

X-ITFormer: Balanced Fusion and Multilevel Auditing for Multimodal Time-Series QA

Leon Felipe Davis Coropuna, Percy Maldonado Quispe, Ana Maria Cuadros Valdivia
Universidad Nacional de San Agustín de Arequipa, Arequipa, Perú

Abstract—LLMs help narrow the semantic gap in time-series analysis by translating numerical patterns into natural-language descriptions. However, their adoption in critical domains is hindered by high computational requirements (models of 7B parameters or more) and by the difficulty of verifying whether predictions rely on the physical signal or on spurious textual correlations. To address these limitations, we propose X-ITFormer, a multimodal framework that builds on ITFormer using a frozen 0.5B-parameter Small Language Model (SLM). It introduces Balanced Fusion, a lightweight pre-attentional normalization mechanism using only 1,794 additional trainable parameters, and a post-hoc X-Module for multi-level auditing through (Level 1) attention-based attribution, (Level 2) k-nearest-neighbor case retrieval, and (Level 3) counterfactual generation. On open-ended queries X-ITFormer leads (0.446 vs. 0.345 Q-HEART and 0.399 ECG-LM) but trails on closed-form tasks (Verify: 0.629 vs. 0.913/0.758; Choose: 0.562 vs. 0.626/0.574); these heterogeneous comparisons do not establish equivalence to larger backbones under a shared protocol. The auditing evaluation shows: 1) attention maps stable under Gaussian noise (cosine similarity 0.9998) with above-chance lead alignment (Top-1 22.81% vs. 8.33%); 2) Balanced Fusion reducing the logged counterfactual mean squared error by 26.3% and the optimization iterations by 38.0% (Wilcoxon, $p < 0.001$) on $N = 1000$ binary ECG-QA samples, at the cost of increased sensitivity to localized adversarial perturbations; and 3) the retrieval module preserving record-level identity and above-random reference-answer agreement in a question-matched cross-record corpus audit. These audits provide partial evidence of signal dependence, rather than a certification of grounding. On TSQA the model outperforms ChatTime only at $L \geq 256$ with the largest advantage at $L = 512$ (+0.0885); on EngineMT-QA, Balanced Fusion yields directionally consistent but statistically non-significant improvements ($p \geq 0.057$, $n = 5$ seeds) at 0.11% of FiLM parameters. A single forward pass uses 2.19 GB VRAM ($B = 1$); SFT cost under \$0.01 USD and emitted 0.0824 kg CO₂eq on ECG-QA. The 0.5B SLM has limited zero-shot reasoning capacity, and performance may degrade on out-of-distribution data.

Keywords—Time series analysis; multimodal learning; multi-modal fusion; question answering; explainable AI; small language models; counterfactual explanations; green AI

I. INTRODUCTION

Time series constitute the backbone of critical systems [1]. However, their analysis faces a fundamental barrier: the *semantic gap*. Unlike natural language, where tokens carry intrinsic meaning, time series operate in continuous numerical spaces devoid of explicit semantics without external context [2]. To close this gap, the integration of Large Language Models (LLMs) has enabled the translation of complex numerical patterns into descriptive and coherent narratives [3].

However, this multimodal convergence introduces new critical risks. Recent architectures such as Time-LLM [4] or ITFormer [5] have improved predictive accuracy, frequently relying on large-scale models (7B+ parameters), which significantly increases memory and computational requirements. Beyond the computational cost, an open problem of *evidence grounding* persists. Because LLM responses may depend strongly on textual priors, it is difficult to determine whether a response is genuinely grounded in the physical evidence of the temporal signal or emerges from correlations learned during pre-training [6]–[9]. This lack of verifiability limits their adoption in high-risk domains such as healthcare, finance, energy systems, smart mobility and transportation, and manufacturing, where errors may translate into economic losses, systemic failures, or risks to human life [10]–[12].

Unlike approaches that depend on massive parameter scaling, we explore whether a properly aligned Small Language Model (SLM) can approximate the deductive capacity of much larger models in domain-specific temporal reasoning tasks. This hypothesis is supported by the scaling laws of [13], which establish that marginal capacity gains diminish as parameters increase beyond the optimal regime for a given task, and by the work of [14], which demonstrates that smaller-scale models trained on sufficient data achieve competitive performance against over-parameterized models. In the context of multi-channel time-series QA, where reasoning operates on bounded morphological patterns and restricted response vocabularies, the required semantic capacity is qualitatively distinct from that needed for general-purpose reasoning, which makes its approximation with smaller-scale architectures plausible.

To address auditability without sacrificing efficiency, we introduce **X-ITFormer**, a multimodal framework aligned with the principles of *Green AI* [15]. Our central question is whether the SLM scaling hypothesis articulated above also permits greater traceability of the decision process.

While ITFormer [5] establishes multimodal fusion through instructional cross-attention, three orthogonal concerns remain relevant design questions that motivate complementary contributions: 1) post-hoc auditing of model behavior, 2) representation-scale control across discrete textual and continuous temporal modalities, and 3) computational efficiency aligned with *Green AI* criteria [15]. X-ITFormer introduces three decoupled additions that target these concerns. First, an X-Module for inspecting model behavior operates alongside the alignment module without modifying the trained parameters, ensuring that the auditing overhead does not interfere with the primary inference path. Second, a pre-attentional normalization mechanism with negligible overhead ($<0.01\%$

of parameters) precedes the cross-attention layer. Third, a frozen 0.5B-parameter backbone is adopted to enable deployment on standard consumer hardware. Balanced Fusion is designed to control representation scales; equalizing norms does not establish equal information contributions, and an exploratory scale and latent-intervention audit distinguishes scale equalization from functional influence. The smaller backbone lowers resource requirements, but its accuracy trade-offs must be assessed separately for each task, and a matched larger-backbone experiment under a common protocol is left for future work.

Our main contributions are:

- Explainability Module (X-Module) for Multilevel Auditing: We implement three complementary auditing tools: 1) attention-based localization; 2) latent case retrieval; and 3) counterfactual sensitivity analysis. Their outputs expose aspects of model behavior without certifying causal explanation or clinical validity.
- Balanced Fusion: We introduce independent per-modality normalization and learnable rescaling with 1,794 trainable parameters. Predictive differences on EngineMT-QA are directionally positive but not statistically significant; correction of textual dominance remains an untested hypothesis.
- Operational Efficiency: We report the memory and energy requirements of a frozen 0.5B-parameter SLM architecture evaluated on consumer hardware. Comparisons with larger models are contextual and do not establish performance equivalence under a common protocol.

II. RELATED WORK

A. Time-Series Analysis and Cross-Modal Alignment

Time-series research has evolved from classical discriminative tasks toward Transformer architectures [1], [16]; patch-based encoders such as PatchTST [17] preserve local semantics, and 2D transformations like TimesNet [18] isolate intra- and inter-period variations. This trend has culminated in Time-Series Foundation Models (TimesFM [19], MOMENT [20]) that establish universal representations for purely numerical domains but lack conversational framing of signal behavior, suffering from the *semantic gap* [2] that limits interactive reasoning.

LLMs have demonstrated competence in general reasoning tasks [3], and multimodal integration began in the visual-linguistic domain [21]. Extending these advances to the temporal domain faces challenges from high resolution and strict sequential dependence [10]: early approaches such as Chat-Time [22] opted for discretization (ignoring mathematical continuity), while continuous-projection models like ITFormer [5] preserve the original granularity without discrete vocabularies.

B. Explainable Time-Series QA

Recent architectures (TempoGPT [23], Q-HEART [24]) focus on accuracy without inference traceability. DiCE [25] extensions to multimodal time-series remain non-trivial, and gradient-based adaptations such as Integrated Gradients [26] or

Attention-SHAP [27] require multiple inference passes. Agent-based prototypes like TimeXL [28] impose 2.29–3.46 s/sample latency with GPT-4o, precluding real-time deployment.

C. Green AI and Identified Gap

The energy implications of deep learning [15], [29], [30] have motivated efficient architectures, and frameworks now quantify computational sustainability [31], [32]. Three limitations persist in the current literature: 1) foundation models lack conversational framing of signal behavior; 2) modality *alignment* is commonly confused with modality *balancing*, leaving the cross-attention vulnerable to dominance bias; and 3) the architectures compared here do not jointly provide continuous projection with all three auditing tools (attention-based attribution, latent retrieval, counterfactual analysis) within a single forward pass plus complementary post-hoc analyses. Table I contrasts representative SOTA architectures along these axes.

III. PROBLEM DEFINITION

Time-series analysis in critical contexts requires evaluating whether responses depend on the observed signal. We formulate three operational audit criteria for this purpose. These criteria organize measurements and expose failures; satisfying an individual criterion does not establish causal or clinically meaningful grounding.

A. Evidence Grounding Definition

Definition 1 (Operational Evidence-Grounding Audit). A prediction a generated from $(\mathbf{X}, q)$ is assessed along three complementary dimensions of signal dependence:

- Locality: The prediction varies systematically with perturbations in regions of high attribution by the model, measured by Comprehensiveness and Locality-RISE (Section IV-C).
- Consistency: Cases from distinct records with compatible questions yield similar predictions, assessed by record-disjoint Label Agreement Rate @ K (LAR@K, Sections IV-C and V-D3a). Identity matches within a record do not establish this criterion.
- Sensitivity: Input perturbations alter the model's decision within a documented optimization budget (Section IV-B3). This measures susceptibility to intervention, not the physiological plausibility or minimality of the perturbation.

Attention maps remain correlational evidence rather than causal explanations [33], [34]. The present results do not establish all three criteria: attribution-faithfulness metrics are close to zero, record-disjoint prediction consistency is task-dependent and Verify agreement falls below its matched random reference, and continuous inversion can exploit adversarial directions. We therefore report an auditing framework with partial evidence of signal dependence, rather than predictions certified as empirically grounded.

TABLE I. FUNCTIONAL COMPARISON AMONG SOTA ARCHITECTURES AND THE PROPOSED MODEL (X-ITFORMER). ENCODING MECHANISMS, REPRESENTATION-SPACE ALIGNMENT STRATEGIES, MODALITY-BALANCING MECHANISMS, AND MULTI-LEVEL EXPLAINABILITY (XAI) CAPABILITIES ARE DISTINGUISHED.

Model	TS	Alignment	Balancing	Explainability (XAI)			External	Efficiency
	Encoder	Strategy	Mechanism	Local	Global	Counterf.	Dependency	(Inference)
ChatTime [22]	Binning + Vocab Exp.	Prompting	N/A	No	No	No	No	Medium
Time-LLM [4]	Reprogramming	Prompting	N/A	No	No	No	No	Low (Large)
TempoGPT [23]	Adaptive Binning	Shared Space	N/A	No	No	No	No	Low (Large)
Q-HEART [24]	CNN+Transf	RAG + Late Fusion	Attention	No	No	No	Yes (RAG)	Low (Large)
TimeXL [28]	Prototypes	Iterative Agent	Prototype Matching	Yes (Text)	Yes (Prototypes)	No	No	Low (Iterative)
ITFormer [5]	Continuous Patching	Cross-Attention	N/A	No	No	No	No	High
X-ITFormer (Prop.)	Continuous Patching	Cross-Attention	Balanced Fusion	Yes (Attention Maps)	Yes (k-NN)	Yes (Iterative Inversion)	No	High (SLM)

B. Evidence Space

Formally, we define *Explainable Time-Series QA* as the joint generation of a textual response and structured auditing evidence. Given a time series $\mathbf{X} \in \mathbb{R}^{L \times V}$ (length L , channels V) and a query q , we seek:

$$f : (\mathbf{X}, q) \mapsto (a, E)$$

where, $E \in \mathcal{E}$ is an evidence tuple $E = (e_{attn}, e_{knn}, e_{cf})$ extracted from three verification mechanisms:

- $e_{attn} = \mathbf{M}_{joint} \in [0, 1]^{V \times L}$: spatio-temporal attention profile whose decision relevance requires perturbation-based assessment.
- $e_{knn} = \{(id_i, s_i)\}_{i=1}^K \subset \mathcal{I} \times [-1, 1]$: set of K reference cases retrieved from the corpus $\mathcal{I}$, where $\mathcal{I}$ is the training corpus and s_i is the cosine similarity in the ABTT-corrected latent space.
- $e_{cf} = \mathbf{X}_{cf} \in \mathbb{R}^{L \times V}$: perturbed signal that alters the model's decision, demonstrating morphological sensitivity.

The space $\mathcal{E}$ is structured as a Cartesian product $\mathcal{E}_{attn} \times \mathcal{E}_{knn} \times \mathcal{E}_{cf}$, where each component captures a complementary aspect of model behavior.

C. Explanatory Quality Criteria

The following criteria guide evaluation of the evidence tuple E ; they are design objectives rather than properties guaranteed by producing the tuple:

- **Attention–Counterfactual Coherence** (design criterion): We hypothesize that regions perturbed in $\mathbf{X}_{cf}$ correlate with high attribution in $\mathbf{M}_{joint}$, explored qualitatively through the case study of Fig. 3.
- **Neighborhood–Decision Consistency**: The retrieved cases e_{knn} must share the same label (exact match for *Verify* and *Choose*) or achieve ROUGE-L-based agreement (for *Query*), as defined by LAR@K (Section IV-C).
- **Perturbation Proximity**: Report the distortion achieved under the optimization budget, distinguishing the logged mean squared error from the final Frobenius norm. Local optimization does not prove a globally minimal perturbation.

These criteria operationalize Definition 1 as complementary diagnostic checks. Their evidential strength must be evaluated separately, and agreement of model predictions need not imply correctness.

D. Functional Architecture

The generative process comprises three stages:

- **Encoding**. Projection to latent spaces: $\mathbf{H}_T = \Phi_T(\mathbf{X})$, $\mathbf{H}_q = \Phi_q(q)$, where Φ_T and Φ_q denote the temporal and textual encoders, respectively.
- **Multimodal Fusion**. Integration via the Instruct-Time Attention mechanism with Balanced Fusion: $\mathbf{H}_{fusion} = \Psi_{BF}(\mathbf{H}_T, \mathbf{H}_q)$ (Section IV), where Ψ_{BF} denotes the composition of the Balanced Fusion pre-attentional normalization and the Instruct-Time Attention block.
- **Response Generation**. The response a is decoded by the frozen LLM Φ_a :

$$a = \Phi_a(\mathbf{H}_{fusion} | q), \quad \Phi_a : \mathbb{R}^{n \times D_{model}} \rightarrow \mathcal{A},$$

where $\mathcal{A}$ is the space of natural language responses. The evidence tuple $E \in \mathcal{E}$ is extracted *post-hoc* by the X-Module (Section IV-B).

This framework extends traditional QA by making the evidence tuple E a mandatory output component, supporting inspection of signal dependence through complementary measurements.

IV. METHODOLOGY

We propose X-ITFormer, a multimodal framework that extends the ITFormer architecture [5] through two complementary contributions: 1) Balanced Fusion (§IV-A), a pre-attentional normalization that stabilizes the cross-modal alignment, and 2) the X-Module (§IV-B), a post-hoc auditing framework that operates without modifying the predictive decision process. As illustrated in Fig. 1, the four-stage inference pipeline (dual encoding $\rightarrow$ preprocessing $\rightarrow$ balanced fusion $\rightarrow$ auditing) is formalized in Section III (Functional Architecture); here we detail its concrete implementation, where the temporal signal is processed by a patch-based encoder and the textual query is embedded via a frozen Small Language Model (SLM).

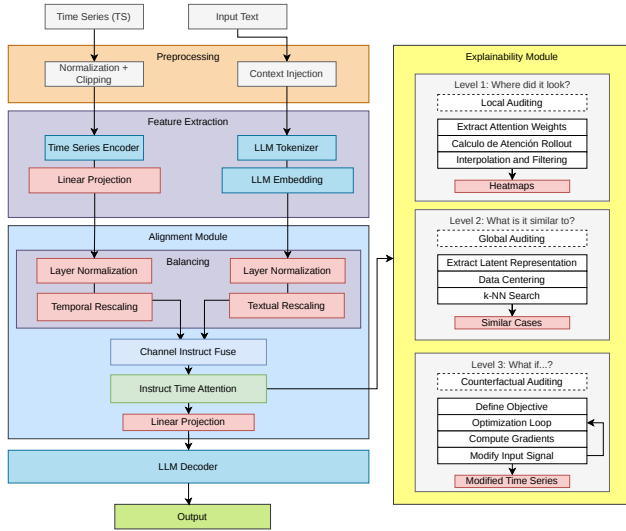


Fig. 1. General architecture of the X-ITFormer pipeline. The flow is divided into four phases: 1) Dual Encoding, where the temporal signal is processed by patches and the text is embedded via an SLM; 2) Preprocessing, which aligns the dimensions of both modalities; 3) Multimodal Fusion, which integrates the representations via a balanced Instruct-Time Attention mechanism; and 4) Explainability, a post-hoc module that audits decisions through attention maps and counterfactual analysis.

A. Balanced Fusion Mechanism

To mitigate the scale discrepancies between textual and temporal representations, X-ITFormer introduces a pre-attentional balancing mechanism based on per-modality independent normalization and learnable scalar modulation. In multimodal scenarios, differences in distribution and magnitude between heterogeneous latent spaces may induce biases in the cross-attention computation, favoring the modality with the larger vector norm.

Given the defined encoded representations, $\mathbf{H}_{ts} \in \mathbb{R}^{L_p \times D_{model}}$ (where L_p is the number of temporal patches) and the textual query $\mathbf{H}_q \in \mathbb{R}^{L_q \times D_{model}}$ (where L_q is the length of textual tokens, assuming $D_{model} = 896$), both are normalized independently using a root-mean-square scheme (RMSNorm [35]), followed by a per-modality scalar rescaling:

$$\begin{aligned}\tilde{\mathbf{H}}_q &= \alpha_{txt} \cdot \text{Norm}_{txt}(\mathbf{H}_q) \\ \tilde{\mathbf{H}}_{ts} &= \alpha_{ts} \cdot \text{Norm}_{ts}(\mathbf{H}_{ts})\end{aligned}$$

where, $\text{Norm}_m(\cdot)$ denotes a per-modality independent normalization function. In this implementation RMSNorm is used, which omits explicit centering and has shown greater stability in modern Transformer architectures. Nevertheless, the mechanism is agnostic to the specific choice of normalization.

Unlike more complex multimodal fusion approaches, such as dynamic *gating* or conditional affine transformations (e.g., FiLM [36]), the proposed mechanism introduces only two learnable scalars (α_{txt} , α_{ts}), enabling dynamic adjustment of the relative contribution of each modality at a minimal parametric cost. This design follows a light modulation philosophy similar to that employed in architectures such as ReZero [37],

where scalar coefficients control information propagation in deep networks.

The module introduces only **1,794** trainable parameters, corresponding to the scale affine parameters of the per-modality normalization (γ_{txt} , γ_{ts}) and the two global scalars. This *overhead* represents a negligible fraction of the total model size ($<0.01\%$), in stark contrast to fusion mechanisms based on MLPs or dense projections.

The global scalar α and RMSNorm affine parameters γ are algebraically composable: $\alpha \cdot (\gamma \odot \mathbf{X}) = (\alpha\gamma) \odot \mathbf{X}$. Their separation permits a shared modality-scale update alongside per-feature updates, a motivation related to redundant parameterization [38]. We do not directly measure gradient-flow improvements or establish that this parameterization prevents instability. The scalars α_{txt} and α_{ts} are initialized at 1.0 and optimized during SFT using AdamW without additional constraints.

Finally, the *Learnable Instruct Tokens* (LIT) $\mathbf{P}_{lit}$ are excluded from the balancing process, preserving their magnitude as invariant anchors. They are concatenated to the balanced textual representation to form the joint input $\mathbf{Q}_{in} = [\mathbf{P}_{lit}; \tilde{\mathbf{H}}_q]$, which together with the temporal memory $\tilde{\mathbf{H}}_{ts}$ feeds the cross-attention block. This attention block formally computes the interaction operator Ψ_{BF} , producing the final fused representation $\mathbf{H}_{fusion}$.

B. Post-Hoc Explainability Framework (X-Module)

The decoding operator $\mathcal{D}$ is functionally implemented through two decoupled components that operate on the same joint representation $\mathbf{H}_{fusion}$, explicitly conditioned by the query q . On one hand, the SLM generates the textual response a autoregressively; on the other, the **X-Module** acts as a *post-hoc* subsystem that extracts the empirical evidence tuple E directly from the internal activations. This unified design supports the **traceability** of the prediction, since the module decomposes model behavior by extracting attention weights, examining latent neighborhoods, and performing counterfactual analysis on exactly the same computational graph that originated the response a , without requiring external approximations or altering the inference process.

1) *Level 1: Local auditing via spatio-temporal attribution*: Local attribution is obtained from the *Instruct-Time Attention* (ITA) block [5], which processes spatial (channels/variables) and temporal features sequentially.

Given the textual query $\mathbf{Q} \in \mathbb{R}^{L_q \times D}$ and the projected memory $\mathbf{M} \in \mathbb{R}^{L_p \times V \times D}$ (where L_q corresponds to instruction tokens, L_p to temporal patches, V to the number of variables, H to the number of attention heads, and D to the model embedding dimension with per-head key dimension $D_k = D/H$), the channel attention first collapses the temporal dimension through a global average, producing weights over the variables:

$$\mathbf{A}_{ch} = \text{Softmax} \left(\frac{\mathbf{Q}_{ch}(\bar{\mathbf{K}}_{ch})^T}{\sqrt{D_k}} \right) \in \mathbb{R}^{B \times H \times L_q \times V}$$

where, $\bar{\mathbf{K}}_{ch}$ corresponds to the projection of the temporally averaged keys. The weighted output of this stage produces an

aggregated memory $\mathbf{M}_{agg} \in \mathbb{R}^{L_p \times D}$, used as *Key* and *Value* in the subsequent temporal attention:

$$\mathbf{A}_{time} = \text{Softmax} \left(\frac{\mathbf{Q}_{time}(\mathbf{K}_{time})^T}{\sqrt{D_k}} \right) \in \mathbb{R}^{B \times H \times L_q \times L_p}$$

To obtain the attribution maps, the attention matrices are aggregated by averaging over the batch (B), attention head (H), and instruction token (L_q) dimensions, excluding padding regions:

$$\bar{\mathbf{a}}_{ch} = \frac{1}{B \cdot H \cdot L_q} \sum_{b=1}^B \sum_{h=1}^H \sum_{l=1}^{L_q} \mathbf{A}_{ch}[b, h, l, :] \in \mathbb{R}^V$$

$$\bar{\mathbf{a}}_{time} = \frac{1}{B \cdot H \cdot L_q} \sum_{b=1}^B \sum_{h=1}^H \sum_{l=1}^{L_q} \mathbf{A}_{time}[b, h, l, :] \in \mathbb{R}^{L_p}$$

Aggregation over L_q yields a single relevance vector per variable and per temporal patch, avoiding the need to analyze the attention of each textual token independently. This procedure is commonly employed in multimodal *grounding* tasks to stabilize attributions [39], [40].

The temporal vector $\bar{\mathbf{a}}_{time}$ is interpolated via an operator Ψ to match the original signal resolution (L), yielding $\hat{\mathbf{a}}_{time} \in \mathbb{R}^L$.

Finally, a joint spatio-temporal map is obtained via the outer product between both vectors:

$$\mathbf{M}_{joint} = \bar{\mathbf{a}}_{ch} \otimes \hat{\mathbf{a}}_{time}$$

This matrix defines the attentional evidence space $\mathbf{M}_{joint} \in \mathcal{E}_{attn} \subset [0, 1]^{V \times L}$. The outer product is a separable approximation [41], not evidence that spatial and temporal contributions are conditionally independent. Its rank is at most one: each lead receives the same temporal profile scaled by its channel weight. It cannot represent, for example, a focal interval in one lead and a different interval in another. Aggregation over heads and query tokens further removes token-specific interactions. Levels 2 and 3 inspect other aspects of behavior but do not restore these lost interactions or validate the map automatically.

If joint localization is required, a non-separable extension would retain the projected channel-patch memory as VL_p distinct keys and extract per-head query-to-memory cross-attention weights of shape $B \times H \times L_q \times (VL_p)$. These can be reshaped to channel-patch maps, aggregated for a specified query, and interpolated in time. The current sequential channel/time attention does not expose this full map: it would require an additional or redesigned attention operation and renewed training/evaluation. The two marginal vectors alone cannot uniquely reconstruct a full joint attention map.

2) *Level 2: Global auditing via latent retrieval*: Global auditing retrieves cases from a multimodal latent representation via k -nearest neighbor search; similarity alone does not validate an analogy. In the backend implementation used for the cross-record audit, the instance embedding is the mean of the $L_{prefix} = 25$ projected ITFormer prefix tokens, $\{\mathbf{h}_p\}_{p=1}^{L_{prefix}}$, before autoregressive decoding:

$$\mathbf{z} = \frac{1}{L_{prefix}} \sum_{p=1}^{L_{prefix}} \mathbf{h}_p \in \mathbb{R}^{D_{model}}.$$

This is the pooled fusion output, not the final textual hidden states of the language model. The embedding depends on the signal and instruction, without including the reference answer.

It is widely known that Transformer latent representations suffer from geometric anisotropy, occupying a narrow cone in vector space where a few dimensions dominate the variance, degrading semantic similarity measurement [42]. To mitigate this phenomenon without incurring the high computational cost of full *whitening*, we apply the *post-hoc All-but-the-Top* (ABTT) transformation [43], [44].

The method proceeds in two phases. First, we apply centering with respect to the global centroid of the training set $\boldsymbol{\mu} \in \mathbb{R}^{D_{model}}$:

$$\tilde{\mathbf{z}} = \mathbf{z} - \boldsymbol{\mu}$$

Subsequently, we eliminate the influence of the c most dominant principal components (which concentrate generic structural information rather than instance-specific semantics). Given the eigenvectors $\mathbf{v}_i$ precomputed *offline* via PCA on the training set, the final isotropic representation $\mathbf{z}'$ is obtained by projecting the global representation outside this dominant subspace:

$$\mathbf{z}' = \tilde{\mathbf{z}} - \sum_{i=1}^c (\tilde{\mathbf{z}}^\top \mathbf{v}_i) \mathbf{v}_i$$

While this $\mathcal{O}(c \cdot D_{model})$ projection offers a reasonable balance between inference efficiency and isotropic correction, we recognize that the construction of intrinsically isotropic models, either through explicit regularization [45] or contrastive training [46], constitutes the current state of the art and the primary direction for future work.

Retrieval uses cosine similarity in the ABTT-corrected latent space. Exact search avoids approximation error in neighbor ranking, which is distinct from semantic relevance or the record-identity Recall@ K reported below. The operational design retrieves an initial $K_{search} = 10K$ candidates and filters matching source identifiers. This bounded candidate pool may return fewer than K eligible neighbors; completeness after filtering requires expanding the search or applying the filter before exact ranking. Importantly, the historical leave-one-out evaluation in Table IX retained other queries from the same record and therefore did not evaluate this record-disjoint design. Section V-D3a formalizes the required evaluation.

3) *Level 3: Counterfactual auditing via iterative inversion:* Sensitivity evaluation is implemented via counterfactual generation by iterative inversion. These perturbations modify the continuous temporal input to probe changes in the model's decision; waveform smoothness does not establish physiological plausibility. Given an original temporal signal $\mathbf{X}_{orig} \in \mathbb{R}^{L \times V}$, a textual prompt, and a target response Y_{target} , we optimize directly over the continuous input space while keeping the model parameters Θ frozen. The objective combines a cross-entropy term $\mathcal{L}_{CE}$ (driving the prediction toward Y_{target}), a proximity term $\mathcal{L}_{Dist} = \frac{1}{LV} \|\mathbf{X}_{cf} - \mathbf{X}_{orig}\|_F^2$ (preserving fidelity to the original signal), and a total-variation term $\mathcal{L}_{TV}$ (an L_1 penalty on adjacent temporal differences, acting as a smoothness regularizer):

$$\mathcal{L}_{total}(\mathbf{X}_{cf}) = \mathcal{L}_{CE} + \lambda_1 \mathcal{L}_{Dist} + \lambda_2 \mathcal{L}_{TV}. \quad (1)$$

The counterfactual signal is initialized as $\mathbf{X}_{cf}^{(0)} = \mathbf{X}_{orig}$ and updated iteratively via the Adam optimizer with learning rate $\eta = 0.1$:

$$\mathbf{X}_{cf}^{(i+1)} = \Pi_{\mathcal{S}} \left(\mathbf{X}_{cf}^{(i)} - \eta \nabla_{\mathbf{X}_{cf}} \mathcal{L}_{total} \right), \quad (2)$$

where, $\Pi_{\mathcal{S}}$ clamps the perturbed input to $\mathcal{S} = [\min(\mathbf{X}_{orig}) - 0.5, \max(\mathbf{X}_{orig}) + 0.5]$ for numerical stability. The optimization budget is 250 updates. Logged target confidence above $\tau_1 = 0.90$ is a stopping criterion, distinct from changing the decoded answer. The implementation also stops when the distance loss is below $\tau_2 = 0.05$ after a grace period ($i > 25$ in zero-based indexing). For binary Verify, response-flip success is exact equality of the final decoded answer to the opposite target answer; it does not require confidence above 0.90. For multi-token targets, logged confidence is the mean teacher-forced probability of the target tokens and is not the probability of the complete generated sequence.

The optimization history stores confidence, distance loss, and TV from the forward pass preceding each update; final responses are decoded after the last update and projection. Consequently, logged terminal metrics and metrics recomputed on the final input must be distinguished. In Tables VI and VII, the field historically named “L2 distance” is the logged $\mathcal{L}_{Dist}$, i.e., mean squared error (MSE), not $\|\mathbf{X}_{cf} - \mathbf{X}_{orig}\|_F$. TV is computed on the perturbed signal, not on perturbation noise. The optimizer seeks a nearby input under the stated objective; it does not guarantee minimum distortion or physiological validity.

C. Explainability Evaluation Metrics

The fidelity of the attribution map $\mathbf{M}_{joint}$ and the consistency of the representation space are evaluated through input-perturbation-based metrics and neighborhood analysis. We rely on four indicators: *Model Confidence* ($CM = P(\hat{y} | \mathbf{X})$) as the reference score; *Comprehensiveness* [47], which measures the drop in CM when the top 20% of attributed positions are masked by the per-channel temporal mean (positive values indicate that the highlighted regions carry decision-relevant information); *Locality-RISE* [48], the difference between the area-under-the-curve of confidence degradation under random occlusion and the curve obtained when positions are occluded

Most Relevant First (MRF) by $\mathbf{M}_{joint}$ (positive values indicate that guided removal degrades confidence faster than chance); and *Label Agreement Rate @ K* (LAR@K), our neighborhood-consistency metric that averages $\mathbb{I}[\text{Agree}(\hat{y}_i, \hat{y})]$ over the K nearest neighbors retrieved in the ABTT-corrected embedding space, with Agree defined as exact match for the *Verify/Choose* tasks and as ROUGE-L agreement for the *Query* task.

D. Training Process

Model training is divided into two phases. In the first, the temporal encoder is trained self-supervised to capture structural regularities of the signal. In the second, multimodal alignment is performed via supervised instruction tuning, keeping the base models frozen and updating only the alignment modules.

1) *Phase 1: Self-supervised pre-training (MSM):* In this phase, the temporal encoder is trained without textual supervision via the *Masked Signal Modeling* (MSM) task. Given a sequence of latent patches $\mathbf{Z}_{ts}$, a binary mask $\mathcal{M}$ is applied with a random masking ratio $r \in [0.6, 0.8]$. The model must reconstruct the masked patches from the visible context.

The loss function is defined as the Mean Squared Error (MSE) between the original and reconstructed patches:

$$\mathcal{L}_{pre} = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} (\mathbf{X}_i - \hat{\mathbf{X}}_i)^2 \quad (3)$$

This pre-training enables initializing the temporal encoder with representations independent of natural language prior to multimodal alignment. It induces structured temporal representations that facilitate subsequent multimodal alignment during the *instruction tuning* phase.

2) *Phase 2: Instruction tuning:* After pre-training, the temporal encoder weights are fixed and supervised training proceeds for multimodal alignment. In this stage, the parameters of the language model (Φ_{llm}) and the temporal encoder (Φ_{ts}) remain frozen. Gradient updates are restricted exclusively to the ITFormer alignment module and its associated linear projectors.

Training uses supervised triplets (Time Series, Question, Answer) and optimizes the autoregressive language modeling loss (causal language modeling):

$$\mathcal{L}_{SFT} = - \sum_{t=1}^T \log P(y_t | y_{<t}, \mathbf{H}_{align}; \Theta_{trainable}) \quad (4)$$

This scheme enables learning the mapping between temporal representations and the semantic space of the language model without modifying the knowledge previously acquired by the base models.

V. RESULTS

We present the results of X-ITFormer (0.5B) on clinical tasks and pattern identification in time-series-based question answering, followed by an assessment of its post-hoc auditing modules and their limitations.

A. Predictive Performance and Generalization

Initial predictive performance was evaluated on the clinical domain via the ECG-QA benchmark [49]. In this dataset, the raw signals consist of multivariate time series with 12 standard leads recorded at 500 Hz over 10 s (5000 samples); for model input they are resampled to 100 Hz (1000 samples), as detailed in Appendix A. X-ITFormer was trained using the union of the training and validation sets (90% of the total), keeping the language model and temporal encoder weights frozen. The official test set (10%) remains completely isolated for final evaluation. The reference model metrics in Table II correspond to the values reported in the literature for this same test set.

The benchmark tasks are organized by level of semantic difficulty: 1) *Verify* (binary yes/no questions), 2) *Choose* (multiple choice among given attributes), and 3) *Query* (open-ended attribute generation). The latter represents the greatest challenge, since the model must identify and write up the findings without explicit options.

As detailed in Table II, X-ITFormer achieves a mean accuracy of 0.446 ± 0.003 ($n = 5$) on the open-ended generation task (*Query*), outperforming ECG-LM (0.399, frozen backbone) and Q-HEART (0.345, LLM adapted via LoRA with external RAG) on this same task, by +0.047 and +0.101 respectively. However, X-ITFormer trails both ECG-LM and Q-HEART on the factual verification tasks, with gaps of -0.129 and -0.284 on *Verify*, and -0.012 and -0.064 on *Choose*. Note that Q-HEART employs augmented retrieval of clinical reports (top-3 from MIMIC-IV) during inference, which changes the information available for factual verification relative to X-ITFormer. These results do not isolate capacity, knowledge access, or adaptation as the cause of the task-specific differences.

X-ITFormer reports higher *Query* accuracy than Q-HEART (0.446 vs. 0.345), but lower *Verify* and *Choose* accuracy, with less than half the backbone parameters (0.5B vs. 1.1B). These task-specific results are contextual comparisons across different adaptation and knowledge-access protocols. The ECG-QA ablation reports positive mean differences on *Query* (+0.0025, from 0.4435 to 0.4460) and *Choose* (+0.0122), but does not establish their statistical significance or identify modality dominance as their cause. The controlled EngineMT-QA ablation is discussed in Section V-C.

To evaluate the generalization capacity on non-clinical domains and measure robustness against sequences of different lengths, the model was compared against ChatTime [22] using the *Time Series Question Answering* (TSQA) set of synthetic univariate series evaluated under different temporal context lengths ($L \in \{64, 128, 256, 512\}$).

Table III reports different trends with sequence length: ChatTime accuracy decreases from 0.8074 at $L = 64$ to 0.6923 at $L = 512$, while X-ITFormer increases from 0.6589 ± 0.0050 to 0.7808 ± 0.0077 . The recovered X-ITFormer artifacts contain five inference executions of one checkpoint, not five independently trained seeds; the reported standard deviations describe repeat-run variability. No significance claim is made from these repetitions. At $L \geq 256$ the reported means exceed ChatTime, with a difference of +0.0885 at $L = 512$. Whether patching explains this pattern requires a controlled analysis. The accompanying supplementary document provides the recovered

Trend/Volatility/Season/Outlier breakdown, sample counts, and provenance limitations (Table X).

ChatTime uses a 7B-parameter backbone with additional pre-training (1M segments) and multi-task fine-tuning (100k instructions), whereas X-ITFormer uses 0.5B parameters (a $14\times$ reduction in backbone size) and 25k TSQA training samples. These different training protocols prevent attributing accuracy differences solely to model size or concluding that 0.5B approximates 7B under matched conditions.

B. Resource Efficiency and Parameter Budget

Computational efficiency was evaluated on a workstation equipped with an NVIDIA RTX 4070 GPU (12 GB VRAM). The basis of our efficiency lies in the architectural design. Of the total **560.93M** parameters in the complete system, the frozen Qwen2.5-0.5B LLM accounts for 88.07% (494.03 M) and the frozen temporal encoder for 3.02% (16.92 M). The alignment module (ITFormer + projectors) holds 47.91 M trainable parameters (8.54% of the total), of which the *Balanced Fusion* mechanism itself contributes only **1,794** parameters (0.0003% of the complete system). This minimalist design preserves a high degree of parameter efficiency. The predictive inference (*forward pass*) with batch size equal to one consumes **2.19 GB** of VRAM and runs in **226 ms**, matching the ITFormer-0.5B baseline exactly (text-only Qwen2.5-0.5B: 0.94 GB, 179 ms). This excludes the post-hoc operations of the explainability module (k -NN search and iterative counterfactual generation). Hardware comparison excludes models with $\geq 7B$ parameters due to memory overflow (OOM on 12 GB VRAM).

To fully align with the methodological guidelines of *Green AI* [15], a transparent environmental assessment requires quantifying the impact of the entire model life cycle. Table IV summarizes the aggregated energy, CO₂eq emissions, and total runtime (sum of pre-training, SFT, and inference) per dataset.

The SFT phase—the most compute-intensive single stage—on ECG-QA required 0.3092 kWh and 0.0824 **kg CO₂eq** for 5 epochs (10 h 26 m); the corresponding figures for EngineMT-QA were 0.0523 kWh and 0.0139 kg, and for ChatTime 0.0189 kWh and 0.0050 kg. The underlying algorithmic effort for SFT, extracted from PyTorch's `total_flops` counter, was 9.349×10^{16} FLOPs for EngineMT-QA and 8.538×10^{17} FLOPs for ECG-QA. Thanks to the efficient 0.5B-parameter backbone, the financial cost of every phase remained below one US cent, which democratizes research reproducibility; exhaustive *Grid Search* was avoided by adopting a cosine scheduler with 10% *warmup*. Energy monitoring of the entire cycle (GPU, CPU, and RAM) was performed via *CodeCarbon* [51], calculating emissions as $\text{CO}_2\text{eq} = E \times \gamma$, where E is the total energy consumed in kWh and $\gamma = 0.266 \text{ kg CO}_2\text{eq/kWh}$ is the regional carbon-intensity factor of the Peruvian grid (MINEM 2024 [52]); the energy values in Table IV are measured, whereas emissions depend on the adopted γ estimate. The contrast between computational efficiency and predictive performance (Section V-A) shows task-specific resource-accuracy trade-offs: on the ECG-QA *Query* task the proposed model (0.5B) reaches 0.446 accuracy (vs. 0.345 Q-HEART and 0.399 ECG-LM); cross-architecture comparisons are reported in Section VI.

TABLE II. DESCRIPTIVE COMPARISON ON THE OFFICIAL ECG-QA TEST SET (10%). BASELINES ARE REPORTED AS PUBLISHED UNDER HETEROGENEOUS ADAPTATION AND KNOWLEDGE-ACCESS PROTOCOLS. METRICS ARE EXACT MATCH ACCURACY BROKEN DOWN BY TASK TYPE. FOR X-ITFORMER, MEAN AND STANDARD DEVIATION OVER $n=5$ INFERENCE EXECUTIONS OF ONE CHECKPOINT ARE REPORTED; THESE ARE NOT INDEPENDENT TRAINING SEEDS, SO THE VALUES ARE DESCRIPTIVE AND NO SIGNIFICANCE TEST IS PERFORMED ACROSS THE TABLE

Method	Configuration			Exact Match Accuracy (10% Test)		
	Backbone	Params	LLM Usage	Verify	Choose	Query
<i>Generative Models: API-Based (Zero-Shot/Few-Shot)</i>						
SE-WRN + Davinci-003 [49]	GPT-3.5	Closed	API	0.750	0.378	0.360
SE-WRN + GPT-4 [49]	GPT-4	Closed	API	0.710	0.481	0.357
SE-WRN + GPT-3.5 [49]	GPT-3.5-Turbo	Closed	API	0.693	0.361	0.311
<i>Generative Models: End-to-End with Frozen or Adapted LLM</i>						
ECG-LM [50]	BioMedGPT	7B	Frozen	0.758	0.574	0.399
X-ITFormer (Ours)	Qwen-2.5	0.5B	Frozen	0.629 ± 0.001	0.562 ± 0.004	0.446 ± 0.003
Q-HEART [24]	LLaMA-1B	1.1B	LoRA	0.913	0.626	0.345

Note: X-ITFormer keeps the Qwen-0.5B backbone and the temporal encoder frozen; only the multimodal alignment modules are trained (ITFormer + Balanced Fusion + projectors, $\sim 48M$ parameters). Q-HEART employs LoRA ($r = 8, \alpha = 32$) on LLaMA-1B and external report retrieval (RAG).

TABLE III. DESCRIPTIVE COMPARISON OF AVERAGE Accuracy BY SEQUENCE LENGTH (L) ON GENERAL TEMPORAL QA TASKS, REPORTED UNDER HETEROGENEOUS TRAINING PROTOCOLS (SEE TEXT). X-ITFORMER VALUES ARE MEAN $\pm$ STANDARD DEVIATION OVER FIVE INFERENCE EXECUTIONS OF ONE CHECKPOINT. THESE ARE NOT INDEPENDENT TRAINING SEEDS. BOLD VALUES INDICATE CASES WHERE X-ITFORMER MATCHES OR EXCEEDS THE CHATTIME BASELINE AT THE SAME SEQUENCE LENGTH; THE DIFFERENCES AT $L \geq 256$ ARE REPORTED WITH EXPLICIT SIGN BUT WITHOUT A FORMAL SIGNIFICANCE TEST, SINCE THE BASELINE DOES NOT REPORT PER-SEED VARIANCE

Len (L)	ChatTime (Base)	X-ITFormer (Ours)	Diff. (Mean)
64	0.8074	0.6589 ± 0.0050	-0.1485
128	0.7893	0.7483 ± 0.0119	-0.0410
256	0.7532	0.7766 ± 0.0068	+0.0234
512	0.6923	0.7808 ± 0.0077	+0.0885

TABLE IV. AGGREGATED *Green AI* AUDIT PER DATASET (SUM OF PRE-TRAINING, SFT, AND INFERENCE OVER THE COMPLETE LIFE CYCLE). ALL PHASES REPORT ENERGY AND EMISSIONS BELOW \$0.01 USD

Dataset	Energy (kWh)	CO ₂ eq (kg)	Cost	Total Time
ECG-QA	0.4272	0.1139	< \$0.01	11h 01m
EngineMT-QA	0.7425	0.1977	< \$0.01	5h 02m
TSQA (ChatTime)	0.1412	0.0376	< \$0.01	1h 33m

C. Internal Mechanism Analysis

The influence of the *Balanced Fusion* mechanism on model behavior was evaluated through three complementary studies: a component ablation analysis on the industrial EngineMT-QA corpus with multiple seeds, a counterfactual sensitivity analysis, and a hyperparameter search for the counterfactual auditing module.

1) *Balanced fusion component ablation*: To isolate the contribution of each element of the proposed mechanism and contextualize its efficiency against higher-capacity alternatives, five configurations are evaluated on the EngineMT-QA corpus. The design covers: 1) the absence of the module (*IT-Former Baseline*); 2) dense conditional modulation *FiLM* [36] as a high-capacity baseline; 3) RMSNorm normalization with affine parameters γ without global scalars (*RMSNorm- γ*); 4)

RMSNorm normalization without affine parameters but with global scalars α (*RMSNorm-core+ α*); and 5) the complete proposed configuration, combining both minimalist elements (*RMSNorm- γ + α*).

To assess variability of the comparisons, each configuration is run with five distinct random seeds ($n = 5$). Table V reports the mean, sample standard deviation, and parametric cost of each metric. To assess the statistical significance of differences between configurations, we apply paired t -tests per seed in Table V, where each configuration shares the same seed set {42, 123, 789, 999, 2024}. For the counterfactual sensitivity study (Table VII, $N=1000$ samples evaluated under both conditions on the same instances), we apply the non-parametric Wilcoxon signed-rank test, which compares paired differences under its distributional assumptions. We acknowledge that with $n=5$ seeds, the statistical power of the paired tests is limited; therefore, we complement them with bootstrap 95% confidence intervals (10,000 resamples) on the mean differences.

The *RMSNorm- γ + α* configuration uses 1,794 parameters versus approximately 1.6M for *FiLM* ($895\times$ more). It has the highest mean on S1, S2, and S3, but none of its paired comparisons against ITFormer reaches the 0.05 threshold ($p \geq 0.057, n = 5$). The observed changes are small and do not establish architectural superiority, equivalence, or absence of a predictive cost. Limited statistical power is one possible explanation; the present experiment cannot distinguish it from a small or negligible effect. The low parameter overhead motivates further investigation, consistent with *Green AI* [15].

The component ablations are descriptive. *RMSNorm-core+ α* has the highest S4 mean (0.5126) and a larger standard deviation (0.0205) than the complete configuration (0.0105); this is not a statistical test of stabilization or synergy. The ECG-QA comparison likewise reports positive mean differences (Choose: +0.0122; Query: +0.0025; Verify: +0.0043), without a significance test. These observations neither establish a domain-general regularizer nor quantify textual dominance. The following scale and latent-intervention audit examines that mechanism without establishing correction of textual dominance.

TABLE V. ABLATION STUDY ON ENGINEMT-QA (MEAN $\pm$ STD, $n = 5$ SEEDS). S1: UNDERSTANDING (ROUGE-L), S2: PERCEPTION (ACC.), S3: REASONING (ACC.), S4: DECISION-MAKING (ROUGE-L). **BOLD** (MEAN COLUMN): BEST MEAN PER TASK, IRRESPECTIVE OF STATISTICAL SIGNIFICANCE. THE SUFFIX *n.s.* AFTER EACH VALUE REPORTS THE PAIRED *t*-TEST PER SEED OF THE **Ours** CONFIGURATION (*RMSNorm- γ + α*) VERSUS *IT-Former* (Baseline): *N.S.* INDICATES $p \geq 0.05$ (STATISTICAL POWER LIMITED BY $n = 5$)

Configuration	S1	S2	S3	S4	Parameters
IT-Former (Baseline)	0.5283 $\pm$ 0.0068	0.5621 $\pm$ 0.0031	0.9734 $\pm$ 0.0053	0.4999 $\pm$ 0.0098	0
FiLM [36]	0.5255 $\pm$ 0.0043	0.5646 $\pm$ 0.0088	0.9764 $\pm$ 0.0031	0.5020 $\pm$ 0.0113	$\sim$ 1.6M
RMSNorm- γ (without α)	0.5262 $\pm$ 0.0044	0.5608 $\pm$ 0.0008	0.9802 $\pm$ 0.0059	0.4974 $\pm$ 0.0163	1,792
RMSNorm-core+ α	0.5233 $\pm$ 0.0055	0.5640 $\pm$ 0.0075	0.9784 $\pm$ 0.0022	0.5126 $\pm$ 0.0205	2
RMSNorm-γ+α (Ours)	0.5311 $\pm$ 0.0062 <i>n.s.</i>	0.5685 $\pm$ 0.0130 <i>n.s.</i>	0.9822 $\pm$ 0.0068 <i>n.s.</i>	0.5087 $\pm$ 0.0105 <i>n.s.</i>	1,794

Significance tests (paired *t*-test, $n = 5$, seeds {42, 123, 789, 999, 2024}): Ours vs. Baseline — S1: $t=2.645$, $p=0.057$; S2: $t=1.168$, $p=0.308$; S3: $t=1.870$, $p=0.135$; S4: $t=1.841$, $p=0.139$. Bootstrap 95% CI of the mean difference (Ours—Baseline): S1 [+0.0010, +0.0047]; S2 [−0.0023, +0.0166]; S3 [+0.0010, +0.0176]; S4 [−0.0006, +0.0167]. The differences are consistent in direction but of small magnitude; we acknowledge the power limitation from $n=5$ seeds.

a) Representation-scale and modality-intervention audit: We measured 512 test queries sampled uniformly without replacement (seed 42), using identical query indices for the two existing ECG-QA checkpoints, 14 ITFormer heads and CPU float32 inference. Per-sample RMS is computed over all token and feature entries, controlling for differing token counts. At the actual Balanced Fusion interface, the mean text/time RMS ratio changes from 0.5673 (record-cluster bootstrap 95% CI [0.5591, 0.5749]) before normalization to 1.0263 ([1.026284, 1.026291]) afterwards. This directly demonstrates scale equalization, not equal semantic influence. At projection outputs, before positional processing, the corresponding ratios are 0.0411 with balancing and 0.0403 without; this is a different measurement location.

To assess alignment sensitivity, we separately zeroed the projected temporal or textual input while leaving the other branch and all weights fixed. We measured $\|z_{\text{zero}} - z\|_2 / \|z\|_2$ on the pooled projected prefix embedding. Mean temporal/textual changes were 0.5050/0.00219 with balancing and 0.5455/0.00216 without. The prefix is substantially more sensitive to temporal zeroing in both configurations. These artificial latent interventions do not measure answer accuracy or the full model’s language reliance: the textual prompt remains available to the LLM during generation. The results therefore support scale control but do not establish correction of textual dominance, equivalent modality contributions, or cluster superiority. Confidence intervals use 2000 record-cluster resamples conditional on the sampled corpus and checkpoints; their different frozen temporal encoders prevent causal attribution of cross-checkpoint differences to balancing.

2) Hyperparameter Ablation for Counterfactual Generation: A sensitivity analysis explored distance and total-variation regularization with $\lambda_1 \in [0.1, 2.0]$ and $\lambda_2 \in [0.1, 5.0]$ on 20 binary samples per configuration. In the recovered artifacts, the recorded success flag denotes a decoded answer change, not necessarily target confidence above 0.90. The two criteria are therefore distinguished when interpreting the reported 100% response-flip rate. This grid explores distortion, smoothness, and optimization effort; it does not establish physiological plausibility or a globally optimal configuration.

a) Morphological fidelity and smoothness analysis: The main selection criterion focused on minimizing the alteration of the original signal (fidelity, measured by lower mean distance), while simultaneously maintaining high temporal

continuity and avoiding the introduction of high-frequency noise (morphological smoothness, measured by low TV).

TABLE VI. REPRESENTATIVE SELECTION FROM THE COUNTERFACTUAL SENSITIVITY ANALYSIS. THE SELECTED CONFIGURATION ($\lambda_1 = 2.0$, $\lambda_2 = 1.5$) HAS THE SMALLEST MEAN LOGGED MSE AMONG THE DISPLAYED ROWS. TV DESCRIBES SIGNAL SMOOTHNESS, NOT VALIDATED PHYSIOLOGICAL PLAUSIBILITY

λ_1 (Dist)	λ_2 (TV)	Logged MSE ($\downarrow$)	TV ($\downarrow$)	Target Prob. ($\uparrow$)
0.1	0.1	0.5714	1.0992	0.9021
0.5	2.0	0.3777	0.6919	0.9001
1.0	1.0	0.3967	0.8311	0.9017
2.0	1.5	0.2927	0.6726	0.8960
2.0	2.0	0.3575	0.6709	0.9001
2.0	5.0	0.5307	0.5156	0.8579

Table VI shows a trade-off between logged MSE and signal TV. Among the displayed configurations, $\lambda_1 = 2.0$, $\lambda_2 = 1.5$ has the lowest mean logged MSE (0.2927) and TV of 0.6726. A stronger TV penalty ($\lambda_2 = 5.0$) lowers TV to 0.5156 but increases MSE to 0.5307. Mean target confidences of 0.8960 and 0.8579 are below 0.90: the recorded answer-flip rate must not be interpreted as universal attainment of that threshold.

b) Efficiency and convergence speed analysis: Optimization effort is measured by the number of updates executed before confidence-based stopping, the distance-based stopping rule, or budget exhaustion. It is not uniformly the time to attain confidence above 0.90. Configurations with low or moderate TV penalty (e.g., $\lambda_2 \leq 1.0$) use fewer steps in the recovered grid, including 71.2 on average for $\lambda_1 = 0.5$, $\lambda_2 = 0.5$.

At $\lambda_1 = 2.0$, $\lambda_2 = 5.0$, the mean number of updates is 182.15, versus 113.75 for the selected $\lambda_1 = 2.0$, $\lambda_2 = 1.5$ configuration. These are observed optimization budgets; they do not demonstrate successful high-confidence convergence in every case or maximum morphological quality.

3) Counterfactual sensitivity analysis: To systematically evaluate the effect of *Balanced Fusion* on the model’s robustness to morphological perturbations, a counterfactual sensitivity study was performed on $N = 1000$ binary verification (*yes/no*) samples from the ECG-QA corpus. Configurations with and without the balancing mechanism were compared, keeping identical the optimization hyperparameters ($\lambda_1 = 2.0$, $\lambda_2 = 1.5$, $\eta = 0.1$, max iter=250).

Table VII separates decoded response changes from logged confidence attainment. Both configurations change the answer to the opposite target in all 1000 cases, whereas $P_{target} > 0.90$ in the last logged forward pass occurs in only 390/1000 cases without balancing and 876/1000 with balancing. The mean logged MSE is 0.328 ± 0.157 with balancing versus 0.445 ± 0.269 without it, a 26.3% reduction. This is an objective-space comparison, not a reduction of the raw Frobenius norm or evidence of clinically valid counterfactuals.

The mean number of executed updates decreases from 210.5 to 130.6 (38.0%); this does not directly measure wall-clock time or energy. Without balancing, 613 runs execute all 250 steps, of which 610 finish with logged confidence at or below 0.90. With balancing, these counts are 126 and 124. Thus the median of 250 without balancing is compatible with a 100% decoded response-flip rate: an answer may change without reaching the confidence stopping threshold. The median with balancing is 118.

The logged MSE maxima are 2.00 without balancing and 1.59 with balancing, with a median of 0.285 in the latter condition. These describe distortion in the normalized input space; their physiological significance has not been established.

These paired results show lower logged distortion and fewer executed updates for this binary inversion procedure with Balanced Fusion. They do not directly measure the contribution of each modality or establish robustness against adversarial inputs. Easier inversion is useful for probing decisions but may also expose exploitable sensitivity.

D. Interpretability and Explainability (XAI)

The explainability of X-ITFormer is evaluated through a post-hoc hierarchical analysis structured in three levels, corresponding to the X-Module defined in Section IV-B: (Level 1) spatio-temporal attribution, (Level 2) global retrieval of analogous cases, and (Level 3) counterfactual auditing. Before proceeding with this formal explanatory audit, we assess the model's dependence on its input modalities via inference-time ablation. This analysis acts as a robustness *pre-test* and does not form part of the X-Module proper, but rather establishes the interpretive basis upon which subsequent levels are interpreted.

1) *Robustness and modal dependency*: Replacing the ECG with Gaussian noise $\mathcal{N}(0, 1)$ reduces accuracy from 54.74% to 41.57%, a drop of 13.17 percentage points (24.1% relative). This supports partial signal dependence, but the substantial residual accuracy is compatible with textual priors, answer-frequency effects, or other shortcuts; the intervention alone cannot isolate them. Removing the instruction reduces accuracy to 8.35% (84.7% relative), showing the importance of task conditioning rather than proving clinically meaningful signal use. Complete signal replacement is distinct from the small additive-noise attention audit ($\sigma = 0.05$) reported below.

2) *Level 1: Spatial alignment analysis*: This analysis evaluates the ability of the *Channel Instruct Fusing* module to concentrate attention on the leads mentioned in the textual instruction. Qualitative inspection of two case studies (instructed target aVL and Lead I) shows that the maximum attention coefficient (α_{max}) always corresponds to the requested lead, but the model also distributes attention to physiologically

correlated adjacent leads—in Case 1 the model attends simultaneously to leads I and V3, while in Case 2 attention spreads between aVL and II. This is consistent with the projective nature of the ECG and provides qualitative evidence that the model distributes attention across structurally related regions of the signal rather than collapsing onto a single source.

A corpus-level audit on $N = 320$ samples (Table VIII) compares the lead named in the instruction with the attention ranking. Uniform random selection gives Top-1 8.33% and Top-3 25.00%; X-ITFormer reports 22.81% and 57.19%, respectively. These descriptive values exceed the uniform reference, but Top-1 agreement is low and instruction alignment does not establish decision-faithful localization. The rank-1 restriction further limits localized interpretation. Higher audit levels provide additional measurements, without compensating automatically for weak attribution fidelity or demonstrating the full criteria of Definition 1.

3) *Level 2: Retrieval module evaluation (k -NN)*: The historical retrieval audit measures record identity retained in multimodal embeddings, using cosine similarity (S_{cos}). It does not measure semantic agreement across independent records.

In illustrative records without focal markers, mean top-5 cosine similarity is 0.435, versus 0.801 in records with explicit markers. These selected cases do not establish corpus-wide structural separability. Retrieval of record ID 8376 under different questions, and the first seven neighbors of query ID 1393 sharing its record, instead demonstrate identity retention. Within-record similarities of $[0.847, 0.950]$ and the drop to 0.784 on changing records cannot isolate temporal identity from question semantics. Table IX reports leave-one-query-out identity retrieval on 23,154 queries from 10,760 record IDs. Other queries from the same record remain eligible, unlike the record-disjoint operational design. This full-corpus audit is separate from the 2,315-query predictive test partition and does not establish held-out cross-record generalization.

a) *Formalized cross-record evaluation protocol*: To isolate semantic retrieval, let r_i be the record identifier of query i and define the eligible candidate set:

$$\mathcal{C}_i = \{j \in \mathcal{I} : r_j \neq r_i, t_j = t_i, u_j = u_i\},$$

where, t is the QA task and u is a canonical question template including the clinical finding or lead being queried. For a generalization evaluation, $\mathcal{I}$ is restricted to training records and queries to record-disjoint test records; ABTT is fitted only on training embeddings. If the historical split is query-based, a new record-disjoint split must be constructed. An audit within the full corpus must instead be identified as transductive. Patient-level exclusion is additionally required when patient identifiers are available and patient generalization is claimed.

For each query, rank all eligible candidates by cosine similarity and retain $K_i = \min(K, |\mathcal{C}_i|)$ neighbors. Filtering only a fixed $10K$ shortlist does not guarantee the top- K eligible neighbors. Report the number of excluded queries with $K_i = 0$ and counts with $K_i < K$. Compute prediction $\text{LAR@}K$ by averaging exact answer agreement over the K_i neighbors for Verify/Choose; for Query, report mean ROUGE-L agreement without converting it to an unspecified binary

TABLE VII. AGGREGATED COUNTERFACTUAL SENSITIVITY ($N = 1000$ SAMPLES). MSE AND TV ARE TERMINAL LOGGED OBJECTIVE VALUES; STEPS COUNT EXECUTED UPDATES. ANSWER-FLIP SUCCESS AND CONFIDENCE-THRESHOLD ATTAINMENT ARE DIFFERENT OUTCOMES

Metric	Without Balancing	With Balancing	Relative change
Logged MSE (mean $\pm$ std)	0.445 $\pm$ 0.269	0.328 $\pm$ 0.157	-26.3%
Logged MSE (median)	0.361	0.285	-21.0%
Iterations (mean $\pm$ std)	210.5 $\pm$ 60.3	130.6 $\pm$ 67.9	-38.0%
Iterations (median)	250.0 (limit)	118.0	-52.8%
Total Variation (TV)	0.741 $\pm$ 0.205	0.708 $\pm$ 0.148	-4.5%
Decoded answer-flip rate	100%	100%	—
Logged $P_{target} > 0.9$	39.0%	87.6%	—
Runs at 250 steps	613/1000	126/1000	—

Paired Wilcoxon tests on all $N = 1000$ instances: logged MSE: $W = 150088$, $p = 5.7 \times 10^{-28}$; executed updates: $W = 46017$, $p = 3.3 \times 10^{-94}$. The statistics were reproduced from the stored per-instance CSVs. TV refers to the perturbed signal in these CSVs; it is distinct from the artifact metric `TV_noise`, so no test using that different quantity is attached to this row. Logged probabilities precede the final update, whereas answer-flip rates validate the subsequently decoded input.

TABLE VIII. TEXT-TO-LEAD ALIGNMENT PREFERENCE (*Lead Alignment*, $N = 320$). THE RANDOM BASELINE CORRESPONDS TO A UNIFORM SELECTION AMONG 12 LEADS

Metric	Random Baseline (Uniform)	X-ITFormer	Relative Improvement
Top-1 Preference	8.33%	22.81%	2.74 $\times$
Top-3 Preference	25.00%	57.19%	2.29 $\times$

threshold. Report agreement of reference answers separately to distinguish model consistency from correctness. Compare against uniformly sampled eligible neighbors under the same task/template restrictions, and report class frequencies and record-clustered bootstrap intervals. Semantic Recall@ K requires a separately defined relevance set and cannot reuse identity Recall@ K .

b) Cross-record retrieval audit: We executed a transductive audit on the available 23,154-query corpus (10,760 ECG record IDs), regenerating embeddings for the balanced and no-balancing checkpoints under identical extraction settings (14 ITFormer heads, BF16 autocast on CUDA with float32 weights/storage). Queries were batched by equal token length to avoid unmasked query padding. ABTT removes one principal component fitted on this audit corpus; it is not a held-out fitting procedure. Eligible neighbors have the same task and exact case/whitespace-normalized instruction, including findings, leads, and answer options. Every matching record ID is excluded before exact cosine ranking. In this corpus the compatible groups also contain distinct record IDs, so selected neighbors do not duplicate a record.

Table X measures *reference-answer agreement*: exact match for Verify/Choose and mean ROUGE-L F1 with stemming for Query. It is a proxy for semantic relevance of retrieved cases, distinct from prediction LAR@ K in Definition 1. A matched random reference is the exact expected agreement under uniform sampling without replacement from the eligible pool. Queries with no eligible neighbor are excluded: 8 Verify, 752 Choose, and 1 Query. Results use $K_i = \min(K, |\mathcal{C}_i|)$; full- K coverage and record-cluster bootstrap intervals (2000 resamples) are supplied with the supplementary audit (Table X). Choose has small pools, with no full- $K = 10$ cases; its $K = 10$ row covers all eligible neighbors and therefore equals the random expectation.

At $K = 5$, X-ITFormer reaches Query reference agreement 0.3478 versus matched random 0.3121, and Verify

agreement 0.5902 versus 0.5703. Both configurations show above-random agreement on these tasks, but balancing does not consistently improve retrieval: its Query difference from the no-balancing checkpoint is -0.0015 (record-cluster 95% CI $[-0.0057, +0.0031]$), while Verify differs by $+0.0067$ ($[+0.0022, +0.0114]$). These are conditional checkpoint comparisons with unadjusted intervals, not evidence of architectural superiority. The frozen temporal-encoder weights differ between checkpoints, so the comparison does not isolate normalization causally.

The available transformed training and test files share 10,754 record IDs; 23,146 audit queries use records also present in training. Identity exclusion at retrieval does not remove training exposure, and these results do not establish unseen-record or patient generalization. Reference-answer agreement is also an indirect relevance proxy, not expert-validated morphological similarity. Historical response files failed a parity spot-check; the completed audit instead uses freshly generated responses for both checkpoints.

c) Cross-record prediction consistency: Fresh responses were generated for both checkpoints using greedy decoding, a 128-new-token limit, and the same extraction and retrieval settings. Table XI reports prediction LAR separately from reference-answer agreement. At $K = 5$, balanced Query LAR is 0.7300 versus random 0.6586, with gain 0.0714 (record-cluster 95% CI $[0.0643, 0.0779]$). Without balancing the corresponding gain is 0.0701 ($[0.0633, 0.0769]$). Balanced Verify LAR is 0.9457 versus random 0.9535, a negative gain of -0.0078 ($[-0.0107, -0.0052]$): high absolute agreement alone is therefore uninformative. Choose LAR is 0.4226 versus random 0.4216, with a small gain 0.0010 ($[0.0005, 0.0014]$) and mean effective $K_i = 2.46$. These results support task-dependent cross-record response consistency, strongest on Query, without establishing uniform grounding or a balancing advantage.

The fresh-response audit ran on an NVIDIA GeForce RTX 4070 without retraining or model downloads. Extraction and generation took approximately 17.7 minutes for the balanced checkpoint and 18.4 minutes without balancing, using BF16 autocast with float32 weights/storage; peak allocated GPU memory was recorded in the audit provenance. These timings are not energy measurements and are separate from the original CodeCarbon results.

TABLE IX. RECORD-IDENTITY RETRIEVAL AT LEVEL 2. RECALL@ K REPORTS THE FRACTION OF QUERIES WHOSE K NEAREST NEIGHBORS IN THE LATENT SPACE CONTAIN AT LEAST ONE EXAMPLE SHARING THE SAME RECORD ID AS THE QUERY (RECORD-IDENTITY MATCH). MRR IS THE MEAN RECIPROCAL RANK OF THE FIRST IDENTITY-MATCHING NEIGHBOR. THE UNIFORM-RECORD REFERENCE IS ILLUSTRATIVE, NOT A QUERY-LEVEL RANDOM-NEIGHBOR BASELINE: RECORD MULTIPLICITIES AFFECT THE LATTER. THESE VALUES DO NOT MEASURE SEMANTIC RELEVANCE

Metric	Value	Uniform-record ref.	Ratio to ref.	Note
Recall@1	0.6919	$\frac{1}{10760}$ (0.00009)	$\times$ 7445	query-level identity match
Recall@5	0.7187	$5 \times \frac{1}{10760}$ (0.00046)	$\times$ 1547	query-level identity match
Recall@10	0.7276	$10 \times \frac{1}{10760}$ (0.00093)	$\times$ 783	query-level identity match
MRR	0.7034	—	—	first identity match rank

Scope: shared ECG inputs can place queries from the same record near each other even without semantic retrieval. The reported Recall@ K and MRR characterize identity preservation only. They neither demonstrate cross-record relevance nor establish the consistency criterion of Definition 1. Ratios to the uniform-record reference must not be interpreted as gains against a matched random-query baseline.

TABLE X. RECORD-DISJOINT, QUESTION-MATCHED RETRIEVAL: REFERENCE-ANSWER AGREEMENT (EXACT MATCH FOR VERIFY/CHOOSE, MEAN ROUGE-L F1 FOR QUERY). $K_i = \min(K, |\mathcal{C}_i|)$; COUNTS EXCLUDE QUERIES WITH NO ELIGIBLE NEIGHBOR. THIS TRANSDUCTIVE AUDIT DOES NOT ESTABLISH GENERALIZATION TO UNSEEN ECG RECORDS.

Task	K	Queries	Mean K_i	No balancing	Balanced	Random
Choose	1	6070	1.00	0.3680	0.3769	0.3364
Choose	5	6070	2.46	0.3364	0.3370	0.3364
Choose	10	6070	2.51	0.3364	0.3364	0.3364
Query	1	7673	1.00	0.3494	0.3548	0.3121
Query	5	7673	5.00	0.3493	0.3478	0.3121
Query	10	7673	9.99	0.3449	0.3424	0.3121
Verify	1	8650	1.00	0.5911	0.6039	0.5703
Verify	5	8650	4.99	0.5835	0.5902	0.5703
Verify	10	8650	9.92	0.5828	0.5842	0.5703

TABLE XI. RECORD-DISJOINT PREDICTION LAR@ K AND MATCHED RANDOM EXPECTATION. QUERY USES CONTINUOUS ROUGE-L F1; VERIFY/CHOOSE USE EXACT AGREEMENT. COVERAGE AND EFFECTIVE K_i ARE AS IN TABLE X.

Task	K	No balancing	Random	Balanced	Random
Query	1	0.7280	0.6429	0.7393	0.6586
Query	5	0.7131	0.6429	0.7300	0.6586
Query	10	0.7080	0.6429	0.7249	0.6586
Verify	1	0.9437	0.9591	0.9351	0.9535
Verify	5	0.9525	0.9591	0.9457	0.9535
Verify	10	0.9566	0.9591	0.9504	0.9535
Choose	1	0.4728	0.4179	0.4954	0.4216
Choose	5	0.4188	0.4179	0.4226	0.4216
Choose	10	0.4179	0.4179	0.4216	0.4216

4) Level 3: Counterfactual Auditing of Model Decision:

To evaluate the decisional behavior of the system beyond attentional correlations, we performed a counterfactual analysis based on optimized adversarial perturbations ($\mathbf{x} + \delta$). The objective is to seek an input modification that changes the decision within a fixed optimization budget; minimum distortion is not guaranteed.

In the **suppression** case (sample 171), the decoded prediction changes from *ischemic in inferolateral leads* to “none”. The stored terminal objective is MSE 0.0735, with target confidence approximately 0.93 and six executed updates (the last zero-based index is five). Recomputing distortion on the final saved input gives MSE 0.0914 and Frobenius norm 33.1096 over 12×1000 normalized values. The difference from the logged MSE reflects the final update and projection; 0.0735 must not be described as a raw L_2 norm. The attention

redistribution visualized in Section V-E is descriptive, without establishing clinically valid removal of pathology.

In the **induction** case (sample 10), optimization toward *non-diagnostic t abnormalities* uses 50 updates, reaches logged MSE 0.5742 and target confidence 0.65, and does not change the decoded answer. These two cases illustrate different optimization outcomes but do not establish a general directional asymmetry. The binary inversion study of Table VII separately reports 100% decoded answer changes and lower confidence-attainment rates. It exposes susceptibility to optimized continuous perturbations, rather than validated physiological counterfactuals or robustness to adversarial inputs.

On the ECG-QA predictive test partition ($N = 2,315$), attention maps under additive Gaussian noise ($\sigma = 0.05$) have cosine similarity 0.9998, and occlusion of the highest-attention channel gives sensitivity of 12.02%. Attention stability alone does not demonstrate prediction robustness. Comprehensiveness and Locality-RISE on $N = 100$ test samples remain close to zero, limiting the decision-faithfulness interpretation of attention [33]. Together with the task-dependent record-disjoint prediction-consistency results, these findings prevent certification of grounding under Definition 1.

E. Qualitative Case Study

To complement the aggregated counterfactual statistics of Table VII ($N=1000$), Fig. 2 and 3 illustrate the per-lead attention redistribution between the original ECG record (sample ID 171) and its counterfactual perturbation on a single representative instance. The comparison highlights the localized segments of the signal whose modification drives the decision inversion with terminal logged MSE 0.0735.

Case study context:

- **Query:** “Which diagnostic symptom does this ECG show, ischemic in inferolateral leads or ischemic in anterior leads, excluding uncertain symptoms? assistant”
- **Ground truth:** ischemic in inferolateral leads
- **Original prediction:** ischemic in inferolateral leads (correct)
- **Counterfactual target:** none (forced inversion)
- **Result after perturbation (logged MSE 0.0735, six Adam updates):** prediction flips to *none*, while the

channel attention top-3 shifts from (aVL, I, aVR) to (aVL, V6, V5), evidencing the localized sensitivity that Level 3 captures.



Fig. 2. Original state (sample ID 171, prediction: *ischemic in inferolateral leads*). The top-3 attended leads (aVL, I, aVR) are highlighted in red on the bar chart.

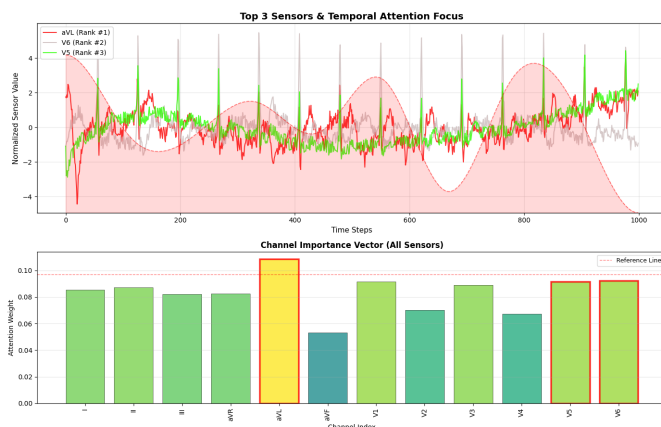


Fig. 3. Counterfactual state (sample ID 171 post-correction: prediction flips to *none* with logged MSE 0.0735). The top-3 attended leads shift to aVL, V6, and V5, evidencing the localized sensitivity that Level 3 captures and the redistribution of attentional focus onto the modified segment.

VI. DISCUSSION AND CONCLUSION

X-ITFormer combines a frozen 0.5B backbone with three post-hoc auditing tools. On ECG-QA it reports higher Query accuracy (0.446) than Q-HEART (0.345) and ECG-LM (0.399), but lower Verify (0.629 vs. 0.913/0.758) and Choose (0.562 vs. 0.626/0.574) accuracy. Verify also falls below the reported zero-shot API baselines. These are heterogeneous comparisons, not evidence of equivalent performance to a 7B model under matched conditions.

1) Capacity and comparison limits: A frozen 0.5B backbone restricts the linguistic and factual capacity available to the alignment module: alignment training cannot update the backbone to acquire domain knowledge. Verify and Choose may require precise discrimination of findings and answer alternatives, whereas Query may benefit from more flexible

generation. Q-HEART additionally adapts its backbone with LoRA and retrieves reports. These factors are plausible explanations for the factual-performance gap, but their causal contributions have not been isolated. A matched larger-backbone experiment would require identical record splits, inputs, knowledge access, adaptation, decoding, and metrics, alongside measured resource costs; that experiment has not been performed. The present contribution is a measured small-backbone implementation with task-specific trade-offs, not demonstrated 0.5B–7B equivalence. **Future work.** A controlled experiment with a $\geq 7B$ backbone under the present protocol—identical record splits, input preprocessing, knowledge access, frozen or adapted status, decoding settings, and metrics, alongside reported resource costs—remains for future work; the present 12 GB workstation could not host the required counterfactual-optimization memory budget. Same-backbone 0.5B–7B comparisons alone would not establish the claim of equivalence between scales; matched-protocol evaluation is required.

2) Fusion and grounding limits: EngineMT-QA paired tests do not establish predictive superiority of Balanced Fusion ($p \geq 0.057$, five seeds), and the scale/intervention audit does not establish correction of modality dominance. Its lower logged inversion MSE and 38.0% fewer updates concern the counterfactual procedure, not universal accuracy or an equivalent reduction in energy/time. Top-1 lead alignment of 22.81%, near-zero attribution-faithfulness metrics, and accuracy of 41.57% after replacing the ECG with noise indicate limited evidence of grounding. The identity audit is supplemented by question-matched, record-disjoint reference-answer agreement above the matched random reference on Query and Verify. The latter is transductive evidence of retrieval relevance, not unseen-record generalization; fresh prediction LAR exceeds its matched random reference on Query but falls below it on Verify. Continuous inversion probes model sensitivity without guaranteeing physiological meaning. The TSQA supplement recovers five inference executions of one checkpoint; their variability does not represent independent training-seed uncertainty. Distribution-shift performance and clinical validity remain unevaluated.

3) Ethical considerations and adversarial robustness: X-ITFormer is a research auditing system and is not validated for autonomous clinical use. Targeted continuous optimization can change decoded answers even when the 0.90 confidence stopping threshold is not reached. This sensitivity raises safety concerns, although it does not by itself establish susceptibility to every form of random noise. Before any safety-critical deployment, defenses requiring evaluation include adversarial training with domain-constrained attacks, noise injection during training, and clipping or rejecting inputs outside validated acquisition ranges. These should be tested against adaptive attacks and ordinary acquisition artifacts, with attention to effects on clinically relevant waveform features and task accuracy. Existing Z-score clipping and optimization-range projection provide numerical bounds, not a demonstrated security guarantee. None of these proposed robustness defenses was evaluated in this study. Physiological constraints, external clinical validation, uncertainty-based escalation, and human review would also be required. Public-benchmark results do not establish deployment readiness; all experiments used publicly available benchmarks without patient-identifiable information.

REFERENCES

- [1] K. Benidis, S. S. Rangapuram, V. Flunkert, Y. Wang, D. Maddix, C. Turkmen, J. Gasthaus, M. Bohlke-Schneider, D. Salinas, L. Stella, F.-X. Aubet, L. Callot, and T. Januschowski, "Deep learning for time series forecasting: Tutorial and literature survey," *ACM Comput. Surv.*, vol. 55, no. 6, Dec. 2022. [Online]. Available: <https://doi.org/10.1145/3533382>
- [2] Y. Kong, Y. Yang, Y. Hwang, W. Du, S. Zohren, Z. Wang, M. Jin, and Q. Wen, "Time-MQA: Time series multi-task question answering with context enhancement," in *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, W. Che, J. Nabende, E. Shutova, and M. T. Pilehvar, Eds. Vienna, Austria: Association for Computational Linguistics, Jul. 2025, pp. 29 736–29 753. [Online]. Available: <https://aclanthology.org/2025.acl-long.1437/>
- [3] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.
- [4] M. Jin, S. Wang, L. Ma, Z. Chu, J. Y. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan, and Q. Wen, "Time-LLM: Time series forecasting by reprogramming large language models," in *International Conference on Learning Representations*, 2024. [Online]. Available: <https://openreview.net/forum?id=Un35WH2AxE>
- [5] Y. Wang, P. Lei, J. Song, Y. Hao, T. Chen, Y. Zhang, L. Jia, Y. Li, and Z. Wei, "Itformer: Bridging time series and natural language for multi-modal qa with large-scale multitask dataset," 2025. [Online]. Available: <https://arxiv.org/abs/2506.20093>
- [6] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann, "Shortcut learning in deep neural networks," *Nature Machine Intelligence*, vol. 2, no. 11, pp. 665–673, Nov. 2020.
- [7] M. T. Ribeiro, S. Singh, and C. Guestrin, "'why should i trust you?': Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 1135–1144. [Online]. Available: <https://doi.org/10.1145/2939672.2939778>
- [8] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, May 2019.
- [9] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A survey on multimodal large language models," *National Science Review*, vol. 11, no. 12, p. nwae403, 12 2024.
- [10] X. Zhang, R. R. Chowdhury, R. K. Gupta, and J. Shang, "Large language models for time series: A survey," in *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, K. Larson, Ed. International Joint Conferences on Artificial Intelligence Organization, 8 2024, pp. 8335–8343, survey Track. [Online]. Available: <https://doi.org/10.24963/ijcai.2024/921>
- [11] Q. Wen, T. Zhou, C. Zhang, W. Chen, Z. Ma, J. Yan, and L. Sun, "Transformers in time series: a survey," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, ser. IJCAI '23, 2023. [Online]. Available: <https://doi.org/10.24963/ijcai.2023/759>
- [12] J. F. Torres, D. Hadjout, A. Sebaa, F. Martínez-Álvarez, and A. Troncoso, "Deep learning for time series forecasting: A survey," *Big Data*, vol. 9, no. 1, pp. 3–21, 2021.
- [13] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," *arXiv preprint arXiv:2001.08361*, 2020.
- [14] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. W. Rae, and L. Sifre, "Training compute-optimal large language models," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.
- [15] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green ai," *Commun. ACM*, vol. 63, no. 12, p. 54–63, Nov. 2020. [Online]. Available: <https://doi.org/10.1145/3381831>
- [16] B. Lim and S. Zohren, "Time-series forecasting with deep learning: a survey," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 379, no. 2194, p. 20200209, 02 2021.
- [17] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *International Conference on Learning Representations*, 2023.
- [18] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "Timesnet: Temporal 2d-variation modeling for general time series analysis," in *International Conference on Learning Representations*, 2023.
- [19] A. Das, W. Kong, R. Sen, and Y. Zhou, "A decoder-only foundation model for time-series forecasting," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 10 148–10 167. [Online]. Available: <https://proceedings.mlr.press/v235/das24c.html>
- [20] M. Goswami, K. Szafer, A. Choudhry, Y. Cai, S. Li, and A. Dubrawski, "MOMENT: A family of open time-series foundation models," in *Proceedings of the 41st International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, R. Salakhutdinov, Z. Kolter, K. Heller, A. Weller, N. Oliver, J. Scarlett, and F. Berkenkamp, Eds., vol. 235. PMLR, 21–27 Jul 2024, pp. 16 115–16 152. [Online]. Available: <https://proceedings.mlr.press/v235/goswami24a.html>
- [21] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds, R. Ring, E. Rutherford, S. Cabi, T. Han, Z. Gong, S. Samangooei, M. Monteiro, J. Menick, S. Borgeaud, A. Brock, A. Nematzadeh, S. Sharifzadeh, M. Binkowski, R. Barreira, O. Vinyals, A. Zisserman, and K. Simonyan, "Flamingo: a visual language model for few-shot learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.
- [22] C. Wang, Q. Qi, J. Wang, H. Sun, Z. Zhuang, J. Wu, L. Zhang, and J. Liao, "Chattime: a unified multimodal time series foundation model bridging numerical and textual data," in *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'25/IAAI'25/EAAI'25. AAAI Press, 2025. [Online]. Available: <https://doi.org/10.1609/aaai.v39i12.33384>
- [23] H. Zhang, C. Yang, J. Han, L. Qin, and X. Wang, "Tempogpt: Enhancing time series reasoning in multi-modal language model via quantizing embedding," *Neurocomputing*, vol. 664, p. 132035, 2026. [Online]. Available: <https://doi.org/10.1016/j.neucom.2025.132035>
- [24] H. M. Pham, J. Tang, A. Saeed, and D. Ma, "Q-heart: Ecg question answering via knowledge-informed multimodal llms," in *Proceedings of the European Conference on Artificial Intelligence (ECAI)*, ser. Frontiers in Artificial Intelligence and Applications, vol. 413. IOS Press, 2025, pp. 4545–4552.
- [25] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, ser. FAT* '20. New York, NY, USA: Association for Computing Machinery, 2020, p. 607–617. [Online]. Available: <https://doi.org/10.1145/3351095.3372850>
- [26] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3319–3328.
- [27] E. Kokalj, B. Škrlić, N. Lavrač, S. Pollak, and M. Robnik-Šikonja, "BERT meets shapley: Extending SHAP explanations to transformer-based classifiers," in *Proceedings of the EACL Hackshop on News Media Content Analysis and Automated Report Generation*, H. Toivonen and M. Boggia, Eds. Online: Association for Computational Linguistics, Apr. 2021, pp. 16–21. [Online]. Available: <https://aclanthology.org/2021.hackashop-1.3/>
- [28] Y. Jiang, W. Yu, G. Lee, D. Song, K. Shin, W. Cheng, Y. Liu, and H. Chen, "Timexl: Explainable multi-modal time series prediction with llm-in-the-loop," in *Advances in Neural Information Processing Systems*, vol. 38, 2025, pp. 150 147–150 183.

- [29] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, A. Korhonen, D. Traum, and L. Màrquez, Eds. Florence, Italy: Association for Computational Linguistics, Jul. 2019, pp. 3645–3650. [Online]. Available: <https://aclanthology.org/P19-1355/>
- [30] D. Patterson, J. Gonzalez, Q. Le, C. Liang, L.-M. Munguia, D. Rothchild, D. So, M. Texier, and J. Dean, "Carbon emissions and large neural network training," *arXiv preprint arXiv:2104.10350*, 2021.
- [31] A. Lacoste, A. Luccioni, V. Schmidt, and T. Dandres, "Quantifying the carbon emissions of machine learning," in *NeurIPS 2019 Workshop on Tackling Climate Change with Machine Learning*, 2019.
- [32] A. S. Luccioni, S. Viguier, and A.-L. Ligozat, "Estimating the carbon footprint of bloom, a 176b parameter language model," *J. Mach. Learn. Res.*, vol. 24, no. 1, Jan. 2023.
- [33] S. Jain and B. C. Wallace, "Attention is not explanation," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 3543–3556.
- [34] S. Wiegrefe and Y. Pinter, "Attention is not not explanation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, 2019, pp. 11–20.
- [35] B. Zhang and R. Sennrich, "Root mean square layer normalization," in *Advances in Neural Information Processing Systems* 32, vol. 32. Curran Associates, Inc., 2019, pp. 12 360–12 371.
- [36] E. Perez, F. Strub, H. de Vries, V. Dumoulin, and A. Courville, "Film: visual reasoning with a general conditioning layer," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence and Thirtieth Innovative Applications of Artificial Intelligence Conference and Eighth AAAI Symposium on Educational Advances in Artificial Intelligence*, ser. AAAI'18/IAAI'18/EAAI'18. AAAI Press, 2018.
- [37] T. Bachlechner, B. P. Majumder, H. Mao, G. Cottrell, and J. McAuley, "Rezero is all you need: fast convergence at large depth," in *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence*, ser. Proceedings of Machine Learning Research, C. de Campos and M. H. Maathuis, Eds., vol. 161. PMLR, 27–30 Jul 2021, pp. 1352–1361. [Online]. Available: <https://proceedings.mlr.press/v161/bachlechner21a.html>
- [38] T. Salimans and D. P. Kingma, "Weight normalization: a simple reparameterization to accelerate training of deep neural networks," in *Proceedings of the 30th International Conference on Neural Information Processing Systems*, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 901–909.
- [39] H. Chefer, S. Gur, and L. Wolf, "Transformer interpretability beyond attention visualization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 782–791.
- [40] A. Kamath, M. Singh, Y. LeCun, G. Synnaeve, I. Misra, and N. Carion, "MDETR - modulated detection for end-to-end multi-modal understanding," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 1780–1790.
- [41] G. Bertasius, H. Wang, and L. Torresani, "Is space-time attention all you need for video understanding?" in *Proceedings of the 38th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, M. Meila and T. Zhang, Eds., vol. 139. PMLR, 18–24 Jul 2021, pp. 813–824. [Online]. Available: <https://proceedings.mlr.press/v139/bertasius21a.html>
- [42] K. Ethayarajh, "How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, K. Inui, J. Jiang, V. Ng, and X. Wan, Eds. Hong Kong, China: Association for Computational Linguistics, Nov. 2019, pp. 55–65. [Online]. Available: <https://aclanthology.org/D19-1006/>
- [43] J. Mu and P. Viswanath, "All-but-the-top: Simple and effective post-processing for word representations," in *International Conference on Learning Representations*, 2018.
- [44] V. Raunak, V. Gupta, and F. Metze, "Effective dimensionality reduction for word embeddings," in *Proceedings of the 4th Workshop on Representation Learning for NLP (RepLanLP-2019)*, I. Augenstein, S. Gella, S. Ruder, K. Kann, B. Can, J. Welbl, A. Conneau, X. Ren, and M. Rei, Eds. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 235–243. [Online]. Available: <https://aclanthology.org/W19-4328/>
- [45] S. Rajasee and M. T. Pilehvar, "A cluster-based approach for improving isotropy in contextual embedding space," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, C. Zong, F. Xia, W. Li, and R. Navigli, Eds. Online: Association for Computational Linguistics, Aug. 2021, pp. 575–584. [Online]. Available: <https://aclanthology.org/2021.acl-short.73/>
- [46] T. Gao, X. Yao, and D. Chen, "SimCSE: Simple contrastive learning of sentence embeddings," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, M.-F. Moens, X. Huang, L. Specia, and S. W.-t. Yih, Eds. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 6894–6910. [Online]. Available: <https://aclanthology.org/2021.emnlp-main.552/>
- [47] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, "ERASER: A benchmark to evaluate rationalized NLP models," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, Eds. Online: Association for Computational Linguistics, Jul. 2020, pp. 4443–4458. [Online]. Available: <https://aclanthology.org/2020.acl-main.408/>
- [48] V. Petsiuk, A. Das, and K. Saenko, "RISE: Randomized input sampling for explanation of black-box models," in *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press, 2018, p. 151.
- [49] J. Oh, G. Lee, S. Bae, J.-m. Kwon, and E. Choi, "Ecg-qa: a comprehensive question answering dataset combined with electrocardiogram," in *Proceedings of the 37th International Conference on Neural Information Processing Systems*, ser. NIPS '23. Red Hook, NY, USA: Curran Associates Inc., 2023.
- [50] K. Yang, M. Hong, J. Zhang, Y. Luo, S. Zhao, O. Zhang, X. Yu, J. Zhou, L. Yang, P. Zhang, M. Qiao, and Z. Nie, "ECG-LM: Understanding electrocardiogram with a large language model," *Health Data Science*, vol. 5, p. 0221, 2025.
- [51] B. Courty, V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld *et al.*, "mlco2/codecarbon: v2.4.1," May 2024. [Online]. Available: <https://doi.org/10.5281/zenodo.11171501>
- [52] Ministerio de Energía y Minas del Perú, "Producción eléctrica nacional con energías renovables representó el 8.1% del total en abril de 2024," Nota de prensa, Dirección General de Electricidad, julio 2024. [Online]. Available: <https://www.gob.pe/institucion/minem/noticias/983429-minem-produccion-electrica-nacional-con-energias-renovables-represento-el-8-1-del-total-en-abril-de-2024>

APPENDIX

A. Technical Specifications per Dataset

Table XII consolidates the structural and training hyperparameters used in each pipeline evaluated in this work.

Implementation. All experiments were implemented in Python 3.10 with PyTorch 2.x (CUDA 12.1) and Transformers 4.37+, executed on a single workstation equipped with an NVIDIA GeForce RTX 4070 (12 GB), an AMD Ryzen 5 7600X (6 cores), and 32 GB RAM under BF16 mixed precision. **Training.** Phase 1 (MAE pre-training of the temporal encoder): AdamW, peak LR 1×10^{-4} , linear decay, weight decay 1×10^{-5} , 30 epochs, batch size 32. Phase 2 (Alignment SFT) keeps both the LLM and the temporal encoder frozen and trains only the ITFormer alignment module and projectors, using AdamW, peak LR 5×10^{-6} , cosine decay with 10% warmup, weight decay 1×10^{-6} , 5 epochs, batch size 4, gradient accumulation of 4 (effective 160), gradient clipping at 1.0. **Preprocessing.** Signals are resampled to 100 Hz,

TABLE XII. TECHNICAL SPECIFICATIONS AND PIPELINE CONFIGURATION PER DATASET. THE ARCHITECTURAL AND TRAINING HYPERPARAMETERS CORRESPOND LITERALLY TO THE ARGUMENTS USED DURING THE TRAINING CYCLE.

Parameter	ECG-QA	TSQA (ChatTime)	EngineMT-QA
<i>(a) Temporal Encoder</i>			
input_len / channels	1000 / 12	512 / 1	600 / 1
patch_len / stride	100 / 100	64 / 64	60 / 60
d_model / e_layers / n_heads	512 / 4 / 8	512 / 4 / 8	512 / 4 / 8
dropout	0.1	0.1	0.1
<i>(b) Alignment Module (ITFormer)</i>			
tt_d_model / tt_n_heads	896 / 14 / 2	896 / 14 / 2	896 / 16 / 2
tt_layers			
tt_dropout / prefix_num	0.1 / 25	0.1 / 25	0.1 / 25
<i>(c) Frozen LLM</i>			
Backbone	Qwen2.5-0.5B-Instruct	Qwen2.5-0.5B-Instruct	Qwen2.5-0.5B-Instruct
<i>(d) Training</i>			
pretrain_epochs / pretrain_bs	30 / 32	30 / 512	10 / 128
sft_epochs / sft_bs / grad. accum.	5 / 4 / 4	5 / 10 / 4	3 / 4 / 4
sft_lr / bal. fusion	5×10^{-6} / true / 1.0	5×10^{-6} / true / 1.0	5×10^{-6} / true / 1.0
fusion_alpha_init			
min/max_mask	0.6 / 0.8	0.6 / 0.8	0.6 / 0.8
<i>(e) Inference</i>			
Decoding / max_new	greedy / 128	greedy / 32	greedy / 128

normalized per lead via Z-score, and clipped to $[-5, +5]$ for BF16 stability; sequences are right-padded to the longest sequence in the batch with masked padding regions; deterministic 90%/10% split with seed 42. **Tokenization and prompt.** We adopt the Qwen2.5 Tokenizer (Fast) with left padding for generation and right padding for training; the 25 $\langle |image_pad| \rangle$ tokens are learned latent queries of the ITFormer module that decouple the LLM context-window usage from the input-signal length, prepended under the ChatML format. **Inference.** Greedy decoding (num_beams=1, do_sample=False),

temperature 1.0, maximum 128 new tokens, repetition penalty 1.0 (disabled). **Reproducibility.** The EngineMT-QA paired ablation uses five seeds {42, 123, 2024, 999, 789}. TSQA metrics are recovered from five inference executions of a single checkpoint, with unrecorded seed fields and identical reported metrics for the last three executions; they are not independent training-seed estimates. The counterfactual study pairs 1000 instances across configurations. Due to CUDA non-determinism, numerical reproducibility across hardware may vary.