Encoder-Free ECG-Language Models

Anonymous ACL submission

Abstract

ECG–Language Models (ELMs) extend recent progress in Multimodal Large Language Models (MLLMs) to automated ECG interpretation. However, most ELMs follow Vision–Language Model (VLM) designs and depend on pretrained ECG encoders, adding architectural and training complexity. Inspired by encoder-free VLMs, we introduce ELF, an encoder-free ELM that replaces the ECG encoder with a single projection layer trained jointly with the LLM. Across five datasets, ELF matches or exceeds state-of-the-art ELMs that use far more complex encoders and training pipelines. We also test whether adding architectural biases to ELF improves performance and find that the single linear projection remains competitive. Finally, we show that ELF, and potentially other ELMs, often rely more on benchmark artifacts and language priors than ECG-derived information, highlighting limitations in current evaluation practices and ELM design. We will open-source all code and data upon acceptance.

1 Introduction

The global volume of over 300 million ECGs recorded annually (Zhu et al., 2020), combined with the expertise required for accurate interpretation and the growing shortage of physicians (AAMC, 2024), has created demand for automated ECG analysis. Recently, the automation of ECG interpretation has advanced from classification-based models (Rajpurkar et al., 2017) to ECG–Language Models (ELMs) (Zhao et al., 2024).

Architecturally, ECG–Language Models (ELMs) are analogous to Vision–Language Models (VLMs) and generate textual responses from queries composed of text or ECG data (Han et al., 2025). ELMs differ from VLMs in that, while images are commonly represented in RGB format for VLMs, ECG data can take multiple forms: raw signal values (Wan et al., 2024), RGB plot images (Liu et al., 2024b), or symbolic representations (Han et al., 2024) (See Appendix A). Each representation requires distinct architectural modifications and corresponding training regimes.

In VLMs, RGB images are typically processed by a large neural network (vision encoder), which is first pretrained on internet-scale image datasets with its own learning objective (Bai et al., 2025). The pretrained encoder is then frozen and paired with a projection layer that maps its output to a latent representation compatible with the Large Language Model (LLM) (Liu et al., 2023). Similarly, many ELM architectures include an ECG encoder that is separately pretrained on large collections of ECG data (Li et al., 2025a). As one might expect, developing an effective encoder, whether for ECG or vision data, is a complex process that requires careful feature engineering and tuning of the architecture, learning objectives, and training procedures.

To address these complexities, Fuyu-8B (Bavishi et al., 2023), an encoder-free VLM, was introduced. Unlike conventional VLMs, Fuyu-8B omits a dedicated vision encoder and instead adds a simple linear projection for patched images. The projection layer and decoder-only transformer are trained jointly in an end-to-end manner using an autoregressive objective. Fuyu-8B demonstrated that a vision encoder is not necessary to achieve competitive performance on VLM benchmarks compared to state-of-the-art (SOTA) VLMs at the time that relied on intricate vision encoders.

Taking inspiration from Fuyu-8B, we introduce ELF, an encoder-free ELM that achieves competitive performances against other SOTA ELMs across five datasets. Unlike Fuyu-8B, which relies on image patching, and prior ELMs that depend on pretrained ECG encoders, we show that directly flattening the ECG signal and mapping it into the LLM embedding space with a single projection layer is sufficient. We summarize our main contributions below:

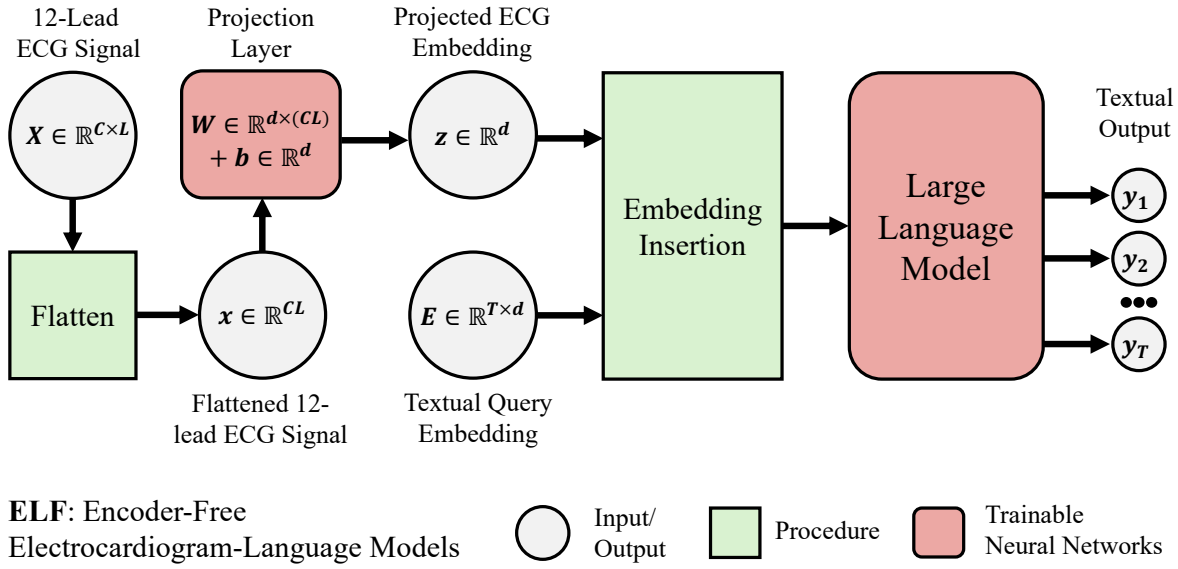

Figure 1: **The Architecture of ELF**. Given an ECG signal $X \in \mathbb{R}^{C \times L}$, where C is the number of leads and L the signal length, we flatten it to $x \in \mathbb{R}^{CL}$ and project it into a d-dimensional latent vector $z = Wx + b$, with $W \in \mathbb{R}^{d \times (CL)}$ and $b \in \mathbb{R}^{d}$. The projected ECG embedding is inserted into the textual query embeddings $E \in \mathbb{R}^{T \times d}$ and processed autoregressively by an LLM to generate outputs $y_{1:T}$.

1. We introduce ELF, an encoder-free ELM that maps ECGs into the LLM embedding space using a single projection layer, achieving competitive performance against far more complex ELMs across five datasets.
2. We introduce inductive biases into ELF via patching and convolutional layers, and find that the original single linear projection performs comparably to these more complex variants.
3. We examine ELF's dependence on the ECG modality and find that performance is often driven more by benchmark artifacts and language priors than by the ECG signal itself.

2 Related Works

2.1 ECG-Language Models

The automation of ECG analysis is advancing toward generative approaches with ELMs (Han et al., 2025). Previous ECG classification systems are limited to outputting probability scores for a fixed set of diagnostic labels (Qiu et al., 2023a; Hannun et al., 2019; Qiu et al., 2023b; Martin et al., 2021; Strodthoff et al., 2021; Liu et al., 2024a; Na et al., 2024; Choi et al., 2023; Jin et al., 2025). In contrast, ELMs can process both text and ECG inputs and generate free-form textual responses conditioned on the given input. Thus, ELMs not only inherit the capabilities of classification systems but also extend them to more general tasks, such as human interaction through text and clinically relevant applications like ECG waveform analysis, patient context inference, and treatment planning. This broad potential motivates the transition from classification-based ECG analysis to generative approaches with ELMs.

Although promising, ELMs remain in their early stages, with most prior works still exploring different ECG representations, architectures, learning objectives, and training regimes (Lan et al., 2025; Zhao et al., 2024; Song et al., 2025). Due to their similarity to VLMs, many studies first train a strong ECG-specific encoder, then pair the pretrained encoder with a pretrained LLM for natural language generation (NLG) (Li et al., 2025a). However, this approach requires substantial ECG data and computational resources to train an effective encoder. To address this limitation, recent works have proposed more efficient designs. MEIT (Wan et al., 2024), for instance, employs a lightweight ECG encoder composed of several 1-D convolutional layers with batch normalization (Ioffe and Szegedy, 2015), ReLU (Fukushima, 1969), and average pooling, trained end-to-end with the LLM. Another approach explores non-learning-based methods, such as applying Byte-Pair Encoding (BPE) (Gage, 1994) to tokenize the ECG signal and directly concatenate it with text query tokens (Han et al., 2024). We benchmark ELF against many of the recent de-

velopments in ELMs and showcase that a *single projection layer* is enough to maintain competitive performances.

2.2 Encoder-Free Vision-Language Models

Fuyu-8B (Bavishi et al., 2023) is one of the first encoder-free VLMs developed. In their work, the authors note that existing VLMs are overly complex, typically requiring separate image encoders, multi-stage training, and intricate connector modules. Fuyu-8B employs a simple decoder-only transformer where image patches are linearly projected into the first transformer layer. When benchmarked against state-of-the-art VLMs at the time, Fuyu-8B achieved competitive performance. We adopt the same motivation in developing ELF.

Since Fuyu-8B, there have been several advances in encoder-free VLMs (Diao et al., 2024; Team, 2025; Diao et al., 2025). However, we observed that the "connector" between the image and the LLM has grown increasingly complex, resembling a vision encoder. This trend undermines the original motivation behind encoder-free VLMs: *simplicity*. Therefore, in our work, we conduct most experiments using the architecture shown in Figure 1, where only a linear projection is used and jointly trained with the LLM. Additional experiments with slight architectural variations are reported in Table 4.

3 Methods

3.1 Datasets

We use preprocessed 12-lead, 5-second ECGs sampled at 250 Hz with 70:30 training/testing splits across five datasets that comprise of ECGs and text: 1) PULSE ECG-Instruct (Liu et al., 2024b), 2) PULSE ECG-Bench (Liu et al., 2024b), 3) ECG-Chat Instruct (Zhao et al., 2024), 4) PTB-XL (Wagner et al., 2020) ECG-QA (Oh et al., 2023), and 5) MIMIC-IV-ECG (Johnson et al., 2023) ECG-QA. These datasets are derived from ECG recordings originating from 1) MIMIC-IV-ECG (Johnson et al., 2023), 2) PTB-XL (Wagner et al., 2020), 3) CODE-15% (Ribeiro et al., 2021), 4) CSN (Zheng et al., 2020), and 5) CPSC (Andres et al., 2022). We note ECG-QA (Oh et al., 2023) contains three primary question types: single-verify, single-choose, and single-query. It also includes comparison-style questions, however we do not consider them in this study.

3.2 ECG-Language Models' Components

We describe the various ELM components utilized throughout the study for comparative baselines below.

ECG Tokenizer We utilize ECG-Byte (Han et al., 2024) as the ECG Tokenizer. The tokenizer is trained for 5000 merges with 300,000 randomly sampled, preprocessed, 10-second ECGs from the MIMIC-IV-ECG dataset. Deviating from the original work (Han et al., 2024) which globally normalizes ECGs using dataset-wide percentiles, we instead employ instance normalization, scaling each ECG individually to the range [0, 1]. We refer to this representation as the ECG Symbol representation.

ECG Encoder We utilize 2 SOTA encoders specifically designed for ECGs (thus utilizing the ECG Signal representation): ST-MEM (Na et al., 2024) and MERL (Liu et al., 2024a). Each encoder is pretrained using the Adam (Kingma and Ba, 2017) optimizer with learning rate 4e-5 and weight decay 1e-4. Training is conducted over 50 epochs with a batch size of 256 on 800,000 5-second instances of the MIMIC-IV-ECG dataset. All encoders are trained using two NVIDIA A6000 (48 GB) GPUs.

Vision Encoder We use the following vision encoders pretrained on internet-scale image data: 1) CLIP[1] (Radford et al., 2021), 2) ViT[2] (Dosovitskiy et al., 2021), and 3) SigLIP[3] (Zhai et al., 2023). Due to the vision encoders being pretrained already, we apply them out of the box for both ECG Stacked Signal and ECG Image representations.

Large Language Models We use the Llama-3.2-1B-Instruct[4] checkpoint (Grattafiori et al., 2024) accessed via the HuggingFace API (Wolf et al., 2020) for all experiments unless stated otherwise. We report results for other LLMs (i.e., Qwen2.5-1.5B-Instruct[5] and gemma-2-2b-it[6]) in Table 3. All LLMs are initialized with their default hyperparameters.

We train all LLMs on 400,000 training examples. In cases where a dataset contains less than 400,000 instances, the full split is used. We use the Adam

[1] `openai/clip-vit-base-patch32`
[2] `google/vit-base-patch16-224-in21k`
[3] `google/siglip-base-patch16-224`
[4] `meta-llama/Llama-3.2-1B-Instruct`
[5] `Qwen/Qwen2.5-1.5B-Instruct`
[6] `google/gemma-2-2b-it`

(Kingma and Ba, 2017) optimizer with learning rate 1e-4, weight decay 1e-2, $\beta_1 = 0.9$, $\beta_2 = 0.99$, and $\epsilon = 1e-8$. We employ LoRA (Hu et al., 2021) with rank 16, $\alpha_{LoRA} = 32$, and dropout 0.05. We use a batch size of 2 with 5 second 12-lead ECGs, max input sequence length of 1024, and warm-up of 500 steps.

Lastly, we use the gpt-5-mini model via the OpenAI API (OpenAI, 2025) and the OpenTSLM model[7] (Langer et al., 2025) to perform zero-shot evaluations on the PTB-XL ECG-QA dataset (Table 2). This experiment aims to establish a baseline against SOTA models in both general vision-language and time-series domains. Although these models are not trained on our PTB-XL ECG-QA splits, this comparison helps contextualize where ELF stands relative to these SOTA models. For gpt-5-mini, we use the ECG image representation and apply the system prompt shown in Appendix C. OpenTSLM is trained on five datasets through a multi-stage curriculum learning framework, including the ECG-QA chain-of-thought (CoT) dataset. Langer et al. (2025) provide a pre- and post-prompt template for their curated ECG-QA CoT dataset. This dataset extends PTB-XL ECG-QA by adding CoT rationales generated with GPT-4o (OpenAI et al., 2024). We use the same pre- and post-prompts applied during OpenTSLM's training and include them in Appendix B.

3.3 Learning Objective

As seen in Figure 1, we define the learning objective of ELF in more detail. Given an ECG signal $X \in \mathbb{R}^{C \times L}$, where C denotes the number of leads and L the signal length, we flatten it into a vector $x \in \mathbb{R}^{CL}$ and project it into a d-dimensional latent representation:

$$z = Wx + b, \qquad W \in \mathbb{R}^{d \times (CL)},\ b \in \mathbb{R}^{d}.$$

The projected ECG embedding z is inserted into the textual query embeddings $E \in \mathbb{R}^{T \times d}$ at the position of a special placeholder token `<signal>`, forming the input sequence to a LLM. T denotes the number of textual tokens. The model is trained autoregressively to maximize the likelihood of generating the response tokens $y_{1:T}$ conditioned on both the ECG signal and the query:

$$\mathcal{L}_{\text{AR}} = -\sum_{t=1}^{T} \log P_\theta\big(y_t \mid y_{<t}, E, z\big),$$

[7] `OpenTSLM/llama-3.2-3b-ecg-sp`

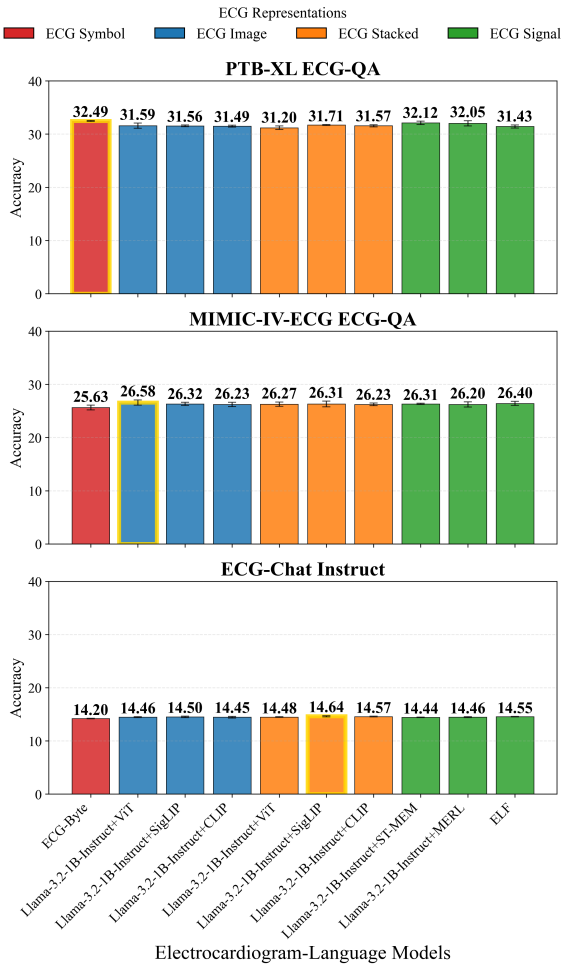

Figure 2: Averaged accuracy across three random seeds for each ELM on three datasets. All ELMs use the Llama-3.2-1B-Instruct backbone; best results are highlighted in yellow.

where θ denotes the model parameters. We mask out all tokens besides the response and end-of-sequence tokens when calculating $\mathcal{L}_{\text{AR}}$.

For other baseline ELMs, we follow the same exact learning objective formulation only differing in ECG representation (i.e., ECG signal, ECG Image, ECG Symbol, ECG Stacked) and their corresponding encoding schemas (i.e., ECG-Byte, ViT, CLIP, SigLIP, ST-MEM, MERL). We also note the projected ECG embedding z for ELF, ECG Encoders, and Vision Encoders all have a sequence length of 1, representing a singular "token". We adopt this design choice based on findings in the VLM literature, where prior work shows that reducing visual tokens, even down to a single token, can still preserve strong performance (Li et al., 2025b; Laurençon et al., 2024). We provide the conversation template and system prompt used throughout the experiments in Appendix B and C respectively.

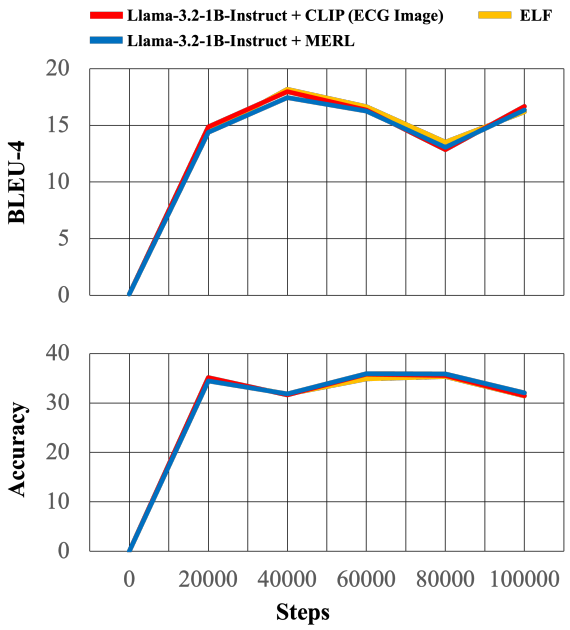

Figure 3: We compare ELF's performance against strong ELMs with dedicated encoders when evaluating across increasing steps trained. All results are averaged across three random seeds and trained/evaluated on the PTB-XL ECG-QA dataset.

3.4 Evaluation

We evaluate all models on 20,000 test examples repeated over three random seeds, reporting performance on the following text generation metrics: BLEU-4 (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005), ROUGE-L (Lin, 2004), and BertScore F1 (Zhang et al., 2020). We also report accuracy, the proportion of generated responses that exactly match the ground truth (i.e., exact string matching). All metrics are in percentage form.

4 Results

4.1 Baselines

We assess the accuracy of ELF with 9 other baseline ELMs across three datasets in Figure 2. We provide a tabular form of Figure 2 with standard deviations and additional metrics in Table 11 of Appendix D.1. We observe that ELMs such as ECG-Byte, Llama-3.2-1B-Instruct + ViT (ECG Image), and Llama-3.2-1B-Instruct + SigLIP (ECG Stacked) achieve the highest accuracies within their respective datasets. However, overall, all ELMs, including ELF, demonstrate comparable performance across the three datasets. Although top-performing models, such as Llama-3.2-1B-Instruct + ViT (ECG Image) and Llama-3.2-1B-Instruct + SigLIP (ECG Stacked), leverage complex encoders and training strategies, we observe ELF remains competitive. Lastly, we include successful and failed generations of ELF across the three datasets in Tables 9 and 10 of the Appendix.

We conduct two additional experiments comparing ELF, Llama-3.2-1B-Instruct + MERL, and Llama-3.2-1B-Instruct + CLIP (ECG Image). Figure 3 reports BLEU-4 and accuracy scores evaluated at different training steps. We observe that all three models follow similar performance trends throughout training. The second experiment, shown in Table 1, reports the BLEU-4 and accuracy scores when training and evaluating on the PULSE ECG-Instruct and PULSE ECG-Bench datasets respectively. In this setting, we observe ELF is able to achieve the highest BLEU-4 and accuracy scores.

Table 1: Mean ± standard deviations over three seeds trained on PULSE ECG-Instruct (Liu et al., 2024b) and evaluated on PULSE ECG-Bench (Liu et al., 2024b).

ELM	ECG Representation	BLEU-4	Accuracy
Llama-3.2-1B-Instruct + CLIP	ECG Image	0.33 ± 0.06	4.82 ± 0.11
Llama-3.2-1B-Instruct + MERL	ECG Signal	0.12 ± 0.01	5.49 ± 0.08
ELF	ECG Signal	**1.25 ± 0.03**	**8.06 ± 0.10**

Table 2: Comparing ELF against gpt-5-mini (OpenAI, 2025) and OpenTSLM (Langer et al., 2025) on the PTB-XL ECG-QA dataset. We report mean ± standard deviations calculated over three seeds.

Models	ECG Representation	BLEU-4	Accuracy
gpt-5-mini	ECG Image	15.26 ± 0.09	24.53 ± 0.11
OpenTSLM	ECG Signal	**20.03 ± 0.08**	18.55 ± 0.09
ELF	ECG Signal	16.18 ± 0.32	**31.43 ± 0.30**

Lastly, we compare ELF against gpt-5-mini (OpenAI, 2025) and OpenTSLM (Langer et al., 2025) on the PTB-XL ECG-QA dataset. Both models are evaluated in a zero-shot setting, and this experiment is intended only to position ELF relative to SOTA vision-language and time-series models. We note that (1) gpt-5-mini is a SOTA vision-language model, so using the RGB ECG image representation, we anticipated strong performance, and (2) OpenTSLM was trained on the ECG-QA-CoT dataset, which is derived from PTB-XL ECG-QA, so potential overlap between its training and our evaluation set suggested it might also perform well. We find that ELF attains the highest accuracy despite its far smaller training data and simpler architecture, while OpenTSLM achieves the best BLEU-4 score.

Table 3: We report performances on swapping the LLM. Reported mean ± standard deviations are over three seeds evaluated on PTB-XL ECG-QA.

ELMs	BLEU-4	Accuracy
Llama-3.2-1B-Instruct + CLIP	**16.68 ± 0.27**	31.49 ± 0.17
Llama-3.2-1B-Instruct + MERL	16.36 ± 0.05	**32.05 ± 0.49**
Llama-3.2-1B-Instruct + ELF	16.18 ± 0.32	31.43 ± 0.30
Qwen2.5-1.5B-Instruct + CLIP	14.62 ± 0.78	34.89 ± 0.36
Qwen2.5-1.5B-Instruct + MERL	**16.58 ± 0.44**	31.74 ± 0.05
Qwen2.5-1.5B-Instruct + ELF	9.70 ± 0.45	**38.81 ± 0.48**
gemma-2-2b-it + CLIP	14.62 ± 0.78	34.89 ± 0.36
gemma-2-2b-it + MERL	**15.45 ± 0.88**	35.10 ± 0.34
gemma-2-2b-it + ELF	15.07 ± 0.68	**36.69 ± 0.37**

4.2 Ablation Study on LLM Variants

We show averaged BLEU-4 and accuracy scores when swapping the LLM for different ELM architectures in Table 3. Llama-3.2-1B-Instruct with CLIP achieves the highest BLEU-4 score, while Qwen2.5-1.5B-Instruct with ELF achieves the highest accuracy. However, overall, the performance gaps across ELMs and their varying LLM backbones remain modest.

Table 4: Observing tradeoffs in performances for ELF variants. Reported mean ± standard deviations are calculated over three seeds evaluated on PTB-XL ECG-QA.

ELF Variants	BLEU-4	Accuracy
Base ELF	16.18 ± 0.32	31.43 ± 0.30
Patch ELF	**16.99 ± 0.10**	31.75 ± 0.22
Conv. ELF	16.22 ± 0.29	**32.31 ± 0.35**

4.3 Ablation Study on ELF Variants

Table 4 examines several architectural variants of ELF. These experiments are motivated by recent encoder-free VLMs (Diao et al., 2024), where techniques such as image patching and adding lightweight modules to the single projection layer (e.g., cross-attention, adaptive pooling) improved visual understanding over Fuyu-8B (Bavishi et al., 2023). However, as noted in subsection 2.2, introducing additional modules undermines the original simplicity that motivates Fuyu-8B. Thus, we present two minimal extensions to ELF: Patch ELF and Conv. ELF. In Table 4, Base ELF refers to the standard single-projection version described in subsection 3.3.

Patch ELF extends Base ELF by applying a patching strategy, partitioning the ECG signal $X \in \mathbb{R}^{C \times L}$ into $N = L/L_p$ non-overlapping patches, where L_p is the patch size, and each patch is represented as $X_i \in \mathbb{R}^{C \times L_p}$ for $i = 1, \ldots, N$. Each patch is flattened into $x_i \in \mathbb{R}^{CL_p}$ and projected using a shared linear projection layer W. We prepare N consecutive placeholder tokens <signal$_i$> for $i = 1, \ldots, N$, and the projected embedding for each patch is inserted into the input sequence at the position of its corresponding <signal$_i$> token. The Patch ELF variant investigates whether dividing the ECG signal into fixed-size patches, producing a sequence of tokens analogous to patching in VLMs, leads to improved performance.

Conv. ELF adds additional complexity to Patch ELF by replacing the linear projection with a series of convolutional layers. Given a sequence of input patches $X_i \in \mathbb{R}^{C \times L_p}$, Conv. ELF applies two 1D convolutions along the temporal dimension L_p, treating the C leads as separate input channels, followed by global average pooling $\mathrm{Pool}()$ and a linear projection W to the LLM hidden dimension:

$$z_i = W\,\mathrm{Pool}(G(X_i)) + b,$$

where G denotes the two convolutional layers. Each patch latent vector z_i is then inserted into the input sequence following the same procedure as Patch ELF. The Conv. ELF variant investigates whether introducing both patching and temporal convolutions, thereby encoding simple inductive biases, can better extract ECG features and improve performance.

Table 4 shows that the Base ELF achieves only marginally lower BLEU-4 and accuracy than the more complex Patch and Conv. ELF variants. The minimal performance differences between the variants suggests that injecting additional inductive biases into ELF does not yield significant performance gains.

Table 5: We selectively train ELF's components and report mean ± standard deviation over three seeds when evaluating on the PTB-XL ECG-QA and ECG-Chat Instruct datasets. ✓indicates the corresponding component was set as trainable.

Dataset	Components		BLEU-4	Accuracy
	LLM	Projection Layer W		
PTB-XL ECG-QA			0.11 ± 0.00	0.00 ± 0.00
	✓		**16.31 ± 0.09**	32.46 ± 0.25
		✓	10.09 ± 0.45	**34.82 ± 0.26**
	✓	✓	16.18 ± 0.32	31.43 ± 0.30
ECG-Chat Instruct			4.26 ± 0.04	0.39 ± 0.03
	✓		25.51 ± 0.19	14.50 ± 0.09
		✓	18.73 ± 0.06	6.96 ± 0.12
	✓	✓	**25.63 ± 0.26**	**14.55 ± 0.04**

4.4 Selectively Training ELF Components

We study the effect of selectively training ELF components in Table 5. Freezing both the LLM and projection layer W causes performance to collapse on both datasets, indicating that at least one trainable component is necessary.

On PTB-XL ECG-QA, accuracy is largely insensitive once either module is trainable, while BLEU-4 depends strongly on which component is updated. Training the LLM with W frozen yields the highest BLEU-4, whereas training only W achieves the best accuracy but substantially lower BLEU-4.

On ECG-Chat Instruct, both metrics depend on the trained components. Training only W underperforms training the LLM (accuracy 6.96 vs. 14.50; BLEU-4 18.73 vs. 25.51), while jointly training both yields the best overall results but is nearly indistinguishable from training the LLM alone.

4.5 ELF's Reaction Under Three Training/Inference Conditions

We observe ELF's accuracy under three training and inference conditions with the PTB-XL ECG-QA and ECG-Chat Instruct datasets in Figure 4 and Table 6 respectively. The three conditions are defined as the following:

ECG Signal: The regular ECG Signal and corresponding textual query is used as input.

Zeros Tensor: A tensor of the same shape as the regular ECG Signal filled with zeros is used as input, alongside the textual query.

Only Text: The ECG Signal is completely omitted and only the textual query is used as input. We note only the input is modified, and the architecture of ELF remains the same.

The motivation behind these conditions is to probe whether ELF leverages the ECG signal modality or relies predominantly on language priors learned during pretraining of the LLM. We also draw inspiration from Cambrian-1 (Tong et al., 2024), which evaluated performance with and without visual input. The results showed less than a 5% difference between the two settings, suggesting that some benchmarks may not meaningfully depend on visual information and instead primarily test the language model's reasoning ability.

Let's first take a look at the results in Figure 4. When trained under normal conditions (i.e., using the ECG Signal), ELF exhibits the expected performance drop when evaluated with the Zeros Tensor or Only Text inputs. However, training and inferencing with the Zeros Tensor still results in relatively high accuracy (28.05%), and training and inferencing with Only Text achieves 30.05%, which is comparable to performance under normal conditions. Additionally, training with Only Text and inferencing with the ECG Signal produces even higher accuracy than the standard setting (37.09%).

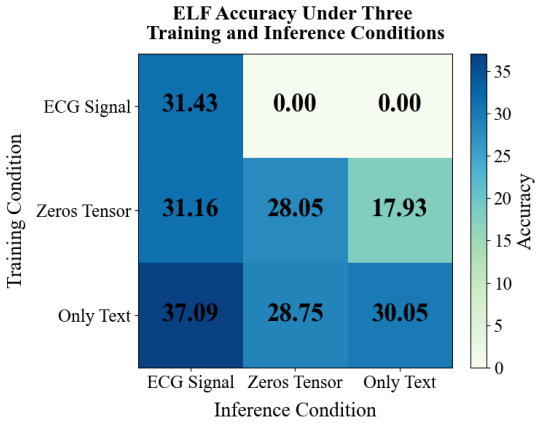

Figure 4: We train and inference ELF under three training and inference conditions: 1) ECG Signal, 2) Zeros Tensor, and 3) Only Text. All reported results are averaged across three random seeds on the PTB-XL ECG-QA dataset.

Table 6: We train ELF with ECG Signals as inputs and report mean ± standard deviations over three seeds when evaluating with Zeros Tensor and Only Text on the ECG-Chat Instruct dataset. We also report the results under the normal inference condition (ECG Signal) for comparison.

Inference Condition	BLEU-4	Accuracy
ECG Signal	**25.63 ± 0.26**	**14.55 ± 0.04**
Zeros Tensor	25.26 ± 0.07	13.95 ± 0.11
Only Text	24.90 ± 0.07	13.74 ± 0.10

This phenomenon seems to hold when ELF is trained on the ECG-Chat Instruct dataset (Table 6). We train ELF with regular ECG Signal inputs and inference under all three conditions. We observe that ELF demonstrates similar performances across all three conditions.

We also evaluated Llama-3.2-1B-Instruct + CLIP (trained on regular ECG images), gpt-5-mini, and OpenTSLM using Zeros Tensor inputs and report the results in Table 7. For clarity, gpt-5-mini and OpenTSLM are not trained; we simply evaluate them out-of-the-box under the Zeros Tensor inference condition. For Llama-3.2-1B-Instruct + CLIP, using the Zeros Tensor during inference leads to

Table 7: We evaluate SOTA vision (gpt-5-mini) and time-series (OpenTSLM) language models under the Zeros Tensor inference condition and report accuracy (mean ± standard deviation across three seeds) on the PTB-XL ECG-QA dataset. We also include results from training the Llama-3.2-1B-Instruct + CLIP ELM on standard ECG images and evaluating it under the Zeros Tensor condition.

Model	ECG Representation	Zeros Tensor	Accuracy
Llama-3.2-1B-Instruct + CLIP	ECG Image		31.49 ± 0.17
		✓	28.31 ± 0.13
gpt-5-mini	ECG Image		24.53 ± 0.11
		✓	14.69 ± 0.08
OpenTSLM	ECG Signal		18.55 ± 0.09
		✓	15.83 ± 0.14

a slight accuracy drop, though the difference between the regular ECG image and Zeros Tensor conditions remains small. OpenTSLM also shows a moderate accuracy reduction, similar to Llama-3.2-1B-Instruct + CLIP. Interestingly, gpt-5-mini shows a larger drop in performance when comparing the accuracy between regular ECG-Image inputs and Zeros Tensor inputs (24.53 $\rightarrow$ 14.69).

After observing the results in Figure 4 and Tables 6 and 7, we draw the following conclusion: current ELMs are not meaningfully leveraging the ECG signal for accurate responses. ELF's performance remains largely unchanged when the ECG signal is omitted (Only Text) or replaced with zeros (Zeros Tensor), and OpenTSLM shows a similarly small change under the Zeros Tensor condition. In contrast, gpt-5-mini experiences a substantial accuracy drop when the ECG input is replaced with a Zeros Tensor, suggesting it may be responding to aspects of the ECG Image itself, as the plotted input would appear as flatlines across all leads. Overall, these results indicate that the text component remains the most consistently informative part of the sequence and that simple perturbation-based evaluations can help reveal whether ELMs are meaningfully engaging with the ECG modality, providing a direction for further investigation.

5 Discussion and Conclusions

In this work, we examined whether the architectural complexity common in recent ELMs is necessary for strong NLG performance. We introduced ELF, an encoder-free ELM that replaces the ECG encoder with a single jointly trained projection layer, and showed that this design attains performance comparable to or better than models that rely on separately pretrained ECG or vision encoders. Across five datasets and multiple backbones, performance gaps between ELF and stronger encoder-based baselines remained modest, suggesting that a single projection is sufficient for current ECG-language tasks.

Our architectural ablations reinforce this finding. Patch ELF and Conv. ELF introduce inductive biases through patching and temporal convolutions, yet yield only minor changes in BLEU-4 and accuracy relative to Base ELF. Swapping LLM backbones shifts absolute performance but does not meaningfully alter the relative behavior of encoder-free and encoder-based ELMs. These results suggest that, under current experimental conditions, overall model capacity and language modeling ability dominate performance, with the choice of ECG connector or encoder playing a minor role. As a result, encoder complexity can often be reduced without sacrificing NLG performance (Li et al., 2025b; Bavishi et al., 2023).

The perturbation experiments highlight a deeper issue regarding how ELMs interact with ECG inputs. When the ECG is zeroed out or removed, ELF retains unexpectedly high accuracy. Llama-3.2-1B-Instruct + CLIP and OpenTSLM show similar behaviors under Zeros Tensor inputs. In contrast, gpt-5-mini exhibits larger degradation for ECG images, indicating some genuine sensitivity to the visual modality. Overall, these patterns suggest that the text component often dominates ELM performance and that existing ECG and text benchmarks may not reliably enforce dependence on the ECG signal (Tong et al., 2024). We view these perturbation tests as a practical diagnostic tool for assessing whether ELMs genuinely use the ECG or rely primarily on language priors.

In conclusion, our findings point to a broader opportunity for the ELM community. Encoder-free designs such as ELF appear sufficient for current ECG-text tasks, yet our perturbation analyses indicate that stronger benchmarks will be necessary to ensure that ELMs faithfully use ECG information. We believe that future progress will benefit from more carefully curated evaluation settings that place clearer demands on ECG understanding, enabling models to demonstrate not only strong language generation but also meaningful engagement with the underlying physiological signal.

Limitations

We highlight two overarching concerns. First, the perturbation experiments indicate that many ELMs do not extract meaningful features from the ECG signal and instead rely heavily on language priors learned during LLM pretraining (Figure 4). Second, progress is limited by the lack of high-quality datasets that demand coherent ECG understanding. Performance on single-verify and single-choose categories from the ECG-QA datasets can be inflated. A model that always predicts the majority choice, such as always answering "yes" or always selecting the first option, can match the underlying label frequency and achieve deceptively high accuracy. More naturalistic ECG and text datasets are available (e.g., ECG-Chat Instruct) but after careful examination, many examples still contain clinically irrelevant text. This is partly due to the fact that most ECG and text datasets rely on proprietary LLMs for generation. Achieving a true "GPT moment" for ELMs will require ECG and text datasets curated with the same rigor that is applied to large-scale text and vision corpora used for general-domain LLMs and MLLMs.

References

AAMC. 2024. The complexities of physician supply and demand: Projections from 2021 to 2036.

Erick Andres, Annie Gu, Amit Shah, Chengyu Liu, Ashish Sharma, Salman Seyedi, Ali Bahrami Rad, Matthew Reyna, and Gari Clifford. 2022. Classification of 12-lead ecgs: The physionet/computing in cardiology challenge 2020.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *IEEvaluation@ACL*.

Rohan Bavishi, Erich Elsen, Curtis Hawthorne, Maxwell Nye, Augustus Odena, and Sagnak Tasirlar. 2023. Fuyu-8b: A multimodal architecture for ai agents.

Seokmin Choi, Sajad Mousavi, Phillip Si, Haben G. Yhdego, Fatemeh Khadem, and Fatemeh Afghah. 2023. Ecgbert: Understanding hidden language of ecgs with self-supervised representation learning. *Preprint*, arXiv:2306.06340.

Haiwen Diao, Yufeng Cui, Xiaotong Li, Yueze Wang, Huchuan Lu, and Xinlong Wang. 2024. Unveiling encoder-free vision-language models. *Preprint*, arXiv:2406.11832.

Haiwen Diao, Xiaotong Li, Yufeng Cui, Yueze Wang, Haoge Deng, Ting Pan, Wenxuan Wang, Huchuan Lu, and Xinlong Wang. 2025. Evev2: Improved baselines for encoder-free vision-language models. *Preprint*, arXiv:2502.06788.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Kunihiko Fukushima. 1969. Visual feature extraction by a multilayered network of analog threshold elements. *IEEE Transactions on Systems Science and Cybernetics*, 5:322–333.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

William Han, Chaojing Duan, Zhepeng Cen, Yihang Yao, Xiaoyu Song, Atharva Mhaskar, Dylan Leong, Michael A. Rosenberg, Emerson Liu, and Ding Zhao. 2025. Signal, image, or symbolic: Exploring the best input representation for electrocardiogram-language models through a unified framework. *Preprint*, arXiv:2505.18847.

William Han, Chaojing Duan, Michael A. Rosenberg, Emerson Liu, and Ding Zhao. 2024. Ecg-byte: A tokenizer for end-to-end generative electrocardiogram language modeling. *Preprint*, arXiv:2412.14373.

Awni Y. Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H. Tison, Codie Bourn, Mintu P. Turakhia, and Andrew Y. Ng. 2019. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature Medicine*, 25:65–69.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Preprint*, arXiv:1502.03167.

Jiarui Jin, Haoyu Wang, Hongyan Li, Jun Li, Jiahui Pan, and Shenda Hong. 2025. Reading your heart: Learning ecg words and sentences via pre-training ecg language model. *Preprint*, arXiv:2502.10707.

Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J. Pollard, Benjamin Moody, Brian Gow, Li-wei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. Mimic-iv, a freely accessible electronic health record dataset. *Scientific Data*, 10.

Diederik P. Kingma and Jimmy Ba. 2017. Adam: A method for stochastic optimization. *Preprint*, arXiv:1412.6980.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. 2025. Gem: Empowering mllm for grounded ecg understanding with time series and images. *Preprint*, arXiv:2503.06073.

Patrick Langer, Thomas Kaar, Max Rosenblattl, Maxwell A. Xu, Winnie Chow, Martin Maritsch, Aradhana Verma, Brian Han, Daniel Seung Kim, Henry Chubb, Scott Ceresnak, Aydin Zahedivash, Alexander Tarlochan Singh Sandhu, Fatima Rodriguez, Daniel McDuff, Elgar Fleisch, Oliver Aalami, Filipe Barata, and Paul Schmiedmayer. 2025. Opentslm: Time-series language models for reasoning over multivariate medical text- and time-series data. *Preprint*, arXiv:2510.02410.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. 2024. What matters when building vision-language models? *Preprint*, arXiv:2405.02246.

Haitao Li, Ziyu Li, Yiheng Mao, Ziyi Liu, Zhoujian Sun, and Zhengxing Huang. 2025a. anyecg-chat: A generalist ecg-mllm for flexible ecg input and multi-task understanding. *Preprint*, arXiv:2506.00942.

Kevin Y. Li, Sachin Goyal, Joao D. Semedo, and J. Zico Kolter. 2025b. Inference optimal vlms need fewer visual tokens and more parameters. *Preprint*, arXiv:2411.03312.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL 2004*.

Che Liu, Zhongwei Wan, Cheng Ouyang, Anand Shah, Wenjia Bai, and Rossella Arcucci. 2024a. Zero-shot ecg classification with multimodal learning and test-time clinical knowledge enhancement. *Preprint*, arXiv:2403.06659.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. 2024b. Teach multimodal llms to comprehend electrocardiographic images. *Preprint*, arXiv:2410.19008.

Harold Martin, Ulyana Morar, Walter Izquierdo, Mercedes Cabrerizo, Anastasio Cabrera, and Malek Adjouadi. 2021. Real-time frequency-independent single-lead and single-beat myocardial infarction detection. *Artificial intelligence in medicine*, 121:102179.

Yeongyeon Na, Minje Park, Yunwon Tae, and Sunghoon Joo. 2024. Guiding masked representation learning to capture spatio-temporal relationship of electrocardiogram. *Preprint*, arXiv:2402.09450.

Jungwoo Oh, Gyubok Lee, Seongsu Bae, Joon myoung Kwon, and Edward Choi. 2023. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *Preprint*, arXiv:2306.15681.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. Gpt-5 system card openai.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *ACL*.

Hung Manh Pham, Jialu Tang, Aaqib Saeed, and Dong Ma. 2025. Q-heart: Ecg question answering via knowledge-informed multimodal llms. *Preprint*, arXiv:2505.06296.

Jielin Qiu, William Han, Jiacheng Zhu, Mengdi Xu, Michael Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. 2023a. Transfer knowledge from natural language to electrocardiography: Can we detect cardiovascular disease through language models? In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 442–453, Dubrovnik, Croatia. Association for Computational Linguistics.

Jielin Qiu, Jiacheng Zhu, Shiqi Liu, William Han, Jingqi Zhang, Chaojing Duan, Michael A. Rosenberg, Emerson Liu, Douglas Weber, and Ding Zhao. 2023b. Automated cardiovascular record retrieval by multimodal learning between electrocardiogram and clinical report. In *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pages 480–497. PMLR.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Pranav Rajpurkar, Awni Y. Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y. Ng. 2017. Cardiologist-level arrhythmia detection with convolutional neural networks. *Preprint*, arXiv:1707.01836.

Antônio H. Ribeiro, Gabriela M.M. Paixao, Emilly M. Lima, Manoel Horta Ribeiro, Marcelo M. Pinto Filho, Paulo R. Gomes, Derick M. Oliveira, Wagner Meira Jr, Thömas B Schon, and Antonio Luiz P. Ribeiro. 2021. Code-15%: a large scale annotated dataset of 12-lead ecgs.

Xiaoyu Song, William Han, Tony Chen, Chaojing Duan, Michael A. Rosenberg, Emerson Liu, and Ding Zhao. 2025. Retrieval-augmented generation for electrocardiogram-language models. *Preprint*, arXiv:2510.00261.

Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. 2021. Deep learning for ecg analysis: Benchmarks and insights from ptb-xl. *IEEE Journal of Biomedical and Health Informatics*, 25:1519–1528.

Chameleon Team. 2025. Chameleon: Mixed-modal early-fusion foundation models. *Preprint*, arXiv:2405.09818.

Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, Ziteng Wang, Rob Fergus, Yann LeCun, and Saining Xie. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Preprint*, arXiv:2406.16860.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. 2020. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154. Number: 1 Publisher: Nature Publishing Group.

Zhongwei Wan, Che Liu, Xin Wang, Chaofan Tao, Hui Shen, Zhenwu Peng, Jie Fu, Rossella Arcucci, Huaxiu Yao, and Mi Zhang. 2024. Meit: Multi-modal electrocardiogram instruction tuning on large language models for report generation. *Preprint*, arXiv:2403.04945.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *Preprint*, arXiv:2303.15343.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. *ArXiv*, abs/1904.09675.

Yubao Zhao, Tian Zhang, Xu Wang, Puyu Han, Tong Chen, Linlin Huang, Youzhu Jin, and Jiaju Kang. 2024. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. *Preprint*, arXiv:2408.08849.

Jianwei Zheng, Jianming Zhang, Sidy Danioko, Hai Yao, Hangyuan Guo, and Cyril Rakovski. 2020. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7:48.

Hongling Zhu, Cheng Cheng, Hang Yin, Xingyi Li, Ping Zuo, Jia Ding, Fan Lin, Jingyi Wang, Beitong Zhou, Yonge Li, Shouxing Hu, Yulong Xiong, Binran Wang, Guohua Wan, Xiaoyun Yang, and Ye Yuan. 2020. Automatic multilabel electrocardiogram diagnosis of heart rhythm or conduction abnormalities with deep learning: a cohort study. *The Lancet Digital Health*, 2:e348–e357.

A ECG Representations

We summarize and modify the definitions of each ECG representation provided in Han et al. (2025).

ECG Signal. The raw ECG signal is denoted as $X_{\text{sig}} \in \mathbb{R}^{C \times L}$, where C represents the number of leads (channels) and L the number of time samples per lead. All subsequent modalities in this work are derived from X_{sig}. Most ELMs utilize the ECG Signal representation (Han et al., 2025; Wan et al., 2024; Zhao et al., 2024; Li et al., 2025a; Pham et al., 2025; Langer et al., 2025).

Stacked ECG Signal. To create a three-channel version of an ECG signal suitable for pretrained vision-encoders, we replicate X_{sig} three times along the channel dimension, forming a stacked signal $X^*_{\text{sig}} \in \mathbb{R}^{3 \times C \times L}$. Previous works (Han et al., 2024, 2025) utilized the Stacked ECG Signal representation as input to pretrained vision-encoders, such as CLIP (Radford et al., 2021), ViT (Dosovitskiy et al., 2021), and SigLIP (Zhai et al., 2023).

ECG Image. An ECG image is generated by plotting X_{sig} and representing it as a tensor $X_{\text{img}} \in \mathbb{R}^{H \times W \times C'}$, where H and W denote the image height and width, and C' is the number of color channels. One can observe that this ECG representation integrates naturally with existing VLM architectures employing vision encoders, as demonstrated by prior adaptations of VLMs for ECG plot images (Liu et al., 2024b). Lan et al. (2025) utilized both the ECG Image and ECG Signal representation.

ECG Symbol. Using the ECG-Byte compression schema (Han et al., 2024), each normalized and discretized ECG signal X_{sig} is transformed into a symbolic sequence over a finite alphabet $\mathcal{A} = \{\text{a}, \text{b}, \ldots, \text{z}\}$. This symbolic sequence is flattened into a one-dimensional array $X_{\text{symb}} \in \mathcal{A}^{C \cdot L}$. A byte-pair encoding (BPE) process (Gage, 1994) is then applied to compress X_{symb} into a sequence of discrete tokens from an extended vocabulary $\mathcal{V}$, yielding the final symbolic representation $X_{\text{ID}} \in \mathcal{V}^m$, where m is the token sequence length.

B Conversation Templates

B.1 Large Language Model Conversation Templates

We provide the conversation templates for Llama-3.2-1B-Instruct, gemma-2-2b-it, and Qwen-2.5-1.5B-Instruct. We also provide the pre- and post-conversation templates used for the OpenTSLM model.

Llama-3.2-1B-Instruct Conversation Template

< |begin_of_text| >< |start_header_id| > system < |end_header_id| >

q_{sys} < |eot_id| >< |start_header_id| > user < |end_header_id| >

< signal >
q_1 < |eot_id| >< |start_header_id| > assistant < |end_header_id| >

y_1 < |eot_id| >< |start_header_id| > user < |end_header_id| >

q_2 < |eot_id| >< |start_header_id| > assistant < |end_header_id| >

y_2 < |eot_id| >
...

q_n < |eot_id| >< |start_header_id| > assistant < |end_header_id| >

y_n < |eot_id| >

gemma-2-2b-it Conversation Template

< bos >< start_of_turn > user
< signal >
q_1 < end_of_turn >
< start_of_turn > model
y_1 < end_of_turn >
< start_of_turn > user
q_2 < end_of_turn >
< start_of_turn > model
y_2 < end_of_turn >

...

< start_of_turn > user
q_n < end_of_turn >
< start_of_turn > model
y_n < end_of_turn >

Qwen-2.5-1.5B-Instruct Conversation Template

$<|\text{im\_start}|>$ system
$q_{\text{sys}} <|\text{im\_end}|>$
$<|\text{im\_start}|>$ user
$<$ signal $>$
$q_1 <|\text{im\_end}|>$
$<|\text{im\_start}|>$ assistant
$y_1 <|\text{im\_end}|>$
$<|\text{im\_start}|>$ user
$q_2 <|\text{im\_end}|>$
$<|\text{im\_start}|>$ assistant
$y_2 <|\text{im\_end}|>$

...

$<|\text{im\_start}|>$ user
$q_n <|\text{im\_end}|>$
$<|\text{im\_start}|>$ assistant
$y_n <|\text{im\_end}|>$

OpenTSLM Pre-Prompt
You are an expert cardiologist analyzing an ECG (electrocardiogram).

Clinical Context: 12-lead ECG recording.

Your task is to examine the ECG signal and answer the following medical question:

Question: q_1

Instructions:
- Begin by analyzing the time series without assuming a specific answer.
- Think step-by-step about what the observed patterns suggest regarding the cardiac condition.
- Write your rationale as a single, natural paragraph — do not use bullet points, numbered steps, or section headings.
- Do **not** mention any final answer until the very end.
- Consider the ECG morphology, intervals, and any abnormalities that relate to the question.

OpenTSLM Post-Prompt
Based on your analysis of the ECG data, provide your answer.
Make sure that your last word is the answer. You MUST end your response with "Answer: "

C System Prompt

We provide the system prompts utilized throughout the study. Namely, we provide the ELM system prompt used for all experiments with the Llama-3.2-1B-Instruct, gemma-2-2b-it, and Qwen-2.5-1.5B-Instruct models. We also provide the system prompt used for the gpt-5-mini experiments.

ELM System Prompt
You are an expert multimodal assistant capable of processing both natural language text and ECG signals. When you receive input, first determine if it is text, ECG data, or both. For ECG signals, interpret them as time-series data representing cardiac activity—analyzing features such as heart rate, rhythm, and potential abnormalities. When both modalities are present, synthesize the information to provide integrated, expert cardiac electrophysiologist-level responses. Your answers should be precise, concise, and informed by clinical signal analysis and natural language understanding. Additionally, if the user asks a general question, you should answer it as a general assistant.

gpt-5-mini System Prompt
You are an expert multimodal assistant with advanced knowledge in **clinical cardiac electrophysiology**.

ECG Signal Analysis
- Given the ECG signal plot and question, provide a concise answer to the question.

Response Style
- Deliver responses that are **precise and concise**.
- Make sure to provide only the answer, no other text or explanations. For example, if the answer is "Yes", just simply respond with "Yes".

D Additional Results

D.1 Tabular Form of Figure 2

We provide the tabular form of Figure 2 with additional metrics in Table 11.

Table 8: Question and answer samples for each category from the testing set of ECG-QA.

ECG-QA Dataset	Category	Question	Answer
PTB-XL	single-verify	Is subendocardial injury in anteroseptal leads detectable from this ECG?	yes
	single-choose	What is the diagnostic symptom displayed in this ECG, subendocardial injury in anteroseptal leads or non-specific st changes, even if there are uncertain symptoms?	subendocardial injury in anteroseptal leads
	single-query	What types of noises are displayed in lead V3 in this ECG waveform?	baseline drift
MIMIC-IV-ECG	single-verify	Does the pr interval shown on this ECG fall within the normal range?	no
	single-choose	What form-related traits are exhibited by this ECG in lead V1?	ST Elevation ST(-T) change
	single-query	Which form-related symptom does this ECG show, inverted T-waves or early R wave transition?	none

Table 9: Successful generations for a given textual query across three datasets.

Dataset	Textual Query	Generated Response
PTB-XL ECG-QA	What are the numerical measurements of this ECG that fall above the normal range?	qt interval p duration qt corrected
MIMIC-IV-ECG ECG-QA	Which form-related symptom does this ECG show, low QRS voltage or poor R wave progression?	none
ECG-Chat Instruct	What is the R wave peak on this ECG?	The R wave peak is 2445 mV

Table 10: Failed generations for a given textual query across three datasets.

Dataset	Textual Query	Ground Truth	Generated Response
PTB-XL ECG-QA	Could you elaborate on the diagnostic symptoms present in this ECG, including those that are uncertain?	non-specific ischemic left ventricular hypertrophy left anterior fascicular block nonspecific intraventricular conduction disturbance (block)	myocardial infarction in anteroseptal leads left anterior fascicular block myocardial infarction in inferior leads
MIMIC-IV-ECG ECG-QA	Which form-related characteristics are visible in this ECG?	premature atrial complexes ST(-T) change repolarization abnormality multifocal premature ventricular complexes ST-T change due to ventricular hypertrophy	ST-T change due to ventricular hypertrophy ST(-T) change QRS changes in anteroseptal leads repolarization abnormality
ECG-Chat Instruct	What should I know about my ECG?	Your ECG shows a sinus rhythm, which is a normal finding, and the overall pattern of the waveforms is consistent with a healthy heart function. There are no abnormalities or irregularities in the various indicators, suggesting that your heart's electrical activity is functioning properly. Based on this, there are no signs of any cardiac disease present.	Your ECG shows a sinus rhythm with a right bundle branch block (RBBB) and a left anterior fascicular block (LAFB). The overall interpretation is an abnormal ECG, indicating potential issues with the heart's electrical conduction system. This pattern could suggest underlying cardiac disease or structural abnormalities in the heart.

Table 11: Mean ± standard deviations over three seeds on the PTB-XL ECG-QA, MIMIC-IV-ECG ECG-QA, and ECG-Chat Instruct datasets.

Dataset	ELM	ECG Representation	BLEU-4	ROUGE-L	METEOR	BertScore F1	Accuracy
PTB-XL ECG-QA	ECG-Byte	ECG Symbol	14.44 ± 0.12	40.89 ± 0.14	25.86 ± 0.16	**92.53 ± 0.08**	**32.49 ± 0.09**
	Llama-3.2-1B-Instruct + ViT	ECG Image	16.29 ± 0.16	41.46 ± 0.45	27.15 ± 0.24	92.47 ± 0.13	31.59 ± 0.48
	Llama-3.2-1B-Instruct + SigLIP		15.71 ± 0.09	41.19 ± 0.10	26.86 ± 0.29	92.48 ± 0.08	31.56 ± 0.14
	Llama-3.2-1B-Instruct + CLIP		**16.68 ± 0.27**	41.26 ± 0.05	26.90 ± 0.15	92.51 ± 0.08	31.49 ± 0.17
	Llama-3.2-1B-Instruct + ViT	ECG Stacked Signal	16.30 ± 0.32	41.02 ± 0.34	26.93 ± 0.39	92.45 ± 0.11	31.20 ± 0.33
	Llama-3.2-1B-Instruct + SigLIP		16.11 ± 0.14	41.68 ± 0.17	27.24 ± 0.23	92.52 ± 0.09	31.71 ± 0.08
	Llama-3.2-1B-Instruct + CLIP		16.15 ± 0.45	40.92 ± 0.38	26.68 ± 0.26	92.42 ± 0.08	31.57 ± 0.21
	Llama-3.2-1B-Instruct + ST-MEM	ECG Signal	16.46 ± 0.37	**41.88 ± 0.35**	**27.29 ± 0.12**	92.50 ± 0.09	32.12 ± 0.31
	Llama-3.2-1B-Instruct + MERL		16.36 ± 0.05	41.64 ± 0.39	27.03 ± 0.18	92.52 ± 0.10	32.05 ± 0.49
	ELF		16.18 ± 0.32	41.28 ± 0.34	26.94 ± 0.16	92.44 ± 0.13	31.43 ± 0.30
MIMIC-IV-ECG ECG-QA	ECG-Byte	ECG Symbol	20.92 ± 0.29	41.57 ± 0.50	**30.73 ± 0.41**	**91.47 ± 0.05**	25.63 ± 0.44
	Llama-3.2-1B-Instruct + ViT	ECG Image	**21.88 ± 0.48**	**41.69 ± 0.33**	28.60 ± 0.19	91.42 ± 0.03	**26.58 ± 0.50**
	Llama-3.2-1B-Instruct + SigLIP		21.41 ± 0.26	41.39 ± 0.18	28.26 ± 0.12	91.36 ± 0.04	26.32 ± 0.29
	Llama-3.2-1B-Instruct + CLIP		20.85 ± 0.09	41.17 ± 0.26	28.09 ± 0.19	91.38 ± 0.04	26.23 ± 0.40
	Llama-3.2-1B-Instruct + ViT	ECG Stacked Signal	21.37 ± 0.35	41.21 ± 0.33	28.10 ± 0.11	91.39 ± 0.03	26.27 ± 0.41
	Llama-3.2-1B-Instruct + SigLIP		21.54 ± 0.09	41.47 ± 0.45	28.35 ± 0.29	91.40 ± 0.03	26.31 ± 0.52
	Llama-3.2-1B-Instruct + CLIP		20.61 ± 0.39	41.15 ± 0.20	28.04 ± 0.12	91.34 ± 0.01	26.23 ± 0.24
	Llama-3.2-1B-Instruct + ST-MEM	ECG Signal	21.39 ± 0.11	41.24 ± 0.09	28.33 ± 0.07	91.38 ± 0.02	26.31 ± 0.11
	Llama-3.2-1B-Instruct + MERL		21.04 ± 0.26	41.14 ± 0.54	27.90 ± 0.42	91.32 ± 0.06	26.20 ± 0.49
	ELF		21.40 ± 0.27	41.31 ± 0.28	28.13 ± 0.23	91.37 ± 0.05	26.40 ± 0.38
ECG-Chat Instruct	ECG-Byte	ECG Symbol	**25.70 ± 0.25**	66.49 ± 0.20	63.34 ± 0.22	95.55 ± 0.03	14.20 ± 0.05
	Llama-3.2-1B-Instruct + ViT	ECG Image	25.58 ± 0.25	66.63 ± 0.11	63.40 ± 0.14	95.55 ± 0.03	14.46 ± 0.09
	Llama-3.2-1B-Instruct + SigLIP		25.53 ± 0.30	66.63 ± 0.21	63.43 ± 0.23	95.55 ± 0.04	14.50 ± 0.14
	Llama-3.2-1B-Instruct + CLIP		25.59 ± 0.33	66.66 ± 0.16	63.45 ± 0.15	95.55 ± 0.03	14.45 ± 0.17
	Llama-3.2-1B-Instruct + ViT	ECG Stacked Signal	25.60 ± 0.22	66.64 ± 0.15	63.40 ± 0.15	95.55 ± 0.03	14.48 ± 0.06
	Llama-3.2-1B-Instruct + SigLIP		25.63 ± 0.28	66.75 ± 0.10	63.52 ± 0.11	95.57 ± 0.03	**14.64 ± 0.11**
	Llama-3.2-1B-Instruct + CLIP		25.64 ± 0.27	**66.79 ± 0.14**	**63.64 ± 0.15**	**95.57 ± 0.02**	14.57 ± 0.07
	Llama-3.2-1B-Instruct + ST-MEM	ECG Signal	25.54 ± 0.26	66.63 ± 0.13	63.44 ± 0.16	95.55 ± 0.03	14.44 ± 0.04
	Llama-3.2-1B-Instruct + MERL		25.61 ± 0.31	66.67 ± 0.15	63.36 ± 0.16	95.55 ± 0.03	14.46 ± 0.09
	ELF		25.63 ± 0.26	66.73 ± 0.16	63.54 ± 0.18	95.56 ± 0.03	14.55 ± 0.04