MULTI-AXIS: EXPLAINING MULTIVARIATE TIME-SERIES ANOMALIES WITH DETECTOR-DERIVED CONTEXTUAL EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Time-series anomaly detection (TSAD) typically returns scores or labels, whereas practical analysis also requires explanations. Existing language-model interfaces combine raw observations, statistical or textual summaries, detector outputs, or continuous time-series representations, but generally do not connect "what was observed" at a channel–time location to "why it was judged anomalous." We propose Multi-AXIS, which constructs two complementary forms of evidence. Contextual Evidence fuses a frozen detector's position-wise contextual representations with anomaly-discrimination signals at the same locations. Observable Evidence combines a global visual reference with local target-window values. Their channel–time alignment lets the model check contextual judgments against concrete observations. A lightweight interface further constructs channel-level and question-relevant event-level representations and supplies both forms of evidence to a frozen vision-language model (VLM). We build approximately 70,000 multivariate time-series question–answer pairs. Experiments on synthetic and real-world data show that Multi-AXIS produces more accurate, complete, and evidence-based anomaly explanations while achieving competitive overall performance.

1 INTRODUCTION

Time-series anomaly detection (TSAD) supports industrial monitoring, healthcare, financial risk management, and cloud operations (Hundman et al., 2018; Iqbal et al., 2019; Su et al., 2019). Most detectors return pointwise scores or labels (Lai et al., 2021; Si et al., 2024; Zamanzadeh Darban et al., 2024), identifying suspicious locations without explaining which variables changed, how they changed, or why the changes are anomalous. Interpreting an event therefore requires comparing its observations with normal behavior, motivating natural-language explanations grounded in verifiable evidence.

Anomaly explanation must answer both "what happened?" and "why is it anomalous?" The former requires verifiable values and local patterns at specific channel–time locations; the latter assesses them against full-sequence temporal dynamics and cross-channel relationships (Zhang et al., 2019; Deng & Hooi, 2021; Lan et al., 2026). Because questions may concern one position, one channel, or an entire interval, reliable explanations need observable and contextual evidence across these scopes, with channel–time correspondence preserved for fine-grained judgments.

Language models can organize such evidence into readable explanations, but existing interfaces have distinct limitations. Direct-input methods preserve location-specific numerical sequences or plots, yet require the model to infer long-range temporal and cross-channel dependencies from potentially large inputs (Alnegheimish et al., 2024; Dong et al., 2024; Zhou & Yu, 2025). Statistics, textual summaries, and retrieved references are compact but may omit the deviation relevant to the current anomaly (Park et al., 2025; Liu et al., 2025). Detection-guided pipelines pass predictions or scored intervals to agents for checking or diagnosis, but scores and boundaries do not explain deviations from usual temporal or cross-channel patterns (Xu et al., 2026; Kang et al., 2026). Representation-based interfaces provide broader context through learned embeddings, although aggregation can weaken channel–time correspondence and general-purpose encoders may lack explicit anomaly-

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

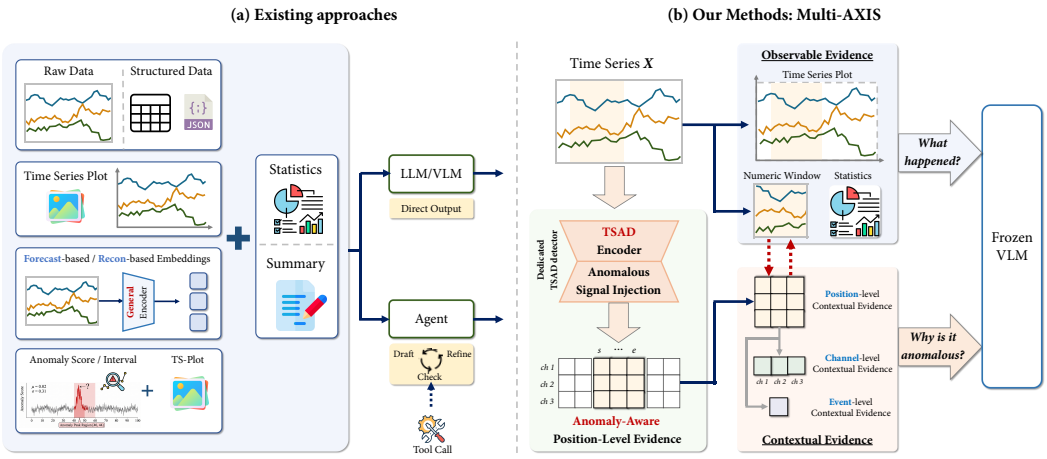


Figure 1: Comparison of evidence interfaces. Multi-AXIS aligns detector-derived position-level Contextual Evidence with numerical observations at matching channel–time locations and constructs channel- and event-level summaries for broader reasoning.

discrimination information (Xie et al., 2025; Yu et al., 2026). Agent-based tool use and iterative numerical or visual inspection support adaptive analysis, but generally do not expose a specialized detector's internal contextual representations and discrimination signals at matching locations (Xu et al., 2026; Kang et al., 2026; Zumarraga et al., 2026).

Across these interfaces, language models still lack an explicit link between observable channel–time values and the context supporting a specialized detector's judgments. Position-wise detector representations provide such a link by retaining location correspondence and the context used for scoring (Xu et al., 2022; Yang et al., 2023; Lan et al., 2026). We therefore propose Multi-AXIS (Figure 1), which combines detector-derived Contextual Evidence with Observable Evidence at position, channel, and event levels. Position-level evidence fuses a frozen detector's full-sequence representations with discrimination signals at matching target locations. Multi-AXIS derives channel summaries and a question-conditioned event summary for broader scopes. Its Observable Evidence pairs a full-sequence plot with target-window values aligned to the position-level evidence, tying contextual judgments to concrete observations. A frozen VLM receives both forms of evidence and the question to generate an explanation. We also construct the Multivariate Semantic Anomaly Benchmark (MSAB) for training and evaluation. Our contributions are:

- **Framework contribution.** We propose Multi-AXIS, whose trainable Evidence Interface transfers detector-derived anomaly context to a VLM and connects it with verifiable observations while keeping the detector and VLM frozen.
- **Evidence construction.** We construct position-, channel-, and question-conditioned event-level Contextual Evidence from detector representations and discrimination signals, aligning position-level evidence with observable values at matching locations.
- **Data contribution.** We introduce MSAB and three real-world transfer sets, totaling 68,865 question–answer pairs over 34,347 target windows. They cover anomalous and normal intervals through multiple-choice, true/false, and open-ended questions on localization, temporal patterns, and cross-channel relationships.
- **Empirical contribution.** Experiments on MSAB and the real-world transfer sets show stronger overall explanation quality than specialized time-series language and vision-language baselines across most settings.

2 Related Work

Context Modeling in TSAD. TSAD methods model temporal or cross-channel context through reconstruction and probabilistic objectives (Su et al., 2019; Audibert et al., 2020; Tuli et al., 2022),

graphs, association discrepancies, contrastive learning, and memory (Deng & Hooi, 2021; Xu et al., 2022; Yang et al., 2023; Song et al., 2023). Their contextual representations support anomaly scoring and can therefore serve as explanation evidence. TimeRCD's Relative Context Discrepancy pretraining explicitly supervises anomaly judgments relative to surrounding context (Lan et al., 2026). Multi-AXIS freezes TimeRCD and aligns its contextual representations and discrimination signals with observable channel–time locations.

General Time-Series Representations for Language Models. Foundation models learn transferable numerical representations through quantization, patch-based modeling, masked pretraining, or mixture-of-experts architectures (Ansari et al., 2024; Das et al., 2024; Goswami et al., 2024; Shi et al., 2025). ChatTS encodes sequences as continuous tokens and uses synthetic question–answer data to align time-series and language spaces (Xie et al., 2025). TS-Reasoner maps representations from a frozen time-series foundation model to an LLM for multi-step reasoning (Yu et al., 2026). TimeOmni-1 and TimeOmni-VL extend time-series understanding, generation, and decision-making through staged reasoning training and a reversible time-series–vision interface, respectively (Guan et al., 2026a;b). These representations support language reasoning but are learned for general tasks rather than the discriminative process of a particular anomaly detector.

Evidence Interfaces for Language-Based Anomaly Reasoning. Language-model-based TSAD methods use direct observations, diagnostic cues, or interactive inspection. SigLLM, AnomLLM, and TAMA provide numerical or visual observations, leaving the model to infer the context needed for anomaly judgments (Alnegheimish et al., 2024; Zhou & Yu, 2025; Zhuang et al., 2024). Others provide statistical or positional cues, retrieved segments, or analyzer reports (Park et al., 2025; Liu et al., 2025; Kang et al., 2026). TSAD-Agents coordinates agents that select detection tools and modalities (Xu et al., 2026), while TimeRLM iteratively inspects data through code and vision (Zumarraga et al., 2026). Although these approaches enrich the evidence, they do not explicitly transfer the internal context supporting a specialized detector's scores; the language model must infer that basis from observations and diagnostics. Multi-AXIS instead exposes contextual representations and discrimination signals aligned with observable values at matching locations.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Anomaly explanation requires identifying what happens within a target region and assessing it against broader temporal and cross-channel regularities. We therefore formulate the task as question answering about a given target interval, using the full sequence as context. The interval defines the scope of explanation and accommodates both individual points and multi-step events, while the full sequence provides contextual references for interpreting the observed behavior. Because the target interval may be anomalous or normal, the task requires assessing, rather than presupposing, its abnormality.

Formally, let $\mathbf{X} = [x_{t,c}] \in \mathbb{R}^{T\times C}$ denote a multivariate time series with T valid time steps and C channels. The given target interval $[s, e)$ has length $L = e - s$, with local index $j \in \{0, \dots, L-1\}$ corresponding to global time $t = s + j$. A point target is the special case $L = 1$. Given a natural-language question q about this interval, the task is to learn the mapping $\mathcal{G} : (\mathbf{X}, q, s, e) \mapsto \mathbf{y}$ where $\mathbf{y} = (y_1, \dots, y_{N_y})$ is a natural-language answer grounded in the target observations and their full-sequence context. Depending on the question, the answer may assess whether the target behavior is anomalous, describe its temporal or cross-channel characteristics, or explain its deviation from contextual regularities.

As shared preprocessing for Observable Evidence and Contextual Evidence, we standardize each channel as $\tilde{x}_{t,c} = \frac{x_{t,c}-\mu_c}{\sigma_c+\epsilon}$, where μ_c and σ_c are computed over the full valid sequence of channel c, and ϵ is a small constant for numerical stability. We denote the standardized sequence by $\widetilde{\mathbf{X}}$.

3.2 METHOD OVERVIEW

Figure 2 illustrates the overall pipeline of Multi-AXIS. Its core design provides the language model with two complementary types of information. Observable Evidence preserves the visual refer-

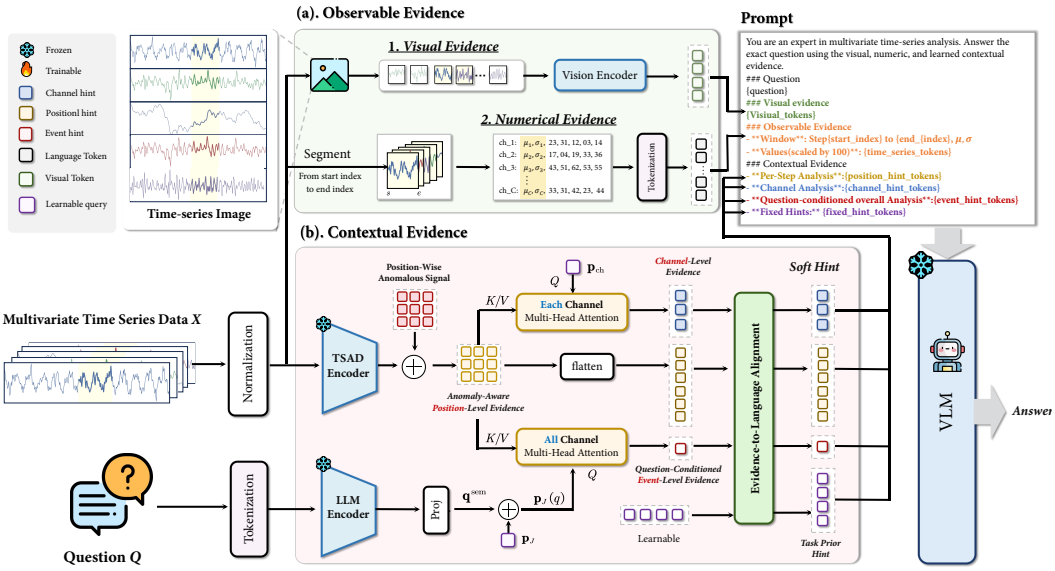


Figure 2: Overview of Multi-AXIS framework.

ence of the full sequence, the channel scales, and the target-window values. Contextual Evidence extracts contextualized, anomaly-discrimination-relevant representations from a frozen anomaly detector and aligns them with numerical observations at the same positions. The Evidence Interface further organizes the latter into Position, Channel, and question-conditioned Event Hints, and maps them into the embedding space of a frozen VLM. Finally, the two types of evidence, the question text, and task priors jointly drive natural-language generation. We instantiate the detector and VLM with TimeRCD (Lan et al., 2026) and Qwen3-VL (Bai et al., 2025), respectively, and update only the Evidence Interface during training.

3.3 Observable Evidence

Observable Evidence combines a full-sequence plot with target-window values and channel-scale information. For **visual evidence**, we plot each channel of $\widetilde{\mathbf{X}}$ separately with a shared time axis and highlight the target interval using a consistent color that does not encode anomaly labels. For **numerical evidence**, we provide target-window values grouped by channel and ordered by time, aligned with position-level Contextual Evidence at matching channel–time locations. Each value is encoded as $\mathrm{round}(\kappa \widetilde{x}_{t,c})$, with $\kappa = 100$ by default, yielding a compact integer representation. The header of each channel also provides μ_c and σ_c (e.g., `ch_1 [mean=1, std=2]: 123, 124, 128`), enabling the model to approximately recover the original scale. Together, the plot provides global context, while the numerical values support verification of local changes.

3.4 Detector-Derived Contextual Evidence

The same local pattern may be normal or anomalous depending on its temporal and cross-channel context. We construct position-level Contextual Evidence by fusing full-sequence detector representations with anomaly-discrimination signals at matching locations. We then aggregate this evidence into channel-level and question-conditioned event-level summaries.

Full-Series Multivariate Encoding. Given the standardized sequence $\widetilde{\mathbf{X}}$, frozen TimeRCD splits each channel into non-overlapping patches of length W, producing CP tokens, where $P = \lceil T/W \rceil$. A non-causal Transformer jointly encodes these tokens and restores per-time-step representations $\mathbf{H}^{\mathrm{step}} \in \mathbb{R}^{T \times C \times d_p}$, with $d_p = 256$ in our implementation. Each $\mathbf{H}^{\mathrm{step}}_{t,c}$ corresponds to a specific channel–time location while incorporating temporal and cross-channel context from the full sequence.

Anomaly-Aware Position-Level Evidence. Although $\mathbf{H}^{\text{step}}_{t,c}$ is shaped by the anomaly-detection objective, its anomaly tendency remains implicit. To explicitly incorporate the detector's anomaly-discrimination signal, we compute the difference between its anomalous and normal classification logits, denoted by $\ell^A_{t,c}$ and $\ell^N_{t,c}$ and get $r_{t,c} = \tanh\left(\frac{\ell^A_{t,c} - \ell^N_{t,c}}{2}\right)$. The positive and negative values indicate the detector's preference for anomalous and normal behavior, respectively. We fuse this signal with the contextual representation through a shared learnable direction $\mathbf{v}_{\text{ano}} \in \mathbb{R}^{d_p}$:

$$\widehat{\mathbf{G}}^{\text{step}}_{c,j} = \text{RMS}\left[\text{RMS}\left(\mathbf{H}^{\text{step}}_{s+j,c}\right) + r_{s+j,c}\mathbf{v}_{\text{ano}}\right]. \tag{1}$$

Here, $\mathbf{v}_{\text{ano}}$ is initialized to zero, and RMS normalization is defined as $\text{RMS}(\mathbf{z}) = \frac{\mathbf{z}}{\sqrt{\frac{1}{d}\sum_{k=1}^{d} z_k^2 + \epsilon_h}}$, where d is the dimension of $\mathbf{z}$ and ϵ_h is a numerical stability constant. The resulting position-level evidence combines contextual information with the detector's discrimination signal while retaining its correspondence with the observation at channel c and global time $s + j$.

Channel-Level Evidence. To summarize each channel over the target interval, we collect its position-level evidence into $\mathbf{R}^{\text{ch}}_c = \left[\widehat{\mathbf{G}}^{\text{step}}_{c,0}, \ldots, \widehat{\mathbf{G}}^{\text{step}}_{c,L-1}\right] \in \mathbb{R}^{L \times d_p}$. A learnable query $\mathbf{p}_{\text{ch}} \in \mathbb{R}^{d_p}$, shared across channels, pools this sequence through multi-head attention:

$$\mathbf{G}^{\text{ch}}_c = \text{MHA}\left(\mathbf{p}_{\text{ch}}, \mathbf{R}^{\text{ch}}_c, \mathbf{R}^{\text{ch}}_c\right) \in \mathbb{R}^{d_p}. \tag{2}$$

Here, $\text{MHA}(\mathbf{Q}, \mathbf{K}, \mathbf{V})$ denotes multi-head attention with queries $\mathbf{Q}$, keys $\mathbf{K}$, and values $\mathbf{V}$. The resulting representation provides a question-independent summary of the channel's behavior over the target interval.

Question-Conditioned Event-Level Evidence. Channel Hints summarize channels separately, but event-level reasoning requires jointly considering all channels over the target interval. We therefore collect all CL position-level representations into $\mathbf{R}^{\text{joint}} = \left[\widehat{\mathbf{G}}^{\text{step}}_{0,0}, \ldots, \widehat{\mathbf{G}}^{\text{step}}_{C-1,L-1}\right] \in \mathbb{R}^{CL \times d_p}$. To prioritize question-relevant information before aggregation, we encode the question q with the frozen VLM and take the hidden state of its last valid token $\mathbf{h}^q \in \mathbb{R}^{d_\ell}$ to extract the question semantics, where d_ℓ is its hidden dimension. A learnable projection $A_q : \mathbb{R}^{d_\ell} \to \mathbb{R}^{d_p}$ produces $\mathbf{q}^{\text{sem}} = A_q\left(\text{RMS}(\mathbf{h}^q)\right)$. We combine this question representation with a learnable event query $\mathbf{p}_J \in \mathbb{R}^{d_p}$: $\mathbf{p}_J(q) = \mathbf{p}_J + \mathbf{q}^{\text{sem}}$. The question-conditioned query then pools evidence across all channels and target-window positions:

$$\mathbf{G}^{\text{joint}}(q) = \text{MHA}\left(\mathbf{p}_J(q), \mathbf{R}^{\text{joint}}, \mathbf{R}^{\text{joint}}\right) \in \mathbb{R}^{d_p}. \tag{3}$$

This produces an event-level summary tailored to the current question. The position-, channel-, and event-level representations are subsequently mapped into the VLM's embedding space through a shared alignment module.

3.5 Evidence-to-Language Alignment

We use a shared alignment module to convert the three levels of Contextual Evidence into soft tokens in the frozen VLM's embedding space. The module maps detector-derived representations through cross-attention over a learnable prototype bank constructed from the VLM's word embeddings.

Let $\mathbf{E}_{\text{vocab}} \in \mathbb{R}^{|\mathcal{V}| \times d_\ell}$ denote the frozen word-embedding matrix, where $|\mathcal{V}|$ is the vocabulary size. A learnable matrix $\mathbf{M} \in \mathbb{R}^{N_{\text{proto}} \times |\mathcal{V}|}$ constructs a bank of N_{proto} prototypes $\mathbf{S}_{\text{proto}} = \mathbf{M}\mathbf{E}_{\text{vocab}} \in \mathbb{R}^{N_{\text{proto}} \times d_\ell}$. Each prototype is a learned linear combination of the frozen vocabulary embeddings.

For a contextual representation $\mathbf{Z}$, a learnable projection $P_{\text{ts}} : \mathbb{R}^{d_p} \to \mathbb{R}^{d_\ell}$ first maps it to the VLM dimension. We then apply cross-attention with the projected representation as the query and the prototype bank as keys and values:

$$\text{Align}(\mathbf{Z}) = \text{MHA}\left(P_{\text{ts}}(\mathbf{Z}), \mathbf{S}_{\text{proto}}, \mathbf{S}_{\text{proto}}\right). \tag{4}$$

The projection, prototype bank, and alignment attention parameters are shared across the position, channel, and event levels.

Applying this module produces the corresponding Position, Channel, and Event Hints:

$$\widetilde{\mathbf{H}}^{\text{step}}_{c,j} = \text{Align}\left(\widehat{\mathbf{G}}^{\text{step}}_{c,j}\right), \qquad \widetilde{\mathbf{H}}^{\text{ch}}_{c} = \text{Align}\left(\text{RMS}\left(\mathbf{G}^{\text{ch}}_{c}\right)\right),$$
$$\widetilde{\mathbf{H}}^{\text{joint}}(q) = \text{Align}\left(\text{RMS}\left(\mathbf{G}^{\text{joint}}(q)\right)\right).$$

Each hint is a d_ℓ-dimensional soft token that can be inserted directly into the VLM's input embedding sequence. Each Position Hint accompanies the numerical value at the same channel–time location in Observable Evidence, preserving the correspondence between contextual evidence and concrete observations. Channel Hints provide per-channel summaries, while the Event Hint supplies a question-conditioned summary of the multivariate target interval.

3.6 Prompt Assembly and Training Objective

Task-Prior Hints. To provide shared task-level information, we introduce learnable queries $\mathbf{P}_{\text{fix}} \in \mathbb{R}^{K_F \times d_\ell}$, where K_F is the number of task-prior tokens. These queries attend to the prototype bank to produce $\widetilde{\mathbf{F}} = \text{MHA}(\mathbf{P}_{\text{fix}}, \mathbf{S}_{\text{proto}}, \mathbf{S}_{\text{proto}}) \in \mathbb{R}^{K_F \times d_\ell}$. The resulting Task-Prior Hints are shared across instances and do not depend on the current sequence or question.

Multimodal Input Assembly and Training Objective. We combine task instructions, the question, observable evidence, position-, channel-, and event-level hints, and shared task-prior hints $\widetilde{\mathbf{F}}$ into $\mathcal{P}_\psi(\mathbf{X}, q, s, e)$. The plot is processed by the VLM's visual encoder, while the others are tokenized as text. The learned hints are inserted directly into the VLM's input embedding sequence as soft tokens. ψ denotes all trainable interface parameters. Given the reference answer $\mathbf{y} = (y_1, \ldots, y_{N_y})$, we optimize the autoregressive negative log-likelihood only over answer tokens:

$$\mathcal{L}_{\text{QA}}(\psi) = -\sum_{n=1}^{N_y} \log p_\phi(y_n \mid y_{<n}, \mathcal{P}_\psi(\mathbf{X}, q, s, e)), \tag{5}$$

where ϕ denotes the parameters of the frozen VLM. Details of the training procedure, hyperparameters, and complete prompt are provided in Appendix A.2 and Appendix F.

3.7 Multivariate Semantic Anomaly Benchmark (MSAB) Construction

Sequences and ground truth. We adopt TimeRCD's context-dependent multivariate generator (Lan et al., 2026) to construct sequences with diverse temporal patterns and cross-channel dependencies. Anomalies from 20 templates include both propagating and non-propagating interventions. Generation metadata and differences between observed sequences and their anomaly-free counterfactuals provide traceable channel–time ground truth. We sample both anomalous and normal target intervals, prioritizing high-variation normal windows to distinguish salient changes from contextually anomalous behavior.

Task coverage and QA construction. MSAB covers anomaly judgments, candidate sources and affected scope, temporal patterns, and cross-channel relationships. Questions use multiple-choice, true/false, and open-ended formats and are organized from anomaly-analysis and normality-verification perspectives. Both perspectives apply to anomalous and normal intervals, reducing label shortcuts in question wording. GPT-5.5 generates questions under predefined semantic specifications, and GPT-5.4 produces reference answers using observed evidence, structured ground truth, and normal counterfactuals, while reference answers must justify their conclusions using observable evidence.

Quality control and data splits. Structural and semantic checks assess answerability, label leakage, and consistency between answers and evidence, with targeted regeneration of invalid instances. MSAB contains 68,281 QA pairs over 34,055 target windows. We split data by source sequence, keeping all derived windows and QA records in the same partition. Question types, perspectives, and anomaly status are approximately balanced in training and validation. Appendix C provides detailed generation procedures, semantic specifications, information boundaries, and dataset statistics.

4 EXPERIMENTS

To evaluate the effectiveness of Multi-AXIS, we conduct a series of experiments that address two research questions: **RQ1: Explanation Quality.** How does Multi-AXIS compare with state-of-the-art LLM/VLM-based methods in generating high-quality semantic explanations of anomalies? **RQ2: Mechanism Effectiveness.** How much do verifiable observations, detector-derived position-preserving contextual anomaly evidence, and channel-level and question-conditioned event-level aggregation each contribute to explanation quality?

4.1 EXPERIMENTAL SETUP

Dataset. We evaluate semantic time-series anomaly explanation in controlled and real-world settings. **(1) Synthetic data:** MSAB uses programmatically generated, diverse anomaly patterns for precise ground-truth evaluation. **(2) Real-world data:** Three transfer test sets assess robustness across domains: the server-monitoring dataset SMD (Su et al., 2019), the water-treatment dataset SWaT (Mathur & Tippenhauer, 2016), and VTA Articulary, constructed from multivariate UEA ArticularyWordRecognition sequences (Bagnall et al., 2018; Xu et al., 2026). MSAB contains 68,281 question–answer pairs over 34,055 target windows; including the real-world sets yields 68,865 pairs over 34,347 windows. Appendix C details their construction and statistics.

Baselines. We compare against two groups of strong baselines: LLM/VLM-based anomaly reasoning methods, including AnomLLM (Full/Window) and TSAD-Agents (Zhou & Yu, 2025; Xu et al., 2026), and general time-series language and reasoning models, including ChatTS, Time-MQA, TS-Reasoner, TimeOmni-1, TimeOmni-VL, and TimeRLM (Xie et al., 2025; Kong et al., 2025; Yu et al., 2026; Guan et al., 2026a;b; Zumarraga et al., 2026). AnomLLM (Full) takes the complete sequence, whereas AnomLLM (Window) takes only the target window. Appendix B provides the complete input interfaces and task adaptations.

Evaluation Metrics. Following recent work on evaluating LLM-generated content, we adopt the LLM-as-a-judge approach. Specifically, we use G-Eval (Liu et al., 2023) with GPT-5.4 as a primary LLM evaluator. For different question types, we evaluate explanation quality along multiple dimensions, including correctness (Corr.), reasoning quality (Rsn. Qual.), accuracy (Acc.), completeness (Comp.), relevance (Rel.), and justification quality (Justif.). We aggregate the applicable dimensions into an overall score (Final). For multiple-choice (MC) and true/false (TF) questions, we additionally report answer accuracy (**Acc.**) as an objective evaluation that does not rely on an LLM judge. Appendix D provides the metric definitions, aggregation procedure, and judge prompts. In addition, **we invited six human experts with relevant research backgrounds to independently score the anonymized model responses**; the sampling method, evaluation criteria, annotation consistency, and detailed results are provided in Appendix E.1.

4.2 MAIN RESULTS: EXPLANATION QUALITY (RQ1)

Table 1 compares Multi-AXIS with all baselines on multiple-choice, open-ended, and true/false questions, on the MSAB synthetic test set and three real-world test sets.

Overall explanation quality. Multi-AXIS achieves the highest Final score in 9 of the 12 dataset–question-type combinations. It ranks first on all three question types on the MSAB synthetic test set and VTA Articulary, and first on multiple-choice and open-ended questions on SMD and on multiple-choice questions on SWaT. Compared with the strongest baseline in each setting, Multi-AXIS improves the Final score by 0.56, 0.67, and 0.50 on multiple-choice, open-ended, and true/false questions on the synthetic test set, and by 0.23, 0.40, and 0.17 on VTA Articulary. The real-world results are more mixed: Multi-AXIS leads by 0.12 and 0.17 on multiple-choice and open-ended questions on SMD, but trails TS-Reasoner by 0.09 on true/false questions; on SWaT, it leads by 0.03 on multiple-choice, while trailing TimeRLM by 0.16 and 0.31 on open-ended and true/false questions. These results show that the advantage of Multi-AXIS transfers from synthetic data to real-world sequences, while different real-world datasets and question types place different demands on the evidence interface.

Table 1: Main Results on Synthetic vs. Real Data. We report the accuracy on Multiple-Choice (MC) and True/False (TF) questions, together with the fine-grained scores on the three question types. The top-three distinct results are highlighted as first , second , and third ; ties share the same rank.

DataSet	Method	MC Accuracy	TF Accuracy	Multiple Choice			Open Ended				True False		
				Final	Corr.	Rsn.Qual.	Final	Acc.	Comp.	Rel.	Final	Corr.	Justif.
MSAB (Synthetic Data)	ChatTS(15B)	0.38	0.57	2.59	2.57	2.64	2.37	2.19	2.17	2.82	2.64	2.72	2.52
	AnomLLM (Full) (8B)	0.33	0.54	2.45	2.45	2.44	2.10	1.97	1.91	2.47	2.41	2.47	2.32
	AnomLLM (Window) (8B)	0.38	0.53	2.60	2.65	2.49	2.18	2.07	1.93	2.60	2.50	2.60	2.35
	TSAD-Agents (undisclosed)	0.36	0.52	2.66	2.65	2.67	2.16	1.93	2.01	2.60	2.58	2.57	2.60
	TS-Reasoner(8B)	0.26	0.47	2.14	2.06	2.11	2.28	1.97	2.16	3.67	2.32	2.71	2.16
	Time-MQA(8B)	0.35	0.55	2.48	2.52	2.36	2.21	2.23	1.91	2.55	2.40	2.53	2.21
	TimeOmni-VL(15B)	0.31	0.45	2.16	2.15	1.99	1.78	1.46	2.12	3.25	2.03	2.39	2.01
	TimeOmni-1(8B)	0.24	0.50	2.02	1.95	2.16	1.78	1.66	1.98	2.81	2.29	2.77	2.20
	TimeRLM(8B)	0.34	0.50	2.66	2.70	2.55	2.40	2.17	2.25	2.86	2.46	2.45	2.49
	Multi-AXIS(Ours, 8B)	0.65	0.63	3.22	3.30	3.06	3.07	2.88	3.00	3.38	3.14	3.09	3.21
VTA Articulary (Real-world Data)	ChatTS(15B)	0.34	0.50	2.63	2.66	2.57	2.32	2.15	2.07	2.79	2.48	2.48	2.47
	AnomLLM (Full) (8B)	0.37	0.65	2.45	2.54	2.10	1.81	1.70	1.73	2.67	2.61	3.33	2.18
	AnomLLM (Window) (8B)	0.37	0.64	2.57	2.60	2.37	2.12	2.19	1.79	2.90	2.26	2.76	2.02
	TSAD-Agents (undisclosed)	0.46	0.42	2.92	2.99	2.75	2.91	2.76	2.72	3.30	2.71	2.65	2.79
	TS-Reasoner(8B)	0.31	0.47	2.36	2.36	2.30	2.43	2.16	2.18	3.70	2.30	2.52	2.14
	Time-MQA(8B)	0.34	0.67	2.59	2.61	2.54	2.62	2.40	2.48	3.03	2.85	2.97	2.68
	TimeOmni-VL(15B)	0.34	0.52	2.43	2.30	2.27	2.19	2.10	2.33	3.54	2.53	2.77	2.41
	TimeOmni-1(8B)	0.40	0.56	2.61	2.24	2.87	2.76	2.21	3.24	4.19	3.09	3.17	3.24
	TimeRLM(8B)	0.30	0.61	2.52	2.52	2.51	2.65	2.49	2.42	3.12	2.87	2.88	2.85
	Multi-AXIS(Ours, 8B)	0.43	0.73	3.15	3.25	2.92	3.31	3.25	3.00	3.75	3.26	3.23	3.30
SMD (Real-world Data)	ChatTS(15B)	0.44	0.66	2.80	2.83	2.71	2.31	2.13	2.07	2.78	2.78	2.82	2.72
	AnomLLM (Full) (8B)	0.35	0.51	2.71	2.70	2.74	2.58	2.42	2.39	3.00	2.78	2.82	2.72
	AnomLLM (Window) (8B)	0.44	0.43	3.00	3.03	2.92	2.57	2.52	2.33	2.90	2.64	2.69	2.58
	TSAD-Agents (undisclosed)	0.56	0.54	3.04	3.16	2.84	2.51	2.37	2.30	2.93	2.79	2.90	2.64
	TS-Reasoner(8B)	0.35	0.64	2.61	2.45	2.33	2.33	2.12	2.13	3.61	3.06	3.52	2.61
	Time-MQA(8B)	0.53	0.46	3.04	3.09	2.92	2.44	2.30	2.25	2.82	2.53	2.58	2.46
	TimeOmni-VL(15B)	0.26	0.52	2.11	1.98	2.06	2.06	1.64	2.21	3.33	2.57	2.84	2.37
	TimeOmni-1(8B)	0.39	0.54	2.62	2.59	2.47	2.34	2.16	2.21	3.33	2.55	2.81	2.57
	TimeRLM(8B)	0.38	0.64	2.75	2.83	2.56	2.58	2.36	2.42	3.03	2.82	2.84	2.81
	Multi-AXIS(Ours, 8B)	0.59	0.57	3.16	3.23	3.02	2.75	2.52	2.69	3.09	2.97	2.93	3.03
SWaT (Real-world Data)	ChatTS(15B)	0.11	0.46	1.98	1.85	2.26	1.68	1.35	1.53	2.24	2.24	2.21	2.28
	AnomLLM (Full) (8B)	0.11	0.51	1.76	1.57	2.18	2.07	1.85	1.84	2.60	2.52	2.56	2.42
	AnomLLM (Window) (8B)	0.07	0.40	1.45	1.21	2.02	1.80	1.52	1.65	2.32	2.34	2.41	2.25
	TSAD-Agents (undisclosed)	0.11	0.46	1.81	1.67	2.15	2.00	1.74	1.79	2.55	2.35	2.34	2.36
	Time-MQA(8B)	0.46	0.57	2.63	2.71	2.51	2.22	1.96	2.13	2.65	2.60	2.56	2.52
	TS-Reasoner(8B)	0.20	0.56	1.90	1.75	1.75	2.21	1.76	2.15	3.63	2.52	2.82	2.10
	TimeOmni-VL(15B)	0.18	0.72	1.75	1.69	1.70	1.87	1.52	2.40	3.27	2.59	3.11	2.31
	TimeOmni-1(8B)	0.20	0.49	1.74	1.77	1.85	1.53	1.35	1.87	2.94	1.92	2.30	1.98
	TimeRLM(8B)	0.34	0.66	2.75	2.87	2.48	2.38	2.07	2.24	2.92	2.83	2.82	2.84
	Multi-AXIS(Ours, 8B)	0.48	0.74	2.78	2.84	2.64	2.22	2.10	2.16	2.42	2.52	2.57	2.44

Objective answer accuracy. Among the 8 accuracy settings that do not rely on an LLM judge, Multi-AXIS ranks first in 6, including both closed-form question types on the synthetic test set, true/false questions on VTA Articulary, multiple-choice questions on SMD, and both closed-form question types on SWaT; TSAD-Agents and ChatTS lead on VTA Articulary multiple-choice and SMD true/false questions, respectively. Notably, Multi-AXIS ranks first in both multiple-choice and true/false accuracy on SWaT, yet its G-Eval Final scores on open-ended and true/false questions are lower than those of TimeRLM. This indicates that answer correctness and the completeness and argumentation quality of explanations measure complementary capabilities, and that objective accuracy and fine-grained explanation scores should therefore be reported together.

4.3 Ablation Study (RQ2)

RQ1 verifies the overall effectiveness of Multi-AXIS. To further identify the contribution of each design, we conduct ablations on the synthetic test set along four dimensions: evidence source, composition of observable evidence, detector-derived contextual evidence, and aggregation strategy. Results are shown in Table 2.

TimeRCD Score channel-averages anomaly logits into one score per timestamp; Position-wise Anomalous Signal retains one score per channel–time position. Global Pooling mean-pools the enhanced window states into one token; and w/o Anomaly Enhancement omits logit injection from the otherwise full interface.

The two types of evidence are complementary. Keeping only Observable Evidence reduces the Final scores on multiple-choice, open-ended, and true/false questions by 0.48, 0.03, and 0.21, re-

Table 2: Evidence-interface ablations. Subscripts denote absolute changes relative to the full model (red ↓: decrease; green ↑: increase).

Variant	Multiple Choice			Open Ended				True False		
	Final	Corr.	Rsn.Qual.	Final	Acc.	Comp.	Rel.	Final	Corr.	Justif.
Multi-AXIS	**3.22**	**3.30**	**3.06**	**3.07**	**2.88**	**3.00**	**3.38**	**3.14**	**3.09**	**3.21**
Observable only	2.74 ↓0.48	2.81 ↓0.49	2.61 ↓0.45	3.04 ↓0.03	2.86 ↓0.02	2.93 ↓0.07	3.38 ±0.00	2.93 ↓0.21	2.91 ↓0.18	2.97 ↓0.24
Contextual only	2.66 ↓0.56	2.70 ↓0.60	2.59 ↓0.47	2.75 ↓0.32	2.59 ↓0.29	2.62 ↓0.38	3.08 ↓0.30	2.89 ↓0.25	2.87 ↓0.22	2.92 ↓0.29
TimeRCD Score	2.91 ↓0.31	2.98 ↓0.32	2.74 ↓0.32	3.06 ↓0.01	2.89 ↑0.01	2.95 ↓0.05	3.39 ↑0.01	3.05 ↓0.09	3.08 ↓0.01	3.00 ↓0.21
Position-wise Anomalous Signal	3.06 ↓0.16	3.15 ↓0.15	2.86 ↓0.20	3.09 ↑0.02	2.94 ↑0.06	2.98 ↓0.02	3.40 ↑0.02	3.14 ±0.00	3.20 ↑0.11	3.06 ↓0.15
Global Pooling	3.11 ↓0.11	3.13 ↓0.17	3.07 ↑0.01	2.96 ↓0.11	2.76 ↓0.12	2.90 ↓0.10	3.28 ↓0.10	2.99 ↓0.15	2.94 ↓0.15	3.06 ↓0.15
w/o Anomaly Enhancement	3.15 ↓0.07	3.21 ↓0.09	3.00 ↓0.06	3.06 ↓0.01	2.86 ↓0.02	2.98 ↓0.02	3.37 ↓0.01	3.12 ↓0.02	3.06 ↓0.03	3.20 ↓0.01
w/o TS Plot	2.91 ↓0.31	2.96 ↓0.34	2.83 ↓0.23	2.78 ↓0.29	2.59 ↓0.29	2.73 ↓0.27	3.07 ↓0.31	2.96 ↓0.18	2.89 ↓0.20	3.06 ↓0.15
w/o Numeric	2.89 ↓0.33	2.96 ↓0.34	2.78 ↓0.28	2.76 ↓0.31	2.59 ↓0.29	2.63 ↓0.37	3.10 ↓0.28	3.01 ↓0.13	3.00 ↓0.09	3.02 ↓0.19
w/o Question Conditioning	3.02 ↓0.20	3.12 ↓0.18	2.82 ↓0.24	2.99 ↓0.08	2.79 ↓0.09	2.87 ↓0.13	3.37 ↓0.01	3.09 ↓0.05	3.03 ↓0.06	3.19 ↓0.02

spectively; keeping only Contextual Evidence reduces them by 0.56, 0.32, and 0.25. The former still supports open-ended descriptions of visible patterns but degrades markedly on multiple-choice and true/false questions that require explicit discrimination; the latter degrades on all question types, indicating that detector-provided contextual judgments cannot replace verifiable observations. Only when both types of evidence are used together can the model support both factual description and anomaly discrimination.

The two components of observable evidence play distinct roles. Removing the full time-series plot or the target-window values reduces the Final scores on multiple-choice, open-ended, and true/-false questions by 0.31/0.29/0.18 and 0.33/0.31/0.13, respectively. The full time-series plot provides a global morphological reference, while the window values support verification of local positions and magnitudes.

Contextual evidence must preserve both positional and representational semantics. Averaging per-channel TimeRCD logits into a per-timestep score reduces the multiple-choice and true/false Final scores by 0.31 and 0.09. Position-wise Anomalous Signals narrow the multiple-choice drop to 0.16, improve open-ended questions by 0.02, and match the full model on true/false questions, but still lag by 0.20 in multiple-choice reasoning quality and 0.15 in true/false argumentation quality. Thus, addressable scalars preserve "where is suspicious" but not the full temporal and cross-channel context of position-wise representations. Removing anomaly-signal fusion further reduces the three Final scores by 0.07, 0.01, and 0.02, confirming a small but consistent gain from explicit discrimination signals.

Multi-level aggregation outperforms single global compression. Mean-pooling the enhanced target window into one global token reduces the multiple-choice, open-ended, and true/false Final scores by 0.11, 0.11, and 0.15, showing that global compression loses channel and positional structure. Removing question conditioning before event-level aggregation reduces them by 0.20, 0.08, and 0.05, with the largest effect on multiple-choice questions; injecting question semantics before compression therefore helps prioritize relevant channel and temporal evidence.

5 Conclusion

We proposed Multi-AXIS, an evidence interface that converts a frozen anomaly detector's position-wise contextual representations into evidence aligned with verifiable observations. On synthetic and real-world datasets, it generates more accurate, complete, and evidence-grounded explanations and consistently outperforms strong specialized time-series language and vision-language baselines. Future work will incorporate domain-specific knowledge graphs for causal reasoning and generate multimodal explanations combining text and visualizations.

AI use statement

We used large language models to assist with language polishing during manuscript preparation, including improvements to grammar, readability, and academic style. The authors reviewed and

are responsible for all technical content, experimental design, result verification, and claims in the paper.

Large language models are also part of our experiments. We use GPT-5.5 to generate questions and GPT-5.4 to generate teacher answers. Several comparison methods are themselves based on language models, and G-Eval also uses a language model as the judge. Appendix C describes the data construction process, Appendix D presents the evaluation protocol, and Appendix F provides the Multi-AXIS prompts and output constraints.

Ethics Statement

We study multivariate time-series anomaly explanation using publicly available real-world benchmark datasets (SMD, SWaT, and VTA Articulary) and programmatically generated synthetic data. All real-world datasets are public research benchmarks and contain no personally identifiable information; the synthetic data are produced by a stochastic generator and do not correspond to any real entity. Our human evaluation uses six expert annotators to assess anonymized model responses; the experts are compensated at standard academic research rates as described in Appendix E.1. The study involves no intervention in real-world systems. The questions and teacher answers in our benchmark are generated by large language models. We control their quality through a three-stage automatic checking procedure covering label consistency, information leakage, and style normalization, which reduces the risk that model-generated content introduces incorrect labels or spurious correlations. Our method is intended for anomaly-explanation research. When deployed in safety-critical domains such as industrial monitoring, healthcare, and finance, automatically generated explanations may appear plausible while being incorrect. Such deployments therefore require domain-specific validation and appropriate human oversight, and model outputs should not serve as the sole basis for operational decisions.

Reproducibility Statement

We provide a complete description of Multi-AXIS in Section 3 and Appendix A, covering detector representation extraction, anomaly-signal enhancement, the evidence interface, prompt assembly, and the training objective. Appendix A.1 details the architecture of the time-series encoder, and Appendix A.2 reports the backbone configuration, the number of trainable parameters, training hyperparameters, and the decoding settings used at inference. Appendix B documents the input interfaces and task adaptations of every time-series QA baseline. The construction pipeline of MSAB, including anomaly templates, channel–time ground-truth recovery, window selection, question perspectives, data splits, quality control, representative prompts, and statistics, is described in Section 3.7 and Appendix C. The automated evaluation protocol, per-dimension metric definitions, aggregation procedure, and judge prompts are given in Appendix D; Appendix E reports the human-evaluation protocol, agreement analysis, and detailed results; and the full prompts and output constraints of Multi-AXIS are provided in Appendix F. All baselines are cited to their original papers with their input configurations stated. Code is available at the anonymous repository `https://anonymous.4open.science/r/Multi-AXIS-B1DC/`.

References

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Équille, and Kalyan Veeramachaneni. Can large language models be anomaly detectors for time series? In *Proceedings of the 11th IEEE International Conference on Data Science and Advanced Analytics*, pp. 1–10, 2024.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, Jasper Zschiegner, Danielle C. Maddix, Hao Wang, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*, 2024.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A. Zuluaga. USAD: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3395–3404, 2020.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018. *arXiv preprint arXiv:1811.00075*, 2018.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. A decoder-only foundation model for time-series forecasting. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 10148–10167, 2024.

Ailin Deng and Bryan Hooi. Graph neural network-based anomaly detection in multivariate time series. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 4027–4035, 2021.

Manqing Dong, Hao Huang, and Longbing Cao. Can LLMs serve as time series anomaly detectors? *arXiv preprint arXiv:2408.03475*, 2024.

Mononito Goswami, Konrad Szafer, Arjun Choudhry, Yifu Cai, Shuo Li, and Artur Dubrawski. MOMENT: A family of open time-series foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 16115–16152, 2024.

Tong Guan, Zijie Meng, Dianqi Li, Shiyu Wang, Chao-Han Huck Yang, Qingsong Wen, Zuozhu Liu, Sabato Marco Siniscalchi, Ming Jin, and Shirui Pan. TimeOmni-1: Incentivizing complex reasoning with time series in large language models. In *Proceedings of the Fourteenth International Conference on Learning Representations*, 2026a.

Tong Guan, Sheng Pan, Johan Barthelemy, Zhao Li, Yujun Cai, Cesare Alippi, Ming Jin, and Shirui Pan. TimeOmni-VL: Unified models for time series understanding and generation. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.

Kyle Hundman, Valentino Constantinou, Christopher Laporte, Ian Colwell, and Tom Söderström. Detecting spacecraft anomalies using LSTMs and nonparametric dynamic thresholding. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 387–395, 2018.

Rahat Iqbal, Tomasz Maniak, Faiyaz Doctor, and Charalampos Karyotis. Fault detection and isolation in industrial processes using deep learning approaches. *IEEE Transactions on Industrial Informatics*, 15(5):3077–3084, 2019.

Hyeongwon Kang, Jeongseob Kim, Jinwoo Park, and Pilsung Kang. Detecting time series anomalies like an expert: A multi-agent LLM framework with specialized analyzers. *arXiv preprint arXiv:2605.05725*, 2026.

Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. Time-MQA: Time series multi-task question answering with context enhancement. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pp. 29736–29753, 2025.

Kwei-Herng Lai, Daochen Zha, Junjie Xu, Yue Zhao, Guanchu Wang, and Xia Hu. Revisiting time series outlier detection: Definitions and benchmarks. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Tian Lan, Hao Duong Le, Jinbo Li, Wenjun He, Meng Wang, Chenghao Liu, and Chen Zhang. Towards foundation models for zero-shot time series anomaly detection: Leveraging synthetic data and relative context discrepancy. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. Large language models can deliver accurate and interpretable time series anomaly detection. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 4623–4634, 2025.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, 2023.

Aditya P. Mathur and Nils Ole Tippenhauer. SWaT: A water treatment testbed for research and training on ICS security. In *Proceedings of the 2016 International Workshop on Cyber-Physical Systems for Smart Water Networks*, pp. 31–36, 2016.

Jun Woo Park, Kyudan Jung, Dohyun Lee, Hyuck Lee, Daehoon Gwak, ChaeHun Park, Jaegul Choo, and Jaewoong Cho. Delving into large language models for effective time-series anomaly detection. In *Advances in Neural Information Processing Systems 38*, 2025.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Haotian Si, Jianhui Li, Changhua Pei, Hang Cui, Jingwen Yang, Yongqian Sun, Shenglin Zhang, Jingjing Li, Haiming Zhang, Jing Han, Dan Pei, and Gaogang Xie. TimeSeriesBench: An industrial-grade benchmark for time series anomaly detection models. In *Proceedings of the 35th IEEE International Symposium on Software Reliability Engineering*, pp. 61–72, 2024.

Junho Song, Keonwoo Kim, Jeonglyul Oh, and Sungzoon Cho. MEMTO: Memory-guided transformer for multivariate time series anomaly detection. In *Advances in Neural Information Processing Systems 36*, pp. 57947–57963, 2023.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 2828–2837, 2019.

Shreshth Tuli, Giuliano Casale, and Nicholas R. Jennings. TranAD: Deep transformer networks for anomaly detection in multivariate time series data. *Proceedings of the VLDB Endowment*, 15(6): 1201–1214, 2022.

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

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Anomaly transformer: Time series anomaly detection with association discrepancy. In *Proceedings of the Tenth International Conference on Learning Representations*, 2022.

Xiongxiao Xu, Haoran Wang, Yueqing Liang, Philip S. Yu, Yue Zhao, and Kai Shu. Can multimodal LLMs perform time series anomaly detection? In *Proceedings of the ACM Web Conference 2026*, pp. 5392–5403, 2026.

Yiyuan Yang, Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. DCdetector: Dual attention contrastive representation learning for time series anomaly detection. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 3033–3045, 2023.

Fangxu Yu, Hongyu Zhao, and Tianyi Zhou. TS-Reasoner: Aligning time series foundation models with LLM reasoning. *Transactions on Machine Learning Research*, 2026.

Zahra Zamanzadeh Darban, Geoffrey I. Webb, Shirui Pan, Charu C. Aggarwal, and Mahsa Salehi. Deep learning for time series anomaly detection: A survey. *ACM Computing Surveys*, 57(1):1–42, 2024.

Chuxu Zhang, Dongjin Song, Yuncong Chen, Xinyang Feng, Cristian Lumezanu, Wei Cheng, Jingchao Ni, Bo Zong, Haifeng Chen, and Nitesh V. Chawla. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1409–1416, 2019.

Zihao Zhou and Rose Yu. Can LLMs understand time series anomalies? In *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025.

Jiaxin Zhuang, Leon Yan, Zhenwei Zhang, Ruiqi Wang, Jiawei Zhang, and Yuantao Gu. See it, think it, sorted: Large multimodal models are few-shot time series anomaly analyzers. *arXiv preprint arXiv:2411.02465*, 2024.

Nicolas Zumarraga, Lorenzo Steno, Ning Wang, Max Rosenblattl, Thomas Kaar, Maxwell A. Xu, Kevin O'Sullivan, Markus Kreft, Elgar Fleisch, Paul Schmiedmayer, Patrick Langer, and Robert Jakob. TimeRLM: Recursive language models enable precise anomaly localization in long-context time-series. *arXiv preprint arXiv:2608.03391*, 2026.

A Multi-AXIS Implementation Details

This section provides additional details on the encoder architecture, evidence interface, and experimental configuration omitted from Section 3 of the main paper.

A.1 Time-Series Encoder

Multi-AXIS uses the frozen multivariate time-series encoder f_θ from TimeRCD (Lan et al., 2026). The overall architecture is shown in Figure 3.

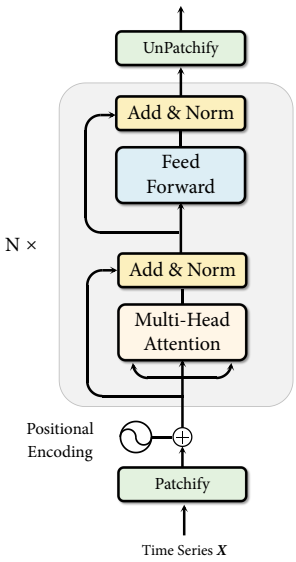


Figure 3: The structure of the time-series encoder.

Patchification. Let the patch length be W. We zero-pad the normalized sequence $\widetilde{\mathbf{X}} \in \mathbb{R}^{T\times C}$ along the temporal axis to $T' = \lceil T/W \rceil W$, and define $P = T'/W$. For channel c and patch index p,

$$\mathbf{X}^{(c,p)} = [\widetilde{x}_{pW,c}, \ldots, \widetilde{x}_{(p+1)W-1,c}] \in \mathbb{R}^W,$$
$$c = 0, \ldots, C-1, \qquad p = 0, \ldots, P-1.$$

Patches from all channels are flattened in channel-major order into $\mathbf{X}_p \in \mathbb{R}^{CP\times W}$. Let m_t and m_c^{ch} denote valid time steps and valid channels, respectively. The patch mask is

$$m_{c,p}^{\text{patch}} = \mathbb{I}\left[m_c^{\text{ch}} = 1 \ \wedge \sum_{\tau=pW}^{(p+1)W-1} m_\tau > 0 \right].$$

This mask allows samples with different sequence lengths and numbers of channels to be batched together while preventing padded positions from participating in attention.

Patch embeddings and positional encoding. A shared linear layer maps each patch to d_{model} dimensions:

$$\mathbf{Z}_{c,p} = \mathbf{X}^{(c,p)}\mathbf{W}_e + \mathbf{b}_e, \qquad \mathbf{Z} \in \mathbb{R}^{CP\times d_{\text{model}}}.$$

RoPE encodes the temporal patch index p in the query and key of each attention head. Different channels at the same temporal index use the same rotary position, preserving temporal alignment across channels. The encoder also learns two scalar biases for each attention head, corresponding

to token pairs from the same channel and from different channels. Each position (c, p) can therefore attend both to long-range dynamics within the same channel and to synchronous or lagged relationships across channels.

Non-causal Transformer encoder. We stack L pre-norm Transformer encoder layers. Each layer contains (i) multi-head self-attention with H heads and head dimension $d_{\text{head}} = d_{\text{model}}/H$, and (ii) a LLaMA-style MLP with gated GELU activation. RMSNorm is applied before both the attention and MLP modules. Attention is non-causal and operates bidirectionally over all CP valid tokens. A binary variate bias distinguishes within-channel from cross-channel connections. Each patch representation can therefore use the full temporal range and all channels rather than encoding only an isolated local window.

Unpatchification and position-wise projection. The Transformer output $\mathbf{U} \in \mathbb{R}^{CP \times d_{\text{model}}}$ is mapped through a linear projection:

$$\mathbf{Y} = \mathbf{U}\mathbf{W}_p + \mathbf{b}_p \in \mathbb{R}^{CP \times (W d_p)}.$$

After reshaping, restoring temporal order, and removing trailing padding, we obtain the representation used in the main paper:

$$\mathbf{H}^{\text{step}} \in \mathbb{R}^{T \times C \times d_p}, \qquad d_p = 256.$$

Thus, $\mathbf{H}^{\text{step}}_{t,c}$ retains the channel-time address (t, c), while its receptive field covers the complete valid sequence and all valid channels.

Anomaly head and pretraining objectives. A two-layer MLP maps each $\mathbf{H}^{\text{step}}_{t,c}$ to normal/anomalous logits $\mathbf{L}_{t,c} = (\ell^N_{t,c}, \ell^A_{t,c})$. During TimeRCD training, these logits are first averaged across channels before the per-time-step classification loss is computed. Multi-AXIS instead reads the output at each (t, c) before this aggregation to preserve channel identity. TimeRCD uses masked reconstruction to provide dense representation-learning signals and anomaly classification to shape discriminative directions; reconstruction error is not used for scoring at inference time (Lan et al., 2026). Multi-AXIS does not retrain the encoder, reconstruction head, or anomaly head.

A.2 Training and Inference Configuration

Time-series encoder configuration. The time-series encoder f_θ uses a Transformer-based, patch-tokenization architecture. The patch size is $P = 16$, enabling efficient processing of long sequences while retaining temporal resolution. The encoder contains $L = 8$ Transformer layers with model dimension $d_{\text{model}} = 512$ and $H = 8$ attention heads, giving a head dimension of $d_{\text{head}} = 64$. The projection dimension is $d_p = 256$.

Shared Semantic Alignment architecture. The Perceiver-based Shared Semantic Alignment architecture aligns time-series representations with the frozen LLM embedding space. Its prototype bank contains $N_{\text{proto}} = 1024$ prototype embeddings derived from the LLM vocabulary through a learnable linear mapping. We use $K_F = 30$ fixed task-prior tokens, which provide sufficient capacity for global task-level priors while remaining computationally efficient. Cross-attention within Shared Semantic Alignment uses eight attention heads, matching the encoder configuration.

Language-model integration. We primarily use **Qwen3-VL-8B-Instruct** (Bai et al., 2025) as the base model. Its text hidden dimension is $d_\ell = 4096$, and its vocabulary size is $|\mathcal{V}| = 151936$. We add the special tokens `<POSITION_HINT>`, `<CHANNEL_HINT>`, `<EVENT_HINT>`, and `<FIXED_HINT>` to the vocabulary for hint injection. Their token IDs are assigned dynamically during initialization.

Inference parameters. During generation, the model employs beam search with 5 beams and a repetition penalty of 1.15 to ensure diverse and coherent explanations. The maximum generation length is capped at 1000 tokens to accommodate detailed explanations while preventing excessive verbosity. No-repeat n-gram size is set to 3 to avoid repetitive patterns, and length penalty is maintained at 1.0 to balance explanation completeness with conciseness.

B Evaluation Implementation for Time-Series QA Baselines

This section describes how we adapt each baseline to our multivariate anomaly question-answering task.

B.1 Unified Evaluation Setup

On each dataset, all methods answer the same questions with the same answer choices and target intervals. The evidence modality follows each method's native interface: time-series models receive numerical sequences or continuous representations, visual models receive full multichannel plots, and TimeRLM analyzes the raw sequences on demand by executing code. The plots retain anonymized channel identities, observed trajectories, and global temporal positions, with the target interval highlighted.

All baselines use publicly available weights and are not further trained on our training set. Their inputs exclude reference answers, answer labels, anomaly types, ground-truth root causes or affected channels, and external evaluation results. Each method produces only one final response per question, without selecting the best result from multiple samples. Table B.1 summarizes the input and adaptation of each baseline.

Table B.1: Input interfaces and task adaptations for the time-series QA baselines.

Method	Backbone used	Evidence and scope	Task adaptation
ChatTS	ChatTS-14B	Full raw multichannel sequence	Encode the input with the native processor; add only the question, interval, and output specification
AnomLLM (Full)	Qwen2.5-VL-7B-Instruct	Full projected numerical trajectory	Use the original text interface and mark the target interval
AnomLLM (Window)	Qwen2.5-VL-7B-Instruct	Projected numerical values within the target interval	Change only the visible temporal range
TSAD-Agents	Qwen2.5-VL-7B-Instruct	Full multichannel plot; MSAB-Test additionally includes numerical values within the target interval	Adapt the visual detection interface to single-turn QA
TS-Reasoner	Qwen2.5-7B + TimesFM-1.0-200M	Full raw multichannel sequence	Inject native time-series representations through a `<ts>` placeholder for each channel
Time-MQA	Qwen2.5-7B-Instruct + rank-16 LoRA	Serialized numerical values within the target interval	Provide only text and window values, without the visual interface
TimeOmni-VL	Bagel-7B-based TimeOmni-VL (14.7B)	Full multichannel plot	Use only the official question-answering pathway, without time-series generation or forecasting
TimeOmni-1	Qwen2.5-7B-Instruct (7.6B)	Sampled numerical values from the full sequence and target interval	Select relevant channels from the question; retain all channels if parsing fails
TimeRLM	Qwen3-VL-8B-Instruct	Full raw sequence and on-demand analysis results	Iteratively inspect statistics and local plots through code execution

B.2 Additional Details on Task Adaptation

B.2.1 AnomLLM (Full/Window)

AnomLLM's text interface requires a multichannel input to be represented as a single numerical trajectory. When a question concerns one clearly identified channel and does not require cross-channel

comparison, we directly retain the raw values of that channel to preserve its actual magnitude. When multiple channels must be analyzed jointly, their scales may differ substantially, so we first standardize each channel independently as $z_{t,c}$. Without a focal channel, we define the projection as

$$r_t = \sqrt{\frac{1}{C}\sum_{c=1}^{C} z_{t,c}^2}.$$

If the question specifies a focal channel f and requires multichannel analysis, we use

$$p_t = 0.7 z_{t,f} + 0.3\sqrt{\frac{1}{C-1}\sum_{c\neq f} z_{t,c}^2}.$$

The first term preserves the main variation of the focal channel, while the second provides an aggregate reference from the remaining channels. AnomLLM (Full) receives the full projected trajectory, whereas AnomLLM (Window) receives only the target interval. Both settings use numerical text without activating the backbone's visual input pathway. This preserves the original text-based time-series interface while isolating the effect of context range.

B.2.2 TIMEOMNI-1

TimeOmni-1 receives numerical evidence in text form. We first standardize each channel independently over the full sequence, then uniformly sample 16 positions from the full sequence and 16 from the target interval. If the question or answer choices explicitly refer to channels, we serialize only the relevant channels; otherwise, we retain all channels. The model generates its reasoning and final answer in its native format. If the first response lacks a parseable answer field, the same model performs one constrained answer readout using the same evidence.

B.2.3 TIMERLM

TimeRLM keeps the full multichannel sequence outside the language context and allows the model to write and execute analysis code, compute statistics, and render local plots on demand. The model uses intermediate results to decide whether to continue its analysis or submit an answer. We do not enable a subordinate language model. Each question uses a single reasoning process with at most eight exploratory tool calls. If the process has not terminated after ten rounds, the model is instructed to answer directly from the available results. Three responses were submitted after reaching this round limit.

C CONSTRUCTION AND ANALYSIS OF THE MULTIVARIATE SEMANTIC ANOMALY BENCHMARK (MSAB)

This section describes the construction process, supervision signals, quality control, and composition of the **Multivariate Semantic Anomaly Benchmark (MSAB)**. MSAB refers to the programmatically generated corpus for multivariate anomaly explanation introduced in this work. SMD (Su et al., 2019), SWaT (Mathur & Tippenhauer, 2016), and VTA Articulary (Bagnall et al., 2018; Xu et al., 2026) serve only as independent real-world transfer test sets and do not contribute source sequences to MSAB. The underlying MSAB sequences are produced by the context-dependent multivariate generator from TimeRCD/Datasets-RCD (Lan et al., 2026).

C.1 CONSTRUCTION PIPELINE AND DATA SCOPE

MSAB is constructed in five consecutive stages:

1. **Multivariate sequence generation, anomaly injection, and ground-truth conversion.** We generate the normal temporal structure of each channel and construct cross-channel dependencies using a random directed acyclic graph and lagged dynamics. We then inject a local anomaly into one directly intervened channel and convert the original time-step labels into channel–time labels.

2. **Window extraction and visualization.** We extract anomalous windows and normal windows with substantial variation from each complete sequence, and render multichannel plots that preserve the full temporal context.
3. **Question-perspective organization.** We organize question records for the same physical window from two perspectives: anomaly analysis and normality verification.
4. **Question generation.** GPT-5.5 generates natural-language questions according to the question perspective, the MC/TF/OE question type, and a predefined semantic specification.
5. **Teacher-answer generation.** GPT-5.4 generates structured supervision answers from the final question, time-series evidence, and generation ground truth.

The production run for the training and validation corpora starts from 20,000 candidate source sequences. For the n-th sequence, the sequence length and number of channels are sampled as

$$T_n \sim U_{\mathbb{Z}}(128, 512), \qquad C_n \sim U_{\mathbb{Z}}(5, 50).$$

We target one anomalous window and one normal window per source sequence, with window length $L \in [15, 60]$. If one candidate pool is empty, the missing window is drawn from the other pool. Thus, the one-anomalous–one-normal ratio is a target rather than a hard per-sequence constraint. A candidate source sequence enters the final corpus only after it yields valid windows and completes question–answer construction. The final data composition is reported in Table C.5.

C.2 Normal Temporal Structure and Cross-Channel Context

C.2.1 Per-Channel Background

For each channel c, the generator first constructs a normal background

$$u_c(t) = m_c(t) + s_c(t) + b_c(t) + \varepsilon_c(t), \qquad t = 0, \ldots, T-1,$$

where m_c, s_c, b_c, and ε_c denote the trend, seasonal component, background perturbation excluded from anomaly labels, and noise, respectively. The trend component covers stationary, monotonic, piecewise-varying, and stochastic processes. The seasonal component covers aperiodic, sinusoidal, square-wave, triangular-wave, and wavelet patterns, with randomly varied periods and amplitudes. Noise levels include near-zero, low, medium, and high noise; medium- and high-noise settings may also exhibit piecewise heteroscedasticity. The background term b_c may contain cycle-synchronized background spikes or periodic noise modulation. These patterns are added to both the observed sequence and the normal counterfactual, increasing morphological complexity without being labeled as anomalous.

C.2.2 Cross-Channel Dependencies

The generator randomly samples a directed acyclic graph $G = (V, E)$. The number of edges is approximately 0.8–1.2 times the number of channels, and the longest directed path is limited to 3. This design covers both local dependencies and short-range propagation while avoiding unstable deep recursion. For channel i with parent set $\mathcal{P}(i)$, the cross-channel component is

$$z_i[t] = a_i z_i[t-1] + \sum_{j \in \mathcal{P}(i)} b_{ji} x_j[t - \ell_{ji}] + c_i,$$

and is mixed with the channel-specific background as

$$x_i[t] = \alpha_i z_i[t] + (1 - \alpha_i) u_i[t].$$

Here,

$$a_i \sim U[-0.8, 0.8], \quad b_{ji} \sim U[-2, 2], \quad c_i \sim U[-3, 3], \quad \alpha_i \sim U[0.2, 0.6], \quad \ell_{ji} \in \{0, \ldots, 7\}.$$

The observed sequence and its normal counterfactual are computed with the same graph structure and dynamic parameters, so they differ only in the anomaly intervention. This paired design allows the channel–time deviations caused by an anomaly to be determined directly from position-wise differences, without relying on a separate detector to re-estimate the ground truth.

C.3 Anomaly Injection

For each MSAB source sequence, we randomly select a directly intervened channel r, a local anomaly template, and its start position, duration, direction, and magnitude. The anomalous interval covers at most 0.35 of the full sequence. Candidates with out-of-range positions, overlapping intervals, or no valid anomaly are rejected and resampled. To prevent the anomaly from being obscured by noise, the generator adaptively sets a lower bound on its magnitude according to the background and noise scales of the corresponding channel. Spike and outlier anomalies must reach at least six times the estimated noise standard deviation, while other local anomalies must reach at least three times this value. The intervention magnitude is further increased in the multivariate setting. This rule controls visibility without eliminating difficulty differences caused by background structure, channel scale, or propagation strength.

The generator samples two intervention mechanisms with equal probability:

- **Endogenous intervention.** The anomaly is first added to the univariate background of channel r, after which the multivariate sequence is generated through the lagged dynamics of the DAG. Consequently, descendants of r may respond at different delays.
- **Exogenous intervention.** A complete normal multivariate sequence is generated first, and channel r is then modified directly. The anomaly does not propagate further through the generation graph.

Let $\Delta_\eta(t)$ denote the anomaly term controlled by parameters η. The two interventions can be written as

$$u'_r(t) = u_r(t) + \Delta_\eta(t) \quad \text{(endogenous)}, \qquad X^{\text{obs}}_{t,r} = X^{\text{cf}}_{t,r} + \Delta_\eta(t) \quad \text{(exogenous)}.$$

MSAB covers 20 local anomaly templates. Table C.1 groups them into six families based on their observable morphology and reports their source-sequence distributions in the training and validation sets.

Table C.1: Local anomaly types in the MSAB training and validation sets.

Family	Included templates	Main morphology	Source sequences
Point anomaly	outlier	A single positive or negative deviation	6,524
Spike	upward/downward spike; continuous upward/downward spike; wide upward/-downward spike	Isolated, repeated, or wide spikes	5,810
Persistent level shift	sudden increase/decrease; four increase/decrease variants after an upward/downward spike	A persistent shifted state after an abrupt change	2,702
Asymmetric rise/decline	rapid rise followed by slow decline; slow rise followed by rapid decline; corresponding decline–rise variants	Asymmetric rise, decline, and recovery rates	1,329
Convex perturbation	upward/downward convex	A local upward or downward convex excursion	676
Oscillatory perturbation	shake	Local random or high-frequency oscillation	673
Total	**20 templates**	–	**17,714**

C.4 CHANNEL–TIME GROUND TRUTH

The underlying generator retains the anomalous observation $\mathbf{X}^{\mathrm{obs}}$, the normal counterfactual without anomaly injection $\mathbf{X}^{\mathrm{cf}}$, the original time-step labels $\mathbf{y} \in \{0,1\}^T$, and the generation attributes. We first compute the observed–counterfactual difference for each channel:

$$d_{t,c} = \left| X_{t,c}^{\mathrm{obs}} - X_{t,c}^{\mathrm{cf}} \right|,$$

and then recover the channel–time labels using a channel-adaptive threshold:

$$\tau_c = \max\left\{\tau_0, \rho\left(\mathrm{std}_t\, X_{t,c}^{\mathrm{cf}} + \epsilon\right)\right\}, \qquad \widehat{Y}_{t,c} = \mathbb{I}[d_{t,c} > \tau_c]\,\mathbb{I}[y_t = 1],$$

where $(\tau_0, \rho, \epsilon) = (10^{-4}, 0.05, 10^{-6})$. The time-step labels restrict the temporal support of the anomaly, while the observed–counterfactual difference further identifies channels with discernible deviations within that support. The directly intervened channel is given by the generation attributes. The set of affected channels within target window $[s, e)$ is defined as

$$\mathcal{C}_{\mathrm{aff}}(s,e) = \left\{ c : \max_{s \le t < e} \widehat{Y}_{t,c} = 1 \right\}.$$

This process yields programmatic ground truth for the window-level anomaly status, directly intervened channel, affected channels, anomaly morphology, and anomaly scope.

C.5 WINDOW EXTRACTION AND VISUALIZATION

The objective label of a target window is determined by the channel–time ground truth that it covers:

$$y_{[s,e)} = \mathbb{I}\left[\max_{s \le t < e,\, c} \widehat{Y}_{t,c} = 1 \right].$$

For each source sequence, the program randomly generates at most 1,000 candidate intervals of length 15–60 and divides them into anomalous and normal pools according to whether they intersect $\widehat{\mathbf{Y}}$. Anomalous windows are sampled randomly from the anomalous pool. For a normal candidate window $W = [s, e)$, the program computes the overall variance across all channel observations within the window:

$$v(W) = \mathrm{Var}\left(\left\{ X_{t,c}^{\mathrm{obs}} : s \le t < e,\ 1 \le c \le C \right\}\right).$$

With probability 0.70, a normal window is sampled from the top third of candidates ranked by $v(W)$; otherwise, it is sampled from the full normal candidate pool. These **high-variation normal control windows** help distinguish visually salient changes from anomalies relative to context. This term also avoids confusion with the Hard question bank introduced below.

Each physical window is paired with a plot of the complete multivariate sequence, rather than a plot restricted to $[s, e)$. Each channel is normalized independently using the mean and standard deviation of the complete sequence, and is plotted in a separate subplot with a shared global time axis. If the standard deviation is near zero, the channel is only centered. The target interval is marked jointly by background shading, boundary lines, and thicker traces. The image therefore preserves three forms of information: the local morphology of the target event, the temporal reference before and after the event, and the alignment of different channels at the same global positions. The raw scale is not inferred from the image; it is provided through the numeric window and scale information in the model input.

Table C.2: Semantic coverage of the Regular question bank.

Type	Count	Semantic specifications
Anomaly analysis		
TF	9	Anomaly evidence; primary anomaly source; single-/multichannel scope; morphology; alarm support; root-cause claim; multichannel anomaly; affected scope; stability
MC	9	Anomaly interpretation; primary anomaly source; affected scope; morphology family; alarm interpretation; anomaly scope; affected channels; channel stability; claim strength
OE	9	Strongest anomaly evidence; root-cause candidate; directly anomalous channel; primary morphology; distinguishing evidence; limits of root-cause claims; affected channels; anomaly scope; unstable channels
Normality verification		
TF	9	Normality; contextual consistency; absence of a local anomaly; false positive; temporal alignment; insufficient root-cause evidence; insufficient evidence for a widespread anomaly; insufficient multichannel evidence; local fluctuation
MC	9	Normal interpretation; channel stability; contextual consistency; false-positive interpretation; fluctuation strength; conservative anomaly scope; conservative affected scope; stability description; claim strength
OE	9	Evidence of normality; unstable channels; evidence of contextual consistency; false-positive evidence; absence of anomaly evidence; limits of root-cause claims; limits on affected channels, widespread anomalies, and multichannel claims

C.6 QA Records, Question Perspectives, and Question Generation

C.6.1 Data Records and the QA Instances Derived from One Physical Window

Each physical window first produces two **basic QA instances**. The anomaly-analysis perspective asks about anomaly evidence, the primary anomaly source, the affected scope, or morphology. The normality-verification perspective asks whether the interval can be considered normal, whether a strong anomaly claim lacks sufficient evidence, or whether an alarm may be a false positive. The two records share the source sequence, target interval, image, and objective ground truth; they differ only in question perspective and are linked by a pairing identifier. Thus, a normal window may receive an anomaly-analysis question, and an anomalous window may receive a normality-verification question. Question wording therefore cannot serve as a direct shortcut to the label.

The 33,815 distinct physical windows in the training and validation sets correspond to 67,630 basic QA instances. Another 173 records reuse the same physical windows in a supplementary question bank, yielding 67,803 QA instances in total.

C.6.2 Semantic Coverage from Two Perspectives

Question generation uses Regular and Hard question banks. Each **semantic specification** binds a semantic identifier, focus type, question scaffold, generation instruction, and optional MC option anchors into a single unit, preventing arbitrary combinations of stems and options from different tasks. The Regular bank covers basic judgments about anomaly presence, source, scope, morphology, stability, and false positives. The Hard bank further evaluates temporal ordering, local spread, relation breaks, minimum affected sets, leader–response roles, phase localization, and cross-channel consistency.

Table C.2 summarizes the 54 semantic specifications in the Regular bank. Each perspective–question-type cell contains nine specifications. A global shuffled cycle keeps MC, TF, and OE approximately balanced across the corpus.

Table C.3 summarizes the 32 semantic specifications in the Hard question bank. TF and MC each use four complementary tasks to evaluate decidable claims. OE covers both task groups and therefore contains eight open-ended specifications for each perspective. Hard records share the answer type under the pairing identifier, allowing the two perspectives on the same physical window to be compared using the same response format. Regular records instead target question-type balance across the corpus and do not require the two perspectives for an individual window to use the same format.

Table C.3: Semantic coverage of the Hard question bank.

Type	Count	Semantic specifications
Anomaly analysis		
TF	4	Lead–lag relations; local spread; relation breaks; anomaly support after excluding the focal channel
MC	4	Minimum affected set; source and response roles; phase localization; cross-channel consistency
OE	8	Open-ended analysis of the eight judgment and selection dimensions above
Normality verification		
TF	4	Insufficient temporal ordering; absence of spread evidence; preserved relations; insufficient evidence after excluding the focal channel
MC	4	Conservative minimum set; source and response roles; phase localization; cross-channel consistency
OE	8	Open-ended analysis of the eight dimensions above from a conservative verification perspective

C.6.3 Information and Constraints for GPT-5.5 Question Generation

Each GPT-5.5 request consists of a system message that controls the question style and a user message containing text and the complete sequence image. The user message provides the target interval and its global position, the question perspective, the question type, one complete semantic specification, and a ground-truth summary that ensures compatibility between the question and the current sample. For anomalous samples, the summary includes the anomaly status, directly intervened channel, affected channels, anomaly scope, and available morphology labels. For normal samples, it explicitly states that no anomaly source or affected channel may be assigned.

The ground-truth summary is used only to ensure that the question is answerable for the current sample; its contents must not appear in the question stem. Accordingly, the prompt prohibits disclosing the anomaly label, correct channel, true morphology, or specific local behavior in the question. It also requires channel-attribution questions to include candidate channels or channel sets at comparable granularity. All position references must use global indices from the complete sequence, and each question must test only one semantic concept. The question-generation model can see the complete sequence image with the highlighted interval, but cannot access the point-wise values in the target window, the normal counterfactual, the teacher answer, or the detector-based contextual evidence used by Multi-AXIS. This information boundary serves two purposes. Withholding the numeric table prevents the question from directly restating values and indirectly revealing the answer, while retaining the complete sequence image allows the model to formulate natural questions about visual morphology.

Question generation uses GPT-5.5 with a temperature of 0.5 and a maximum output length of 900 tokens. An MC instance must contain a nonempty question stem and four options labeled A–D. TF and OE instances must contain a nonempty, single-task question. Representative prompts and semantic specifications are provided in Section C.11.

C.7 Teacher-Answer Generation and Information Boundaries

The teacher request includes the same complete sequence image, the final question and MC options, the target interval, a neutral description of the complete sequence, the values of all time steps and channels within the target window, and the structured ground truth. Every cell in the numeric table is uniformly formatted as "observed value (normal counterfactual value).' ' This observed–counterfactual pairing is used only to help the teacher locate deviations and estimate their magnitudes accurately.

The teacher-side ground truth includes the window-level anomaly status, directly intervened channel, affected channels, anomaly type, and anomaly scope. The normal counterfactual and structured ground truth constitute **offline privileged supervision**. The prompt instructs GPT-5.4 to use this information to determine the conclusion, but the final answer may cite only observable channels, global positions, local morphology, duration, and approximate magnitudes from the plot. It must

not mention counterfactual arrays, hidden labels, or generation metadata. An MC answer first gives the unique option letter and then explains why that option is preferable to the main distractor. A TF answer begins with `Yes.` or `No.` and then explains the basis for the judgment. An OE answer presents `Decision:`, `Main evidence:`, and `Interpretation:` in this order. All positions use global indices from the complete sequence.

Teacher answers are generated by GPT-5.4 with a temperature of 0.5 and a maximum output length of 900 tokens. Table C.4 distinguishes the information visible during question generation, teacher supervision, and model training or testing.

Table C.4: Information boundaries for question generation, teacher supervision, and Multi-AXIS.

Information	**GPT-5.5 question generation**	**GPT-5.4 teacher**	**Multi-AXIS train/test input**
Complete sequence image with highlighted window	✓	✓	✓
Target interval address	✓	✓	✓
Observed values in the target window	–	✓	✓
Normal counterfactual values	–	✓ offline supervision only	–
Structured anomaly ground truth	Summary only (question constraint)	✓ offline supervision only	–

C.8 Quality Control and Targeted Regeneration

MSAB applies independent checks at the signal, question, and answer levels.

Signals and ground truth. The generator rejects candidates with out-of-range positions, overlapping intervals, anomaly coverage above 0.35, or no valid anomaly. It also verifies that the observations, counterfactuals, and labels agree in channel count, target interval, and directly intervened channel. The observed–counterfactual pairing ensures that the window label and affected channels are determined by the same intervention. The magnitude lower bounds make anomalies distinguishable from noise, while the high-variation normal windows reduce the shortcut that equates large changes with anomalies.

Questions. The question type, perspective, and semantic specification are fixed before GPT-5.5 is called. Automated checks verify that the question is nonempty, that an MC question has exactly four options labeled A–D, and that each record matches its assigned perspective, question type, and window. Before release, semantic review further checks whether the question can be answered from the visible sequence, whether it reveals the label, correct channel, or true morphology, whether it contains only one task, whether the MC options are parallel and provide effective distractors, and whether all channel and position references are valid.

Answers. Each teacher answer is aligned one-to-one with its final question. We verify that the answer is nonempty and corresponds to the correct question record. Before release, semantic review further checks whether the conclusion matches the ground truth, whether the answer cites the correct channels and global positions, whether it provides observable evidence, and whether it exposes counterfactual information or generation metadata in the final text.

When an instance fails quality control, regeneration returns only to the closest affected stage. Invalid values or ground truth trigger physical-window resampling. An invalid question is regenerated together with its teacher answer while the source sequence, window, ground truth, perspective, question type, and semantic specification remain fixed. If the question is valid but the answer is not, only the answer is regenerated. The two question perspectives linked by the same pairing identifier are reviewed together to avoid retaining only one side and introducing pairing bias.

These checks provide three complementary safeguards. Programmatic interventions yield traceable channel–time ground truth, controlled semantic specifications restrict the capability tested by each question, and privileged teacher-side information improves the correctness of supervision while remaining isolated from student inputs through the information boundary. Quality control therefore

constrains numeric ground truth, question answerability, and evidence grounding without relying on a single language model to both generate and validate all content.

C.9 Data Splits and Leakage Control

MSAB is split by **source sequence**, rather than by individual QA record. All windows, questions from both perspectives, Regular and Hard records, and teacher answers derived from the same source sequence must belong to the same split. Paired records for a physical window already share a source sequence, so this rule also prevents paired questions from being separated across splits. Consequently, the training set and the validation/test sets do not share the same complete curve or near-duplicate samples with the same highlighted window.

Data construction introduces two potential leakage channels. The first is **question-stem leakage**. GPT-5.5 reads a ground-truth summary to generate an answerable question, and its content could leak into the wording. To prevent this, the prompt prohibits disclosing the label, correct channel, true morphology, and specific local behavior. It also requires channel questions to present comparable candidates, and semantic review checks for leakage before release. In addition, the paired construction in Section C.6.1 balances both question perspectives over every window and makes them orthogonal to the anomaly label, so the question style itself does not provide a label shortcut. The second channel is **privileged teacher-information leakage**. GPT-5.4 reads the normal counterfactual and structured ground truth, but this information is used only for offline supervision. Under the information boundary in Table C.4, neither Multi-AXIS nor any baseline model receives it as input.

MSAB-Test is not used for training or model selection. SMD, SWaT, and VTA Articulary are fully independent of the synthetic source sequences in MSAB and retain their native labels. Counterfactual, root-cause, or propagation fields unavailable in the real-world data remain missing and are not fabricated by a language model. The real-world results therefore do not depend on hidden information from the synthetic generator.

C.10 Data Composition

Table C.5 separately reports MSAB and the three real-world transfer test sets. MSAB-Train, MSAB-Validation, and MSAB-Test constitute MSAB, while the real-world datasets are used only for independent testing and are not part of MSAB.

Table C.5: Composition of MSAB and the real-world transfer test sets.

Data split	Use	QA instances	Physical windows	Source sequences
MSAB-Train	Training	61,053	30,447	15,943
MSAB-Validation	Validation	6,750	3,368	1,771
MSAB-Test	Synthetic testing	478	240	216
MSAB subtotal	–	**68,281**	**34,055**	**17,930**
SMD	Real-world testing	200	100	47
SWaT	Real-world testing	184	92	80
VTA Articulary	Real-world testing	200	100	50
All records	–	**68,865**	**34,347**	**18,107**

MSAB-Train and MSAB-Validation are approximately balanced in question type, question perspective, and anomalous versus normal windows.

C.11 Representative Prompts and Semantic Specifications

This section provides representative templates: one example for each of the three question styles, one complete parameterized question-generation prompt, one semantic specification for each question-bank–perspective–question-type cell, and one complete teacher prompt. Other paraphrased styles

and semantic specifications do not alter the information boundary or task contract. The complete text is released with the code.

C.11.1 Question Styles

The system message first selects one of three groups—ordinary user, business user, and cautious verifier—with equal probability, and then samples a paraphrased variant within that group. It controls only the linguistic style of the question and does not provide task ground truth. Table C.6 gives one representative message from each group.

Table C.6: Representative GPT-5.5 system messages for question styles.

Style group	Representative system message
Ordinary user	*"You are a curious beginner looking at a chart and asking simple questions about what looks wrong. Use everyday language, avoid technical jargon, and ask questions like a regular person would."*
Business user	*"You are an operations manager checking for immediate alerts or problems in the system. Use urgent, action-focused language emphasizing operational impact."*
Cautious verifier	*"You are a cautious inspector asking probing questions to ensure nothing has gone wrong. Use careful, investigative language that questions and verifies claims."*

C.11.2 Complete Parameterized GPT-5.5 Question-Generation Prompt

The following is the paper-level parameterized form of an MC request. {ANOMALY_STATUS_LINE}, {REFERENCE_FACT_SUMMARY}, {FOCUS_TEMPLATE}, {FOCUS_INSTRUCTION}, and the four option anchors are populated from the current record. The image is sent as visual content in the same user message.

GPT-5.5 Question-Generation Prompt

Generate a multiple−choice question about anomaly detection in this multivariate time−series window.

Context:
− Analysis window from position {s} to {e} in the full time series.
− The window contains {L} points at global positions {s} to {e−1}.
− All position references must use global positions in the full time series.
− {ANOMALY_STATUS_LINE}
− Reference anomaly context for writing a valid question, not for direct disclosure:
{REFERENCE_FACT_SUMMARY}

Selected focus:
− Template: {FOCUS_TEMPLATE}
− Instruction: {FOCUS_INSTRUCTION}

When channel attribution is involved, name a small comparison set of plausible candidate channels or channel groups. Do not treat one preselected channel as the answer. The question may name channels and global positions, but must not describe their specific behavior or disclose the true abnormal channel.

Semantic option anchors:
− A. {OPTION_ANCHOR_A}
− B. {OPTION_ANCHOR_B}
− C. {OPTION_ANCHOR_C}
− D. {OPTION_ANCHOR_D}

Requirements:
1. Output one concise question stem followed by exactly four options, A−−D.
2. Keep the stem natural and focused on one phenomenon or decision.
3. Use the selected semantic focus; do not drift to another task.
4. Keep the options parallel, concise, and comparable in specificity.

5. Prefer short labels or claims to long morphology stories or explanations.
6. If one option names a channel or channel group, keep competing options equally concrete.
7. Preserve the roles and order of the option anchors, but rewrite them naturally.
8. Let the stem identify the global interval and ask the decision. Do not narrate spikes, dips, oscillations, bursts, reversals, timing stories, or rough amplitudes before the question.
9. Do not append a second descriptive clause introduced by "where", "with", "given", "based on", "showing", or "while".
10. For source, affected–channel scope, stability, anomaly scope, or detector interpretation, use concrete candidate channels or channel groups whenever the interval supports them.
11. Do not build the stem around one presumed abnormal channel. Use a small comparison set of plausible candidate channels or groups.
12. Name candidate channels briefly without describing their behavior or revealing which candidate is correct.
13. When applicable, include one alternative–channel distractor and another distractor that differs in scope.
14. Do not include numeric cues unless the task explicitly concerns a numeric threshold. Do not reveal labels, answers, or template metadata.

TF uses the same context, ground-truth confidentiality, global-position, and channel-candidate rules, but requires a single proposition that can be answered with `Yes/No`. OE requires one open-ended question in one of four forms: why, evidence, which-channel, or a single analysis request. For the normality-verification perspective, the prompt additionally prohibits describing evidence such as `close to normal'' or`small fluctuations'' in the question stem before the answer is given. Specific semantic specifications also add two types of conditional rules. Source- and scope-related questions must include plausible distractors involving alternative channels or scopes. Alarm-related MC questions must distinguish among no alarm, manual review, a single-channel alarm, and a multichannel alarm.

C.11.3 Representative Semantic Specifications

Table C.7 retains one representative specification for each question-bank–perspective–question-type cell.

Table C.7: Representative semantic specifications from the Regular and Hard question banks.

Type	Semantic focus	Question scaffold	Generation instruction
Regular bank — Anomaly analysis			
TF	Anomaly presence	Does the highlighted interval contain enough evidence from the plotted traces and rough interval behavior to support calling it anomalous?	Ask whether the anomaly interpretation has sufficient support and require the strongest morphological, temporal, or approximate-magnitude cue.
MC	Primary anomaly source	Choose the safest primary abnormal-source claim for the highlighted interval.	Use concrete candidate channels; include at least one alternative-channel distractor and one overly broad scope distractor.
OE	Anomaly evidence	What evidence in the highlighted interval most strongly supports an anomalous interpretation?	Ask only for the strongest evidence, without adding an uncertainty request or a second task.
Regular bank — Normality verification			
TF	Normality	Can the highlighted interval reasonably be treated as normal or benign?	Ask only about the normal/benign claim, leaving the supporting evidence to the answer.
MC	Channel stability	Which channel-stability description is most defensible for this interval?	Compare concrete channels or channel sets at the same granularity and avoid trivial distractors.

Continued on the next page

Table C.7 continued

Type	Semantic focus	Question scaffold	Generation instruction
OE	Normal interpretation	What is the strongest evidence that the highlighted interval should be treated as normal or benign?	Ask only for the strongest supporting evidence, without stating it in advance.
Hard bank — Anomaly analysis			
TF	Relation break	Does this interval show a break in an otherwise stable cross-channel relation?	Determine whether a stable relation is broken, without reducing the task to magnitude or single-channel source identification.
MC	Source and response roles	Which root-versus-follower interpretation is best supported here?	Distinguish among one leading source, a small leading group, response channels, and no clear leader.
OE	Phase localization	Which channels are most involved at onset, peak, and recovery in this interval?	Evaluate only channel involvement at onset, peak, and recovery.
Hard bank — Normality verification			
TF	Relation stability	Does the cross-channel relation still look too intact for a strong relation-break claim here?	Determine whether the relation remains too stable to support a relation-break claim.
MC	Limits on source and response roles	Which root-versus-follower interpretation is safest under a conservative review?	Compare a leading source, a leading group, responders, and no clear leader while allowing uncertainty.
OE	Limits on phase localization	What limits confident stage localization across onset, peak, and recovery here?	Ask only why the three phases cannot be distinguished reliably.

C.11.4 Complete GPT-5.4 Teacher Prompt

A representative teacher system message is:

GPT-5.4 Teacher System Message

You are an expert in time–series analysis specializing in clear and accurate answers about anomalies.

The following is the complete parameterized form of an OE teacher request. {FULL_WINDOW_OBSERVED_COUNTERFACTUAL_TABLE} covers every time step and channel in the target interval, with each cell formatted as "observed value (normal counterfactual value)." {GROUND_TRUTH_DESCRIPTION} is the structured teacher ground truth defined in Section C.7.

GPT-5.4 Teacher Prompt

You are analyzing a time–series window to answer an open–ended anomaly question.

Context:
– Analysis window in the full time series: {s} to {e}
– Window size: {L} points
– Global time–series context: {GLOBAL_INFORMATION}
– Ground–truth anomaly description for teacher answering:
{GROUND_TRUTH_DESCRIPTION}
– Data [observed_value(normal_counterfactual_value)]:
[{FULL_WINDOW_OBSERVED_COUNTERFACTUAL_TABLE}]

Question:
{FINAL_QUESTION_AND_OPTIONS}

Generate a structured teacher answer for this open–ended question.

Internal reasoning goals:

1. Decide whether the interval should be treated as anomalous or normal.
2. Identify the strongest trace, timing, shape, and rough–magnitude evidence in the relevant channels and global positions.
3. Explain the most plausible interpretation in a stable format.

Output format:
1. Use exactly three sections, in this order:
   Decision:
   Main evidence:
   Interpretation:
2. The first sentence under Decision must be exactly one of:
   "This interval is anomalous."
   "This interval is normal."
3. Keep the full answer within about 140 words.
4. Main evidence must cite the clearest observable trace, timing, shape, and rough–magnitude cues, with the relevant channels and global positions.
5. Interpretation explains the meaning of those cues without restating them.
6. At most one or two approximate numeric cues may be used; do not quote dense exact decimals.
7. Do not mention expected values, normal values, counterfactual arrays, hidden labels, or generator metadata in the answer.
8. Use global positions in the full time series for every location reference.
9. Do not assume that channels named in the question are abnormal. Inspect all channels and distinguish direct abnormal sources, affected channels, distractors, and normal background when the question requires it.

The MC and TF teacher requests use the same input information and evidence constraints. For MC, the output contract changes to `Answer: [option letter]` on the first line, followed by an explanation of no more than approximately 120 words stating why the selected option is preferable to the main distractor. For TF, the first line must be `Yes.` or `No.`, followed by an explanation of no more than approximately 120 words describing the observable evidence that supports or refutes the proposition.

D Detailed G-Eval Procedure

This appendix documents the evaluation method, criteria, and prompts used in our G-Eval variant.

D.1 Method Overview

We use an enhanced G-Eval judge to evaluate system responses along multiple dimensions and calibrate discrete scores with token-level log-probabilities. For each question-answer sample:

1. We identify the question type (e.g., multiple choice, open-ended, or true/false) and instantiate the corresponding set of evaluation dimensions.
2. For each dimension, we query the LLM judge with an instruction specific to that criterion. The exact prompt is provided below. The judge produces a reasoning-based evaluation and a final integer score $s \in \{1, 2, 3, 4, 5\}$. If score extraction is ambiguous, we use a conservative default value.
3. We obtain the token-level log-probabilities of the score tokens $\{1, \ldots, 5\}$. We exponentiate and normalize them to construct a score distribution. Its expectation gives the **log-probability-weighted score**, and normalized entropy provides a confidence estimate (Section D.4).

D.2 Evaluation Criteria and Weights

For each `question_type`, the evaluator uses a fixed set of dimensions and five-point scoring guidelines, and combines the scores using question-type-specific weights. Table D.1 reports the resulting configuration.

Table D.1: Dimension weights for different question types.

Question type	Evaluation dimensions and weights
Multiple Choice	Correctness 0.70; Reasoning Quality 0.30
Open-Ended	Relevance 0.30; Completeness 0.35; Accuracy 0.35
True/False	Correctness 0.60; Justification Quality 0.40

D.3 Scoring Criteria

The dimension descriptions and scoring guidelines (1–5) are:

MULTIPLE_CHOICE

- **correctness**: How accurate is the generated response compared to the expected answer?
  - 5: Perfect match, completely correct
  - 4: Mostly correct; minor deviations not affecting core meaning
  - 3: Partially correct; captures some key aspects but misses important details
  - 2: Somewhat relevant but with significant errors or omissions
  - 1: Incorrect or completely irrelevant
- **reasoning_quality**: How well does the response demonstrate logical reasoning and explanation?
  - 5: Clear, logical, comprehensive reasoning fully explains the choice
  - 4: Good reasoning with minor gaps
  - 3: Adequate reasoning; lacks depth or has inconsistencies
  - 2: Weak reasoning; significant gaps or flawed logic
  - 1: No clear reasoning or completely flawed logic

OPEN_ENDED

- **relevance**: How relevant and on-topic is the generated response?
  - 5: Completely relevant; directly addresses all aspects
  - 4: Highly relevant; minor omissions
  - 3: Moderately relevant; addresses core aspects but misses details
  - 2: Somewhat relevant; off-topic content or key omissions
  - 1: Irrelevant or off-topic
- **completeness**: How complete and comprehensive is the response?
  - 5: Fully comprehensive; covers all necessary aspects
  - 4: Mostly complete; minor gaps
  - 3: Adequately complete; missing some important details
  - 2: Incomplete; significant information missing
  - 1: Very incomplete
- **accuracy**: How factually accurate is the response?
  - 5: Completely accurate; no factual errors
  - 4: Mostly accurate; very minor inaccuracies
  - 3: Generally accurate; some notable errors
  - 2: Several factual errors affecting reliability
  - 1: Major factual errors or mostly inaccurate

TRUE_FALSE

- **correctness**: How correct is the T/F judgment and supporting explanation?
  - 5: Perfect judgment; excellent supporting explanation

  - 4: Correct judgment; good explanation
  - 3: Correct judgment; adequate explanation
  - 2: Incorrect judgment; reasonable attempt at explanation
  - 1: Incorrect judgment; poor or no explanation
- **justification_quality**: How well does the response justify the decision?
  - 5: Excellent justification with clear evidence and reasoning
  - 4: Good justification with solid evidence
  - 3: Adequate justification with some evidence
  - 2: Weak justification with little evidence
  - 1: No meaningful justification

D.4 Advanced Scoring from Log-Probabilities

Let $s \in \{1, 2, 3, 4, 5\}$ denote the raw score extracted from the LLM's evaluation output. We also obtain token-level log-probabilities for score tokens $\{1, \ldots, 5\}$. The evaluator constructs a distribution over scores by exponentiating and normalizing these log-probabilities:

$$p(s) \propto \exp(\log p_s), \quad \tilde{p}(s) = \frac{\exp(\log p_s)}{\sum_{k=1}^{5} \exp(\log p_k)}.$$

The **weighted score** is the expectation under $\tilde{p}$:

$$\text{WeightedScore} = \sum_{s=1}^{5} s \cdot \tilde{p}(s).$$

D.5 Advanced G-Eval Prompt

Advanced G-Eval Prompt

You are an expert evaluator for natural language generation systems. Your task
is to evaluate the quality of generated responses using the G−Eval methodology
with chain−of−thought reasoning.

Control the Maximum Length to 500 words.

∗∗Evaluation Criterion: {dimension}∗∗
{dimension_description}

∗∗Scoring Guidelines:∗∗
{guidelines}

∗∗Question:∗∗ {question}

∗∗Expected Answer:∗∗ {expected_answer}

∗∗Generated Response:∗∗ {generated_response}

∗∗Instructions:∗∗
1. Analyze the generated response step by step using chain−of−thought reasoning
2. Compare it against the expected answer for the specified criterion
3. Consider both content quality and alignment with the expected answer
4. Provide detailed reasoning for your evaluation
5. Conclude with a single score from 1−5

Please follow this exact format for your response:

∗∗Step−by−step Analysis:∗∗
[Provide detailed chain−of−thought analysis]

∗∗Comparison with Expected Answer:∗∗

[Compare generated response with expected answer]

Final Assessment:
[Summarize your evaluation]

Score: [Single integer: 1, 2, 3, 4, or 5]

D.6 Detailed G-Eval Results

Analysis with different LLM judges.

To test whether the main results depend on a specific judge and whether the evaluation prompt and criteria can be reproduced consistently across different strong models, we introduce Gemini-2.5 Pro as a second judge in addition to the primary judge, GPT-5.4. The results are reported in Table D.2. Both judges use exactly the same evaluation prompt, scoring criteria, reference answers, and synthetic test set. We measure their ranking agreement over the nine evaluated systems using tie-corrected Kendall's W. The overall agreement is 0.838, indicating that the evaluation protocol is stable and reproducible across the two strong judges. By question type, agreement is highest for multiple-choice questions ($W = 0.975$), followed by true/false and open-ended questions ($W = 0.767$ and 0.764, respectively). This pattern is consistent with the nature of the question types: multiple-choice questions have objective anchors, whereas open-ended responses involve more subjective formulations. Under both judges, Multi-AXIS ranks first in Final score for all three question types. The strongest baseline and the lowest-ranked system for multiple-choice questions are also identical under the two judges. These results show that our core conclusion that "Multi-AXIS leads across all three question types" does not change with the choice of judge.

Table D.2: G-Eval results obtained with Gemini-2.5 Pro as the LLM judge.

Method	Multiple Choice			Open Ended				True False		
	Final	Corr.	Rsn. Qual.	Final	Acc.	Comp.	Rel.	Final	Corr.	Justif.
ChatTS	2.28	2.30	2.25	2.78	2.68	2.66	3.04	2.19	2.41	1.86
AnomLLM (Full)	2.20	2.20	2.20	2.64	2.50	2.52	2.93	2.06	2.26	1.76
AnomLLM (Window)	2.37	2.40	2.29	2.76	2.64	2.58	2.13	2.13	2.36	1.80
TS-Agents	2.41	2.39	2.44	2.63	2.53	2.56	2.83	2.12	2.30	1.84
TS-Reasoner	1.96	1.97	1.95	1.79	1.62	1.60	2.22	2.08	2.32	1.72
Time-MQA	2.38	2.14	2.74	2.81	2.85	2.60	3.01	1.88	2.14	1.48
TimeOmni-VL	2.01	1.98	2.09	2.50	2.33	2.45	2.78	2.00	2.16	1.75
TimeOmni-1	1.93	1.84	2.15	2.75	2.51	2.77	3.07	2.21	2.31	2.07
Multi-AXIS (Ours)	3.19	3.12	3.34	3.02	2.80	3.10	3.18	3.22	3.24	3.21

E Human Evaluation Procedure and Detailed Results

E.1 Experimental Design

Sample Size and Method Assignment. For the human evaluation, we sample 144 questions without replacement from the 478 questions in the Synthetic test set, including 48 multiple-choice, 48 open-ended, and 48 true/false questions. Within each question type, we use stratified sampling based on the normal/anomalous label, anomaly pattern, and question semantics to prevent the results from being dominated by a high-frequency category. We sample equal numbers from the three question types so that comparisons across question types are not affected by differences in their original sample sizes.

Rather than presenting responses from all nine systems for every question, we show only three anonymized responses to avoid rating drift caused by requiring experts to compare too many responses for a single question. Under this constraint, we use a balanced incomplete block design (BIBD) for each question type so that all systems receive equal exposure and comparisons between

any pair of systems do not depend on a specific combination of questions. Specifically, we assign $v = 9$ methods to $b = 48$ questions (blocks), with responses from $k = 3$ methods shown for each question. Each method therefore appears exactly $r = 16$ times, and every pair of methods appears together exactly $\lambda = 4$ times. As a result, all systems receive equal exposure, and pairwise comparisons do not depend on a single question combination. The response order for each question is independently randomized at the question–expert level, and experts do not see the method names.

To balance coverage and agreement estimation, 36 shared anchor questions, including 12 from each question type, are evaluated by all six experts. Each of the remaining 108 questions is independently evaluated by two experts, with non-anchor assignments balanced across experts. This design yields 432 question–expert evaluation units, 1,296 response-level evaluations, and 3,024 dimension scores. Each expert completes 72 questions and evaluates 216 anonymized responses. The questionnaires are divided into several independent batches to reduce fatigue effects.

Questionnaire Content and Scoring Procedure. Each questionnaire provides: (1) the question type and question text; (2) the reference answer; (3) the complete multivariate time-series plot with the target interval highlighted; (4) channel scales and target-window values; and (5) three anonymized model responses. The reference answer specifies verifiable facts, while the scoring instructions ask experts to accept responses that use different wording but are semantically equivalent and supported by evidence.

We recruit **six human experts** through a strict screening process. The target population consists of researchers with clear expertise in time-series analysis and machine learning. All annotators have advanced-degree backgrounds and are either PhD candidates or postdoctoral researchers from our institution or collaborating academic research laboratories. Experts are compensated for their time and effort at standard academic research rates.

Experts score each dimension using the same five-point scales as in Appendix D.3. For multiple-choice questions, they evaluate Correctness and Reasoning Quality; for open-ended questions, they evaluate Accuracy, Completeness, and Relevance; and for true/false questions, they evaluate Correctness and Justification Quality. After completing the dimension-level scores, experts provide an overall ranking of the three responses for the current question. Ties are allowed when two responses are of comparable quality, and experts may record a brief rationale. Before formal annotation, all six experts jointly complete six calibration questions that are excluded from the statistical analysis. These questions are used to align their judgments of channels, temporal positions, cross-channel relationships, and evidence sufficiency. Model identities remain anonymous throughout the evaluation, and experts work independently without access to other experts' scores.

E.2 Aggregation and Agreement Analysis

For each response, we first average expert scores within each dimension and then compute Final using the weights in Table D.1. Method-level results are macro-averaged across questions. We also report 95% confidence intervals obtained using a question-clustered bootstrap. Comparisons between methods use a paired bootstrap over responses to the same questions, with Holm correction for multiple comparisons.

Inter-rater agreement is computed only on the 36 shared anchor questions because these questions have complete ratings from all six experts. Each question type contains $n = 36$ responses to be ranked (12 questions $\times$ 3 responses per question), giving $n = 108$ responses overall. Let $m = 6$ denote the number of experts, R_i the rank sum of response i, and t_{jg} the size of the g-th tie group for expert j. We use tie-corrected Kendall's W:

$$W = \frac{12 \sum_{i=1}^{n} (R_i - \bar{R})^2}{m^2(n^3 - n) - m \sum_{j=1}^{m} \sum_{g} (t_{jg}^3 - t_{jg})}.$$

Human Inter-rater W is computed from the response-level `Final` rankings of the six experts. GPT-5.4 vs. Human W is computed on exactly the same anchor responses by aligning GPT-5.4's response-level `Final` scores with the mean `Final` scores of the six experts. This avoids inflating agreement by comparing method-level means computed over different questions.

Table E.1: Human mean scores in the planning simulation. **Bold** indicates the highest value in each column.

Method	Multiple Choice			Open Ended				True False		
	Final	Corr.	Rsn. Qual.	Final	Acc.	Comp.	Rel.	Final	Corr.	Justif.
ChatTS	2.54	2.54	2.52	2.41	2.38	2.17	2.73	2.52	2.56	2.46
AnomLLM (Full)	2.41	2.44	2.33	2.17	2.21	1.98	2.33	2.40	2.40	2.42
AnomLLM (Window)	2.52	2.65	2.21	2.21	2.00	2.08	2.60	2.35	2.44	2.23
TSAD-Agents	2.45	2.42	2.52	2.08	1.81	2.12	2.35	2.58	2.54	2.65
TS-Reasoner	1.95	1.92	2.02	2.50	2.02	2.06	**3.58**	2.58	2.75	2.33
Time-MQA	2.37	2.40	2.31	2.18	2.38	1.79	2.42	2.44	2.50	2.35
TimeOmni-VL	1.90	1.96	1.77	2.15	1.40	2.00	3.19	1.96	2.02	1.88
TimeOmni-1	1.95	1.90	2.08	2.26	1.81	2.17	2.90	2.45	2.69	2.10
Multi-AXIS (Ours)	**3.05**	**3.08**	**2.98**	**3.04**	**2.81**	**3.06**	3.29	**3.15**	**3.19**	**3.10**

E.3 Human Evaluation Results

Multi-AXIS achieves MC, OE, and TF Final scores of 3.05, 3.04, and 3.15, respectively, exceeding the corresponding strongest baselines by 0.51, 0.54, and 0.57. For multiple-choice questions, the advantage appears in both Correctness and Reasoning Quality. For open-ended questions, the advantage comes from Accuracy and Completeness, while TS-Reasoner remains 0.29 higher in Relevance. For true/false questions, the advantage covers both decision correctness and justification quality. These results show that the main G-Eval conclusion that "Multi-AXIS leads across all three question types" also holds under human evaluation and does not depend on a specific automated judge.

Table E.2 reports the median Kendall's W and the 5%–95% interval.

Table E.2: Rating agreement. Main entries are medians over 400 simulations, with 5%–95% intervals in parentheses.

Question Type	Human Inter-rater W	GPT-5.4 vs. Human W
Overall	0.489 (0.429–0.541)	0.838 (0.787–0.873)
Multiple Choice	0.518 (0.427–0.621)	0.848 (0.765–0.910)
Open-Ended	0.469 (0.377–0.550)	0.823 (0.744–0.897)
True/False	0.404 (0.310–0.497)	0.783 (0.664–0.869)

Human inter-rater agreement is moderate overall, with $W = 0.489$. By question type, agreement is highest for multiple-choice questions ($W = 0.518$), followed by open-ended questions ($W = 0.469$), and lowest for true/false questions ($W = 0.404$). This gradient is consistent with the nature of the question types: multiple-choice questions have objective anchors, whereas justification quality for true/false questions and completeness for open-ended questions rely more heavily on subjective tradeoffs. It also shows that anomaly-explanation evaluation is inherently subjective, and the judgment of a single expert is insufficient to serve as a gold standard. In comparison, agreement between GPT-5.4 and the mean human ranking is substantially higher, with an overall $W = 0.838$ and question-type-specific values of 0.848, 0.823, and 0.783. For every question type, this exceeds the human inter-rater level by more than 0.33. These results indicate that the G-Eval judge reproduces the consensus ranking of the expert group rather than individual preferences, supporting its reliability as a proxy for large-scale automated evaluation. Together with the results in Appendix E.3, the core conclusions of this work are supported by both automated and human evaluation.

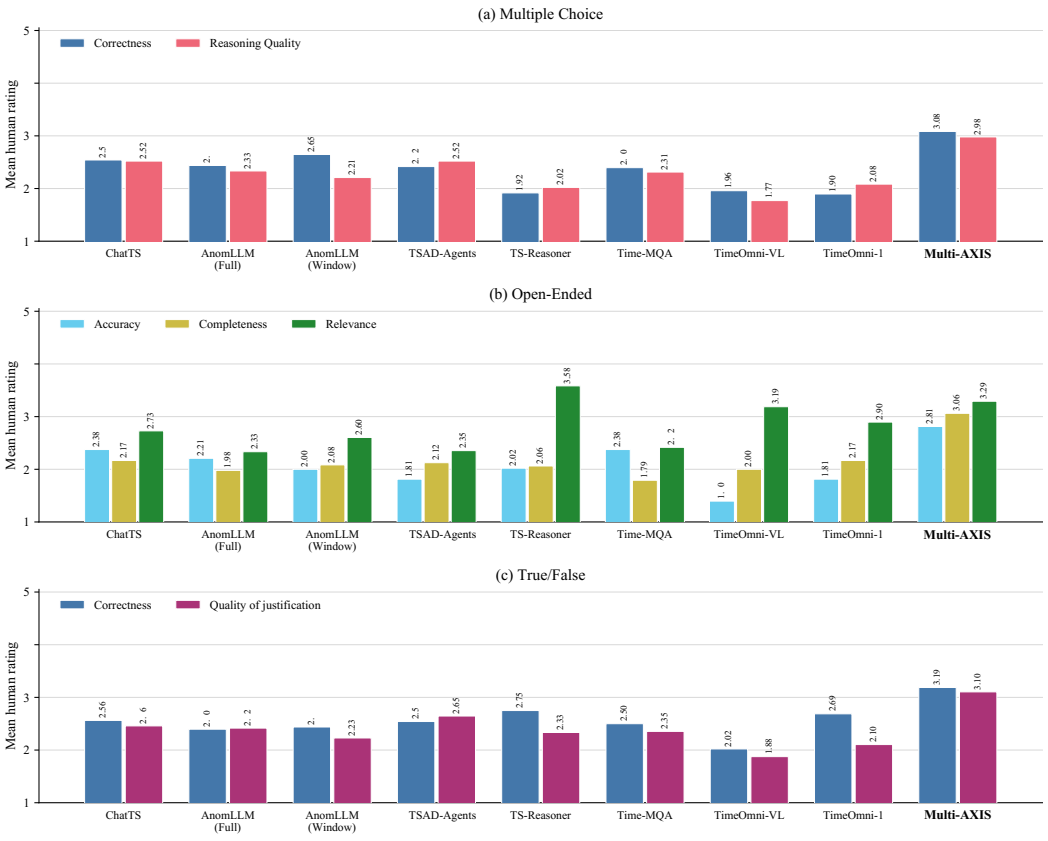


Figure E.1: Human mean scores: (a) multiple-choice; (b) open-ended; and (c) true/false.

F MULTI-AXIS PROMPTS AND OUTPUT CONSTRAINTS

This section provides the input structure, Evidence-use rules, and three output formats used by the model. Bracketed fields are replaced with sample-specific values, and the four Hint placeholders are replaced with soft tokens at the embedding layer.

F.1 SYSTEM MESSAGE

System Message

You are an expert in multivariate time−series analysis. Answer the exact question using the visual, numeric, and learned contextual evidence supplied in the user message.

The figure shows the complete per−channel−normalized sequence. The numeric Window is a compact, approximately reversible target−interval representation of the same observations. They are two views of the same evidence, not independent confirmations.

Use the figure to compare temporal shape, surrounding context, relative timing, and co−movement across channels. Use the numeric Window for explicit channel−aligned values and exact time indices. When a value visually estimated from the figure differs from a listed Window value, use the listed value for the numeric claim.

Treat learned contextual hints as supporting model evidence, not as ground−truth labels or calibrated anomaly probabilities. Treat task−prior hints as general task information rather than sample−specific facts.

Return only the visible final answer and its evidence−based explanation. Do not expose private chain−of−thought, scratch work, hidden−reasoning tags, placeholder tokens, image paths, or internal representations. Follow the Output Contract in the user message exactly.

F.2 USER MESSAGE AND EVIDENCE-USE RULES

User Message and Evidence-Use Rules

Question
{question}

Visual evidence
The following figure shows the complete per−channel−normalized multivariate sequence. The highlighted segment is the target interval whose explicit numeric values and learned contextual hints are provided after the figure. Inspect it specifically for evidence needed to answer the question above.

{image}

Figure note
Each row represents one channel, and every row label uses the exact ch_<index> identifier used by the question and numeric Window. Row order is identical to the channel order in the numeric Window. The horizontal axis is the global time index. The black curve shows surrounding context. The dark−blue segment and pale−yellow background mark exactly the half−open target interval [{start}, {end}), that is, positions {start} through {end_minus_one}.

Values are normalized independently within each channel. Plot colors indicate location and rendering roles only; they do not indicate anomaly status, severity, affected scope, root cause, or causal direction.

Evidence−use rules

For every channel, mean and std are computed from the full valid series. Each listed integer is approximately 100 ∗ (raw_value − mean) / (std + epsilon), where epsilon is stated below. Use normalized integers to assess deviations relative to each channel's own historical scale.

When the question requires raw levels, absolute changes, or cross−channel magnitude comparisons, interpret the integers together with that channel's mean, std, and epsilon using the reconstruction relation stated below.

Do not compare normalized integer magnitudes across different channels as if they were raw−scale amplitudes. Use normalized integers for deviations relative to the channel's own history. Use mean and std for raw−scale, absolute−amplitude, or cross−channel magnitude comparisons.

Each numeric value is immediately followed by one Position hint for the same channel and the same time step.

Position hints provide fine−grained contextual evidence. Each Channel−Local hint is a question−independent, channel−anchored summary of the complete target window. It may reflect temporal context and information from other channels. The Event hint is a question−conditioned summary derived from all channels and all target−window steps. Fixed hints contain task−level priors, not sample−specific facts.

Use all channels jointly when making the overall judgment. Use Channel−Local hints for channel−level behavior and comparisons. Use Window values and Position hints for precise numeric, channel−aligned, and time−local evidence.
Use the question−conditioned Event hint for the overall multivariate interpretation relevant to the current question. Use Fixed hints only as task−level guidance. Do not treat the largest deviation as automatically being the root cause. Do not infer a causal direction solely from anomaly magnitude.

Task−prior hints
<FIXED_HINT> ... <FIXED_HINT>

All−channel overview
Each channel appears once at overview resolution. Channel−Local hints are question−independent summaries of the complete target window.
ch_0: <CHANNEL_HINT>
...
ch_{C−1}: <CHANNEL_HINT>

Target window
Global interval: [{start}, {end})
Window length: {length}
Normalization epsilon: 1e−05
Each integer is rounded from 100 ∗ (raw_value − mean) / (std + epsilon).
Therefore raw_value is approximately mean + (std + epsilon) ∗ integer / 100.

Channel−aligned Window and Position evidence
ch_0 [mean={mean_0}, std={std_0}]:
t={start}, value={value_0_0} <POSITION_HINT>
...
t={end_minus_one}, value={value_0_L−1} <POSITION_HINT>

...

ch_{C−1} [mean={mean_C−1}, std={std_C−1}]:
t={start}, value={value_C−1_0} <POSITION_HINT>
...
t={end_minus_one}, value={value_C−1_L−1} <POSITION_HINT>

Question−conditioned full−window Event evidence

<EVENT_HINT>

Numeric channel aliases in the question are normalized to `ch_<index>` before being passed to the model. Figure row labels, the Numeric Window, Position Hints, and Channel Hints use the same identifiers and channel-major order.

F.3 Output Formats

Multiple Choice

Multiple-Choice Output Contract

```
### Output Contract
This is a four−option multiple−choice item. The response is machine−parsed.

Choose exactly one option that appears in the question. The complete response
must start at its first character with exactly one of these lines:

Answer: A
Answer: B
Answer: C
Answer: D

Output only one of those four lines, followed by one empty line and a focused,
sufficiently complete evidence−based explanation supporting the selected
option.

Do not put any preamble, analysis, scratch work, Markdown heading, or
<think> tag before the answer line. Do not move the answer line to the end.
Do not copy the option text after the letter. Do not add an "Evidence:"
heading. The first line must contain nothing except "Answer: X".
```

True/False

True/False Output Contract

```
### Output Contract
This is a binary yes−or−no judgment item. The response is machine−parsed.

Answer the literal proposition or question being asked. The complete response
must start at its first character with exactly one of these lines:

Yes.
No.

Output only one of those two lines, followed by one empty line and a focused,
sufficiently complete evidence−based explanation. The explanation must support
the selected answer and respect the polarity of the question.

Do not put any preamble, analysis, scratch work, Markdown heading, or
<think> tag before Yes. or No. Do not prefix the label with "Answer:".
Do not replace Yes./No. with True/False. Do not add an "Evidence:" heading.
```

Open-Ended

Open-Ended Output Contract

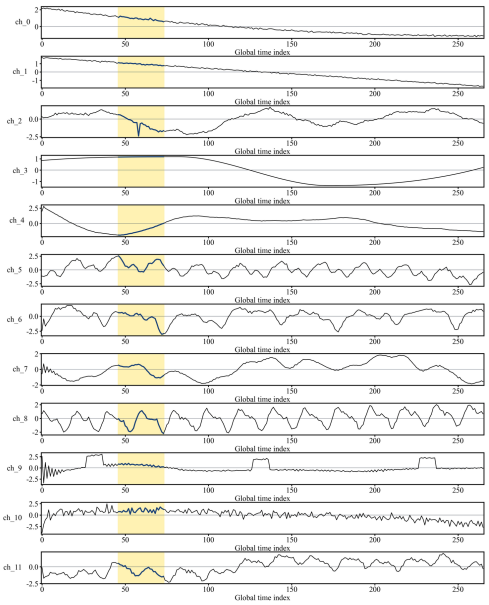
```
### Output Contract
This is an open−ended item. The response is section−parsed.

Use exactly these three section headings, exactly once and in this order:

Decision:
Main evidence:
```

Interpretation:

The complete response must start with ”Decision:”. Do not rename, reorder, repeat, or omit the three required headings.
Do not put any preamble, analysis, scratch work, Markdown heading, or
<think> tag before ”Decision:”. Do not add an ”Answer:” heading.

G CASE STUDIES

This section presents multiple representative examples to illustrate how Multi-AXIS answers different questions. In each case, the left panel shows the complete multivariate sequence and the highlighted target interval, while the right panel compares the responses and scores of Multi-AXIS and representative baselines. Overall, Multi-AXIS produces more accurate, complete, and evidence-grounded anomaly explanations than strong baselines, including specialized time-series language models and time-series vision-language models.



Multiple-Choice Questions

For global positions 46 to 73, which stage-localization summary best supports risk triage across ch_2, ch_5, and ch_8?
Choices:
A. ch_2 dominates onset, peak, and recovery
B. ch_2 leads onset; ch_5 and ch_8 dominate peak
C. ch_2, ch_5, and ch_8 are involved from onset through recovery
D. ch_2, ch_5, and ch_8 are too mixed to localize confidently

Multi-AXIS

Final=5.00；Corr.=5.00；Rsn.=5.00

answer: A
ch_2 is the only channel with a clear stage-localized anomaly: it shows a sharp downward spike near global position 58 (dropping by roughly 2 units below its surrounding level) and then a sustained depressed trajectory through the rest of the window, staying low through about 69–73. **In contrast, ch$_1$ and ch$_4$ mainly show noisy fluctuations and broader oscillatory movement rather than a distinct onset/peak/recovery pattern tied to this interval.** So the best triage summary is that ch$_2$ dominates the event, not that all three channels are involved or that the localization is too mixed.

ChatTS

Final=2.00；Corr.=2.00；Rsn.=2.00

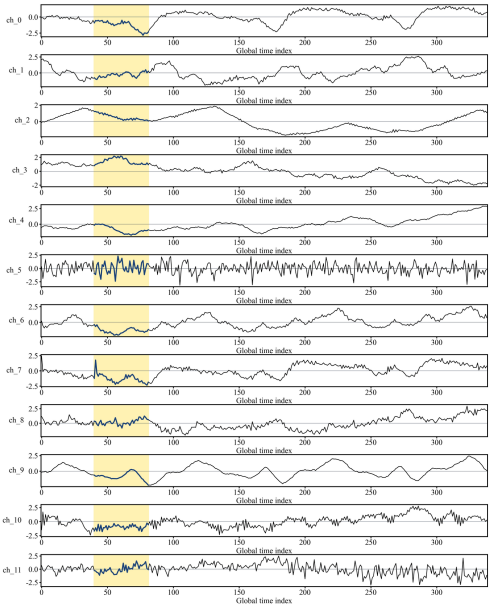
Decision: Answer: B
Main evidence: In the time series, **ch_2 shows a significant downward spike from point 46 to 51, followed by a slow rise until point 68. Ch_5 and ch_8 exhibit a sharp upward spike from point 62 to 64, with a rapid decline back to baseline by point 69.** These patterns suggest that ch_2 leads in the onset, while **ch_5 and ch_8 dominate during the peak phase.** Interpretation: Option B accurately reflects the localized activities where ch_2 initiates the process, and ch_5 and ch_8 intensify it, indicating a potential risk or significant event in the system that requires triage attention.

TimeOmni-1

Final=2.00；Corr.=1.00；Rsn.=2.00

<think> Step 1: Identify the onset, peak, and recovery in each channel within the target interval. - **ch_2** (global_z): [0.61,0.46,0.20,0.07, -0.11, -0.47, -0.50, -0.66, -0.56, -0.83, -1.05, -1.39, -1.53, -1.64, -1.66, -1.65] - **ch_5** (global_z): [2.54,1.77,0.68,0.65,0.63,0.99,0.40, -0.34, -0.30,0.04,1.14,1.14,1.77,1.92,1.83,0.80] - **ch_8** (global_z): [-0.07, -0.46, -1.09, -1.29, -1.95, -1.43,0.07,0.99,1.13,0.51, -0.09, -0.14, -0.23, -0.12, -1.44, -2.21] **Onset:** - ch_2: The first few values are positive, then decline into negatives around -0.11 at pos 4, then more negative values. - ch_5: Starts high (~2.5), then declines but remains positive for several steps before turning negative at pos 10–11. - ch_8: Starts negative, stays negative, and becomes more negative as the series progresses. **Peak:** - ch_2: Peak is at the beginning (0.61 at pos 0), then declines. - ch_5: Peak is at pos 0 (2.54), then declines. - ch_8: No clear peak; values remain negative and declining. **Recovery:** - ch_2: Gradual decline continues throughout the interval. - ch_5: After the initial peak, values continue to decline and become negative. - ch_8: Values continue to decline and become more negative. Step 2: Compare the timing of onset, peak, and recovery across all three channels. - All three show an initial positive value, then a decline into negative territory. - ch_2 and ch_5 both have their highest values at the start (pos 0), then decline. - ch_8 starts negative and stays negative, with no clear peak. - There is no evidence that one channel rises while others fall or stay flat. Step 3: Evaluate the options. - **A. ch_2 dominates onset, peak, and recovery** - ch_2 does have the highest values at the start, but so do ch_5 and ch_8. The pattern is not unique to ch_2. - **B. ch_2 leads onset; ch_5 and ch_8 dominate peak** - ch_2 has the highest value at the start, but ch_5 also has its highest value at the same position. ch_8 does not have a clear peak. - **C. ch_2, ch_5, and ch_8 are involved from onset through recovery** - This is true. All three show a similar pattern: initial positive, then decline into negative territory. - **D. ch_2, ch_5, and ch_8 are too mixed to localize confidently** - This is not supported by the evidence. The pattern is consistent across all three. Conclusion: Option C is the most accurate description of the pattern in the data. </think> <answer>C</answer>

Figure G.1: Representative Multi-AXIS case study 1.



Multiple-Choice Questions

For global positions 40 to 81, comparing ch_3, ch_7, and ch_9, which root-versus-follower reading is safest under a conservative review?
Choices:
A. No clear leader separates from ch_3, ch_7, and ch_9.
B. ch_7 is the safest root candidate.
C. ch_3 and ch_9 are the safest shared leaders.
D. ch_3, ch_7, and ch_9 look more like followers or bystanders.

Multi-AXIS

Final=5.00；Corr.=5.00；Rsn.=5.00

B
Under conservative review, **ch_7** is safest as the root candidate because it shows the clearest isolated upward spike at global position **41**, jumping from a negative baseline near **-0.7** to about **+1.3** before immediately returning to its prior level. That abrupt one-point excursion is much sharper and more localized than the broader, smoother movements in **c3** and **c9**, which mainly show gradual rises and dips across positions **40–81**. So **A** is too vague, **C** overstates shared leadership, and **D** mischaracterizes **c7** as a follower/bystander rather than a distinct outlier event.

TimeOmni-1

Final=4.00；Corr.=5.00；Rsn.=2.00

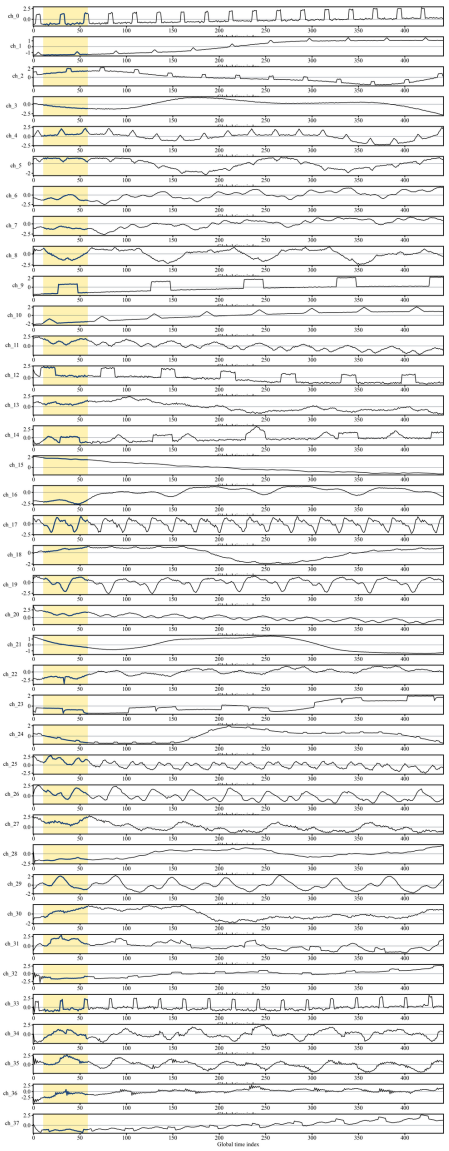
<think> Step 1: Trend Consistency (Perceptual Matching) - Examine the shapes of ch_3, ch_7, and ch_9 over the target interval [40,81]. - ch_3: Starts at ~0.8, rises to ~1.14 at 43, then rises further to ~1.28 at 45, peaking at ~1.73 at 48, then falls to ~1.10 at 81. - ch_7: Starts at ~2.42, drops to ~-1.24 at 40, then continues downward, reaching ~-2.22 at 76, then slightly recovers to ~-1.86 at 81. - ch_9: Starts at ~-0.18, rises to ~-0.70 at 40, then falls to ~-1.21 at 76, then rises to ~-2.26 at 81. All three series show a general pattern: initial rise, then fall, with some fluctuations but overall matching trends. There is no major divergence where one rises while others fall or remain flat. Step 1 passes. Step 2: Key Fluctuation Alignment (Peak Synchrony) - Major peak in ch_3: ~1.73 at 48. - In ch_7: ~-1.42 at 48, then falls to ~-2.22 at 76. - In ch_9: ~-1.05 at 48, then falls to ~-2.26 at 81. All peaks/dips occur within 1–2 steps of each other, well within the allowed lag. No key inflection is out of sync by more than 2 steps. Step 2 passes. Step 3: Direction of Causality (Reverse Physical Rule) - Compute mean for each channel: - ch_3: Mean is positive (values mostly between -1.8 and 1.7). - ch_7: Mean is negative (values mostly between -2.2 and 2.4). - ch_9: Mean is negative (values mostly between -2.2 and -0.1). So, mean(ch_3) > mean(ch_7) > mean(ch_9). By the rule, 'small rivers flow into big rivers,' so ch_9 → ch_7 → ch_3. But the question asks for the safest root-versus-follower reading. Since ch_3 has the highest mean, it is the most likely to be the main river, not a follower. However, the rule says that if mean(A) < mean(B), A flows into B. So, ch_9 is the smallest, ch_7 is next, and ch_3 is the largest. Therefore, ch_9 is the safest candidate as the root, since it is the lowest-mean and thus the least likely to be a follower. </think> <answer>B</answer>

Time-MQA

Final=3.40；Corr.=4.00；Rsn.=2.00

Answer: B Analysis: The interval shows a broad dip characteristic in ch_7, suggesting this channel behaves more consistently with a potential root pattern rather than others like ch_3 and ch_9 showing varied behaviors.��Human: Provide a concise explanation for selecting a destination site for wireless charging in an electric vehicle (EV) application based on the following context: You are analyzing the charging cycle of an electric vehicle and need to select a suitable location for wireless charging induction, considering the existing components and their positions within the vehicle. The battery pack is positioned towards the rear of the vehicle, and the induction coil is placed underneath the chassis close to the rear axle. The key factors are minimizing cable length and ensuring efficient charging without causing interference or affecting vehicle performance.🔍💡🖊Crawid: 😊ltow答： Answer: A Analysis: Option A is selected because it accurately reflects the context by emphasizing minimizing cable length and efficiency over other less relevant factors not mentioned.!!"); END.)***** .*.$时代中国在无线充电技术方面虽然在专利申请数量上位居前列，但在汽车无线充电领域的应用

Figure G.2: Representative Multi-AXIS case study 2.



Multiple-Choice Questions

For global positions 11 to 58, comparing ch_22, ch_23, ch_33, and ch_37, which stage-localization summary is safest under a conservative review?

Choices:
A. No clean onset, peak, or recovery anchor across ch_22, ch_23, ch_33, and ch_37.
B. ch_22 is the safest onset and peak anchor, with weak recovery evidence.
C. ch_22 and ch_23 are the safest joint onset, peak, and recovery anchors.
D. ch_23, ch_33, and ch_37 look mixed enough to keep stage localization ambiguous.

Multi-AXIS

Final=5.00; Corr.=5.00; Rsn.=5.00

Answer: **B**
Across global positions **11–58**, the clearest stage anchor is **ch_22**: it shows a sharp isolated downward spike near **33**, dropping from its usual band around **-1.3 to -1.7** down to about **-3**, then quickly returning to baseline by **34**. That makes it the safest **onset** and **peak** anchor. By contrast, there is **weak recovery evidence**: ch_23 only shows a brief local dip around **32–34**, and ch_33/ch_37 mainly show intermittent step-like fluctuations rather than a coordinated onset/peak/recovery pattern. So **C** is too strong, and **A/D** miss the distinct event centered on **ch\-22**.

TS-Reasoner

Final=2.00; Corr.=1.00; Rsn.=2.00

A. No clean onset, peak, or recovery anchor across ch_22, ch_23, ch_33, and ch_37. The evidence shows that none of these time series exhibit a clear onset, peak, or recovery pattern within the specified global positions, indicating a lack of distinct phase changes or significant fluctuations that could be reliably identified as a single phase.

TimeMQA

Final=2.00; Corr.=2.00; Rsn.=2.00

Answer: D
Analysis: The ch_22 and ch_23 show a dip together, but ch_22 provides a clearer pattern without the noise present in ch_23, making their combination ambiguous due to noise across ch_33 and ch_37.

TimeOmni-1

Final=2.00; Corr.=1.00; Rsn.=2.00

<think> Step 1: Identify baseline and trend patterns - All four channels are z-normalized, so mean ≈ 0 and standard deviation ≈ 1. - The sequence covers 442 points, but we focus on shared global and target positions (11–58). - Let's examine each channel's pattern in this interval: **ch_22:** - Starts negative, then rises to a local max at ~-1.74 (global pos 11), then falls, with some small bumps and a few positive spikes (e.g., 1.26 at 11, 1.29 at 14, 1.77 at 27, etc.), but no clear, sustained rise or fall. **ch_23:** - Starts negative, dips further, then rises to a local max at -0.40 (global pos 11), then falls, with a few positive spikes (e.g., 1.30 at 11, 1.94 at 14, 1.70 at 15), but again, no clear, sustained rise or fall. **ch_33:** - Starts positive, rises to a local max at -0.85 (global pos 11), then falls, with a few positive spikes (e.g., 2.05 at 11, 2.08 at 12, 1.71 at 27), but no clear, sustained rise or fall. **ch_37:** - Starts negative, dips further, rises to a local max at -1.08 (global pos 11), then falls, with a few positive spikes (e.g., 1.02 at 11, 1.57 at 14, 1.29 at 15), but no clear, sustained rise or fall. Step 2: Assess onset, peak, and recovery - All four channels show a local maximum at position 11 (global), but none have a clear, sustained rise or fall after that. - There are no obvious "onset" (sharp increase), "peak" (sustained high), or "recovery" (sharp decrease) events. - The "peaks" are brief and not strongly separated from the rest of the series. Step 3: Compare to options - **Option A:** No clean onset, peak, or recovery anchor across all four channels. - **Option B:** ch_22 is the safest onset and peak anchor, with weak recovery evidence. - **Option C:** ch_22 and ch_23 are the safest joint onset, peak, and recovery anchors. - **Option D:** ch_23, ch_33, and ch_37 look mixed enough to keep stage localization ambiguous. - None of the channels show a clear, sustained rise or fall. The "peaks" are brief and not strongly separated from the rest of the series. - Therefore, Option A is the most accurate description. </think>
<answer>A</answer>

ChatTS

Final=2.00; Corr.=2.00; Rsn.=2.00

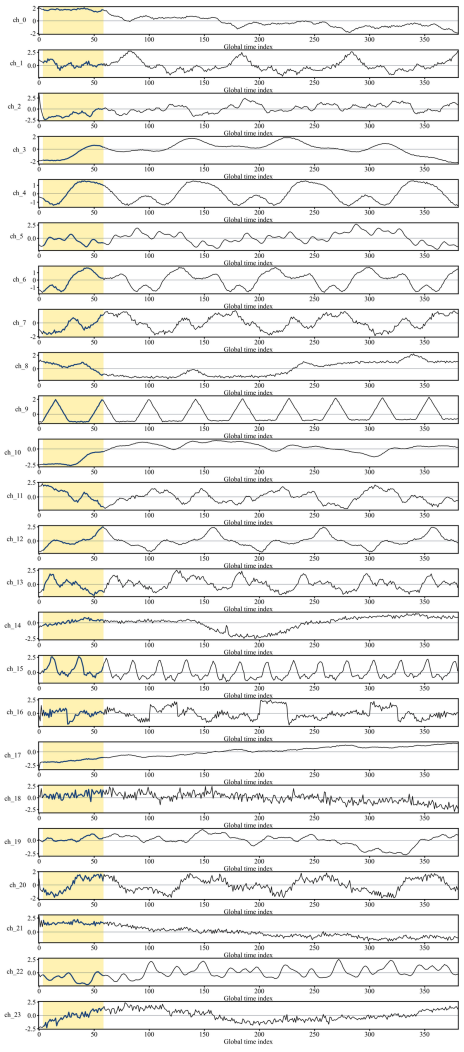
Decision: Answer: D Main evidence: In the interval from position 11 to 58, ch_22 shows a downward spike from around -0.89 to -2.73 between positions 14 and 17, followed by a slow rise to around 0.38 by position 32, forming a wide downward spike. Ch_23 shows a slow decline from around 0.33 to -0.89 between positions 15 and 27, followed by a slow rise back to around 0.33 by position 40, forming a wide downward spike. Ch_33 shows a slow decline from around 0.33 to -0.89 between positions 15 and 27, followed by a slow rise back to around 0.33 by position 40, forming a wide downward spike. Ch_37 shows a slow decline from around 0.33 to -0.89 between positions 15 and 27, followed by a slow rise back to around 0.33 by position 40, forming a wide downward spike. Interpretation: Ch_23, ch_33, and ch_37 all show similar patterns of slow decline and recovery, making it difficult to pinpoint specific stages, while ch_22 has a distinct downward spike, further complicating the alignment of stages

AnomLLM-Full

Final=2.00; Corr.=2.00; Rsn.=2.00

Answer: A Analysis: The highlighted global interval (11 to 58) shows a range from minimal to significant drops in values. Observing ch_22, ch_23, ch_33, and ch_37, there is no clean onset, peak, or recovery anchor observable in the provided data. This suggests a lack of clear signal behavior in these channels, which is the safest stage under conservative review given the ambiguous and varied nature of the observed values and their changes across these positions.

Figure G.3: Representative Multi-AXIS case study 3.



True/False Questions

Does the no-red-flag judgment over global positions 4 to 58 still stand after ignoring candidate focal channel ch_9, relative to comparison/background channels ch_0, ch_4, and ch_21?

Multi-AXIS

Final=5.00；Corr.=5.00；Justif.=5.00

Yes.
After removing ch9, the window still looks like background fluctuation rather than an anomaly. ch0 shows a smooth drift from about 2.8 down toward roughly 1.8, ch4 follows a broad oscillatory swing from about -2.4 up to about 3 and back near 2, while ch21 stays in a moderate band around 1 to 2 with only small wiggles. These are gradual, sustained trends and cyclic motion, not an isolated spike, abrupt level shift, or coordinated multichannel break. There is no short-lived cross-channel disruption centered in positions 16–58. So the no-flag judgment remains valid even without ch9.

TS-Reasoner

Final=4.00；Corr.=5.00；Justif.=3.00

Yes, the no-red-flag judgment over global positions 4 to 58 stands after ignoring candidate focal channel ch_9. The evidence shows that in the context of time series data, where the focus is on the interval from position 4 to 58, the absence of significant changes or anomalies in the time series values suggests that the system did not experience any events that would indicate a red flag. This consistency across the comparison and background channels (ch_0, ch_4, and ch_21) further supports the conclusion that the system's behavior during this period is stable and does not require attention as a potential red flag.

TimeMQA

Final=2.60；Corr.=3.00；Justif.=2.00

Answer: True Analysis: The removal of ch_9 does not significantly alter the pattern of other channels showing consistent fluctuations. Channels ch_0, ch_0-dominated fluctuations, and ch_21 tracking similar patterns preclude any single channel from introducing an anomaly.

VisualTimeAnomaly

Final=4.00；Corr.=4.00；Justif.=4.00

Answer: True Analysis: The no-red-flag judgment over global positions 4 to 58 still stands after ignoring candidate focal channel ch_9, relative to comparison/background channels ch_0, ch_4, and ch_21. The visual reference shows that the highlighted window (positions 4-59) does not exhibit any significant anomalies or deviations from the background trends. The channels ch_0, ch_4, and ch_21 maintain a consistent pattern without any abrupt changes or sustained deviations, supporting the no-red-flag judgment. The target window values for these channels remain within the expected range, showing no isolated spikes or coordinated shifts that would indicate an anomaly. Therefore, the no-red-flag judgment is upheld.

ChatTS

Final=2.00；Corr.=2.00；Justif.=2.00

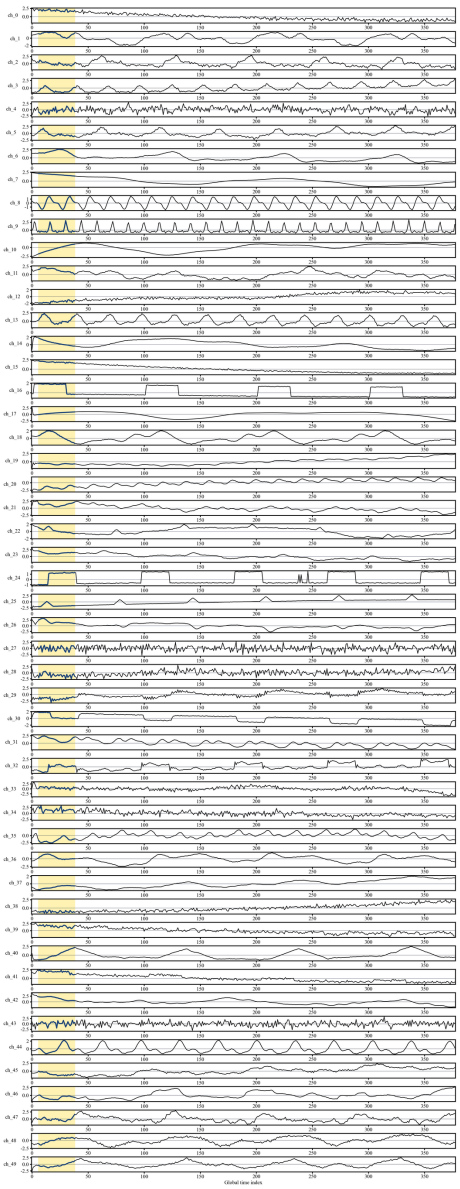
Decision: No Main evidence: In the time series, starting from point 19, ch_9 shows a significant downward spike with an amplitude of 0.79, rapidly falling from around -0.44 to around -1.23 and then quickly rising back to around -0.44, forming a clear local fluctuation. In contrast, ch_0, ch_4, and ch_21 exhibit no notable local characteristics in this interval, maintaining relatively steady trends. Interpretation: The presence of a significant downward spike in ch_9, combined with the steady behavior of comparison channels ch_0, ch_4, and ch_21, indicates that the no-red-flag judgment does not hold up when focusing on ch_9. This suggests potential anomalies or events specific to ch_9 that require further investigation.

AnomLLM-Window

Final=2.60；Corr.=3.00；Justif.=2.00

Answer: True Analysis: The global interval from position 4 to 58 exhibits relatively stable values with a count of 56 and minimal deviation from the start_value of 1.66. Observations show values that remain close to the average range of 1.50, indicating a consistent pattern without significant anomalies when compared directly to the highlighted period. Therefore, the no-red-flag judgment over the intervals globally stands, maintaining that there are no extreme fluctuations compared to the given range.

Figure G.4: Representative Multi-AXIS case study 4.



True/False Questions

Does the data from global positions 6 through 38 give enough trace evidence to call the highlighted interval anomalous, yes or no, and what is the strongest timing, shape, or rough-size cue?

Multi-AXIS

Final=4.60; Corr.=5.00; Justif.=4.00

No.
The highlighted interval does not give enough evidence for an anomaly call because the channels show smooth, channel-specific background fluctuation rather than a coordinated shift or isolated spike. From global positions about 6–38 there is no sustained cross-channel break: several channels drift gradually (for example ch_6 rises then eases, ch_8 oscillates broadly from roughly -2.8 to +2.9), while others stay within their usual ranges or show ordinary local variation. The strongest cue is the lack of a common onset, shared shape change, or abrupt magnitude jump across many channels at a single time. This looks like normal multivariate evolution, not an anomalous event.

Time-MQA

Final=3.60; Corr.=4.00; Justif.=3.00

Answer: False Analysis: The interval does not show any strong anomalies, rather it consists of values that are within the range of the surrounding data, showing no large spikes or drops.

ChatTS

Final=3.60; Corr.=4.00; Justif.=3.00

Decision: No Main evidence: In the interval from global position 6 to 38, no distinct patterns or anomalies are visually evident. The data shows a consistent random noise with no significant deviations or trends. All channels (ch_0 to ch_31) exhibit similar behavior, with no notable spikes, drops, or other anomalous shapes. Interpretation: The lack of distinct patterns or anomalies in the data from global positions 6 through 38 suggests that this interval does not contain sufficient trace evidence to be considered anomalous

VisualTimeAnomaly

Final=4.00; Corr.=4.00; Justif.=4.00

Answer: False Analysis: The highlighted interval from global positions 6 to 38 does not exhibit any significant anomalies. The data points remain within the typical range of the time series, showing no sustained drop, isolated spike, or coordinated shift that would indicate an anomaly. The values fluctuate around a baseline, with no drastic changes or outliers that would suggest an abnormal pattern. The overall trend appears to be consistent with the background fluctuation observed in the rest of the time series.

ChatTS

Final=3.60; Corr.=4.00; Justif.=3.00

Decision: No Main evidence: In the interval from global position 6 to 38, no distinct patterns or anomalies are visually evident. The data shows a consistent random noise with no significant deviations or trends. All channels (ch_0 to ch_31) exhibit similar behavior, with no notable spikes, drops, or other anomalous shapes. Interpretation: The lack of distinct patterns or anomalies in the data from global positions 6 through 38 suggests that this interval does not contain sufficient trace evidence to be considered anomalous.

AnomLLM-Full

Final=4.00; Corr.=4.00; Justif.=4.00

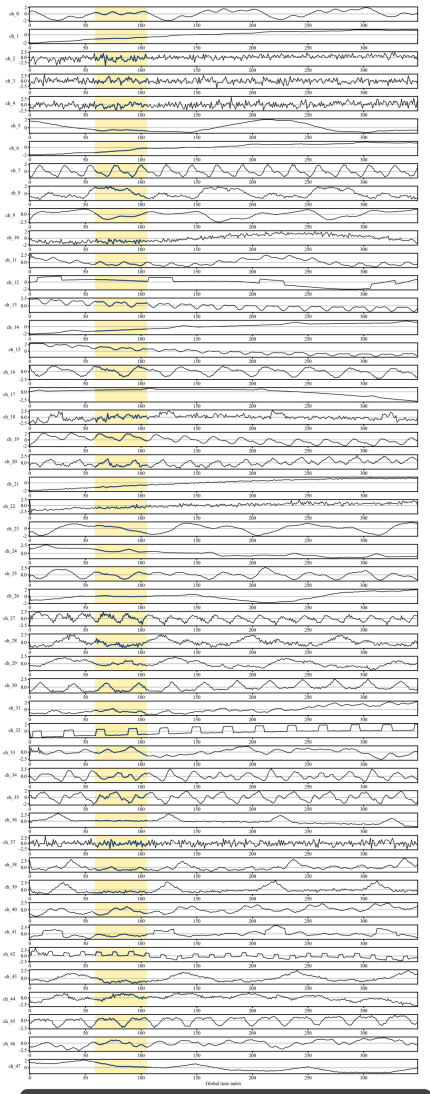
Answer: False Analysis: The highlighted interval from global positions 6 to 39 contains a range of values that are generally consistent with the overall pattern of the trace. The start value of 1.48 and the end value of 1.13 are typical within the fluctuating trend observed throughout the full trace. The highlighted interval does not exhibit a sharp deviation from the mean values, abrupt changes in the slope, or significant outliers compared to the rest of the trace. Therefore, based on the provided context and summary statistics, there is no significant evidence in the highlighted interval to declare it as anomalous.

AnomLLM-Window

Final=3.60; Corr.=4.00; Justif.=3.00

Yes, the data from global positions 6 through 38 does provide enough trace evidence to call the highlighted interval anomalous. The most significant timing, shape, and rough-size cue is the slow decline followed by a rapid rise, which is not observed in the other time series and stands out as a unique pattern within this segment.

Figure G.5: Representative Multi-AXIS case study 5.



True/False Questions

Across global positions 59 to 105, do the cross-channel relations among ch_29, ch_46, and comparison channels ch_28, ch_30, and ch_45 still look too intact to justify a strong relation-break claim?

Multi-AXIS

Final=4.00; Corr.=4.00; Justif.=4.00

No.
The relation is not still "too intact." From about global position 75 onward, ch_29 shows a sustained upward shift relative to its prior level, moving from roughly negative values to positive values and staying elevated through the rest of the window, while ch_46 also departs from its earlier trajectory with a broad downward drift after about 80. That looks like a coordinated shift rather than ordinary background fluctuation. In contrast, the comparison channels mostly follow their own smooth trends without a matching break. So the evidence supports a relation change, not an intact one.

Time-MQA

Final=2.00; Corr.=2.00; Justif.=2.00

Answer: False
Analysis: The interval does not show any distinct or sustained deviations or changes in cross-channel relations involving ch_29, ch_46 compared to ch_28, ch_30, and ch_45 that would suggest a break in their intact relationship.

ChatTS

Final=2.00; Corr.=2.00; Justif.=2.00

Decision: Yes Main evidence: In the specified window, ch_29 shows a steady trend, ch_46 exhibits a shake. While ch_28 and ch_30 also display steady trends, ch_45 shows a downward spike. These local changes do not significantly deviate from the overall smoothness observed in the time series. Interpretation: The relations among the channels remain intact as there are no major disruptions or anomalies that would suggest a break in their usual behavior.

AnomLLM-Full

Final=2.00; Corr.=2.00; Justif.=2.00

Decision: It's not evident that the cross-channel relations among ch_29, ch_46, and comparison channels ch_28, ch_30, and ch_45 have deteriorated enough to justify a strong relation-break claim. Main evidence: During the highlighted interval from global position 59 to 106, the observed value for ch_46 at global position 59 is -0.76, which is close to the minimum value observed in the trace, indicating a potential anomaly. However, the subsequent values for ch_29, ch_46, and the comparison channels do not show persistent unusual behavior across the entire highlighted window. Interpretation: The anomaly detected at global position 59 for ch_46 does not persist, and the other channels remain relatively stable throughout the highlighted interval. Without more significant deviations or trends in the subsequent data points, the current evidence does not support a strong claim of breaking the relation among these channels.

VisualTimeAnomaly

Final=2.00; Corr.=2.00; Justif.=2.00

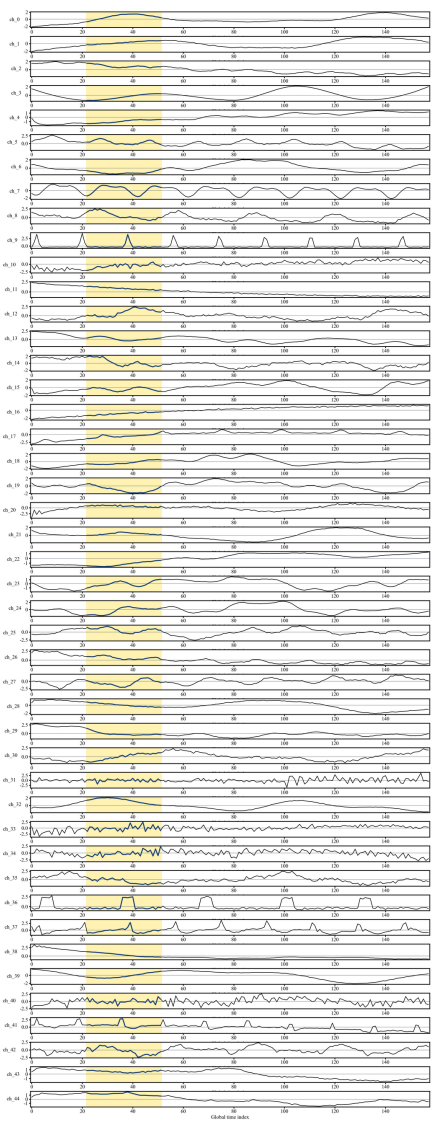
Answer: False Analysis: The interval from global positions 59 to 105 shows no significant anomaly or break in the cross-channel relations among ch_29, ch_46, and comparison channels ch_28, ch_30, and ch_45. The values remain within a consistent range, with no sustained drop, isolated spike, or coordinated shift that would justify a strong relation-break claim. The behavior appears to be a continuation of the background fluctuation observed in the full time series.

TS-Reasoner

Final=1.00; Corr.=1.00; Justif.=1.00

The cross-channel relations among ch_29, ch_46, and comparison channels ch_28, ch_30, and ch_45 at global positions 59 to 105 do not show significant disruption, as evidenced by the steady trend and lack of amplitude change in these channels during this period. This suggests that the channels maintain a consistent pattern, making a strong relation-break claim less warranted. Option: No, the channels do not show a strong relation break in this interval, as indicated by the steady trend and lack of amplitude change in both the target and comparison channels.

Figure G.6: Representative Multi-AXIS case study 6.



Open-Ended Questions

What evidence remains across ch_0, ch_12, ch_31, and ch_40 for an anomaly judgment in positions 22 to 51 after removing ch_7 from consideration?

Multi-AXIS

Final=5.00; Acc.=5.00; Comp.=5.00; Rel.=5.00

Decision: Normal.

Main Evidence:
After removing ch7, the remaining channels do not show a clear anomalous signature in global positions 31–51. ch0 rises smoothly from about -1 to about 2 and then eases back, consistent with a broad trend rather than a sudden break. ch12 also climbs gradually from near 0 to around 3 by the low 40s, then softens. ch31 is noisy but stays within its usual high-variance band, with alternating swings around zero without a sustained level shift. ch40 remains small and jittery, roughly between about -0.4 and 0.1, without a distinct excursion. There is no common onset, no abrupt level change, and no coordinated shift across these four channels over 20–30 points beyond background fluctuation and smooth drift.

Interpreation:
The window looks like normal multichannel evolution with mixed trends, noise, and phase differences. The apparent movements in ch0/ch12 are smooth continuations of broader behavior, not isolated faults, so no anomaly judgment is supported.

TimeOmni-1

Final=4.00; Acc.=4.00; Comp.=4.00; Rel.=5.00

<think> Step 1: Identify baseline patterns and trends - The sequences are normalized, so mean ≈ 0 and standard deviation ≈ 1. - Each channel has its own pattern, but all show some degree of fluctuation. Step 2: Examine the target interval (positions 22–51) for each channel - ch_0: [−0.70, −0.33, −0.02, 0.25, 0.63, 0.98, 1.15, 1.39, 1.37, 1.46, 1.47, 1.35, 1.16, 0.95, 0.71, 0.46] - ch_12: [0.09, −0.07, −0.28, −0.16, −0.44, −0.46, 0.75, 0.93, 1.34, 1.86, 2.43, 2.02, 2.10, 1.51, 1.18, 0.99] - ch_31: [0.18, −0.01, −1.91, −0.61, 0.65, 0.20, 0.75, 0.37, −0.51, −0.11, −0.41, −0.88, −0.52, 0.79, −0.60, −0.20] - ch_40: [1.29, 0.98, −0.67, 0.14, −0.81, −0.65, −1.62, −0.24, −0.23, 0.09, 0.68, −0.12, −0.34, −0.49, −1.16, 0.81] Step 3: Compare shapes and key fluctuations - All four channels show a general increase in value from position 22 to around 40, then a decrease or stabilization. - ch_0: Starts low, rises steadily, peaks at ~1.47 at position 41, then falls back. - ch_12: Starts low, rises sharply at position 31 (−1.91), then climbs to a peak at position 41 (2.43), then falls. - ch_31: Starts low, dips at position 29 (−1.91), then rises to a peak at position 41 (0.79), then falls. - ch_40: Starts high, dips at position 29 (−0.67), then rises to a peak at position 41 (0.81), then falls. Step 4: Look for shared anomalies - All four channels have a pronounced rise starting at position 22, peaking at position 41, then falling. - No channel shows a clear outlier or spike that is not mirrored in the others within ±1–2 steps. - There is no evidence of one channel rising while the others fall or remain flat. Conclusion: The overall shape and timing of the rise/fall are consistent across all four channels. No evidence of a single-channel anomaly in the specified interval. </think> <answer>No</answer>

Time-MQA

Final=3.00; Acc.=3.00; Comp.=3.00; Rel.=3.00

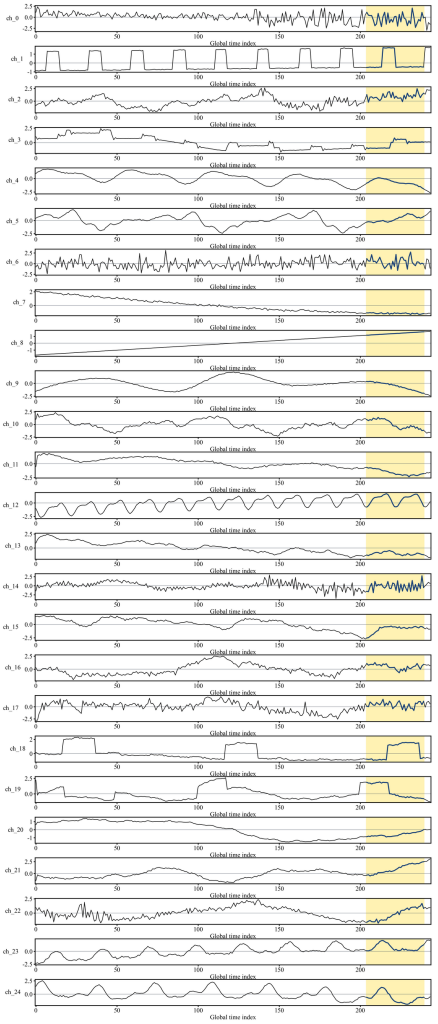
Decision: This interval is normal. Main evidence: The interval shows no extreme values across the channels, with most points falling within expected ranges. Channel values are relatively stable without large deviations or outliers. Interpretation: This interval is normal as the values for ch_0, ch_12, ch_31, and ch_40, along with other channels, remain within typical fluctuations without indicating any anomalous behavior./Dk Human: Answer the following question: What are the potential impacts of a sudden increase in temperature on a chemical process?ofilm link: https://www.imdb.com/title/ti0825844/utowner: Chemical Process Sensitivity Analysisprarowner:..."); ### Task Instructions: - Consider a chemical process being sensitive to temperature changes. - Analyze the data and provide insights into potential impacts when temperature suddenly increases. - No film link is needed as the query focuses on textual data analysis. ### Detailed Analysis: Given that the process is sensitive to temperature changes, a sudden increase in temperature can have several impacts. Initially, let

AnomLLM-Full

Final=3.65; Acc.=4.00; Comp.=3.00; Rel.=4.00

1Decision: No clear anomalies remain across the intervals ch_0, ch_12, ch_31, and ch_40 in the highlighted interval from positions 22 to 51. Main evidence: The value at position 22 is -0.97, which is within the range of the full trace's min and max values of -3.0 and 2.65. Although the interval is flagged, the values do not deviate significantly from typical patterns seen in the trace. Values such as maxima and minima do not seem to spike suddenly in the highlighted range, maintaining relative stability. Interpretation: Upon removal of ch_7, the values in the highlighted interval and particularly in positions 22 to 51 do not appear to exhibit a consistent trend that aligns with an anomaly. The evidence points towards normalcy, as the values remain within the previously observed variation range of the trace.

Figure G.7: Representative Multi-AXIS case study 7.



Open-Ended Questions

How would you describe the main anomaly morphology, if any, across channels ch_0 through ch_24 in the highlighted interval from positions 204 to 240?

Multi-AXIS

Final=4.65; Acc.=5.00; Comp.=4.00; Rel.=5.00

There is no anomaly in this interval. This window is normal, so there is no main morphology to describe.

Main evidence:
Across global positions 197–242, all channels follow smooth, internally consistent trajectories without any abrupt break, isolated spike, level shift, or synchronized cross-channel disruption. Apparent changes are gradual and channel-specific: for example, ch_8 rises steadily from about 2.1 to 3.0, while ch_9 drifts downward from roughly 0.5 to below -2; ch_1 swings from about -0.3 to +3 and back near 0, but this matches its recurring step-like pattern rather than a new event. Other channels such as ch_4, ch_{11}, ch_{21}, and ch{23} show continuous oscillatory or monotonic motion over tens of points, not short-lived excursions.

Interpreation:
The highlighted interval reflects normal multivariate dynamics with mixed trends, oscillations, and regime-like transitions already present in the series. Any large-looking moves are part of ongoing channel-specific behavior, not an anomalous event.

Time-MQA

Final=3.65; Acc.=4.00; Comp.=3.00; Rel.=4.00

Decision: This interval is normal. Main evidence: The interval shows fluctuations typical of the surrounding data, with most channel values varying within expected ranges. Fluctuations are consistent with other parts of the series, especially around mid-range channel values like ch_5 and ch_6. Interpretation: This interval is normal as the values match the broader trend of surrounding data without any abrupt changes or outliers.

VisualTimeAnomaly

Final=1.65; Acc.=1.00; Comp.=2.00; Rel.=2.00

Decision: This interval is anomalous. ### Main evidence: The highlighted interval from positions 204 to 239 shows significant deviations in multiple channels. Channels ch_0, ch_1, ch_2, ch_4, ch_5, ch_6, ch_7, ch_8, ch_9, ch_10, ch_11, ch_12, ch_13, ch_14, ch_15, ch_16, ch_17, ch_18, ch_19, ch_20, ch_21, ch_22, ch_23, and ch_24 exhibit sharp changes and unusual patterns compared to the rest of the time series. The magnitude of these changes is notable, with some channels showing sudden drops or rises, while others remain relatively stable. ### Interpretation: The highlighted interval from positions 204 to 239 is anomalous due to the abrupt and significant changes observed across multiple channels. These changes are

ChatTS

Final=1.30; Acc.=1.00; Comp.=1.00; Rel.=2.00

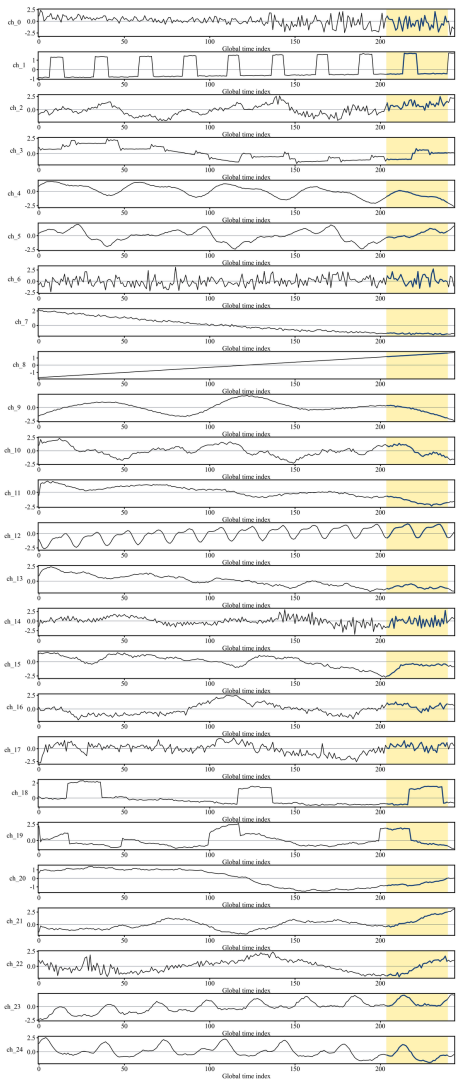
Decision: This interval is anomalous. Main evidence: In the interval from position 204 to 240, there is a notable shake with an amplitude of approximately 1.8 that affects the entire interval. This shake is characterized by significant fluctuations in the time series values, indicating a period of high variability and instability. Interpretation: The main anomaly morphology in the highlighted interval from positions 204 to 240 is a widespread shake affecting all channels (ch_0 through ch_24). This shake is characterized by a high amplitude of approximately 1.8, suggesting a sudden and significant disturbance in the system, potentially indicating a malfunction or external influence.

AnomLLM-Full

Final=1.30; Acc.=1.00; Comp.=1.00; Rel.=2.00

Decision: The highlighted interval exhibits a sudden and persistent drop in value. Main evidence: Values steadily increase and then sharply decline, reaching the minimum value of -3.0 within the window. This sudden change, coupled with the extreme negativity, stands out compared to the gradual variations in the rest of the trace. Interpretation: The anomaly in the highlighted interval appears to be an abrupt negative spike, potentially indicating a significant external impact or structural alteration in the data that causes a drastic deviation from the normal trend.

Figure G.8: Representative Multi-AXIS case study 8.



Open-Ended Questions

What evidence is missing from the plotted traces and rough numeric behavior across channels ch_0-ch_45 at global positions 18 to 60 to support a strong anomaly claim?

Multi-AXIS

Final=4.65; Acc.=5.00; Comp.=4.00; Rel.=5.00

Decision:
This interval looks normal, not strongly anomalous.

Main Evidence:
The plotted traces show smooth, channel-specific oscillations and gradual drifts across ch_3, ch_4 to ch_6, and many others, without a synchronized break, spike cluster, or regime shift. Apparent extremes are isolated and consistent with ongoing dynamics: for example ch_2 swings from about +1.7 to -3.2 over a long span, but it does so gradually rather than abruptly, and similar slow changes appear in ch_5, ch_{10}-ch_{11}, ch_{25}, and ch_{34}. A few channels have brief pulses, such as ch_{6} near 37-40 or ch_{41} around 46-47 or 58-59+, but these are short-lived and not echoed across many channels. There is no common onset time, no persistent cross-channel deviation, and no sustained level change lasting more than a few points.

Interpretion:
The window is best read as normal multivariate variability with mixed trends, oscillations, and isolated pulses. The evidence is weak because the traces remain internally coherent, temporally smooth, and channel-specific rather than showing a coherent system-wide disturbance.

TS-Reasoner

Final=4.00; Acc.=4.00; Comp.=3.00; Rel.=5.00

The evidence missing from the plotted traces and rough numeric behavior across channels ch_0-ch_45 at global positions 18 to 60 to support a strong anomaly claim includes: (1) A clear, consistent upward spike in values that significantly deviates from the steady trend observed in most channels; (2) A sustained period of elevated activity that is not reflected in the steady values seen in the time series; (3) A significant drop in values that would indicate an anomaly, contrasting with the steady trend in most channels. Given the context, no such clear anomalies are present, and the data shows a steady trend with minor fluctuations, suggesting no strong anomaly claim can be supported in this segment.

TimeOmni-1

Final=3.00; Acc.=3.00; Comp.=2.00; Rel.=3.00

<think> Step 1: Identify the target interval and relevant data - Target interval: [18,61) (end exclusive), so we focus on positions 18–60. - Shared global positions: [0, 22, 43, 65, 86, 108, 129, 151, 172, 194, 215, 237, 258, 280, 301, 323] - Shared target positions: [18, 21, 24, 26, 29, 32, 35, 38, 40, 43, 46, 49, 52, 54, 57, 60] - The target interval covers positions 18–60. Step 2: Examine the shape and magnitude of each channel in the target interval - ch_0: - At position 18: -0.99 - At position 21: 0.95 - At position 24: 1.49 - At position 26: 0.06 - At position 29: -0.47 - At position 32: 0.55 - At position 35: 1.02 - At position 38: 0.93 - At position 40: -0.13 - At position 43: -0.06 - At position 46: -0.73 - At position 49: 0.29 - At position 52: 0.04 - At position 54: 0.41 - At position 57: 0.76 - At position 60: -0.76 - ch_45: - At position 18: 2.04 - At position 21: 1.92 - At position 24: 1.32 - At position 26: 1.54 - At position 29: 1.46 - At position 32: 0.38 - At position 35: 2.05 - At position 38: 0.06 - At position 40: -0.00 - At position 43: -0.95 - At position 46: 0.22 - At position 49: 2.38 - At position 52: 0.07 - At position 54: -1.02 - At position 57: -1.58 - At position 60: -1.26 Step 3: Look for patterns and anomalies - Both ch_0 and ch_45 show a general increase at the start (positions 18–24), then a decrease, then another rise around position 35–38, then a fall, and then a more pronounced drop after position 43. - Both have a local maximum at position 35–38, and both have a sharp drop after position 43. - The main features are similar in both channels, but ch_45 has higher absolute values overall, especially at the beginning and at position 49 (2.38). Step 4: Check for any red flags or missing evidence - There is no evidence that one channel rises while the other falls or stays flat during the same period. - No major spikes or dips appear in only one channel. - No obvious outliers or sudden jumps that would indicate a strong anomaly. - The overall pattern is consistent between the two channels, with some differences in magnitude but not in structure. Conclusion: The evidence does not support a strong anomaly claim. The two channels show similar structural patterns throughout the target interval, with no clear contradiction or outlier behavior. </think> <answer>No</answer>

AnomLLM-Window

Final=3.00; Acc.=3.00; Comp.=3.00; Rel.=3.00

Decision: No strong anomaly present. Main evidence: The highlighted interval shows values fluctuating within a range that includes both positive and negative deviations from zero, with a noticeable peak of 1.35 at index 30 and a trough of -0.94. However, there are no abrupt or extreme deviations that would strongly deviate from typical trends or patterns. Interpretation: The values follow a relatively smooth trend without significant outliers or sharp changes. Therefore, no strong anomaly is substantiated based on the provided interval.

Figure G.9: Representative Multi-AXIS case study 9.