CLARTÉ - Channel-adaptive Lightweight Aggregation for Retrieval and Temporal Explainability

A needle in a timeseries haystack

Martin Royer[1], Maxence Gagnant[1], Martin Gonzalez[1], and Milad Leyli-Abadi[1] (✉)[2]

[1] IRT SystemX, Palaiseau, France
[2] {name.surname}@irt-systemx.fr

Abstract. Multivariate subsequence search is a core primitive for time series analysis, yet aggregating univariate distances with no effort to differentiate signal from noise can lead to failure. When informative channels are sparse, noise from irrelevant dimensions drowns out genuine matches. We propose two lightweight, training-free methods that address this "needle in a haystack" problem by producing not just match locations but also dimension attributions, both operating at the same asymptotic cost as the standard efficient search cost.
We integrate these into CLARTE, a two-stage pipeline that chains dimension-adaptive retrieval to a contrastive explanation stage where a text-LLM grounded in a fault-free reference produces a structured per-channel pattern description for each attributed channel.
Experiments show that contrasted (local, per-position) aggregation outperforms global strategies when background processes recur coherently; that historical reference segments are decisive for text-LLM interpretation quality while visual and signal-tokenized alternatives do not benefit from additional context; and that the full pipeline correctly localizes and interprets faults against a pool of 47 confounders on the LBNL FCU dataset.

Keywords: Multivariate Time Series · Subsequence Search · Channel Attribution · Explainability · LLM Reasoning

1 Introduction

Consider an engineer who has captured a reference fault event: five minutes of sensor readings from a machine with an early-stage heat-exchanger fouling, across 24 channels of data. Only 3 of those channels carry the fault signature; the remaining 21 carry ongoing and recurring operational patterns (pump cycles, thermal rhythms, ventilation schedules) that appear dozens of times throughout six months of continuous recordings and are unrelated to the fault.

Any method that aggregates over all channels indiscriminately produces false positives at every occurrence of the dominant background process rather than

at the fault: the globally best-matching channel may be tracking anything but the event of interest. How to find fault recurrences and understand them? Which channels bear the signature, what temporal pattern does each exhibit (a temperature spike, a pressure drop, a flow-rate flatline), and when a maintenance logbook of past events is available, which fault type does the historical records associate with each occurrence? This is a needle-in-a-haystack problem with a multivariate twist: the needle exists only in an *unknown subset* of channels, and treating all channels equally drowns the signal in noise. Even once the needle is located, it remains opaque: it is simply a match score with no account of what distinguishes the matched window from the surrounding hay.

The structure of that vignette recurs across multiple domains: a maintenance engineer scanning months of plant sensor readings for a known fault signature [16]; a cardiologist retrieving a reference arrhythmia episode from ambulatory recordings [2,13]; a physiotherapist locating a movement prototype in days of inertial data [10,11]; an analyst matching a price pattern against a multi-year tick record [3]. In every case the task is the same: given a short multivariate event (let us call it a *query*), locate its occurrences in a long multivariate target; and the difficulty is shared: most channels track background rhythms (respiration, periodic price drift, cyclic movement) that recur far more often than the event of interest, while only an unknown subset bear the event signature.

The de facto method for subsequence search in a multivariate signal is to compute a distance profile via MASS: Mueen's Algorithm for Similarity Search [14] independently on each channel. MASS efficiently computes the z-normalized Euclidean distance between a *query* time series and every subsequence of a longer series using FFT-based convolution. One then aggregates the scores, typically by summing or averaging into a single composite profile whose minimizer is the best match. This works if all channels carry signal, but under equal-weight aggregation, irrelevant channels contribute noise that can overwhelm the signal from the informative ones, leading to missing matches or false positives; a failure mode shown to render up to 60% of motif discoveries spurious [19].

Several lines of work acknowledge that channel relevance is query-dependent, yet none simultaneously addresses channel discovery, efficient retrieval, and textual interpretation. In the matrix profile literature, mSTAMP [26] and LAMA [18] perform dimension-adaptive motif discovery in the self-join setting and leave channel-count selection as an open problem. Channel-aware anomaly detection methods learn channel importance from labeled data [15] but do not address subsequence search. Time series language models struggle with temporal localization as context length grows [29], motivating a dedicated retrieval backbone. On the interpretation side, instruction-tuned LLMs and TS-LLMs have been applied to time-series question answering [23] and anomaly description [25]; VLMs have been used to characterize anomalies from plotted series images [24,9], with TAMA [28] providing normal-window images as visual reference. None of these approaches grounds interpretation in a formally retrieved match position, with respect to a reference window from a fault-free region.

We introduce CLARTÉ (**C**hannel-adaptive **L**ightweight **A**ggregation for **R**etrieval and **T**emporal **E**xplainability), a training-free two-stage pipeline for explainable multivariate subsequence search running at the $O(nd \log n)$ cost of per-dimension MASS. Stage 1 computes per-dimension MASS profiles and aggregates them via two complementary strategies: *weighted aggregation*, which assigns each channel a query-adaptive importance score from the shape of its distance profile, and *contrasted aggregation*, which selects the best channel subset independently at each candidate position. Stage 2 exploits the rank-1 match position in two ways: a deterministic logbook lookup derives a fault label; and a contrastive LLM call anchored on a fault-free reference window drawn from the same signal, produces a structured per-channel description P (pattern label and signed amplitude deviation). Unlike embedding-based retrievers [22] or channel-cluster learners [15], CLARTÉ requires no training data: the only inputs are the query, the target signal, an optional logbook of timestamped fault records, and a pre-trained LLM for Stage 2.

In this paper we contribute the following:

- **Two training-free, query-adaptive aggregation strategies for multivariate MASS.** A *weighted* strategy assigns each channel a global importance score from the discriminativeness of its distance profile; a *contrasted* strategy selects the best channel subset independently at each candidate position. Both yield interpretable channel attributions alongside the match. We characterize their complementary failure modes on a controlled synthetic benchmark and introduce the global/local distinction as a principled design axis for multivariate subsequence search.
- **A model-agnostic Stage 2 framework for contrastive fault explanation.** Given the top match position, a structured output $P = \{(i, \psi_i, a_i)\}$ describes each informative channel: pattern label from a fixed vocabulary Ψ and signed amplitude deviation, relative to a fault-free reference window anchored to the same signal. We propose and evaluate three implementations (text-LLM, visual prompting, TS-LLM), and show that a text-LLM augmented with structured numerical context consistently outperforms visual and signal-tokenized alternatives when historical reference segments are available.
- **An evaluation protocol and end-to-end demonstration on LBNL FCU.** Since structured per-channel output lacks annotated ground truth, we evaluate Stage 2 via binary Ψ-pattern detection on synthetically probed FCU segments, a necessary-condition proxy for full labelling, showing that contrastive access to T is decisive for text-based reasoning but not for visual or TS-LLM approaches. The full pipeline is then exercised end-to-end on four fault types against a pool of 47 confounders, with controlled oracle and random baselines isolating the contribution of each stage.

A detailed survey of related work is provided in the Appendix (§5.1).

2 The CLARTÉ pipeline

Problem setting. Let $Q \in \mathbb{R}^{m \times d}$ be a multivariate event of interest (*query*) and $T \in \mathbb{R}^{n \times d}$ a target signal with $n \gg m$. When available, a *logbook* $\mathcal{L} = \{(s_k, \ell_k)\}_{k=1}^{K}$ records past fault events annotated in T: $s_k \in \{1, \ldots, n\}$ is the start position and $\ell_k \in \Lambda$ its fault label. Stage 1 does not require $\mathcal{L}$; it is used solely by Stage 2 for fault label derivation. We assume that there exist a ground-truth position $t^* \in \{1, \ldots, n-m+1\}$ and an *informative channel set* $\mathcal{I}^* \subsetneq \{1, \ldots, d\}$ such that $Q_{\mathcal{I}^*}$ is a noisy transformation of $T[t^* : t^*+m]_{\mathcal{I}^*}$; channels outside $\mathcal{I}^*$ carry noise independent of the target. Both t^* and $\mathcal{I}^*$ are unknown.

CLARTÉ. We propose a two-stage pipeline for explainable multivariate subsequence search; see Figure 1. Stage 1 takes Q and T and produces a ranked list of match candidates, each annotated with a per-channel importance attribution. Stage 2 takes the rank-1 match, derives a fault label $\hat{\ell} \in \Lambda \cup \{\texttt{normal}\}$ from the logbook, and produces a structured per-channel description P of how the fault manifests relative to a fault-free reference window sampled from T.

The pipeline jointly pursues four objectives:

1. **Event retrieval**: estimate $\hat{t} \approx t^*$, a ranked list of candidate positions in T;
2. **Dimension attribution**: for each candidate, identify $\hat{\mathcal{I}} \approx \mathcal{I}^*$, the subset of channels driving the match;
3. **Fault derivation** (requires $\mathcal{L}$): assign a fault label $\hat{\ell} \in \Lambda \cup \{\texttt{normal}\}$ by looking up the logbook entry nearest to $\hat{t}$ within a tolerance τ; return `normal` if no entry falls within τ;
4. **Fault interpretation**: for each channel $i \in \hat{\mathcal{I}}$, assign a pattern label $\psi_i \in \Psi$ and a signed amplitude deviation a_i relative to normal behaviour in T.

Objectives (1) and (2) are coupled: aggregating over all d channels is suboptimal when $|\mathcal{I}^*| \ll d$, but selecting channels requires knowing where the true match is. Stage 1 solves them jointly. Objective (4) is downstream: the interpretation is only meaningful in the right channels and against a quantitative baseline; the channel attribution $\hat{\mathcal{I}}$ and the match position $\hat{t}_1$ that anchor normal reference windows are both provided by Stage 1.

2.1 Stage 1: Dimension-Adaptive Retrieval

Given Q and T, we compute per-dimension distance profiles $\pi_i(t) = \mathrm{MASS}(Q_i, T_i)$ for each dimension $i = 1, \ldots, d$, where $\pi_i(t)$ is the z-normalized Euclidean distance between Q_i and the subsequence of T_i starting at position t, for $t = 1, \ldots, n-m+1$. These d profiles are aggregated into a single composite distance profile $D(t)$ using one of two strategies.

Weighted aggregation. Each dimension receives a global weight $w_i = (\bar{\pi}_i - \min \pi_i)/(\sigma_{\pi_i} + \epsilon)$, reflecting how discriminative its distance profile is. Dimensions with sharp minima against a high background receive large weights, whereas flat and non-discriminative profiles are heavily penalized and receive close to zero weight. The composite profile is $D(t) = \sum_i \hat{w}_i \pi_i(t)$ where $\hat{w} = w/\|w\|_1$.

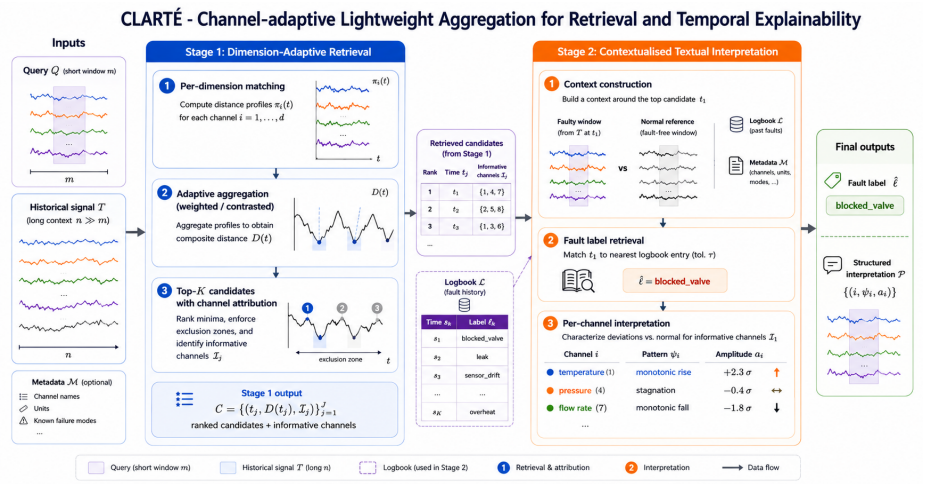

Fig. 1: Overview of the CLARTE pipeline. Stage 1 performs explainable multivariate subsequence retrieval and identifies informative channels. Stage 2 derives the fault label $\hat{\ell}$ from the logbook and produces a structured per-channel description P by contrasting Q against a fault-free reference window drawn from T.

Contrasted aggregation. At each position t, the d per-dimension distances are sorted in ascending order. Cumulative sums $S_k(t) = \sum_{r=1}^{k} \pi_{(r)}(t)$ are computed for $k = 1, \ldots, d$, where $\pi_{(r)}(t)$ denotes the r-th smallest distance at position t. Each S_k profile is z-normalized across all timestamps, yielding $Z_k(t)$. The composite profile is $D(t) = \min_k Z_k(t)$, and the *support* $K(t) = \arg\min_k Z_k(t)$ records how many dimensions participate in the match at that position.

Output. Stage 1 produces a set of match candidates

$$\mathcal{C} = \{(t_j, D(t_j), \mathcal{I}_j)\}_{j=1}^{J}, \tag{1}$$

ranked by ascending $D(t_j)$ with an exclusion zone enforced between candidates. Each candidate carries a *dimension attribution* $\mathcal{I}_j \subseteq \{1, \ldots, d\}$: the subset of dimensions identified as informative for that match. In the weighted strategy, $\mathcal{I}_j$ is the set of dimensions with weights exceeding the uniform baseline $1/d$; in the contrasted strategy, $\mathcal{I}_j$ is the $K(t_j)$ best-scoring dimensions at position t_j, naturally varying across candidates.

2.2 Stage 2: Contextualised Textual Interpretation

Stage 2 exploits the rank-1 match position $\hat{t}_1$ in two complementary ways, and produces two outputs accordingly.

The first is a fault label $\hat{\ell}$, retrieved deterministically from $\mathcal{L}$ the logbook. The second is a per-channel description P of how the fault manifests relative to

normal behaviour. We use that *context* to make visible what stands out in the *query*: rather than characterising absolute signal values, we answer the question of what in Q deviates from how T normally looks. To characterize that behaviour, we use a *normal reference window* $R \in \mathbb{R}^{m \times d}$, a fault-free segment drawn from T.

Fault label derivation. Given $\hat{t}_1$ and $\mathcal{L}$, the fault label is derived deterministically:

$$\hat{\ell} = \begin{cases} \ell_{k^*} & \text{if } \min_k |s_k - \hat{t}_1| \leq \tau, \\ \texttt{normal} & \text{otherwise,} \end{cases} \qquad k^* = \arg\min_k |s_k - \hat{t}_1|, \tag{2}$$

where τ is a tolerance (default: $m/4$).

Per-channel contextualised description. Let

$$f_2 : \big(Q,\ R,\ \mathcal{I}_1\big) \longmapsto P \tag{3}$$

where $Q \in \mathbb{R}^{m \times d}$ is the current faulty event; $R \in \mathbb{R}^{m \times d}$ is a normal reference window (a fault-free segment drawn from T); $\mathcal{I}_1 \subseteq \{1, \ldots, d\}$ is the channel attribution from Stage 1, identifying the channels that carry the fault signature common to Q and $T[\hat{t}_1 : \hat{t}_1 + m]$.

The output is a *per-channel description* $P = \{(i,\ \psi_i,\ a_i)\}_{i \in \mathcal{I}_1}$, where $\psi_i \in \Psi$ is a pattern label drawn from a fixed vocabulary Ψ (Table 1) and $a_i \in \mathbb{R}$ is a signed amplitude deviation of Q_i relative to R_i.

Label	Description
`peak` / `trough`	sharp local maximum / minimum
`monotonic_rise` / `monotonic_fall`	sustained directional drift
`level_shift_up` / `level_shift_down`	abrupt step change
`piecewise_const`	generic level change (covers piecewise-constant signals)
`stagnation`	variance collapse, flatline
`oscillation`	periodic pattern
`noise_burst`	high-frequency power increase
`normal`	no anomaly relative to baseline

Table 1: Pattern vocabulary Ψ used in the per-channel output P.

Implementation. Equation (3) is model-agnostic: any system that accepts $(Q, R, \mathcal{I}_1)$ and returns P is a valid Stage 2 implementation. Our primary implementation is a **text-LLM** pipeline. Per-channel statistics (e.g. amplitude z-score, trend slope, variance ratio, lag-1 autocorrelation) are derived from $Q_{\mathcal{I}_1}$ and $R_{\mathcal{I}_1}$, both z-scored against the global baseline of T. These are complemented by per-channel natural-language descriptions produced by a TS-LLM (ChatTS [23]) from the raw signal values, grounding the symbolic statistics in signal-level language. The combined context is serialised as JSON and fed to an

instruction-tuned LLM with a schema-constrained prompt requesting P. This design requires no time-series-specific architecture for the reasoning step and benefits from the TS-LLM's signal perception without inheriting its limitations for structured output.

Two alternative implementations: **visual prompting** (rendering Q and R as overlaid line plots passed to a VLM) and **TS-LLM prompting** (tokenising Q and R directly into a TS-LLM with event/reference channel labelling), are evaluated as baselines in (§3.2).

3 Experiments

We structure the evaluation in three parts. Part I (§3.1) evaluates Stage 1 (dimension-adaptive retrieval) in isolation, where ground truth is fully controlled and the question is purely whether the method finds the right position and identifies the right dimensions. Part II (§3.2) measures how much more explainable is a timeseries signal by contextualisation with respect to historical data. Stage 2 requires realistic signals with natural background variability: overly clean data trivialises contrastive pattern recognition, so Stage 2 is evaluated on the LBNL FCU dataset [7], a realistic building-sensor collection. Part III (§3.3) shows an end-to-end evaluation that bridges both stages: faults are injected into FCU to supply t^* ground truth while the signal's natural complexity provides a realistic backdrop for Stage 2.

3.1 Stage 1: Dimension-Adaptive Retrieval

A short query event (length $m \in [150, 300]$, d channels) is planted at a known position in a long target signal ($n = 5\,000$). Query channels in $\mathcal{I}^*$ carry a piecewise-constant synthetic signal; the remaining $d - |\mathcal{I}^*|$ channels are modified according to the condition. We evaluate six specific data conditions (Table 5 in the Appendix §5.2), each designed to probe a qualitatively distinct failure mode of dimension-naive aggregation, with 50 seeds per condition (300 samples total).

We evaluate Stage 1 on controlled synthetic data where ground truth is fully known, comparing the aggregators from (§ 2) against the following baselines:

- `euclidean`: equal-weight mean of per-channel distance profiles.
- `mv_mass`: equal-weight mean of MASS profiles per-channel (z-normalized).
- `lama`: single-query adaptation of LAMA [18]; per-position dimension sorting by MASS distance with k^* set as the largest gap in the sorted profile (extent criterion approximated for the AB-join setting; no cross-position normalization).
- `mstamp_mdl`: mSTAMP AB-join [26] with global dim ordering and a Minimum Description Length (MDL) stopping rule; unfortunately we find the rule selects $k^* = 1$ on all samples because the global-minimum dim ordering makes the cumulative mean non-decreasing.
- `oracle_dim`: MASS restricted to the true $\mathcal{I}^*$, acting as a hard upper bound.

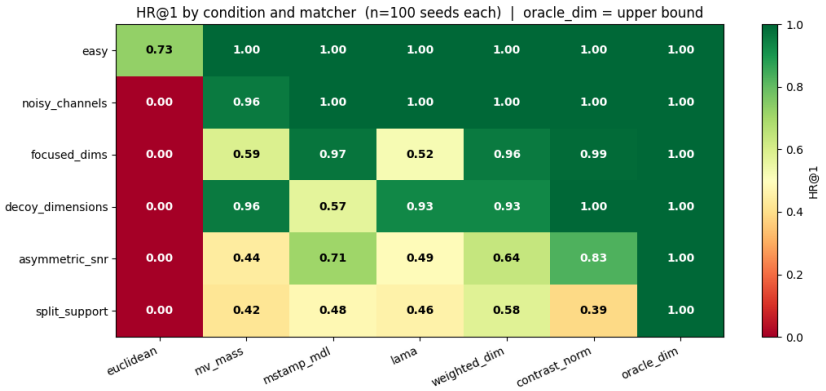

Fig. 2: HR@1 by condition and retrieval approach (50 seeds each, $n = 5\,000$). `oracle_dim` (rightmost column) is the hard upper bound; higher is better.

These methods fall into two families: *global* aggregators (`mv_mass`, `weighted_dim`, `mstamp_mdl`), whose dimension weights or ranks are fixed once from the full profile and applied uniformly at every t; and *local* aggregators (`contrast_norm`, `lama`), which re-evaluate the contributing channel subset independently at each candidate position t.

We report two metrics: **HR@1**, the fraction of queries whose top-1 candidate falls within $\lfloor m/2 \rfloor$ of the ground-truth start; and **Dim F1**, the F1 of the predicted channel set $\hat{\mathcal{I}}$ against $\mathcal{I}^*$. For `weighted_dim` and `mstamp_mdl` (*global* methods), $\hat{\mathcal{I}}$ is fixed independently of position: for `weighted_dim`, $\hat{\mathcal{I}}$ contains channels whoses normalized weight is above $1/d$; for `mstamp_mdl`, $\hat{\mathcal{I}}$ contains the greedy-ordered k^* dims selected by MDL. For `contrast_norm` and `lama` (*local* methods), $\hat{\mathcal{I}}$ is the $k^*(\hat{t})$ channels with the lowest per-dim MASS distance at the predicted position $\hat{t}$.

Figure 2 shows HR@1 for each (condition, retriever) cell. `contrast_norm` achieves the highest mean HR@1 across the six conditions (0.88), followed by `weighted_dim` (0.84) and `mstamp_mdl` (0.81). No single method dominates on all conditions: `mv_mass` is competitive on decoy dimensions; `mstamp_mdl` matches `contrast_norm` on easy, noisy channels, and focused dims. The asymmetric-SNR condition is where methods separate most clearly, because it simultaneously penalizes global-weight aggregation (`weighted_dim`), equal-channel aggregation (`mv_mass`), and single-dim selection (`mstamp_mdl`), while leaving room for `contrast_norm` to approach the oracle. Those results show that `contrast_norm` with its per-position channel selection is a good asset to approach the oracle upper bound: it maintains near-oracle retrieval across noise regimes where global-weighting and equal-aggregation strategies degrade, while remaining computationally competitive with all baselines. The `mstamp_mdl` and `weighted_dim` strategies also prove to be viable candidates for dimension selection in the needle in a timeseries haystack problem.

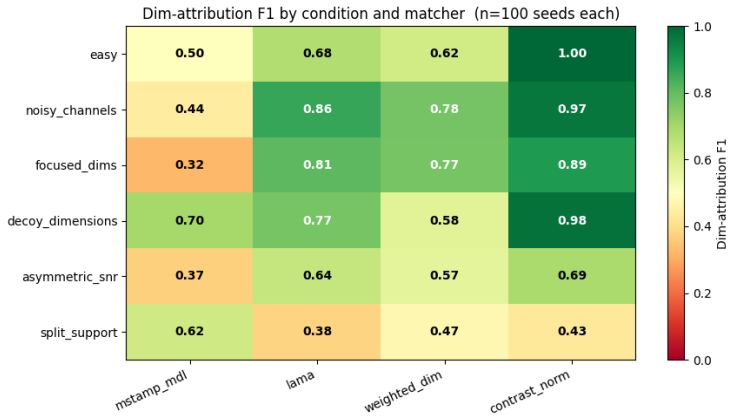

Fig. 3: Dim-attribution F1 by condition and retrieval approach (50 seeds each, $n = 5\,000$). `euclidean` and `mv_mass` are excluded as they do not produce channel attributions.

Figure 3 shows dim-attribution F1 per (condition, retriever) cell; the picture largely mirrors HR@1, with two instructive divergences. First, `mstamp_mdl` achieves non-trivial HR@1 on several conditions yet has structurally low F1 throughout: its MDL stopping rule selects $k^* = 1$ while $n_{\text{inf}} = 4$, yielding high precision but near-zero recall regardless of retrieval outcome. Second, under *asymmetric SNR*, `contrast_norm` maintains both retrieval and attribution quality whereas `weighted_dim` degrades on both in parallel - its global weights systematically under-represent low-amplitude informative channels at both the retrieval and the attribution threshold. The *split-support* condition is the only one where all methods show uniformly low F1: the chimeric channel structure is irrecoverable without knowing the partition beforehand, reflected equally in attribution and retrieval.

Additional details about these six synthetic presets, supporting details about this experiment, and extended experiments are given in the Appendix Sections 5.2, 5.3.

`contrast_norm` proves to be the most robust choice when background processes are coherent and recurring: local per-position selection resists the failure mode that reliably misleads global aggregators. `mstamp_mdl`'s noise robustness is a different property: minimal aggregation rather than adaptive selection, and comes at the cost of near-zero attribution recall; it may be preferable only when channel count is very low or the background is structurally noisy.

3.2 Stage 2: Contextualised Textual Interpretation

This section isolates the model-selection question for Stage 2: *given the informative channels $\mathcal{I}^*$, which LLM backend best exploits a historical reference to characterise a time series event?* Channel attribution is treated as oracle throughout

so that results reflect only the model's capacity to leverage historical context, not the quality of upstream retrieval.

Directly evaluating the full structured output $P = \{(i, \psi_i, a_i)\}$ would require annotated ground truth per channel; instead we use binary Ψ-pattern classification as a proxy: each model decides whether a pattern from the vocabulary Ψ (`peak`/`trough`, `monotonic_rise`/`monotonic_fall`, `noise_burst`) is present in Q. Binary detection is a necessary condition for structured description: a model that cannot identify a pattern type cannot reliably assign it a label ψ_i or a signed deviation a_i. Each strategy is tested in two variants: *with context* ($N \geqslant 1$ reference segments from T are provided alongside the query) and *without* ($N = 0$), to measure how much performance depends on the contrastive reference. We report per-pattern accuracy and true-negative rate (TNR): the fraction of clean segments correctly left unflagged, measuring robustness to false alarms on naturally noisy signals.

We use the LBNL FCU dataset [7], high-quality simulated fan-coil-unit sensor signals, and introduce controlled plausibility probes covering three pattern types from the vocabulary Ψ (Table 1): `peak`/`trough` (a sharp, non-physically-plausible spike absent from the historical baseline); `monotonic_rise`/`monotonic_fall` (a sustained directional drift inconsistent with stationary reference behaviour); and `noise_burst` (a high-frequency power increase distinguishable from natural measurement noise by contrast with T). Each probe type is synthetically inserted into otherwise fault-free segments, ensuring ground-truth labels remain unambiguous. The binary evaluation collapses the sign distinction within each type (e.g., rise vs. fall, peak vs. trough), since the task is detection, not direction. Crucially, the original FCU signals exhibit natural non-stationary behavior: transient spikes, slow drifts, and measurement noise, making the task genuinely challenging: models must identify contextual inconsistency relative to T, not merely flag statistical outliers.

We use "Fault Free" segments only; each 20-min signal at 1 Hz yields 2,800 non-overlapping 1,000-point windows. Probes are inserted into half; each run draws 150 clean and 150 corrupted per pattern type (450 queries per condition, balanced).

Text-LLM (Mistral-Small-3.2 [12], primary): Per-channel statistics (amplitude z-score, trend slope, variance ratio, lag-1 autocorrelation) are derived from Q and from each reference segment, both z-scored against the reference baseline. The JSON context is augmented with per-channel natural-language descriptions produced by ChatTS, grounding the symbolic statistics in signal-level language. The statistic and description blocks are serialised as JSON and fed to an instruction-tuned LLM with a schema-constrained prompt.

Visual prompting (VLM: Qwen3-VL-30B-A3B-Instruct [1]): One subplot per channel shows Q_i overlaid on a $\pm 1\sigma$ band derived from the reference segments R_i. The figure is rendered with matplotlib[3] and fed as a visual input to the VLM with a schema-constrained prompt.

[3] `https://matplotlib.org/`

TS-LLM prompting (ChatTS [23]): Q and the reference segments are tokenised and fed to ChatTS in its native multi-content message format. The model reasons directly over raw channel values without pre-computed statistics.

Additional details are provided in Appendix §5.4 and §5.5.

Model name	Overall accuracy	Accuracy spike	Accuracy drift	Accuracy noise	TNR
ChatTS	59	73	58	46	65
Qwen3-VL	59	67	49	62	18
Mistral-small	60	68	51	62	23

Table 2: Performance without historical context (all values in %).

Without context (Table 2), all three models score 59–60%, barely above the 50% chance level. Spike detection is strongest (> 67%) and drift hardest (46–58%), consistent with drift requiring contextual comparison rather than local shape recognition. TNR diverges widely (18–65%): ChatTS is cautious and rarely over-flags; Qwen3-VL and Mistral-small are strongly biased toward the anomalous class.

Model name	Overall accuracy	Accuracy spike	Accuracy drift	Accuracy noise	TNR
ChatTS	50	56	48	47	56
Qwen3-VL	59	64	49	63	32
Mistral-small	65	71	67	57	73

Table 3: Performance with 3 reference segments (all values in %).

Adding context (Table 3, Figure 4) separates the models sharply. Mistral-small improves steadily from 62% (1 segment) to 69% (7 segments) with TNR recovering to 73%, confirming that structured text reasoning integrates additional context coherently. ChatTS drops from 59% to 50% as context is added and Qwen3-VL shows no consistent trend, indicating that multi-segment visual or signal-tokenised input actively degrades rather than aids reasoning. These results establish Mistral-small as the preferred Stage 2 backend and confirm that raw visual encoding is insufficient for context-aware plausibility assessment.

3.3 End-to-End Evaluation

The E2E experiment exercises the full CLARTÉ pipeline on a scenario where neither t^* nor $\mathcal{I}^*$ are given: Stage 1 must recover both before Stage 2 can char-

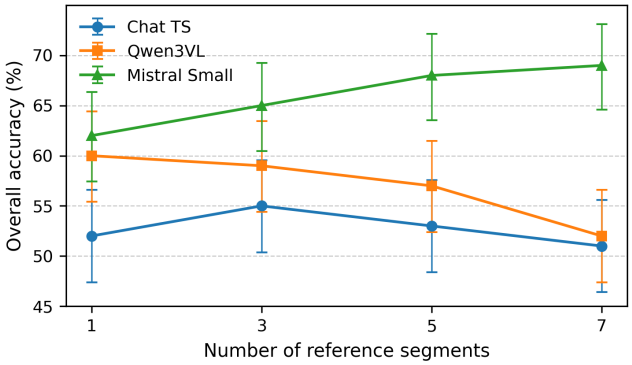

Fig. 4: Overall answer accuracy depending on the number of provided reference segments (1000-long historical signals). 95% confidence intervals have been calculated using Wilson confidence interval method.

acterise anything. Let $T \in \mathbb{R}^{n \times d}$ be one month of fault-free building sensor data ($d = 29$, $n \approx 8\,640$ at 5-minute resolution) from a Fan Coil Unit. One target fault window F_1 of length $m = 288$ is injected at a known position t^*, and $N = 5$ confounder windows from other fault types are injected at random non-overlapping positions. The query is $Q = F_1 + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0,\ 0.1 \cdot \mathrm{std}(F_1))$ applied channel-wise. A logbook $\mathcal{L}$ records injection positions and fault labels, available at inference time. Results are averaged over 10 seeds per fault (40 runs per competitor).

We use the LBNL Fan Coil Unit (FCU) dataset [7] again ($d = 29$ channels) and since $\mathcal{I}^*$ is not directly observable, we construct a data-driven proxy: **shape** $\mathcal{I}^*$, the mean $(1 - \rho_P)$ over $n_s = 300$ paired fault/fault-free windows of length m, where ρ_P is the Pearson correlation of z-normalised windows. The top-3 channels by this score will also serve as the oracle attribution. Four fault types spanning distinct physical subsystems are evaluated: **Filter restriction (50%):** increased air resistance reduces discharge airflow; the fan compensates with a speed increase. **Cooling valve stuck fully open:** the chilled-water valve stays open regardless of command; the coil overcools. **Zone temperature sensor bias (+4°C):** a biased zone reading propagates into the control loop, triggering spurious setpoint adjustments. **Outdoor damper stuck closed:** outdoor airflow collapses while the damper command remains active.

We compare the following pipelines that use the text-LLM Stage 2 backend (Mistral-Small-3.2 augmented with ChatTS pre-annotation) and differ only in Stage 1: `contrast_norm`, `weighted_dim`, `oracle` where shape $\mathcal{I}^*$ dims is fed directly, and `random` that uniformly selects random position and dimensions. Comparing oracle to `contrast_norm` isolates the cost of Stage 1 attribution imperfection while holding Stage 2 fixed; random vs. `contrast_norm` quantifies the effect of channel selection quality while holding context fixed.

We use the following metrics: **M1 HR@5** measures the fraction of runs where any top-5 candidate falls within $\tau = m/4 = 72$ steps of t^*, **M2 ρ-shape** is the Spearman rank correlation between Stage 1 importance scores and per-channel shape discriminability $\mathcal{I}^*$, **M3 Fault Acc.** is the logbook hit rate at rank-1, and **M4 Channel** ($\in [0, 3]$) is a deterministic per-channel judge for Stage 2. It uses ground-truth labels ψ_c and amplitude signs derived from the fault-free reference at t^*; 1 point if the predicted ψ_i matches, 0.5 if only the sign agrees, 0 otherwise; evaluated on the top-3 channels by $\mathcal{I}^*$.

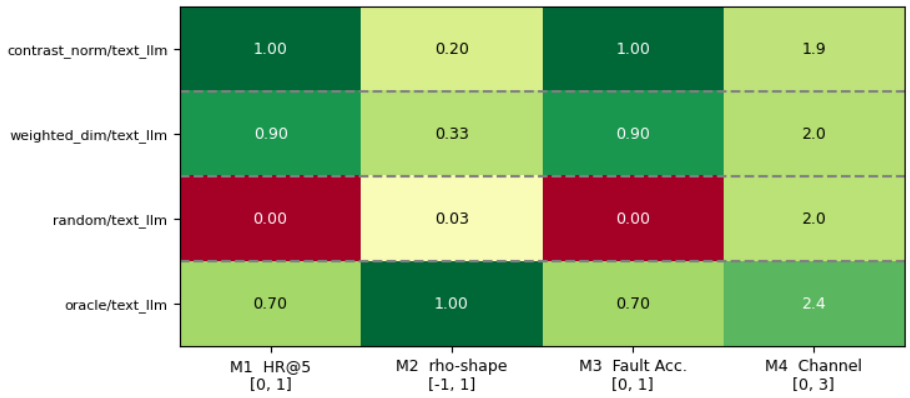

Fig. 5: End-to-end evaluation on the LBNL FCU dataset [7], macro-averaged over 4 fault types and 10 seeds (40 runs per competitor). The *oracle* row gives the Stage 1 ceiling; *random* gives the floor. Colours: M1, M3 in $[0, 1]$; M2 mapped from $[-1, 1]$ to $[0, 1]$; M4 divided by 3.

Results are summarised in Figure 5. `contrast_norm` achieves perfect retrieval across all faults and seeds. The `oracle`, despite receiving the exact three most informative channels, scores M1 = 0.70: it fails on the outdoor-damper fault (HR@5 = 0.1), where the dominant effect is a step change that saturates within the query window, making even the ideal channel set insufficient for subsequence matching. `weighted_dim` scores 0.9 on both (failing only on that same fault). On dimension attributions (M2), globally weighted attribution aligns better on average with the true fault signature $\mathcal{I}^*$, even when it does not always find the right timestamp. Contrastive per-position selection excels at retrieval precisely because it re-evaluates the channel subset locally, but this makes its global attribution noisier. On Stage 2 description quality (M4), the **oracle** leads over all retrieval competitors: having the genuinely anomalous channels is the main driver of description accuracy. M4 and M1/M3 are thus complementary: good retrieval does not guarantee good description, but finding the right channels, as the oracle does, remains decisive for Stage 2 quality.

4 Conclusion

We presented CLARTÉ, a training-free two-stage pipeline for explainable multivariate subsequence search. Stage 1 aggregates per-channel MASS distance profiles via two complementary strategies: a weighted (global) approach and a contrasted (local, per-position) approach, both yielding channel attributions alongside match positions. Stage 2 grounds a text-LLM call in a fault-free reference window anchored to the rank-1 match, producing a structured per-channel pattern description.

Results and limitations. Contrasted aggregation achieves the highest mean HR@1 (0.88) and is the most robust Stage 1 strategy under coherent background processes; weighted aggregation (0.84) is the strongest global alternative; mSTAMP's MDL rule forces $k^* = 1$, collapsing attribution recall. A structural failure mode remains when informative channels are split across disjoint subsets: no method recovers attribution without knowing the partition. Without historical context all Stage 2 models score near chance; with reference segments Mistral-small improves to 69% while ChatTS and Qwen3-VL show no consistent benefit, confirming that structured text reasoning is the only approach that exploits contrastive context. Stage 2 evaluation is indirect: the binary detection proxy uses ChatTS-derived labels rather than expert annotation, and the pattern vocabulary Ψ is fixed. End-to-end on LBNL FCU, the full pipeline recovers fault positions and attributions against a pool of 47 confounders; a focused experiment rather than a large-scale benchmark (4 faults, 10 seeds, 40 runs per competitor).

Perspectives. The most direct extension is to replace the binary Stage 2 evaluation with expert-annotated per-channel labels, enabling a full assessment of the structured output P. On the retrieval side, the k-selection criterion for contrasted aggregation is currently a z-score heuristic; a natural follow-up would be learning it or replacing it with a significance test anchored to the MSig framework [19]. Longer-term directions include multi-query or multi-event detection (locating several distinct fault types simultaneously), integration with model-based diagnosis pipelines for validation against physics-based channel attributions, and extension to the variable-length query setting.

References

1. Bai, S., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bijlani, N., Villarroel, M.: Motif-based morphology signatures for interpretable ECG screening and monitoring. In: Proceedings of the IEEE Engineering in Medicine and Biology Conference (EMBC) (2026), arXiv:2606.00107
3. Cartwright, E., Crane, M., Ruskin, H.J.: Financial time series: market analysis techniques based on matrix profiles. Engineering Proceedings **5**(1), 45 (2021). https://doi.org/10.3390/engproc2021005045

4. Chan, S., Hang, Y., Tong, C., Acquah, A., Schonfeldt, A., Gershuny, J., Doherty, A.: CAPTURE-24: A large dataset of wrist-worn activity tracker data collected in the wild for human activity recognition. Scientific Data **11**(1), 1135 (2024). https://doi.org/10.1038/s41597-024-03960-3
5. Daswani, M., Bellaiche, M.M., Wilson, M., Ivanov, D., Papkov, M., Schnider, E., Tang, J., Lamerigts, K., Botea, G., Sanchez, M.A., Patel, Y., Prabhakara, S., Shetty, S., Telang, U.: Plots unlock time-series understanding in multimodal models. arXiv preprint arXiv:2410.02637 (2024)
6. d'Hondt, J.E., Papapetrou, O., Palpanas, T., Paparrizos, J.: MS-Index: Fast top-k subsequence search for multivariate time series under Euclidean distance. arXiv preprint arXiv:2512.14723 (2025)
7. Granderson, J., Lin, G., Chen, Y., Casillas, A., Im, P., Jung, S., Benne, K., Ling, J., Gorthala, R., Wen, J., Chen, Z., Huang, S., Vrabie, D.: Lbnl fault detection and diagnostics datasets. Open Energy Data Initiative (OEDI), Lawrence Berkeley National Laboratory (2022). https://doi.org/10.25984/1881324, https://data.openei.org/submissions/5763, accessed: 2026-06-09
8. Guerrini, V., Germain, T., Truong, C., Oudre, L., Boniol, P.: Time series motif discovery: A comprehensive evaluation. Proceedings of the VLDB Endowment (2025), hAL: hal-05218910
9. He, Z., Alnegheimish, S., Reimherr, M.: Harnessing vision-language models for time series anomaly detection. In: Proceedings of the AAAI Conference on Artificial Intelligence (2026)
10. Kobsar, D., Ferber, R.: Wearable sensor data to track subject-specific movement patterns related to clinical outcomes using a machine learning approach. Sensors **18**(9), 2828 (2018). https://doi.org/10.3390/s18092828
11. Komitova, R., Raabe, D., Rein, R., Memmert, D.: Time series data mining for sport data: a review. International Journal of Computer Science in Sport **21**(2), 17–31 (2022). https://doi.org/10.2478/ijcss-2022-0008
12. Liu, A.H., Khandelwal, K., Subramanian, S., Jouault, V., Rastogi, A., Sadé, A., Jeffares, A., Jiang, A., Cahill, A., Gavaudan, A., et al.: Ministral 3. arXiv preprint arXiv:2601.08584 (2026)
13. Moody, G.B., Mark, R.G.: The impact of the MIT-BIH arrhythmia database. IEEE Engineering in Medicine and Biology Magazine **20**(3), 45–50 (2001). https://doi.org/10.1109/51.932724
14. Mueen, A., Zhong, S., Zhu, Y., Yeh, M., Kamgar, K., Viswanathan, K., Gupta, C., Keogh, E.: The fastest similarity search algorithm for time series subsequences under euclidean distance (August 2022), http://www.cs.unm.edu/~mueen/FastestSimilaritySearch.html
15. Murad, M.M.N., et al.: Cluster-aware causal mixer for online anomaly detection in multivariate time series. arXiv preprint arXiv:2506.00188 (2025)
16. Nilsson, F., Bouguelia, M.R., Rögnvaldsson, T.: Practical joint human-machine exploration of industrial time series using the matrix profile. Data Mining and Knowledge Discovery **37**, 1–38 (2023). https://doi.org/10.1007/s10618-022-00871-y
17. Pelok, B., d'Hondt, J.E.: MULISSE: Variable-length similarity search for multivariate time series. In: Proceedings of the 41st IEEE International Conference on Data Engineering. ICDE (2025)
18. Schäfer, P., Leser, U.: Discovering leitmotifs in multidimensional time series. Proceedings of the VLDB Endowment **18**(2), 377–389 (2025)
19. Silva, M.G., Madeira, S.C., Henriques, R.: On why and how statistical significance criteria can guide multivariate time series motif analysis. Pattern Recognition Letters (2026). https://doi.org/10.1016/j.patrec.2026.04.020

20. Wang, J., Zhu, Y., Li, S., Wan, D., Zhang, P.: Multivariate time series similarity searching. The Scientific World Journal **2014**, 851017 (2014). https://doi.org/10.1155/2014/851017
21. Wu, X., Qiu, X., Li, Z., Wang, Y., Hu, J., Guo, C., Xiong, H., Yang, B.: CATCH: Channel-aware multivariate time series anomaly detection via frequency patching. In: International Conference on Learning Representations (2025), https://arxiv.org/abs/2410.12261
22. Xiao, M., Jiang, Z., Qian, L., Chen, Z., He, Y., Xu, Y., Jiang, Y., Li, D., Weng, R.L., Peng, M., Huang, J., Ananiadou, S., Xie, Q.: Retrieval-augmented large language models for financial time series forecasting. arXiv preprint arXiv:2502.05878 (2025)
23. Xie, Z., Li, Z., He, X., Xu, L., Wen, X., Zhang, T., Chen, J., Shi, R., Pei, D.: ChatTS: Aligning time series with LLMs via synthetic data for enhanced understanding and reasoning. In: Proceedings of the VLDB Endowment. vol. 18, pp. 2385–2398 (2025)
24. Xu, X., Wang, H., Liang, Y., Yu, P.S., Zhao, Y., Shu, K.: Can multimodal LLMs perform time series anomaly detection? In: Proceedings of the ACM Web Conference (WWW) (2026)
25. Yang, Y., Liu, Z., Song, L., Ying, K., Wang, Z., Bamford, T., Vyetrenko, S., Bian, J., Wen, Q.: Time-RA: Towards time series reasoning for anomaly diagnosis with LLM feedback. arXiv preprint arXiv:2507.15066 (2025), kDD '26
26. Yeh, C.C.M., Kavantzas, N., Keogh, E.: Matrix profile VI: Meaningful multidimensional motif discovery. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). pp. 565–574 (2017). https://doi.org/10.1109/ICDM.2017.66
27. Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Silva, D.F., Mueen, A., Keogh, E.: Matrix profile I: All pairs similarity joins for time series: A unifying view that includes motifs, discords and shapelets. In: Proceedings of the IEEE International Conference on Data Mining (ICDM). pp. 1317–1322 (2016)
28. Zhuang, J., Yan, L., Zhang, Z., Wang, R., Zhang, J., Gu, Y.: See it, think it, sorted: Large multimodal models are few-shot time series anomaly analyzers. arXiv preprint arXiv:2411.02465 (2024)
29. Zumarraga, N., Kaar, T., Wang, N., Xu, M.A., Rosenblattl, M., Kreft, M., O'Sullivan, K., Schmiedmayer, P., Langer, P., Jakob, R.: TS-Haystack: A multi-scale retrieval benchmark for time series language models. arXiv preprint arXiv:2602.14200 (2026)

5 Appendix

Symbol	Description
Signals	
$Q \in \mathbb{R}^{m \times d}$	Query event
$T \in \mathbb{R}^{n \times d}$	Target signal ($n \gg m$)
m, n, d	Query length, target length, number of channels
$T[t : t + m]$	Subsequence of T of length m starting at position t
$R \in \mathbb{R}^{m \times d}$	Normal reference window sampled from T
$\mathcal{M}$	Optional domain metadata (channel names, units, failure modes)
Logbook	
$\mathcal{L} = \{(s_k, \ell_k)\}_{k=1}^{K}$	Logbook of K past fault records
$s_k \in \{1, \ldots, n\}$	Start position of past event k in T
$\ell_k \in \Lambda$	Fault label of past event k; Λ is the label set
Ground truth	
t^*	True match position in T
$\mathcal{I}^* \subsetneq \{1, \ldots, d\}$	True informative channel set
Stage 1: distance profiles	
$\pi_i(t)$	Per-channel MASS distance profile for channel i
w_i, $\hat{w}_i$	Raw and ℓ_1-normalized channel weight (weighted aggregation)
Stage 1: output	
$\mathcal{C} = \{(t_j, D(t_j), \mathcal{I}_j)\}_{j=1}^{J}$	Ranked candidate set
$\hat{t}_1$	Top-1 predicted match position
$\hat{\mathcal{I}}$, $\mathcal{I}_j$	Predicted informative channels (global / per candidate j)
Stage 2: outputs	
$\hat{\ell} \in \Lambda \cup \{\texttt{normal}\}$	Derived fault label
$P = \{(i, \psi_i, a_i)\}_{i \in \mathcal{I}_1}$	Per-channel description
$\psi_i \in \Psi$	Pattern label (vocabulary in Table 1)
$a_i \in \mathbb{R}$	Signed amplitude deviation of Q_i relative to R_i

Table 4: Notation used throughout the main paper.

5.1 Related Work

We organize related work along three axes: multivariate subsequence search and motif discovery, dimension selection for multivariate time series, and LLM-based temporal reasoning.

Multivariate Subsequence Search and Motif Discovery

Matrix profile and MASS. The matrix profile [27] provides a data structure that records, for every subsequence in a time series, the distance to its nearest neighbor. MASS (Mueen's Algorithm for Similarity Search) [14] is the $O(n \log n)$ FFT-based subroutine that computes a single *distance profile*-the distance from one query subsequence to every position in a longer series under z-normalized Euclidean distance. The full matrix profile is obtained by computing a distance profile for every subsequence, yielding $O(n^2 \log n)$ overall cost. Our work builds directly on MASS as the per-dimension primitive, but restricts attention to the *single-query* setting (one reference event matched against a target signal), avoiding the quadratic cost of full matrix profile construction.

Multivariate matrix profile (mSTAMP). Yeh et al. [26] showed that naively searching for motifs across *all* dimensions of a multivariate time series fails in most realistic scenarios, and introduced mSTAMP, which discovers multi-dimensional motifs by sorting per-dimension distances and selecting the best k-dimensional subset at each position. This per-position dimension-sorting idea is conceptually close to our *contrasted* aggregation, but mSTAMP operates in the self-join (all-pairs) setting at $O(n^2 d)$ cost and uses MDL-based model selection to choose k across motif pairs. We adapt the dimension-sorting concept to the single-query case and replace MDL with a lightweight z-normalization heuristic that makes different values of k comparable at negligible additional cost.

Weighted BORDA voting. Wang et al. [20] proposed a dimension-combination method for MTS similarity search: per-dimension similarities are computed independently, then synthesized via weighted BORDA voting using PCA eigenvalues as weights. This is conceptually close to our *weighted* profile, but with a key difference: their weights are derived from PCA eigenvalues (a global property of the reference signal, independent of the query), whereas our profile-contrast weights adapt to each specific query by measuring how discriminative each dimension's distance profile is. Their approach also operates in PCA-transformed space, whereas ours works in the original feature space, preserving per-dimension interpretability.

Dimension Selection for Multivariate Time Series

LAMA. Schäfer and Leser [18] introduced LAMA, a leitmotif discovery algorithm that jointly performs dimension selection and motif mining. LAMA selects each subsequence as a candidate and determines the most suitable dimensions by sorting k-NN distances in ascending order, selecting those with the lowest distances. The paper explicitly criticizes approaches that perform dimension selection and pattern discovery independently. Our contrasted method shares the per-position dimension-sorting mechanism and per-dimension z-normalization, but differs in two key respects:

- (i) we replace extent-based k-selection with a z-score stopping criterion on the cumulative sorted profile, making subset sizes directly comparable across positions without a separate optimization pass, and

– (ii) we propose a complementary profile-contrast weighting strategy, absent from LAMA's framework, that assigns query-adaptive per-channel importance weights.

We include a single-query adaptation of LAMA's extent stopping as a direct competitor in our Stage 1 benchmark (§3.1), comparing the two stopping criteria under identical per-dimension MASS inputs.

MS-Index. d'Hondt et al. [6] introduced MS-Index, a nearest-neighbor MTS subsequence search algorithm under Euclidean distance that supports ad-hoc selection of query channels and achieves sublinear query-time scaling in the number of channels. Crucially, MS-Index assumes the relevant channels are *specified at query time* by the user. Our methods address a complementary problem: *discovering* which channels are informative when this is not known a priori. The two approaches are naturally composable-our Stage 1 could identify informative channels, which are then passed to MS-Index for efficient retrieval over large MTS databases.

Variable-length subsequence search. MULISSE [17] extends the univariate ULISSE algorithm to multivariate time series, enabling exact variable-length k-NN subsequence search via multivariate symbolic approximations with refined node splitting. Like MS-Index, it assumes query channels are specified by the user; unlike MS-Index, it supports flexible query lengths and operates on symbolic rather than continuous representations, which precludes direct comparison under z-normalized Euclidean distance but broadens its applicability to settings where the query length is not fixed in advance.

Statistical significance of motifs. Silva et al. [19] proposed MSig, a statistical framework for assessing motif significance in multivariate time series using binomial testing, showing that up to 60% of discoveries by state-of-the-art algorithms are spurious under normative search conditions; Guerrini et al. [8] corroborate this finding across a comprehensive independent evaluation of motif discovery methods. Our contrasted method's support $K(t)$ – the number of dimensions selected at the predicted position – is a direct input to MSig's binomial test: given d total dimensions and an empirical per-dim null, MSig would assess whether $K(t)$ agreeing dimensions at position t is significant. The composability is architecturally clean: our per-dim MASS profiles are precisely the inputs MSig requires. The main open problem is the null distribution in the AB-join (single-query) setting, where MSig's self-join assumptions do not directly apply; deriving it from the empirical distribution of per-dim profiles is a concrete integration path left to future work.

LLM-Based Temporal Reasoning

Time series language models. ChatTS [23] treats time series as a modality analogous to images, aligning multivariate time series with LLMs via synthetic training data. While effective at understanding and reasoning about short segments, ChatTS and similar TS-LLMs operate end-to-end without explicit channel attribution, making their explanations difficult to verify. Our pipeline provides a structured decomposition: a transparent numeric stage identifies *where* and *which channels*, and the TS-LLM stage explains *why*-grounding textual interpretation in reproducible dimension attributions.

Visual prompting for time series. Recent work has shown that VLMs can reason over plotted time series without time-series-specific training. Daswani et al. [5] show that feeding line-plot images to GPT-4o and Gemini outperforms text serialisation by up to 150% on real-world sensor classification tasks (fall detection, activity recognition), providing empirical grounding for plot-as-input pipelines. Xu et al. [24] systematically benchmark eight MLLMs on a 12.4k-image anomaly dataset with variate-wise (per-channel) evaluation, finding that MLLMs excel at coarse-grained anomalies but underperform traditional methods on fine-grained localisation. VLM4TS [9] chains a lightweight ViT screener with a VLM verifier in a zero-shot two-stage architecture. TAMA [28] supplies normal-window images as few-shot in-context examples so the VLM can characterise what makes the anomalous window stand out: the closest prior work to our Stage 2 contrastive framing. The key distinction from our approach is that none of these methods grounds the VLM call in a formally retrieved match position: the anomalous window is assumed to be given, and reference windows are user-supplied examples rather than windows anchored to a Stage 1 match position in a specific target signal.

Long-context retrieval. TS-Haystack [29] introduced a needle-in-a-haystack retrieval benchmark on the Capture24 accelerometer dataset [4], showing that existing TSLMs struggle with temporal localization as context length increases. This motivates our pipeline design: rather than relying on the TS-LLM's internal attention for localization, we use MASS-based retrieval (Stage 1) as a dedicated, scalable backbone, freeing the LLM to focus on reasoning over the retrieved candidates.

Anomaly reasoning. Time-RA [25] reformulated anomaly detection as a generative reasoning task, introducing RATs40K, a benchmark of $\sim$40,000 multivariate anomaly samples with structured (Observation $\rightarrow$ Thought $\rightarrow$ Action) annotations across 10 domains. We leverage this benchmark to evaluate our end-to-end pipeline: Stage 1 localizes and attributes the anomaly; Stage 2 generates explanations evaluated against RATs40K's ground-truth reasoning.

Channel-aware anomaly detection. Murad et al. [15] address the same "some channels are more equal than others" intuition for anomaly detection, grouping channels into clusters via a learned causal mixer. CATCH [21] takes a more fine-grained approach, learning per-frequency-band channel masks within a Channel

Fusion Module that suppresses irrelevant channels at different frequency resolutions – also from labeled data. Our approach is the unsupervised, training-free counterpart for subsequence search: rather than learning channel groups from labeled data, we infer per-query channel importance directly from the distance profiles.

5.2 Stage1 Synthetic Benchmark: and Multi-Condition details.

Here are details about the synthetic presets described in Table 5.

Condition	Noise ratio	Signal structure
Easy	0%	Control: all channels informative, equal amplitude
Noisy channels	60%	Majority of channels replaced by independent AR(1) noise
Focused dims	85%	Equal-amplitude signal; high noise channel fraction
Decoy dimensions	adversarial	Noise channels amplitude-matched to mimic signal envelope
Asymmetric SNR	∼75%	Signal amplitude varies across informative channels
Split support	chimera	Query event split across two disjoint channel subsets

Table 5: Six preset conditions for the synthetic benchmark.

Easy. All d channels are informative and carry equal-amplitude signal. `euclidean` already fails here (HR@1 = 0.72) because raw Euclidean distances are not amplitude-normalized, making the aggregation sensitive to scale differences across seeds. All MASS-based methods achieve HR@1 = 1.00.

Noisy channels. 60% of channels are replaced by AR(1) noise. `euclidean` collapses to 0; dimension-aware methods are unaffected (`mv_mass` 0.98, all selective methods 1.00) because per-channel z-normalization suppresses the noise channels' contribution.

Focused dims. 85% noise, equal-amplitude piecewise-constant signal on the informative channels. `mv_mass` degrades to 0.56 because the noise channels outweigh the signal in the equal-weight aggregation. Selective methods (`contrast_norm` 0.98, `weighted_dim` 0.96, `mstamp_mdl` 0.96) correctly concentrate on the informative subset.

Decoy dimensions. Noise channels are shaped to mimic the signal in raw-amplitude space, creating adversarial false peaks. `mv_mass` is surprisingly robust here (0.94): the decoy signal does not replicate the query's z-normalized waveform, so MASS-normalized profiles are less affected than raw distances. `mstamp_mdl` drops to 0.66, likely because its global-minimum dim ordering ranks decoy channels ahead of genuinely informative ones.

Asymmetric SNR. Informative channels carry signals of *varying amplitude*, some strong, some weak, while noise channels are amplitude-matched. This is the key discriminating condition: `contrast_norm` achieves 0.90 by adapting its channel selection per-position, whereas `weighted_dim` (0.66) is penalised because its global weights under-represent low-amplitude informative channels, and `mv_mass` (0.44) degrades as the signal-to-noise contrast varies.

Split support (chimera). The query event is a chimera: two sub-patterns occupy disjoint channel subsets. No single channel-selection strategy simultaneously covers both subsets, and all unsupervised methods plateau below 0.55 (`mstamp_mdl` 0.52, `weighted_dim` 0.54, `contrast_norm` 0.40, `mv_mass` 0.42). `oracle_dim` reaches 1.00, confirming the task is solvable when the full channel set is used: the difficulty is purely in unsupervised dim selection.

5.3 Stage 1 Synthetic Benchmark: additionnal experiments.

Statistical Validation and Runtime for HR@1 by Condition and Retriever: Figure 6, Table 6 Friedman test with post-hoc Nemenyi analysis ($\alpha = 0.05$, CD= 0.435) across the pooled $6 \times 50 = 300$ samples from the multi-condition sweep. The Friedman test rejects the null ($p \ll 0.001$). Post-hoc Nemenyi reveals two clear results: `euclidean` (mean rank 5.33) is significantly worse than every other method which is evidently expected; and `contrast_norm` (MR 3.05) is the only unsupervised method not significantly different from `oracle_dim` (MR 2.69, gap = 0.36 < CD). `mv_mass` (MR 3.52) is significantly worse than `contrast_norm` and `oracle_dim` but indistinguishable from `mstamp_mdl` and `weighted_dim`, which form a middle tier.

Mean wall-clock time per query (single CPU core): all MASS-based methods use FFT-accelerated computation (via STUMPY) and are 2-26× faster than `euclidean`, which computes distance profiles via naive sliding-sum. `oracle_dim` is fastest because it processes only $n_{\text{inf}} \ll d$ channels.

HR@1 vs Noise Ratio under Asymmetric SNR. The multi-condition heatmap in Section 3.1 shows that the asymmetric-SNR condition is the setting where methods separate most clearly. This section details continuous degradation under that condition alone, sweeping the noise ratio across a wide range to characterize the rate of degradation and the dim-attribution behavior of each method.

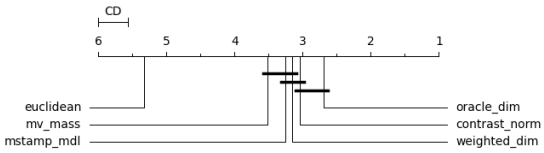

Fig. 6: Critical-difference diagram (Friedman + Nemenyi, $\alpha = 0.05$). Lower average rank is better; connected groups are not significantly different.

Retriever	Mean (s)
`euclidean`	0.1678
`mv_mass`	0.0139
`mstamp_mdl`	0.0146
`lama`	0.0167
`weighted_dim`	0.0153
`contrast_norm`	0.0188
`oracle_dim`	0.0053

Table 6: Mean per-query compute time, averaged over all conditions (50 seeds each).

Setup. We fix $n_{\text{inf}} = 4$ informative channels and sweep the total channel count $d \in \{5, 6, 8, 10, 12, 15, 20, 30\}$, giving noise ratios $(d - n_{\text{inf}})/d$ from 0.20 to 0.87. All other settings follow Section 3.1: asymmetric per-channel amplitude on informative dims, amplitude-matched AR(1) noise on the rest, 50 seeds per value.

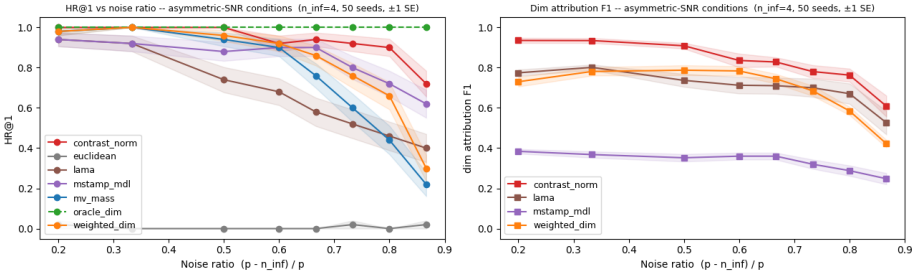

Fig. 7: HR@1 (left) and dim-attribution F1 (right) vs noise ratio under asymmetric-SNR conditions ($n_{\text{inf}} = 4$, d swept from 5 to 30, 50 seeds, ±1 SE). Dashed = `oracle_dim` upper bound.

HR@1. `contrast_norm` degrades last: it remains close to the oracle through noise ratio 0.60 before declining, separating clearly from `weighted_dim` (which is penalized by amplitude variability across informative channels) and `mv_mass` (equal-weight aggregation). `mstamp_mdl` tracks `mv_mass` in the HR@1 curve because its global-minimum dim ordering reduces to single-best-dimension MASS under asymmetric SNR; the globally sharpest channel is not always the most informative one.

Dim-attribution F1. `mstamp_mdl` has structurally low recall at all noise levels because it always outputs one predicted dim ($k^* = 1$) while $n_{\text{inf}} = 4$ - high precision

but systematically low recall. `contrast_norm` and `weighted_dim` both improve in dim F1 as noise ratio increases: with fewer informative channels relative to d, the channels that do carry signal stand out more sharply in the cumulative profile, making attribution easier. This counter-intuitive effect reflects a detection boundary: extreme noise ratios make the few surviving informative channels unambiguous, even as retrieval itself becomes harder.

5.4 Stage 2 : Description of prompts and visual inputs of baselines

The baseline prompts instruct models to classify the signal segment as either normal (no anomaly of the specified type present) or abnormal (the target anomaly is present). For instance, for the "noise" anomaly, the decision is binary: normal vs noise. The three prompts diverge only in how the input is provided to the model: ChatTS and Qwen3-VL receive the time series as raw inputs via the model call, while Mistral-small processes a textual representation of the signal — where numerical values are explicitly formatted and embedded in the prompt template (the "signal" variable).

Listing 1.1: Prompt template used for ChatTS baseline.

```
You are an expert in time series anomaly detection and
    categorization. We provide you of a fraction of a
    signal.
You are asked to determine whether the signal has the
    anomaly "{anomaly}" or not.

Below the description of the two labels to choose from :
{anomaly_types}

Give a short answer then indicate the correct label.
```

Listing 1.2: Prompt template used for Qwen3-VL baseline.

```
You are an expert in time series anomaly detection and
    categorization. We provide you the image of a
    fraction of a signal.
You are asked to determine whether the signal has the
    anomaly {anomaly} or not.

Below the description of the two labels to choose from :
{anomaly_types}

Give a short answer then indicate the correct label.
```

Listing 1.3: Prompt template used for Mistral-small baseline.

```
You are an expert in time series anomaly detection and
    categorization. We provide you the values of a
    fraction of a signal.
You are asked to determine whether the signal has the
    anomaly {anomaly} or not.

Below the description of the two labels to choose from :
{anomaly_types}

Here are the values of the signal to analyse :
{signal}

Give a short answer then indicate the correct label.
```

5.5 Stage 2 : Description of prompts and visual inputs of context-powered models

Prompts for the context-powered models are similar to the previous ones, with a special instruction to leverage the normal sequences in order to reason by contrast, and thus determine whether an anomaly is present in the original segment or not.

Figure 8 shows an example of a visual input with both the segment to classify and the reference context. It is visible that the segment of interest has a normal shape as it looks similar to the reference ones. Therefore it is possible for a vision language model to infere that the correct category to return is "normal".

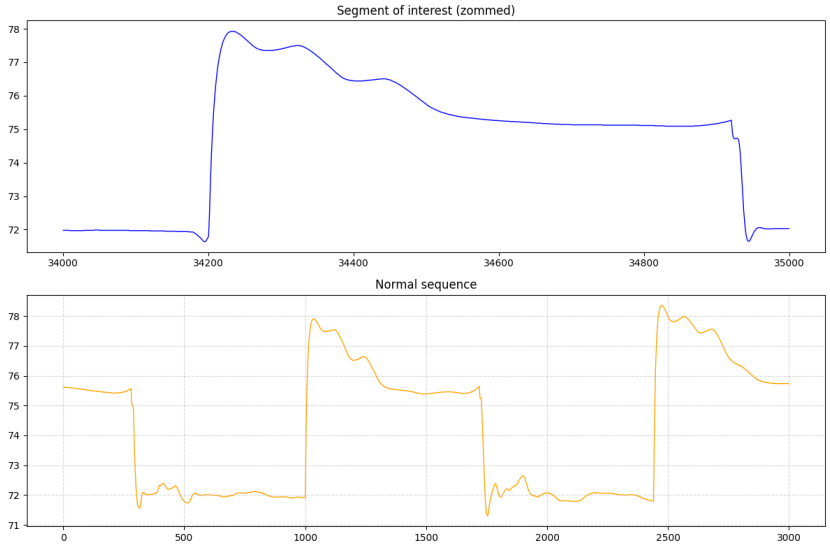

Fig. 8: Example of visual input for the context-powered visual language model. The above plot shows the signal to categorize, and below in the reference context which is here 3 segments long.