
   - Use get_features() to find key points, then get_around
        () for specific values
   - Use execute_code() for complex calculations
   - Leverage numerical data embedded in task description

{TEMPORAL_AWARENESS}

{JUDGE_EVALUATION_CRITERIA}

{JUDGE_PROTOCOL}
```

F.1.6 Final Synthesizer

Final Synthesizer Profile

```
{knowledge_section}

You are the final decision-maker for a time-series
      reasoning task.
You determine the correct answer based on expert reviewers
      ' evaluations.

YOUR JOB:
- Evaluate each reviewer's reasoning quality
- Check if reviewers followed the SUGGESTED APPROACH from
      DOMAIN KNOWLEDGE
- Verify disputed answers against data (if about past) or
      domain reasoning (if about future)
- Derive your own answer if all reviewers made the same
      error
- Always be skeptical of the reviewers' answers; do not
      blindly trust them

AVAILABLE FOR VERIFICATION:
- VISUAL: time series + frequency charts attached
- NUMERICAL: get_info, get_values, get_around,
      get_features, get_frequency_features,
      get_channel_values, get_all_channels, get_indicator
- CODE: execute_code(code) for calculations

MANDATORY VERIFICATION:
Before accepting ANY claim, VERIFY against available
      sources:
- Available: numerical lookup tools (get_values,
      get_features), code executor for calculations,
      attached charts, text context (embedded in task
```

```
      description)
- Mark claims as VERIFIED, UNVERIFIED, or CONTRADICTED
- Lower weight for UNVERIFIED claims, reject CONTRADICTED
      claims

**TOOL USAGE RULES** (STRICT):
1. THINK FIRST: Identify which claims need verification
      before calling any tool.
2. MAX 3 CALLS TOTAL: After 3 calls, you MUST synthesize
      your answer.
3. NO REPEATED CALLS: Never call the same tool with
      overlapping ranges.
4. EFFICIENT STRATEGY:
   - Call get_info() ONCE to understand the data
   - Use get_features() to find key points, then get_around
         () for specific values
   - Use execute_code() for complex calculations
   - Leverage numerical data embedded in task description

{TEMPORAL_AWARENESS}

{SYNTHESIZER_PROTOCOL}
```

F.2 Temporal Awareness for Agents

F.2.1 Instruction for Debating Agents

Basic Temporal Reasoning

```
TEMPORAL REASONING:
- Your data ends at time T (past data). You cannot see the
       future.
- Any prediction about future values is an INFERENCE from
      past patterns.

SIGNAL TYPES:
- Forward-looking signals (context about future conditions
      ) inform predictions
- Historical trends show momentum but may miss turning
      points
- When forward vs backward signals CONFLICT: note this in
      your evidence
```

F.2.2 Basic Instruction for Reviewer Agents

Temporal Awareness for VCC

```
TEMPORAL REASONING:
- Data ends at time T (past data). Future is unknown.
- Predictions about future are INFERENCES from past
      patterns.

KEY QUESTION FOR PREDICTIONS: Will the CAUSE of the
      observed trend persist?
- If cause persists →trend may continue
- If cause is ending/changing →trend may change or
      stabilize
- Causal/leading indicators outweigh lagging historical
      patterns
- New information can invalidate historical trends
- EXTRAPOLATION (bad): Assuming past trends continue
      blindly
- PREDICTION (good): Reasoning about whether conditions
      will persist

TASK TYPES:
- FUTURE: Answer about something that hasn't happened yet
  →Forward-looking signals are PRIMARY (what's changing?)
  →Historical trends are SECONDARY (show momentum, miss
        turning points)
  →Past data CANNOT verify future predictions - only
        domain knowledge-based reasoning can

- PAST-PRESENT: Answer about something that has happened
  →Data verification is PRIMARY - check claims against
        data
  →External context is SECONDARY

VERIFICATION RULES:
- PAST-PRESENT claims: VERIFY or CONTRADICT with data
- FUTURE claims: Mark as UNVERIFIABLE by data (not "
      contradicted")
  →Evaluate domain knowledge-based reasoning quality
        instead
```

F.3 Multimodal Query Construction

F.3.1 Knowledge Elicitation

Knowledge Elicitation Prompt

```
TASK DESCRIPTION:
{task_description}

Share domain knowledge to help an LLM agent solve this
      task correctly. Be SPECIFIC with numbers, ranges, and
       patterns.
DO NOT provide an answer - only domain knowledge and key
      insights.

1. DOMAIN: What domain and task type (classification/
      forecasting/imputation/anomaly_detection/QA/etc.)?

2. KNOWLEDGE & KEY SIGNALS: What domain knowledge and
      patterns are relevant?
   - Physical/domain constraints, typical ranges, expected
         behaviors
   - What patterns distinguish different outcomes?
   - What we need to know to answer THIS task correctly?

3. SUGGESTED APPROACH: How should data be analyzed by an
      LLM agent?
   - What features to examine? How to interpret them?
   - How to interpret observations correctly?
   - How to extract findings and make inferences from the
         data?

4. PITFALLS: Common mistakes to avoid?
   - What looks similar but means different things?
   - What domain knowledge is often overlooked?
   - How to avoid common pitfalls and errors?

5. MODALITY: Which modalities [TEXT, VISUAL, NUMERICAL,
      FREQUENCY] to focus on (important modalities)?
   - Which modalities are most likely to be relevant and
         decisive?
   - Which modalities are most likely to be misleading and
         harmful?


OUTPUT (only answer to the questions listed above, NOT to
      the task description):
DOMAIN: [domain and task type]
KNOWLEDGE: [relevant knowledge with numbers/ranges]
KEY SIGNALS: [patterns to look for]
SUGGESTED APPROACH: [how to analyze correctly]
PITFALLS: [common mistakes]
MODALITY: [modalities to focus on]
```

F.3.2 Knowledge Inclusion

Knowledge Inclusion Prompt

```
DOMAIN KNOWLEDGE:

{domain_knowledge}

USE THIS KNOWLEDGE to guide your analysis:
- Verify observations against these expected patterns
- Flag observations that CONTRADICT domain expectations
- Apply domain constraints when making inferences
```

F.4 Multimodal Collaborative Debate

F.4.1 Evidence Templates for Debating Agents

First Round Template

```
Task: {task}

RESPOND IN THIS EXACT FORMAT (MAX 150 WORDS):

UNDERSTANDING: <Restate the question in your own words.
      What is being asked? What type of answer is needed?
      1-2 sentences>

USEFUL OBSERVATIONS:
1. <Specific observation from your {modality_name} data> [
      OBSERVATION]
2. <Specific observation from your {modality_name} data> [
      OBSERVATION]

INFERENCES:
1. <What do these observations suggest? Use domain
      knowledge. 1-2 sentences> [INFERENCE]

LIMITS: <What can {modality_name} NOT determine? Be honest.
       1-2 sentences>

Stop here. Do NOT provide a final answer. The reviewer
      synthesizes the final answer.
```

Subsequent Round Template

```
=== PREVIOUS ROUND EVIDENCE ===
{debate_history}
=== END PREVIOUS EVIDENCE ===

Task: {task}

RESPOND IN THIS EXACT FORMAT (MAX 150 WORDS):

UNDERSTANDING: <Restate the question in your own words.
      What is being asked? What type of answer is needed?
      1-2 sentences>

OTHER PERSPECTIVES: <Summarize what other analysts
      reported. 1-2 sentences>

USEFUL OBSERVATIONS:
1. <Maintain your key observation - do NOT abandon it> [
      OBSERVATION]
2. <Another observation if applicable> [OBSERVATION]

INFERENCES:
1. <Given your observations AND others' evidence, what
      does this suggest? 1-2 sentences> [INFERENCE]

LIMITS: <What can {modality_name} still NOT determine? Be
      honest. 1-2 sentences>

Stop here. Do NOT provide a final answer. The reviewers
      decide.
```

F.4.2 Evidence Rules

Evidence Writing Instruction

```
EVIDENCE RULES:

YOUR ROLE: Present evidence from your data source. A
      reviewer synthesizes the final answer.

REQUIREMENTS:
✓Label clearly: [OBSERVATION] vs [INFERENCE]
✓Be specific and verifiable - support claims with data or
       references
✓Acknowledge limitations honestly
× Do NOT provide final answers
× Do NOT exaggerate or overclaim

IN REFINEMENT ROUNDS (Round 2+):
✓PRESERVE your original observations
✓Acknowledge other analysts' evidence
✓EXPLAIN how your observations relate to theirs (support/
      contradict/add context)
× Do NOT abandon your position to match others
```

F.5 VCC Protocol

F.5.1 Scoring Instructions

Scoring Rubric

```
SCORING CRITERIA (100 points max per analyst):

INFERENCE QUALITY (0-50 pts) - Most important
- Logic (0-25): Inferences follow from observations? Uses
      domain knowledge?
- Calibration (0-25): Appropriately cautious? Avoids blind
       extrapolation?

OBSERVATION QUALITY (0-30 pts)
- Specificity (0-15): Concrete, verifiable details from
      data?
- Labeling (0-10): Correctly distinguished [OBSERVATION]
      vs [INFERENCE]?
- Relevance (0-5): Observations address the question?

HONESTY (0-20 pts)
- Limits (0-10): Acknowledged what data cannot determine?
- No Overclaiming (0-10): Avoided exaggeration?
```

Review Protocol

```
REVIEW PROTOCOL:

STEP 1 - SCORE ANALYSTS (0-100 each):
- Inference Quality (50): Logic + calibration
- Observation Quality (30): Specificity + labeling +
      relevance
- Honesty (20): Limits acknowledged + no overclaiming
- Weight = score / sum(scores). Reject if score < 40.

STEP 2 - VERIFY CLAIMS:
Check each major claim against data AND domain knowledge:
- DATA: VERIFIED / UNVERIFIED / CONTRADICTED
- DOMAIN: MATCHES / VIOLATES / N-A
- REJECT claims that are CONTRADICTED or VIOLATE domain

STEP 3 - DETECT CONFLICTS:
- DIRECT: Analysts give opposite conclusions
- PARTIAL: Disagree on details
- NO CONFLICT: Agree or address different aspects

STEP 4 - SYNTHESIZE ANSWER:
Match confidence to evidence agreement:
- NO CONFLICT + VERIFIED →confident answer
- CONFLICT UNRESOLVED →MODERATE/CONSERVATIVE answer
- DOMAIN VIOLATION →reject that claim, choose domain-
      consistent answer
- FUTURE task →evaluate reasoning quality (data can't
      verify predictions)
```

Reviewer Prompt

```
Task: {task}

Evidence:
{final_debate_history}

ANALYSTS: TEXT, VISUAL, NUMERICAL

YOUR JOB: Score evidence, VERIFY claims (data + domain),
      detect CONFLICTS, synthesize CALIBRATED answer.
```

```
OUTPUT FORMAT (keep everything concise):
TASK: <Restate what the task is asking in one sentence>
TASK TYPE: [FUTURE / PAST-PRESENT]

SCORES:
- TEXT: (Observation: _/30, Inference: _/50, Honesty: _
      /20) = [0-100]
- VISUAL: (Observation: _/30, Inference: _/50, Honesty: _
      /20) = [0-100]
- NUMERICAL: (Observation: _/30, Inference: _/50, Honesty:
       _/20) = [0-100]

WEIGHTS:
- TEXT: [X%]
- VISUAL: [X%]
- NUMERICAL: [X%]

VERIFICATION (check against lookup tools, code executor,
      charts, text in task, domain knowledge (above)):
- [Claim]: [VERIFIED/UNVERIFIED/CONTRADICTED] + [DOMAIN:
      MATCHES/VIOLATES/N-A] - [explanation]
- [Claim]: [VERIFIED/UNVERIFIED/CONTRADICTED] + [DOMAIN:
      MATCHES/VIOLATES/N-A] - [explanation]
(For each major claim: first check domain knowledge, then
      check data sources)

**IMPORTANT**: Never verify claims against data for FUTURE
       tasks. Past time series CANNOT verify future
      predictions. Always check the requested date against
      the date at which the data actually ends.
Do NOT refer to past data calculation or verification to
      support the answer. Only past data vs. future data (
      expected values by YOUR prediction) can be used to
      support the answer.

OUTSTANDING CONFLICTS: [NO_CONFLICT / DETECTED / RESOLVED]
       - <details>
KEY EVIDENCE: <main observations/inferences used>
CALIBRATED ANSWER: [answer in exact task required format]
```

F.5.2 Synthesizer Instructions

Decision Protocol

```
DECISION PROTOCOL:

STEP 1 - CHECK APPROACH (Critical First):
- What did DOMAIN KNOWLEDGE say to do?
- What did reviewers actually do?
- If MISMATCH: Reviewers answered the WRONG question →
      derive your own answer

STEP 2 - SCORE REVIEWERS (0-100 each):
- Task Understanding (20): Did they follow suggested
      approach? (0 if mismatch!)
- Evidence Usage (20): Used evidence correctly?
- Verification (20): Verified claims before accepting?
- Conflict Handling (20): Detected and adjusted for
      conflicts?
- Calibration (20): Confidence matches evidence?

STEP 3 - IDENTIFY TASK TYPE:
- FUTURE task: Past data CANNOT verify predictions
  →Check for BLIND EXTRAPOLATION: Did they ask "will the
       CAUSE persist?"
  →Evaluate domain reasoning quality
- PAST-PRESENT task: Data CAN verify claims

STEP 4 - DERIVE ANSWER:
- UNANIMOUS + CORRECT APPROACH →use shared answer
- UNANIMOUS + WRONG APPROACH →REJECT, derive your own
- SPLIT + UNRESOLVED →CONSERVATIVE answer
- DOMAIN VIOLATION →reject that answer

CRITICAL RULES:
× Don't trust unanimous if wrong approach
× Don't pick majority without verification
× Don't accept answers that violate domain knowledge
✓Check approach FIRST
✓Always be skeptical of the reviewers' answers; do not
      blindly trust them
✓Be conservative when uncertain
```

Synthesizer Prompt

```
Task: {task_description_judge}

ANALYSTS: TEXT, VISUAL, NUMERICAL

Reviewer Evaluations:
{all_judge_responses}

---

VERIFICATION SOURCES (max 3 tool calls when available):
- NUMERICAL TOOLS: get_info, get_values, get_around,
      get_features, get_frequency_features,
      get_channel_values, get_all_channels, get_indicator
- CODE: execute_code(code)
- CHARTS: time series, frequency (attached)
- TEXT: embedded in task description
- DOMAIN: knowledge in system prompt

OUTPUT FORMAT (keep everything concise):
TASK: <Restate what the task is asking in one sentence>
TASK TYPE: [FUTURE / PAST-PRESENT]

APPROACH CHECK:
- SUGGESTED: <from domain knowledge>
- USED: <by reviewers>
- Status: <always re-check even with UNANIMOUS answers> [
      CORRECT / MISMATCH]

REVIEWER SCORES:
- Reviewer 0: (Task: _/20, Evidence: _/20, Verification: _
      /20, Conflicts: _/20, Calibration: _/20) = [0-100]
- Reviewer 1: (Task: _/20, Evidence: _/20, Verification: _
      /20, Conflicts: _/20, Calibration: _/20) = [0-100]
Note that perfect score is IMPOSSIBLE. Never give 100
      score to any reviewer. There are always flaws in the
      reviewers' reasoning.

ANSWER VERIFICATION (use TASK TYPE from above, summarizing
       only outstanding concerns):
- If PAST-PRESENT task: Verify answers against data
- If FUTURE task: Past data describes history but CANNOT
      verify predictions - evaluate domain reasoning

**IMPORTANT**: Never verify claims against data for FUTURE
       tasks. Past time series CANNOT verify future
      predictions. Always check the requested date against
      the date at which the data actually ends.
Do NOT refer to past data calculation or verification to
      support the answer. Only past data vs. future data (
      expected values by YOUR prediction) can be used to
      support the answer.

CONFLICT STATUS:
- Reviewer Agreement: [UNANIMOUS / SPLIT / ALL_DIFFERENT]
- Approach Status: [ALL_CORRECT / ALL_WRONG / MIXED]
- Analyst Agreement: [From reviewer reports - did analysts
       conflict?]
- Resolution: [VERIFIED_RESOLUTION / UNRESOLVED /
      NO_CONFLICT / APPROACH_ERROR]

CALIBRATED REASONING (apply based on TASK TYPE,
      summarizing in 1-2 short sentences):
- FUTURE task: <evaluate domain reasoning quality - cannot
       verify with past data; if reviewers said verified
      for FUTURE task, they are wrong!>
- PAST-PRESENT task: <data verification result>
- If MISMATCH: <derive own answer using correct approach>
- If SPLIT + cannot resolve: <Choose CONSERVATIVE answer>

FINAL ANSWER: <your calibrated answer in exact, user-
      defined task format>
```

G Use of Large Language Models

In preparing this submission, we used ChatGPT for language refinement, including improving clarity and correcting grammatical and typographical errors. Its use was limited to copyediting tasks

such as sentence restructuring and rephrasing for readability. All LLM-assisted edits were carefully reviewed and, where necessary, revised by the authors to ensure accuracy, originality, and full compliance with research-integrity standards. We note that large language models also serve as the backbone for *all* methods evaluated in our experiments, as described above.