NeuroSonic: Conditional Flow Matching for EEG-to-Speech Reconstruction

Wenhao Gao[1⋆], Yifan Wang[1⋆], Yijia Ma[2], Carl Yang[3], Wen Li[2], and Chenyu You[1⋆⋆]

[1] Stony Brook University, Stony Brook, NY, USA
chenyu.you@stonybrook.edu
[2] University of Texas Health Center at Houston, Houston, TX, USA
[3] Emory University, Atlanta, GA, USA

Abstract. Reconstructing continuous speech from scalp electroencephalography (EEG) remains fundamentally challenging. EEG provides a weak, spatially diffuse, and highly variable measurement of distributed cortical activity, whereas speech is organized as a coherent acoustic trajectory with strong harmonic and temporal structure. The resulting mismatch makes waveform regression unstable and causes stochastic multi-step generation to be sensitive to artifact-dependent conditioning and subject variability. We introduce **NeuroSonic**, a conditional flow-matching framework for EEG-to-speech reconstruction. Instead of predicting waveforms directly or refining them through stochastic denoising, NeuroSonic learns a deterministic probability-flow velocity field that transports a noise-corrupted acoustic state toward clean speech under EEG conditioning. EEG and audio are embedded into a shared token space and processed by a time-conditioned gated Transformer that parameterizes the transport ordinary differential equation. This formulation models trajectory evolution explicitly while avoiding iterative stochastic sampling. We evaluate NeuroSonic on the CineBrain and EAV benchmarks under cross-subject evaluation. Across both datasets, the proposed method improves distributional realism, spectral fidelity, and perceptual quality over representative GAN-, diffusion-, and mean-flow baselines, with up to a 26.3% gain in overall perceptual quality. The performance gap is most evident in artifact-heavy segments, where conditioning variability is strongest. These findings indicate that deterministic conditional transport provides a stable and effective formulation for EEG-driven speech reconstruction. Code is available at here.

Keywords: EEG-to-Audio Reconstruction · Neural Speech Decoding · Conditional Flow Matching.

1 Introduction

Reconstructing continuous speech from scalp electroencephalography (EEG) entails coupling two signals with markedly different structure. EEG recordings are

⋆ Equal contribution.
⋆⋆ Corresponding author.

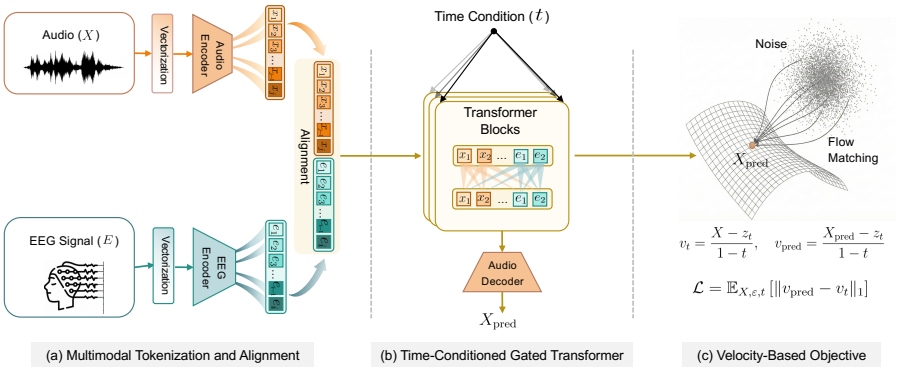

Fig. 1. Overview of NeuroSonic. (a) EEG and audio signals are partitioned into patches, $\{E_i\}$ and $\{X_j\}$, and projected through modality-specific encoders $f_E(\cdot)$ and $f_A(\cdot)$ into a shared latent space for joint modeling. (b) A time-conditioned gated Transformer processes the combined sequence together with a corrupted acoustic state z_t, obtained by interpolating clean audio with Gaussian noise ϵ at time t, along the flow-matching path. Adaptive layer normalization and RMS-stabilized attention are used to preserve stable feature scaling across interpolation times. (c) The velocity-based objective trains the predicted velocity v_{pred}, computed from the predicted clean state X_{pred}, to match the target transport velocity v_t governing acoustic transport under EEG conditioning.

low-amplitude, spatially diffuse projections of distributed cortical sources [19,1]. They exhibit substantial variability across subjects and sessions and are susceptible to motion and physiological artifacts [27,17,24]. In contrast, speech evolves along a highly organized acoustic trajectory characterized by harmonic structure and temporal coherence. The mapping from neural measurements to acoustic realizations is therefore indirect, temporally misaligned, and strongly confounded by nuisance variability. Although EEG-based systems have achieved promising results for constrained vocabulary classification [13], reconstructing natural, continuous speech with high fidelity remains unresolved. Recent EEG foundation models improve transferable representations [5], but continuous speech reconstruction remains unresolved.

Generative modeling offers a principled alternative to discrete decoding [16,4,3,30,29]. However, prevailing paradigms do not fully align with scalp EEG. GAN-based synthesis can become unstable when the conditioning signal is weak or highly variable [8,9]. Diffusion models improve optimization behavior but rely on multi-step stochastic sampling and assume a consistent corruption schedule across timesteps [10,18,23,21]. Under EEG conditioning, these assumptions are challenged by artifact-dependent noise patterns and inter-subject heterogeneity, which can accumulate across sampling steps and degrade reconstruction consistency.

These motivate us to seek a formulation that models acoustic trajectory evolution directly while remaining stable under heterogeneous conditioning. Flow

Matching (FM) provides a continuous-time generative framework in which a neural network learns a velocity field transporting a probability path between distributions [15]. Recent work explores rectified-flow-based latent synthesis to align EEG and speech representations for speech-driven clinical analysis [26]. By parameterizing deterministic probability flows rather than stochastic refinement chains, FM removes the need for iterative denoising and enables conditioning to act on the transport dynamics themselves. Recent EEG generation work further suggests that flow matching is effective for preserving continuous temporal and spectral structure in neural signals [25]. This perspective is suited to speech reconstruction, where temporal coherence is intrinsic to the signal structure.

In this work, we formulate EEG-to-speech reconstruction as conditional acoustic transport. Instead of predicting waveforms in a single step or refining them through stochastic sampling, we learn a deterministic velocity field that maps corrupted acoustic states toward clean speech under EEG conditioning. Building on this formulation, we introduce **NeuroSonic**. As illustrated in Fig. 1, EEG and audio signals are partitioned into patch-level representations and embedded into a shared latent space. A time-conditioned gated Transformer processes the joint sequence to parameterize the probability-flow ordinary differential equation governing acoustic evolution. This design enables global cross-modal interaction while stabilizing feature dynamics across interpolation times, leading to robust reconstruction under artifact corruption and cross-subject variability. (1) We reformulate EEG-to-speech reconstruction as a deterministic, trajectory-aware inverse problem via conditional flow matching [15]. (2) We propose a multimodal tokenization scheme and a time-conditioned Transformer architecture that align neural representations with acoustic dynamics within a shared latent space. (3) We demonstrate consistent improvements over representative GAN-, diffusion-, and mean-flow baselines on public EEG-audio benchmarks, particularly under cross-subject evaluation and artifact-heavy conditions.

2 Method

2.1 Preliminary: Flow Matching for Conditional Transport

Flow Matching formulates generative modeling as learning a continuous-time transport between probability distributions [15]. Let p_0 denote a simple prior and $p_1 = p_{\text{data}}$ the target distribution. FM defines a probability path $\{p_t(x)\}_{t\in[0,1]}$ connecting the two and learns a velocity field that transports samples along this path. Under the linear interpolation path,

$$x_t = (1-t)x_0 + tx_1, \qquad v(x_0, x_1, t) = x_1 - x_0, \qquad \frac{\mathrm{d}x_t}{\mathrm{d}t} = v_\theta(x_t, t), \tag{1}$$

where $x_0 \sim p_0$ and $x_1 \sim p_{\text{data}}$. The neural velocity field v_θ is trained by regressing toward the closed-form target velocity:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{x_0,x_1,t}\left[\|v_\theta(x_t,t) - (x_1 - x_0)\|_1\right]. \tag{2}$$

Integrating the probability flow ODE transports a prior sample to the data manifold at $t = 1$. This deterministic transport formulation removes the need for stochastic denoising and serves as the basis for conditional acoustic modeling.

2.2 NeuroSonic

Conditional Acoustic Transport. We cast EEG-to-speech reconstruction as conditional transport of acoustic trajectories. Given paired EEG–audio samples (E, X), we construct a corrupted acoustic state:

$$z_t = tX + (1-t)\varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{3}$$

and learn a velocity field that transports z_t toward clean speech under EEG conditioning. An overview of the architecture is shown in Fig. 1. EEG and corrupted audio tokens are jointly processed to predict the probability-flow ordinary differential equation governing acoustic evolution. At inference, the learned ODE is integrated from $t = 0$ to $t = 1$ using a fixed-step Heun solver, yielding deterministic reconstruction conditioned on neural activity.

Multimodal Tokenization and Alignment. Let $E \in \mathbb{R}^{C\times T_1}$ and $X \in \mathbb{R}^{T_2}$. EEG and acoustic signals are partitioned into non-overlapping patches: $E \in \mathbb{R}^{C\times T_1}$ and $X \in \mathbb{R}^{T_2}$. Each patch is projected into a shared latent space:

$$e_i = f_E(\mathrm{vec}(E_i)), \qquad x_j = f_A(\mathrm{vec}(X_j)), \tag{4}$$

with embedding dimension d. Learnable modality embeddings and positional encodings are incorporated:

$$\tilde{e}_i = e_i + \tau_E + p_i, \qquad \tilde{x}_j = x_j + \tau_A + p_j. \tag{5}$$

The resulting sequence:

$$Z = [\{\tilde{e}_i\}; \{\tilde{x}_j\}] \in \mathbb{R}^{(N_E+N_A)\times d} \tag{6}$$

enables global cross-modal interaction. Aggregating information through self-attention implicitly attenuates localized motion artifacts and low-SNR perturbations in scalp EEG.

Time-Conditioned Gated Transformer. The sequence Z is processed by L pre-normalized Transformer blocks conditioned on interpolation time t:

$$Z' = Z + g_{\mathrm{msa}} \cdot \mathrm{MSA}(\mathrm{AdaLN}(Z; t)), \tag{7}$$

$$Z = Z' + g_{\mathrm{mlp}} \cdot \mathrm{MLP}(\mathrm{AdaLN}(Z'; t)). \tag{8}$$

where $\mathrm{AdaLN}(U; t) = \gamma_t \odot \mathrm{LN}(U) + \beta_t$, with γ_t, β_t from the time embedding. Global multi-head self-attention is defined as

$$\mathrm{MSA}(Z) = \mathrm{Concat}_{i=1}^{h}\left(\mathrm{Softmax}\left(\frac{Q_i K_i^\top}{\sqrt{d_h}}\right) V_i\right) W^O, \tag{9}$$

Table 1. Objective evaluation of EEG-conditioned speech reconstruction under cross-subject evaluation. Lower values indicate better performance for FAD, LSD, SC, and inference time (seconds). Results are reported as mean ± standard deviation. Best values for each metric are shown in bold.

Dataset	Method	Metrics			
		FAD (↓)	LSD (↓)	SC (↓)	Time (s) (↓)
Cine	MF	173.65 ± 0.26	69.78 ± 0.11	1.34 ± 0.09	0.04
	GAN	57.12 ± 3.49	15.13 ± 0.14	1.25 ± 0.11	**0.02**
	DM	72.56 ± 1.69	22.08 ± 0.07	1.12 ± 0.04	2.00
	Ours	**39.06 ± 0.52**	**14.24 ± 0.31**	**0.64 ±0.04**	0.86
EAV	MF	85.27 ± 0.80	29.25 ± 0.64	1.49 ± 0.06	0.04
	GAN	39.47 ± 0.83	15.71 ± 0.34	1.00 ± 0.01	**0.02**
	DM	15.87 ± 6.78	19.47 ± 0.25	1.25 ± 0.10	2.08
	Ours	**11.64 ± 1.17**	**12.98 ± 0.16**	**0.28 ± 0.02**	1.40

where $d_h = d/h$. Time-dependent interpolation induces feature distribution shifts across t, potentially destabilizing attention logits [28]. To control this effect, we apply per-head RMS normalization to query and key:

$$Q \leftarrow \mathrm{RMSNorm}(Q), \qquad K \leftarrow \mathrm{RMSNorm}(K). \tag{10}$$

The network outputs $X_{\mathrm{pred}} = \mathrm{net}(z_t, t, E)$, from which velocities are derived.

Velocity-Based Objective. Under the manifold assumption [2], clean acoustic signals lie on a low-dimensional structure. Rather than regressing waveforms directly, we supervise transport dynamics in velocity space [14]. Given $\varepsilon = \frac{z_t - tX}{1-t}$, and $v_t = \frac{X - z_t}{1-t}$, the predicted velocity is $v_{\mathrm{pred}} = \frac{X_{\mathrm{pred}} - z_t}{1-t}$. The final objective is

$$\mathcal{L} = \mathbb{E}_{X,\varepsilon,t}\left[\|v_{\mathrm{pred}} - v_t\|_1\right]. \tag{11}$$

Supervising transport in velocity space anchors learning on clean acoustic states and improves robustness under low-SNR and heterogeneous neural conditioning.

3 Experiments

3.1 Dataset

We evaluate NeuroSonic on two publicly available EEG–audio datasets that span controlled conversational recordings and naturalistic audiovisual stimulation. After preprocessing, the combined corpus contains data from 48 subjects, totaling approximately 60 hours of synchronized EEG-audio recordings (49,200 paired segments). CineBrain [6] provides simultaneously recorded EEG and fMRI during continuous audiovisual presentation. The accompanying audio includes speech as well as environmental sounds, yielding acoustically complex reconstruction targets. We follow the original protocol to temporally align EEG signals with the audio stream and reorganize continuous recordings into matched segments. EAV [12] consists of conversational interactions with synchronized EEG, audio,

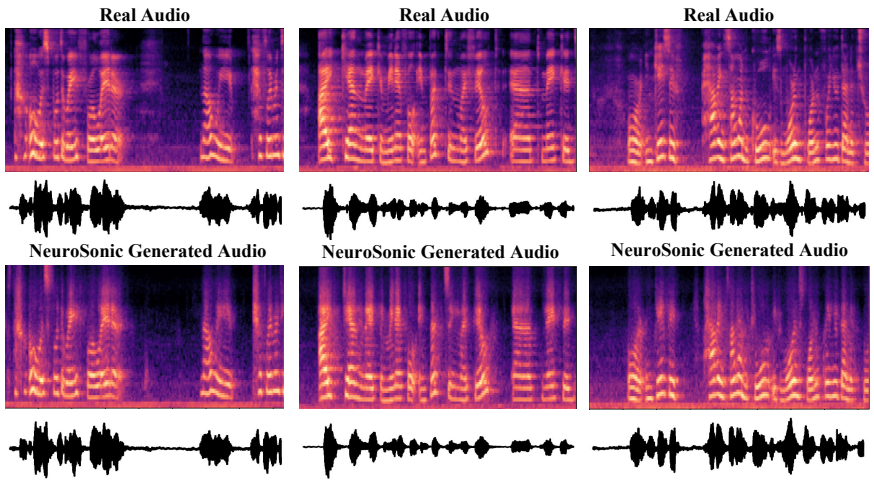

Fig. 2. Comparison of ground-truth speech and NeuroSonic reconstructions. For each example, the reference mel-spectrogram and waveform are shown on top, with the EEG-conditioned reconstruction below. The reconstructed signals exhibit coherent formant trajectories and temporal modulation patterns consistent with the reference.

and video from 42 participants. Compared with CineBrain, EAV contains cleaner speech structure but stronger subject-specific variability arising from spontaneous dialogue and articulation differences.

For both datasets, preprocessing strictly follows the setting in [6,12]. EEG signals undergo standard artifact removal procedures, including MRI-related artifact correction when applicable, 0.1–30 Hz band-pass filtering, 50 Hz notch filtering, and ICA-based removal of ocular, muscular, and cardiac components. All reported results are obtained under cross-subject evaluation, ensuring that test subjects are not observed during training.

3.2 Implementation Details

Setup. NeuroSonic uses a multimodal Transformer with 16 blocks, hidden size 1024, and 16 attention heads. Each block employs RMS-normalized self-attention and a gated MLP with a 4× expansion ratio. To improve robustness to motion artifacts without over-regularizing early feature formation, dropout is applied selectively: attention, projection, and feed-forward dropout are enabled only in the middle blocks, while the earliest and latest blocks remain dropout-free. Models are trained for 400 epochs with batch size 32 using AdamW, cosine learning-rate scheduling, and EMA tracking on an NVIDIA GeForce RTX5090 (32 GB). Inference integrates the learned probability-flow ODE using 100 fixed Heun steps. Dataset-specific window lengths and channel counts follow the original preprocessing protocols [6,12].

Baselines and Evaluation. We compare NeuroSonic with three representative generative paradigms for continuous signal synthesis: GANs [11], diffusion mod-

Table 2. Perceptual evaluation using DNSMOS. Higher values indicate better perceptual quality. Results are reported as mean ± standard deviation. Best results among learned models are shown in bold.

Dataset	Method	DNSMOS		
		SIG (↑)	BAK (↑)	OVRL (↑)
Cine	GT	2.41	1.75	1.67
	MF	1.20 ± 0.01	1.10 ± 0.01	1.10 ± 0.01
	GAN	1.26 ± 0.03	1.29 ± 0.01	1.14 ± 0.01
	DM	1.21 ± 0.00	1.33 ± 0.01	1.07 ± 0.00
	Ours	**1.95 ± 0.04**	**1.64 ± 0.04**	**1.44 ± 0.01**
EAV	GT	3.32	2.77	2.47
	MF	1.19 ± 0.01	1.12 ± 0.01	1.08 ± 0.01
	GAN	1.98 ± 0.01	2.62 ± 0.05	1.45 ± 0.01
	DM	2.92 ± 0.04	2.95 ± 0.10	2.29 ± 0.11
	Ours	**3.31 ± 0.02**	**3.07 ± 0.04**	**2.59 ± 0.03**

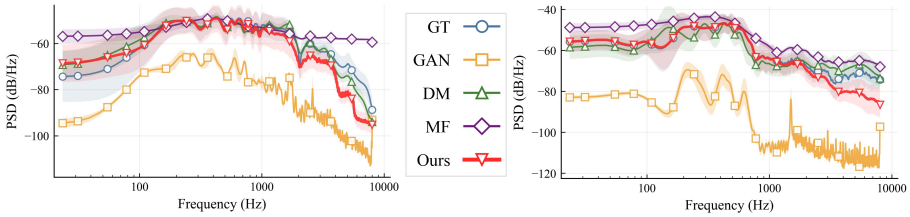

Fig. 3. Power spectral density (PSD) of reconstructed audio on the Cine dataset (*left*) and the EAV dataset (*right*). Ground-truth audio (GT) is shown in blue. NeuroSonic (red) more closely follows the ground-truth spectrum in the low-frequency band and maintains consistent spectral behavior across datasets. GAN outputs exhibit broader spectral deviations, while diffusion models show increased energy in higher-frequency regions.

els [22], and mean flows [7]. Each baseline is adapted to EEG conditioning in the simplest direct form: the GAN generator maps EEG features to waveform outputs; the diffusion model introduces EEG embeddings through cross-attention; and the mean-flow baseline concatenates EEG temporal embeddings as global conditioning. Reconstruction quality is evaluated from four complementary perspectives: distributional alignment via Fréchet Audio Distance (FAD↓), spectral fidelity via Log-Spectral Distance (LSD↓) and Spectral Convergence (SC↓), inference time (seconds↓), and perceptual quality using DNSMOS↑ [20]. Together, these metrics reflect statistical realism, harmonic structure, computational efficiency, and subjective intelligibility.

3.3 Result

Distributional and spectral fidelity. Table 1 reports objective reconstruction quality. NeuroSonic attains the best FAD and LSD on both datasets and yields a substantial reduction in spectral convergence error, indicating improvements that go beyond matching marginal audio statistics and extend to fine-grained

Table 3. Ablation study of clean-state velocity supervision. The x-loss variant replaces velocity supervision with direct waveform regression. Lower values indicate better performance for FAD, LSD, and SC; higher values indicate better perceptual quality. Best values per metric are shown in bold.

Dataset	Method	Metrics					
		FAD (↓)	LSD (↓)	SC (↓)	SIG (↑)	BAK (↑)	OVRL (↑)
Cine	x-loss	**32.23**	14.27	0.91	1.45	1.30	1.18
	Ours	39.06	**14.24**	**0.64**	**1.95**	**1.64**	**1.44**
EAV	x-loss	12.14	13.45	0.90	3.07	2.66	2.28
	Ours	**11.64**	**12.98**	**0.28**	**3.31**	**3.07**	**2.59**

spectral structure. The advantage is most pronounced on Cine, whose audio contains substantial background content and heterogeneous acoustic events, where stochastic baselines are more affected by artifact- and subject-dependent conditioning.

Fig. 3 further illustrates this trend: NeuroSonic aligns most closely with the ground-truth PSD in the low-frequency band that dominates perceived speech quality, while avoiding the high-frequency over-emphasis observed in diffusion baselines and the broadband distortion typical of GAN outputs.

Human-perceptual quality. Table 2 summarizes DNSMOS scores [20]. NeuroSonic achieves the highest SIG and OVRL on both datasets and shows consistent improvement in BAK, suggesting that the reconstructions reduce background interference while preserving intelligible speech structure. On EAV, the reconstruction slightly exceeds the recorded reference in OVRL, driven primarily by higher BAK (3.07 vs. 2.77). A plausible explanation is that EEG reflects neural representations related to speech intent and perception rather than the full acoustic mixture: conditioning on EEG provides weak support for non-linguistic background components, effectively suppressing them in the generated waveform.

Qualitative examples in Fig. 2 are consistent with the perceptual gains: NeuroSonic preserves coherent formant trajectories and temporal modulation patterns without the over-smoothing artifacts typically introduced by regression-style objectives.

Ablation. Table 3 studies the effect of clean-state velocity formulation. The direct x-loss variant achieves competitive FAD, suggesting that endpoint regression can approximate coarse distributional properties. However, it consistently degrades LSD, SC, and all DNSMOS components across both datasets, indicating weaker harmonic organization and temporal coherence. This divergence highlights a key distinction: matching endpoints does not constrain the path from noisy to clean states, whereas velocity supervision explicitly trains the transport dynamics. By anchoring learning on clean-state velocities, NeuroSonic enforces a coherent evolution along the acoustic manifold, which is reflected in improved spectral structure and higher perceived quality.

4 Conclusion

We presented NeuroSonic, a conditional flow-matching approach to reconstruct continuous speech from scalp EEG. By reframing EEG-to-speech reconstruction as deterministic conditional transport, the model learns a probability-flow velocity field that maps corrupted acoustic states to clean speech in a single ODE integration, avoiding stochastic sampling chains that are prone to artifact- and subject-dependent variability. Across two public datasets under cross-subject evaluation, NeuroSonic improves distributional realism, spectral fidelity, and perceptual quality over representative GAN-, diffusion-, and mean-flow baselines. The ablation study further shows that clean-state velocity supervision is essential for preserving spectro-temporal structure, even when endpoint regression can match coarse statistics. These results suggest that trajectory-based conditional transport is a principled and stable direction for neural speech reconstruction, and motivates future work on richer linguistic objectives and broader naturalistic settings.

Disclosure of Interests. The authors declare that they have no competing interests related to this work.

References

1. Bréchet, L., Brunet, D., Birot, G., Gruetter, R., Michel, C.M., Jorge, J.: Capturing the spatiotemporal dynamics of self-generated, task-initiated thoughts with eeg and fmri. Neuroimage **194**, 82–92 (2019)
2. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
3. Chen, N., Liu, F., You, C., Zhou, P., Zou, Y.: Adaptive bi-directional attention: Exploring multi-granularity representations for machine reading comprehension. In: ICASSP. IEEE (2021)
4. Chen, N., You, C., Zou, Y.: Self-supervised dialogue learning for spoken conversational question answering. arXiv preprint arXiv:2106.02182 (2021)
5. Chen, Y., Ren, K., Song, K., Wang, Y., Wang, Y., Li, D., Qiu, L.: Eegformer: Towards transferable and interpretable large-scale eeg foundation model. arXiv preprint arXiv:2401.10278 (2024)
6. Gao, J., Liu, Y., Yang, B., Feng, J., Fu, Y.: Cinebrain: A large-scale multi-modal brain dataset during naturalistic audiovisual narrative processing. arXiv preprint arXiv:2503.06940 (2025)
7. Geng, Z., Deng, M., Bai, X., Kolter, J.Z., He, K.: Mean flows for one-step generative modeling. arXiv preprint arXiv:2505.13447 (2025)
8. Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. Advances in neural information processing systems **27** (2014)
9. Han, K., Xiong, Y., You, C., Khosravi, P., Sun, S., Yan, X., Duncan, J.S., Xie, X.: Medgen3d: A deep generative framework for paired 3d image and mask generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2023)
10. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

11. Kong, J., Kim, J., Bae, J.: HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. Advances in neural information processing systems **33**, 17022–17033 (2020)
12. Lee, M.H., Shomanov, A., Begim, B., Kabidenova, Z., Nyssanbay, A., Yazici, A., Lee, S.W.: EAV: EEG-audio-video dataset for emotion recognition in conversational contexts. Scientific data **11**(1), 1026 (2024)
13. Lee, Y.E., Lee, S.H., Kim, S.H., Lee, S.W.: Towards voice reconstruction from eeg during imagined speech. In: Proceedings of the AAAI conference on artificial intelligence (2023)
14. Li, T., He, K.: Back to basics: Let denoising generative models denoise. arXiv preprint arXiv:2511.13720 (2025)
15. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
16. Liu, F., Wu, X., You, C., Ge, S., Zou, Y., Sun, X.: Aligning source visual and target language domains for unpaired video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence (2021)
17. Ma, J., Yang, B., Qiu, W., Li, Y., Gao, S., Xia, X.: A large eeg dataset for studying cross-session variability in motor imagery brain-computer interface. Scientific Data **9**(1), 531 (2022)
18. Ma, J., Zhu, Y., You, C., Wang, B.: Pre-trained diffusion models for plug-and-play medical image enhancement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention (2023)
19. Nunez, P.L., Srinivasan, R.: Electric fields of the brain: the neurophysics of EEG. Oxford university press (2006)
20. Reddy, C.K., Gopal, V., Cutler, R.: DNSMOS: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In: IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 6493–6497 (2021)
21. Ren, Q., Wang, Y., Guo, L., Zhang, W., Fan, Z., You, C.: Scale where it matters: Training-free localized scaling for diffusion models. arXiv preprint arXiv:2511.19917 (2025)
22. Schneider, F., Kamal, O., Jin, Z., Schölkopf, B.: Mo\^ usai: Text-to-music generation with long-context latent diffusion. arXiv preprint arXiv:2301.11757 (2023)
23. Sun, S., Wang, Y., Zhang, H., Xiong, Y., Ren, Q., Fang, R., Xie, X., You, C.: Ouroboros: Single-step diffusion models for cycle-consistent forward and inverse rendering. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (2025)
24. Tran, X.T., Vo, T.N., Vu, S.T., Tran, T.T., Nguyen, M.D., Do, T., Lin, C.T.: Inter-and intra-subject variability in eeg: A systematic survey. arXiv preprint arXiv:2602.01019 (2026)
25. Wang, Y., Ma, Y., Li, W., You, C.: Let eeg models learn eeg. arXiv preprint arXiv:2605.21280 (2026)
26. Xiang, S., Ling, H., Wu, M.: Cross-modal alignment and rectified flow-based latent representation synthesis for enhanced speech-driven alzheimer's disease detection. Bioengineering **13**(3), 370 (2026)
27. Xu, L., Xu, M., Ke, Y., An, X., Liu, S., Ming, D.: Cross-dataset variability problem in eeg decoding with deep learning. Frontiers in human neuroscience **14**, 103 (2020)
28. Yang, D., Zhang, Y., Yu, X., Hou, L., Tao, X., Wan, P., Qi, X., Liao, R.: Stable velocity: A variance perspective on flow matching. arXiv preprint arXiv:2602.05435 (2026)

29. You, C., Dai, H., Min, Y., Sekhon, J.S., Joshi, S., Duncan, J.S.: Uncovering memorization effect in the presence of spurious correlations. Nature Communications (2025)
30. You, C., Mint, Y., Dai, W., Sekhon, J.S., Staib, L., Duncan, J.S.: Calibrating multi-modal representations: A pursuit of group robustness without annotations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)