EVL-ECG: Efficient ECG Interpretation With Multi-Aspect Heterogeneous Knowledge Distillation

Nguyen Hong Dang [1 2] **Nhi Ngoc-Yen Nguyen** [2] **Huy-Hieu Pham** [2 3 4]

Abstract

High-fidelity ECG interpretation is increasingly reliant on massive foundation models, yet their deployment in clinical edge-care remains hindered by extreme computational demands. While knowledge distillation (KD) is a promising solution, traditional methods fail to capture the complex spatio-temporal dependencies of ECG signals when transferring knowledge across heterogeneous architectures. In this paper, we propose EVL-ECG, a framework specifically designed for cross-architecture distillation of cardiac diagnostic logic. EVL-ECG introduces three ECG-aware innovations: (1) Multi-Head Cross-Attention Alignment, which harmonizes architectural discrepancies to preserve fine-grained morphological features; (2) Optimal Transport-based Visual Feature Matching, utilizing optimal transport to maintain global structural relationships across ECG leads despite mismatched token representations; and (3) Geometric Intra-Architecture Relation Matching, which distills the latent diagnostic reasoning of the teacher model. Evaluations across ECG benchmarks demonstrate that EVL-ECG yields improvements of up to 2.4% AUC and 1.1% clinical accuracy over existing baselines. Notably, EVL-ECG establishes an efficient 2B-parameter ECG foundation model, suitable for resource-constrained clinical environments.

[1]Hanoi University of Science and Technology, Hanoi, Vietnam [2]VinUni-Illinois Smart Health Center, VinUniversity, Hanoi, Vietnam [3]College of Engineering and Computer Science, VinUniversity, Hanoi, Vietnam [4]Center for Innovations in Health Sciences, VinUniversity, Hanoi, Vietnam. Correspondence to: Huy-Hieu Pham <hieu.ph@vinuni.edu.vn>.

Proceedings of the ICML 2026 Workshop on Structured Data for Health, Seoul, South Korea. PMLR 306, 2026.

1. Introduction

Vision-Language Models (VLMs) have advanced Electrocardiogram (ECG) interpretation from simple classification to generative clinical reporting (Kaplan Berkaya et al., 2018; Rajpurkar et al., 2017; Khunte et al., 2024; Liu et al., 2024b; Lan et al., 2025; Zhao et al., 2025). However, their massive architectures prohibit real-time clinical deployment. While Knowledge Distillation (KD) offers a compression solution (Wang et al., 2020; Cai et al., 2024; Feng et al., 2024), distilling from a VLM teacher to a smaller student faces two critical barriers: severe vocabulary mismatch due to tokenizer heterogeneity (Gu et al., 2024; Feng et al., 2025), and unbalanced visual tokens caused by differing visual encoders.

Existing KD methods treat cross-tokenizer and visual-token mismatches in isolation (Truong et al., 2025; Wan et al., 2024; Boizard et al., 2025; Cui et al., 2024; Feng et al., 2025; 2024; Cai et al., 2024), restricting the use of modern, efficient small language models as backbones. To bridge this gap, we propose EVL-ECG, a Multi-Aspect Heterogeneous KD framework that employs Multi-Head Cross-Attention to adaptively aggregate dense teacher representations into a cohesive feature space. This architecture utilizes Optimal Transport regularization to maintain vital global spatial and morphological ECG characteristics through soft visual token alignment, while simultaneously applying Geometric Intra-Architecture Relation Matching to distill complex diagnostic logic via distance- and angle-wise potentials, ensuring the student model masters the intricate clinical patterns inherent in ECG interpretation. We also discuss the related works in Appendix A. Our main contributions are:

1. We propose EVL-ECG, a unified heterogeneous knowledge distillation framework that resolves tokenizer and visual-token mismatches between massive teachers and efficient students through a Multi-Head Cross-Attention alignment and Optimal Transport regularization, ensuring the preservation of critical morphological features in ECG images.
2. We introduce a Geometric Intra-Architecture Relation Matching module that distills the teacher's internal diagnostic logic via distance- and angle-wise potentials,

enabling the student model to inherit sophisticated reasoning patterns for complex cardiac analysis while maintaining a resource-friendly size.

3. Extensive evaluations across multiple ECG interpretation benchmarks demonstrate that EVL-ECG significantly outperforms state-of-the-art distillation methods and proprietary VLMs highlighting its robust clinical fidelity and strong generalizability to diverse diagnostic scenarios.

2. Method

This section details our proposed EVL-ECG framework in Figure 1, which facilitates multi-level heterogeneous knowledge distillation between a large teacher and an efficient student VLM.

2.1. Optimal Transport-based Visual Feature Matching

In ECG images, spatial layout intrinsically encodes clinical information such as the standard 12-lead grid where specific patches correspond to distinct cardiac axes. Standard point-wise matching ignores this global spatial topology. To preserve the teacher's collective geometry of visual descriptors, ensuring the student accurately maps the spatial progression of leads and waveforms without positional confusion, we employ OT regularization.

Let $\mathbf{t}_i, \mathbf{s}_j \in \mathbb{R}^{D_s}$ denote the L_2-normalized visual tokens of the teacher (linearly projected to the student's dimension) and student, respectively. We treat the V_t teacher and V_s student tokens as uniform empirical distributions: $\mu = \frac{1}{V_t}\sum_i \delta_{\mathbf{t}_i}$ and $\nu = \frac{1}{V_s}\sum_j \delta_{\mathbf{s}_j}$. Defining the transport cost as the squared Euclidean distance $C_{ij} = \|\mathbf{t}_i - \mathbf{s}_j\|_2^2$, we solve for the optimal transport plan $\mathbf{P}^*_\varepsilon \in \mathbb{R}_+^{V_t \times V_s}$ that minimizes the entropically regularized Sinkhorn distance:

$$\mathbf{P}^*_\varepsilon = \arg\min_{\mathbf{P}\in\Pi(\mu,\nu)} \langle \mathbf{P}, C\rangle - \varepsilon\mathcal{H}(\mathbf{P})$$

where $\Pi(\mu,\nu)$ is the set of valid couplings with marginals μ and ν, and $\mathcal{H}(\mathbf{P})$ is the entropy of the transport plan. The OT-based distillation loss is then computed as the expected cost under this optimal coupling:

$$\mathcal{L}_{ot} = \sum_{i=1}^{V_t}\sum_{j=1}^{V_s} P^*_{\varepsilon,ij}\|\mathbf{t}_i - \mathbf{s}_j\|_2^2$$

By minimizing $\mathcal{L}_{ot}$, the student aligns its visual feature map with the teacher's distribution, actively preventing the spatial misalignment of critical diagnostic regions such as misinterpreting precordial leads as limb leads.

2.2. Multi-Head Cross-Attention Alignment

To bridge the sequence length disparity between the teacher's dense representations $H_t \in \mathbb{R}^{B\times L_t\times D_t}$, which capture high-resolution temporal ECG morphologies, and the student's compact latent space $H_s \in \mathbb{R}^{B\times L_s\times D_s}$, we employ Multi-Head Cross-Attention (MHCA).

Treating the student's hidden states as queries and the teacher's states as keys and values allows the student to adaptively isolate and aggregate diagnostically relevant features, such as localized ectopic beats or subtle ST-segment deviations.

Rather than enforcing strict positional matching, the student dynamically attends to the most critical ECG segments across the sequence. The aligned teacher representation is obtained by computing

$$\hat{H}_t = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)W^O$$

where the student features H_s serve as the query and the teacher features H_t act as both key and value, optimal h detailed in Table 4. For each attention head, the operation is defined as

$$\text{head}_i = \text{Softmax}\left(\frac{(H_sW_i^Q)(H_tW_i^K)^\top}{\sqrt{d_k}}\right)(H_tW_i^V)$$

where W_i^Q, W_i^K, and W_i^V are learnable weight matrices. The final alignment loss ensures the student accurately reconstructs these clinical features by minimizing the mean squared error between its states and the attention-weighted teacher context:

$$\mathcal{L}_{\text{mhca}} = \frac{1}{B\cdot L_s}\sum_{b=1}^{B}\sum_{i=1}^{L_s}\|H_s^{(b,i)} - \hat{H}_t^{(b,i)}\|_2^2$$

Theoretical Motivation. We demonstrate that the attention mechanism mathematically can be interpreted as an Entropic Barycentric Projection under an OT framework. In ECG interpretation, enforcing a strict 1-to-1 alignment is ill-posed when sequences differ in length ($L_s \neq L_t$) or when clinically relevant waveforms such as P waves, QRS complexes, and ST segments are temporally shifted or vary across leads. Further theoretical details are provided in Appendix B.

2.3. Geometric Intra-Architecture Relation Matching

ECG interpretation relies fundamentally on the structural topology and temporal relationships between distinct waveform segments such as P-wave to QRS complex intervals, or ST-segment orientations. While point-wise alignment aids token reconstruction, it overlooks this global diagnostic logic. Inspired by relational KD (Park et al., 2019), we introduce a Geometric Intra-Architecture Relation Matching module to ensure the student maintains the teacher's internal geometric reasoning regarding the ECG signal's morphology.

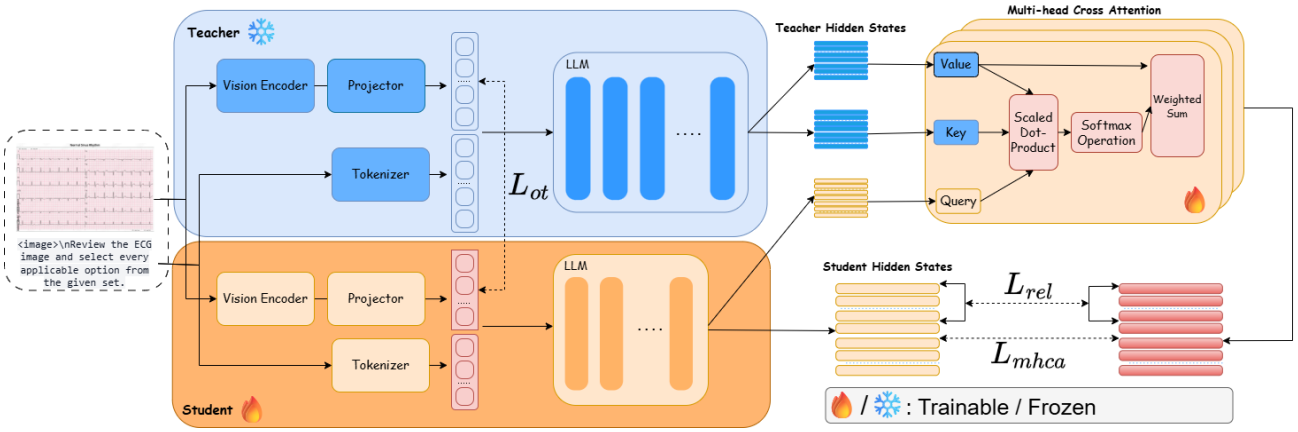

Figure 1. Overview of the proposed **EVL-ECG** distillation framework. Our approach is tailored to capture the complex temporal and spatial dependencies of ECG signals. The framework employs multi-head cross-attention to align heterogeneous ECG representations between a large-scale teacher and an efficient student. Furthermore, it integrates OT-based visual matching to preserve the global structural patterns of ECG leads, while Geometric Intra-Architecture Relation Matching distills the underlying diagnostic logic essential for cardiac arrhythmia detection.

We define two relation potentials for hidden state sequences H: a distance-wise potential ψ_{D} mean-normalized pairwise Euclidean distance to capture magnitude and interval-based signal relationships, and an angle-wise potential ψ_{A} pairwise cosine similarity to capture the directional morphology and electrical axis of the ECG features. The distillation loss aligns these structural potentials between the projected teacher $\hat{H}_t$ and student H_s across batch B and sequence length L_s:

$$\mathcal{L}_k = \frac{1}{B \cdot L_s^2} \sum_{b=1}^{B} \sum_{i,j=1}^{L_s} \Big\| \psi_k(\hat{H}_t^{(b,i)}, \hat{H}_t^{(b,j)}) - \psi_k(H_s^{(b,i)}, H_s^{(b,j)}) \Big\|^2 \quad (1)$$

where $k \in \{\mathrm{D}, \mathrm{A}\}$ denotes the distance and angle metrics. The total relational loss, $\mathcal{L}_{\mathrm{rel}} = \frac{1}{2}(\mathcal{L}_{\mathrm{D}} + \mathcal{L}_{\mathrm{A}})$, acts as a structural regularizer. By matching this geometric manifold, the student learns the teacher's clinical clustering, which is vital for recognizing complex, multi-segment cardiac abnormalities. To further illustrate our findings, we include visualizations in Appendix D. We also provide clinical insights and limitations of our method in Appendix F.

2.4. Total Distillation Objective

The final objective function integrates the cross-architecture alignment and the geometric constraints. By combining the MHCA-based reconstruction with relation matching, the student model effectively inherits both the fine-grained diagnostic features and the global reasoning patterns of the teacher model:

$$\mathcal{L}_{total} = (1-\alpha)\mathcal{L}_{CE} + \alpha \cdot (\lambda_m \mathcal{L}_{mhca} + \lambda_r \mathcal{L}_{rel} + \lambda_{ot} \mathcal{L}_{ot})$$

Best hyper-parameters after the searching process are provided in Table 4 from Appendix C.1.

3. Experiments

3.1. Datasets and Metrics

The training of EVL-ECG leverages a large-scale dataset, which is ECGInstruct from PULSE (Liu et al., 2024b) with 1,156,110 conversations. This data comprises four primary sources: PTB-XL (Wagner et al., 2022), ECG-QA (Oh et al., 2023), MIMIC-IV-ECG (Gow et al., 2023), and CODE-15% (Ribeiro et al., 2021).

To evaluate the clinical diagnostic capabilities of EVL-ECG, we utilize ECG-Bench (Liu et al., 2024b), a standardized dataset designed for the assessment of VLMs in the cardiac domain, which compresses from multiple sources. Following PULSE (Liu et al., 2024b), we utilize multi-label classification metrics, including Macro AUC, Macro F1, and Hamming Loss, to evaluate the datasets PTB-XL Super, CODE-15%, and CPSC-2018 (Ng et al., 2018), where multiple correct labels may exist. For the ECG-QA, CSN (Zheng et al., 2020), MMMU-ECG (Liu et al., 2024b) and G12EC (Ng et al., 2018) datasets, we adopt accuracy as the evaluation metric. Evaluation datasets statistics are reported in Appendix C.2.

3.2. Baselines

We compare our method against established methods including other knowledge distillation methods, domain-specific methods and MLLMs with different size. We consider three proprietary MLLMs to establish a high-level performance benchmark: GPT-4o (OpenAI, 2024), Gemini 1.5 Pro (Reid

Table 1. Comparison of EVL-ECG with state-of-the-art domain-specific methods, proprietary MLLMs, and open-source MLLMs across ECG benchmarks. ↓ indicates that lower values correspond to superior model performance.

Datasets	**PTB-XL-Super**			**CODE-15%**			**ECG-QA**	**CPSC-2018**			**CSN**	**G12EC**	**MMMU-ECG**
Metric	AUC	F1	HL ↓	AUC	F1	HL ↓	Accuracy	AUC	F1	HL ↓	Accuracy	Accuracy	Accuracy
Random	50.3	33.2	50.1	48.8	15.0	32.1	16.2	51.2	15.1	28.8	11.6	12.1	24.2
Proprietary VLMs													
GPT-4o	55.6	28.3	26.2	59.9	24.9	15.7	35.2	50.9	10.6	18.2	57.5	49.2	43.5
Gemini 1.5 Pro	50.7	15.3	27.9	56.7	20.0	15.9	33.2	50.1	7.4	20.5	50.5	36.0	40.0
Claude 3.5 Sonnet	54.0	27.5	29.6	58.3	20.3	17.8	34.2	52.8	11.5	18.9	51.5	51.4	42.0
Open-source VLMs													
LLaVA-Med	50.0	12.3	28.1	69.2	27.0	33.4	29.5	50.0	2.5	20.2	13.8	14.1	27.0
LLaVA-OneVision-7B	49.8	11.4	34.5	58.7	17.0	20.6	20.4	49.6	8.0	28.3	23.3	25.7	26.0
LLaVA-1.6-Vicuna-13B	50.0	20.1	38.3	53.0	3.6	16.6	22.0	50.0	19.3	62.8	31.4	35.0	38.0
LLaVA-1.6-34B	50.2	19.9	36.0	57.2	12.8	16.6	22.4	49.6	19.3	62.8	44.3	45.9	31.0
Domain-specific VLMs/KD Methods													
ECG-GPT	69.5	53.9	20.1	68.9	40.1	17.4	N/A	**69.3**	**44.0**	**9.9**	N/A	N/A	N/A
SFT	74.6	62.8	16.3	85.0	76.7	7.0	63.7	66.4	32.7	10.8	89.3	76.4	47.7
ULD (Boizard et al., 2025)	73.4	61.7	16.9	84.3	76.4	7.2	62.5	65.5	28.2	11.6	86.3	75.6	46.5
MultiLevelOT (Cui et al., 2024)	72.3	59.4	17.1	85.6	74.7	7.4	63.1	66.6	29.1	12.1	87.9	76.1	44.5
DSKD (Zhang et al., 2024)	73.0	60.3	18.8	80.8	75.4	8.0	60.5	65.3	31.3	12.1	88.4	75.8	46.5
EM-KD (Feng et al., 2025)	74.3	62.9	16.4	84.9	77.4	6.8	63.9	66.3	32.9	11.2	89.2	76.0	47.4
EVL-ECG (Ours)	**75.2**	**63.3**	**16.0**	**86.4**	**78.4**	**6.5**	**64.8**	66.8	33.6	12.3	**89.7**	**77.1**	**48.5**

et al., 2024), and Claude 3.5 Sonnet (Anthropic, 2024). For open-source VLMs, we select a range of open-source models to ensure comprehensive coverage across different visual components. These include the general-purpose LLaVA-1.6 (Liu et al., 2024a), LLaVA-OneVision-7B (Li et al., 2024), and the domain-specific LLaVA-Med (Li et al., 2023). With domain-specific methods and KD baselines, we evaluate our model against ECG-GPT (Khunte et al., 2024), supervised fine-tuning (SFT) and other state-of-the-art KD frameworks, which are ULD (Boizard et al., 2025), MultiLevel-OT (Cui et al., 2024), DSKD (Zhang et al., 2024) and EM-KD (Feng et al., 2025). More implementation details are reported in the Appendix C.1 due to page limitation.

3.3. Main Results

EVL-ECG demonstrates a clear performance advantage over both high-tier proprietary models and open-source VLMs across benchmarks, as detailed in Table 1. While general-purpose models struggle with intricate temporal ECG patterns, evidenced by EVL-ECG's 75.2 AUC on PTB-XL-Super compared to GPT-4o's 55.6, our model also introduces a unified knowledge distillation framework that simultaneously resolves tokenizer and visual token mismatches. This framework allows EVL-ECG to consistently outperform established KD baselines, while maintaining greater cross-dataset versatility than specialized models like ECG-GPT. Beyond aggregate metrics, we observe that EVL-ECG yields more consistent predictions across correlated leads, indicating improved modeling of inter-lead dependencies that are critical for clinical diagnosis. In particular, the model shows enhanced sensitivity to subtle morphological variations, such as ST-segment deviations and irregular QRS complexes, which are commonly associated with clinically significant cardiac abnormalities. We also report additional experiments and loss components contributions in table 6 from Appendix E respectively.

4. Conclusion and Future Works

This work presents EVL-ECG, an efficient Vision-Language Model achieving state-of-the-art performance in ECG interpretation through a specialized hierarchical distillation framework. By bridging heterogeneous architectures via the MHCA module and ensuring structural integrity with OT-based and geometric matching, our model captures the complex diagnostic logic required for high-fidelity clinical accuracy. This enables the deployment of sophisticated reasoning patterns in resource-constrained medical environments. Future research will focus on enhancing model robustness against real-world clinical noise, such as baseline wander and electrode artifacts, to ensure reliability in ambulatory settings.

Acknowledgment

This work was funded by Vingroup Joint Stock Company (Vingroup JSC), Vingroup, and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2021.DA00128.

References

Anthropic. Claude 3.5 sonnet, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet.

Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., and Zhu, K. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Boizard, N., Haddad, K. E., HUDELOT, C., and Colombo, P. Towards cross-tokenizer distillation: the universal logit distillation loss for LLMs. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=bwRxXiGO9A.

Cai, Y., Goswami, M., Choudhry, A., Srinivasan, A., and Dubrawski, A. JoLT: Jointly learned representations of language and time-series. In *Deep Generative Models for Health Workshop NeurIPS 2023*, 2023. URL https://openreview.net/forum?id=UVF1AMBj9u.

Cai, Y., Zhang, J., He, H., He, X., Tong, A., Gan, Z., Wang, C., and Bai, X. Llava-kd: A framework of distilling multimodal large language models. *arXiv preprint arXiv:2410.16236*, 2024.

Cui, X., Zhu, M., Qin, Y., Xie, L., Zhou, W., and Li, H. Multi-level optimal transport for universal cross-tokenizer knowledge distillation on language models. *arXiv preprint arXiv:2412.14528*, 2024.

Feng, Q., Li, W., Lin, T., and Chen, X. Align-kd: Distilling cross-modal alignment knowledge for mobile vision-language model. *arXiv preprint arXiv:2412.01282*, 2024. URL https://arxiv.org/abs/2412.01282.

Feng, Z., Yang, S., Duan, B., Yang, W., and Wang, J. Emkd:distilling efficient multimodal large language model with unbalanced vision tokens. In *Association for the Advancement of Artificial Intelligence*, 2025.

Gow, B., Pollard, T., Nathanson, L. A., Johnson, A., Moody, B., Fernandes, C., Greenbaum, N., Waks, J. W., Eslami, P., Carbonati, T., Chaudhari, A., Herbst, E., Moukheiber, D., Berkowitz, S., Mark, R., and Horng, S. MIMIC-IV-ECG: Diagnostic Electrocardiogram Matched Subset. *PhysioNet*, 2023. URL https://doi.org/10.13026/4nqg-sb35.

Gu, Y., Dong, L., Wei, F., and Huang, M. Minillm: Knowledge distillation of large language models. In *Proceedings of ICLR*, 2024.

He, K., Mao, R., Lin, Q., Ruan, Y., Lan, X., Feng, M., and Cambria, E. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Inf. Fusion*, 118 (C), June 2025. ISSN 1566-2535. doi: 10.1016/j.inffus.2025.102963. URL https://doi.org/10.1016/j.inffus.2025.102963.

Hinton, G., Vinyals, O., and Dean, J. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Kaplan Berkaya, S., Uysal, A. K., Sora Gunal, E., Ergin, S., Gunal, S., and Gulmezoglu, M. B. A survey on ecg analysis. *Biomedical Signal Processing and Control*, 43:216–235, 2018. ISSN 1746-8094. URL https://www.sciencedirect.com/science/article/pii/S1746809418300636.

Khunte, A., Sangha, V., Oikonomou, E., Dhingra, L., Aminorroaya, A., Coppi, A., Shankar, S., Mortazavi, B., Bhatt, D., Krumholz, H., Nadkarni, G., Vaid, A., and Khera, R. Automated diagnostic reports from images of electrocardiograms at the point-of-care. *medRxiv : the preprint server for health sciences*, 02 2024.

Lan, X., Wu, F., He, K., Zhao, Q., Hong, S., and Feng, M. Gem: Empowering mllm for grounded ecg understanding with time series and images. *arXiv preprint arXiv:2503.06073*, 2025.

Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Li, Y., Liu, Z., and Li, C. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., and Gao, J. Llava-med: training a large language-and-vision assistant for biomedicine in one day. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2023.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024a. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Liu, R., Bai, Y., Yue, X., and Zhang, P. Teach multimodal llms to comprehend electrocardiographic images. *arXiv preprint arXiv:2410.19008*, 2024b. URL https://arxiv.org/abs/2410.19008.

Ng, E. Y. K., Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., and Li, J. An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. *Journal of Medical Imaging and Health Informatics*, 2018. URL https://api.semanticscholar.org/CorpusID:70024401.

Oh, J., Lee, G., Bae, S., Kwon, J.-m., and Choi, E. Ecg-qa: A comprehensive question answering dataset combined with electrocardiogram. *Advances in Neural Information Processing Systems*, 36:66277–66288, 2023.

OpenAI. Gpt-4o contributions, 2024. URL https://openai.com/gpt-4o-contributions/.

Park, W., Kim, D., Lu, Y., and Cho, M. Relational knowledge distillation. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Rajpurkar, P., Hannun, A. Y., Haghpanahi, M., Bourn, C., and Ng, A. Y. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836*, 2017. URL https://arxiv.org/abs/1707.01836.

Reid, M., Savinov, N., Teplyashin, D., et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Ribeiro, A. H., Paixao, G. M., Lima, E. M., Horta Ribeiro, M., Pinto Filho, M. M., Gomes, P. R., Oliveira, D. M., Meira Jr, W., Schon, T. B., and Ribeiro, A. L. P. Code-15%: a large scale annotated dataset of 12-lead ecgs, June 2021. URL https://doi.org/10.5281/zenodo.4916206.

Truong, M.-P., Vu, H. A., Vu, T., Diep, N. T. N., Van, L. N., Nguyen, T. H., and Le, T. Emo: Embedding model distillation via intra-model relation and optimal transport alignments. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 2025. URL https://aclanthology.org/2025.emnlp-main.385/.

Wagner, P., Strodthoff, N., Bousseljot, R.-D., Samek, W., and Schaeffter, T. PTB-XL, a large publicly available electrocardiography dataset. *PhysioNet*, 2022. URL https://doi.org/10.13026/kfzx-aw45.

Wan, F., Huang, X., Cai, D., Quan, X., Bi, W., and Shi, S. Knowledge fusion of large language models. In *International Conference on Learning Representations*, 2024.

Wan, Z., Liu, C., Wang, X., Tao, C., Shen, H., Xiong, J., Arcucci, R., Yao, H., and Zhang, M. MEIT: Multimodal electrocardiogram instruction tuning on large language models for report generation. In Che, W., Nabende, J., Shutova, E., and Pilehvar, M. T. (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 14510–14527, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.749. URL https://aclanthology.org/2025.findings-acl.749/.

Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., and Zhou, M. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, 2020.

Yang, K., Hong, M., Zhang, J., Luo, Y., Zhao, S., Zhang, O., Yu, X., Zhou, J., Yang, L., Zhang, P., Qiao, M., and Nie, Z. Ecg-lm: Understanding electrocardiogram with a large language model. *Health Data Science*, 5:0221, 2025. doi: 10.34133/hds.0221. URL https://spj.science.org/doi/abs/10.34133/hds.0221.

Yu, H., Guo, P., and Sano, A. Zero-shot ecg diagnosis with large language models and retrieval-augmented generation. In Hegselmann, S., Parziale, A., Shanmugam, D., Tang, S., Asiedu, M. N., Chang, S., Hartvigsen, T., and Singh, H. (eds.), *Proceedings of the 3rd Machine Learning for Health Symposium*, volume 225 of *Proceedings of Machine Learning Research*, pp. 650–663. PMLR, 10 Dec 2023. URL https://proceedings.mlr.press/v225/yu23b.html.

Zhang, S., Zhang, X., Sun, Z., Chen, Y., and Xu, J. Dual-space knowledge distillation for large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 18164–18181. Association for Computational Linguistics, November 2024. URL https://aclanthology.org/2024.emnlp-main.1010/.

Zhao, Y., Kang, J., Zhang, T., Han, P., and Chen, T. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11209476.

Zheng, J., Chu, H., Struppa, D. C., Zhang, J., Yacoub, S. M., El-Askary, H. M., Chang, A., Ehwerhemuepha, L., Abudayyeh, I., Barrett, A., Fu, G., Yao, H., Li, D., Guo, H., and Rakovski, C. Optimal multi-stage arrhythmia classification approach. *Scientific Reports*, 10, 2020. URL https://api.semanticscholar.org/CorpusID:211170003.

A. Related Works

A.1. Multimodal Large Language Models

The rapid evolution of MLLMs, such as the Qwen-VL series (Bai et al., 2025) and LLaVA-OneVision (Li et al., 2024), has expanded LLMs' sensory scope by using learnable projectors to translate visual features into linguistic workspaces. This multimodality is particularly critical in healthcare (He et al., 2025), where clinical decision-making inherently relies on integrating multidimensional data like medical imaging, physiological time-series, and narrative reports (Yang et al., 2025).

A.2. Knowledge Distillation

Knowledge distillation (KD) compresses models by transferring capabilities from a teacher to a student (Hinton et al., 2015). For VLMs, KD strategies primarily divide into LLM-style and MLLM-style. LLM-style approaches refine text generation and reduce hallucinations (Gu et al., 2024), with recent advances utilizing Optimal Transport for cross-tokenizer and sequence alignment (Boizard et al., 2025; Cui et al., 2024; Wan et al., 2024; Zhang et al., 2024). Conversely, MLLM-style methods prioritize visual feature and cross-modal consistency (Cai et al., 2024; Feng et al., 2024). Notably, frameworks like EM-KD (Feng et al., 2025) address general distillation scenarios with unbalanced vision tokens.

A.3. ECG Analysis

AI-driven ECG interpretation has shifted from retrieval-augmented zero-shot models (Yu et al., 2023) toward complex multimodal alignment. Frameworks like JoLT (Cai et al., 2023) and ECG-CoCa (Zhao et al., 2025) synchronize 1D ECG signals with text, whereas PULSE (Liu et al., 2024b) synthesizes ECG images to leverage VLM architectures. Recent developments further incorporate instruction-tuning for clinical report generation (Wan et al., 2025) and unified architectures like GEM (Lan et al., 2025), which process temporal signals and visual waveforms simultaneously to mimic a cardiologist's workflow.

B. Theoretical Motivation of MHCA

Let the student queries $Q_m^{(i)}$ and teacher keys $K_n^{(i)}$ represent empirical samples from latent feature distributions. For a given head i, define the transport cost between the m-th student token and n-th teacher token as the negative inner product $C_{mn}^{(i)} = -Q_m^{(i)}(K_n^{(i)})^\top$. Then, the cross-attention mechanism computes the optimal conditional transport plan π^* that minimizes the Entropic Optimal Transport objective, and the resulting context vector head$_m^{(i)}$ is the exact Barycentric Projection of the teacher's value space onto the student's coordinate system.

Consider the m-th student token. We seek a probability distribution or conditional transport plan $\pi_m \in \Delta^{L_t-1}$ over the teacher tokens $n \in \{1, \dots, L_t\}$ that minimizes the expected transport cost from the student to the teacher, while maintaining a degree of smoothness. We formulate this as an entropy-regularized optimization problem: $\pi_m^* = \arg\min_{\pi_m \in \Delta^{L_t-1}} \left(\sum_{n=1}^{L_t} \pi_{mn} C_{mn}^{(i)} - \epsilon \mathcal{H}(\pi_m) \right)$ where Δ^{L_t-1} is the (L_t-1)-dimensional probability simplex, $C_{mn}^{(i)} = -Q_m^{(i)}(K_n^{(i)})^\top$ is the geometric cost function, $\mathcal{H}(\pi_m) = -\sum_n \pi_{mn} \log \pi_{mn}$ is the Shannon entropy acting as a regularizer, and $\epsilon > 0$ is the regularization coefficient. To solve this constrained optimization problem, we introduce a Lagrange multiplier λ for the simplex constraint $\sum_{n=1}^{L_t} \pi_{mn} = 1$:

$$\mathcal{L}(\pi_m, \lambda) = \sum_{n=1}^{L_t} \pi_{mn} C_{mn}^{(i)} + \epsilon \sum_{n=1}^{L_t} \pi_{mn} \log \pi_{mn} + \lambda \left(\sum_{n=1}^{L_t} \pi_{mn} - 1 \right) \tag{2}$$

Taking the partial derivative with respect to π_{mn} and setting it to zero yields:

$$\frac{\partial \mathcal{L}}{\partial \pi_{mn}} = C_{mn}^{(i)} + \epsilon(1 + \log \pi_{mn}) + \lambda = 0 \tag{3}$$

Solving for π_{mn}, we obtain $\pi_{mn}^* = \exp\left(-\frac{C_{mn}^{(i)}}{\epsilon} - \frac{\lambda}{\epsilon} - 1\right) \propto \exp\left(-\frac{C_{mn}^{(i)}}{\epsilon}\right)$ and applying the sum-to-one constraint $\sum_j \pi_{mj}^* = 1$ to find the normalizing constant gives:

$$\pi_{mn}^* = \frac{\exp\left(-C_{mn}^{(i)}/\epsilon\right)}{\sum_{j=1}^{L_t} \exp\left(-C_{mj}^{(i)}/\epsilon\right)}$$

By substituting our defined cost $C_{mn}^{(i)} = -Q_m^{(i)}(K_n^{(i)})^\top$ and setting the entropy regularization coefficient to $\epsilon = \sqrt{d_k}$, we exactly recover the attention weights from the original formulation:

$$\pi_{mn}^* = \frac{\exp\left(\frac{Q_m^{(i)}(K_n^{(i)})^\top}{\sqrt{d_k}}\right)}{\sum_{j=1}^{L_t}\exp\left(\frac{Q_m^{(i)}(K_j^{(i)})^\top}{\sqrt{d_k}}\right)} \equiv w_{mn}^{(i)}$$

In optimal transport, once the optimal plan π^* is found, mapping the source distribution (teacher) to the target support (student) in a metric space is achieved via the barycentric projection. The projection of the teacher's features as values $V^{(i)}$ onto the m-th student token is defined as the expectation under the optimal transport plan:

$$\mathcal{P}_{\pi^*}(V^{(i)})_m = \mathbb{E}_{n\sim\pi_m^*}\left[V_n^{(i)}\right] = \sum_{n=1}^{L_t}\pi_{mn}^* V_n^{(i)} = \sum_{n=1}^{L_t} w_{mn}^{(i)} V_n^{(i)} = \text{head}_m^{(i)}$$

Thus, our final loss function $\mathcal{L}_{mhca}$ is mathematically equivalent to minimizing the L_2 distance between the student's representation and the optimal entropic barycentric projection of the teacher's representation.

C. Experiment Details

C.1. Implementation Details

We employ **Qwen3-VL-2B-Instruct** (Bai et al., 2025) as the student and **PULSE-7B** (Liu et al., 2024b) as the teacher (details in Tables 2 and 3). Training involves two phases: (1) **Supervised fine-tuning** for 54,195 steps with batch size 64 for domain adaptation, establishing a baseline checkpoint; and (2) **Knowledge Distillation** applied to this initialization. All baseline KD methods are trained in 1 epoch with **three different runs** and reported average score, start from this same checkpoint for fair comparison. We use **full-model fine-tuning** across both phases to ensure the student thoroughly learns complex ECG diagnostic patterns. Models are implemented in **PyTorch** (Python 3.11) and trained on a single **NVIDIA H100 (80GB)** using `bf16` mixed-precision.

Table 2. Training configurations for the SFT phase.

Settings	**Qwen3-VL-2B-Instruct**
Steps	54195
Learning Rate	2×10^{-4}
Projector LR	5×10^{-4}
Global Batch Size	64
BF16 Setting	True
LR Scheduler	Cosine
Warmup Ratio	0.03
Weight Decay	0.01
Model Max Length	4096
Min Pixels/Max Pixels	784/50176

Hyperparameters We fixed the Sinkhorn regularization (λ) parameter to 0.1, a widely adopted standard value in Optimal Transport implementations to ensure numerical stability. The weight hyper-parameters are selected from the following ranges: $\alpha \in [0.1, 0.3, 0.5, 0.7, 1.0]$, $\lambda_m \in [0.1, 0.3, 0.5, 0.7, 1.0]$, $\lambda_r \in [0.01, 0.03, 0.05, 0.07]$, $\lambda_{ot} \in [0.1, 0.3, 0.5]$ and $h \in [4, 8, 16]$.

Knowledge Distillation Baselines To the best of our knowledge, our work is the first unified framework to simultaneously address the dual alignment challenge: the discrepancy in both text tokenizers and visual token counts in VLMs. Consequently, we select and adapt state-of-the-art baselines from both cross-tokenizer LLM distillation and cross-visual token VLM distillation to compare with our framework, which are DSKD (Zhang et al., 2024), ULD (Boizard et al., 2025), MultiLevelOT (Cui et al., 2024) and EM-KD (Feng et al., 2025).

Table 3. Training configurations for the knowledge distillation phase, also for all the KD baseline methods.

Settings	**Qwen3-VL-2B-Instruct**
Epoch	1
Learning Rate	2×10^{-4}
Projector LR	5×10^{-4}
Global Batch Size	48
BF16 Setting	True
LR Scheduler	Cosine
Warmup Ratio	0.03
Weight Decay	0.01
Model Max Length	4096
Min Pixels/Max Pixels	784/50176

Table 4. Optimal hyperparameters for our proposed framework.

Hyperparameters	α	λ_m	λ_r	λ_{ot}	h
Best Values	0.3	1.0	0.03	0.1	8

C.2. Datasets

Table 5. Overview of evaluation datasets in ECGBench. This collection contains both in-domain and out-of-domain problems across tasks with diverse answer types.

Evaluation Dataset	**Task**	**Type**	**# Samples**	**In-Domain?**
PTB-XL Super	Abnormality Detection	Close-ended	2,082	YES
CODE-15%	Abnormality Detection	Close-ended	1,400	YES
ECG-QA	Abnormality Detection	Close-ended	1,317	YES
CPSC 2018	Abnormality Detection	Close-ended	2,061	NO
CSN	Abnormality Detection	MCQ (8-option)	1,611	NO
G12EC	Abnormality Detection	MCQ (8-option)	2,026	NO
MMMU ECG	Multimodal Understanding	MCQ (4-option)	200	NO

For the training dataset, we utilize ECGInstruct as we mentioned in the previous sections. To evaluate the clinical diagnostic capabilities of EVL-ECG, we utilize ECGBench, which contains both repurposed tasks from different existing datasets. Table 5 shows the details of each evaluation dataset.

D. Visualizations

Figure 2 visualizes the Sinkhorn Transport Plan for visual feature alignment. By minimizing this transport cost, the student model is forced to accurately replicate the teacher's precise spatial understanding of ECG topology. This ensures that critical clinical features, such as specific lead placements and localized PQRST waveforms, are correctly mapped without spatial misalignment, facilitating robust diagnostic knowledge transfer.

Singular Value Decomposition of the feature matrices reveals a rapid decay for both models, indicating that core ECG diagnostic patterns which are primary waveforms and cardiac axes occupy a low-rank subspace. While the teacher's higher singular values reflect a richer capture of subtle clinical nuances, the structurally similar decay curves confirm the student effectively learns to prioritize the same critical cardiac features.

E. Analysis

Impact of Loss Components. The ablation study presented in Table 6 quantifies the incremental performance gains provided by each proposed loss component across three distinct ECG benchmarks. Starting from a baseline of 85.0% AUC on CODE-15% and 63.7% accuracy on ECG-QA, the introduction of the OT loss ($\mathcal{L}_{ot}$) establishes a foundation for

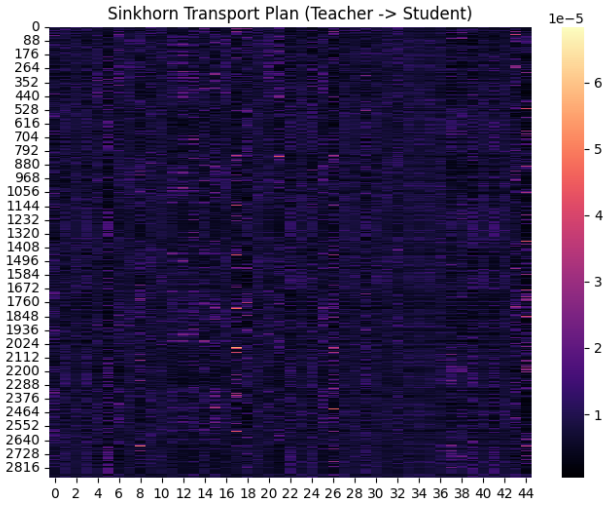

Figure 2. Heatmap showing the Sinkhorn Transport Plan between teacher and student features. The intensity of each cell represents the transport cost between the corresponding feature representations, with brighter colors indicating higher similarity and lower costs.

Table 6. Ablation study of proposed loss components across different ECG benchmarks.

Method	$\mathcal{L}_{ot}$	$\mathcal{L}_{mhca}$	$\mathcal{L}_{rel}$	**CODE-15%**			**ECG-QA**	**G12EC**
				AUC	F1	HL	Acc	Acc
Baseline				85.0	76.7	7.0	63.7	76.4
Components	✓			85.4	77.1	7.1	64.1	76.7
	✓	✓		85.9	77.8	6.8	64.5	76.8
	✓	✓	✓	**86.4**	**78.4**	**6.5**	**64.8**	**77.1**

geometric alignment, marginally improving most metrics. A more substantial performance leap occurs with the integration of the MHCA reconstruction loss ($\mathcal{L}_{mhca}$), which elevates the CODE-15% AUC to 85.9 and G12EC accuracy to 76.8%, underscoring its efficacy in aligning cross-architecture feature representations. The full configuration, which incorporates the relational matching loss ($\mathcal{L}_{rel}$), achieves the superior result in every category, reaching peak values of 86.4 AUC, 78.4 F1, and a record-low 6.5 Hamming Loss on CODE-15%, while also maximizing accuracy on ECG-QA and G12EC.

Computational Analysis. We evaluate the computational efficiency of the proposed EVL-ECG method by comparing its per-step training time and peak GPU memory (VRAM) usage against different KD method baselines. As summarized in Table 7, EVL-ECG maintains a competitive computational footprint, requiring only 8.54 seconds per step and 56.4 GB of VRAM. While slightly higher than the baseline ULD (8.37s, 55.7 GB), EVL-ECG remains significantly more efficient than EM-KD, which exhibits the highest overhead at 9.13s and 77.4 GB. These results demonstrate that EVL-ECG provides an effective trade-off, achieving its performance gains without necessitating the prohibitive hardware requirements or extended training durations seen in more complex distillation frameworks.

Table 7. Computation time and GPU memory consumption of different KD methods.

Method	**ULD**	**MultiLevelOT**	**DSKD**	**EM-KD**	**EVL-ECG**
Time/step (s)	8.37	8.64	8.48	9.13	8.54
VRAM (GB)	55.7	58.1	57.4	77.4	56.4

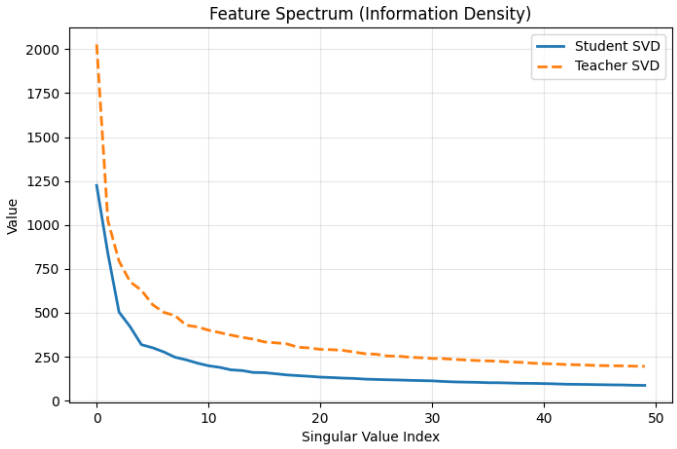

Figure 3. Comparison of the Feature Spectrum via Singular Value Decomposition (SVD). The plot depicts the singular value decay for both Student and Teacher models, reflecting the information density and rank distribution of their respective feature representations.

F. Clinical Insights and Limitations

Clinical Insights. EVL-ECG is designed to preserve clinically meaningful structure in ECG interpretation rather than merely improve aggregate prediction scores. In particular, the combination of cross-attention alignment, optimal transport matching, and relational distillation helps the student model retain inter-lead consistency and waveform geometry, which are essential for recognizing patterns such as ST-segment deviation, QRS morphology changes, and rhythm irregularities. This is especially valuable in ECG analysis because many clinically important abnormalities are expressed through subtle temporal shifts or lead-wise correlations that can be missed by compact models trained with point-wise supervision alone.

Clinical Limitations. Despite these advantages, EVL-ECG remains a decision-support model rather than a clinical diagnostic system. Its performance is evaluated on curated benchmarks, and real-world ECGs may contain noisy signals, missing leads, acquisition artifacts, or patient-specific confounders that can reduce reliability. In addition, while the model improves structural alignment, it does not provide explicit causal explanations or guaranteed uncertainty calibration for every prediction. Therefore, deployment in clinical settings should be accompanied by clinician oversight, external validation on hospital-specific data, and careful assessment under distribution shift.