

# PowerMTQA: Dual-Stream Multimodal Temporal Question Answering for Power-System Customer Service

Guorui Wang, Jiang Jiang, Jinchun Ma, and Dingxiao Liu (✉)

**Abstract** Power-system customer service requires answering user queries by jointly interpreting chart images, textual instructions, and high-precision temporal measurements. Existing multimodal large language models are effective at visual-language interaction but remain weak at exact temporal reasoning and reliability-aware decision support. We propose PowerMTQA, a dual-stream multimodal temporal question-answering framework that combines a Qwen-based visual-language backbone with a dedicated multichannel time-series branch, prototype reprogramming, explicit cross-attention fusion, and auxiliary anomaly/risk supervision. Experiments on a unified dataset built from four public electricity and power-related time-series sources show that explicit temporal modeling is critical: dual-stream variants reduce the validation generative loss from 24.5611 for a vision-only baseline to 0.0809 in the single-channel setting, while multichannel modeling and LoRA further improve the best loss to 0.0721 with 0.9788 token accuracy. On answer-level semantic consistency evaluated by an external judge, PowerMTQA reaches 0.82 accuracy, outperforming the visual-only and text-only baselines by 15 and 39 percentage points, respectively. These results indicate that combining chart understanding with structured temporal reasoning substantially improves factual consistency and practical reliability in power-service question answering.

**Keywords** multimodal large language model, time-series reasoning, power-system customer service, question answering, cross-attention fusion

## 1 Introduction

Power-system customer service is no longer limited to answering routine billing questions or providing static outage

notices. In modern utility operations, customer-facing intelligence must explain abnormal consumption patterns, interpret load and voltage fluctuations, summarize device alarms, and provide cautious recommendations under operational constraints. The challenge lies in integrating domain-specific knowledge, temporal data, and high reliability requirements. The answer space is shaped by power-engineering terminology, service regulations, equipment status, and utility workflows. Many service questions can only be resolved after inspecting multivariate time-series records such as load, current, voltage, temperature, or transformer indicators over a recent window. At the same time, the response must remain reliable and auditable, because a fluent but incorrect explanation may mislead users, trigger unnecessary service actions, or conceal genuine risk.

Recent progress in large language models (LLMs) and multimodal large language models (MLLMs) has significantly improved conversational interaction, instruction following, and joint visual-language reasoning<sup>[1–8]</sup>. In parallel, the time-series community has proposed increasingly capable long-sequence architectures, including Informer, Autoformer, PatchTST, and TimesNet<sup>[9–12]</sup>, while emerging time-series foundation models such as Chronos, Lag-Llama, and Time-LLM have shown that numerical signals can be aligned with the reasoning capacity of language models<sup>[13–15]</sup>. These developments suggest a promising direction for intelligent power service: instead of treating charts, text, and sensor values as isolated data products, they can be modeled jointly within a unified question-answering framework.

However, directly deploying a general-purpose MLLM in this setting remains unsatisfactory. A curve image may convey rough trend shape, but it does not preserve the precision required for threshold checking, anomaly localization, and multivariable comparison. Conversely, converting long numeric sequences into raw text tokens is both inefficient and lossy, and often fails to preserve stable quantitative regularities<sup>[9–13]</sup>. Moreover, real customer-service deployment requires more than natural-sounding answers: the model should expose structured signals, such as anomaly presence or risk level, so that downstream systems can audit, filter, or escalate uncertain outputs.

• Guorui Wang is with the Economics and Management School, Wuhan University, Wuhan 430073, China.

• Jiang Jiang is with the Enterprise Architecture and Digitalization Department, Guangdong Power Grid Corporation, Guangzhou 510180, China.

• Jinchun Ma is with ZhipuAI, Beijing 100190, China.

• Dingxiao Liu is with ZhipuAI, Beijing 100190, China. E-mail: dingxiao.liu@aminer.cn.

\* To whom correspondence should be addressed. E-mail: dingxiao.liu@aminer.cn.

Manuscript received: /-/; revised: /-/; accepted: /-/

This paper therefore focuses on a practical yet underexplored problem: **multimodal temporal question answering for power-system customer service**. We define the task around a time window aligned across three modalities:

- an image modality  $I$ , typically a line chart, multivariable curve chart, or diagnostic rendering of the same temporal slice;
- a text modality  $X$ , including the user question, task instruction, compact statistical context, and, when available, lightweight threshold descriptors;
- a multichannel time-series tensor  $S \in \mathbb{R}^{C \times T}$ , containing one target sequence and several support variables.

Given these inputs, the model produces a customer-service style natural-language response  $Y$ , and optionally predicts structured labels  $Z$ , such as `anomaly` and `risk_level`. This formulation differs from standard visual question answering because the image is not the sole evidence source; it also differs from time-series forecasting because the target is not merely a future numeric point or distribution, but an interpretable service-oriented explanation. The task therefore requires the model to integrate visual cues, precise numerical evidence, and domain-specific instructions within a single reasoning loop.

From an operational perspective, this abstraction matches typical utility scenarios. A user may ask why the electricity bill increased over the past week, whether a sudden load spike indicates unsafe behavior, or whether abnormal transformer temperature suggests urgent maintenance. In each case, the answer depends on joint reasoning over chart morphology, high-precision temporal values, service knowledge, and, when available, operating-threshold context. A useful model should therefore support both qualitative explanation and structured screening judgment.

Existing pipelines remain insufficient for this setting. Many multimodal assistants are strong at interpreting natural images, optical character recognition (OCR)-rich screenshots, and visually summarized charts. Yet chart images are compressed views of underlying temporal processes. Important operational decisions in power systems often depend on exact value comparisons, localized threshold violations, and cross-channel dependencies. A vision-only model may detect that a curve rises sharply, but it may fail to determine whether the rise exceeds a safety threshold, whether it co-occurs with voltage sag, or whether the event is unusual for the device category. This gap is especially problematic in customer service, where the answer must often justify itself with accurate trends and concrete conditions rather than impressionistic visual language.

A naive alternative is to serialize the numeric sequence into text and append it to the prompt. This approach scales poorly as the number of channels or time steps increases, and it forces the language model to infer quantitative structure from tokenized numbers. Prior work on time-series modeling has shown that explicit patching, normalization, and reprogramming are better suited to preserving temporal regularities than raw numeric concatenation<sup>[9–13, 16]</sup>. For power-service applications, where both window length and variable count may grow, a more structured numerical interface is needed.

Generative answers are attractive because they provide natural and user-friendly explanations. Nevertheless, in safety-sensitive domains, free-form generation alone is not enough. A customer-service system must frequently answer questions that contain implicit binary or ordinal judgments: whether the observed pattern is abnormal, whether the operational risk is low or high, and whether escalation is advisable. If these judgments remain hidden inside the language model’s internal state, they are difficult to verify and even harder to integrate with downstream service rules. A more deployable architecture should therefore couple generative output with explicit auxiliary predictions.

We propose **PowerMTQA** (A Dual-Stream Multimodal Temporal Reasoning and Question Answering Model for Power Service), a dual-stream architecture that combines a visual-language backbone with an explicit time-series branch. Concretely, PowerMTQA keeps a Qwen-VL-family backbone for image-text understanding<sup>[5–8]</sup>, introduces a multichannel time-series encoder inspired by Time-LLM-style reprogramming<sup>[13]</sup>, injects temporal tokens into language hidden states through cross-attention, and optimizes the model with a joint objective consisting of generative question answering and auxiliary anomaly/risk classification.

The goal of this study is not to replace operational dispatch systems, supervisory control tools, or formal protection logic. Rather, the model is positioned as an intelligent customer-service layer that can summarize evidence, explain temporal behavior in natural language, and provide structured cues for downstream verification. In a real deployment, the model can be combined with retrieval-augmented generation (RAG), regulation databases, and live tool calling for outage or ticket lookup<sup>[17, 18]</sup>. The contribution of this paper is therefore centered on the **model-side problem**: how to endow a multimodal conversational system with stronger temporal understanding while preserving interpretability and engineering compatibility.

This framing is particularly valuable for power utilities

because customer questions often mix operational diagnosis with conversational expectations. Users do not ask for raw tensors or classification labels; they ask for explanations, causes, and suggestions. At the same time, the service provider cannot rely on style alone. PowerMTQA is designed to bridge this gap by combining multimodal interaction, temporal representation learning, and structured reliability signals within one architecture.

The main contributions of this paper are summarized as follows.

- (1) **Task formulation.** We formulate power-system intelligent customer service as a multimodal temporal question-answering problem that unifies chart interpretation, textual instruction following, and precise multichannel time-series reasoning within a single window-based interface.
- (2) **Model architecture.** We propose PowerMTQA, a dual-stream multimodal generation framework that augments a visual-language backbone with a dedicated time-series branch and explicit cross-attention fusion, enabling high-fidelity temporal evidence to participate in answer generation.
- (3) **Reliability-oriented learning objective.** We introduce auxiliary anomaly and risk supervision on top of generative training so that the model learns not only how to respond fluently, but also how to expose structured service-relevant judgments.
- (4) **Reproducible experimental planning.** We organize the overall method around modular components, including multichannel temporal input design, prototype-based reprogramming, parameter-efficient tuning, and ablation-friendly fusion, thereby providing a clear path for systematic validation.

## 2 Related Works

### 2.1 Multimodal Vision-Language Models

Multimodal large models have rapidly evolved from loosely coupled vision-language systems into unified instruction-following assistants capable of image understanding, grounded dialogue, and multi-image reasoning. Representative designs differ in how visual information is bridged into the language model. Flamingo introduces a Perceiver Resampler and interleaved cross-attention layers to connect visual tokens with a frozen language backbone<sup>[3]</sup>. BLIP-2 proposes the Q-Former as a lightweight querying module that maps frozen vision features into a language-friendly representation<sup>[4]</sup>. LLaVA demonstrates that strong multimodal instruction tuning can be obtained by pairing CLIP-style vi-

sion features with a conversational decoder<sup>[5]</sup>. Qwen-VL and its later generations further emphasize multilingual capability, robust document and chart understanding, and scalable image-text interaction<sup>[6–8]</sup>.

These models are highly relevant to the present study because power-service scenarios often include chart images, dashboard snapshots, and textual questions. Their strong instruction-following and visual grounding abilities make them attractive backbones for applied multimodal systems. However, most of them are optimized for general-purpose visual understanding rather than precision-oriented temporal reasoning. In chart-heavy operational settings, a model may correctly identify a rising or oscillating pattern from the image while still missing exact threshold crossings, subtle multichannel dependencies, or statistical context unavailable in the rendered figure. Therefore, although visual-language models provide a strong starting point, they do not by themselves solve the high-fidelity reasoning problem required for power-service analysis.

A second issue concerns interface compatibility. Many multimodal architectures are designed around image patches or region features, whereas time-series evidence is continuous, multichannel, and often far longer than a typical image token sequence. This mismatch motivates a design in which the visual-language backbone is preserved for chart and text understanding, while a separate numerical branch is introduced for structured temporal information.

### 2.2 Time-Series Foundation Models

The time-series community has made major progress in long-horizon sequence modeling over the past several years. Informer reduces the complexity of Transformer attention through ProbSparse attention and sequence distillation, enabling more efficient long-sequence forecasting<sup>[10]</sup>. Autoformer incorporates decomposition and autocorrelation mechanisms to better capture trend and periodic structure<sup>[11]</sup>. PatchTST shows that patch tokenization and channel-independent processing are highly effective for long-term forecasting, demonstrating that temporal patching can play a role analogous to tokenization in language modeling<sup>[9]</sup>. TimesNet reformulates one-dimensional temporal variation into a two-dimensional representation and proposes a unified backbone for a range of time-series tasks<sup>[12]</sup>.

These methods establish an important premise for our work: raw continuous sequences benefit from dedicated architectural biases rather than being treated as ordinary text. This is especially true in power systems, where temporal

signals display multiple scales, abrupt anomalies, distribution shift across devices, and variable-specific dynamics. Reversible instance normalization (RevIN) addresses one practical challenge in this setting by reducing the harmful effect of distribution shift<sup>[16]</sup>. Because utility data may span different customers, transformers, and environmental conditions, such normalization becomes a natural ingredient for stable modeling.

Recent work has further pushed toward time-series foundation models. Chronos discretizes rescaled time series and trains language-model-style objectives on the resulting token stream<sup>[14]</sup>. Lag-Llama explores pretraining for probabilistic time-series forecasting under a foundation-model perspective<sup>[15]</sup>. A particularly relevant line of work is Time-LLM, which treats a frozen LLM as a reasoning engine and adapts time-series inputs through prompt engineering and prototype-based reprogramming<sup>[13]</sup>. Instead of retraining the entire language model, Time-LLM uses patch embeddings, text prototypes, and prompt-as-prefix statistics to align numerical sequences with the linguistic space. This idea is highly influential for the present study because it offers a principled way to connect continuous numerical signals with a large language backbone. Recent multimodal time-series surveys further stress that performance gains are most reliable when complementary modalities are aligned through explicit interfaces rather than flattened into a single textual stream<sup>[19]</sup>.

Nevertheless, the objective of Time-LLM is primarily forecasting, not multimodal customer-service explanation. It does not consider chart images, visual-language instruction tuning, or structured service outputs such as anomaly and risk labels. Our work inherits the reprogramming philosophy but extends it into a dual-stream multimodal question-answering setting.

### 2.3 Chart Understanding and Numerical Reasoning

Chart understanding occupies a distinctive position between vision and structured data analysis. Benchmarks such as ChartQA highlight that answering chart-related questions requires both visual interpretation and logical reasoning over plotted values<sup>[20]</sup>. However, chart-question answering datasets typically focus on the figure itself and do not provide the full raw numerical stream needed for precise anomaly detection or multivariate diagnosis. In operational domains, the chart is often only a compact rendering of a richer temporal record.

This distinction matters for power-service applications. A user may ask whether a voltage drop is correlated with a

load spike or whether recent fluctuations indicate risky behavior. A chart-only interface can suggest shape-level correspondence, but fine-grained reasoning may depend on exact values, temporal lags, or support variables that are not legible in the image. Thus, chart understanding is necessary but insufficient. The problem is not merely to see the chart, but to jointly reason over the chart as a human-readable view and the underlying numerical signals as machine-precise evidence.

The broader multimodal literature also points to a persistent modality gap between continuous numerical data and discrete language tokens. When numeric sequences are flattened into prompt text, token budgets increase rapidly while quantitative structure becomes less explicit. Reprogramming, patch embedding, and explicit cross-modal fusion offer more scalable alternatives. Our method builds on this observation by treating chart images and time-series values as complementary rather than interchangeable modalities.

### 2.4 Reliable Dialogue with Retrieval and Tools

Knowledge-intensive dialogue systems have long been known to suffer from factual errors when they rely only on parametric memory. Retrieval-augmented generation (RAG) addresses this issue by retrieving external documents and conditioning generation on the retrieved evidence<sup>[17]</sup>. ReAct and related tool-augmented prompting frameworks further improve controllability by interleaving reasoning with external actions such as search or application programming interface (API) calls<sup>[18]</sup>. These ideas are particularly important for domain-specific service systems, where compliance, traceability, and access to live operational data matter as much as linguistic quality.

Power-system customer service naturally benefits from this line of work. Utility regulations, frequently asked questions, outage notices, ticket histories, and device records often reside outside the model parameters. At the same time, many customer questions require access to current measurements or service status. Therefore, an end-to-end deployable assistant should ideally combine (i) a strong multimodal reasoning core, (ii) external knowledge retrieval, and (iii) tool invocation for real-time lookup.

Our focus in this paper is narrower but complementary: we address the modeling challenge of temporal multimodal reasoning itself. In other words, we study how the model should process chart, text, and multichannel numerical input before it is connected to RAG or external tools. This modular view is practical because it separates two distinct reliability sources: one from better evidence representation inside the

model, and another from better external grounding at the system level.

## 2.5 LLMs in Power Systems

The application of LLMs to power and energy systems is attracting increasing attention, including tasks such as forecasting, report generation, rule interpretation, simulation assistance, and decision support. Recent surveys summarize a rapidly growing literature on power-oriented large models, covering forecasting, fault diagnosis, regulation understanding, decision support, and engineering assistance<sup>[21, 22]</sup>. Existing discussions in the area generally emphasize either text-centric capabilities, such as summarizing documents and assisting with engineering language, or purely numerical tasks, such as prediction and operational analytics. Recent domain studies also show that reliability, grounding, and safety constraints remain major barriers when general-purpose LLMs are adapted to high-stakes energy workflows<sup>[23]</sup>. What remains insufficiently studied is therefore the intersection of these capabilities: conversational explanation grounded in multimodal temporal evidence.

Power-system customer service is a representative case of this unmet need. The system must explain what happened, why it matters, whether it is abnormal, and what action is reasonable, often based on mixed evidence sources. A model limited to text reasoning may not fully exploit temporal measurements; a model limited to forecasting may not communicate the result in a service-ready format; and a model limited to chart interpretation may miss precise numerical conditions. This creates a methodological gap that motivates a unified architecture explicitly designed for chart-plus-text-plus-time-series reasoning.

## 2.6 Research Gap and Positioning

The reviewed literature provides several ingredients but does not yet cover the setting studied here. Visual-language models offer strong conversational and image-understanding priors, yet they do not preserve high-precision temporal evidence on their own. Time-series foundation models contribute principled mechanisms for handling continuous signals, and in our case they motivate the extension of prototype-based reprogramming from forecasting to multimodal question answering. Research on trustworthy dialogue further suggests the value of structured outputs that can support downstream verification, escalation, or filtering. Existing power-oriented LLM studies, meanwhile, rarely focus on the concrete service scenario in which explanation quality and reliability must be

handled together. Compared with recent surveys of multimodal time-series analysis and power-oriented large models, our target setting is narrower but operationally more specific: customer-service question answering grounded simultaneously in chart images, textual instructions, and aligned multichannel temporal signals<sup>[19, 21, 22]</sup>.

PowerMTQA is positioned at the intersection of these lines of work. Its main novelty is not merely the coexistence of modalities, but the explicit separation of image-text understanding from high-fidelity temporal encoding, followed by structured fusion and service-oriented auxiliary screening. It combines a visual-language backbone, an explicit temporal stream, and auxiliary service-oriented outputs to address a customer-service setting that is only partially covered by prior research.

## 3 Method

### 3.1 Task Formulation

We formulate intelligent power-service dialogue as a multimodal temporal question-answering task. Each training or inference instance is constructed from a time window aligned across three modalities:

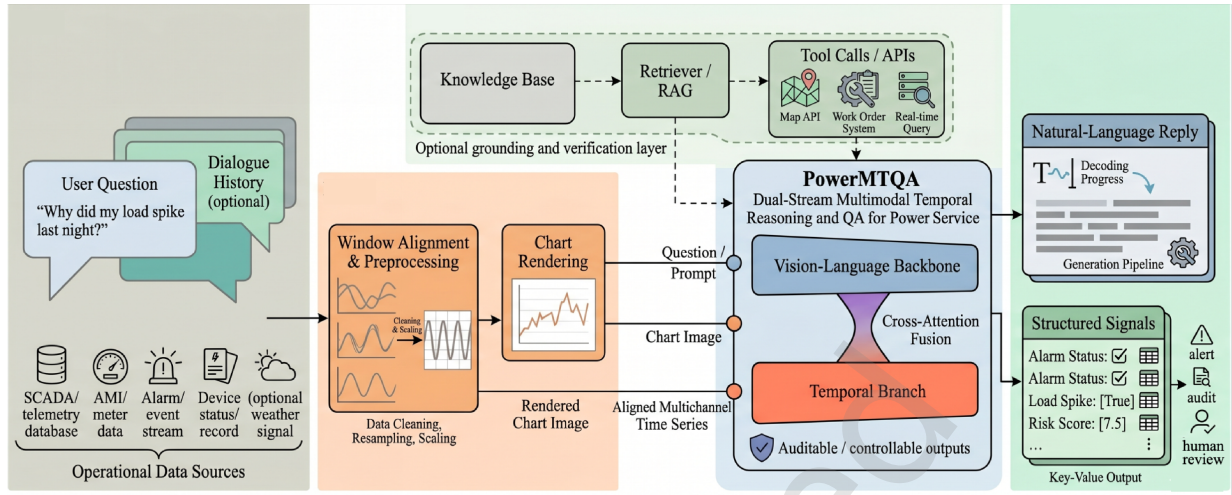
- image modality  $I$ : an offline-rendered chart, such as a single-variable trend plot, a multivariable overlay chart, or a diagnostic figure with threshold annotations;
- text modality  $X$ : the user question together with a prompt prefix containing dataset context, service instruction, statistical summaries, and, when available, lightweight threshold descriptors;
- time-series modality  $S \in \mathbb{R}^{C \times T}$ : one target historical sequence and zero or more support sequences, where  $C$  is the number of channels and  $T$  is the window length.

The model outputs (i) a natural-language answer  $Y$  suitable for customer-service interaction and (ii) optional structured labels  $Z$ , such as `anomaly` and `risk_level`.

From a systems viewpoint, the model sits inside a larger service pipeline, as illustrated in Fig. 1. User questions, chart renderings, and aligned temporal data are passed into the multimodal reasoning core; external retrieval or tool modules may optionally provide additional context such as outage records, ticket status, or knowledge-base snippets. The result is an answer that can be shown to the user, along with structured outputs for internal screening or audit.

### 3.2 PowerMTQA Overview

PowerMTQA is a dual-stream multimodal generation framework that preserves a native visual-language backbone while adding a dedicated time-series reasoning branch. The overall



**Fig. 1** Overall framework of intelligent power-system customer service from a system perspective.

architecture is shown in Fig. 2, which is organized into four modules:

- (1) **Visual-language backbone.** A Qwen-VL-family open-source MLLM encodes the chart image and question text, producing language hidden states  $H_{\text{lang}} \in \mathbb{R}^{L \times D}$  per sample.
- (2) **Multichannel temporal branch.** The raw time series is normalized, patchified, embedded, and reprogrammed into temporal tokens  $H_{\text{time}} \in \mathbb{R}^{N \times D}$  per sample that match the language hidden size.
- (3) **Explicit fusion layer.** Cross-attention injects temporal evidence into the language stream, yielding fused representations that remain compatible with the backbone decoding interface.
- (4) **Output heads.** The backbone `lm_head` generates the final answer, while auxiliary classification heads predict anomaly and risk labels.

This design preserves the original image-text interface of the backbone and avoids forcing the temporal signal into plain-text serialization. It also keeps the temporal components modular, which is useful for ablations, parameter-efficient tuning, and future system integration.

### 3.3 Vision-Language Backbone

Given the image  $I$  and textual input  $X$ , the visual encoder converts the image into latent visual tokens, and the multimodal decoder integrates those tokens with the text sequence under the backbone's native multimodal prompting mechanism. The resulting hidden states are denoted by

$$H_{\text{lang}} \in \mathbb{R}^{B \times L \times D},$$

where  $B$  is the batch size,  $L$  is the sequence length after multimodal tokenization, and  $D$  is the hidden dimension of

the language backbone.

A key design decision is non-intrusive extension. We do not replace the original image encoder, text decoder, or vocabulary projection. Instead, the temporal reasoning capability is added externally, after the backbone has already constructed a strong joint visual-language representation. This strategy improves engineering stability and reduces the risk of damaging the pretrained multimodal alignment.

### 3.4 Temporal Encoding

#### 3.4.1 Instance Normalization

Power-system temporal signals vary across devices, customers, and datasets. Their value ranges may differ significantly, and distribution shift is common. To mitigate this issue, we apply instance-wise normalization to each channel of each sample, following the practical intuition of RevIN and related time-series methods [13, 16]:

$$\hat{x} = \frac{x - \mu}{\sigma + \epsilon},$$

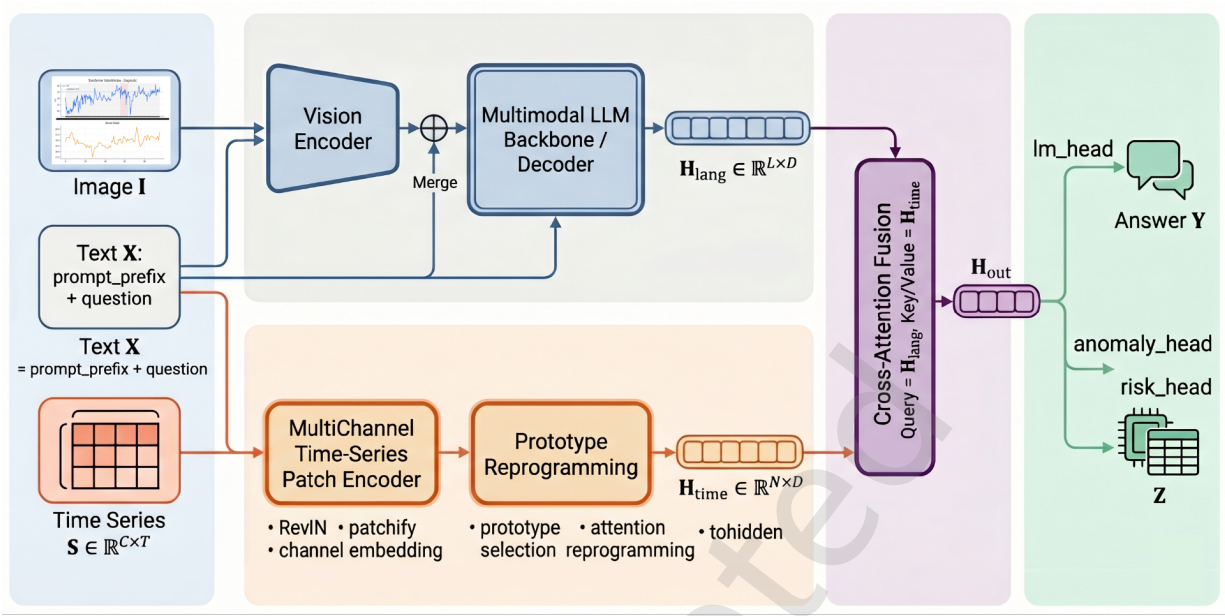
where  $\mu$  and  $\sigma$  are the mean and standard deviation computed over the time dimension of a single channel within a single sample, and  $\epsilon$  is a small constant for numerical stability.

This step improves comparability across samples and stabilizes subsequent temporal embedding.

#### 3.4.2 Patch Embedding

After normalization, the multichannel tensor is segmented into overlapping temporal patches, as shown in Fig. 3. For patch length  $L_p$  and stride  $s$ , the input tensor is transformed from

$$S \in \mathbb{R}^{B \times C \times T}$$



**Fig. 2** Overall architecture of PowerMTQA from a model perspective.

to

$$P \in \mathbb{R}^{B \times C \times N_p \times L_p},$$

where  $N_p$  is the number of patches per channel.

Each patch then passes through LayerNorm + Linear to produce a low-dimensional patch embedding. In addition, we introduce a learnable channel embedding  $e_{\text{channel}}$  so that the model can distinguish the target variable from support variables and preserve channel identity after token flattening.

The result is a temporal token tensor that keeps both local temporal structure and channel attribution. Compared with raw numerical serialization, this representation is more compact, more structured, and better aligned with downstream attention mechanisms.

### 3.5 Prototype Reprogramming

Inspired by Time-LLM<sup>[13]</sup>, we bridge the modality gap between continuous numeric values and the discrete token space through prototype reprogramming, the process depicted in Fig. 4. Let the low-dimensional patch embeddings be the query sequence. We first sample seed vectors from the backbone's input embedding matrix and use a trainable selector to compose a compact prototype set

$$E' \in \mathbb{R}^{M \times D},$$

where  $M$  is the number of prototypes and  $D$  is the hidden dimension of the backbone embedding space. The temporal patch sequence acts as the query stream,

$$Q_{\text{patch}} \in \mathbb{R}^{N \times D_p},$$

where  $D_p$  is the temporal-branch embedding dimension after patch embedding. The patch tokens then attend to these prototypes through multi-head attention (MHA):

$$R = \text{MHA}(Q_{\text{patch}}, E', E').$$

Because the attention module uses the patch sequence as query, the attended output  $R$  remains in the temporal-branch dimension  $D_p$ . It is then projected into the backbone hidden dimension  $D$ , producing temporal hidden states

$$H_{\text{time}} \in \mathbb{R}^{B \times N \times D},$$

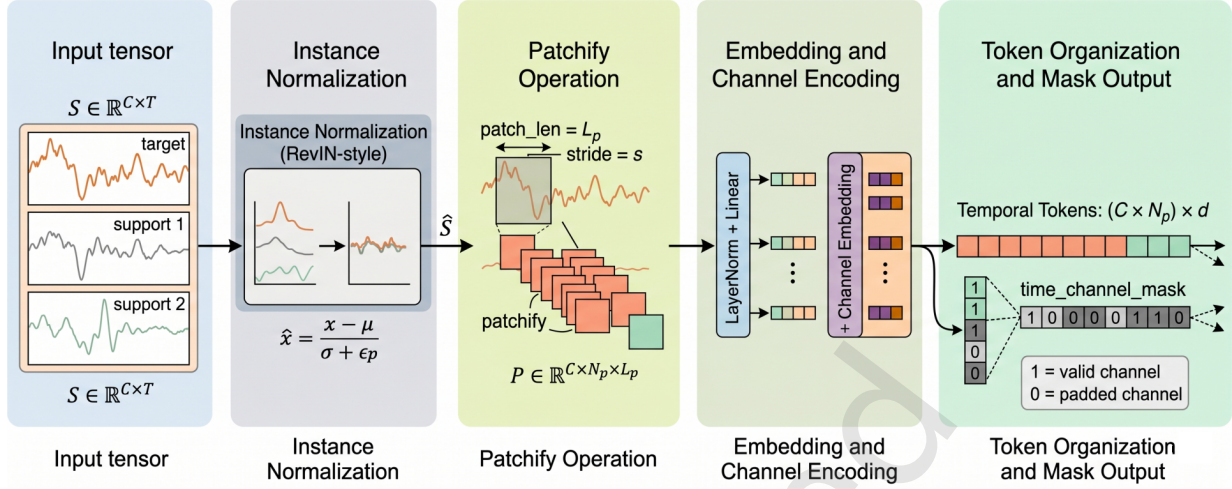
where  $N = C \cdot N_p$  after flattening the channel and patch axes.

This design offers two practical advantages. The temporal branch does not need to imitate the full tokenization behavior of the language model; instead, it aligns with the language space through a compact learned interface. Because the prototypes are derived from the backbone embedding space, the reprogrammed tokens are also more compatible with the backbone's downstream attention dynamics than arbitrary standalone numerical embeddings.

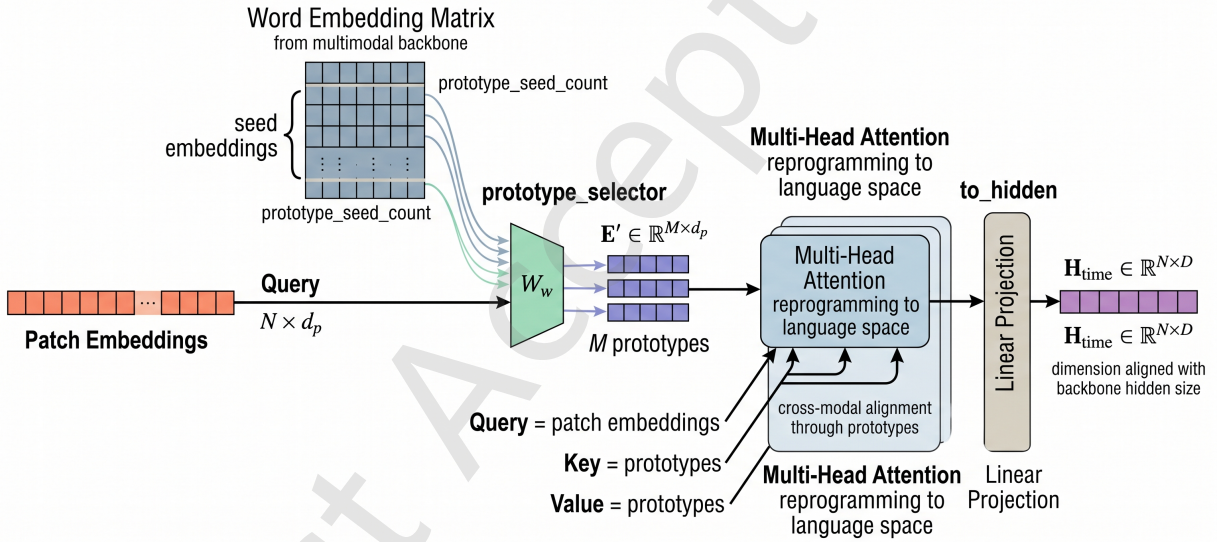
### 3.6 Cross-Attention Fusion

Once the backbone produces  $H_{\text{lang}}$  and the temporal branch produces  $H_{\text{time}}$ , PowerMTQA performs explicit cross-modal fusion, the structure of which is illustrated in Fig. 5. The query stream is first stabilized by layer normalization (LN):

$$H_{\text{attn}} = \text{MHA}(\text{LN}(H_{\text{lang}}), \text{LN}(H_{\text{time}}), \text{LN}(H_{\text{time}})).$$



**Fig. 3** Multichannel temporal patch encoder with normalization, patchification, and channel embedding.



**Fig. 4** Prototype reprogramming from temporal patches to language-compatible hidden tokens.

The attention output is injected through a gated residual update:

$$H_{\text{fused}} = H_{\text{lang}} + \tanh(g) \cdot H_{\text{attn}},$$

where  $g$  is a learnable scalar gate. A feed-forward network (FFN) further refines the fused representation:

$$H_{\text{out}} = H_{\text{fused}} + \text{FFN}(H_{\text{fused}}).$$

This fusion strategy is chosen deliberately over direct token concatenation. Concatenation would enlarge the input context and force the backbone to absorb temporal information implicitly from the start of decoding. In contrast, post-backbone cross-attention allows the pretrained multimodal model to first establish a stable visual-language interpretation and then consult the temporal memory explicitly. This separation improves modularity and makes the contribution

of temporal evidence easier to analyze.

### 3.6.1 Channel Masking

The number of support variables can differ across samples. Therefore, we maintain a `time_channel_mask` and expand it into a token-level `key_padding_mask` during fusion. Tokens originating from padded channels are masked out so that the cross-attention mechanism only attends to valid temporal evidence. This is essential for stable batching and prevents artificial zero-padding channels from contaminating the fused representation.

## 3.7 Prompt Prefix and Sample Construction

Raw time-series values alone rarely provide enough context for a robust service explanation. Following the prompt-prefix philosophy of Time-LLM<sup>[13]</sup>, we prepend a compact

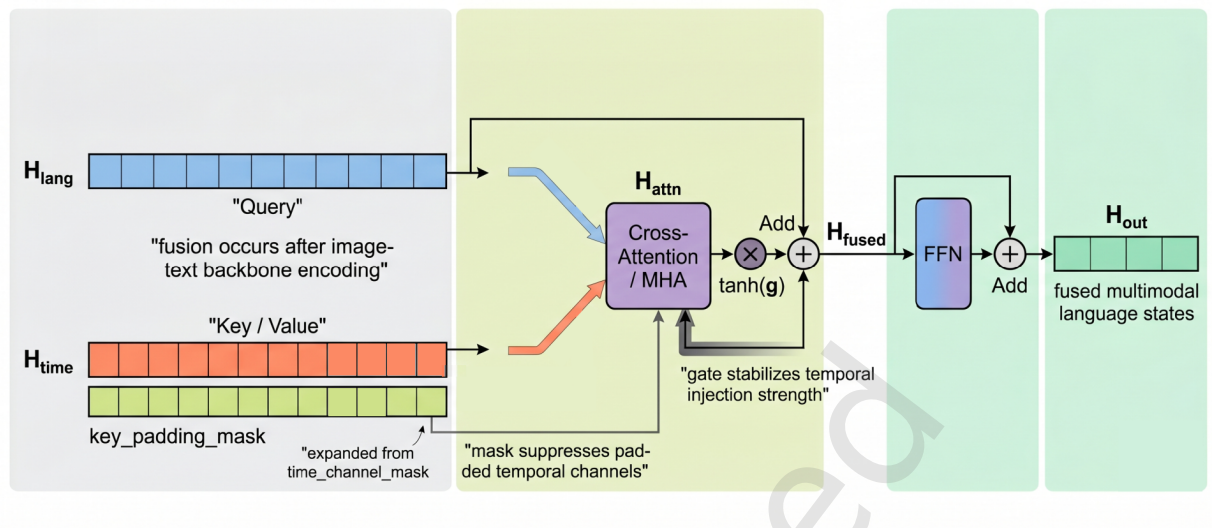


Fig. 5 Cross-attention fusion with gating and key-padding mask for variable channel counts.

textual summary to the user question. The prefix includes three core components:

- **Dataset or scenario context**, such as residential load monitoring or transformer oil temperature analysis;
- **Task instruction**, specifying whether the model should explain a trend, identify abnormality, compare variables, or provide service suggestions;
- **Statistical summary**, including descriptive features explicitly used in the current implementation, namely minimum, maximum, median, trend descriptor, and top-k lags.

When threshold metadata are available, we further serialize lightweight threshold descriptors in the prefix, such as lower or upper operating bounds, margin to the nearest threshold, and recent violation counts. These text summaries do not replace the original time series. Instead, they act as structured hints that reduce the burden on the model to infer all statistics implicitly from raw sequences and provide operating context that may not be recoverable from normalized values alone. This design is particularly important for anomaly- and risk-related supervision, whose semantics may partly depend on absolute operating limits.

To ensure reproducibility and efficient training, chart images are generated offline rather than rendered inside the data loader, following the sample construction pipeline in Fig. 6. Each window is therefore converted into a consistent multimodal training instance containing the image path, prompt prefix, question, answer, raw historical values, support series, and structured labels. The three modalities are synchronized to the same time span so that image cues and numerical evidence correspond exactly.

### 3.8 Training Objectives

#### 3.8.1 Generative Objective

Using teacher forcing, we optimize the next-token cross-entropy (CE) over the assistant answer span:

$$\mathcal{L}_{\text{gen}} = \text{CE}(y_{1:T-1}, \hat{y}_{2:T}),$$

where  $y_{1:T-1}$  denotes the input answer tokens,  $\hat{y}_{2:T}$  denotes the corresponding next-token predictions of the model, and  $T$  is the answer length. Only the answer tokens contribute to this loss; system prompts, user messages, padding tokens, and non-target regions are masked out.

#### 3.8.2 Auxiliary Objectives

To expose service-critical judgments explicitly, we add two auxiliary classification heads on top of the generative objective, forming the joint training scheme in Fig. 7:

- **anomaly\_head** for binary abnormality screening;
- **risk\_head** for multi-class risk-level screening.

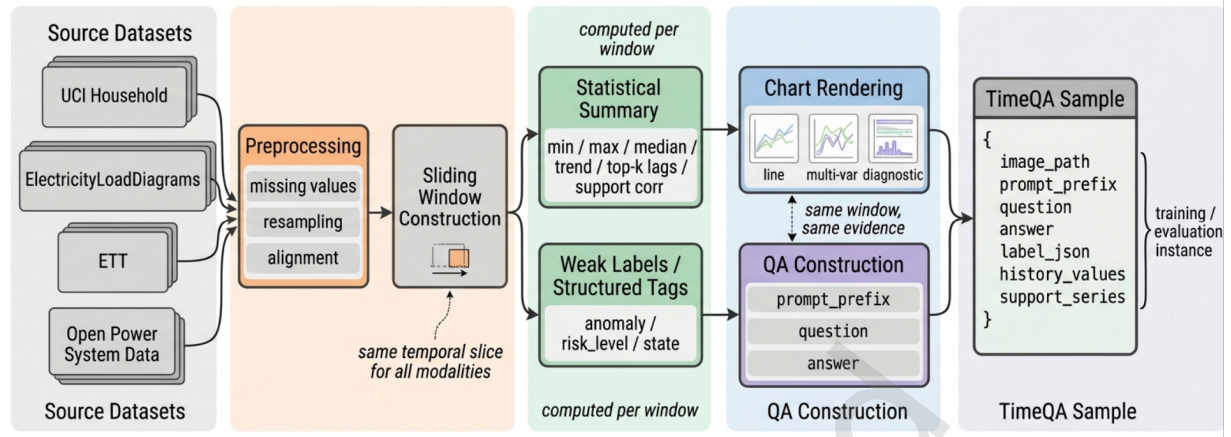
In the current implementation, these heads operate on a pooled representation obtained by attention-mask-weighted averaging over the fused hidden states. The pooled vector is then fed into two linear classifiers for anomaly and risk screening, respectively. The auxiliary loss is defined as

$$\mathcal{L}_{\text{aux}} = \frac{1}{2}(\mathcal{L}_{\text{anom}} + \mathcal{L}_{\text{risk}}),$$

where  $\mathcal{L}_{\text{anom}}$  and  $\mathcal{L}_{\text{risk}}$  are the cross-entropy losses of the anomaly and risk classification heads, respectively. The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{gen}} + \lambda \mathcal{L}_{\text{aux}},$$

where  $\lambda$  is a balancing weight that controls the contribution of the auxiliary supervision relative to generative training.



**Fig. 6** Multimodal sample construction pipeline from raw datasets to aligned multimodal question-answering instances.

This formulation is well suited to power-service scenarios. The generative term encourages user-friendly explanations, while the auxiliary heads encourage the model to internalize and expose categorical decisions that are useful for system-side validation. These outputs are intended as coarse, system-facing screening signals rather than standalone protection decisions. Their predictions are conditioned on the joint visual, temporal, and textual evidence and, when available, explicit threshold descriptors. In deployment, they can be used to trigger escalation rules, warning banners, or confidence-dependent post-processing without claiming that the model can infer all device-specific alarm rules solely from generic multimodal context. Practically, this design aims to reduce three common failure modes: fluent but trend-only answers that miss localized spikes or threshold violations, explanations that describe the curve correctly but understate operational severity, and plausible narratives whose wording is inconsistent with system-side screening logic. When the generated answer and the auxiliary predictions disagree, the structured heads provide a natural basis for conservative templates, warning prompts, or human escalation.

### 3.9 Parameter-Efficient Fine-Tuning

To reduce training cost and preserve pretrained capability, we adopt parameter-efficient fine-tuning rather than full-model updating. In the current configuration, low-rank adaptation (LoRA)<sup>[24]</sup> is attached to the attention projection layers of the multimodal backbone, specifically  $q\_proj$ ,  $k\_proj$ ,  $v\_proj$ , and  $o\_proj$ . The temporal branch, fusion layer, and auxiliary heads are trained directly.

Several implementation choices are important in practice. The temporal branch is intentionally lightweight compared with the backbone so that the additional reasoning path remains economical. Fusion occurs after the backbone's visual-

language encoding stage, which leaves the original multimodal processor and decoding interface largely unchanged. Offline chart generation also improves experimental reproducibility and avoids a training bottleneck caused by repeated plotting.

### 3.10 Design Rationale

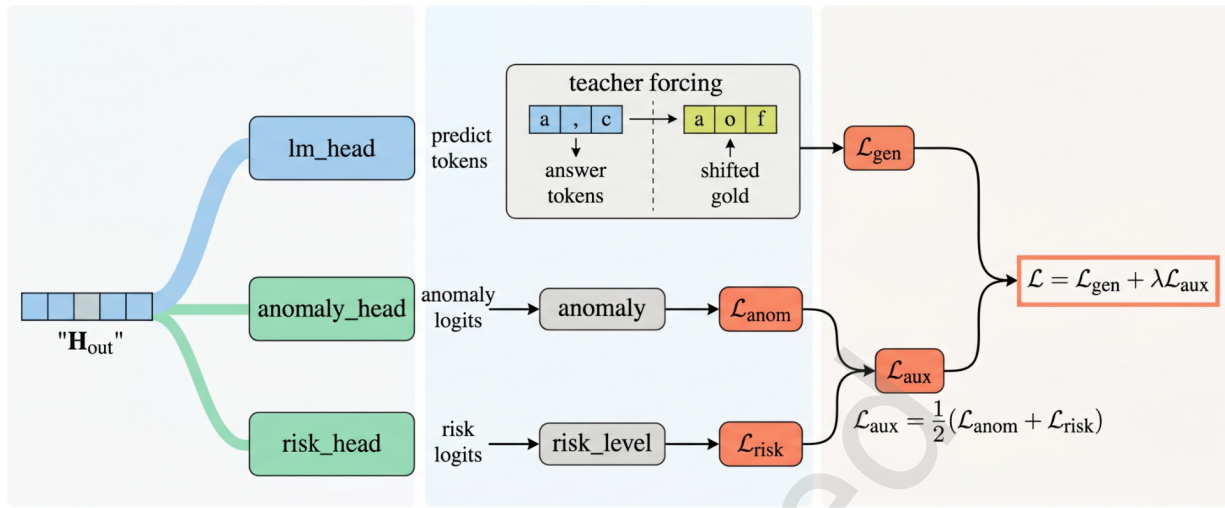
The proposed architecture is motivated by the specific demands of power-system customer service. A pure visual-language backbone is strong at interpreting the rendered chart and the user's wording but weak at exact temporal computation. A pure time-series model is strong at numerical pattern extraction but weak at conversational explanation. PowerMTQA combines both strengths without collapsing one modality into the other.

More importantly, the model is designed to remain deployable. Its auxiliary outputs support auditing, its modular structure supports ablation and maintenance, and its backbone compatibility supports reuse of strong open-source multimodal models. As a result, PowerMTQA serves not only as a research architecture for multimodal temporal reasoning, but also as a practical blueprint for trustworthy power-service assistants.

## 4 Experiments

### 4.1 Experimental Objectives

This section evaluates whether the proposed dual-stream design improves multimodal temporal question answering for power-service scenarios. The experiments focus on three issues: whether an explicit time-series branch yields clear gains over a vision-only baseline, whether multichannel temporal modeling, parameter-efficient tuning, and auxiliary supervision contribute incremental benefits, and whether the re-



**Fig. 7** Joint training objective with generative loss and auxiliary anomaly/risk supervision.

sulting answers are more semantically consistent with the ground-truth answers under third-party judgment.

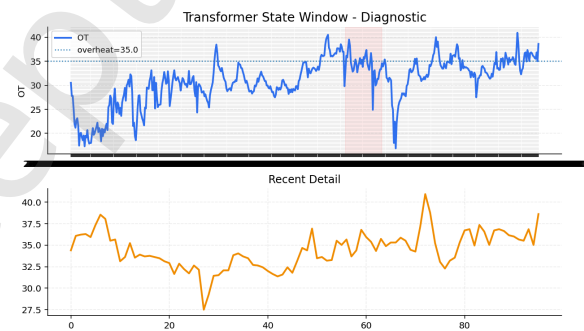
At the present stage, the experiments should still be interpreted as model-side validation rather than the final end-to-end service evaluation. Answer-level semantic consistency evaluation has been introduced for the most representative settings, but a fully exhaustive third-party assessment across all ablation variants has not yet been completed.

## 4.2 Dataset and Split

The experiments use the unified PowerMTQA dataset constructed from four public electricity and power-related time-series sources: University of California, Irvine (UCI) Individual Household Electric Power Consumption<sup>[25]</sup>, UCI ElectricityLoadDiagrams20112014<sup>[26]</sup>, Electricity Transformer Temperature (ETT)<sup>[10]</sup>, and Open Power System Data<sup>[27]</sup>. Each sample is organized as an aligned multimodal instance containing:

- an offline-rendered chart image;
- a text prompt composed of task instruction, dataset context, compact statistical summary, and, when available, lightweight threshold descriptors;
- a target time series together with optional support channels;
- a natural-language answer and structured weak labels.

Because more than one question answering (QA) pair can in principle be derived from a single temporal window, the window rather than the individual QA pair is the natural unit for reporting dataset scale and for reasoning about split integrity. For this reason, Table 1 reports both the number of windows and the number of QA samples. At the same time, the present dataset remains relatively small, and a larger



**Fig. 8** Example of an offline-rendered chart from the ETT subset used as the visual input to PowerMTQA.

**Table 1** Dataset split used in the current experiments.

| Split      | Windows | QA Samples |
|------------|---------|------------|
| Train      | 24      | 240        |
| Validation | 10      | 100        |
| Test       | 10      | 100        |

source-balanced benchmark with finer-grained source-wise accounting is still needed for a more comprehensive evaluation.

Fig. 8 shows an example of the offline-rendered chart image used in the multimodal sample construction pipeline. This example comes from the ETT subset and illustrates the visual evidence provided to the model together with the aligned textual and temporal inputs.

For the current multichannel experiment set, the dataset summary is reported in Table 1. All experiments reported in this section use the same validation split of 100 QA samples. The test split is reserved for later large-scale evaluation.

### 4.3 Implementation Details

The final model studied in this paper is PowerMTQA, which is built on the local Qwen3.5-2B multimodal backbone while keeping the native visual branch unchanged. Its temporal branch uses patch length 16, patch stride 8, patch embedding dimension 16, 64 prototypes, 2048 prototype seed candidates, 8 temporal attention heads, 8 fusion heads, and a maximum of 16 channels.

Unless otherwise specified in the ablation subsection, the main optimization settings of PowerMTQA are as follows:

- learning rate  $5 \times 10^{-5}$ ;
- weight decay 0.01;
- batch size and evaluation batch size of 1;
- 20 training epochs for all reported experiments;
- LoRA with rank 8, alpha 16, dropout 0.05, and target group attn;
- auxiliary-loss weight 0.5.

The images are rendered offline before training, so the data loader only reads aligned multimodal samples rather than plotting figures on the fly. This design improves reproducibility and avoids unnecessary visualization overhead during optimization.

### 4.4 Evaluation Protocol

The currently logged automatic metrics are mainly used to diagnose optimization behavior, including validation loss, token-level accuracy, and perplexity. For the full PowerMTQA model, we further distinguish total validation loss, generation validation loss, and auxiliary validation loss under the multitask objective.

These metrics are useful for comparing convergence behavior, but they are not the final answer-quality criterion of this work. The final evaluation should emphasize semantic consistency between the generated answers and the ground-truth answers. For all compared settings, the current validation predictions have already been exported on the same 100-sample validation split and can be used for third-party scoring, pairwise preference judgment, or rubric-based answer grading. In the present revision, the GLM-4.7 result should therefore be interpreted as a single-judge external consistency check rather than a fully standardized benchmark. A more exhaustive protocol with released judge prompts, repeated trials, and cross-judge comparison remains an important direction for future evaluation.

### 4.5 Training Dynamics

Fig. 9 reports the 20-epoch training trajectories used in the current experiments. This horizon is sufficient to capture

both the early convergence behavior of the simpler ablation settings and the optimization fluctuations that emerge in the full PowerMTQA configuration. Presenting the complete 20-epoch runs therefore provides a direct and consistent view of the training dynamics across all compared models.

For the simpler ablation settings, the plotted curves correspond to the standard generative training objective. For the full PowerMTQA model, the figure separately visualizes total loss, generation loss, and auxiliary loss because the multitask objective is not directly comparable to the single-objective settings.

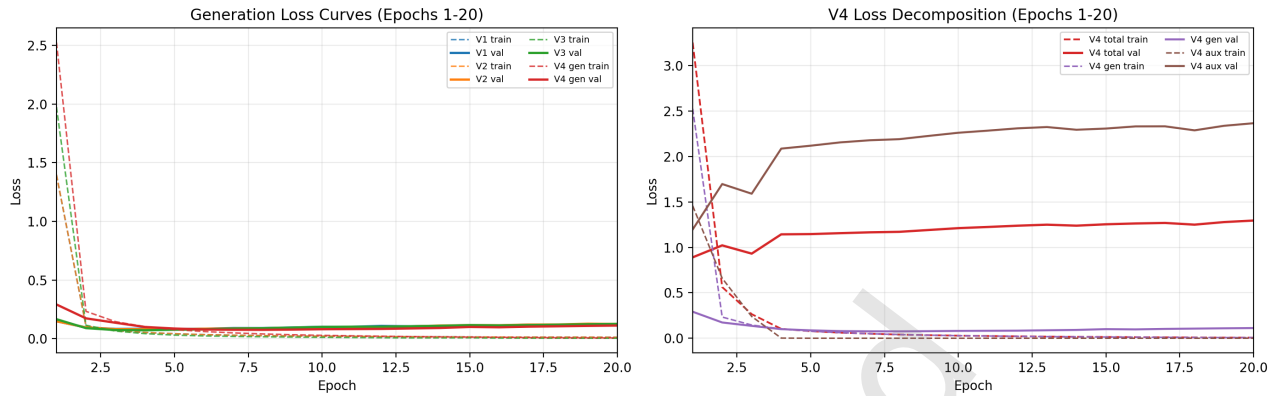
The 20-epoch curves support three preliminary observations. The vision-only baseline is clearly much weaker than all dual-stream settings, which reinforces the need for explicit temporal input. The dual-stream ablations also converge quickly in the early epochs, with their best validation losses appearing near epochs 4–5, indicating that the main performance gains are obtained very early in training. PowerMTQA achieves a competitive generation loss in the early stage, but its multitask optimization is visibly less stable than that of the simpler ablation settings. Within the 20-epoch training window, this instability remains evident in the later phase, which suggests that the current PowerMTQA configuration likely suffers from an overly aggressive optimization trajectory and should be revisited through learning-rate control, gradient clipping, or auxiliary-loss rebalancing.

### 4.6 Main Validation Results

Table 2 reports the currently available validation results. Because PowerMTQA is optimized with a multitask objective, the value reported in the *Validation Generative Loss* column for PowerMTQA is the best generation validation loss (`val_gen_loss`) rather than the total validation loss. This makes the QA-oriented comparison more interpretable relative to the simpler baselines and ablations.

The current results support three immediate conclusions. The gap between the vision-only baseline and the dual-stream single-channel setting is very large, showing that chart-only reasoning is insufficient for this task. Moving from single-channel temporal input to multichannel modeling yields a small but consistent improvement. Adding LoRA gives the best QA-oriented validation result among the simpler ablation runs. PowerMTQA remains competitive on the generation objective, but its multitask setting introduces an additional optimization trade-off and already exhibits clear late-stage instability within the 20-epoch runs.

Training Curves of PowerMTQA Variants (First 20 Epochs)



**Fig. 9** Training loss curves over the first 20 epochs. The left panel compares the generative loss trajectories of the baseline and ablation settings during the stable optimization stage, while the right panel decomposes the full PowerMTQA objective into total, generation, and auxiliary losses in the same interval.

**Table 2** Main validation results across the baseline, ablation settings, and the full PowerMTQA model.

| Model Setting                     | Temporal Branch | Multi-channel | LoRA | Auxiliary Heads | Selected Epoch | Validation Generative Loss | Token Accuracy |
|-----------------------------------|-----------------|---------------|------|-----------------|----------------|----------------------------|----------------|
| Vision-only baseline              | ✗               | ✗             | ✗    | ✗               | —              | 24.5611                    | 0.0290         |
| Dual-stream, single-channel       | ✓               | ✗             | ✗    | ✗               | 5              | 0.0809                     | 0.9774         |
| Dual-stream, multi-channel        | ✓               | ✓             | ✗    | ✗               | 5              | 0.0798                     | 0.9777         |
| Dual-stream, multi-channel + LoRA | ✓               | ✓             | ✓    | ✗               | 4              | 0.0721                     | 0.9788         |
| PowerMTQA                         | ✓               | ✓             | ✓    | ✓               | 8              | 0.0758*                    | 0.9785         |

\*For PowerMTQA, the reported generative loss corresponds to the best `val_gen_loss` rather than the total validation loss, because the total loss additionally includes the auxiliary classification term.

#### 4.7 Ablation Study

To support ablation analysis, we organize the study into five comparable settings:

- **Vision-only baseline:** the model receives only the rendered chart and question text, without the explicit time-series branch.
- **Dual-stream, single-channel:** the visual-language backbone is augmented with a temporal branch that uses only the target series.
- **Dual-stream, multi-channel:** the temporal branch is extended to use the target series together with support variables and channel-aware encoding.
- **Dual-stream, multi-channel + LoRA:** the multi-channel dual-stream model is further enhanced with parameter-efficient tuning on the backbone attention projections.
- **PowerMTQA:** the full model adds auxiliary anomaly and risk supervision on top of the multi-channel LoRA-enabled dual-stream architecture.

Based on this organization, Table 3 summarizes the incremental observations obtained from the current ablation runs.

For completeness, Table 4 reports the currently available PowerMTQA multitask validation statistics under two views:

the checkpoint with the best total loss and the checkpoint with the best generation loss.

These results indicate that the current PowerMTQA configuration successfully preserves QA generation quality in the early training stage, but the multitask objective still introduces a clear optimization trade-off. In particular, the later epochs exhibit noticeable instability, and the auxiliary loss remains comparatively large under the best generation-oriented checkpoint. Therefore, the current PowerMTQA training setup should be treated as a promising but still incomplete multitask configuration rather than the final optimized form of the model.

#### 4.8 Answer-Level Semantic Consistency Evaluation

Beyond token-level metrics, we further examine whether the generated answers are semantically consistent with the ground-truth answers. To directly test whether the proposed dual-stream design is better than a naive time-series textualization strategy, this answer-level assessment includes three representative models: the Qwen3.5-2B visual-only baseline, a Qwen3.5-2B text-only baseline that serializes the time series into the prompt, and the full PowerMTQA model. The evaluation is conducted on the same 100-sample validation split, where each sample is paired with a human

**Table 3** Incremental observations from the ablation study.

| Step                                                   | Incremental Change                                                   | Observation                                                                                                                                   |
|--------------------------------------------------------|----------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| Baseline → single-channel dual-stream                  | Add an explicit temporal branch to the visual-language backbone      | Validation loss drops sharply, indicating that direct temporal evidence is essential.                                                         |
| Single-channel dual-stream → multi-channel dual-stream | Replace single-channel modeling with multichannel temporal encoding  | Validation loss improves slightly, suggesting that support variables provide useful complementary evidence.                                   |
| Multi-channel LoRA                                     | Add parameter-efficient tuning on the backbone attention projections | The best QA-oriented validation result is achieved in the current completed ablation runs.                                                    |
| Multi-channel + LoRA → PowerMTQA                       | Add auxiliary anomaly/risk supervision                               | Generation quality remains strong in the early stage, but multitask optimization becomes noticeably less stable and later shows loss rebound. |

**Table 4** PowerMTQA multitask validation statistics under two checkpoint-selection views.

| Checkpoint View      | Epoch | Total Val Loss | Gen Val Loss | Aux Val Loss | Token Accuracy |
|----------------------|-------|----------------|--------------|--------------|----------------|
| Best total loss      | 1     | 0.8902         | 0.2918       | 1.1967       | 0.9424         |
| Best generation loss | 8     | 1.1709         | 0.0758       | 2.1903       | 0.9785         |

**Table 5** Answer-level semantic consistency accuracy under GLM-4.7 evaluation.

| Model                  | Accuracy |
|------------------------|----------|
| Qwen3.5-2B visual-only | 0.67     |
| Qwen3.5-2B text-only   | 0.43     |
| PowerMTQA              | 0.82     |

customer-service reference answer accumulated in the historical customer-service-style QA dataset and treated as the standard answer label. We used GLM-4.7 as an external judge to compare the model output against the corresponding human reference answer. Instead of exposing the full judge prompt, we summarize the released evaluation rule at the criterion level: the judge checks whether the generated answer is (i) semantically consistent with the human reference answer, (ii) correct in temporal data analysis, and (iii) compliant with customer-service tone and response norms. A prediction is counted as correct only when all three criteria are satisfied simultaneously; if any one criterion fails, the sample is marked as incorrect under an AND-style decision rule.

The aggregate results are summarized in Table 5.

The answer-level evaluation now supports two complementary conclusions. First, PowerMTQA outperforms the visual-only baseline by 15 percentage points, confirming that explicit temporal evidence is necessary for reliable power-service question answering. Second, PowerMTQA outperforms the text-only baseline by 39 percentage points, showing that the gain does not come merely from exposing the raw numeric sequence to the language model. Instead, the result supports the original motivation of the method: directly serializing time series into prompt text is an inefficient and lossy interface, whereas the dedicated temporal branch

preserves numerical structure much more effectively. Interestingly, the text-only baseline also underperforms the visual-only baseline in this setting, which further suggests that naive textualization alone is not a strong substitute for either chart-based grounding or structured temporal encoding. Because the judgment rule is intentionally strict, the reported accuracy should be interpreted as a conservative answer-level reliability indicator rather than a lenient semantic-similarity score.

To further illustrate the qualitative difference between the two models, we examine one representative validation example. In this sample, the ground-truth answer emphasizes an overall upward trend together with the approximate maximum and minimum values. The visual-only baseline incorrectly interprets the curve as staying near zero with almost no fluctuation, whereas PowerMTQA correctly identifies the upward trend and recovers the key numerical range. Although the PowerMTQA answer is still shorter than the ground-truth answer and does not mention the localized high-load interval, it preserves the main factual content much better than the baseline.

**Ground-truth answer:** “The window shows an overall upward trend, with a maximum of about 4.764 and a minimum of about 0.392. High-load activity is mainly concentrated around steps 425–426.”

**Qwen3.5-2B visual-only baseline:** “From the figure, the household’s `Global_active_power` remains near zero over the 0–512 time-step range with almost no fluctuation. This suggests that the household consumed almost no electricity during

this period and exhibited a very stable, very low load pattern.”

**PowerMTQA:** “The household shows an overall upward trend, with a maximum load of about 4.764 and a minimum of about 0.392.”

This case suggests that the advantage of PowerMTQA is not merely stylistic refinement. Instead, the explicit temporal branch helps the model anchor its response to trend direction and quantitative magnitude, which are precisely the aspects most vulnerable to distortion under image-only reasoning.

## 5 Discussion

### 5.1 Main Findings

The current experimental results suggest that the main benefit of PowerMTQA comes from introducing an explicit temporal reasoning path into a multimodal question-answering framework. Compared with the visual-only baseline, the dual-stream settings achieve substantially better validation behavior, and the answer-level semantic evaluation further shows that the full PowerMTQA model produces answers that are more consistent with the ground-truth answers. This pattern indicates that chart images alone are insufficient for reliable power-service question answering when the task depends on precise trend direction, magnitude estimation, and localized temporal evidence.

Another important observation is that the gains are not limited to stylistic improvement. The representative case analysis shows that the visual-only baseline can produce fluent but factually distorted responses, whereas PowerMTQA is better able to recover key numerical attributes such as overall trend and approximate extrema. This is consistent with the intended design of the model: the visual branch preserves chart understanding, while the temporal branch injects higher-fidelity numerical evidence into answer generation.

### 5.2 Design Implications

The ablation results point to several practical implications for multimodal temporal reasoning. The largest gain comes from adding an explicit temporal branch, which narrows the gap between visual impression and numerical evidence. Multi-channel temporal encoding brings a further, though smaller, improvement, suggesting that support variables are useful when they are aligned with the target series and properly encoded. LoRA also remains effective in this setting and improves answer quality without requiring full-model fine-tuning.

At the same time, the current results also show that extending the model from a strong generative configuration to a multitask configuration is not automatically beneficial. The full PowerMTQA model remains competitive on the generation-oriented metric and achieves the best answer-level semantic consistency among the evaluated settings, but its multitask formulation still introduces additional optimization difficulty. This suggests that multitask learning is promising in principle, yet still sensitive to loss balancing and optimization stability in practice.

### 5.3 Limitations

Several limitations should be considered when interpreting the present study. Due to cost constraints, the current experiments do not include larger-parameter models. Although the present comparison is sufficient to validate the effectiveness of PowerMTQA relative to the selected baseline, it does not yet establish how the proposed method scales against stronger multimodal backbones with substantially larger capacity.

The current dataset construction also relies heavily on open-source power datasets, and much of the sample generation pipeline is automated. This design is suitable for early-stage research and controlled experimentation, but it may not fully capture the diversity, ambiguity, and business specificity of real-world power-service scenarios. Broader practical deployment will therefore require datasets supported by more extensive human expert involvement and covering a wider range of business domains, operational contexts, and service tasks. It may also inherit bias from the source datasets, prompt templates, and weak-label generation rules, which can affect both the style and the difficulty distribution of the resulting QA pairs.

Another limitation concerns threshold grounding. Explicit equipment-specific thresholds are not uniformly available across all public datasets used in this study, and their operational semantics may vary across devices and scenarios. Therefore, the current anomaly- and risk-related outputs should be interpreted as coarse screening signals rather than fully threshold-grounded protection judgments. A more complete treatment of device-specific anomaly diagnosis will require richer threshold metadata and stronger expert annotation in future dataset construction. A related limitation is split transparency: when multiple QA pairs are generated from overlapping windows, the split protocol should be documented at the window level to avoid hidden leakage risks. The present revision clarifies this issue conceptually, but a larger public release with finer-grained overlap auditing is still necessary.

## 5.4 Future Works

The next stage of this work should expand in three directions. One priority is evaluation on stronger multimodal backbones with larger parameter scales so that the scalability of PowerMTQA can be assessed more rigorously. Such experiments are important for determining whether the observed gains stem only from the current backbone setting or remain valid when transferred to models with substantially stronger pre-trained capabilities.

Another priority is dataset construction with deeper human expert participation. Although the current pipeline is effective for early-stage experimentation, broader application in power-service scenarios will require expert-validated datasets that cover a wider range of business domains, operational conditions, service intents, and industry-specific reasoning patterns. This extension should also include more explicit source-wise statistics, tighter split auditing, and dedicated bias analysis across datasets and question types.

The framework should also be integrated more tightly with retrieval, knowledge grounding, and tool invocation so that multimodal temporal reasoning can be combined with external factual support in a deployable power-service assistant.

The present results support the central claim of this study: multimodal power-service question answering benefits substantially from combining chart understanding with explicit temporal reasoning. While the current implementation is not yet the final optimized form of the model, it already demonstrates that PowerMTQA provides a practical and effective direction for trustworthy temporal-aware customer-service intelligence.

## Reference

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. [Online]. Available: <https://arxiv.org/abs/1706.03762>
- [2] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, Learning transferable visual models from natural language supervision, in *International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.
- [3] J. 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, Y. Du, J. Wulff, K. Simonyan, S. M. A. Eslami, A. Zisserman *et al.*, Flamingo: a visual language model for few-shot learning, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. [Online]. Available: <https://arxiv.org/abs/2204.14198>
- [4] J. Li, D. Li, S. Savarese, and S. Hoi, Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, in *International Conference on Machine Learning (ICML)*. PMLR, 2023, pp. 19 730–19 742.
- [5] H. Liu, C. Li, Q. Wu, and Y. J. Lee, Visual instruction tuning, *Advances in neural information processing systems*, vol. 36, pp. 34 892–34 916, 2023.
- [6] S. Bai *et al.*, Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, *arXiv*, 2023. [Online]. Available: <https://arxiv.org/abs/2308.12966>
- [7] —, Qwen2.5-vl technical report, *arXiv*, 2025. [Online]. Available: <https://arxiv.org/abs/2502.13923>
- [8] —, Qwen3-vl technical report, *arXiv*, 2025. [Online]. Available: <https://arxiv.org/abs/2511.21631>
- [9] 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 (ICLR)*, 2023. [Online]. Available: <https://arxiv.org/abs/2211.14730>
- [10] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, Informer: Beyond efficient transformer for long sequence time-series forecasting, in *AAAI Conference on Artificial Intelligence (AAAI)*, 2021. [Online]. Available: <https://arxiv.org/abs/2012.07436>
- [11] H. Wu, J. Xu, J. Wang, and M. Long, Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting, in *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. [Online]. Available: <https://arxiv.org/abs/2106.13008>
- [12] 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 (ICLR)*, 2023. [Online]. Available: <https://arxiv.org/abs/2210.02186>
- [13] M. Jin, S. Wang, L. Ma, Z. Chu, J. Zhang, X. Shi, P.-Y. Chen, Y. Liang, Y.-F. Li, S. Pan *et al.*, Time-llm: Time series forecasting by reprogramming large language models, in *International conference on learning representations*, vol. 2024, 2024, pp. 23 857–23 880.
- [14] A. F. Ansari *et al.*, Chronos: Learning the language of time series, *arXiv*, 2024. [Online]. Available: <https://arxiv.org/abs/2403.07815>
- [15] V. Ekambaram *et al.*, Lag-llama: Towards foundation models for probabilistic time series forecasting, *arXiv*, 2023. [Online]. Available: <https://arxiv.org/abs/2310.08278>
- [16] T. Kim, J. Kim, Y. Tae, C. Park, J. Choi, and J. Choo, Reversible instance normalization for accurate time-series forecasting against distribution shift, in *International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: <https://openreview.net/forum?id=cGDAkQo1C0p>
- [17] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. Yih, T. Rocktäschel,

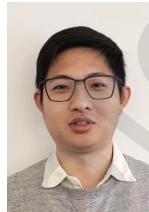
- S. Riedel, and D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, *arXiv*, 2020. [Online]. Available: <https://arxiv.org/abs/2005.11401>
- [18] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, ReAct: Synergizing reasoning and acting in language models, in *International Conference on Learning Representations (ICLR)*, 2023. [Online]. Available: <https://arxiv.org/abs/2210.03629>
- [19] H. Liu, H. Kamarthi, Z. Zhao, S. Xu, S. Wang, Q. Wen, T. Hartvigsen, F. Wang, and B. A. Prakash, How can time series analysis benefit from multiple modalities? a survey and outlook, *arXiv*, 2025. [Online]. Available: <https://arxiv.org/abs/2503.11835>
- [20] A. Masry, X. Do, A. Tan, S. Joty *et al.*, Chartqa: A benchmark for question answering about charts with visual and logical reasoning, in *Findings of the Association for Computational Linguistics: ACL*, 2022. [Online]. Available: <https://aclanthology.org/2022.findings-acl.177/>
- [21] Q. Yao, F. Fang, Y. Chen, J. Liu, H. Mo, and Y. Ao, Ai large models for power system: A survey and outlook, *IET Smart Energy Systems*, vol. 1, no. 1, pp. 3–21, 2025. [Online]. Available: <https://doi.org/10.1049/ses2.70000>
- [22] F. Amjad, T. Korotko, and A. Rosin, Review of LLMs applications in electrical power and energy systems, *IEEE Access*, vol. 13, pp. 150 951–150 969, 2025. [Online]. Available: <https://doi.org/10.1109/ACCESS.2025.3599922>
- [23] Y. Xiang, L. Tan, G. Qiu, Z. Tang, and J. Liu, Large language model for secure operation of power systems, *Smart Energy System Research*, vol. 1, no. 1, p. 10005, 2025.
- [24] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, Lora: Low-rank adaptation of large language models, *arXiv*, 2021. [Online]. Available: <https://arxiv.org/abs/2106.09685>
- [25] G. Hebrail and A. Berard, Individual household electric power consumption, UCI Machine Learning Repository (Dataset), 2012. [Online]. Available: <https://doi.org/10.24432/C58K54>
- [26] Electricityloaddiagrams20112014, UCI Machine Learning Repository (Dataset), 2015. [Online]. Available: <https://archive.ics.uci.edu/dataset/321/electricityloaddiagrams20112014>
- [27] F. Wiese, I. Schlecht, W. Bunke, C. Gerbaulet, L. Hirth, M. Jahn, F. Kunz, C. Lorenz, B. Müller, J. Reimann, and A. Zerrahn, Open power system data – frictionless data for electricity system modelling, *Applied Energy*, vol. 236, pp. 401–409, 2019. [Online]. Available: <https://arxiv.org/abs/1812.10405>



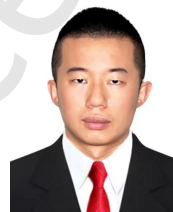
**Guorui Wang** Guorui Wang is currently a Ph.D. candidate at the Economics and Management School of Wuhan University. His research interests focus on marketing management and enterprise digital transformation.



**Jiang Jiang** Jiang Jiang is a staff member at the Enterprise Architecture and Digitalization Department of Guangdong Power Grid Corporation. His research interests mainly focus on enterprise digital transformation and large language models.



**Jinchen Ma** Jinchen Ma serves as a Product Manager at Zhipu AI. He received his master's degree from the University of Duisburg-Essen (UDE), Germany. His work focuses on bridging advanced AI technologies and practical product applications, with particular interests in large language models and intelligent product innovation.



**Dingxiao Liu** Dr. Dingxiao Liu serves as Vice President at ZhipuAI. A recognized expert in cognitive intelligence, he focuses on the frontier of LLMs and knowledge graphs. Dr. Liu has been instrumental in building specialized AI models for key sectors such as energy, industrial intelligence, and IoT. By bridging the gap between foundational research and practical application, he has pioneered the implementation of advanced AI architectures that redefine efficiency in large-scale enterprise environments.

## Author biography