Structured Frequency-Domain Evidence for LLM-Based Time-Series Anomaly Detection

Jungwook Seo[1], Sangwon Son[1], Minjeong Kim[1], Seungmin Han[1], Seojin Yoo[2], Sungyong Baik[1,2,†]
[1]Department of Artificial Intelligence, Hanyang University
[2]Department of Data Science, Hanyang University
Seoul, Republic of Korea
{zungwooker, swson, mjkim0720, handsomemin, dbtjwls0821, dsybaik}@hanyang.ac.kr

Abstract

Time-series anomalies can appear not only as pointwise deviations but also as changes in recurring temporal structure, such as shifted periodicity or localized oscillatory fluctuations. However, existing LLM-based time-series anomaly detection methods mainly expose time-domain evidence through indexed values, plots, or de-seasonalized representations, leaving spectral structure implicit. We propose an evidence-augmented zero-shot TSAD framework that preserves indexed de-seasonalized observations while adding compact frequency-domain evidence computed with the Fast Fourier Transform (FFT). The evidence is constructed at two resolutions: global frequency-domain evidence summarizes sequence-level periodic context, while local frequency-domain evidence captures time-localized spectral departures. Experiments on AnomLLM with InternVL2-LLaMA3-76B, Qwen2.5-VL-72B-Instruct, Gemini-2.5-Flash, and GPT-4o, together with evaluation on the TSB-AD-U subset, show that explicit frequency-domain evidence improves LLM-based TSAD baselines. These results suggest that frequency-domain evidence can complement indexed and de-seasonalized time-domain inputs for zero-shot LLM-based TSAD.

1 Introduction

Beyond isolated spikes or pointwise deviations, anomalous intervals can appear as subtle changes in recurring temporal structure, such as altered cycles, shifted periodic components, or unstable local fluctuations. These patterns are naturally frequency-relevant: detecting them may require evidence about how the periodic structure of a sequence changes over time.

This issue is critical in realistic time-series anomaly detection (TSAD) (Blázquez-García et al., 2020; Pang et al., 2020), where anomaly labels, anomaly categories, and dataset-specific priors are often unavailable before deployment (Han et al., 2022). A zero-shot setting therefore requires the detector to localize anomalous intervals directly from the input sequence, without relying on benchmark-specific taxonomies, historical examples, or auxiliary supervision. This is stricter than training-free inference alone, since a method can avoid task-specific model training while still using external priors.

Recent studies have investigated large language models (LLMs) for TSAD, showing that LLMs can localize anomalous intervals without task-specific training from numerical sequences and, in some cases, visual context (Alnegheimish et al., 2024b; Liu et al., 2025; Zhou and Yu, 2025; Park et al., 2025). Existing LLM-based methods expose time series through textualized numerical sequences, visualized plots, or transformed inputs such as de-seasonalized sequences (Alnegheimish et al., 2024b; Liu et al., 2025; Zhuang et al., 2024), while LLM-TSAD preserves temporal indices to improve interval-level localization (Park et al., 2025). These designs help LLMs reason over temporal order and interval boundaries, but they do not explicitly summarize spectral properties such as dominant periods, spectral energy distribution, or time-localized frequency changes.

Frequency-domain modeling has long been useful in TSAD for capturing periodic structure, long-range dependencies, and anomalies that are weakly expressed in local time-domain values (Zhang et al., 2022; Lu et al., 2024; Wang et al., 2024; Nam et al., 2024). However, existing frequency-aware TSAD methods typically encode spectral information inside model-specific architectures, reconstruction objectives, or supervised pipelines (Zhang et al., 2022; Lu et al., 2024). This creates a representation-level evidence gap: spectral information may be useful for anomaly localization, but it is rarely exposed as explicit input evidence that a general-

†Corresponding author.

Probe	Metric	Indexed Text	Text + Plot	De-seasonalized	Global Freq.	Local Freq.
Single sinusoid	Top-1 Recovery (10%)	24.5	23.5	14.5	**95.5**	52.0
Multi sinusoid	Top-k Recovery (10%)	24.7	18.1	34.6	**91.9**	43.0
Local frequency change	Detection Acc.	11.5	80.5	88.0	8.5	**90.5**

Table 1: **Summary of frequency-reasoning probes.** Top-1 and Top-k Recovery use a 10% relative-error tolerance. Global Frequency and Local Frequency denote indexed text augmented with global and local frequency summaries, respectively.

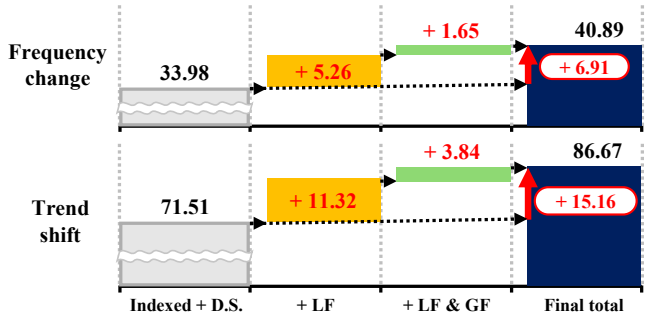

Figure 1: **Stepwise F1 gains from local (LF) and global (GF) frequency-domain evidence on representative hard anomaly types.** The waterfall plot shows incremental gains over the indexed de-seasonalized baseline.

purpose LLM can directly condition on.

We address this gap by constructing frequency-domain evidence at the input level. The proposed framework augments indexed de-seasonalized observations with compact global and local frequency-domain evidence computed from the input sequence using the Fast Fourier Transform (FFT). Global evidence captures the sequence-level periodic regime, while local evidence captures spectral behavior within overlapping temporal windows. The resulting input representation allows the LLM to condition on indexed time-domain observations and explicit frequency-domain evidence without using predefined anomaly categories, auxiliary supervision, or task-specific model training.

Our contributions are as follows:

- We identify a representation-level evidence gap in LLM-based TSAD: existing inputs help represent temporal order and interval boundaries, but leave frequency-relevant cues largely implicit.
- We propose a training-free evidence-augmented framework that constructs explicit global and local FFT-based summaries directly from the input sequence.
- We validate the framework on the Anom-LLM benchmark with multiple multimodal LLMs and the TSB-AD-U subset protocol used by LLM-TSAD, showing improvements over strong LLM-based TSAD baselines.

2 Can LLMs Infer Frequency Structure from Time-Domain Inputs?

In time-series anomaly detection (TSAD), anomalous intervals may remain within a plausible value range while changing their period, frequency composition, or oscillatory regularity. Such cases are difficult to identify from pointwise values alone because the anomalous pattern is expressed through changes in temporal structure rather than through large amplitude deviations. This raises a representation-level question: can LLMs reliably recover frequency structure when it is provided only implicitly through time-domain inputs?

To examine this question, we conduct controlled probing experiments on synthetic sinusoidal sequences. These probes are not intended as TSAD benchmarks, but as diagnostic tasks that isolate frequency reasoning from dataset-specific anomaly patterns. Table 1 summarizes three probes: *single sinusoid*, *multi sinusoid*, and *local frequency change*. The first two probes evaluate sequence-level frequency recovery: the model must identify the dominant frequency in a single-component signal or recover the dominant components in a mixture of sinusoids. The third probe evaluates localized frequency-change detection: the model must identify whether a segment contains a change in oscillatory behavior.

As shown in Table 1, global frequency-domain evidence substantially improves sequence-level frequency recovery, achieving the best results on both

the single- and multi-sinusoid probes. This result indicates that full-sequence spectral summaries provide periodic information that the LLM does not reliably recover from indexed values, plots, or de-seasonalized time-domain inputs alone. In particular, global summaries make the dominant periodic regime explicit, which is essential for reasoning about frequency components that persist across the sequence.

However, sequence-level frequency recovery alone is insufficient for anomaly localization. The same global evidence performs poorly on the local frequency-change probe because a full-sequence spectrum aggregates spectral content over time and therefore does not indicate where a frequency change occurs. Local frequency-domain evidence complements this limitation by exposing window-level changes in dominant frequency and spectral entropy, leading to the highest detection accuracy on the localized probe. Together, these results motivate a two-level evidence design: global summaries provide sequence-level periodic context, while local summaries expose time-localized spectral departures.

Figure 1 shows that this diagnostic pattern also appears in TSAD performance. Starting from the indexed de-seasonalized baseline, local frequency-domain evidence improves frequency-relevant anomaly types by exposing window-level spectral changes, while global evidence provides additional sequence-level context. This connection between the probes and TSAD results supports the use of both evidence levels in the final framework.

3 Preliminaries

3.1 LLM-based TSAD

We formulate time-series anomaly detection (TSAD) with a language model as an interval prediction task. Let $\mathbf{x} = (x_1, \ldots, x_T)$ denote a univariate time series of length T, where x_t is the observation at time index t. The goal is to predict a set of anomalous intervals $\hat{\mathcal{A}} = \{(\hat{s}_m, \hat{e}_m)\}_{m=1}^{M}$, where $\hat{s}_m$ and $\hat{e}_m$ are the predicted start and end indices of the m-th anomalous interval, and M is the number of predicted intervals.

Following LLM-TSAD (Park et al., 2025), the input sequence is provided as indexed text, where each observation x_t is paired with its time index t. When visual input is enabled, a time-series plot is additionally provided as an auxiliary modality. The plot is not used as a replacement for indexed text, but as an additional representation in multimodal prompting.

3.2 De-seasonalization-Based Prompting

LLM-TSAD (Park et al., 2025) combines index-aware prompting with de-seasonalization. Given an additive decomposition

$$x_t = s_t + \tau_t + r_t, \tag{1}$$

where s_t denotes the seasonal component, τ_t the trend component, and r_t the residual component at time t, de-seasonalization constructs

$$\tilde{x}_t = x_t - s_t. \tag{2}$$

Here, $\tilde{x}_t$ denotes the de-seasonalized value provided to the LLM. This preprocessing step removes dominant recurring patterns before prompting.

However, de-seasonalization does not explicitly represent frequency changes. It removes a stable seasonal component, but does not describe where local oscillatory behavior changes or how such changes relate to the sequence-level periodic structure. This limitation motivates adding explicit frequency-domain evidence to the LLM input.

3.3 Frequency-Domain Representation

We construct frequency-domain descriptors by applying the Fast Fourier Transform (FFT) to the input time series. Given $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_T)$, we compute the frequency-domain descriptors of its mean-centered version. Let F_k denote the Fourier coefficient at nonzero one-sided frequency bin $k \in \{1, \ldots, K\}$, and define the power spectrum as $P_k = |F_k|^2$. Excluding the zero-frequency component ensures that the summaries describe oscillatory structure rather than the average signal level.

Dominant frequency and period. The dominant frequency is the frequency with the largest spectral power. Let $k^\star = \arg\max_{1 \le k \le K} P_k$. Then $f_{\mathrm{dom}} = f_{k^\star}$, and the corresponding dominant period is $T_{\mathrm{dom}} = 1/f_{\mathrm{dom}}$ when $f_{\mathrm{dom}} > 0$.

Spectral entropy. Spectral entropy measures whether spectral power is concentrated around a few frequencies or diffusely spread across many frequencies:

$$H = -\frac{1}{\log K} \sum_{k=1}^{K} p_k \log p_k, \qquad p_k = \frac{P_k}{\sum_{r=1}^{K} P_r}. \tag{3}$$

Low entropy indicates concentrated periodic structure, while high entropy indicates a more diffuse spectrum.

Spectral peaks and energy ratio. Spectral peaks are the frequency bins with the largest spectral power and summarize the main periodic components beyond a single dominant frequency. We also compare lower- and higher-frequency energy:

$$R_{\mathrm{LH}} = \frac{\sum_{k\in\mathcal{L}} P_k}{\sum_{k\in\mathcal{H}} P_k}, \tag{4}$$

where $\mathcal{L}$ and $\mathcal{H}$ denote the lower and higher halves of the frequency bins excluding the zero-frequency component. This ratio summarizes whether the spectrum is dominated by slower or faster oscillations.

We call individual FFT-derived quantities, such as dominant period and spectral entropy, frequency-domain descriptors. A frequency-domain summary consists of descriptors computed at a specific resolution, either global or local. Once incorporated into the input representation, these summaries serve as frequency-domain evidence rather than anomaly scores or boundary predictions.

4 Proposed Method

4.1 Overview

We propose an LLM-based time-series anomaly detection framework augmented with explicit frequency-domain evidence. The framework is based on the observation that indexed time-domain representations, even after de-seasonalization, do not explicitly expose frequency structure to the LLM. Rather than expecting the LLM to infer periodic structure implicitly from sequence values or plots, we construct explicit frequency-domain evidence from the same de-seasonalized sequence used for interval prediction.

Following LLM-TSAD (Park et al., 2025), the time-domain input is represented as an indexed de-seasonalized sequence, denoted by $\tilde{\mathbf{x}} = (\tilde{x}_1, \ldots, \tilde{x}_T)$. All components in our framework are computed from $\tilde{\mathbf{x}}$. This design preserves the localization benefit of de-seasonalized indexed representations while adding explicit frequency-domain evidence that is not directly available from the de-seasonalized values alone.

The framework constructs two complementary forms of frequency-domain evidence. We first compute global frequency-domain evidence to summarize the sequence-level periodic regime. This global view provides the reference needed to interpret whether a frequency pattern is typical for the full sequence. We then compute local windowed frequency-domain evidence to expose time-localized spectral departures relative to this sequence-level context. Thus, global frequency-domain evidence describes the overall periodic reference, while local frequency-domain evidence identifies candidate regions where frequency behavior changes. Figure 2 illustrates the overall pipeline.

The final interval prediction is performed from an evidence package containing the indexed de-seasonalized sequence, global frequency-domain evidence, and local frequency-domain evidence. When visual input is enabled, a time-series plot is additionally included as an auxiliary modality. These components are not used as standalone anomaly decisions. Instead, they provide structured support for integrating de-seasonalized time-domain observations, visual cues, and frequency-domain evidence during interval prediction.

4.2 Global Frequency-Domain Summary

Global frequency-domain evidence provides sequence-level periodic context. It is computed from the indexed de-seasonalized sequence $\tilde{\mathbf{x}}$. Using the frequency-domain quantities defined in Section 3.3, we compute a global one-sided power spectrum over the full sequence and extract four compact descriptors: the dominant period, the strongest spectral peaks, global spectral entropy, and the low/high-frequency energy ratio.

The dominant period and strongest peaks summarize the main periodic components of the sequence. Global spectral entropy measures whether the sequence-level spectrum is concentrated around a few dominant frequencies or distributed across many frequencies. The low/high-frequency energy ratio summarizes whether the sequence is dominated by slower long-period variation or by stronger high-frequency variation.

The global summary is not associated with temporal location. Instead, it serves as a sequence-level periodic reference. This reference is used to interpret whether local spectral departures are unusual relative to the overall periodic regime.

4.3 Local Windowed Frequency-Domain Summary

Local frequency-domain evidence represents frequency behavior at a time-localized resolution.

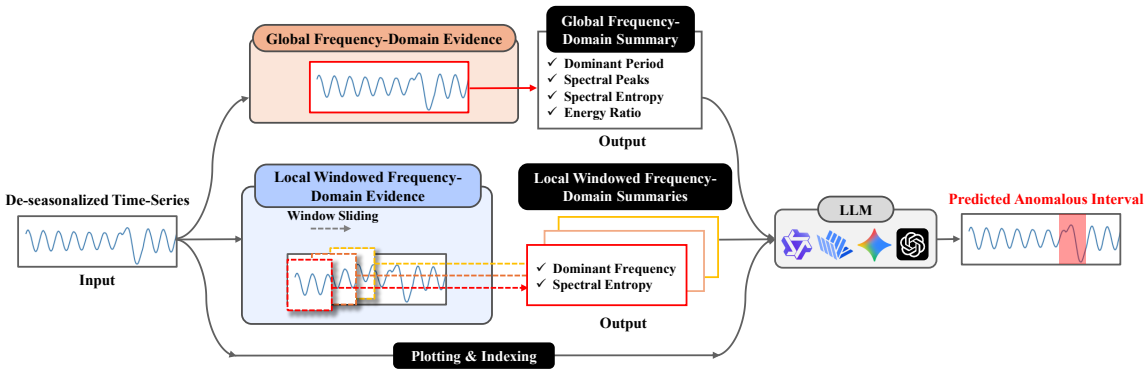

Figure 2: **Overview of the proposed framework.** Indexed de-seasonalized values, global frequency-domain evidence, and local windowed frequency-domain evidence are combined in a single prompt for interval prediction.

A global spectrum describes the overall periodic regime, but it cannot indicate where a local rhythm change occurs. We therefore divide the indexed de-seasonalized sequence $\tilde{\mathbf{x}}$ into overlapping windows and compute a compact subset of the descriptors for each window.

Given window length L and stride S, the j-th window is

$$\tilde{\mathbf{w}}^{(j)} = (\tilde{x}_t)_{t=s_j}^{s_j+L-1}, \quad s_j = jS. \tag{5}$$

for all valid windows satisfying $s_j + L - 1 < T$.

For each window, we compute the dominant local frequency and local spectral entropy. Each local window is represented as

$$\left(s_j,\ s_j + L - 1,\ f_{\text{dom}}^{(j)},\ H^{(j)}\right). \tag{6}$$

The index range provides coarse temporal support, while the spectral descriptors summarize the local rhythm. These tuples are used as auxiliary evidence for candidate spectral departures, not as anomaly boundaries.

4.4 Final Evidence Integration

The final prediction stage integrates the indexed de-seasonalized sequence with the constructed frequency-domain evidence. For each input sequence, we provide the LLM with an evidence package containing the global and local frequency-domain evidence, and indexed de-seasonalized values. When visual input is enabled, the corresponding time-series plot is also included.

The global summary provides sequence-level periodic context, while the local summaries provide time-localized spectral cues. Neither source is treated as a standalone anomaly decision. Instead, they are used as auxiliary evidence that complements the indexed de-seasonalized sequence, which remains the main basis for interval boundary selection.

The final stage formulates TSAD as interval prediction over the indexed sequence. The LLM integrates the indexed de-seasonalized values with global and local frequency evidence to produce the final anomalous interval set. The complete evidence serialization format and output template are provided in Appendix C.

5 Experiment

5.1 Experiment Setup

Datasets. We evaluate our method on two benchmarks. The main evaluation uses AnomLLM (Zhou and Yu, 2025), following the dataset configuration and preprocessing protocol of LLM-TSAD (Park et al., 2025). AnomLLM contains controlled time-series instances with annotated anomaly intervals across representative anomaly types, including point, range, trend, and frequency anomalies. To further examine whether the proposed evidence representation remains useful beyond this controlled anomaly-type setting, we also evaluate on the same eight-category TSB-AD-U (Liu and Paparrizos, 2024) evaluation subset used by LLM-TSAD. This subset-level evaluation is used as an additional practical comparison setting, rather than as exhaustive coverage of the full TSB-AD-U benchmark.

Evaluation models. We evaluate the proposed method with four vision-language models: InternVL2-LLaMA3-76B (Chen et al., 2024), Qwen2.5-VL-72B-Instruct (Yang et al., 2024), Gemini-2.5-Flash (Comanici et al., 2025), and GPT-4o (Hurst et al., 2024). These models include both open-source and proprietary multimodal LLMs, allowing us to test whether the proposed evidence representation is useful across different model families. For the TSB-AD-U evaluation subset, we

LLMs	Methods	Standard Metrics			Affiliation Metrics		
		Prec.	Recall	F1	Prec.	Recall	F1
Naïve baseline (Always Normal Predictor)		31.75	31.75	31.75	31.75	31.75	31.75
InternVL2 (LLaMA3-76B)	*Existing Prompts (AnomLLM)*						
	0shot-Text	12.94	21.15	13.38	22.12	28.38	24.10
	0shot-Vision	22.33	46.19	23.92	50.93	60.94	55.53
	1shot-Vision-CoT	33.72	36.28	33.67	53.62	55.55	53.93
	LLM-TSAD (0shot-Text)	27.67	52.29	29.21	54.31	60.71	55.04
	LLM-TSAD (0shot-Text+Vision)	30.00	59.35	34.66	58.14	71.79	62.90
	Ours (0shot-Text)	**44.66**	**66.83**	**46.64**	**73.14**	**74.01**	**71.93**
	Ours (0shot-Text+Vision)	**36.99**	**72.28**	**42.15**	**69.54**	**78.45**	**72.41**
Qwen-2.5 (VL-72B-Instruct)	*Existing Prompts (AnomLLM)*						
	0shot-Text	25.99	25.66	25.49	45.69	42.44	43.11
	0shot-Vision	45.78	51.00	46.65	69.50	68.99	68.78
	1shot-Vision-CoT	27.09	29.08	27.23	39.52	40.07	39.38
	LLM-TSAD (0shot-Text)	**72.47**	54.09	56.13	82.91	78.40	78.96
	LLM-TSAD (0shot-Text+Vision)	72.95	69.78	67.56	84.51	81.52	81.87
	Ours (0shot-Text)	68.80	**68.54**	**64.38**	**87.22**	**85.44**	**85.39**
	Ours (0shot-Text+Vision)	**77.93**	**77.20**	**73.70**	**93.88**	**90.64**	**91.39**
Gemini-2.5 (Flash)	*Existing Prompts (AnomLLM)*						
	0shot-Text	15.69	12.83	13.37	42.64	38.31	38.85
	0shot-Vision	52.07	70.67	57.28	81.39	79.30	79.55
	1shot-Vision-CoT	46.64	61.54	50.96	75.51	74.80	74.55
	LLM-TSAD (0shot-Text)	**69.54**	61.95	62.87	77.21	75.43	75.58
	LLM-TSAD (0shot-Text+Vision)	**80.88**	76.44	76.62	88.55	86.90	87.12
	Ours (0shot-Text)	65.91	**72.22**	**64.38**	**82.20**	**82.58**	**82.02**
	Ours (0shot-Text+Vision)	78.48	**84.72**	**77.97**	**92.66**	**92.80**	**92.48**
GPT-4o	*Existing Prompts (AnomLLM)*						
	0shot-Text	19.69	17.62	17.76	46.35	45.51	44.73
	0shot-Vision	39.54	50.60	42.03	62.19	61.99	61.68
	1shot-Vision-CoT	31.48	38.50	32.93	56.43	55.84	55.39
	LLM-TSAD (0shot-Text)	**65.52**	53.45	54.40	75.65	73.88	73.47
	LLM-TSAD (0shot-Text+Vision)	**76.94**	68.42	70.04	83.10	78.85	80.11
	Ours (0shot-Text)	58.65	**63.79**	**55.51**	**79.76**	**78.73**	**78.56**
	Ours (0shot-Text+Vision)	72.30	**75.75**	**70.20**	**89.51**	**86.71**	**87.35**

Table 2: **Main anomaly detection performance on AnomLLM across four multimodal LLMs.** We report standard and affiliation-based precision, recall, and F1. For each model block and input modality group, the best result is shown in **bold**, and the second-best result is underlined.

use Gemini-2.5-Flash to compare our method with Gemini-based LLM baselines under the same backbone.

Baselines. On AnomLLM, we compare with representative AnomLLM prompting variants (Zhou and Yu, 2025) and LLM-TSAD (Park et al., 2025), our main baseline. LLM-TSAD is most relevant because our method preserves its indexed deseasonalized formulation while adding FFT-based global and local frequency-domain evidence. We also report the naïve always-normal predictor as a dataset-level prior.

For the TSB-AD-U subset, we follow the compact LLM-TSAD protocol and report the strongest representative non-LLM baseline from each family, together with Gemini-2.5-Flash-based LLM baselines.

Metrics. Following LLM-TSAD (Park et al., 2025), we report both standard metrics and affiliation metrics. For each metric family, we report precision, recall, and F1. Standard metrics evaluate label-level overlap between predicted and ground-truth anomaly intervals, providing a strict measure of interval prediction accuracy. Affiliation metrics provide a complementary distance-aware assessment of temporal agreement between predicted and ground-truth intervals. All scores are aggregated over evaluation instances following the same averaging protocol as LLM-TSAD (Park et al., 2025).

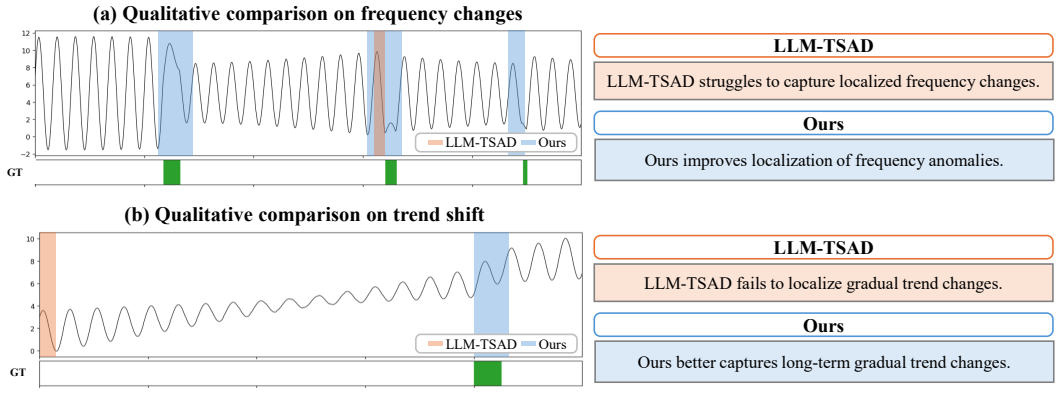

Figure 3: **Qualitative comparison between LLM-TSAD and our method using Qwen2.5-VL-72B-Instruct.** The examples show frequency-change and trend-shift anomalies from AnomLLM.

5.2 Main Results

Quantitative results. Table 2 reports the main results on AnomLLM across four multimodal LLMs. Compared with LLM-TSAD, our method consistently improves both standard F1 and affiliation F1 across all four evaluated models and both input settings. These results show that the proposed frequency-domain evidence improves both strict interval overlap and distance-aware temporal agreement.

The magnitude of improvement varies across backbone and modality configurations, reflecting differences in how multimodal LLMs integrate indexed numerical values, visual plots, and structured frequency-domain evidence. For InternVL2, the text-only setting achieves the strongest standard F1, while the text+vision setting yields the highest affiliation F1 within the same backbone. This pattern suggests that visual context does not uniformly affect strict boundary overlap, but can still support distance-aware temporal agreement under affiliation-based evaluation. Overall, frequency-domain summaries act as complementary evidence for LLM-based TSAD, with backbone-specific modality effects.

Qualitative results. Figure 3 shows representative cases where anomalies involve localized rhythm changes or gradual temporal structure shifts rather than isolated pointwise deviations. Compared with LLM-TSAD, our method produces intervals that better align with the ground-truth regions, illustrating how explicit frequency-domain evidence can complement indexed and de-seasonalized time-domain inputs. These examples support the interpretation of the quantitative results, but are used only as qualitative evidence rather than standalone proof of overall superiority.

Model	Setting	F1	Affi. F1
InternVL2	LLM-TSAD	34.66	62.90
	+ LF	<u>40.31</u>	<u>72.27</u>
	+ LF + GF	**42.15**	**72.41**
Qwen-2.5	LLM-TSAD	67.56	81.87
	+ LF	<u>71.62</u>	**92.57**
	+ LF + GF	**73.70**	<u>91.39</u>
Gemini-2.5	LLM-TSAD	<u>76.62</u>	<u>87.12</u>
	+ LF	62.86	80.44
	+ LF + GF	**77.97**	**92.48**
GPT-4o	LLM-TSAD	<u>70.04</u>	80.11
	+ LF	67.40	**89.02**
	+ LF + GF	**70.20**	<u>87.35</u>

Table 3: **Component-to-final ablation.** LF and GF denote local and global frequency-domain evidence, respectively, with LF+GF representing the final frequency-augmented configuration. The best and second-best results within each model are shown in **bold** and <u>underlined</u>.

5.3 Ablation Studies

Component-to-final ablation. Table 3 evaluates how local frequency-domain evidence (LF) and global frequency-domain evidence (GF) contribute to the final prediction. LF-only is an intermediate configuration that exposes window-level spectral departures without sequence-level periodic context. While LF improves affiliation F1 for most models, the Gemini-2.5 result indicates that local evidence alone can be insufficient or less reliable for stable interval localization. Adding GF on top of LF improves standard F1 for all models and restores or further improves affiliation F1 in several cases, suggesting that global periodic context helps calibrate local spectral departures. Overall, the ablation supports the use of LF and GF as complementary evidence in the final configuration.

Group	Method	Standard Metrics			Affiliation Metrics		
		Prec.	Recall	F1	Prec.	Recall	F1
Non-LLM baseline	ML: SR (Ren et al., 2019)	32.61	40.77	30.57	66.80	**95.11**	75.94
	DL: USAD (Audibert et al., 2020)	23.92	34.54	24.46	59.84	61.60	55.24
	FM: Chronos (Ansari et al., 2024)	23.58	**55.65**	25.08	64.87	96.33	75.90
LLM baseline	AnomLLM (Gemini-2.5-Flash)	12.28	16.93	6.82	40.37	33.75	30.97
	LLM-TSAD (Gemini-2.5-Flash)	43.05	43.49	36.85	73.08	93.88	80.15
Ours	**Ours (Gemini-2.5-Flash)**	**46.51**	46.19	**40.26**	**74.62**	92.71	**80.35**

Table 4: **Overall performance on the eight-category TSB-AD-U subset used by LLM-TSAD.** We compare our method with representative non-LLM and Gemini-based LLM baselines.

Config.	Freq.	Point	Range	Trend	Overall
Indexed	23.99	**88.13**	**86.39**	60.28	64.70
+D.S.†	33.98	77.51	85.58	71.51	67.15
+D.S.+LF	39.24	81.55	82.85	82.83	71.62
+D.S.+LF+GF	**40.89**	83.30	83.94	**86.67**	**73.70**

Table 5: **Type-wise standard F1 across evidence configurations.** We report standard F1 by anomaly type using Qwen2.5-VL-72B-Instruct. D.S., LF, and GF denote de-seasonalization, local frequency-domain evidence, and global frequency-domain evidence, respectively. The best and second-best results for each anomaly type are shown in **bold** and underlined, respectively. † denotes our reproduced result.

Type-wise analysis. Table 5 breaks down standard F1 by anomaly type. The largest gains appear for frequency changes and trend shifts, where anomalies reflect changes in rhythm, periodicity, or long-term temporal structure rather than isolated value deviations. Although trend shifts are not frequency anomalies in the strict sense, they can induce low-frequency energy changes or alter local spectral distributions, making them partially observable through FFT-based local and global summaries. This aligns with the roles of LF and GF: local evidence captures window-level spectral changes, while global evidence provides sequence-level periodic context.

For point and range anomalies, the indexed de-seasonalized sequence already exposes strong amplitude- and boundary-level cues, so frequency-domain summaries provide less additional information. Overall, the type-wise results support a targeted interpretation of our method: frequency-domain evidence is most useful when anomaly localization requires spectral or long-range temporal context beyond local time-domain values.

Evaluation on TSB-AD-U. Table 4 reports results on the same eight-category TSB-AD-U subset used by LLM-TSAD: NEK, TAO, MSL, Power, Daphnet, YAHOO, SED, and TODS. This setting follows the prior subset-level protocol to provide a cross-benchmark robustness check against the strongest available LLM-based TSAD baseline, rather than an exhaustive evaluation of the full benchmark.

With the same Gemini-2.5-Flash backbone, our method improves standard F1 over Gemini-based LLM-TSAD while maintaining comparable affiliation-based performance. This supports the usefulness of the proposed frequency-domain summaries beyond AnomLLM, within the scope of the subset-level evaluation.

6 Conclusion

We proposed an evidence-augmented framework for zero-shot LLM-based time-series anomaly detection that complements indexed and de-seasonalized time-domain inputs with compact global and local frequency-domain evidence. Global evidence provides sequence-level periodic context, while local evidence exposes time-localized spectral departures; both are used as auxiliary evidence rather than standalone anomaly decisions. Our probing experiments suggest that such frequency-relevant structure is not always reliably recovered from indexed values or plots alone. Experiments on AnomLLM with multiple multimodal LLMs and evaluation on the TSB-AD-U evaluation subset show that explicit frequency-domain evidence can improve strong LLM-based baselines, especially for anomalies involving rhythm, periodicity, or gradual temporal changes. These findings motivate adaptive frequency-evidence construction for longer or multivariate time series.

7 Limitations

While frequency-domain evidence improves LLM-based TSAD in our experiments, zero-shot anomaly localization remains challenging. We do not directly analyze which parts of the provided evidence the LLM relies on when producing its interval predictions. Therefore, our results support the usefulness of structured frequency-domain evidence at the input level, but do not fully explain the internal reasoning process of the model. Future work may investigate attribution or intervention-based analyses to better understand how LLMs use time-domain, visual, and frequency-domain evidence during anomaly localization.

References

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. 2024a. Can large language models be anomaly detectors for time series? In *IEEE*.

Sarah Alnegheimish, Linh Nguyen, Laure Berti-Equille, and Kalyan Veeramachaneni. 2024b. Large language models can be zero-shot anomaly detectors for time series? In *DSAA*.

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, Michael W. Mahoney, Kari Torkkola, Andrew Gordon Wilson, Michael Bohlke-Schneider, and Yuyang Wang. 2024. Chronos: Learning the language of time series. *Transactions on Machine Learning Research*.

Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. 2020. Usad: Unsupervised anomaly detection on multivariate time series. In *KDD*.

Ane Blázquez-García, Angel Conde, Usue Mori, and Jose A. Lozano. 2020. A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, and others Lu, Lewei. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *CVPR*.

Robert B. Cleveland, William S. Cleveland, Jean E. McRae, and Irma Terpenning. 1990. Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, and others Rosen, Evan. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Abhimanyu Das, Weihao Kong, Rajat Sen, and Yichen Zhou. 2024. A decoder-only foundation model for time-series forecasting. In *ICML*.

Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. 2022. Adbench: Anomaly detection benchmark. In *NeurIPS*.

Jeff Hawkins, Subutai Ahmad, and Donna Dubinsky. 2010. Hierarchical temporal memory including htm cortical learning algorithms. White paper, Numenta, Inc.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, and others Radford, Alec. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Daehyun Kim, Sungyong Baik, and Tae Hyun Kim. 2023. Sanflow: Semantic-aware normalizing flow for anomaly detection. In *NeurIPS*.

Yi Kun, Zhang Qi, Fan Wei, Wang Shoujin, Wang Pengyang, He Hui, An Ning, Lian Defu, Cao Longbing, and Niu Zhedong. 2023. Frequency-domain mlps are more effective learners in time series forecasting. In *NeurIPS*.

Jun Liu, Chaoyun Zhang, Jiaxu Qian, Minghua Ma, Si Qin, Chetan Bansal, Qingwei Lin, Saravan Rajmohan, and Dongmei Zhang. 2025. Large language models can deliver accurate and interpretable time series anomaly detection. In *KDD*.

Qinghua Liu and John Paparrizos. 2024. The elephant in the room: Towards a reliable time-series anomaly detection benchmark. In *NeurIPS*.

Xu Liu, Junfeng Hu, Yuan Li, Shizhe Diao, Yuxuan Liang, Bryan Hooi, and Roger Zimmerman. 2023. Unitime: A language-empowered unified model for cross-domain time series forecasting. In *WWW*.

Yi-Xiang Lu, Xiao-Bo Jin, Jian Chen, Dong-Jie Liu, and Guang-Gang Geng. 2024. F-se-lstm: A time series anomaly detection method with frequency domain information. *arXiv preprint arXiv:2412.02474*.

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

Youngeun Nam, Susik Yoon, Yooju Shin, Minyoung Bae, Hwanjun Song, Jae-Gil Lee, and Byung Suk Lee. 2024. Breaking the time-frequency granularity discrepancy in time-series anomaly detection. In *WWW*.

Guansong Pang, Chunhua Shen, Longbing Cao, and Anton van den Hengel. 2020. Deep learning for anomaly detection: A review. *ACM Computing Surveys*.

Junwoo Park, Kyudan Jung, Dohyun Lee, Hyuck Lee, Daehoon Gwak, ChaeHun Park, Jaegul Choo, and Jaewoong Cho. 2025. Delving into large language models for effective time-series anomaly detection. In *NeurIPS*.

Hansheng Ren, Bixiong Xu, Yujing Wang, Chao Yi, Congrui Huang, Xiaoyu Kou, Tony Xing, Mao Yang, Jie Tong, and Qi Zhang. 2019. Time-series anomaly detection service at microsoft. In *KDD*.

Jungwook Seo, Minjeong Kim, Younkwan Lee, Seungho Shin, and Sungyong Baik. 2026. Anomaly as non-conformity via training-free graph laplacian energy minimization. In *CVPR*.

Ya Su, Youjian Zhao, Chenhao Niu, Rong Liu, Wei Sun, and Dan Pei. 2019. Robust anomaly detection for multivariate time series through stochastic recurrent neural network. In *KDD*.

Swee Chuan Tan, Kai Ming Ting, and Tony Fei Liu. 2011. Fast anomaly detection for streaming data. In *IJCAI*.

Sean J. Taylor and Benjamin Letham. 2018. Forecasting at scale. *The American Statistician*.

Zexin Wang, Changhua Pei, Minghua Ma, Xin Wang, Zhihan Li, Dan Pei, Saravan Rajmohan, Dongmei Zhang, Qingwei Lin, and others Zhang, Haiming. 2024. Revisiting vae for unsupervised time series anomaly detection: A frequency perspective. In *WWW*.

Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Juanmin Wang, and Mingsheng Long. 2023. Temporal 2d-variation modeling for general time series analysis. In *ICLR*.

Jiehui Xu, Haixu Wu, Jianmin Wang, and Mingsheng Long. 2022. Anomaly transformer: Time series anomaly detection with association discrepancy. In *ICLR*.

Hao Xue and Flora D. Salim. 2022. Promptcast: A new prompt-based learning paradigm for time series forecasting. *TKDE*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Chin-Chia Michael Yeh, Yan Zhu, Liudmila Ulanova, Nurjahan Begum, Yifei Ding, Hoang Anh Dau, Diego Furtado Silva, Abdullah Mueen, and Eamonn Keogh. 2018. Time series joins, motifs, discords and shapelets: A unifying view that exploits the matrix profile. *Data Mining and Knowledge Discovery*.

Chaoli Zhang, Tian Zhou, Qingsong Wen, and Liang Sun. 2022. Tfad: A decomposition time series anomaly detection architecture with time-frequency analysis. In *CIKM*.

Liangwei Nathan Zheng, Chang Dong, Wei Emma Zhang, Lin Yue, Miao Xu, Olaf Maennel, and Weitong Chen. 2025. Understanding why large language models can be ineffective in time series analysis: The impact of modality alignment. In *KDD*.

Zihao Zhou and Rose Yu. 2025. Can llms understand time series anomalies? In *ICLR*.

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

Table of Contents

A Related Work

Time-series anomaly detection (TSAD) commonly uses reconstruction, forecasting, or representation-learning objectives. Classic paradigms include statistical decomposition like STL (Cleveland et al., 1990), decomposable forecasting models such as Prophet (Taylor and Letham, 2018), and bio-inspired online learning like HTM (Hawkins et al., 2010). Additionally, algorithmic approaches such as Matrix Profiles (Yeh et al., 2018) and ensemble-based HS-Trees (Tan et al., 2011) capture complex sequence representations. Deep learning models further extend this paradigm by modeling probabilistic latent structures, adversarial reconstruction errors, temporal dependencies, and frequency-aware representations (Su et al., 2019; Audibert et al., 2020; Xu et al., 2022; Zhang et al., 2022; Wu et al., 2023; Kun et al., 2023). Related efforts to characterize normality also appear in visual anomaly detection, including density-based and non-conformity-based formulations (Seo et al., 2026; Kim et al., 2023). Within TSAD, however, most conventional approaches remain dataset-specific and typically require sufficient anomaly-free training data, which limits their applicability in dynamic or previously unseen environments.

Recent progress in foundation models motivates a shift from task-specific TSAD models toward more generalizable approaches. Large-scale pre-trained time-series models show that generic sequence modeling can support numerical forecasting across diverse domains (Ansari et al., 2024; Das et al., 2024). In parallel, LLM-based approaches examine whether time-series signals can be transformed into textual representations and analyzed through in-context reasoning. By using textualization, prompting, and retrieval-based demonstrations, these methods aim to support zero-shot anomaly detection while also producing interpretable explanations (Alnegheimish et al., 2024a; Liu et al., 2025).

Although LLM-based methods improve flexibility and interpretability, applying language models directly to continuous numerical signals remains challenging (Ming et al., 2024). Since LLMs are primarily trained on discrete semantic tokens, naively serialized time-series inputs may not preserve fine-grained local variations, periodic structures, or subtle frequency changes in a form that the model can reliably interpret (Zhou and Yu, 2025; Zheng et al., 2025). As a result, existing prompting-based detectors may become sensitive to visually salient amplitude changes while being less reliable for contextual anomalies, phase shifts, or frequency-domain irregularities (Nam et al., 2024). Moreover, when anomaly localization relies only on textual indices, the model must implicitly track positions and compare numerical values, which can impose an additional reasoning burden on the LLM.

Recent work mitigates these issues by combining statistical preprocessing with structured prompting. For example, de-seasonalization can simplify the input signal and expose residual deviations(Xue and Salim, 2022; Liu et al., 2023), while index-aware prompting provides explicit positional information (Park et al., 2025). However, suppressing seasonal components is not the same as explicitly representing frequency structure. Although de-seasonalization can help localization, it does not directly describe how local oscillatory behavior changes or how such changes relate to the sequence-level periodic regime. This makes preprocessing-based simplification incomplete for anomalies expressed through rhythm, periodicity, or transient frequency changes.

Our evidence-augmented framework addresses this limitation by preserving the indexed time-domain signal while adding compact frequency-domain evidence. Local frequency-domain summaries expose time-localized spectral departures, while global frequency-domain summaries provide sequence-level periodic context. Together, these components form a structured interface that allows zero-shot LLM-based TSAD to use both time-domain and frequency-domain evidence.

Provider	Model	Endpoint identifier
OpenAI	GPT-4o	`openai/gpt-4o`
Google	Gemini 2.5 Flash	`google/gemini-2.5-flash`
Qwen	Qwen 2.5 VL	`qwen/qwen2.5-vl-72b-instruct`

Table 6: API-based model endpoints for our experiments.

B Computational Resources and Experimental Setup

In this section, we provide details regarding the hardware and software environments for preprocessing, model inference, and server execution.

B.1 Computing Infrastructure

Most preprocessing steps, including data cleaning, de-seasonalization, and frequency-domain evidence construction, run on a local workstation. The hardware specifications are as follows:

- **CPU:** Intel(R) Core(TM) i9-10980XE CPU @ 3.00GHz (18 cores, 36 threads).
- **GPU:** NVIDIA GeForce RTX 4090 (24GB VRAM).
- **Usage:** This infrastructure supports dataset preprocessing, FFT-based evidence construction, and the multimodal inference pipeline.

For the InternVL2 (Chen et al., 2024) evaluation, we use one NVIDIA B200 GPU through the Runpod cloud service. In this setting, the corresponding FFT-based evidence construction and model inference run on the same cloud server.

B.2 API-based Model Specifications

For API-based LLM inference, we use the model endpoints listed in Table 6. The inference pipeline accesses these endpoints through OpenRouter.

B.3 Artifact Use

We use AnomLLM (Zhou and Yu, 2025) and the TSB-AD-U (Liu and Paparrizos, 2024) subset only for research evaluation, consistent with their intended use as anomaly-detection benchmarks. We access API-based and open-source LLMs only through their documented inference interfaces. Derived preprocessing outputs, including de-seasonalized sequences and FFT-based evidence summaries, support only anomaly-detection evaluation and are not redistributed.

B.4 Data Privacy and Content

We use publicly available time-series anomaly detection benchmarks, AnomLLM and the TSB-AD-U subset, only for research evaluation. The data used in our experiments consist of numerical time-series values and anomaly interval annotations rather than natural-language user content. We do not identify personally identifying information or offensive textual content in the benchmark inputs for evaluation. We collect no additional personal data and make no attempt to re-identify individuals or link the data to external sources.

B.5 Data Statistics

We use the AnomLLM benchmark following the dataset configuration and preprocessing protocol of LLM-TSAD. We also evaluate on the eight-category TSB-AD-U subset used by LLM-TSAD, consisting of NEK, TAO, MSL, Power, Daphnet, YAHOO, SED, and TODS. Since our method is evaluated in a zero-shot setting, no train/dev split is used for model training or prompt tuning.

C Prompt Templates

We report the prompt template for our frequency-augmented LLM-based TSAD framework. The full template is shown in Figure 4. Ellipses (. . .) indicate omitted consecutive numeric entries for readability; in the actual prompt, these entries are provided explicitly. For prompt serialization, we use zero-based indices to match the benchmark input format, although the mathematical notation in the main text follows the conventional one-based form. Following the LLM-TSAD (Park et al., 2025) evaluation protocol, the prompt specifies that each sequence contains up to five anomalous intervals. This constraint is used only to match the prior benchmark protocol and is applied consistently across the LLM-based methods in the comparison; no anomaly-type labels or instance-specific demonstrations are provided during inference. For text-only settings, the image-related instruction is omitted from the prompt, while the remaining evidence serialization and output format are kept unchanged.

D Frequency-Reasoning Probe Setup

We use three synthetic probing experiments to examine whether different input representations expose frequency-relevant information to an LLM. These probes serve as diagnostic representation

```
I will provide you with time-series value data recorded at hourly intervals, along with a plotted time-series image.

<global_frequency_analysis>
The following global frequency-domain features summarize the overall frequency structure of the entire time series.

Use global frequency-domain evidence to understand the baseline periodic structure of the sequence.
Do not use global frequency-domain evidence alone to determine the anomaly interval.

Dominant frequency: 0.0310 (period ~= 32.3 steps)
Top spectral peaks: 0.0310, 0.0300, 0.0290
Spectral entropy (normalized): 0.4374
Low/High frequency energy ratio: 1149.75
</global_frequency_analysis>

<local_frequency_analysis>
The following local/windowed frequency-domain features summarize frequency behavior across time windows.

Use local frequency-domain evidence to identify where frequency behavior changes within the sequence.
Localized changes in dominant frequency or spectral entropy may support point, burst, oscillatory, or frequency anomalies.

Window [0-63]: dominant_freq=0.0312, spectral_entropy=0.1295
Window [32-95]: dominant_freq=0.0312, spectral_entropy=0.1400
...
Window [864-927]: dominant_freq=0.0312, spectral_entropy=0.0958
Window [896-959]: dominant_freq=0.0312, spectral_entropy=0.0913
</local_frequency_analysis>

Here is time-series data in (index, value) format:
<history>
(0, 7.13)
(1, 7.52)
...
(998, -0.77)
(999, -0.14)
</history>

Assume there are up to 5 anomalies.

Detect ranges of anomalies in this time series, considering the plotted image if it is useful.
The index of the time series starts from 0 to 999.
List one by one, in JSON format.
If there are no anomalies, answer with an empty list [].
Do not say anything other than the answer.

Output template:
[{"start": ..., "end": ...}, {"start": ..., "end": ...}, ...]
```

Figure 4: Prompt template used in the proposed frequency-augmented LLM-based TSAD framework.

tests that motivate explicit frequency-domain evidence rather than full TSAD benchmarks.

D.1 Common Setup

Each probe uses 200 synthetic samples. All evaluations use Qwen2.5-VL-72B-Instruct (Yang et al., 2024). For each sample, we compare five input representations: indexed text, indexed text with a plot, de-seasonalized text, indexed text augmented with global frequency-domain evidence, and indexed text augmented with local frequency-domain evidence. Global frequency-domain evidence comes from a frequency-domain summary over the full sequence. Local frequency-domain evidence comes from window-level frequency-domain summaries with sliding windows of length 64 and stride 32. When constructing frequency-domain summaries, we exclude periods shorter than 4 samples or longer than 0.8 times the analyzed sequence or window length.

D.2 Single-Sinusoid Probe

Each sequence follows

$$x_t = A\sin(2\pi t/P + \phi) + \epsilon_t,$$

where we sample the sequence length from $\{128, 256, 512\}$, the period P from $\{8, 12, 16, 24, 32, 48, 64\}$, the amplitude A from $[0.5, 2.0]$, and the Gaussian noise standard deviation from $[0.02, 0.15]$. The model outputs the dominant period. We report Top-1 Recovery with a 10% relative-error tolerance, where a prediction is counted as correct if

$$|\hat{P} - P|/P \leq 0.10.$$

D.3 Multi-Sinusoid Probe

Each sequence follows a sum of 2 to 4 sinusoidal components:

$$x_t = \sum_i A_i \sin(2\pi t/P_i + \phi_i) + \epsilon_t.$$

We sample the sequence length from $\{256, 512\}$ and component periods without replacement from $\{8, 12, 16, 24, 32, 48, 64\}$. Amplitude and noise ranges follow the single-sinusoid probe. The model outputs the dominant period and the top-k periods. We report Top-k Recovery with a 10% relative-error tolerance, computed as the fraction of true periods matched by any predicted period and then macro-averaged over samples.

D.4 Local Frequency-Change Probe

Each sequence contains one contiguous segment whose frequency differs from the background while the amplitude range remains comparable:

$$x_t = A \sin(2\pi t / P_s + \phi_s) + \epsilon_t,$$

where $(P_s, \phi_s) = (P_{\text{bg}}, \phi_{\text{bg}})$ outside the changed-frequency segment and $(P_s, \phi_s) = (P_{\text{anom}}, \phi_{\text{anom}})$ inside the segment. We sample the sequence length from $\{256, 512\}$ and the background period from $\{16, 24, 32\}$. We set the anomalous period to $P_{\text{bg}} r$, where $r \in \{0.4, 0.5, 2.0, 3.0\}$, with clipping to a valid range. We sample the changed-frequency segment length from 20 to 80 time steps and place it away from the first and last 10% of the sequence. We sample amplitude from $[0.8, 1.2]$ and the Gaussian noise standard deviation from $[0.02, 0.10]$. The model determines whether the sequence contains a local frequency-change segment. The reported metric is binary detection accuracy, computed from the predicted anomaly-presence flag.

D.5 De-seasonalization

For the de-seasonalized representation, we construct a seasonal template from an oracle period and subtract the tiled template from the original sequence. The oracle period serves only to define a strong diagnostic de-seasonalized representation. For the local frequency-change probe, this period is the background period. For the single- and multi-sinusoid probes, it is the ground-truth dominant period. This oracle-period construction is used only in the synthetic probing experiments and is not used in the AnomLLM (Zhou and Yu, 2025)or TSB-AD-U (Liu and Paparrizos, 2024) benchmark evaluations.

E Implementation Details

For local windowed frequency-domain evidence, we use a window length of 64 for the AnomLLM benchmark (Zhou and Yu, 2025) and 128 for the TSB-AD-U subset (Liu and Paparrizos, 2024). Since the input sequences in the TSB-AD-U subset are approximately twice as long as those in AnomLLM under our evaluation setting, we double the window length to keep the relative temporal scale of local windowed frequency-domain summaries comparable across benchmarks. In both benchmarks, the window stride is set to half of the corresponding window length. For the global frequency-domain summary, we retain the top three nonzero-frequency spectral peaks with the largest spectral power across all experiments.

Model	Configuration	F1	Affi. F1
Qwen2.5-VL-72B	Baseline, no frequency descriptors	68.33	82.78
Qwen2.5-VL-72B	Naive descriptor serialization	67.24	85.62
Qwen2.5-VL-72B	Structured GF/LF evidence	**73.97**	**92.53**
Gemini-2.5-Flash	Baseline, no frequency descriptors	75.84	86.76
Gemini-2.5-Flash	Naive descriptor serialization	66.24	83.99
Gemini-2.5-Flash	Structured GF/LF evidence	**77.27**	**92.09**

Table 7: Same-LLM prompt-schema ablation on the balanced 1/4-scale diagnostic subset. All conditions keep the LLM, instances, image, indexed history, output format, descriptor values and order, and block position fixed; only the representation schema changes.

F Additional Experimental Analyses

This section consolidates the controlled analyses that complement the main evaluation. The analyses target three questions that follow directly from the motivation of the paper: whether raw FFT descriptors alone explain the gains, whether global and local descriptors play distinct roles, and whether the observed behavior remains stable under evidence perturbations, temporal-resolution changes, and backbone changes.

F.1 Controlled Diagnostic Subset

Several analyses below use a balanced 1/4-scale subset of AnomLLM. This subset serves as a controlled diagnostic setting for comparisons that require many matched LLM calls. Every condition uses the same instances and preserves anomaly-type balance, which isolates the changed evidence representation while keeping repeated diagnostic inference tractable. The subset-scale analyses do not replace the full-scale results in the main paper; they test narrower properties of the representation under matched conditions.

Method	LLM	Same desc.	F1	Affi. F1
FFT-ZScore	No	Yes	34.63	75.19
FFT-ZScore top 5%	No	Yes	21.28	52.61
FFT-IsolationForest	No	Yes	23.09	55.86
FFT-CUSUM	No	Yes	41.04	78.64
LLM-TSAD, paired run	Yes	No	67.14	82.57
Ours LF+GF	Yes	Yes	**73.70**	**91.39**

Table 8: Label-free non-LLM controls that use the same FFT-derived descriptors, preprocessing, and evaluation protocol. Thresholds are fixed a priori. The LLM-TSAD row is the reproduced paired diagnostic run used for these controls.

F.2 Structured Evidence versus Naive Descriptor Serialization

Table 7 tests whether the descriptor values themselves are sufficient. The naive condition keeps the same LLM, instances, image, indexed history, output format, descriptor values and order, and block position. It only replaces the GF/LF evidence schema with a neutral descriptor serialization. Structured GF/LF evidence yields higher standard and affiliation F1 than naive serialization for both backbones. This controlled gap indicates that raw descriptor values alone do not account for the result in this setting. The finding supports the representation-level motivation: frequency information becomes more useful when the prompt distinguishes sequence-level periodic context from time-localized spectral evidence.

F.3 Same-Feature Non-LLM Controls

Table 8 evaluates label-free classical controls that use the same FFT-derived descriptors and preprocessing without an LLM. Among these controls, FFT-CUSUM gives the highest standard F1 at 41.04 and affiliation F1 at 78.64, while LF+GF reaches 73.70 and 91.39 under the same evaluation protocol. These controls do not establish that an LLM is necessary for every frequency anomaly. They show more narrowly that the tested direct descriptor detectors do not reproduce the LF+GF result. This distinction is consistent with the method design, where FFT summaries provide auxiliary evidence and the LLM still performs interval prediction from the indexed sequence.

F.4 Component-Level Descriptor Ablation

Table 9 isolates individual global and local descriptors on the balanced diagnostic subset. The descriptors show different metric profiles. Global dominant frequency gives the highest standard F1 among single descriptors, while local spectral entropy gives the highest affiliation F1 among single descriptors. No individual descriptor gives the strongest value on both metrics, and the full LF+GF configuration gives a balanced overall profile. This pattern matches the intended division of evidence: global descriptors summarize sequence-level periodic structure, whereas local descriptors expose interval-level variation. The result therefore supports complementary roles rather than a claim that every descriptor contributes uniformly.

Added evidence	F1	Affi. F1
LLM-TSAD baseline	68.33	82.78
Global dominant frequency	**74.01**	89.45
Global top-k spectral peaks	72.52	87.50
Global spectral entropy	71.91	86.61
Global low/high energy ratio	72.54	87.03
Local dominant frequency	72.70	87.91
Local spectral entropy	71.58	92.25
Full LF+GF	73.97	**92.53**

Table 9: Component-level descriptor ablation on the balanced 1/4-scale diagnostic subset with Qwen2.5-VL-72B-Instruct. Each condition adds one descriptor type to the same baseline prompt; the final row uses the full structured LF+GF evidence.

Condition	F1	Δ F1	Affi. F1	Δ Affi.
Correct GF/LF evidence	73.70	–	91.39	–
Local evidence shuffled	71.50	-2.20	87.84	-3.55
Global+local mismatched donor	70.57	-3.13	86.87	-4.52
Remove frequency evidence	67.14	-6.56	82.57	-8.82

Table 10: Evidence-level perturbation with Qwen2.5-VL-72B-Instruct. The prompt remains matched except for the frequency-evidence blocks. The removal condition corresponds to the paired LLM-TSAD diagnostic run.

F.5 Evidence-Level Perturbation

Table 10 changes only the frequency-evidence blocks while keeping the remaining prompt matched. Shuffling local evidence reduces F1 from 73.70 to 71.50, using a mismatched global and local donor reduces it to 70.57, and removing frequency evidence reduces it to 67.14. Affiliation F1 follows the same ordering. The graded degradation is consistent with sensitivity to sequence-specific frequency information rather than a generic benefit from adding more text. This analysis remains an evidence-level intervention and does not identify the internal reasoning mechanism of the LLM.

F.6 Local Window and Stride Sensitivity

Table 11 evaluates local-window lengths of 32, 64, and 128 with stride fixed to half of the window

Window L	Stride S	F1	Affi. F1
32	16	73.31	91.36
64	32	**73.70**	**91.39**
128	64	72.77	90.49

Table 11: Full-scale sensitivity to local-window length with stride fixed to half of the window length.

Type	Indexed	LLM-TSAD (+D.S.)	Ours LF+GF	Δ vs. LLM-TSAD
Frequency	23.99	33.98	40.89	+6.91
Point	88.13	77.51	83.30	+5.79
Range	86.39	85.58	83.94	-1.64
Trend	60.28	71.51	86.67	+15.16
Overall	64.70	67.15	73.70	+6.55

Table 12: Type-wise standard F1 with the de-seasonalized LLM-TSAD configuration as the direct reference for the frequency-evidence setting.

length. Standard F1 ranges from 72.77 to 73.70, and affiliation F1 ranges from 90.49 to 91.39. The variation stays below one F1 point across this four-fold window range. This result suggests that the main result does not depend on a single narrowly selected temporal resolution.

F.7 Type-Wise Reference and False-Positive Audit

For the frequency-evidence configuration, the direct reference is the de-seasonalized LLM-TSAD baseline because LF+GF augments that representation. Table 12 shows gains of +6.91 F1 for frequency anomalies, +5.79 for point anomalies, and +15.16 for trend anomalies, with a -1.64 change for range anomalies. The pattern supports a targeted interpretation rather than uniform dominance: frequency evidence contributes most when periodic, spectral, or long-range temporal structure is informative, while amplitude-defined categories can already contain strong time-domain cues.

Table 13 checks whether the additional evidence simply induces broader anomaly prediction. On fully normal sequences, the any-prediction false alarm rate increases from 0.20 to 2.17. In the non-frequency-dominant categories, raw precision increases from 74.63 to 82.22, while both the spurious interval ratio and the number of sequences with a spurious interval decrease. The audit therefore shows a mixed but bounded effect: fully normal inputs become somewhat more sensitive, while the non-frequency-dominant anomaly categories do not show broad over-detection under these measures.

Analysis subset	Metric	LLM-TSAD	Ours LF+GF
Fully normal sequences	Any-prediction false alarm rate	0.20	2.17
Non-frequency-dominant categories	Raw precision	74.63	82.22
Non-frequency-dominant categories	Spurious interval ratio	3.85	2.81
Non-frequency-dominant categories	Sequences with spurious interval	7.28	5.86

Table 13: False-positive audit reported under the same evaluation protocol. Values follow the scale used by the evaluation audit.

Backbone	Metric	LLM-TSAD	Ours	Δ 95% CI
Qwen2.5-VL-72B	Std. F1	67.14	73.70	+6.56 [5.29, 7.72]
Qwen2.5-VL-72B	Affi. F1	82.57	91.39	+8.82 [7.47, 10.09]
GPT-4o	Std. F1	69.03	70.20	+1.17 [-0.19, 2.45]
GPT-4o	Affi. F1	79.98	87.35	+7.37 [5.97, 8.69]
Gemini-2.5-Flash	Std. F1	76.62	77.97	+1.35 [0.04, 2.60]
Gemini-2.5-Flash	Affi. F1	87.12	92.48	+5.36 [4.28, 6.46]
InternVL2-Llama3-76B	Std. F1	38.85	42.15	+3.30 [2.58, 4.03]
InternVL2-Llama3-76B	Affi. F1	69.93	72.41	+2.47 [1.86, 3.13]

Table 14: Paired bootstrap 95% confidence intervals over evaluation instances. The LLM-TSAD values in this table are reproduced paired diagnostics for the confidence-interval analysis and do not replace the main-paper baseline values.

F.8 Paired Bootstrap Confidence Intervals

Table 14 reports paired bootstrap 95% confidence intervals over evaluation instances. Affiliation-F1 intervals remain above zero for all four backbones. Standard-F1 intervals remain above zero for Qwen2.5-VL-72B, Gemini-2.5-Flash, and InternVL2-Llama3-76B, while the GPT-4o interval includes zero. This result supports a more limited claim than uniform improvement: affiliation gains are stable in this paired analysis, whereas standard-F1 gains vary more by backbone. The baseline values in this table come from reproduced paired diagnostic runs and do not replace the main-paper baseline values.

F.9 Newer Backbone Robustness Check

The main evaluation keeps the original backbone set to remain comparable with prior LLM-TSAD settings. Table 15 adds a paired robustness check on GLM-4.6V, Gemini-3-Flash, and GPT-5.1 using the same balanced 1/4-scale subset. The average changes are +2.83 standard F1 and +2.35 affiliation F1. This diagnostic does not serve as a full-scale model leaderboard. It supports the narrower observation that the structured frequency-evidence design remains useful on the evaluated newer backbone set.

Model	Baseline F1 / Affi.	Ours F1 / Affi.	Δ F1	Δ Affi.
GLM-4.6V	51.58 / 75.10	57.78 / 76.87	+6.20	+1.77
Gemini-3-Flash	73.85 / 80.51	75.10 / 82.64	+1.25	+2.13
GPT-5.1	75.77 / 85.64	76.80 / 88.79	+1.03	+3.15
Average	–	–	+2.83	+2.35

Table 15: Paired robustness check on three newer backbones using the same balanced 1/4-scale diagnostic subset. This check complements rather than replaces the full-scale main evaluation.

Setting	Groups	Method	Std.	Affi.	Δ Std.	Δ Affi.
Original	8	LLM-TSAD	36.85	80.15	–	–
Original	8	Ours	40.26	80.35	+3.41	+0.20
Additional	8	LLM-TSAD	13.51	68.46	–	–
Additional	8	Ours	16.14	70.00	+2.63	+1.54
Combined	16	LLM-TSAD	25.18	74.31	–	–
Combined	16	Ours	28.20	75.18	+3.02	+0.87

Table 16: Expanded TSB-AD-U text+vision evaluation. The original eight-group subset follows the submission protocol, and the additional subset contains eight further groups. The combined row averages all 16 groups.

F.10 Expanded TSB-AD-U Subset Evaluation

Table 16 extends the original eight-group TSB-AD-U evaluation with eight additional groups under text+vision inference. The method improves the reported averages on both the original and additional subsets. Across all 16 groups, standard F1 changes from 25.18 to 28.20 and affiliation F1 changes from 74.31 to 75.18. The affiliation change is modest, so the result is best interpreted as a cross-subset robustness check rather than evidence of uniform gains on every TSB-AD-U group.

G Prompt Overhead and Evaluation Details

G.1 Prompt-Length Audit

Config.	Prompts	Avg. tokens	P95	Max	Local windows / Δ
LLM-TSAD	1,600	8,137	8,137	8,137	0 / –
Ours-LF	1,600	8,854	8,854	8,854	30 / +717 (+8.8%)
Ours-LF+GF	1,600	8,970	8,971	8,973	30 / +834 (+10.2%)

Table 17: Prompt-length audit over all 4,800 prompts. Each configuration contains the same 1,000-point indexed history; LF and GF add the structured frequency-evidence text.

Table 17 summarizes the token audit over 4,800 prompts. Because the baseline already contains the same 1,000-point indexed history, LF+GF increases the average text length from 8,137 to 8,970 tokens, corresponding to +834 tokens or +10.2%. The LF-only condition adds +717 tokens or +8.8%. These values quantify representation overhead rather than runtime overhead. Latency and provider-specific API cost can vary with serving conditions, so we do not infer them from token counts alone.

G.2 Decoding, Parsing, and Metric Computation

Item	Value
Decoding	`temperature=0`, max output tokens = 512, request seed = 0
Retry	At most 4 attempts; 30 s wait on HTTP 503; otherwise exponential backoff
Parsing	JSON array between the first `[` and the last `]`, with `start` and `end` fields
Invalid output	Retry; if all retries fail, assign all-zero metrics
Standard F1	Point-wise scikit-learn precision, recall, and F1
Affiliation F1	`affiliation` package
Aggregation	Macro-average over instances, then categories

Table 18: Evaluation and response-processing details used for the reproducibility analysis.

Table 18 specifies the response-processing and evaluation pipeline used in the reproducibility analysis. Decoding uses temperature 0, a maximum of 512 output tokens, and request seed 0. A request receives at most four attempts. HTTP 503 responses use a 30-second wait, while other retryable failures use exponential backoff. Parsing extracts the JSON array between the first opening bracket and the last closing bracket and requires `start` and `end` fields. If all retries fail, the instance receives all-zero metrics. Standard precision, recall, and F1 use point-wise scikit-learn computation, affiliation metrics use the `affiliation` package, and aggregation takes a macro-average over instances and then categories.