

Latent-Level Uncertainty Resolution for Stable EEG-to-Text Decoding under Cross-Subject Variability

Sejin Park^{1,*} Jedong Lee^{2,*} Yunhui So^{2,*} Byung Hyung Kim^{1,2}

¹AFFCTIV

²Inha University

*Equal contribution

<https://github.com/affctivai/LURS>

Abstract

Autoregressive generators often rely on fixed conditioning representations, so uncertainty in noisy inputs can be amplified during rollout under distribution shift. We address this by resolving uncertainty in latent space before decoding: Latent-Level Uncertainty Resolution (LURS) encodes inputs as stochastic latent sequences and learns a denoising refinement operator. At inference, a single deterministic refinement step, without reference tokens or iterative sampling, feeds a refined latent memory to the decoder via cross-attention. In teacher-forced and teacher-forcing-free decoding on open-vocabulary EEG-to-text, LURS improves generation quality and substantially reduces cross-subject variability, while remaining competitive in shared-subject settings. Ablations and stability analysis attribute gains to latent refinement rather than decoder changes or hyperparameter effects.

1 Introduction

Autoregressive sequence generation commonly conditions decoding on a fixed encoder-side representation. When the input is noisy or ambiguous, uncertainty left in this representation can be amplified during token-level rollout, especially under distribution shift (Holtzman et al., 2020; Zablotskaia et al., 2023; Bengio et al., 2015; Taori et al., 2020; Ovadia et al., 2019). This suggests that stability depends not only on the decoder, but also on whether uncertainty is resolved before generation begins.

We propose Latent-Level Uncertainty Resolution (LURS), which resolves input uncertainty in latent space prior to autoregressive decoding. LURS encodes an observation as a stochastic latent sequence, learns a denoising-based refinement operator under semantic conditioning, and at inference applies a single deterministic refinement step without reference tokens or iterative sampling. The refined latent sequence is then used as cross-attention

memory for a pretrained sequence-to-sequence decoder, separating latent-level uncertainty resolution from token-level rollout.

We evaluate LURS on open-vocabulary EEG-to-text decoding, where neural noise and inter-subject variability create strong input uncertainty (Wang and Ji, 2022; Duan et al., 2023). Across teacher-forced and teacher-forcing-free settings, LURS improves cross-subject generation quality and stability, while remaining competitive under shared-subject evaluation. Ablations and stability analysis show that these gains arise from denoising-trained latent refinement rather than decoder changes or diffusion hyperparameter tuning.

Contributions. We make three contributions: (i) we frame robust sequence generation under distribution shift as latent-level uncertainty resolution before decoding; (ii) we introduce a denoising-trained latent refinement module that uses a single deterministic inference step; and (iii) we validate the approach on EEG-to-text decoding under shared- and cross-subject, teacher-forced and teacher-forcing-free protocols.

2 Related Work

Autoregressive generation and uncertainty.

Autoregressive models are widely used for conditional text generation, but free decoding can suffer from degeneration and error amplification when the conditioning signal is imperfect (Holtzman et al., 2020; Zhang et al., 2022; Bengio et al., 2015). Existing work models uncertainty mainly through sampling, entropy control, calibration, or Bayesian and ensemble methods (Welleck et al., 2020; Zablotskaia et al., 2023; Kim et al., 2025). These methods operate largely at the token level and generally assume that the conditioning representation is reliable. Our work instead targets upstream uncertainty before rollout begins.

Latent and diffusion-based generation. Latent variable models represent semantic variability through stochastic representations (Kwon et al., 2025), while diffusion models learn denoising processes that refine noisy states toward coherent samples (Ho et al., 2020; Li et al., 2022; Rombach et al., 2022). Most diffusion-based sequence models emphasize sample quality or diversity through iterative inference-time sampling. In contrast, LURS uses denoising as a one-step latent refinement mechanism to stabilize the conditioning memory for autoregressive decoding under distribution shift.

Distribution shift and EEG-to-text decoding. Robustness under distribution shift is challenging because shifts can perturb intermediate representations and degrade uncertainty estimates (Taori et al., 2020; Ovadia et al., 2019; Kong et al., 2023). EEG-to-text decoding is a natural stress test: EEG signals are noisy, subject-dependent, and difficult to align with language. Prior work combines EEG encoders with pretrained language models or improves discretization, alignment, and representation learning (Wang and Ji, 2022; Feng et al., 2023; Duan et al., 2023; Zhou et al., 2024b), while recent analyses highlight robustness and evaluation concerns (Jo et al., 2026). We build on this setting to test whether resolving uncertainty in the conditioning representation improves stable generation under free decoding and cross-subject shift.

3 Problem Formulation

We study conditional sequence generation from uncertain observations. Given an input x , the model produces an output sequence $y = (y_1, \dots, y_T)$ of length T . We use $t \in \{1, \dots, T\}$ to index autoregressive token positions and reserve τ for diffusion timesteps introduced later (Section 4.2).

Noisy or ambiguous inputs may admit multiple plausible semantic interpretations, and distribution shift at inference (e.g., subject-level variability in EEG) can perturb the conditioning signal. Our goal is to formalize how such uncertainty propagates through autoregressive decoding, and how resolving it before rollout improves stability.

3.1 Autoregressive Generation with Conditioning Memory

Autoregressive generation factorizes the conditional distribution as $p(y | x) = \prod_{t=1}^T p_\psi(y_t | y_{<t}, \mathbf{M}(x))$, where $\mathbf{M}(x)$ denotes an encoder-side conditioning memory computed once before de-

coding, and ψ denotes decoder parameters. The decoder repeatedly queries the same fixed memory (e.g., via cross-attention), so perturbations in $\mathbf{M}(x)$ affect token distributions at every timestep and can accumulate over the rollout, especially under distribution shift.

3.2 Latent Sequences for Uncertain Inputs

To represent uncertainty explicitly, we introduce a latent sequence $\mathbf{Z} \in \mathbb{R}^{L \times d_z}$ with an amortized posterior $q_\phi(\mathbf{Z} | x)$. This posterior captures multiple plausible semantic representations induced by a noisy observation x .

We write $p_\psi(y | \mathbf{Z})$ to denote decoding conditioned on a representation derived from $\mathbf{Z}$ (made explicit in Section 3.3). Conditional generation can then be expressed as

$$p(y | x) \approx \int p_\psi(y | \mathbf{Z}) q_\phi(\mathbf{Z} | x) d\mathbf{Z} \quad (1)$$

$$= \mathbb{E}_{q_\phi(\mathbf{Z} | x)} [p_\psi(y | \mathbf{Z})].$$

In practice, decoding conditions on a single latent realization $\mathbf{Z}_0 \sim q_\phi(\mathbf{Z} | x)$ or a point estimate. When $\mathbf{Z}_0$ is noisy or weakly aligned with linguistic structure, residual uncertainty enters the conditioning signal and can be amplified by autoregressive rollout.

3.3 Uncertainty Resolution Prior to Decoding

We resolve uncertainty in latent space before token generation begins. Given an initial latent sample $\mathbf{Z}_0 \sim q_\phi(\mathbf{Z} | x)$, a refinement operator R_θ produces a refined latent sequence

$$\mathbf{Z}^* = R_\theta(\mathbf{Z}_0). \quad (2)$$

The refined latent sequence is projected to an encoder-side memory

$$\mathbf{M}(x) = \text{Proj}_{\text{enc}}(\mathbf{Z}^*), \quad (3)$$

which is held fixed throughout decoding. Autoregressive generation then follows

$$p(y | x) \approx p_\psi(y | \mathbf{M}(x))$$

$$= \prod_{t=1}^T p_\psi(y_t | y_{<t}, \mathbf{M}(x)). \quad (4)$$

This formulation cleanly separates latent-level uncertainty resolution from token-level autoregressive rollout. Appendix B formalizes how reducing perturbations in $\mathbf{Z}^*$ and $\mathbf{M}(x)$ improves stability of the induced sequence distribution.

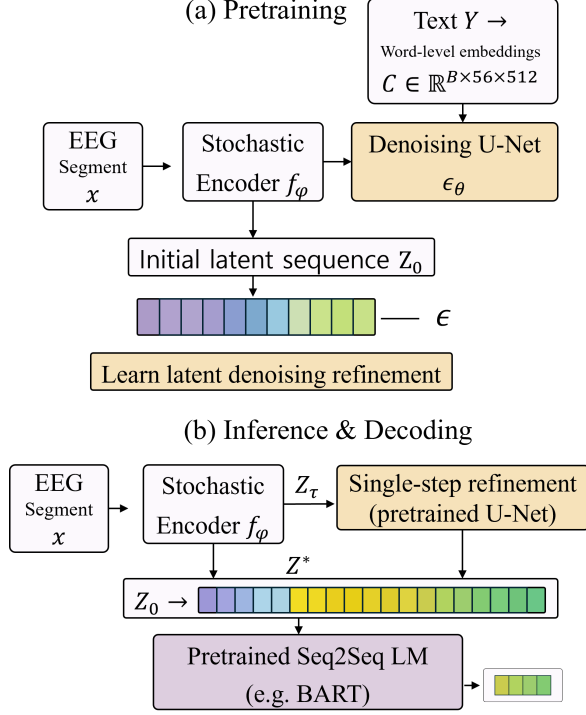


Figure 1: Overview of LURS. (a) Pretraining: a latent refinement operator is learned under semantic conditioning. (b) Inference and decoding: a refined latent sequence conditions autoregressive generation.

4 Method: Latent-Level Uncertainty Resolution

This section presents LURS, a latent-level uncertainty resolution framework for sequence generation from uncertain inputs. Experiments instantiate the input as a sentence-aligned EEG segment, while the decoding interface applies to any modality that yields a latent sequence. LURS follows the pipeline $x \rightarrow \mathbf{Z}_0 \rightarrow \mathbf{Z}^* \rightarrow y$: a stochastic encoder defines a posterior over a latent sequence $\mathbf{Z} \in \mathbb{R}^{L \times d_z}$, produces an initial sample $\mathbf{Z}_0$, refines it to $\mathbf{Z}^*$ prior to decoding, and a pretrained language decoder generates y by attending to the refined latent memory (Figure 1). Latent refinement is learned via diffusion-style denoising in representation space, while inference applies a single deterministic refinement pass at a fixed diffusion index τ_{ref} without iterative reverse sampling.

4.1 Stochastic EEG Latent Encoding

This subsection defines a stochastic latent encoder for uncertain observations; experiments instantiate the observation as a sentence-aligned EEG segment x . EEG signals exhibit trial-level noise and subject-dependent variability (Garrett et al., 2013;

Minamoto et al., 2017; Schirrmester et al., 2017; Choi et al., 2024). The encoder models this uncertainty by mapping x to a factorized diagonal Gaussian posterior over a latent sequence $\mathbf{Z} \in \mathbb{R}^{L \times d_z}$.

An encoder f_ϕ outputs unconstrained distribution parameters:

$$\begin{aligned} (\boldsymbol{\mu}_\phi(x), \tilde{\boldsymbol{\sigma}}_\phi(x)) &= f_\phi(x), \\ \boldsymbol{\sigma}_\phi(x) &= g(\tilde{\boldsymbol{\sigma}}_\phi(x)), \end{aligned} \quad (5)$$

where $\boldsymbol{\mu}_\phi(x), \boldsymbol{\sigma}_\phi(x) \in \mathbb{R}^{L \times d_z}$, and $g(\cdot)$ enforces $\boldsymbol{\sigma}_\phi(x) > \mathbf{0}$.

The posterior factorizes across positions with diagonal covariance:

$$q_\phi(\mathbf{Z} | x) = \prod_{\ell=1}^L \mathcal{N}(\mathbf{z}_\ell; \boldsymbol{\mu}_{\phi,\ell}(x), \text{diag}(\boldsymbol{\sigma}_{\phi,\ell}^2(x))). \quad (6)$$

A latent sample is obtained via reparameterization:

$$\mathbf{Z}_0 = \boldsymbol{\mu}_\phi(x) + \boldsymbol{\sigma}_\phi(x) \odot \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \quad (7)$$

Unless otherwise stated, experiments use one sample per instance; a deterministic variant replaces $\mathbf{Z}_0$ with $\boldsymbol{\mu}_\phi(x)$.

4.2 Text-Conditioned Latent Refinement

We learn a latent refinement operator that resolves representation uncertainty prior to decoding. During pretraining, paired text provides semantic conditioning, while inference performs refinement without reference text. Refinement is learned via diffusion-style denoising in latent space, but inference applies only a single deterministic denoising pass without iterative reverse sampling. To avoid notational collision, we use $t \in \{1, \dots, T\}$ for autoregressive token positions (Section 3.1) and reserve $\tau \in \{1, \dots, N\}$ exclusively for diffusion timesteps.

Text-conditioned denoising pretraining. We construct a conditioning sequence using the pretrained BART-large token embedding table (1024-d) (Lewis et al., 2020). Token embeddings are linearly projected to dimension d_z :

$$\mathbf{C} = \text{Proj}(\mathbf{E}^{\text{BART}}) \in \mathbb{R}^{L \times d_z}.$$

The conditioning sequence $\mathbf{C}$ is used only during denoising-based pretraining through cross-attention in the U-Net; at inference, we set $\mathbf{C} = \mathbf{0}$.

Let $\{\beta_\tau\}_{\tau=1}^N$ be a variance schedule, $\alpha_\tau = 1 - \beta_\tau$, and $\bar{\alpha}_\tau = \prod_{i=1}^\tau \alpha_i$. Given an initial latent

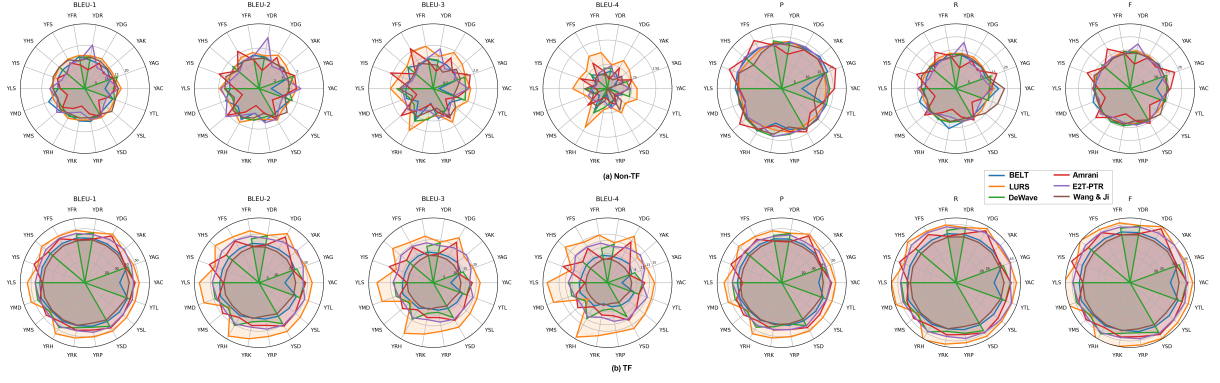


Figure 2: Cross-subject radar visualization of decoding performance. The more circular profiles observed for LURS under Non-TF (Orange) decoding indicate reduced sensitivity to subject-level perturbations. In the radar plots, larger radii correspond to higher average performance across subjects, while increased circularity reflects lower inter-subject variability and reduced coefficients of variation.

sample $\mathbf{Z}_0 \sim q_\phi(\mathbf{Z} \mid x)$, the forward diffusion process produces a corrupted latent sequence:

$$\mathbf{Z}_\tau = \sqrt{\bar{\alpha}_\tau} \mathbf{Z}_0 + \sqrt{1 - \bar{\alpha}_\tau} \epsilon, \quad (8)$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\tau \in \{1, \dots, N\}$. A denoising U-Net ϵ_θ predicts the injected noise conditioned on $\mathbf{C}$:

$$\hat{\epsilon} = \epsilon_\theta(\mathbf{Z}_\tau, \tau, \mathbf{C}), \quad (9)$$

implemented via cross-attention where latent features act as queries and text embeddings as keys and values. Training minimizes the mean-squared error

$$\mathcal{L}_{\text{pre}} = \mathbb{E}_{(x,y), \mathbf{Z}_0 \sim q_\phi(\cdot \mid x), \tau, \epsilon} \left[\|\hat{\epsilon} - \epsilon\|_F^2 \right]. \quad (10)$$

Single-step deterministic refinement at inference. At inference, reference text is unavailable and we apply a null condition $\mathbf{C} = \mathbf{0}$. Refinement is defined as a single x_0 -estimate obtained from one denoiser evaluation. Fix a small diffusion index $\tau_{\text{ref}} \in \{1, \dots, N\}$ (we use $\tau_{\text{ref}} = 1$ in all experiments). Given $\mathbf{Z}_0$, we form a deterministic forward-mean perturbation without sampling additional noise:

$$\tilde{\mathbf{Z}}_{\tau_{\text{ref}}} := \sqrt{\bar{\alpha}_{\tau_{\text{ref}}}} \mathbf{Z}_0. \quad (11)$$

We then predict noise under the null condition and convert it into an x_0 -estimate:

$$\mathbf{Z}^* := \frac{\tilde{\mathbf{Z}}_{\tau_{\text{ref}}} - \sqrt{1 - \bar{\alpha}_{\tau_{\text{ref}}}} \hat{\epsilon}_{\text{ref}}}{\sqrt{\bar{\alpha}_{\tau_{\text{ref}}}}}, \quad (12)$$

where $\hat{\epsilon}_{\text{ref}} := \epsilon_\theta(\tilde{\mathbf{Z}}_{\tau_{\text{ref}}}, \tau_{\text{ref}}, \mathbf{0})$. This defines a deterministic refinement operator $R_\theta(\cdot)$ via $\mathbf{Z}^* =$

$R_\theta(\mathbf{Z}_0)$. The refinement map is deterministic conditioned on $\mathbf{Z}_0$; the only stochasticity arises from sampling $\mathbf{Z}_0 \sim q_\phi(\mathbf{Z} \mid x)$. Appendix A provides an interpretation of how denoising-based pretraining shapes representation stability.

This single-step approach is theoretically grounded in the stability bound derived in Proposition 4.1 and Corollary B.3. By mapping the initial stochastic sample $\mathbf{Z}_0$ toward the high-density region of the conditioned manifold, the refinement operator R_θ acts as a contraction map that reduces the variance of the conditioning signal. As proven in Appendix B, reducing perturbations in the latent space directly minimizes the KL divergence between induced sequence distributions, effectively stabilizing the autoregressive rollout against input noise and subject-level shifts. Empirically, this deterministic pass provides sufficient stabilization for cross-attention memory without the computational overhead of iterative reverse sampling.

4.3 EEG-Conditioned Text Generation

Given a refined latent sequence $\mathbf{Z}^* \in \mathbb{R}^{L \times d_z}$, a pretrained sequence-to-sequence language model generates an output $y = (y_1, \dots, y_T)$ via standard encoder-decoder attention over the refined latent memory. The refined latent sequence is first projected to encoder-side memory, $\mathbf{M} = \text{Proj}_{\text{enc}}(\mathbf{Z}^*) \in \mathbb{R}^{L \times d_{\text{model}}}$, which is treated as fixed conditioning memory throughout decoding.

Autoregressive generation then follows $p_\psi(y \mid \mathbf{M}) = \prod_{t=1}^T p_\psi(y_t \mid y_{<t}, \mathbf{M})$, where ψ denotes decoder parameters. Teacher-forced training supplies ground-truth prefixes $y_{<t}$, while teacher-forcing-free decoding generates tokens autoregressively

using previously generated outputs. In both cases, the conditioning memory $\mathbf{M}$ remains fixed for the entire sequence.

4.4 Training Objectives and Stability

Training proceeds in two stages: denoising-based pretraining for latent refinement, followed by latent-conditioned sequence learning for generation.

In Stage 1 (latent refinement pretraining), given paired data (x, y) , the encoder samples $\mathbf{Z}_0 \sim q_\phi(\mathbf{Z} \mid x)$, constructs $\mathbf{Z}_T$ via forward diffusion, and minimizes the denoising objective $\mathcal{L}_{\text{pre}}$.

In Stage 2 (latent-conditioned generation), the decoder is trained by conditioning on $\mathbf{M} = \text{Proj}_{\text{enc}}(\mathbf{Z}^*)$ and minimizing the negative log-likelihood $\mathcal{L}_{\text{text}} = -\mathbb{E}_{(x,y)} \sum_{t=1}^T \log p_\psi(y_t \mid y_{<t}, \mathbf{M})$. Optimization first trains latent refinement with $\mathcal{L}_{\text{pre}}$ and then trains text generation with $\mathcal{L}_{\text{text}}$.

Autoregressive decoding conditions on the same encoder-side memory $\mathbf{M}$ at every timestep. Perturbations in $\mathbf{M}$ can alter token distributions at each step and accumulate along the rollout.

Proposition 4.1 (Stability bound under memory perturbations). *Let $p_\psi(y \mid \mathbf{M}) = \prod_{t=1}^T p_\psi(y_t \mid y_{<t}, \mathbf{M})$ be an autoregressive decoder conditioned on encoder-side memory $\mathbf{M}$. Assume there exists $C_{\text{lip}} > 0$ such that for any timestep t and history $y_{<t}$,*

$$\text{KL}(p_\psi(\cdot \mid y_{<t}, \mathbf{M}) \parallel p_\psi(\cdot \mid y_{<t}, \mathbf{M}')) \leq C_{\text{lip}} \|\mathbf{M} - \mathbf{M}'\|_F. \quad (13)$$

Then the induced sequence distributions satisfy

$$\text{KL}(p_\psi(\cdot \mid \mathbf{M}) \parallel p_\psi(\cdot \mid \mathbf{M}')) \leq T C_{\text{lip}} \|\mathbf{M} - \mathbf{M}'\|_F. \quad (14)$$

Appendix B provides a full proof and a corollary linking perturbation reduction from $\mathbf{Z}_0$ to $\mathbf{Z}^*$ with reduced sensitivity in the induced sequence distribution.

5 Experiments

This section evaluates latent-level uncertainty resolution on ZuCo EEG-to-text decoding under shared- and cross-subject settings with teacher-forced (TF) and teacher-forcing-free (Non-TF) decoding. We follow the standard word-aligned setup with $L = 56$ and $d_z = 512$, and share the decoding interface across methods to isolate latent refinement effects (Hollenstein et al., 2018; Wang and Ji, 2022; Duan et al., 2023; Zhou et al., 2024b).

Model	BLEU $\uparrow$				ROUGE-1 $\uparrow$		
	1	2	3	4	P	R	F1
Cross-subject (LOSO, Non-TF): Mean$\uparrow$ / CV$\downarrow$							
DeWave	10.44 (64.12)	2.86 (64.60)	1.03 (66.03)	0.42 (77.32)	13.06 (63.91)	9.54 (64.02)	10.30 (63.94)
Wang & Ji	14.56 (6.44)	3.84 (9.83)	1.22 (17.64)	0.43 (53.87)	16.81 (9.75)	13.57 (13.20)	13.97 (5.06)
E2T-PTR	13.55 (19.13)	4.04 (22.39)	1.30 (18.46)	0.52 (21.46)	17.73 (5.48)	12.50 (14.98)	13.90 (8.53)
Amrani et al.	11.69 (18.60)	3.63 (30.48)	1.33 (34.95)	0.52 (42.77)	17.89 (15.07)	12.02 (24.41)	13.68 (20.77)
BELT	13.84 (12.56)	3.74 (17.31)	1.19 (27.10)	0.44 (42.53)	17.17 (15.46)	13.22 (9.85)	13.85 (7.32)
LURS	15.39 (3.78)	4.29 (8.15)	1.69 (15.59)	0.89 (23.78)	17.20 (4.09)	13.44 (3.39)	14.07 (2.72)
Cross-subject (LOSO, TF): Mean$\uparrow$ / CV$\downarrow$							
DeWave	32.42 (64.03)	19.69 (64.47)	12.13 (65.17)	7.63 (65.96)	31.90 (64.16)	29.70 (64.03)	30.72 (64.07)
Wang & Ji	39.74 (1.89)	22.05 (2.36)	12.36 (3.08)	7.02 (4.22)	38.90 (1.48)	35.80 (1.79)	37.18 (1.41)
E2T-PTR	46.24 (2.16)	28.97 (3.83)	18.47 (5.64)	12.02 (7.36)	44.54 (2.36)	43.71 (2.96)	44.06 (1.79)
Amrani et al.	44.75 (7.04)	27.15 (12.10)	16.69 (17.73)	10.46 (22.93)	44.02 (8.80)	41.04 (7.35)	42.41 (8.04)
BELT	41.19 (5.47)	23.58 (8.55)	13.63 (11.10)	8.10 (12.98)	40.40 (4.07)	37.28 (5.41)	38.68 (5.09)
LURS	50.00 (6.31)	32.88 (10.60)	22.03 (15.02)	14.94 (18.81)	49.78 (6.75)	46.21 (7.01)	47.87 (6.91)

Table 1: Cross-subject (LOSO) performance and stability. Mean scores with CV (%) in parentheses. LURS achieves statistically significant improvements with lower cross-subject variability.

5.1 Evaluation protocols and decoding regimes

5.1.1 Shared- and Cross-Subject Evaluation

Shared-subject evaluation measures sentence-level generalization with overlapping subjects in training and testing. We use the standard 80/10/10 train validation test split following prior work (Wang and Ji, 2022; Duan et al., 2023; Zhou et al., 2024b).

Cross-subject evaluation assesses robustness under subject-level distribution shift using a leave-one-subject-out (LOSO) protocol. For each subject, the model is trained on the remaining 17 subjects and evaluated on the held-out subject, with results aggregated across all 18 subjects. Since subject variability perturbs the encoder-side conditioning memory $\mathbf{M}$ consumed via cross-attention, LOSO directly probes robustness to memory perturbations.

5.1.2 Teacher-Forced vs. Non-Teacher-Forced Decoding

Results are reported under both TF and Non-TF decoding. TF conditions generation on ground-truth prefixes, whereas Non-TF performs fully autoregressive generation while keeping the conditioning memory $\mathbf{M}$ fixed throughout the rollout (Section 4.3). Non-TF serves as the primary eval-

Model	BLEU $\uparrow$				ROUGE-1 $\uparrow$		
	1	2	3	4	P	R	F1
Shared-subject (Non-TF)							
Wang & Ji	13.72	3.62	1.38	0.60	16.67	12.10	13.16
DeWave	11.76	3.26	1.35	0.61	14.70	10.50	11.37
E2T-PTR*	14.78	4.62	1.93	0.86	17.11	12.97	14.38
BELT*	12.55	3.43	1.31	0.52	18.72	12.48	13.57
Amrani et al.*	10.39	2.64	1.21	0.54	16.74	10.14	12.00
LURS	15.79	4.38	1.82	0.95	16.63	13.34	14.05
Shared-subject (TF)							
Wang & Ji*	40.86	23.48	12.84	6.68	40.54	37.47	38.91
DeWave	42.33	25.06	14.73	8.55	34.55	31.40	32.78
E2T-PTR*	40.76	23.81	13.78	7.97	39.61	37.28	38.39
BELT*	41.36	24.27	14.02	8.29	40.79	37.94	39.24
Amrani et al.*	42.62	25.39	15.04	9.00	42.18	39.50	40.77
BELT-2	43.06	25.57	15.05	9.09	30.28	34.12	31.99
LURS	42.98	25.63	15.16	8.92	42.45	39.76	41.02
LURS-con	42.79	25.64	15.40	9.47	42.13	39.73	40.87

Table 2: Shared-subject evaluation. The suffix "-con" denotes the variant trained with an additional contrastive objective.

uation setting, as it matches inference-time usage and exposes instability amplified by autoregressive rollout.

5.2 Baselines

Comparisons include representative open-vocabulary EEG-to-text decoders, including Wang & Ji (Wang and Ji, 2022), DeWave (Duan et al., 2023), E2T-PTR (Wang et al., 2024), BELT (Zhou et al., 2024b), Amrani et al. (Amrani et al., 2026), and BELT-2 (Zhou et al., 2024a) evaluated under the same decoding interface and protocols. Methods marked with * denote our reimplementation under a unified setup.

5.3 Metrics and Cross-Subject Stability

Generation quality is assessed using BLEU-1/2/3/4 and ROUGE-1 precision, recall, and F1 (Papineni et al., 2002; Lin, 2004). To quantify robustness under subject-level distribution shift, we additionally report the coefficient of variation (CV) across held-out subjects, reported as a percentage: $100 \times \frac{\text{Std}}{\text{Mean}}$ over subject-level scores.

Average decoding performance reflects overall generation quality but fails to capture robustness under subject-level distribution shift. Our theoretical analysis (Proposition 4.1 and Appendix B) shows that perturbations in the conditioning memory $\mathbf{M}$ propagate through the autoregressive rollout, amplifying sequence-level sensitivity. Accordingly, CV across held-out subjects serves as a measure of cross-subject stability.

5.4 Quantitative Results

5.4.1 Cross-Subject Evaluation under Distribution Shift

Table 1 reports performance and stability under subject-level distribution shift using LOSO evaluation. Under Non-TF decoding (our primary setting), LURS achieves the highest mean BLEU-1/2/3/4 and ROUGE-1 F1. Compared to the strongest baseline, ROUGE-1 F1 increases from 13.97 to 14.07 and BLEU-1 from 14.56 to 15.39.

More importantly, LURS substantially reduces cross-subject variability. For ROUGE-1 F1, the coefficient of variation (CV) decreases to 2.72% compared to 5.06% for the lowest-CV baseline, and for BLEU-1 to 3.78% compared to 6.44% (Figure 2). Although the mean improvement in ROUGE-1 F1 is modest, subject-wise statistical tests under LOSO Non-TF confirm consistent gains across the subject population. LURS shows statistically significant improvements in ROUGE-1 F1 ($p = 0.021$, Holm-corrected; median difference $+0.18$, 95% CI $[+0.06, +0.31]$) and BLEU-1 ($p = 0.013$, Holm-corrected; median difference $+0.74$, 95% CI $[+0.28, +1.19]$). These results indicate that the observed improvements reflect population-level robustness rather than isolated favorable cases.

Overall, LURS attains the lowest CV in 6 of 7 metrics under cross-subject Non-TF decoding, indicating more stable generation across unseen subjects. This supports our central hypothesis that latent refinement stabilizes the conditioning memory $\mathbf{M}$ under distribution shift, thereby reducing rollout sensitivity in teacher-forcing-free autoregressive decoding.

Under TF decoding, LURS achieves the best mean performance across all metrics. We report TF results for completeness, while emphasizing Non-TF as the more realistic setting, since access to ground-truth prefixes can mask sensitivity to perturbations in $\mathbf{M}$.

5.4.2 Shared-Subject Evaluation

Table 2 reports shared-subject performance. Under Non-TF decoding, LURS achieves the best BLEU-1 and BLEU-4 and the highest ROUGE recall, while E2T-PTR attains slightly higher BLEU-2/3 and ROUGE-1 F1. Under TF decoding, the plain LURS model achieves the best BLEU-2/3 and ROUGE-1 P/R/F1 among the non-contrastive variants, while BELT-2 attains slightly higher BLEU-1

Ablation	BLEU $\uparrow$				ROUGE-1 $\uparrow$		
	1	2	3	4	P	R	F1
Full	15.79	4.38	1.82	0.95	16.63	13.34	14.05
No pretrained U-Net reuse	5.99	0.90	0.13	0.00	4.83	8.06	5.70
w/o diffusion pretraining	12.02	3.26	1.23	0.52	17.71	11.41	13.01
w/o U-Net in decoding	14.67	3.94	1.50	0.67	17.05	12.73	13.60

Table 3: Component ablations on shared-subject Non-TF decoding.

Noise	Non-TF $\uparrow$		TF $\uparrow$	
	BLEU-1	ROUGE-F1	BLEU-1	ROUGE-F1
Linear	15.14	13.76	42.17	40.57
Sqrt	15.79	14.05	42.98	41.02
Cosine	15.28	13.63	42.58	40.80

Table 4: Noise schedule sensitivity.

and BLEU-4. The contrastive variant LURS-con further improves BLEU-2/3/4, but slightly lowers ROUGE scores compared with LURS. Overall, improvements under cross-subject shift do not come at the expense of within-protocol performance.

5.5 Ablations on Latent-level Uncertainty Resolution

We ablate key components to identify which elements of latent-level uncertainty resolution contribute to stable generation. Table 3 reports shared-subject Non-TF results.

Reuse of the denoising-pretrained U-Net is critical. When the U-Net used at decoding does not reuse denoising-pretrained parameters, performance collapses (BLEU-1: 15.79 $\rightarrow$ 5.99; ROUGE-1 F1: 14.05 $\rightarrow$ 5.70). This shows that denoising-based pretraining is essential for learning a refinement operator that produces an effective conditioning memory for cross-attention.

Diffusion-based denoising pretraining improves free decoding. Removing diffusion pretraining degrades performance (BLEU-1: 15.79 $\rightarrow$ 12.02; F1: 14.05 $\rightarrow$ 13.01), indicating that the denoising objective provides a strong training signal for semantically structured latent refinement.

Inference-time latent refinement matters under Non-TF rollout. Removing U-Net refinement at decoding and conditioning directly on $\mathbf{Z}_0$ reduces scores (BLEU-1: 15.79 $\rightarrow$ 14.67; F1: 14.05 $\rightarrow$ 13.60), supporting the importance of resolving residual latent uncertainty before autoregressive rollout.

# steps N	Non-TF $\uparrow$		TF $\uparrow$	
	BLEU-1	ROUGE-F1	BLEU-1	ROUGE-F1
250	14.79	13.50	42.81	40.98
750	15.21	13.82	43.44	41.71
1000	15.79	14.05	42.98	41.02
1500	15.20	13.76	43.04	41.26

Table 5: Sensitivity to the number of diffusion steps.

5.6 Sensitivity to Diffusion Hyperparameters

To assess whether performance improvements arise from the proposed design rather than carefully tuned diffusion hyperparameters, we evaluate LURS across different forward noise schedules (Table 4) and diffusion step counts (Table 5) during denoising-based pretraining. Performance remains stable across a wide range of settings under both TF and Non-TF decoding, indicating that LURS does not rely on specific diffusion hyperparameter choices. The observed improvements are attributable to denoising-trained latent refinement that resolves uncertainty prior to autoregressive decoding.

5.7 Comparison with Explicit Latent Regularization

A natural question is whether the stabilizing effect of latent refinement can be achieved through explicit latent regularization, such as a Kullback–Leibler (KL) penalty that encourages the posterior to match a fixed prior. To examine this possibility, we augment the denoising pretraining objective with an additional KL divergence term of varying strength and evaluate its effect under Non-TF decoding, which most strongly exposes autoregressive instability.

Table 7 reports the results. Introducing an explicit KL regularizer does not improve generation quality and often degrades performance, particularly under teacher-forcing-free decoding. In contrast, LURS without explicit KL regularization achieves higher BLEU and ROUGE scores. These results indicate that the benefit of LURS does not arise from constraining latent representations toward a fixed prior. Instead, denoising-based latent refinement more effectively stabilizes the conditioning representation prior to autoregressive decoding.

5.8 Qualitative Analysis

Table 6 presents qualitative examples illustrating the behavior of LURS under both TF and Non-TF decoding in the shared-subject setting, with ad-

Table 6: Translation results on the unseen EEG waves, where **bold** denotes a correct match between ground truth and our prediction. Underline denotes a fuzzy match with similar semantic meanings.

#	Ground Truth	Prediction
(1)	He is a prominent member of the Bush family, the younger brother of President George W. Bush and the second son of former President George H. W. Bush and Barbara Bush.	(Non-TF) He was a member of the U.S. Army during World War II, and served as an <u>aide</u> to General George S. Patton .
(2)	The eldest son of William Parker, he was born in <u>Norwich</u> , in St. Saviour's parish.	(Non-TF) He is the son of George W. Bush and Jacqueline Kennedy Onassis, who both served as U.S. presidents from 1961 to 1963. (Non-TF) He was born in <u>London</u> , England and raised in New York City.
(3)	He was born in Overisel, Michigan.	(TF) was born in <u>Newland</u> , Germany
(4)	A classy item by a legend who may have nothing left to prove but still has the chops and <u>drive</u> to show how its done.	(Non-TF) The <u>film</u> is a <u>classic</u> , and it's easy to see why people <u>love it</u> so much (TF) few,, a talented who is have been to to prove. his has a ability to the to make it much characters.
(5)	Kennedy's other decorations of the <u>Second World War</u> include the Purple Heart, <u>Asiatic-Pacific Campaign Medal</u> , and the World War II Victory Medal.	(Non-TF) He served in the Army during World War II and was a <u>member of the United States Marine Corps</u> .
(6)	He was first <u>appointed</u> to fill the <u>Senate seat</u> of Ernest <u>Lundeen</u> who had <u>died</u> in office.	(Non-TF) He was <u>elected</u> to the U.S. <u>House of Representatives</u> in 1894, and served until his <u>death</u> in 1901.

Table 7: Comparison of performance metrics across different KL regularization settings.

Metric	LURS	$\beta_{KL} = 0.01$	$\beta_{KL} = 0.001$
BLEU-1 $\uparrow$	15.79	13.83	13.54
BLEU-2 $\uparrow$	4.38	3.94	3.81
BLEU-3 $\uparrow$	1.82	1.62	1.55
BLEU-4 $\uparrow$	0.95	0.81	0.73
ROUGE-P $\uparrow$	16.63	16.72	16.16
ROUGE-R $\uparrow$	13.34	12.17	11.89
ROUGE-F1 $\uparrow$	14.05	13.05	13.21

ditional examples provided in Appendix D. These cases complement the quantitative analysis by highlighting stable and coherent generation beyond metric-level comparisons.

Although LURS does not aim to reproduce reference sentences verbatim, it consistently preserves semantic content, syntactic structure, and concept-level correspondence. In contrast, TF decoding often produces broken syntax, repetition, or incomplete sentences, whereas Non-TF decoding with LURS yields outputs that are more fluent, coherent, and semantically faithful. This contrast reflects the inherent instability of autoregressive generation when models rely directly on token-level supervision rather than on a stabilized conditioning representation.

The examples in Table 6 illustrate several representative behaviors. LURS preserves core semantic relations such as affiliation and role structure even when surface forms differ, maintains common syntactic templates such as birth or event descriptions, and retains sentiment and evaluative tone through paraphrasing and conceptual substitution. The model also generalizes historical and political concepts by preserving high-level intent while varying specific details.

Across these examples, LURS generates novel

sentences that convey the same intent or topic as the reference without copying surface wording. This behavior reflects the probabilistic nature of latent diffusion, which supports semantic diversity and expressive flexibility. Importantly, the qualitative results suggest that latent refinement stabilizes the conditioning signal during autoregressive decoding. By generating text from a semantically refined latent representation rather than directly propagating token-level predictions, LURS mitigates error amplification and reduces degenerate rollouts.

While these examples are drawn from the shared-subject setting, they provide insight into the mechanisms underlying the improved stability observed in quantitative evaluations. Together, the qualitative and quantitative results support latent refinement as a key factor in stable and interpretable EEG-to-text generation.

6 Conclusion

We introduced LURS, a framework that resolves uncertainty in latent space before autoregressive decoding. Experiments on open-vocabulary EEG-to-text decoding show that latent refinement improves generation quality and reduces cross-subject variability, especially under teacher-forcing-free inference. These results suggest that stabilizing the conditioning representation before token rollout is an effective direction for robust sequence generation under distribution shift.

7 Limitations

This study is confined to open-vocabulary EEG-to-text decoding on ZuCo, so generality to other neural modalities, datasets, languages, and real-time BCI settings remains unverified. Real-time use in

particular differs from our setting in more than the data, since the word-aligned setup we inherit supplies sentence boundaries and word-level timing that a continuous online BCI would not provide. The refinement operator itself assumes only a latent sequence, so applying LURS to other corpora such as inner-speech EEG or MEG reading data is a natural next step, though relaxing the word-alignment assumption remains necessary for online use.

LURS improves stability and coarse semantic alignment rather than exact sentence recovery; generated text may still substitute entities, locations, and factual details. We characterize this as the level at which alignment currently operates rather than as incidental noise. For instance, in several cases in Table 6 and Appendix D, the model produces a fluent, well-formed sentence in the correct register while realizing a different referent from the target. Named entities carry much of the intended content, so recovering them reliably is necessary before the approach can support communication rather than topic-level alignment alone.

Our evaluation also relies on automatic overlap metrics and cross-subject CV. Because the overlap metrics compare surface forms, they cannot distinguish different degrees of semantic proximity. Metrics that capture semantic and referential correspondence, such as embedding-based similarity or entity-level precision and recall, are needed to characterize what the model actually preserves. A separate question is how much of the reported performance is attributable to the EEG input rather than to the decoder’s language prior, which absolute overlap scores do not separate. Quantifying this contribution requires evaluation designed for that purpose.

Finally, the current model uses a large pretrained decoder and denoising module, totaling approximately 930M trainable parameters, of which 493M belong to the denoising U-Net. The single-step design reduces inference-time sampling cost rather than parameter or training cost, leaving lightweight deployment-oriented variants for future work.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.RS-2026-25520944, No.RS-2026-25548560).

References

- Hamza Amrani, Daniela Micucci, and Paolo Napoletano. 2026. Deep representation learning for open vocabulary electroencephalography-to-text decoding. *IEEE Journal of Biomedical and Health Informatics*, 30(4):3593–3600.
- Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 28.
- HyoSeon Choi, Dahoon Choi, Netiwit Kaongoen, and Byung Hyung Kim. 2024. Detecting concept shifts under different levels of self-awareness on emotion labeling. In *International Conference on Pattern Recognition (ICPR)*, pages 276–291. Springer.
- Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-teng Lin. 2023. DeWave: Discrete encoding of EEG waves for EEG to text translation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pages 9907–9918.
- Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 31:3874–3883.
- Douglas D Garrett, Gregory R Samanez-Larkin, Stuart WS MacDonald, Ulman Lindenberger, Anthony R McIntosh, and Cheryl L Grady. 2013. Moment-to-moment brain signal variability: a next frontier in human brain mapping? *Neuroscience & Biobehavioral Reviews*, 37(4):610–624.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851.
- Nora Hollenstein, Jonathan Rotsztein, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. *Scientific Data*, 5:180291.
- Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2020. ZuCo 2.0: A dataset of physiological recordings during natural reading and annotation. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 138–146.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations (ICLR)*.
- Hyejeong Jo, Yiqian Yang, Juhyeok Han, Yiqun Duan, Hui Xiong, and Won Hee Lee. 2026. Evaluating EEG-to-text models through noise-based performance analysis. *Scientific Reports*, 16(1):350.

- Minsu Kim, Sangryul Kim, and James Thorne. 2025. From evidence to belief: A Bayesian epistemology approach to language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10578–10611.
- Lingkai Kong, Harshavardhan Kamarthi, Peng Chen, B. Aditya Prakash, and Chao Zhang. 2023. Uncertainty quantification in deep learning. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*, pages 5809–5810.
- Bum Chul Kwon, Ben Shapira, Moshiko Raboh, Shreyans Sethi, Shruti Murarka, Joseph A Morrone, Jianying Hu, and Parthasarathy Suryanarayanan. 2025. STAR-VAE: Latent variable transformers for scalable and controllable molecular generation.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880.
- Xiang Lisa Li, John Thickstun, Ishaan Gulrajani, Percy Liang, and Tatsunori B. Hashimoto. 2022. Diffusion-LM improves controllable text generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 4328–4343.
- Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81.
- Takehiro Minamoto, Hiroyuki Tsubomi, and Naoyuki Osaka. 2017. Neural mechanisms of individual differences in working memory capacity: Observations from functional neuroimaging studies. *Current Directions in Psychological Science*, 26(4):335–345.
- Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. 2019. Can you trust your model’s uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. 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.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695.
- Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggenberger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. 2017. Deep learning with convolutional neural networks for EEG decoding and visualization. *Human brain mapping*, 38(11):5391–5420.
- Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 18583–18599.
- Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. 2024. Enhancing EEG-to-text decoding through transferable representations from pre-trained contrastive EEG-text masked autoencoder. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7278–7292.
- Zhenhailong Wang and Heng Ji. 2022. 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.
- Sean Welleck, Ilia Kulikov, Jaedeok Kim, Richard Yuanzhe Pang, and Kyunghyun Cho. 2020. Consistency of a recurrent language model with respect to incomplete decoding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5553–5568.
- Polina Zablotskaia, Du Phan, Joshua Maynez, Shashi Narayan, Jie Ren, and Jeremiah Liu. 2023. On uncertainty calibration and selective generation in probabilistic neural summarization: A benchmark study. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2980–2992.
- Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, and 1 others. 2022. OPT: Open pre-trained transformer language models.
- Jinzhao Zhou, Yiqun Duan, Fred Chang, Thomas Do, Yu-Kai Wang, and Chin-Teng Lin. 2024a. BELT-2: Bootstrapping EEG-to-language representation alignment for multi-task brain decoding.
- Jinzhao Zhou, Yiqun Duan, Yu-Cheng Chang, Yu-Kai Wang, and Chin-Teng Lin. 2024b. BELT: Bootstrapped EEG-to-language training by natural language supervision. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 32:3278–3288.

A Interpretation of Latent Diffusion Effects

This appendix provides an interpretation of how latent diffusion influences representation stability in the proposed framework. The discussion is intended to clarify the effect of diffusion-based pretraining on latent refinement, rather than to establish a general theoretical equivalence between diffusion and explicit regularization.

A.1 Latent Smoothing via Diffusion Noise

During pretraining, an input observation x is mapped to a stochastic latent sequence $\mathbf{Z}_0 \sim q_\phi(\mathbf{Z} \mid x)$. A forward diffusion process perturbs this latent sequence as

$$\mathbf{Z}_\tau = \sqrt{\bar{\alpha}_\tau} \mathbf{Z}_0 + \sqrt{1 - \bar{\alpha}_\tau} \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \quad (15)$$

where $\tau \in \{1, \dots, N\}$ and $\bar{\alpha}_\tau$ follow the same noise schedule defined in Section 4.2. This process introduces Gaussian noise over multiple diffusion steps, producing corrupted latent representations that preserve coarse semantic structure while differing in fine-scale perturbations.

The denoising network $\epsilon_\theta(\cdot)$ is trained to recover structured latent sequences from these noisy variants under semantic conditioning $\mathbf{C}$ (defined in Section 4.2). Rather than enforcing proximity to a fixed prior distribution, diffusion-based pretraining promotes robustness by discouraging overly sharp or brittle latent configurations. As a result, nearby latent samples tend to be mapped to similar refined representations, reducing sensitivity to small perturbations in the initial latent sequence $\mathbf{Z}_0$.

A.2 Interpretation through Score-Based Denoising

Diffusion models admit an interpretation through score-based denoising, where the learned denoiser ϵ_θ approximates the gradient of the log-density of corrupted samples. Under this view, denoising guides latent representations toward regions of higher conditional density, without explicitly constraining them to match a predefined prior.

This perspective suggests that diffusion implicitly shapes the geometry of the latent space by favoring smooth, semantically aligned representations conditioned on $\mathbf{C}$. While this effect resembles the stabilizing role often attributed to KL regularization in variational models, diffusion does not impose an explicit distributional constraint. Instead,

structure emerges through iterative noise injection and denoising within the representation space.

A.3 Discussion

Taken together, these observations indicate that latent diffusion exerts a stabilizing influence by smoothing latent variability and promoting semantically coherent representations. This effect arises from denoising dynamics learned during pretraining, rather than from explicit prior matching. In the proposed framework, diffusion-based latent refinement offers a practical mechanism for improving representation stability prior to autoregressive decoding.

B Stability Analysis Under Latent Perturbations

This appendix analyzes the sensitivity of autoregressive sequence generation to perturbations in the conditioning memory. Section 4.3 defines encoder-side memory $\mathbf{M} = \text{Proj}_{\text{enc}}(\mathbf{Z}^*) \in \mathbb{R}^{L \times d_{\text{model}}}$, which is consumed by the decoder through cross-attention at every decoding step. The analysis characterizes how differences between two memories $\mathbf{M}$ and $\mathbf{M}'$ translate into differences between the induced sequence distributions.

Setup. Let $p_\psi(\mathbf{y} \mid \mathbf{M})$ denote the autoregressive decoder distribution:

$$p_\psi(\mathbf{y} \mid \mathbf{M}) = \prod_{t=1}^T p_\psi(y_t \mid y_{<t}, \mathbf{M}), \quad (16)$$

where $\mathbf{y} = (y_1, \dots, y_T)$ and T denotes the generated sequence length.

Proposition B.1 (Chain rule decomposition for memory sensitivity). *For any two conditioning memories $\mathbf{M}$ and $\mathbf{M}'$, the sequence-level KL divergence decomposes as*

$$\begin{aligned} & \text{KL}(p_\psi(\cdot \mid \mathbf{M}) \parallel p_\psi(\cdot \mid \mathbf{M}')) \\ &= \sum_{t=1}^T \mathbb{E}_{\mathbf{y}_{<t} \sim p_\psi(\cdot \mid \mathbf{M})} \left[\text{KL}(p_\psi(\cdot \mid \mathbf{y}_{<t}, \mathbf{M}) \parallel p_\psi(\cdot \mid \mathbf{y}_{<t}, \mathbf{M}')) \right]. \end{aligned} \quad (17)$$

Proof. Write $\mathbf{y} = (\mathbf{y}_{<t}, y_t, \mathbf{y}_{>t})$ and expand the KL divergence:

$$\text{KL}(p_\psi(\cdot | \mathbf{M}) \| p_\psi(\cdot | \mathbf{M}')) = \mathbb{E}_{\mathbf{y} \sim p_\psi(\cdot | \mathbf{M})} \left[\log \frac{p_\psi(\mathbf{y} | \mathbf{M})}{p_\psi(\mathbf{y} | \mathbf{M}')} \right]. \quad (18)$$

Using the autoregressive factorization in Eq. (16),

$$\log \frac{p_\psi(\mathbf{y} | \mathbf{M})}{p_\psi(\mathbf{y} | \mathbf{M}')} = \sum_{t=1}^T \log \frac{p_\psi(y_t | \mathbf{y}_{<t}, \mathbf{M})}{p_\psi(y_t | \mathbf{y}_{<t}, \mathbf{M}')}.$$
(19)

Substituting Eq. (19) into Eq. (18) and exchanging summation and expectation yields

$$\begin{aligned} \text{KL}(p_\psi(\cdot | \mathbf{M}) \| p_\psi(\cdot | \mathbf{M}')) &= \sum_{t=1}^T \mathbb{E}_{\mathbf{y} \sim p_\psi(\cdot | \mathbf{M})} \left[\log \frac{p_\psi(y_t | \mathbf{y}_{<t}, \mathbf{M})}{p_\psi(y_t | \mathbf{y}_{<t}, \mathbf{M}')} \right]. \end{aligned} \quad (20)$$

Conditioning on the random history $\mathbf{y}_{<t}$ gives

$$\begin{aligned} &\mathbb{E}_{\mathbf{y} \sim p_\psi(\cdot | \mathbf{M})} \left[\log \frac{p_\psi(y_t | \mathbf{y}_{<t}, \mathbf{M})}{p_\psi(y_t | \mathbf{y}_{<t}, \mathbf{M}')} \right] \\ &= \mathbb{E}_{\mathbf{y}_{<t} \sim p_\psi(\cdot | \mathbf{M})} \left[\mathbb{E}_{y_t \sim p_\psi(\cdot | \mathbf{y}_{<t}, \mathbf{M})} \left[\log \frac{p_\psi(y_t | \mathbf{y}_{<t}, \mathbf{M})}{p_\psi(y_t | \mathbf{y}_{<t}, \mathbf{M}')} \right] \right] \\ &= \mathbb{E}_{\mathbf{y}_{<t} \sim p_\psi(\cdot | \mathbf{M})} \left[\text{KL}(p_\psi(\cdot | \mathbf{y}_{<t}, \mathbf{M}) \| p_\psi(\cdot | \mathbf{y}_{<t}, \mathbf{M}')) \right]. \end{aligned} \quad (21)$$

which substituted into Eq. (20) yields Eq. (17). $\square$

Proposition B.2 (A Lipschitz-type bound). *Assume the decoder satisfies the following history-wise bound for some constant $C_{\text{lip}} > 0$:*

$$\text{KL}(p_\psi(\cdot | \mathbf{y}_{<t}, \mathbf{M}) \| p_\psi(\cdot | \mathbf{y}_{<t}, \mathbf{M}')) \leq C_{\text{lip}} \|\mathbf{M} - \mathbf{M}'\|_F, \quad \forall t, \forall \mathbf{y}_{<t}. \quad (22)$$

Then the induced sequence distributions satisfy

$$\text{KL}(p_\psi(\cdot | \mathbf{M}) \| p_\psi(\cdot | \mathbf{M}')) \leq T C_{\text{lip}} \|\mathbf{M} - \mathbf{M}'\|_F. \quad (23)$$

Proof. Apply Proposition B.1 and upper bound each summand using Eq. (22):

$$\begin{aligned} &\text{KL}(p_\psi(\cdot | \mathbf{M}) \| p_\psi(\cdot | \mathbf{M}')) \\ &= \sum_{t=1}^T \mathbb{E}_{\mathbf{y}_{<t} \sim p_\psi(\cdot | \mathbf{M})} \left[\text{KL}(p_\psi(\cdot | \mathbf{y}_{<t}, \mathbf{M}) \| p_\psi(\cdot | \mathbf{y}_{<t}, \mathbf{M}')) \right] \\ &\leq \sum_{t=1}^T \mathbb{E}_{\mathbf{y}_{<t} \sim p_\psi(\cdot | \mathbf{M})} [C_{\text{lip}} \|\mathbf{M} - \mathbf{M}'\|_F] \\ &= T C_{\text{lip}} \|\mathbf{M} - \mathbf{M}'\|_F. \end{aligned} \quad (24)$$

$\square$

Connection to latent refinement. Section 4.2 defines inference-time refinement via the single-step denoised estimate in Eq. (12), i.e., $\mathbf{Z}^* = R_\theta(\mathbf{Z}_0)$ for a fixed τ_{ref} , and memory projection $\mathbf{M} = \text{Proj}_{\text{enc}}(\mathbf{Z}^*)$. The following corollary links refinement-induced perturbation reduction in latent space to stability in the induced sequence distribution.

Corollary B.3 (Perturbation reduction through refinement). *Let $\mathbf{Z}_0$ and $\mathbf{Z}'_0$ denote two initial latent sequences. Define $\mathbf{Z}^* = R_\theta(\mathbf{Z}_0)$ and $\mathbf{Z}'^* = R_\theta(\mathbf{Z}'_0)$ with R_θ defined by Eq. (12), and define $\mathbf{M} = \text{Proj}_{\text{enc}}(\mathbf{Z}^*)$ and $\mathbf{M}' = \text{Proj}_{\text{enc}}(\mathbf{Z}'^*)$. Assume a contraction property for some $\rho \in (0, 1)$:*

$$\|R_\theta(\mathbf{Z}_0) - R_\theta(\mathbf{Z}'_0)\|_F \leq \rho \|\mathbf{Z}_0 - \mathbf{Z}'_0\|_F. \quad (25)$$

Then

$$\|\mathbf{M} - \mathbf{M}'\|_F \leq \|\text{Proj}_{\text{enc}}\| \rho \|\mathbf{Z}_0 - \mathbf{Z}'_0\|_F, \quad (26)$$

and under the assumption in Proposition B.2,

$$\begin{aligned} &\text{KL}(p_\psi(\cdot | \mathbf{M}) \| p_\psi(\cdot | \mathbf{M}')) \leq \\ &T C_{\text{lip}} \|\text{Proj}_{\text{enc}}\| \rho \|\mathbf{Z}_0 - \mathbf{Z}'_0\|_F. \end{aligned} \quad (27)$$

Proof. Since Proj_{enc} is a linear projection applied token-wise,

$$\|\mathbf{M} - \mathbf{M}'\|_F = \|\text{Proj}_{\text{enc}}(\mathbf{Z}^* - \mathbf{Z}'^*)\|_F \leq \|\text{Proj}_{\text{enc}}\| \|\mathbf{Z}^* - \mathbf{Z}'^*\|_F, \quad (28)$$

where $\|\text{Proj}_{\text{enc}}\|$ denotes the operator norm induced by the Frobenius norm. Apply Eq. (25) to obtain Eq. (26). Combine Eq. (26) with Proposition B.2 to obtain Eq. (27). $\square$

C Implementation Details

During pretraining, the EEG encoder produces latent representations $z \in \mathbb{R}^{B \times 56 \times 512}$ via the reparameterization trick. A forward diffusion process with 1,000 diffusion steps under a sqrt noise schedule is applied, and a denoising U-Net is trained to recover the original latent. The U-Net includes Transformer layers with 8 attention heads and applies cross-attention between EEG and text embeddings. The encoder and U-Net share weights across pretraining and decoding. The full model contains approximately **930 M** trainable parameters, comprising the EEG encoder (**30 M**), the denoising U-Net (**493 M**), and the BART-large decoder (406 M). Pretraining uses a batch size of 32, learning rate of $2e-5$, and runs for up to 100 epochs with early stopping. In the main decoding stage, training proceeds in two phases: first, the language model is frozen and only the encoder and U-Net are updated; second, all components are jointly trained. Both phases use a batch size of 32, learning rate of $2e-5$, and early stopping, with epoch limits of 100 and 50 respectively. All experiments run on Intel Xeon Gold 6330 CPUs and NVIDIA RTX A6000, A6000 ADA GPUs. A single training run (pretraining followed by both decoding phases) takes approximately **12 GPU hours** on one A6000. The model is implemented in PyTorch, and evaluation is performed using the nltk and rouge libraries. All scientific artifacts used in this work are publicly released for research use: ZuCo (Hollenstein et al., 2018, 2020) is released for non-commercial research use; BART-large is distributed via the HuggingFace Transformers library under the Apache License 2.0. Our use of these artifacts is consistent with their intended research purpose.

D Qualitative Analysis of Generation Behavior

This appendix provides additional qualitative examples to illustrate typical generation behaviors of the proposed method under latent uncertainty. Rather than demonstrating exact sentence recovery, the examples are intended to characterize how semantic structure, entity-level information, and syntactic templates are preserved or transformed during generation. Both successful paraphrasing and characteristic failure modes are included to clarify the inductive biases and limitations of the model. Examples are shown under both teacher-forced (TF) and teacher-forcing-free (Non-TF) decoding when

applicable.

- Target:** An amateurish, quasi-improvised acting exercise shot on ugly digital **video**.
Prediction: interesting actor, unreligiousprovised, performance, through a, **video**.
 $\Rightarrow$ The lexical item **video** is preserved, and semantically related terms such as actor and performance indicate partial thematic alignment, despite degraded grammatical structure.
- Target:** When Bush was six **years old**, the **family moved to Houston, Texas**.
Prediction: he was a **years old**, he **family moved to New, Texas**.
 $\Rightarrow$ Although the named entity and location differ, the age-related phrase and family relocation template are preserved.
- Target:** An action/thriller of the finest kind, evoking memories of Day of the Jackal, The French Connection, and Heat.
Prediction: The movie is a mixture of horror and comedy.
 $\Rightarrow$ The predicted sentence substitutes genres while retaining a multi-genre descriptive structure, reflecting partial thematic reinterpretation.
- Target:** However, **the U.S. Navy** accepted him in September of that year.
Prediction: He was elected to **the U.S.** House of Representatives in 1894, serving from 1895 until his death in 1904.
 $\Rightarrow$ Institutional structure is preserved through substitution, replacing military admission with political election while maintaining selection-related semantics.
- Target:** In the spring of 1941, Kennedy volunteered for the **U.S.** Army, but was rejected.
Prediction: In 1962, he was elected to the **U.S.** House of Representatives as a member of the Democratic Party.
 $\Rightarrow$ Temporal markers and institutional context are preserved, while the underlying event is substituted, indicating concept-level restructuring.
- Target:** Every individual will see the movie through the prism of beliefs, but the sense that

- peace is possible remains.
Prediction: The film is a romantic comedy about the power of love.
 ⇒ Abstract affective meaning is retained through paraphrasing, replacing peace with a related notion of emotional resolution.
7. **Target:** Mary Lilian Lucy Josepha Monique Bael **was born in** Highbury, London, England, one of eight children of Henri Bael.
Prediction: He was born in New York City, the son of a Russian Jewish immigrant father and an American mother.
 ⇒ Birthplace and parentage structure is preserved, despite changes in name and location.
 8. **Target:** Everything its title **implies**, a standard-issue crime drama from the Tinseltown assembly line.
Prediction: about characters **implies** movie Hollywood drama.
 ⇒ Despite strong disfluency, key topical cues such as **implies** and Hollywood drama are retained.
 9. **Target:** He also created the Defense Intelligence **Agency**.
Prediction: the National **Intelligence Agency**.
 ⇒ Partial entity-level recall is observed through preservation of institutional keywords.
 10. **Target:** He was first appointed to fill the Senate **seat** of Ernest Lundeen.
Prediction: was elected to the position **seat** of John Hemy.
 ⇒ Structural elements such as succession and office holding are preserved with altered entities.
 11. **Target:** He is the great great great grandfather of **President George W. Bush**.
Prediction: he served as Chief of Staff to **President George W. Bush**.
 ⇒ The named entity is preserved, while familial relations are replaced by institutional association.
 12. **Target:** After earning **his degree**, Bush **went to work** in a finance-related division.
Prediction: after **his degree** he **went to work** in an insurance department.
 ⇒ Career progression structure is preserved, with substitution of the specific professional domain.
 13. **Target:** **He was** the son of a blacksmith and **was born in New York**.
Prediction: **He was born in New York City**.
 ⇒ Core biographical facts are preserved with reduced detail.
 14. **Target:** Joseph H. Ball was an American politician.
Prediction: He was elected to the United States House of Representatives.
 ⇒ Occupational category is preserved through contextual substitution.
 15. **Target:** He served in the **United States Senate** as a Republican.
Prediction: he served in the **United States Army**.
 ⇒ National affiliation is preserved, while institutional role is substituted.
 16. **Target:** He later became an educator at the **University** of the District of Columbia.
Prediction: He graduated from Princeton University.
 ⇒ Academic context is preserved, with role substitution between employment and education.
 17. **Target:** The picture does not know it is a comedy.
Prediction: the film is unclear.
 ⇒ Abstract evaluative meaning is partially preserved through synonym substitution.
 18. **Target:** He took a job in real estate with a self-made American millionaire.
Prediction: he was a stockbroker in New York City.
 ⇒ Economic activity and national context are preserved with domain substitution.