SiDe: Long-Sequence Electrophysiological Generation with Raw-Signal Detail Recovery

Anonymous authors
Paper under double-blind review

Abstract

Electrophysiological recordings often become long sequences due to long temporal spans or high sampling rates. Yet most biosignal generation studies remain in short-window settings, and scaling them to long sequences is not a simple extrapolation. Without compression, modeling a large number of raw time points in the full signal space quickly becomes computationally expensive; with front-end latent compression, high-frequency statistics and local signal details may be suppressed before generation begins. We study where compression should occur: long-range structure can be modeled in compressed tokens, but the final raw-space velocity should not be forced to recover local detail only through these tokens. We propose SiDe, a Signal-space Detailer that keeps compression in a conditional Transformer backbone while providing an output-side raw-signal path for velocity prediction. On pose-conditioned EMG generation, SiDe reduces rFID from 7.363 to 2.414 compared with a JiT-style baseline, and better preserves frequency profiles and common EMG statistical feature distributions. The generated EMG also follows pose-speed manipulations and matches real-EMG degradation trends under controlled perturbations. Additional experiments on Sleep-EDF EEG and PTB-XL ECG further characterize where raw-signal detail recovery helps across modalities.

1 Introduction

Diffusion and flow models have brought generative modeling into electrophysiological signals, including EEG, ECG, and EMG, with applications in data augmentation, cross-modal generation, signal imputation, denoising, and conditional synthesis (Chen et al., 2025; Shome et al., 2024; Yao et al., 2025b; Wang et al., 2026a; Luo & Cai, 2025; Liu et al., 2026; Jiang et al., 2026; Wang et al., 2026b). Most existing work, however, still operates in a short-sequence regime, where the temporal window is short or the signal has been substantially downsampled and preprocessed; representative settings are summarized in Appendix A.

Many real electrophysiological recordings are long sequences, with high sampling rates, long temporal contexts, or both. kHz-level sEMG and ECoG expose fine local waveforms and high-frequency structure, while PSG and wearable recordings often require modeling dependencies over much longer contexts (Salter et al., 2024; Sivakumar et al., 2024). This makes long-sequence electrophysiological generation more than a linear extrapolation of short-window biosignal synthesis: the model must generate many raw time points while preserving signal fidelity, conditional consistency, and computational scalability.

This regime creates a tension between computation and signal fidelity. A full raw-space backbone, such as a 1D U-Net or DiT over very long sequences, preserves raw information but quickly becomes expensive as sampling rate and window length grow (Peebles & Xie, 2023; Yao et al., 2025a). The latent recipe from image generation offers a natural alternative, since mature VAE infrastructure lets image diffusion models operate efficiently in compressed spaces (Rombach et al., 2022; Zheng et al., 2025). For electrophysiological signals, however, this transfer is not automatically safe: high-frequency components in EMG, EEG, and some ECG structures often carry physiological meaning (De Luca, 1997; Uhlhaas & Singer, 2006; Goldberger et al., 1981), and the field does not yet have image-scale validated VAEs, RAEs, or perceptual representations (Oquab et al., 2023; Zhang et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

2018; He et al., 2022). Latent compression should therefore be treated as a modeling assumption to diagnose, not a free component to reuse.

The central question is not simply whether to compress, but where compression should occur. To control computation, the backbone can model long-range structure and conditional dependencies in a compressed token space. Yet the local waveforms and high-frequency statistics required by the final output should remain aligned with raw space. Following recent pixel-space diffusion and flow designs (Li & He, 2026; Ma et al., 2026b;a; Yu et al., 2026; Chen et al., 2026; Wang et al., 2025), we view the failure mode in which the final velocity must recover many raw time points only through an overly narrow representation as a **bottlenecked velocity path**. We adapt this output-side recovery perspective to long-sequence electrophysiological signal generation.

Driven by this view, we propose **SiDe (Signal-space Detailer)**, a signal-space architecture shown in Figure 1. SiDe uses a compressed Transformer backbone to model long-range temporal structure and conditional dependencies, and connects it to a local detailer that combines the noisy raw signal with the backbone representation to predict the final velocity in raw space. Thus, SiDe keeps compression where it is computationally useful, while preserving an output-side path for raw-signal detail recovery.

We evaluate SiDe on long-window pose-conditioned EMG generation with EMG2Pose (Salter et al., 2024). SiDe improves rFID from 7.363 to 2.414 over a JiT-style fixed-detailer baseline, better preserves raw-signal statistics such as log-PSD, waveform length, and zero-crossing rate, and responds consistently to pose-speed manipulations. Perturbation-matching experiments further show that generated EMG matches real-EMG degradation trends under controlled perturbations. Additional experiments on Sleep-EDF EEG and PTB-XL ECG examine the generality and boundary of this design across electrophysiological modalities (Kemp et al., 2000; Wagner et al., 2020).

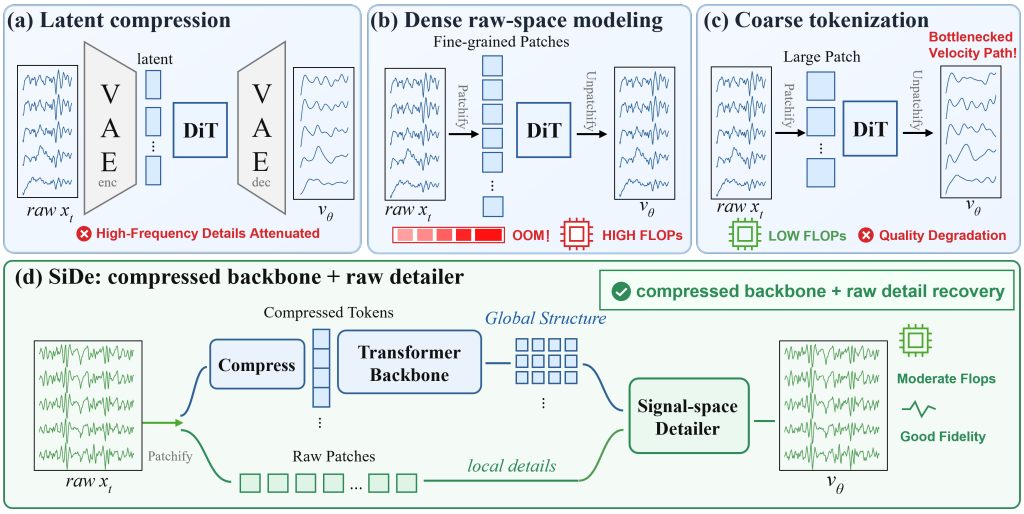


Figure 1: Overview of long-sequence electrophysiological generation routes. SiDe keeps a compressed backbone while preserving a raw signal-space recovery path.

2 Why Long-Sequence Electrophysiological Signals Are Different

The design space of long-sequence electrophysiological signal generation. We first specify the generation setting considered in this paper. By *long-sequence electrophysiological signal generation*, we refer to tasks in which a single generated sample contains many raw time points. Given channels C, sampling rate f_s, and window duration L, the raw output size is $N = C \times f_s \times L$. In our main setting, kHz-level sampling and long temporal windows expose local waveforms, high-frequency components, and long-range conditional dependencies simultaneously. This regime changes both the information to preserve and the computation needed to preserve it.

A direct scaling route is to keep the diffusion or flow backbone in raw signal space. This preserves full raw-signal information, but its cost grows with N, and fixed-patch Transformers also increase token count as $M = N/p$. Similar compute gaps are observed in high-resolution image generation, where compressed token or latent representations are much cheaper than full pixel-space backbones (Li & He, 2026). Other routes reduce the single-forward burden but introduce different assumptions: chunk-wise or autoregressive generation raises sampling cost, cross-chunk continuity, and error-accumulation issues, while progressive or multiscale generation assumes that physiological signals can be reliably decomposed into coarse structure and residual detail.

Latent or tokenized generation offers the most transferable compression route from image and audio generation. It reduces backbone compute by placing an autoencoder or tokenizer before the generative model, but this also delegates part of the target distribution to the compressor: if spectra, local waveforms, or physiological statistics are shifted before generation begins, the downstream model learns an already altered target. We therefore study where compression can be introduced while preserving the raw-signal statistics required for generation fidelity.

A normalized-spectrum view. We next view long electrophysiological signals in normalized frequency, measured in cycles/sample rather than Hz. This describes how much variation must be reconstructed between adjacent raw samples and allows datasets with different sampling rates to be compared on the same discrete scale.

Figure 2 compares ImageNet images with EMG, EEG, and ECG using normalized spectra and cumulative spectral energy; dataset and fitting details are provided in Appendix B. Natural images provide a low-frequency-dominated reference, whereas EMG and EEG retain substantially more energy at high normalized frequencies. ECG is smoother overall, although sharp events such as QRS complexes still carry high-frequency components. Thus, long-sequence electrophysiological generation jointly requires long-context modeling and the preservation of local waveforms and high-frequency statistics, especially for EMG and EEG. Latent compression and output-side recovery should therefore be treated as modeling assumptions to validate, rather than default engineering choices.

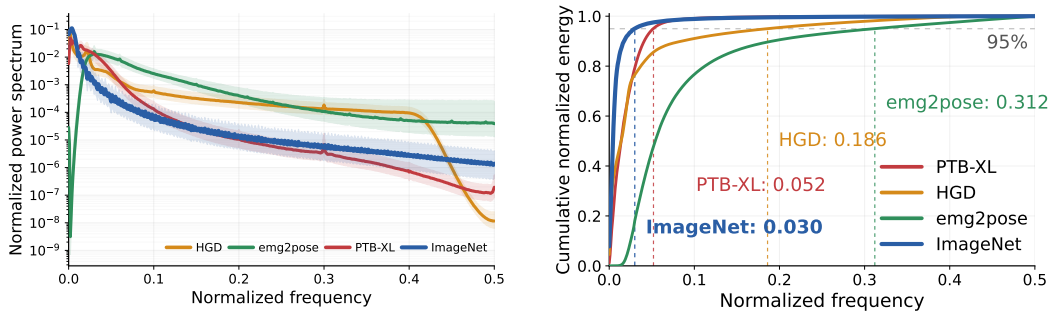


Figure 2: Normalized spectral diagnostics across images and electrophysiological signals. Left: normalized power spectra. Right: cumulative normalized spectral energy.

VAE reconstruction as a pre-generation diagnostic for latent routes. We therefore use VAE reconstruction as a pre-generation diagnostic for latent routes. Specifically, we test whether a given compressor provides a sufficiently faithful front end for the long-sequence EMG setting studied here. The resulting diagnosis is specific to the evaluated compressors and setting. Reconstruction quality can reveal whether the latent representation has already shifted away from raw-signal statistics before generation begins; generation quality is evaluated separately in downstream experiments.

We train VAEs with overall compression ratios of 4×, 8×, and 16× on 12000-sample EMG windows. Beyond reconstruction MSE, we evaluate signal-level distribution metrics on non-overlapping 1000-sample segments: feature MMD over time-domain features, band-wise log-PSD, and channel correlation; logPSD L1 over the 20–500 Hz EMG band; and Wasserstein distances for RMS, MAV, waveform length, and ZCR.

Even the mildest $4\times$ VAE has low reconstruction MSE (0.0292), but its logPSD L1 is 0.3753, over $11\times$ the train-vs-val reference of 0.0336; its feature MMD is about $2\times$ the reference, and waveform length and ZCR are shifted by about $3.3\times$ and $4.5\times$, respectively. Higher compression worsens these shifts. Thus, in this long-sequence EMG setting, latent compression should be diagnosed at the signal level before being treated as free infrastructure. Full VAE diagnostics are provided in Appendix C.

3 Velocity-Path Bottlenecks in Long-Sequence Signal Generation

3.1 High-Dimensional Updates and the Velocity Bottleneck

We refer to the problem considered here as a high-dimensional update problem. We view a signal sample as $x \in \mathbb{R}^N$. For electrophysiological signals, the raw output size can be written as $N = C \times f_s \times L$, where high sampling rates and long temporal windows both increase the number of raw time points in a sample. Similar issues arise in high-resolution images, long audio, and other continuous signals as the raw output space grows. The question in this section is how a flow or diffusion model can remain computationally scalable while still recovering the local details needed at each raw coordinate.

Although a model may be trained to predict clean data, noise, or velocity, the sampler ultimately needs a velocity-like update direction in the raw output space, e.g.,

$$x_{t-\Delta t} = x_t + \Delta t\, u_\theta(x_t, t, c), \qquad u_\theta(x_t, t, c) \in \mathbb{R}^N.$$

Thus, the key architectural issue is how the prediction target and information path jointly determine the amount of local raw information that reaches the final velocity output.

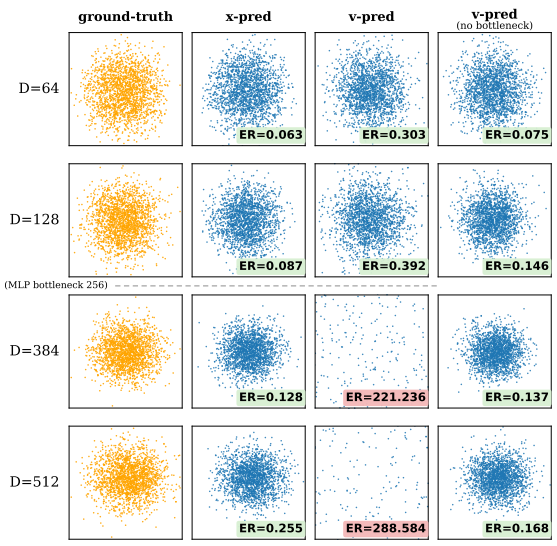
We say that a model has a *bottlenecked velocity path* if the local information from the noisy input x_t to the final velocity prediction must pass entirely through an intermediate representation that is much smaller than the local raw observation. In EMG2Pose, for example, a local raw patch contains $80 \times 16 = 1280$ observed dimensions, while the backbone token hidden size is 384. Directly predicting velocity from this token therefore creates a hard bottleneck: the model must recover a high-dimensional local update from a much narrower representation. Even before this hard boundary is reached, increasing patch size can still hurt performance, as each token must represent a longer raw segment; this is consistent with diffusion Transformer results showing that smaller patches often improve quality at the cost of higher FLOPs and memory (Peebles & Xie, 2023).

We next test this path-based view empirically, then use it to reinterpret x-prediction and learned detailers.

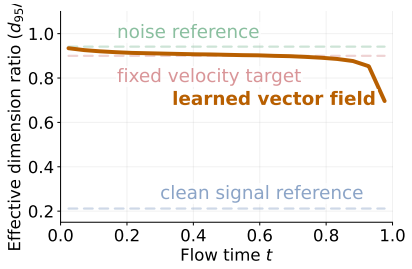
3.2 Diagnosing Bottlenecked Velocity Paths

Controlled bottleneck experiments. We first test the velocity-path bottleneck in a controlled setting. We construct ellipsoid datasets in increasing observed dimensions and scale their intrinsic dimension accordingly, so that the velocity target must be recovered across an increasingly large raw coordinate space. Under the same flow-matching setup, Figure 3(a) compares bottlenecked direct v-prediction, x-prediction, and v-prediction with a raw side path. Direct v-prediction degrades once the observed dimension exceeds the bottleneck capacity; x-prediction remains stable; and adding a raw side path makes v-prediction learnable again. This comparison raises a simple question: if v-prediction also succeeds once the bottleneck is removed, what exactly makes x-prediction more robust in the bottlenecked setting? We will return to this question in Sec. 3.3.

Effective dimension of velocity targets. We next ask whether real EMG velocity targets are high-dimensional enough for this bottleneck to matter. Since flow or diffusion moves samples from full-space Gaussian noise toward a data manifold, the local velocity field may still have broad support in raw coordinates. We therefore form x_0, ϵ, and the linear flow-matching target $v = \epsilon - x_0$ on EMG patches, and measure PCA95 effective dimension for both target-side quantities and the learned velocity field $\hat{v}$ at different time bins.



(a)



(b)

Pred.	Dim.	Patch size 5	6	8	10	12
v	320	**1.77**	**1.85**	**1.92**	**1.96**	**13.07**
	256	**1.86**	**1.87**	**2.03**	**23.39**	**40.81**
	192	**2.04**	**2.17**	**34.81**	**45.28**	**46.06**
	160	**2.76**	**19.40**	**45.33**	**46.46**	**49.34**
	128	**26.03**	**42.99**	**47.95**	**48.84**	**52.74**
x	128	**2.41**	**3.09**	**3.24**	**3.53**	**4.44**
Proposed	128	**1.96**	**1.93**	**1.88**	**1.79**	**1.80**

(c)

Figure 3: Diagnostics for bottlenecked velocity paths. Left: ellipsoid generation with direct v-prediction, x-prediction, and v-prediction with a raw side path. ER is the mean ellipsoid residual $|\sum_i (z_i/r_i)^2 - 1|$. Right: GrabMyo FID over patch size and hidden width.

Figure 3(b) shows that ϵ and v remain high-dimensional at the patch level, rather than quantities that can be reliably recovered from a strongly compressed token. The learned field varies with time, but the architectural constraint is set by the difficult regions, motivating a non-bottlenecked path to local velocity details.

Bottleneck validation on real EMG. Finally, we test the same failure mode in real EMG generation. On GrabMyo (Pradhan et al., 2022), following the EMGFlow recipe (Jiang et al., 2026), we scan DiT-style models over patch size and hidden width, varying the ratio between local observation dimension and token capacity. We compare direct v-prediction, x-prediction, and a proposed detailer with a raw output-side path.

Figure 3(c) shows a sharp boundary: when hidden width falls below the local patch dimension, direct v-prediction degrades abruptly in FID. x-prediction remains usable but weakens with larger patches, while the proposed detailer stays nearly stable. Together with the ellipsoid results, this systematic boundary across patch-size and hidden-width settings shows that the bottleneck persists in real EMG generation.

3.3 From x-Prediction to Learned Detailers

The three diagnostics above reaffirm that whether the final velocity path is bottlenecked can be a decisive factor in high-dimensional generation, consistent with related diagnoses in recent visual generative models (Zheng et al., 2025). This naturally raises the question of whether the robustness of x-prediction may partly come from avoiding such a bottleneck in a particular form. JiT-style x-prediction is often explained by the fact that clean data lie on a low-dimensional manifold (Li & He, 2026; Ma et al., 2026b). For our purpose, an additional point is crucial: if the model predicts x_θ, the sampling velocity can be recovered through

$$v_\theta = \frac{x_\theta - x_t}{1 - t}.$$

Thus, even when the backbone operates in compressed token space, the final velocity still receives the noisy raw input through a fixed residual path. In this sense, x-prediction acts as a fixed output-side detailer.

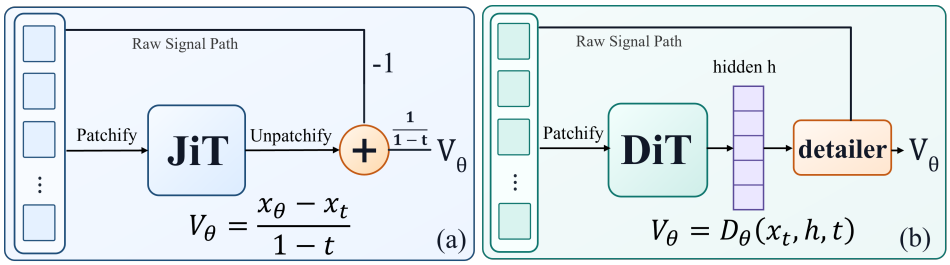


Figure 4: Schematic view of the information paths induced by x-prediction and output-side detailer.

This fixed algebraic detailer is useful but limited. Near the high-noise end, $1/(1-t)$ can amplify errors in x_θ, and the same formula is applied to all timesteps and patches without adapting to local activity or spectral statistics. This motivates replacing the fixed residual rule with a learned output-side recovery module.

Recent raw-space methods can be viewed through a similar lens. DECO, PixelDiT, DiP, and PixNerd (Ma et al., 2026a; Yu et al., 2026; Chen et al., 2026; Wang et al., 2025) differ in decoder design, but are consistent with a common pattern: a compressed backbone models global structure, while a local decoder or detailer restores output-space details near the final prediction. SiDe adapts this idea to long-sequence electrophysiological generation with a lightweight 1D signal-space detailer, which combines the noisy raw EMG patch with the backbone representation to predict the final velocity in raw signal space.

4 Method

Overview. Given pose condition c, diffusion time t, and noisy EMG sample x_t, SiDe predicts a flow-matching velocity $v_\theta(x_t, t, c)$ in raw EMG space. The model has two components: a compressed conditional Transformer backbone for long-range temporal structure and pose-dependent dynamics, and a lightweight signal-space detailer for local raw-space recovery.

We first patch-embed the noisy EMG and pose condition, and compute a compact token representation. The final velocity is then predicted by a local detailer:

$$h = B_\theta(P_x(x_t), P_c(c), t), \qquad \hat{v} = D_\theta(x_t, h, t),$$

where P_x and P_c denote EMG and pose patch embeddings. The backbone aggregates temporal context, pose information, and diffusion-time conditioning in token space, while the detailer explicitly receives x_t. Thus, the output velocity need not be reconstructed only from compressed tokens.

Compressed conditional Transformer backbone. The backbone is a 1D conditional DiT operating on compressed patch tokens. In the main EMG2Pose setting, EMG patch size 80 gives 150 target tokens per window, while pose patch size 120 gives 100 condition tokens. Each EMG token represents an 80×16 raw patch with a 384-dimensional hidden state, giving about $3.3\times$ local dimensional compression before output-side recovery. The backbone models long-range temporal structure and pose-dependent dynamics in this compact token space, using a more advanced DiT configuration following recent high-dimensional diffusion backbones (Li & He, 2026; Yao et al., 2025a); architectural details are given in Appendix D.

Pose conditions enter through token-level cross-attention and global AdaLN conditioning. Because EMG and pose use different patch sizes, their token indices are not directly aligned. We therefore assign RoPE positions by physical time centers rather than token indices, keeping cross-attention temporally aligned while leaving local raw-space recovery to the signal-space detailer.

Local signal-space detailer. The detailer provides a non-bottlenecked output-side path from the noisy raw signal to the final raw-space velocity. For each noisy EMG patch x_t^i and the corresponding backbone token h_i, it predicts a local velocity patch $\hat{v}^i = D_\theta(x_t^i, h_i, t)$. The same detailer is shared across patches, and local predictions are concatenated into the full velocity.

Table 1: Main EMG generation results on the test-user split, averaged over three random seeds. Decoder-based metrics use a frozen EMG2Pose decoder; signal-level metrics follow Table 5.

	Decoder-based			Signal-level					
Model	rFID ↓	Ang. ↓	Lm. ↓	Feat. MMD ↓	PSD L1 ↓	RMS ↓	MAV ↓	WL ↓	ZCR ↓
SiDe	**2.414**	**0.1882**	**13.498**	**0.00297**	**0.1354**	**0.0801**	**0.0529**	**13.667**	**0.0178**
JiT	7.363	0.2120	15.699	0.01250	0.6562	0.2151	0.1376	40.286	0.0639
JiT-768	5.306	0.2028	14.803	0.01980	0.3178	0.2741	0.1691	59.319	0.0462
LFM4x	6.195	0.1926	14.119	0.00390	0.3455	0.1080	0.0657	14.570	0.0358
Real baseline	1.262	0.2110	15.705	0.00213	0.0479	0.0352	0.0247	8.345	0.0043

The module is local by design: cross-patch interaction and pose alignment are handled by the Transformer backbone, while the detailer converts each conditional token into raw-space waveform corrections. We instantiate it as a lightweight patch-space dilated ResNet with about 1.5M parameters. We inject h_i and t through AdaGN modulation, so recovery depends on both the noisy raw patch and the global conditional representation. Architecturally, we use a kernel-9 input stem and kernel-5 residual blocks with dilated convolutions, expanding the local receptive field at low cost. This targets EMG morphology, high-frequency spectral shape, and raw-space alignment.

5 Experiments

5.1 Main Task and Comparison Setup

We evaluate long-window pose-conditioned EMG generation on EMG2Pose (Salter et al., 2024). Given a continuous pose trajectory, the model generates a 16-channel EMG window of length 12000, corresponding to 6 seconds at 2 kHz. We use the user-generalization split; preprocessing and implementation details are in Appendix D.

Our main comparison isolates the output-side recovery route: JiT-style x_0-prediction uses a fixed algebraic raw path, whereas SiDe uses a learned signal-space detailer. We further strengthen the alternatives with JiT-768, which roughly matches SiDe's theoretical MACs using 3.3× the parameter count, and LFM4x, a favorable latent-flow baseline using the least compressed 4× VAE from Sec. 2, a same-scale backbone, and 300 target tokens, twice those of JiT and SiDe.

Other variants are reserved for mechanism and feasibility analysis: raw-space scaling baselines diagnose compute cost, bottlenecked variants test the velocity-path hypothesis, and adapted image detailers test whether a generic learned decoder is sufficient. Appendix E reports the scaling diagnostics, and Sec. 5.4 reports the bottleneck and detailer ablations.

5.2 Main EMG Generation Results

Decoder-based quality. Table 1 reports the main generation results on the test-user split. We use the official Meta open-source `regression_vemg2pose` checkpoint as the frozen EMG2Pose regression model g, and compute rFID from pooled decoder features. Appendix D.2 further evaluates two additional official regression checkpoints to test sensitivity to this decoder choice. Conditional consistency is measured by $E_{\text{cond}} = d(g(\hat{x}), y)$, where d is angle MAE or landmark distance. The real-data baseline reports train-test rFID and the decoder error $d(g(x), y)$ on real test EMG.

SiDe gives the best decoder-based results among generated models. It reduces rFID from 7.363 to 2.414 over JiT and remains better than JiT-768, even though JiT-768 is enlarged to roughly match SiDe's theoretical MACs. SiDe also gives the lowest conditional errors, while LFM4x is competitive on Cond metrics but does not close the rFID gap. This suggests that learned output-side recovery improves decoder-level realism and pose consistency beyond simply increasing JiT capacity or moving generation into a favorable latent setting.

Signal fidelity beyond decoders. Decoder-based metrics measure whether generated EMG can be interpreted by a downstream model. We therefore complement them with signal-level diagnostics from Sec. 2, covering feature distributions, spectral shape, amplitude statistics, and temporal complexity.

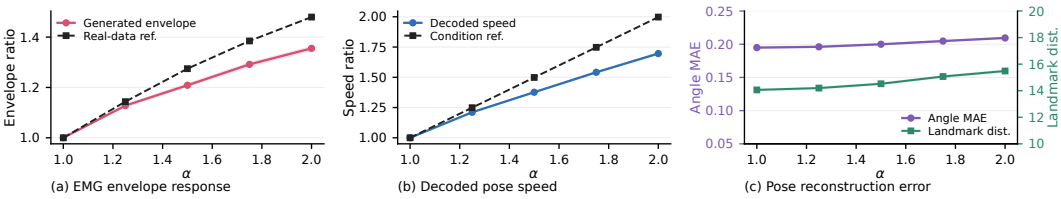


Figure 5: Conditional controllability under pose-speed scaling. Generated EMG responses change coherently with the manipulated motion speed.

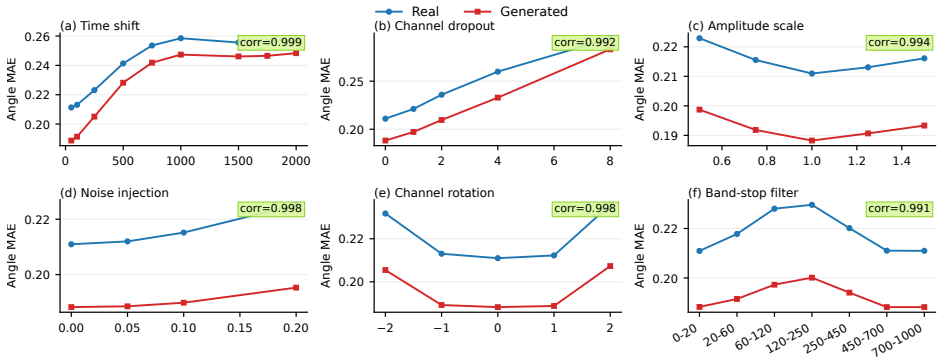


Figure 6: Degradation matching between synthetic and real EMG under controlled perturbations.

SiDe is also the best generated model across all signal-level metrics in Table 1. JiT-768 improves PSD L1 over JiT but worsens several distributional statistics, indicating that more JiT capacity does not by itself recover local signal fidelity. The favorable LFM4x baseline improves several raw statistics over JiT, but still trails SiDe in every signal-level metric despite using the least compressed VAE and twice as many target tokens. Additional diagnostic PSD plots are provided in Appendix D. These results support learned signal-space detail recovery over both fixed algebraic recovery and the strengthened latent route.

5.3 Conditional Controllability

A useful conditional generator should behave as a controllable simulator whose output responds coherently to manipulated pose conditions. We test this behavior through controlled changes to the conditioning trajectory.

In the speed-scaling experiment, we time-warp the conditioning trajectory, i.e., $\mathrm{pose}(t) \mapsto \mathrm{pose}(\alpha t)$, and generate EMG under the modified trajectory. Figure 5 shows that, when the pose motion is accelerated, the generated EMG envelope increases accordingly, matching the speed-envelope relationship observed in real data. The pose trajectory recovered by the EMG2Pose decoder also shows the corresponding speed change, while sample quality, mainly measured by decoded-pose error, degrades only mildly as speed increases. This suggests that the model learns a controllable relation between motion dynamics and EMG activity, rather than only memorizing static pose-signal pairs.

We further evaluate degradation matching. We apply the same perturbations to real and synthetic EMG, including time shift, channel dropout, amplitude scaling, noise injection, and channel rotation, and compare the downstream decoder error curves. As shown in Figure 6, the degradation curves of synthetic and real EMG are highly correlated (> 0.99). This test jointly evaluates whether synthetic EMG remains decodable in the clean condition and reproduces the failure behavior of real EMG under controlled perturbations. The result supports using generated EMG as a realistic test signal for offline decoder evaluation and robustness analysis.

Table 2: Ablations on velocity-path bottlenecks, local receptive field, and adapted image-domain detailers.

Abl.	Var.	rFID ↓	Ang. ↓	Lm. ↓	PSD ↓
Path	JiT-v	53.48	0.4833	34.93	2.685
Path	Bottlenecked	50.14	0.4116	31.63	2.786
RF	ResNet-3	2.391	0.1915	13.65	0.1421
RF	ResNet-4	2.396	0.1899	13.47	0.1100
RF	SiDe	**2.369**	**0.1894**	**13.53**	**0.0825**

Model	rFID ↓	Ang. ↓	Lm. ↓	PSD ↓
DeCo	4.163	0.2064	15.12	0.1655
PixelDiT	2.709	0.2048	14.97	0.1259
PixNerd	6.818	0.2096	15.36	0.4121
DiP	2.770	0.1921	13.76	0.1498
SiDe	**2.369**	**0.1894**	**13.53**	**0.0825**

5.4 Ablation

This section separates two design choices in SiDe: whether the final velocity benefits from a non-bottlenecked raw x_t path, and whether the output module benefits from a 1D signal-space recovery bias. Unless otherwise specified, ablations are reported on the validation split.

Raw path bottleneck. We first disrupt the information path from noisy raw signal to final velocity. *Direct v-pred, JiT config* uses the same compressed-token backbone as the JiT-style baseline but predicts velocity directly; *SiDe, raw path bottlenecked 4×* keeps the SiDe framework but inserts a 4× downsample/upsample bottleneck into the raw path. Both variants degrade substantially in Table 2, showing that compressed tokens alone are insufficient for stable local velocity recovery in this setting, and that a strongly compressed raw path can reintroduce the bottleneck.

Detailer design. We next vary the detailer while keeping the backbone and training protocol fixed. Increasing depth alone gives limited gains, whereas larger kernels and dilation improve logPSD fidelity, supporting the need for a sufficiently large local receptive field. We also compare against 1D adaptations of several image-domain detailers under a shared interface: each receives a noisy raw patch and its corresponding Transformer token. These adaptations isolate the output-side mechanisms that transfer to 1D signals, while the original methods additionally rely on image-specific structures or priors. SiDe performs best overall, indicating that local convolution, dilation, and raw-path conditioning contribute jointly to its gains.

The Sleep-EDF EEG and PTB-XL ECG results in Appendix F indicate that SiDe is most useful when high-frequency and local statistics dominate. Its smaller ECG gains reflect the emphasis of PTB-XL evaluation on low-frequency morphology and event-level measurements, whereas SiDe targets local signal detail and high-frequency statistics.

6 Conclusion

This paper studies where compression should occur in long-sequence electrophysiological signal generation. SiDe keeps compression in a conditional Transformer backbone for long-range structure and pose-dependent dynamics, while preserving an output-side raw-signal path through a lightweight signal-space detailer. On long-window pose-conditioned EMG generation, SiDe improves over JiT across decoder-based and signal-level metrics, and retains this advantage over compute-matched JiT-768 and a favorable LFM4x latent-flow baseline. Conditional manipulation and degradation matching further show that generated EMG responds coherently to changed conditions and exhibits downstream failure modes close to real EMG. Experiments beyond EMG indicate that this design is most useful when local and high-frequency statistics matter, with weaker gains when evaluation is dominated by low-frequency morphology. Overall, our results support a simple design principle: keep compression in the backbone, but preserve an output-side path for signal recovery.

AI Use Statement

In this work, we used generative AI tools to assist with method implementation, translation, feedback on experimental design, interpretation of experimental results, and figure creation. Specifically, generative AI tools implemented most of the software code based on detailed specifications provided

by the authors. The core implementation was manually inspected and verified by the authors. Generative AI tools were also used to translate an author-written Chinese manuscript into English, after which the resulting text was reviewed and partially revised by the authors. In addition, generative AI tools participated in the creation of figures, which were reviewed and revised by the authors as necessary.

Additionally, generative AI tools were used for brainstorming, literature search and summarization, explanation of relevant literature, suggestions of experimental parameters, suggestions regarding the structure of the paper, formatting of references, and language editing.

All AI-assisted outputs were reviewed by the authors as appropriate to their role in the research workflow. The authors take full responsibility for the methodology, implementation, experiments, figures, interpretation of results, and final content of this work, including any text, claims, code, figures, or other artifacts produced with the assistance of generative AI.

Reproducibility Statement

The model formulation is described in Sec. 4. Appendix D provides the data processing, architecture, training, sampling, and evaluation protocols for the main EMG2Pose experiments. The VAE and latent-flow baselines are documented in Appendix C, while random-seed variability, evaluator sensitivity, and compute and memory measurements are reported in Appendices D.1, D.2, and E, respectively. The Sleep-EDF and PTB-XL experimental settings are described in Appendix F. The main experiments were conducted on a single NVIDIA GeForce RTX 5090 GPU. We will publicly release the implementation, configuration files, and evaluation code upon acceptance.

References

Zehua Chen, Yuyang Miao, Liyuan Wang, Luyun Fan, Danilo P. Mandic, and Jun Zhu. Versatile cardiovascular signal generation with a unified diffusion transformer. *Nature Machine Intelligence*, 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01147-y.

Zhennan Chen, Junwei Zhu, Xu Chen, Jiangning Zhang, Xiaobin Hu, Hanzhen Zhao, Chengjie Wang, Jian Yang, and Ying Tai. Dip: Taming diffusion models in pixel space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36136–36146, June 2026.

Xuewei Cheng, Ke Huang, Yi Zou, and Shujie Ma. Sleepegan: A gan-enhanced ensemble deep learning model for imbalanced classification of sleep stages. *Biomedical Signal Processing and Control*, 92:106020, 2024. ISSN 17468094. doi: 10.1016/j.bspc.2024.106020.

Carlo J. De Luca. The use of surface electromyography in biomechanics. *Journal of Applied Biomechanics*, 13(2):135–163, 1997. doi: 10.1123/jab.13.2.135.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. IEEE, 2009.

Ary L. Goldberger, V. Bhargava, V. Froelicher, and J. Covell. Effect of myocardial infarction on high-frequency QRS potentials. *Circulation*, 64(1):34–42, 1981. doi: 10.1161/01.CIR.64.1.34.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Boxuan Jiang, Chenyun Dai, and Can Han. Emgflow: Robust and efficient surface electromyography synthesis via flow matching, 2026.

B. Kemp, A.H. Zwinderman, B. Tuk, H.A.C. Kamphuisen, and J.J.L. Oberye. Analysis of a sleep-dependent neuronal feedback loop: the slow-wave microcontinuity of the eeg. *IEEE Transactions on Biomedical Engineering*, 47(9):1185–1194, 2000. ISSN 1558-2531. doi: 10.1109/10.867928.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 36115–36125, June 2026.

Chen Liu, Can Han, Weishi Xu, Yaqi Wang, and Dahong Qian. Sasg-da: Sparse-aware semantic-guided diffusion augmentation for myoelectric gesture recognition. *IEEE Journal of Biomedical and Health Informatics*, pp. 1–14, 2026. ISSN 2168-2208. doi: 10.1109/JBHI.2026.3670388.

Tian-jian Luo and Zikun Cai. Diffusion models-based motor imagery eeg sample augmentation via mixup strategy. *Expert Systems with Applications*, 262:125585, 2025. ISSN 09574174. doi: 10.1016/j.eswa.2024.125585.

Zehong Ma, Longhui Wei, Shuai Wang, Shiliang Zhang, and Qi Tian. Deco: Frequency-decoupled pixel diffusion for end-to-end image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 43600–43610, June 2026a.

Zehong Ma, Ruihan Xu, and Shiliang Zhang. Pixelgen: Pixel diffusion beats latent diffusion with perceptual loss, 2026b.

Yuta Nambu, Masahiro Kohjima, and Ryuji Yamamoto. Cardioflow: Learning to generate ecg from ppg with rectified flow. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10888856.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Ashirbad Pradhan, Jiayuan He, and Ning Jiang. Multi-day dataset of forearm and wrist electromyogram for hand gesture recognition and biometrics. *Scientific Data*, 9(1):733, 2022. doi: 10.1038/s41597-022-01836-y.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Sasha Salter, Richard Warren, Collin Schlager, Adrian Spurr, Shangchen Han, Rohin Bhasin, Yujun Cai, Peter Walkington, Anuoluwapo Bolarinwa, Robert Wang, Nathan Danielson, Josh Merel, Eftychios A. Pnevmatikakis, and Jesse D. Marshall. emg2pose: A large and diverse benchmark for surface electromyographic hand pose estimation. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, 2017. doi: 10.1002/hbm.23730.

Debaditya Shome, Pritam Sarkar, and Ali Etemad. Region-disentangled diffusion model for high-fidelity ppg-to-ecg translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38 (13):15009–15019, 2024. ISSN 2374-3468. doi: 10.1609/aaai.v38i13.29422.

Viswanath Sivakumar, Jeffrey Seely, Alan Du, Sean R. Bittner, Adam Berenzweig, Anuoluwapo Bolarinwa, Alexandre Gramfort, and Michael I. Mandel. emg2qwerty: A large dataset with baselines for touch typing using surface electromyography. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2024.

Peter J. Uhlhaas and Wolf Singer. Neural synchrony in brain disorders: Relevance for cognitive dysfunctions and pathophysiology. *Neuron*, 52(1):155–168, 2006. doi: 10.1016/j.neuron.2006.09.020.

Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. Ptb-xl, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020. ISSN 2052-4463. doi: 10.1038/s41597-020-0495-6.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Shijian Li, and Gang Pan. Eegdiffuser: Label-guided eeg signals synthesis via diffusion model for bci applications. *Neurocomputing*, 670:132636, 2026a. ISSN 09252312. doi: 10.1016/j.neucom.2026.132636.

Shuai Wang, Ziteng Gao, Chenhui Zhu, Weilin Huang, and Limin Wang. Pixnerd: Pixel neural field diffusion. In *The Fourteenth International Conference on Learning Representations*, 2025.

Xiaoda Wang, Minxiao Wang, Kaiqiao Han, Defu Cao, Ching Chang, Yidan Shi, Runze Yan, Xiao Luo, Yan Liu, Xiao Hu, Yizhou Sun, Wei Wang, and Carl Yang. Pg-lrf: Physiology-guided latent rectified flow for electro-hemodynamic ppg-to-ecg generation, 2026b.

Jingfeng Yao, Bin Yang, and Xinggang Wang. Reconstruction vs. generation: Taming optimization dilemma in latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15703–15712, 2025a.

Weiheng Yao, Zhihan Lyu, Mufti Mahmud, Ning Zhong, Baiying Lei, and Shuqiang Wang. Catd: Unified representation learning for eeg-to-fmri cross-modal generation. *IEEE Transactions on Medical Imaging*, 44(7):2757–2767, 2025b. ISSN 1558-254X. doi: 10.1109/TMI.2025.3550206.

Yongsheng Yu, Wei Xiong, Weili Nie, Yichen Sheng, Shiqiu Liu, and Jiebo Luo. Pixeldit: Pixel diffusion transformers for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14273–14282, June 2026.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 586–595, 2018.

Boyang Zheng, Nanye Ma, Shengbang Tong, and Saining Xie. Diffusion transformers with representation autoencoders. In *The Fourteenth International Conference on Learning Representations*, 2025.

Table 3: Representative raw dimensional regimes in prior biosignal generation settings. The counts are approximate and intended to indicate scale.

Modality	Typical generation setting	Channels	Sampling/window regime	Approx. raw scalars
EEG	Trial- or epoch-level synthesis/augmentation	1–64	100–256 Hz, 3–30 s	0.8k–60k
ECG	ECG generation or PPG-to-ECG translation	1–12	100–500 Hz, 4–10 s	0.5k–12k
EMG	Gesture-level sEMG synthesis/augmentation	12–16	2 kHz or 2048 Hz, 0.1–1 s	0.1k–12k

A Representative Dimensional Regimes in Biosignal Generation

This appendix summarizes representative dimensional regimes in prior biosignal generation work. The goal is not to provide an exhaustive survey, but to make explicit the scale difference between common short-window settings and the long-sequence regime studied in this paper.

EEG generation and augmentation. Prior EEG generation studies often operate on short trial- or epoch-level segments, such as Sleep-EDF epoch synthesis or label-conditioned BCI EEG augmentation (Wang et al., 2026a; Cheng et al., 2024; Luo & Cai, 2025). These settings typically use 1–64 channels, sampling rates around 100–256 Hz, and windows of 3–30 s. The resulting raw sample count is usually on the order of 0.8k–30k scalars, with larger cases reaching roughly 60k scalars.

ECG generation. ECG generation and translation studies often focus on 12-lead PTB-XL-style ECG windows or wearable PPG-to-ECG translation (Wang et al., 2026b; Shome et al., 2024; Nambu et al., 2025). These settings typically involve 1–12 channels, 4–10 s windows, and sampling rates from 100–500 Hz, often with downsampling to 100 Hz. Their raw sample count is therefore commonly around 0.5k–12k scalars.

EMG generation and augmentation. Existing EMG generation work is usually evaluated on gesture-level sEMG synthesis or augmentation benchmarks (Liu et al., 2026; Jiang et al., 2026). Although EMG is often sampled at 2 kHz or 2048 Hz, the windows are usually short, around 0.1–1 s, and channel counts are commonly 12–16. This yields roughly 0.1k–12k raw scalars per sample.

B Additional Normalized-Spectrum Diagnostics

Datasets and spectrum rule. The normalized-spectrum comparison in Figure 2 uses ImageNet validation images (Deng et al., 2009), EMG2Pose sEMG (Salter et al., 2024), the High-Gamma Dataset (HGD) EEG (Schirrmeister et al., 2017), and PTB-XL ECG (Wagner et al., 2020). ImageNet provides a natural-image reference; EMG2Pose, HGD, and PTB-XL provide electrophysiological signals with different sampling rates and local waveform structure. For images, we convert center crops to grayscale, z-score each crop, apply a 2D Hann window, compute FFT2 power, radially average the spectrum, and normalize by total spectral energy. For 1D signals, we z-score each segment, apply the corresponding Hann-windowed FFT power spectrum, average over channels when applicable, and normalize by total energy. All spectra are plotted against normalized frequency in cycles/sample. Under this rule, the radially averaged ImageNet spectrum is close to $1/f^{2.2}$, consistent with its role as a low-frequency-dominated reference.

ImageNet SDVAE reconstruction diagnostic. We additionally test whether a mature image-domain VAE preserves the same normalized spectral statistics. We use the common Stable Diffusion VAE checkpoint `vae-ft-mse-840000-ema-pruned.ckpt` from `stabilityai/sd-vae-ft-mse`. ImageNet validation images are preprocessed with the same center-crop rule used for the ImageNet spectrum in Figure 2. We reconstruct each crop deterministically by taking the encoder posterior mode, without posterior sampling, and save both the original crop and reconstruction as lossless PNGs. This yields 48627 paired images; 1373 validation images are skipped because their short side is below 256 pixels. The pixel-level reconstruction error is moderate for this setting, with mean MSE 0.0052673 in RGB $[0, 1]$ space and mean PSNR 24.7792 dB.

We then recompute spectra for the original crops and reconstructions using the same ImageNet spectral rule: grayscale conversion, per-image z-score, 2D Hann window, FFT2, radial averaging,

and total-energy normalization. Figure 7 shows the median spectrum with interquartile ranges. The mean-spectrum L1 is small (2.08×10^{-5}), but the mean log-spectrum L1 remains 0.3988, with the visible gap concentrated toward high normalized frequencies. Thus, even a well-engineered image VAE can introduce measurable high-frequency spectral shifts while remaining visually useful. This supports our main caution: latent compression may be acceptable for perceptual image generation, but its spectral effects should be diagnosed before being transferred to electrophysiological signals where high-frequency statistics are part of the target distribution.

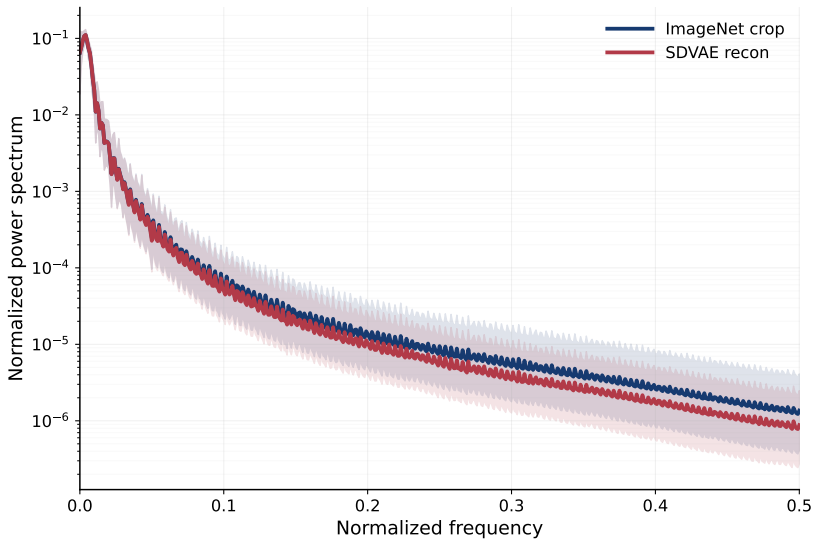


Figure 7: ImageNet validation spectrum before and after deterministic SDVAE reconstruction. Median normalized spectra and interquartile ranges are computed from 48627 original-reconstruction crop pairs.

C VAE Reconstruction Diagnostics

This appendix provides the full reconstruction diagnostics for the VAE compressors discussed in Sec. 2. The compression ratios are overall ratios, not temporal-only ratios.

C.1 VAE Architecture and Training Details

We use a 1D convolutional residual VAE as the EMG compressor. The input is a 16×12000 EMG window, corresponding to a 6 s segment sampled at 2 kHz. The encoder consists of a 1D convolutional stem, strided temporal downsampling stages, and residual blocks. The decoder mirrors this design with nearest-neighbor temporal upsampling, 1D convolutions, and residual blocks to recover the original temporal length. Each residual block uses GroupNorm, SiLU activations, and a residual skip connection. Both the VAE input and output are in EMG waveform space, and the decoder directly reconstructs a 16×12000 EMG signal.

The encoder and decoder widths are $(48, 96, 160, 224, 224)$. The encoder uses $(1, 1, 2, 2, 2)$ residual blocks across stages, and the decoder uses $(1, 1, 1, 1, 0)$ residual blocks. The encoder predicts the posterior mean and log-variance, and the latent channel count is fixed to 32. Different compression ratios are obtained by changing the temporal downsampling factor. Since the latent channel count expands from 16 to 32, the overall compression ratio is the temporal downsampling factor divided by 2. Table 4 summarizes the resulting latent shapes and parameter counts.

Table 4: VAE compressor configurations. Compression denotes the overall scalar compression ratio relative to the 16×12000 EMG input.

Compression	Temporal downsample	Raw shape	Latent shape	Parameters
$4\times$	$8\times$	16×12000	32×1500	3.18M
$8\times$	$16\times$	16×12000	32×750	3.22M
$16\times$	$32\times$	16×12000	32×375	3.27M

The VAE is trained only on the EMG2Pose training split. The validation split is used for reconstruction diagnostics, and the test split is not used for VAE training or checkpoint selection. Preprocessing follows the main generation experiments: EMG is percentile-clamped at 0.999 and then z-scored per channel. During VAE training, we use the same light channel-rotation augmentation as the main experiments, with probability 0.15 and candidate rotations $[-1, 0, 1]$. Validation reconstructions do not use random augmentation.

The training objective combines a time-domain L1 reconstruction loss, a multi-resolution STFT loss, and KL regularization:

$$\mathcal{L} = \mathcal{L}_{\mathrm{L1}}(x, \hat{x}) + \lambda_{\mathrm{stft}} \mathcal{L}_{\text{MR-STFT}}(x, \hat{x}) + \beta_t \mathcal{L}_{\mathrm{KL}}.$$

The MR-STFT loss uses FFT sizes $[128, 256, 512]$, hop sizes $[32, 64, 128]$, and window lengths $[128, 256, 512]$. The VAEs used for the main diagnostics and for LFM4x use $\lambda_{\mathrm{stft}} = 0.05$. The KL weight follows a linear warmup from 0 to its target value. The target β values for the $4\times$, $8\times$, and $16\times$ VAEs are 3×10^{-3}, 1.2×10^{-3}, and 5×10^{-4}, with warmup lengths of 2000, 20000, and 20000 steps, respectively.

We found VAE training to be sensitive to numerical precision and posterior parameterization. Early runs with bf16 AMP were less stable, especially when the KL term increased, and could produce abnormal amplification in the posterior mean or reconstruction. We therefore train the VAEs in FP32. We also avoid hard clamping as the main constraint on posterior statistics: hard clamps, as well as unconstrained parameterizations, gave inferior reconstruction behavior in early experiments. The final configuration uses a smooth bounded parameterization for the posterior statistics, with the log-variance constrained to $[-6, 2]$. For validation diagnostics, we use deterministic reconstruction from the posterior mean, $\hat{x} = D(\mu)$, to avoid injecting sampling noise into the compressor evaluation.

All VAEs are trained for 80000 steps with batch size 64 using AdamW, weight decay 1×10^{-4}, and betas $(0.9, 0.99)$. The learning rate uses warmup followed by cosine decay. The $4\times$ VAE uses a peak learning rate of 1×10^{-4}, while the $8\times$ and $16\times$ VAEs use 2×10^{-4}; the minimum learning rate is 2×10^{-5} for all three models. We use gradient clipping with maximum norm 1.0 and maintain an EMA with decay 0.999. All diagnostics report the final 80000-step checkpoint rather than selecting the best checkpoint by validation metrics.

The three VAEs are close to a validation plateau at 80000 steps. From 60000 to 80000 steps, the $4\times$ VAE reconstruction MSE decreases from 0.03195 to 0.02924, and logPSD L1 decreases from 0.3899 to 0.3753. The $8\times$ VAE MSE decreases from 0.04988 to 0.04852, while logPSD L1 remains essentially unchanged at 0.4316. The $16\times$ VAE MSE decreases from 0.1160 to 0.1126, and logPSD L1 decreases from 0.7068 to 0.7007. These trends indicate that the reported checkpoints are not under-trained in an obvious way; longer training may change the numbers slightly, but is unlikely to alter the main trend across compression ratios.

C.2 VAE Reconstruction Diagnostics

Table 5 reports the reconstruction diagnostics used to evaluate the compressor before training a latent generator. In addition to pointwise reconstruction MSE, we use the same signal-level metrics as the main generation experiments. The train-vs-val row compares real training and validation windows under the same metrics, and serves as a reference for the natural split-level variation in the data.

The $4\times$ VAE has low reconstruction MSE, but its signal-level statistics remain noticeably shifted from real EMG. Its logPSD L1 is 0.3753, more than $11\times$ the train-vs-val reference of 0.0336. Its feature MMD is about $2.1\times$ the reference, while waveform length and ZCR Wasserstein distances are about $3.3\times$ and $4.5\times$ the reference, respectively. Higher compression further increases these

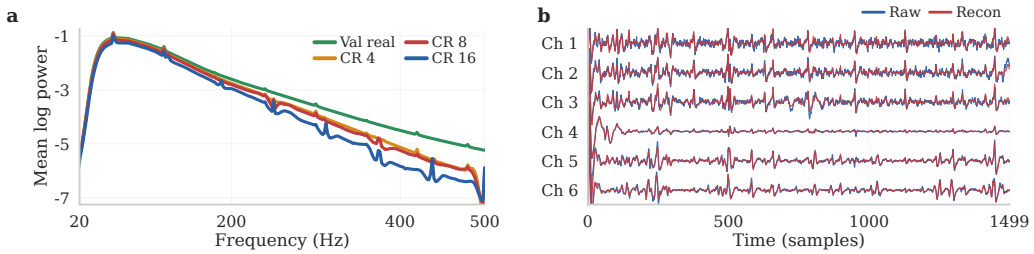


Figure 8: VAE reconstruction diagnostics for long-window EMG. Mean log-PSD profiles and representative waveform overlays show that the VAE preserves coarse waveform structure but introduces systematic spectral shifts, which become larger at higher compression ratios.

deviations. This indicates that the main limitation of the VAE is not only pointwise error; the compressor has already changed spectral and local-complexity statistics before the generative model is trained.

Table 5: VAE reconstruction diagnostics under different overall compression ratios.

VAE compression	KL ↓	Recon MSE ↓	Feature MMD ↓	logPSD L1 ↓	RMS W ↓	MAV W ↓	WL W ↓	ZCR W ↓
4×	**2.3390**	**0.0292**	**0.00437**	**0.3753**	**0.0173**	**0.0121**	**23.757**	**0.0333**
8×	3.1990	0.0485	0.00797	0.4317	0.0251	0.0197	31.349	0.0553
16×	3.6918	0.1126	0.01902	0.7007	0.0527	0.0422	47.112	0.0625
Train vs Val	–	–	0.00205	0.0336	0.0574	0.0411	7.264	0.00747

C.3 Spectral Diagnostics

Figure 8 provides the visual counterpart to Table 5. The waveform overlays show that the VAEs can preserve coarse temporal structure, especially at 4× compression. However, the mean log-PSD profiles reveal a systematic spectral shift in the reconstructed signals, and the shift becomes larger as compression increases. This trend is consistent with the changes in logPSD L1, waveform length, and ZCR in Table 5.

This diagnostic highlights why EMG reconstruction quality cannot be summarized by pointwise error alone. A compressor may reconstruct the large-scale waveform while attenuating or reshaping mid- and high-frequency components. In the long-window EMG2Pose setting, these local and spectral statistics are part of the target distribution that the generator must reproduce. We therefore treat signal-level reconstruction diagnostics as a prerequisite before using a VAE as the front end of a latent generative model.

C.4 Latent-Flow Baseline Construction

LFM4x is a latent-flow baseline built on the 4× VAE. We use the least compressed VAE because it has the best reconstruction diagnostics among the three compressors in Table 5. This makes LFM4x a favorable setting for the latent route.

We first encode every EMG2Pose training window with the 4× VAE encoder and cache the posterior mean. Each raw EMG window of shape 16×12000 is mapped to a latent tensor of shape 32×1500. Random rotation augmentation is disabled when building the latent cache, so that the cached representation is deterministic. We then normalize the latent cache per channel. For latent channel c, the mean and standard deviation are computed over all training windows and latent time positions:

$$\mu_c = \mathbb{E}_{n,t}[z_{n,c,t}], \qquad \sigma_c = \sqrt{\mathrm{Var}_{n,t}(z_{n,c,t})}.$$

The latent flow model is trained on normalized latents,

$$\tilde{z}_{n,c,t} = \frac{z_{n,c,t} - \mu_c}{\sigma_c + \epsilon}.$$

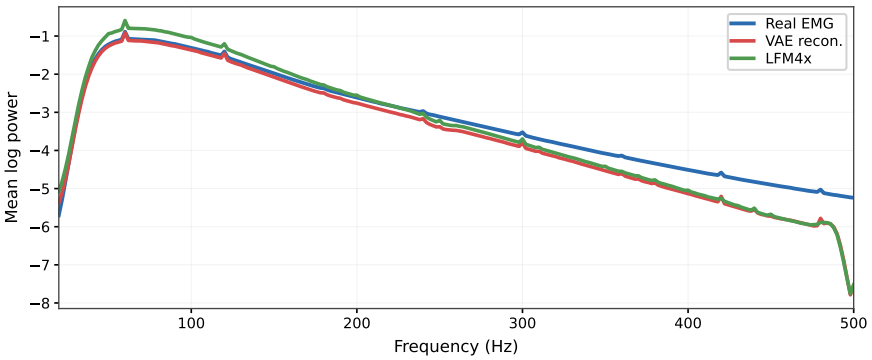


Figure 9: PSD diagnostic for the LFM4x baseline. The mean log-PSD of LFM4x generated EMG follows the 4× VAE reconstruction spectrum, especially at high frequencies, while the VAE reconstruction spectrum remains shifted from real EMG. This indicates that decoded LFM4x quality is constrained by the spectral fidelity of the VAE.

At sampling time, generated latents are first mapped back with the same statistics and are then decoded by the fixed VAE decoder.

The flow backbone uses the same scale and training protocol as the main DiT/JiT models. Pose conditioning follows the same temporal alignment as the raw-space models: the pose trajectory has length 12000 and patch size 120, giving 100 condition tokens. On the target side, the 4× latent has temporal length 1500 and latent patch size 5, giving $1500/5 = 300$ target tokens. This is twice the 150 raw EMG target tokens used by JiT and SiDe.

The fixed VAE decoder maps generated latents back to normalized EMG waveforms of shape 16×12000. The downstream evaluation protocol is then identical to the main experiments. For decoder-based EMG2Pose metrics, generated EMG is inverted from z-score space to the clamp-only raw-amplitude space before being passed to the frozen EMG2Pose decoder; real EMG is evaluated in the same clamp-only space. For signal-level distribution metrics, generated and real EMG are compared in normalized EMG space.

Figure 9 further diagnoses the source of LFM4x errors by comparing the mean log-PSD of real EMG, 4× VAE reconstructions, and LFM4x generated samples. LFM4x follows the VAE reconstruction spectrum more closely than the real EMG spectrum, especially in the high-frequency range. This indicates that the latent flow can model the VAE-induced spectral distribution and can reproduce the high-frequency statistics allowed by the decoder. The remaining gap is therefore not simply a failure of the flow model to learn latent dynamics; it is also inherited from the spectral bias of the fixed VAE decoder.

This observation explains why LFM4x remains below SiDe on signal-level metrics despite using the least compressed VAE, a same-scale backbone, and twice as many target tokens. If the compressor has already shifted the reconstruction spectrum away from real EMG, a latent generator that accurately fits the VAE latent distribution will still decode into a biased waveform distribution. The quality ceiling of a latent-flow baseline is therefore constrained not only by the flow backbone, but also by the spectral fidelity of the VAE.

D Main EMG2Pose Experimental Details

This appendix gives the implementation details for the main long-window pose-conditioned EMG experiments. The goal is to make the data processing, architecture, training, sampling, and evaluation protocol reproducible without repeating the motivation and analysis in the main text.

Dataset and preprocessing. The main experiments use EMG2Pose (Salter et al., 2024) for long-window pose-conditioned EMG generation. Each sample contains 16-channel EMG with length 12000, corresponding to 6 s at 2000 Hz, and the window stride is also 12000. The condition is the time-aligned 20-dimensional joint-angle trajectory. We use the official user-generalization

split. EMG is percentile-clamped at 0.999 and then z-scored per channel. During training, we use the EMG2Pose rotation augmentation with probability 0.15 and candidate rotations $[-1, 0, 1]$. For decoder-based metrics, generated EMG is inverted from z-score space to the clamped raw-amplitude space before being passed to the frozen EMG2Pose decoder; real EMG is evaluated in the same clamp-only space. We verified that this clamp-only protocol does not meaningfully degrade the frozen decoder's validation/test performance.

Backbone and tokenization. The generator uses a 1D continuous conditional DiT. EMG is the target stream and the joint-angle trajectory is the pose-conditioning stream. The EMG patch size is 80, giving 150 target tokens per window; the pose patch size is 120, giving 100 condition tokens. Pose conditions are injected by token-level cross-attention and global AdaLN conditioning. The transformer has hidden size 384, depth 8, 8 attention heads, MLP ratio 2.6666667, SwiGLU MLPs, RMSNorm, QK normalization, and RoPE with $\theta = 10000.0$. For RoPE, positions are assigned by physical time centers rather than token indices: an EMG token covering $[ip_x, (i+1)p_x)$ is placed at $(i + 0.5)p_x$, and a pose token covering $[jp_c, (j+1)p_c)$ is placed at $(j + 0.5)p_c$.

Signal-space detailer. SiDe uses a patch-space dilated ResNet detailer to recover the final velocity in raw EMG space. The detailer has 128 hidden channels, a kernel-9 input stem, and four kernel-5 residual blocks with dilation pattern $[1, 1, 2, 4]$. It uses GroupNorm and injects token/time conditions through AdaGN modulation.

Training and sampling. Training uses flow matching with velocity prediction and velocity loss, with MSE as the base loss. The diffusion time is sampled from a logit-normal distribution with $\mu = -0.8$, $\sigma = 1.0$, and $\epsilon = 10^{-6}$. Classifier-free guidance dropout is 0.1. Models are trained for 200000 steps with batch size 128 using AdamW, learning rate 3.0×10^{-4}, weight decay 1.0×10^{-3}, and betas $[0.9, 0.99]$. The learning rate follows cosine annealing with $t_{\max} = 200000$, 3000 warmup steps, and minimum learning rate 1.0×10^{-6}. Training uses bfloat16 AMP, gradient clipping with maximum norm 1.0, and EMA with decay 0.9995 starting from step 5000. Unless otherwise specified, main experiments and ablations use three random seeds and report the mean. We evaluate the last checkpoint rather than selecting by validation metrics. Sampling uses EMA weights, a 25-step Heun sampler, and classifier-free guidance weight 2.0.

Evaluation. Decoder-based metrics use the frozen official EMG2Pose regression checkpoint as the evaluation model g. For rFID, real and generated EMG are fed into g, and decoder features are average-pooled over time before computing

$$\mathrm{rFID} = \|\mu_r - \mu_g\|_2^2 + \mathrm{Tr}\left(\Sigma_r + \Sigma_g - 2(\Sigma_r\Sigma_g)^{1/2}\right).$$

For the real-data reference row, we compute this FID between randomly sampled training EMG windows and real test EMG windows with the same sample count. This row therefore estimates the split-level FID between independent real EMG samples, rather than a true reconstruction-FID lower bound; in principle, the rFID lower bound for paired reconstruction would be zero. Conditional consistency uses the Cond error defined in the main text, $E_{\mathrm{cond}} = d(g(\hat{x}), y)$, where d is either angle MAE or landmark distance. For signal-level metrics, each 12000-sample window is split into non-overlapping 1000-sample segments. Feature MMD uses RMS, MAV, waveform length, ZCR, slope sign changes, kurtosis, band-wise log-PSD, and channel correlation. For real and generated feature sets X and Y, we compute the unbiased RBF-kernel MMD,

$$\mathrm{MMD}^2(X, Y) = \frac{1}{n(n-1)} \sum_{i \neq j} k(x_i, x_j) + \frac{1}{m(m-1)} \sum_{i \neq j} k(y_i, y_j) - \frac{2}{nm} \sum_{i,j} k(x_i, y_j),$$

where $k(a, b) = \exp(-\gamma\|a-b\|_2^2)$. The bandwidth is set by the median pairwise distance after merging real and generated features, using at most 4096 feature vectors. logPSD L1 uses an orthonormal rFFT and averages the absolute difference between real and generated mean log-power profiles over the 20–500 Hz band. For each scalar statistic s, the Wasserstein metric is $W_1(\{s(x_r)\}, \{s(x_g)\})$ over real and generated segments.

D.1 Random-Seed Variability

The main EMG2Pose comparisons use three independent training seeds. For each run, we evaluate the EMA weights from the final 200000-step checkpoint on the same test-user split, without validation-based checkpoint selection. The main table reports the mean across these runs. Table 6 supplements the SiDe row with its standard deviation across seeds.

Table 6: SiDe generation quality on the EMG2Pose test-user split, reported as mean ± standard deviation over three independent training seeds. Each run uses the final 200000-step checkpoint.

	Decoder-based			Signal-level					
Model	rFID ↓	Ang. ↓	Lm. ↓	Feat. MMD ↓	PSD L1 ↓	RMS ↓	MAV ↓	WL ↓	ZCR ↓
SiDe	2.414 ± 0.176	0.1882 ± 0.0005	13.498 ± 0.051	0.00297 ± 0.00020	0.1354 ± 0.0026	0.0801 ± 0.0015	0.0529 ± 0.0010	13.667 ± 0.044	0.0178 ± 0.0010

The conditional metrics have small run-to-run variation. The relative variation is larger for rFID and feature MMD, which estimate distributional discrepancies from generated samples, but remains limited in scale across the three SiDe runs.

To summarize seed sensitivity across generator families while accounting for the different metric scales, we compute the coefficient of variation separately for each generated model and then average it across SiDe, JiT, JiT-768, and LFM4x. For model m and metric k,

$$\mathrm{CV}_{m,k} = 100\frac{s_{m,k}}{\bar{x}_{m,k}}, \qquad \overline{\mathrm{CV}}_k = \frac{1}{4}\sum_{m=1}^{4} \mathrm{CV}_{m,k},$$

where $s_{m,k}$ and $\bar{x}_{m,k}$ are the standard deviation and mean over three seeds.

Table 7: Random-seed sensitivity of the main EMG2Pose generators. For each metric, we compute the coefficient of variation over three seeds within each generator and report its average across SiDe, JiT, JiT-768, and LFM4x.

	Decoder-based			Signal-level					
Summary	rFID	Ang.	Lm.	Feat. MMD	PSD L1	RMS	MAV	WL	ZCR
Mean CV	4.776%	0.271%	0.209%	7.316%	4.858%	4.916%	4.451%	5.948%	3.613%

We observe broadly similar CVs across generators and therefore report their average as a more stable summary of seed sensitivity. These CVs are substantially smaller than the corresponding relative differences between generator means, supporting the robustness of the main conclusions. This analysis is based on three seeds because of computational constraints, and its statistical resolution is therefore limited.

D.2 Sensitivity to the Frozen EMG2Pose Regressor

The main table uses the official `regression_vemg2pose` checkpoint as the frozen decoder g. To check whether the decoder-based conclusion depends on this single regressor, we additionally evaluate the same generated test samples with two other official EMG2Pose regression checkpoints, `regression_emg2pose` and `regression_neuropose`. The split, generated samples, preprocessing, and Cond error definition are otherwise unchanged.

Table 8 reports the results. The rFID values should be compared within each regressor, not across regressors, because different regressors have different hidden feature scales. Under both additional regressors, SiDe is the best generated model on all three decoder-based metrics. This supports the main conclusion that the decoder-based advantage of SiDe is not an artifact of only using the `regression_vemg2pose` checkpoint.

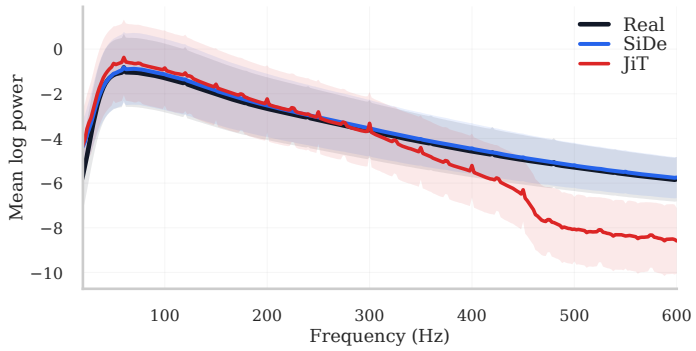


Figure 10: Mean log-PSD overlay for real EMG, SiDe, and JiT on the test-user split.

Table 9: Scaling and feasibility diagnostics on the EMG2Pose (Salter et al., 2024) training setting. FLOPs are theoretical single-forward costs, counting multiply-add as 2 FLOPs; memory is measured on a single GeForce RTX 5090. SiDe uses detailer activation checkpointing.

Model	Tokens	Hidden	Parameters	FLOPs	Memory	Quality
unet1d	none	none	30.00M	43.55 TFLOPs	OOM	N.A
Large-hidden DiT	300	768	114.64M	5.82 TFLOPs	27.59 GiB (near OOM)	N.A
Small-patch DiT	800	384	29.14M	4.64 TFLOPs	OOM	N.A
JiT	150	384	29.94M	817.09 GFLOPs	7.81 GiB	weak
SiDe	150	384	30.92M	2.92 TFLOPs	17.47 GiB	fair

Table 8: Supplementary decoder-based evaluation with two additional official EMG2Pose regression checkpoints: left, `regression_emg2pose`; right, `regression_neuropose`. rFID is comparable within each regressor but not across different frozen regressors.

Model	rFID ↓	Ang. ↓	Lm. ↓
SiDe	**48.3294**	**0.1900**	**13.661**
JiT	429.8623	0.2337	18.250
JiT-768	171.2258	0.2092	15.751
LFM4x	479.6170	0.2125	16.367
Real base.	75.9673	0.2172	16.201

Model	rFID ↓	Ang. ↓	Lm. ↓
SiDe	**0.0086**	**0.2128**	**15.664**
JiT	0.0486	0.2470	18.481
JiT-768	0.0723	0.2421	17.894
LFM4x	0.0398	0.2302	16.817
Real base.	0.0097	0.2263	17.354

E Scaling and Feasibility Diagnostics

Table 9 reports the compute and memory cost of several scaling-oriented designs under the EMG2Pose (Salter et al., 2024) setting. These models are used to characterize feasibility boundaries rather than as primary quality baselines. Directly scaling a full raw-space backbone or substantially reducing patch size quickly leads to high FLOPs and activation memory; increasing hidden width also substantially increases parameter count and memory pressure.

In contrast, JiT and SiDe both keep the token count small. JiT is computationally lighter, but relies on fixed algebraic recovery. SiDe adds a local detailer and raw-space computation, so its FLOPs and memory are higher, but it remains trainable on a single GPU while preserving a learnable recovery path from raw x_t to the final velocity.

Table 10: Sleep-EDF subject-wise split and 10-epoch window statistics.

Split	Subjects	Records	Epochs	10-epoch windows
Train	54	106	138527	68815
Val	8	16	19566	9715
Test	16	31	37376	18557

F Supplementary Experiments on PTB-XL and Sleep-EDF

F.1 Sleep-EDF EEG Generation

We use Sleep-EDF Expanded Sleep Cassette (Kemp et al., 2000) to test whether output-side detail recovery also helps long-horizon EEG synthesis. Sleep staging is annotated at 30 s resolution, but real sleep recordings contain continuous stage structure across consecutive epochs. We therefore construct a label-conditioned generation task in which the model receives 10 consecutive sleep-stage labels and generates the corresponding two-channel EEG segment. Each sample contains 10 epochs, each epoch has 3000 samples at 100 Hz, and the generated window has shape 2×30000, corresponding to 5 minutes of EEG. This task requires the model to follow the label sequence while preserving local spectral and waveform structure within each epoch.

Data and preprocessing. We use the EEG Fpz-Cz and EEG Pz-Oz channels. The labels are W, N1, N2, N3, and REM, with the original S3/S4 labels merged into N3. Movement and unknown stages are removed. Following common sleep-staging preprocessing, wake epochs are cropped to 30 minutes before the first sleep epoch and 30 minutes after the last sleep epoch. Each record is normalized by per-record, per-channel z-score statistics. Because Sleep-EDF does not have a single universally adopted cross-subject split for generation, we use a fixed subject-wise split and keep both nights from the same subject in the same split. Training and evaluation use 10-epoch windows with stride 2 epochs.

Models. All models use the same label-sequence tokenizer. The condition length is 10, with one sleep-stage token per epoch and an epoch-position embedding. Conditions are injected through cross-attention and AdaLN. We compare three same-scale models: JiT, a JiT-style x-prediction baseline with a linear unpatchifier; DiT, a direct v-prediction baseline with the same linear output path; and SiDe, which uses the same v-prediction objective but replaces the linear unpatchifier with a lightweight dilated ResNet detailer. For Sleep-EDF, the detailer uses three residual blocks and hidden width 24, reflecting the two-channel EEG setting while retaining dilation for local temporal context. All three models are optimized with a v-loss. The EEG patch size is 300, so each 10-epoch window gives 100 target tokens. Since each local EEG patch has 300×2 raw dimensions while the hidden size is 256, the direct v-prediction DiT provides a bottlenecked-velocity-path comparison. The models have about 12.8M parameters. Training uses classifier-free guidance dropout 0.1; sampling uses a 25-step Heun sampler with guidance weight 1.25.

Metrics. Generation is performed at the 10-epoch-window level and evaluated after splitting each window into 30 s epochs. We report logPSD L1 between real and generated mean log-power spectra over 0–50 Hz. We also use a frozen Sleep-EDF classifier to compute FID from 256-dimensional intermediate features, inception score (IS), and classifier accuracy score (CAS). CAS compares the classifier prediction for each generated epoch against the conditioning sleep-stage label; the main table reports balanced accuracy, and Table 12 reports per-class accuracy. The classifier is trained only on the real training split. On the validation split, its accuracy is 0.7897 and balanced accuracy is 0.8095, with per-class accuracies 0.8665 for W, 0.5934 for N1, 0.7207 for N2, 0.9867 for N3, and 0.8800 for REM. Because FID, IS, and CAS depend on this frozen classifier, we use them as relative model-comparison metrics rather than as sufficient evidence of clinical realism.

Results. Table 11 reports validation results. The real-data baseline compares a random sample of 9715 training windows against the full validation split, using the same windowing and epoch-level evaluation protocol. This baseline is not an upper bound on generation quality, but it indicates the scale of natural split-level variation. SiDe closely matches the real-data baseline in logPSD

Table 11: Sleep-EDF validation results. Generation is performed on 10-epoch windows and metrics are aggregated at the epoch level.

Model / Baseline	Windows	Epochs	logPSD L1 ↓	FID ↓	IS ↑	CAS bAcc ↑
Train sample vs. val	9715	97150	0.1614	3.6888	2.9447	0.8095
JiT	9715	97150	1.9382	27.5507	2.3154	0.5643
DiT	9715	97150	1.7568	97.8850	1.2931	0.3558
SiDe	9715	97150	**0.1494**	**6.5254**	**2.7221**	**0.8110**

Table 12: Per-class CAS on generated Sleep-EDF epochs.

Model	W	N1	N2	N3	REM
JiT	0.7145	0.1530	0.7737	0.9953	0.1852
DiT	0.9975	0.0535	0.0661	0.6620	0.0000
SiDe	**0.8912**	**0.6123**	**0.8116**	0.9795	**0.7606**

L1 and substantially improves classifier-based realism and label controllability over both baselines. Compared with JiT, SiDe reduces logPSD L1 from 1.9382 to 0.1494 and improves CAS balanced accuracy from 0.5643 to 0.8110. The direct v-prediction DiT performs poorly, especially in FID, IS, and CAS, consistent with the bottlenecked-velocity-path failure mode.

Per-class CAS in Table 12 shows where the controllability improves. JiT reaches high accuracy on N3 but is weak on N1 and REM, suggesting a bias toward more easily separable sleep-stage patterns. DiT shows stronger collapse, with near-failure on N1, N2, and REM. In contrast, SiDe improves N1, N2, and REM while keeping high W and N3 accuracy. This suggests that the generated EEG follows the input label sequence more evenly instead of falling back to a small set of dominant spectral patterns.

Overall, the Sleep-EDF experiment supports the same design principle as the EMG2Pose experiment. The Transformer backbone models the 10-epoch label sequence and cross-epoch dependencies, while the local signal-space detailer recovers epoch-level spectral shape and waveform detail. This separation is useful for EEG because sleep-stage information appears at multiple temporal and spectral scales, including delta activity in N3, spindle- and K-complex-related structure in N2, and weaker mixed patterns in N1 and REM.

F.2 PTB-XL ECG Generation

We next evaluate SiDe on unconditional long-record ECG generation using PTB-XL (Wagner et al., 2020). This experiment is intended as a boundary case rather than a setting where SiDe is expected to dominate every metric. Each sample is a 12-lead, 500 Hz, 10 s ECG record, giving $12 \times 5000 = 60000$ scalar values. The model does not use diagnostic labels, rhythm labels, patient metadata, or other conditions. Unlike beat-level ECG synthesis, lower-sampling-rate window generation, or latent-space ECG generation, this setting directly models the full 500 Hz multi-lead waveform.

Data and preprocessing. We follow the PTB-XL stratified folds: folds 1–8 for training, fold 9 for validation, and fold 10 for testing. Preprocessing only applies z-score normalization using the training-set per-channel mean and standard deviation. We do not apply additional filtering, clipping, or resampling. The main table reports validation results. Each model generates 2183 ECG records, matching the validation split size. As a real-data reference, we sample the same number of training ECGs and compare them with the validation split under the same metrics.

Models. The main comparison uses same-scale SiDe and JiT-style models. Both use patch size 50, hidden or embedding size 256, 8 attention heads, and 100k training steps. Unless otherwise specified, we run three random seeds and report the mean. The last checkpoint is used for evaluation, without validation-based checkpoint selection.

Metrics. We evaluate both signal-level fidelity and ECG event-level morphology. Event metrics are computed on lead II. R peaks are extracted using NeuroKit cleaning and peak detection; sam-

Table 13: PTB-XL 500 Hz unconditional ECG generation on the validation split. Lower is better except for valid ratios.

Model	Beat ↑	Delin. ↑	HR W ↓	RR med. ↓	RR pool ↓	QRS W ↓	QTc W ↓	PR W ↓	PSD L1 ↓	AC L1 ↓
Train-vs-val	0.960	0.832	0.669	0.0075	0.0054	0.992	2.237	1.247	0.045	0.0176
SiDe	0.820	0.793	**1.952**	**0.0185**	0.0268	3.051	8.334	**3.622**	**0.288**	0.0175
JiT	**0.852**	**0.831**	1.987	0.0211	**0.0179**	**2.864**	**8.246**	3.747	1.232	**0.0135**

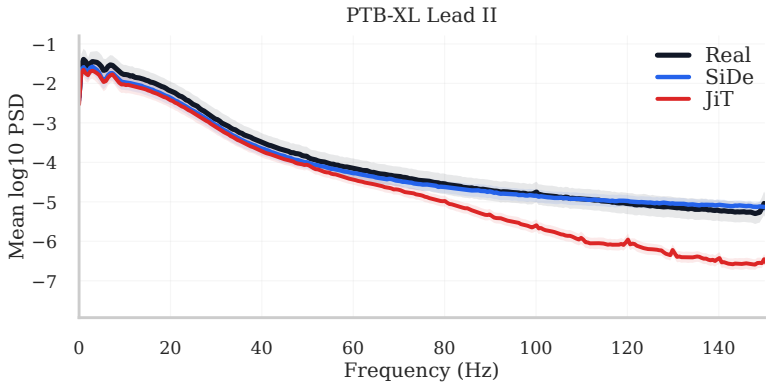


Figure 11: Lead-II PSD comparison for PTB-XL validation generation. SiDe better matches the validation spectrum at higher frequencies, whereas the low-frequency profiles of SiDe and JiT are more similar.

ples are marked invalid if detection fails, fewer than three R peaks are found, or RR intervals fall outside reasonable ranges. The valid beat ratio uses all generated samples as the denominator. HR Wasserstein, sample-level median RR Wasserstein, and pooled RR Wasserstein are computed only over samples with valid R peaks. Delineation metrics further require successful DWT-based fiducial extraction. The valid delineation ratio again uses all generated samples as the denominator, while QRS, QTc, and PR Wasserstein distances aggregate sample-level median intervals only from successfully delineated samples. We also report logPSD L1 over 0.5–150 Hz and autocorrelation L1.

Boundary interpretation. SiDe substantially improves spectral fidelity: its logPSD L1 is 0.288, compared with 1.232 for JiT. Figure 11 visualizes this difference on lead II: the main separation appears above roughly 40 Hz, while low-frequency profiles are more similar. Thus, output-side signal recovery still helps the ECG spectrum. However, this spectral gain does not uniformly transfer to standard ECG morphology metrics. JiT is better on valid beat ratio, valid delineation ratio, QRS W, QTc W, and autocorrelation L1; its valid delineation ratio of 0.831 is close to the train-vs-val reference of 0.832.

This result clarifies a boundary of the method. Many ECG metrics are dominated by low-frequency morphology, event timing, and fiducial extraction. R-peak detection, RR intervals, and QRS/QTc/PR measurements first pass the waveform through a nonlinear event detector before comparing distributions. This evaluation can be relatively insensitive to improvements in high-frequency spectral shape. In this PTB-XL setting, the JiT-style fixed raw path is already sufficient for many event-level metrics, while SiDe mainly improves logPSD fidelity. This is consistent with our broader conclusion: output-side detail recovery is most beneficial when high-frequency and local statistics are directly tied to task quality, as in EMG2Pose and Sleep-EDF, and its gains are more metric-dependent when evaluation is dominated by low-frequency ECG morphology.

G GENERATED SAMPLE VISUALIZATIONS

G.1 EMG

Figure 12 shows representative generated and real EMG windows from the EMG2Pose test-user split. The plots are shown in the normalized model space used for training and are intended as waveform-level visual checks rather than pointwise reconstructions.

G.2 EEG

Figure 13 shows representative generated and real Sleep-EDF EEG epochs. We show one 30 s epoch for each sleep stage used in the conditional generation task. For N2 and N3, examples are selected by simple morphology-based screening to favor visually interpretable stage-specific structure: spindle- or K-complex-like events for N2, and sustained frontal slow waves for N3.

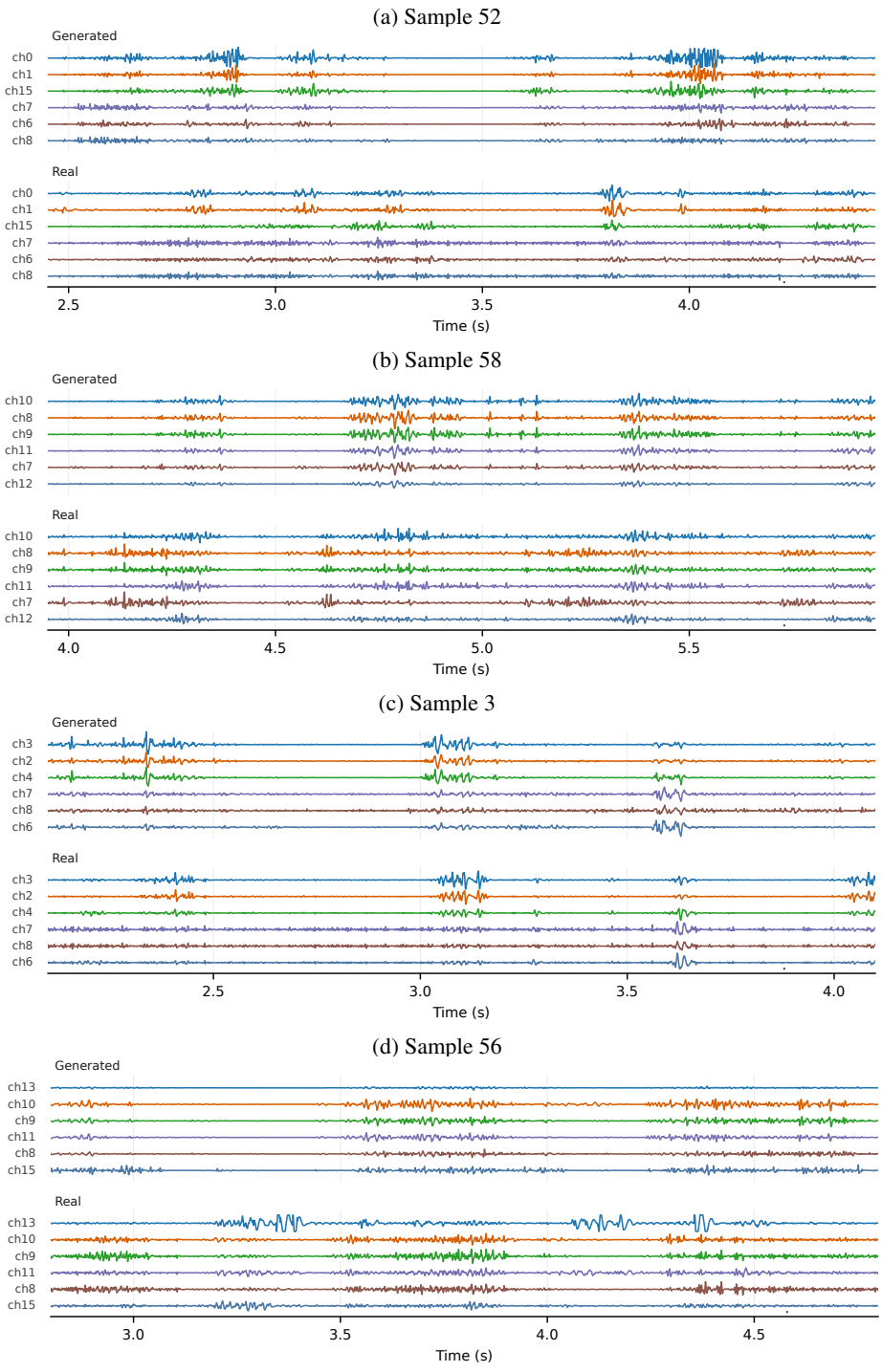


Figure 12: Representative generated and real EMG waveforms on EMG2Pose. Each panel shows six selected channels over a 2 s window in normalized model space.

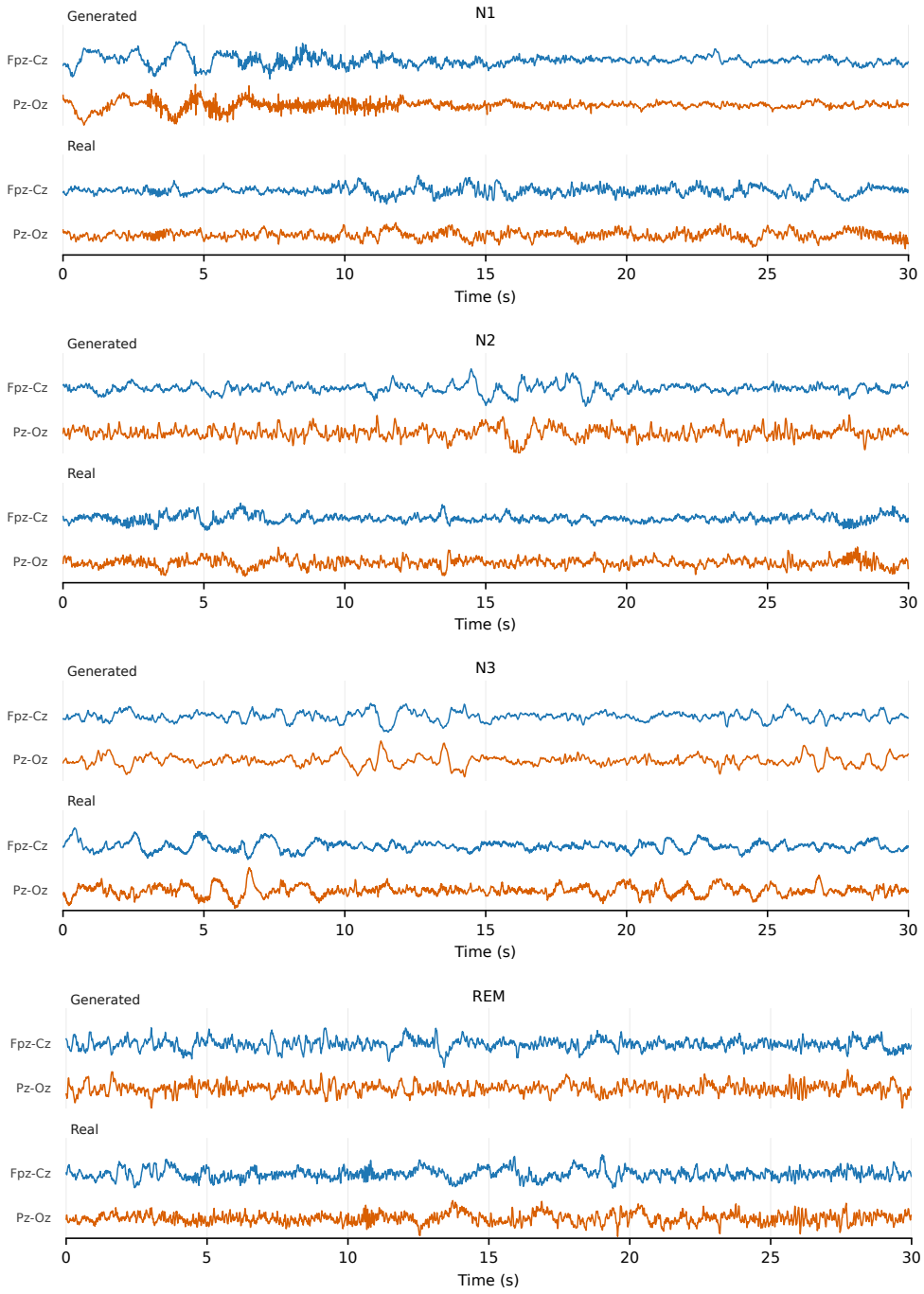


Figure 13: Representative generated and real EEG epochs on Sleep-EDF. Each panel shows the Fpz-Cz and Pz-Oz channels over a 30 s epoch in normalized model space.