△ DELTA: Language Diffusion-based EEG-to-Text Architecture

MinGyu Jeon[*]
MODULABS
jkmcoma7@gmail.com

HyoBin Kim[*]
Sungkyunkwan University
hyobinkim@gmail.com

Abstract

Electroencephalogram (EEG)-to-text remains challenging due to high-dimensional noise, subject variability, and error accumulation in autoregressive decoding. We introduce DELTA, which pairs a Residual Vector Quantization (RVQ) EEG tokenizer with a masked language diffusion model (LLaDA). RVQ discretizes continuous EEG into multi-layer tokens to reduce noise and individual differences, while LLaDA reconstructs sentences via non-sequential denoising. On ZuCo, DELTA improves semantic alignment by up to 5.37 points over autoregressive baselines, achieving BLEU-1 21.9 and ROUGE-1 F 17.2 under word-level conditions. These results enable reliable text generation from small EEG-text datasets and point toward scalable multimodal EEG-language models.

Keywords: EEG-to-Text, Residual Vector Quantization, Language Diffusion Model, Discrete Tokenization, Multimodal Brain-Language Learning

1 Introduction

Brain-Computer Interfaces (BCIs) offer a vital communication channel for individuals with severe neuromuscular disorders by translating brain signals into commands for external devicesChaudhary et al. [2016]. While early BCI research relied on invasive methods Metzger et al. [2023], the field shifted to non-invasive EEG Moses et al. [2021], initially focusing on classification tasks Autthasan et al. [2024]. More recently, Large Language Models (LLMs) have driven significant progress in EEG-to-Text translation Wang and Ji [2022], Duan et al. [2023], Mishra et al. [2024].

Despite these advancements, existing EEG-to-Text models face significant challenges in terms of reliability. The teacher-forcing technique used in the evaluation process of many studies can overestimate a model's true performance Jo et al. [2024], and performance measured without teacher-forcing suggests that current models have not yet sufficiently overcome the inherent low signal-to-noise ratio (SNR) problem of EEG data Jiang et al. [2019].

The limitations of previous research are evident from two perspectives: signal processing and generative models. First, from a signal processing standpoint, continuous EEG signals are inherently unstable. Early studies opened new possibilities with an Open-Vocabulary approach that directly mapped EEG waveforms to language models Wang and Ji [2022]. However, due to the high inter-subject variability and extreme noise characteristic of EEG, such direct mapping methods can yield results akin to processing random values rather than extracting meaningful information Jo et al. [2024]. Second, from a generative model perspective, conventional Autoregressive (AR) methods are vulnerable to error accumulation. AR models like BART generate text sequentially, using the token generated in the previous step as input for the next Lewis et al. [2019], Raffel et al. [2020], Zhang et al. [2020]. This structure is susceptible to "error accumulation," where a small initial error can

[*] These authors contributed equally to this work.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Foundation Models for the Brain and Body.

cascade and severely degrade the quality of the entire output Wang et al. [2022]. This issue becomes a critical limitation, especially in environments with noisy EEG signals, where the probability of inferring an incorrect token is higher.

To address the instability in signal processing and the error accumulation in autoregressive models, this paper proposes a novel EEG-to-Text framework. The proposed model integrates an RVQ (Residual Vector Quantization)-based EEG tokenizer with a Diffusion Model-based non-autoregressive language model. First, to tackle the high noise and inter-subject variability of EEG, we transform the continuous waveforms into a robust discrete representation using RVQ. The hierarchical quantization process of RVQ sequentially encodes the signal from its core features, effectively filtering out noise and establishing a stable foundation for extracting consistent semantic information from unstable brainwaves Zeghidour et al. [2021], Défossez et al. [2022]. Next, to fundamentally solve the error accumulation problem of AR models, we introduce the concept of restoration from diffusion models into the text generation process Ho et al. [2020], Austin et al. [2021]. Instead of generating text sequentially, this approach restores the final output by progressively denoising the entire sentence structure. This inherently prevents the sequential error propagation where an error in one step directly affects the next, enabling stable sentence generation even from noisy EEG inputs Nie et al. [2025].

The key contributions of this study are as follows: Methodological Innovation: We propose the first EEG-to-Text framework that combines an RVQ tokenizer and a diffusion-based language model, addressing challenges in both signal processing and text generation. Paradigm Shift: We shift the paradigm from the conventional direct translation approach of 'EEG-to-Text' to a restoration approach, significantly improving the stability of the generation process. Performance and Robustness Validation: We demonstrate that the proposed model achieves significant performance improvements over existing SOTA models under conditions identical to real-world inference, without relying on teacher-forcing.

2 Method

In this section, we describe the construction and training of our proposed EEG-to-Text generation framework. The framework consists of two stages: (1) RVQ-based EEG tokenizer that converts EEG signals into discrete tokens, and (2) LLaDA-based language diffusion model that generates text given the discrete tokens. The two modules are trained in stages and finally combined to generate sentences from EEG signals. Figure 1 shows the overall structure schematically.

2.1 Stage 1: RVQ-based EEG Tokenizer

2.1.1 EEG Feature Extraction and Quantization

The EEG tokenizer takes continuous, multichannel EEG signals as input, extracts latent features, and quantizes them into a sequence of discrete tokens using the RVQ. In this study, we constructed an extended input tensor of 105 channels × 8 frequencies, equivalent to 840 channels, by separating 8 additional frequency bands from the original 105 channels. We then converted it into a compressed feature representation Z via a one-dimensional convolutional encoder. The RVQ module then quantizes Z into M codebooks, generating M code indices $\{q_1, q_2, \dots, q_M\}$.

RVQ obtains a high-resolution discrete representation by applying hierarchical quantization over M codebooks. The final quantization vector z_q is the sum of the selected code vectors from each codebook:

$$z_q = \sum_{m=1}^{M} c_{(m,q_m)} \tag{1}$$

2.1.2 RVQ Training Objective

The tokenizer is trained as a VQ-VAE Van Den Oord et al. [2017] by minimizing the standard objective, which combines a reconstruction loss with codebook and commitment losses to ensure a robust discrete representation:

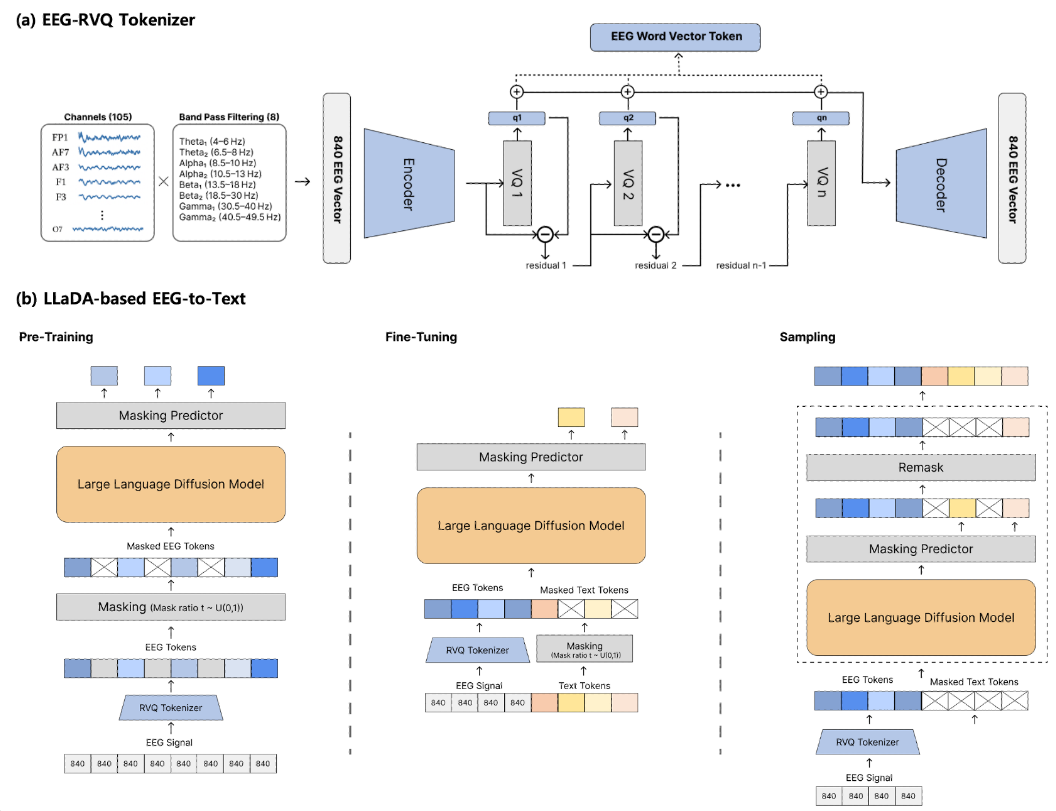

Figure 1: The DELTA framework: (a) An RVQ-based tokenizer discretizes EEG signals. (b) A diffusion model is then pre-trained on EEG tokens, fine-tuned for EEG-to-Text generation, and used for inference.

$$\mathcal{L}_{\text{VQ-VAE}} = \underbrace{\text{MSE}(x, \hat{x})}_{\text{Reconstruction Loss}} + \underbrace{\text{MSE}(\text{sg}(z_e), z_q)}_{\text{Codebook Loss}} + \beta \cdot \underbrace{\text{MSE}(z_e, \text{sg}(z_q))}_{\text{Commitment Loss}} \tag{2}$$

2.2 Stage 2: LLaDA-based Text Generation

LLaDA is a Large Language Diffusion Model, a new paradigm of language model that generates text through a stochastic masking and denoising process, unlike autoregressive models. In this study, we apply LLaDA's diffusion-based language generation technique to the EEG-to-Text problem by structuring its training and inference into distinct phases.

2.2.1 EEG Pre-Training

First, we perform a pre-training step using only the discrete EEG tokens obtained from the RVQ tokenizer, following the approach of Nie et al. Nie et al. [2025]. This stage involves a large-scale random masking and restoration process on the EEG tokens themselves, without using any text data. By learning to restore the masked EEG tokens from their noised version, LLaDA learns the inherent distribution and structure of the EEG token space. The objective is to minimize the following loss function:

$$\mathcal{L}_{\text{Pre-train}} = \mathbb{E}_{t, e_0 \sim p(e)}[\text{CE}(f_\theta(e_t, t), e_0)] \tag{3}$$

Here, e_0 is the original sequence of EEG tokens, and e_t is the corrupted version of e_0 at a random timestep t. The model f_θ is trained to predict e_0 from e_t and t. This process equips the model with diffusion-based learning capabilities tailored for EEG signals before it learns the cross-modal task of text generation.

2.2.2 EEG-to-Text Supervised Fine-Tuning (SFT)

After pre-training, the model is fine-tuned for the main task of generating text conditioned on EEG signals. In this stage, the model f_θ is given the discrete EEG tokens $\hat{Q}$ as a conditional prompt. The model's objective is to learn the probability distribution $P_\theta(Y|\hat{Q})$ to reconstruct the original text sequence Y from a partially masked or noised version of it. The training process minimizes the following loss function:

$$\mathcal{L}_{\text{LLaDA}} = \mathbb{E}_{t,x_0 \sim p(x)}[\text{CE}(f_\theta(x_t, t, \hat{Q}), x_0)] \tag{4}$$

Here, x_0 represents the original, clean sequence of text tokens. t is a randomly sampled timestep, and x_t is the noised version of x_0 at that timestep. The model f_θ is trained to predict the original text x_0 given the corrupted text x_t, the timestep t, and the conditional EEG tokens $\hat{Q}$. The loss is the cross-entropy (CE) between the model's prediction and the original text.

2.2.3 Inference

During inference, the model generates a sentence conditioned solely on the given EEG tokens $\hat{Q}$. The process begins with a sequence of tokens that are all completely masked, representing maximum uncertainty. The model then iteratively applies a reverse diffusion (denoising) process for a fixed number of steps. In each step, the model refines its prediction for the entire sequence of text tokens, gradually filling in the masked positions to form a coherent sentence. This non-autoregressive, parallel reconstruction approach is advantageous as it minimizes the cumulative errors common in sequential prediction. An incorrectly predicted token in one step can be corrected in subsequent steps, making the generation process more robust to the noisy and variable nature of EEG signals.

3 Experiments

3.1 Datasets

In this study, we used an integrated set of EEG-text pairs from ZuCo 1.0 Hollenstein et al. [2018] and ZuCo 2.0 Hollenstein et al. [2019]. Both datasets provide raw signals recorded from 128 channels of EEG, sampled at 500 Hz, over a frequency band of approximately 0.1-100 Hz, while subjects read English sentences naturally (Normal Reading, NR) or performed a specific task (Task-Specific Reading, TSR). After subsequent denoising, 105 channels of EEG are selected as active channels, and frequency and time series characteristics are extracted for each channel, normalized to the range of 0 to 1.

Preprocessing was performed for each word read by the subject to extract EEG segments corresponding to every sentence separately. Sentence-specific EEG sequences with missing values (NaN) or extreme noise were removed. Finally, out of a total of 25,616 sentences, 20,492 (approximately 80%) were assigned to the training set, 2,562 (approximately 10%) to the validation set, and 2,562 (approximately 10%) to the test set. We applied a unique sentence-based segmentation to avoid the same sentence appearing twice. By covering ZuCo data, which comprises various reading tasks (NR, TSR) and text sources, we have created an environment that enables us to validate the open lexical mapping of noisy EEG signals to natural language text in multiple ways.

3.2 Experimental setups

All experiments were conducted on a single NVIDIA L40S 40GB GPU using PyTorch. We used the AdamW optimizer with a batch size of 32 and employed early stopping to prevent overfitting. To manage computational resources, we fine-tuned the LLaDA-8B model in our DELTA framework using the Quantized Low-Rank Adapter (QLoRA) technique for both the EEG token pre-training and final generation stages Dettmers et al. [2023]. For comparison, we benchmarked our approach against autoregressive models, including BART Lewis et al. [2019], Pegasus Zhang et al. [2020], and T5 Raffel et al. [2020], which were also trained on our RVQ-tokenized EEG data. Model performance was evaluated on the test set using BLEU-N (N = 1, 2, 3, 4) Papineni et al. [2002], ROUGE-1 Lin [2004], and WER Klakow and Peters [2002] scores.

Table 1: Evaluation results on sentence- and word-level features.

Source	Method	**BLEU-N (%)**				**ROUGE-1 (%)**			**WER (%)**
		N=1	N=2	N=3	N=4	P	R	F	
Sentence-level features	BART	11.02	1.08	0.42	0.28	7.85	8.04	6.54	142.1
	Pegasus	6.20	0.55	0.24	0.16	4.90	5.05	4.03	147.8
	T5	12.50	1.18	0.48	0.31	8.6	8.72	7.15	141.0
	DELTA (proposed)	14.82	1.36	0.56	0.37	9.12	9.24	7.26	139.71
Word-level features	BART	13.69	2.97	0.82	0.32	11.98	13.43	11.87	108.43
	Pegasus	8.47	2.48	0.81	0.25	0	0	0	99.69
	T5	16.64	5.80	1.96	0.81	12.28	12.88	11.85	111.13
	DELTA (proposed)	21.93	6.43	2.01	0.76	18.88	18.86	17.24	110.03

3.3 Qualitative Analysis

As shown in Table 1, our proposed DELTA model demonstrates superior performance. On **sentence-level features**, DELTA significantly outperforms autoregressive models in BLEU and ROUGE scores (e.g., 14.82 BLEU-1, 7.26 ROUGE-1 F), proving its effectiveness in restoring global context. However, its WER is high (139.71) as the sentence-level aggregation disrupts the word order information crucial for the metric.

With **word-level features**, all models improve, but DELTA's lead widens. It surpasses the next-best models, T5 and BART, by over 5 percentage points in BLEU-1 (21.93) and ROUGE-1 F (17.24), respectively. Despite being slightly behind in BLEU-4, DELTA's overall performance confirms its strong semantic restoration capabilities.

3.4 Case Study

A closer look at generation examples reveals two key behaviors. First, successful generations can be nearly identical to the ground truth, accurately capturing complex structures. More notably, even in semantic failures, the model preserves the original's syntactic framework. For example, one prediction incorrectly altered a sentence's subject but retained its core grammatical structure (e.g., `"(Name) (Dates) was a (Profession)."`). This indicates that our model effectively learns syntax, even when it fails to decode the correct semantic content.

4 Conclusion

In this study, we propose a DELTA framework that combines an RVQ-based EEG tokenizer with the LLaDA to convert continuous electroencephalogram (EEG) signals into discrete tokens with minimal information loss and generate rich contextual text through a non-sequential diffusion process to achieve superior performance on key metrics such as BLEU and ROUGE compared to existing autoregressive-based methods. These results demonstrate that language diffusion models are a promising alternative for brain-signal-based natural language generation. In the future, we plan to extend the model's generality through large-scale pre-training and integrate other brain signals, such as MEG and multimodal inputs, further to improve the accuracy and applicability of non-invasive brain-language interfaces.

Acknowledgments and Disclosure of Funding

This research was supported by Brian Impact Foundation, a non-profit organization dedicated to the advancement of science and technology for all.

References

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in neural information processing systems*, 34:17981–17993, 2021.

Phairot Autthasan, Rattanaphon Chaisaen, Huy Phan, Maarten De Vos, and Theerawit Wilaiprasitporn. Mixnet: Joining force of classical and modern approaches toward the comprehensive pipeline in motor imagery eeg classification. *IEEE Internet of Things Journal*, 11(18):29736–29749, 2024.

Ujwal Chaudhary, Niels Birbaumer, and Antonio Ramos-Murguialday. Brain–computer interfaces for communication and rehabilitation. *Nature Reviews Neurology*, 12(9):513–525, 2016.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *arXiv preprint arXiv:2210.13438*, 2022.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *Advances in neural information processing systems*, 36:10088–10115, 2023.

Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. Dewave: discrete eeg waves encoding for brain dynamics to text translation. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018.

Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*, 2019.

Xiang Jiang, Gui-Bin Bian, and Zean Tian. Removal of artifacts from eeg signals: a review. *Sensors*, 19(5):987, 2019.

Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. Are eeg-to-text models working? *arXiv preprint arXiv:2405.06459*, 2024.

Dietrich Klakow and Jochen Peters. Testing the correlation of word error rate and perplexity. *Speech Communication*, 38(1-2):19–28, 2002.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*, 2019.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Sean L Metzger, Kaylo T Littlejohn, Alexander B Silva, David A Moses, Margaret P Seaton, Ran Wang, Maximilian E Dougherty, Jessie R Liu, Peter Wu, Michael A Berger, et al. A high-performance neuroprosthesis for speech decoding and avatar control. *Nature*, 620(7976): 1037–1046, 2023.

Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. Thought2text: Text generation from eeg signal using large language models (llms). *arXiv preprint arXiv:2410.07507*, 2024.

David A Moses, Sean L Metzger, Jessie R Liu, Gopala K Anumanchipalli, Joseph G Makin, Pengfei F Sun, Josh Chartier, Maximilian E Dougherty, Patricia M Liu, Gary M Abrams, et al. Neuroprosthesis for decoding speech in a paralyzed person with anarthria. *New England Journal of Medicine*, 385(3):217–227, 2021.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhenhailong Wang and Heng Ji. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5350–5358, 2022.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International conference on machine learning*, pages 11328–11339. PMLR, 2020.