

A Survey on Bridging EEG Signals and Generative AI: From Image and Text to Beyond

Shreya Shukla, Jose Torres, Abhijit Mishra, Jacek Gwizdka, Shounak Roychowdhury

School of Information, University of Texas at Austin

{shreya.shukla, jtorres1221, abhijitmishra, jacekg, shounak.roychowdhury}@utexas.edu

Abstract

Integration of Brain-Computer Interfaces (BCIs) and Generative Artificial Intelligence (GenAI) has opened new frontiers in brain signal decoding, enabling assistive communication, neural representation learning, and multimodal integration. BCIs, particularly those leveraging Electroencephalography (EEG), provide a non-invasive means of translating neural activity into meaningful outputs. Recent advances in deep learning, including Generative Adversarial Networks (GANs) and Transformer-based Large Language Models (LLMs), have significantly improved EEG-based generation of images, text, and speech. This paper provides a literature review of the state-of-the-art in EEG-based multimodal generation, focusing on (i) EEG-to-image generation through GANs, Variational Autoencoders (VAEs), and Diffusion Models, and (ii) EEG-to-text generation leveraging Transformer based language models and contrastive learning methods. Additionally, we discuss the emerging domain of *EEG-to-speech synthesis*, an evolving multimodal frontier. We highlight key datasets, use cases, challenges, and EEG feature encoding methods that underpin generative approaches. By providing a structured overview of EEG-based generative AI, this survey aims to equip researchers and practitioners with insights to advance neural decoding, enhance assistive technologies, and expand the frontiers of brain-computer interaction.

1 Introduction & Motivation

The convergence of Brain-Computer Interfaces (BCIs) and Generative Artificial Intelligence (GenAI) is transforming human-computer interaction by enabling direct brain-to-device communication. These advancements have enabled applications in assistive communication for individuals with disabilities, cognitive neuroscience, mental health assessment, augmented reality (AR)/virtual reality (VR), and neural art generation. Electroencephalography (EEG), a widely used non-invasive

neural recording technique, **enables both passive and active Brain-Computer Interfaces (BCIs) and holds potential for applications in real-time adaptive human-computer interaction** (Zander et al., 2010; Wolpaw and Boulay, 2010). Recent advancements in deep learning and generative models have significantly improved the decoding of EEG signals, enabling the translation of neural activity into text, images, and speech. Specifically, Generative AI, including Generative Adversarial Networks (GANs) (Goodfellow et al., 2014) and Transformers (Vaswani et al., 2017), has significantly advanced brain decoding, facilitating visual reconstruction, language generation, and speech synthesis (Bai et al., 2023; Srivastava and Shinde, 2020; Lee et al., 2023b). GANs improve cross-subject classification and EEG data augmentation (Song et al., 2021), while Transformer-based architectures and multimodal deep learning frameworks (Liu et al., 2024a; Wang and Ji, 2022) enhance EEG-to-text translation and semantic decoding (Ali et al., 2024), pushing the boundaries of brain-signal interpretation.

In light of recent breakthroughs in *Generative AI*, this survey provides a scope review of recent advancements in EEG-based generative AI, with a focus on two primary directions. The first explores *how brain signals can be used to generate or reconstruct visual stimuli*, utilizing models such as GANs and Diffusion Models to decode perceptual representations. The second investigates the application of *deep learning for EEG-to-text translation*, where recurrent neural network and Transformers (Vaswani et al., 2017) based language models, and contrastive learning techniques play a crucial role in learning linguistic representation. The survey also examines emerging trends in *speech decoding from EEG signals and multimodal integration considerations* surrounding the use of generative AI for brain signal interpretation. Through this, we hope to provide a structured understanding of

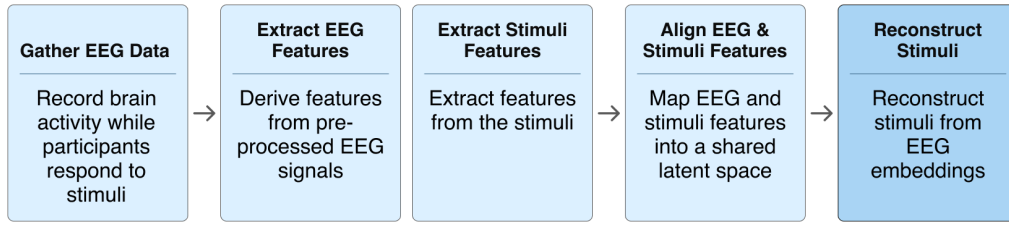


Figure 1: General Steps from EEG Data Gathering to Stimuli Reconstruction (Image, Text, or Sound)

EEG-based generative AI to researchers and practitioners, offering insights to drive innovation in neural decoding, assistive technology, and brain-computer interaction.

Before proceeding, we would like to highlight that Figure 1 provides a high-level overview of the EEG-to-stimuli generation pipeline, illustrating the key stages from neural data acquisition to the generation of text, images, or audio. Additionally, Table 1 summarizes key datasets that incorporate EEG and other modalities, serving as a valuable resource for researchers exploring multimodal neural decoding. Finally, Figure 2 centralizes a detailed overview of the techniques and approaches covered in this survey, contextualizing their purpose and applications in EEG-based generative modeling.

2 Related Work

Several surveys have explored EEG-based brain-computer interfaces (BCIs) and deep learning techniques for neural decoding. (Chen et al., 2022) provides a broad overview of EEG-based BCI applications, discussing signal processing methodologies and traditional machine learning techniques but lacks an in-depth analysis of generative models. (Gong et al., 2021) and (Weng et al., 2024) review deep learning techniques, particularly CNNs and RNNs, for EEG classification, emotion recognition, and cognitive state monitoring, yet they focus more on supervised learning approaches rather than self-supervised and generative modeling. More recently, (Murad and Rahimi, 2024) examined EEG-to-text decoding but primarily covered classification-based approaches without exploring the role of large language models (LLMs) or multimodal integration with vision-based EEG applications. Additionally, (Sun and Mou, 2023) discusses self-supervised learning (SSL) techniques for EEG but does not emphasize their applicability to generative tasks, such as EEG-to-image or EEG-to-text generation. Our survey specifically examines the intersection of EEG and generative AI, focusing on recent ap-

proches for EEG-to-text and EEG-to-image translation. While previous studies focus on EEG feature extraction and classification, our work highlights how self-supervised learning, contrastive learning, and multimodal alignment improve EEG-based generation.

3 EEG-to-Image Generation

This section explores regenerating images from visually evoked brain signals via EEG. It covers use cases, concerns addressed, techniques employed, and EEG feature encoding methods for image generation used by surveyed studies.

3.1 Use Cases and Addressed Concerns

Surveyed studies address key challenges like low signal-to-noise ratio of EEG signals (Bai et al., 2023; Lan et al., 2023; Zeng et al., 2023a), limited information and individual differences in EEG signals (Bai et al., 2023), lower performance on natural object images compared to digits and characters (Mishra et al., 2023) and small dataset sizes (Singh et al., 2023). Additionally, some efforts explore alternatives to supervised learning (Li et al., 2020; Song et al., 2023), since it demands large amount of data. Song et al. (2023) addresses concerns regarding convolution layers applied separately along temporal and spatial dimensions, which disrupts the correlation between channels and hinders the spatial properties of brain activity. Overall, these approaches aim to enhance the training, performance, and interpretation of brain data (Li et al., 2024).

Kavasidis et al. (2017), Song et al. (2023) and Mishra et al. (2023) extract class-specific EEG encodings that contain discriminative information to improve image generation quality, while (Nemrodov et al., 2018) focus on utilizing spatiotemporal EEG information to determine the neural correlates of facial identity representations and (Khaleghi et al., 2022) map EEG signals to visual saliency maps corresponding to each image. Other Com-

Dataset	Stimuli Type	Channels/Electrodes	Stimuli Details	Subjects
Zuco 1.0 (Hollenstein et al., 2018)	Text	128	Sentences from Stanford Sentiment Treebank, Wikipedia corpus	12
Zuco 2.0 (Hollenstein et al., 2019)	Text	128	Expanded subjects with similar content as Zuco 1.0	18
Alice (Bhattachali et al., 2020)	Text	61 + 1 ground	2,129 words, 84 sentences from <i>Alice in Wonderland</i>	52
Envisioned Speech (Kumar et al., 2018)	Imagined Speech	14	20 text stimuli (digits, characters), 10 objects	23
Alljoined (Xu et al., 2024)	Image	64	10,000 images per participant from 80 MS-COCO categories	8
ImageNet EEG (Spampinato et al., 2017)	Image	128	40 ImageNet classes, 50 images/class, 2000 total	6
DM-RE2I (Zeng et al., 2023b)	Image	32	200 ImageNet images across 26 subjects	26
Texture Perception (Orima and Motoyoshi, 2021)	Image	19	166 grayscale natural texture images	15
THINGS-EEG (Grootswagers et al., 2022)	Image	64	22,248 images across 1854 object concepts	50
DCAE (Zeng et al., 2023b)	Image	32	200 ImageNet images (cats, dogs, flowers, pandas)	26
ThoughtViz (Tirupattur et al., 2018)	Imagined Objects	14	EEG recorded while participants imagined digits, characters, and objects	23
OCED (Kaneshiro et al., 2015)	Image	128	12 images per 6 object categories	10
NMED-T (Losorelli et al., 2017)	Music	128	10 songs (4:30-5:00 mins) with tempos 56-150 BPM	20
NMED-H (Kaneshiro et al., 2016)	Music	125	4 versions of 4 songs, total 16 stimuli	48
KARA ONE (Zhao and Rudzicz, 2015)	Text, Audio, Speech	64	Rest state, stimulus, imagined speech, speaking task	12
Japanese Speech EEG (Mizuno et al., 2024)	Audio	64	503 spoken sentences (male/female speaker)	1
Phrase/Word Speech EEG (Park et al., 2024)	Audio	64	Audio of 13 words/phrases, followed by speech replication	10

Table 1: EEG-Based Datasets from Surveyed Studies with Text, Image and Audio/Speech/Music Stimuli

mon strategies include projecting neural signals into a shared subspace with image embeddings (Shimizu and Srinivasan, 2022), generating class-specific EEG encodings as latent representations (Mishra et al., 2023), and decoding multi-level perceptual information from EEG signals to produce multi-grained outputs (Lan et al., 2023).

Additionally, research efforts focus on enhancing the generalizability of feature extraction pipelines across datasets (Singh et al., 2024), evaluating the performance of different channels (Sugimoto et al., 2024), and incorporating attention modules to highlight the significance of each channel or frequency band (Li et al., 2024).

3.2 Techniques Used Across Studies

Various computer vision generative models are employed to reconstruct images from EEG signals. These include **Variational Autoencoders** (Kavasidis et al., 2017; Wakita et al., 2021), **Generative Adversarial Networks (GANs)** (Kavasidis et al., 2017; Khaleghi et al., 2022; Mishra et al., 2023; Singh et al., 2024; Li et al., 2024), and conditional GANs (Singh et al., 2023; Ahmadiéh et al., 2024). **Diffusion models**, including prior diffusion models that refine EEG embeddings into image priors (Shimizu and Srinivasan, 2022), as well as pre-trained diffusion models such as Stable Diffusion (Bai et al., 2023), are also commonly used. Additionally, diffusion modules based on U-net architecture have been used in (Zeng et al., 2023a; Lan et al., 2023) to further enhance EEG-to-image reconstruction.

Contrastive learning is another popular approach to align multimodal embeddings, employed in studies (Singh et al., 2023; Lan et al., 2023; Song et al., 2023; Sugimoto et al., 2024) to obtain dis-

criminative features from EEG signals and align the two modalities by constraining their cosine similarity (Song et al., 2023). Furthermore, **attention mechanisms** are integrated into various models (Mishra et al., 2023; Song et al., 2023; Li et al., 2024) to enhance image quality, capture spatial correlations that reflect brain activity inferred from EEG data, and determine the relative importance of individual EEG channels.

3.3 EEG Feature Encoding Techniques

In EEG-to-image reconstruction, the process typically begins with an encoder identifying the latent feature space of EEG signals, followed by a decoder that converts these features into an image. Long Short-Term Memory (LSTM)-based architectures are widely used due to their effectiveness in capturing **temporal dependencies** in EEG signals. Kavasidis et al. (2017) employs an LSTM network to generate a compact and class-discriminative feature vector, which is also used for object recognition. Similarly, (Singh et al., 2023) integrates LSTM with a triplet-loss-based contrastive learning approach to enhance **feature discrimination**. Singh et al. (2024) extends this approach by incorporating both CNN and LSTM architectures trained under EEG label supervision with triplet loss, further improving discriminative feature learning. Additionally, (Ahmadiéh et al., 2024) uses LSTM to extract EEG features across two dimensions (EEG channels and signal duration) and **enhances feature generation** through various regression methods, including polynomial regression, neural network regression, and type-1 and type-2 fuzzy regression.

Several studies also leverage convolutional architectures to capture **spatial dependencies** in EEG.

Techniques (EEG-Text)	Studies and main use cases
CNNs	• Generate medical reports (Biswal et al., 2019) • Translate user's active intent to text represented by morse code (Srivastava and Shinde, 2020) • Folded ensemble technique to minimize the computational complexity and solve all the class imbalance issues to enhance the accuracy of text generation (Rathod et al., 2024)
LSTMs	• Translate user's active intent to text represented by morse code (Srivastava and Shinde, 2020) • Folded ensemble technique to minimize the computational complexity and solve all the class imbalance issues to enhance the accuracy of text generation (Rathod et al., 2024)
LLMs	• EEG-to-text seq-to-seq decoding and zero-shot sentence sentiment classification on natural reading tasks (Wang and Ji, 2022) • Improving accuracy of open-vocabulary EEG-to-text decoding (Liu et al., 2024a) • Bridge the semantic gap between EEG and Text (Wang et al., 2024) • Open vocabulary EEG decoding incorporating a subject-dependent representation learning module (Amrani et al., 2024) • Ensuring cross-modal semantic consistency between EEG and Text (Tao et al., 2024) • Capture global and local contextual information and long-term dependencies • (Chen et al., 2025) • Use visual stimuli rather than text to circumvent the complexities of language processing (Mishra et al., 2024)
Transformer	• EEG-to-text seq-to-seq decoding and zero-shot sentence sentiment classification on natural reading tasks (Wang and Ji, 2022) • Open vocabulary EEG decoding incorporating a subject-dependent representation learning module (Amrani et al., 2024) • Ensuring cross-modal semantic consistency between EEG and Text (Tao et al., 2024) • Improving accuracy of open-vocabulary EEG-to-text decoding (Liu et al., 2024a) • Bridge the semantic gap between EEG and Text using LLMs (Wang et al., 2024) • Recalibrates subject-dependent EEG representation to the semantic-dependent EEG representation (Feng et al., 2023) • Open vocabulary EEG-to-Text translation tasks with or without word-level markers (Duan et al., 2023) • Decoding EEG Speech Perception with Transformers and VAE-based Data Augmentation (Yu-Hao Chen et al., 2025)
Gated Recurrent Units	• Open vocabulary EEG decoding incorporating a subject-dependent representation learning module (Amrani et al., 2024) • Capture global and local contextual information and long-term dependencies. (Chen et al., 2025)
Recurrent Neural Networks	• Translate active intention into text format based on Morse code Yang et al. (2023)
Attention Mechanism	• Generate medical reports (Biswal et al., 2019) • Capture global and local contextual information and long-term dependencies (Chen et al., 2025)
Contrastive Learning	• Recalibrates subject-dependent EEG representation to the semantic-dependent EEG representation Feng et al. (2023) • Ensuring cross-modal semantic consistency between EEG and Text (Tao et al., 2024) • Bridge the semantic gap between EEG and Text using LLMs (Wang et al., 2024)
Masked Signal Modeling	• Improving accuracy of open-vocabulary EEG-to-text decoding (Liu et al., 2024a) • Ensuring cross-modal semantic consistency between EEG and Text (Tao et al., 2024) • Bridge the semantic gap between EEG and Text using LLMs (Wang et al., 2024)

Techniques (EEG-Image)	Studies and main use cases
CNNs	• Semi-supervised cross-modal image generation (Li et al., 2020) • Visual Saliency and Image Reconstruction from EEG Signals (Khaleghi et al., 2022) • Map EEG signals to the visual saliency maps corresponding to each image (Song et al., 2023) • Demonstrate the generalizability of feature extraction pipeline across three different datasets (Singh et al., 2024) • Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion (Li et al., 2024) • Reconstructing images using the same semantics as the corresponding EEG (Zeng et al., 2023a) • Zero-shot framework to project neural signals from different sources into the shared subspace (Shimizu and Srinivasan, 2022)
LSTMs	• Extracting visual class discriminative information from EEG data (Kavassidis et al., 2017) • Framework for synthesizing the images using small-size EEG datasets (Singh et al., 2023) • Image reconstruction using generative adversarial and deep fuzzy neural network (Ahmadi et al., 2024) • Demonstrate the generalizability of feature extraction pipeline across three different datasets (Singh et al., 2024)
Generative Adversarial Networks (GANs)	• Extracting visual class discriminative information from EEG data (Kavassidis et al., 2017) • Map EEG signals to the visual saliency maps corresponding to each image (Khaleghi et al., 2022) (Mishra et al., 2023) • Generates images along with producing class-specific EEG encoding as a latent representation (Singh et al., 2024) • Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion (Li et al., 2024) • Framework for synthesizing the images using small-size EEG datasets (Singh et al., 2023) • Image reconstruction using generative adversarial and deep fuzzy neural network (Ahmadi et al., 2024)
Variational Autoencoder	• Generating high-quality images directly from brain EEG signals, without the need to translate thoughts into text (Bai et al., 2023) • Photorealistic Reconstruction of Visual Texture From EEG Signals (Wakita et al., 2021) • Extracting visual class discriminative information from EEG data (Kavassidis et al., 2017)
Diffusion Models	• Zero-shot framework to project neural signals from different sources into the shared subspace (Shimizu and Srinivasan, 2022) • Generating high-quality images directly from brain EEG signals, without the need to translate thoughts into text (Bai et al., 2023) • Reconstructing images using the same semantics as the corresponding EEG (Zeng et al., 2023a) • Multi-level perceptual information decoding to draw multi grained outputs from given EEG (Lan et al., 2023)
Attention Mechanism	• Generates images along with producing class-specific EEG encoding as a latent representation (Mishra et al., 2023) • Self-supervised framework to decode natural images for object recognition (Song et al., 2023) • Visual Decoding and Reconstruction via EEG Embeddings with Guided Diffusion (Li et al., 2024)
Contrastive Learning	• Framework for synthesizing the images using small-size EEG datasets (Singh et al., 2023) • Multi-level perceptual information decoding to draw multi grained outputs from given EEG (Lan et al., 2023) • Self-supervised framework to decode natural images for object recognition (Song et al., 2023) • Generating perceptual and cognitive contents using EEG data (Sugimoto et al., 2024) • Self-supervised framework to decode natural images for object recognition (Song et al., 2023)
Masked Signal Modeling	• Generating high-quality images directly from brain EEG signals, without the need to translate thoughts into text (Bai et al., 2023)

Techniques (EEG-Audio/Speech/Music)	Studies and main use cases
CNNs	• Show how acoustic features are related to EEG signals recorded during speech perception and production (Krishna et al., 2021) • Reconstructing music stimuli to be perceived and identified independently (Ramirez-Aristizabal and Kello, 2022)
Transformer	• Investigate the potential to reconstruct speech from EEG signals, including the corresponding speaker's characteristics (Mizuno et al., 2024) • Utilizing deep features from EEG data for emotional music composition (Jiang et al., 2024)
Gated Recurrent Units	• Show how acoustic features are related to EEG signals recorded during speech perception and production (Krishna et al., 2021) • Convert EEG of imagined speech into user's own voice (Lee et al., 2023b) • EEG-based Talking-face Generation (Park et al., 2024)
Generative Adversarial Networks (GANs)	• Convert EEG of imagined speech into user's own voice (Lee et al., 2023b) • EEG-based Talking-face Generation (Park et al., 2024)
Automatic Speech Recognition	• Convert EEG of imagined speech into user's own voice (Lee et al., 2023b) • EEG-based Talking-face Generation (Park et al., 2024)
Latent Diffusion Models	• Reconstructing naturalistic music from EEG without need for manual pre-processing and channel selection (Postolache et al., 2024)
Attention Module	• Show how acoustic features are related to EEG signals recorded during speech perception and production (Krishna et al., 2021)

Figure 2: Techniques and References of Surveyed Studies for EEG to Text, Image and Beyond

Li et al. (2020) uses a three-layer feedforward neural network to project EEG signals into semantic features. Wakita et al. (2021) adopts a 1D convolutional encoder-decoder as part of a multimodal variational autoencoder (VAE) to obtain mean and variance vectors for EEG signal representation. Mishra et al. (2023) uses a convolutional encoder-decoder framework enhanced with an attention module to focus more on channels with important features instead of using all the features with equal weights. Similarly, Sugimoto et al. (2024) implements EEGNet (Lawhern et al., 2018), a compact convolutional network, as an EEG encoder, while Li et al. (2024) uses Sinc-EEGNet (Bria et al., 2021), an architecture incorporating a sinc-based convolution layer, depth-wise convolution, and separable convolution to extract EEG features. It also integrates an attention mechanism to identify the **most relevant frequency bands and channels** for signal-based classification.

Graph-based techniques have been explored for EEG feature extraction. Khaleghi et al. (2022) constructs functional graph connectivity-based embed-

dings from EEG signals, which are then processed using a Geometric Deep Network (GDN) to derive feature vectors. Song et al. (2023) integrates temporal-spatial convolution with plug-and-play spatial modules, leveraging self-attention and graph attention mechanisms to extract EEG features more effectively.

To integrate temporal and spatial feature extraction mechanisms to improve EEG-based image reconstruction, Zeng et al. (2023a) develops a framework inspired by EEGChannelNet (Palazzo et al., 2020) and ResNet-18, combining spatial, temporal, and temporal-spatial blocks with a multi-kernel residual block. Shimizu and Srinivasan (2022) uses a time-series-inspired architecture with a channel-wise transformed encoder and temporal-spatial convolution to extract **rich latent EEG representations**.

Self-supervised learning and contrastive learning have been applied to enhance EEG feature extraction. Bai et al. (2023) uses masked signal modeling, where EEG tokens are partially masked, and a 1D convolutional layer transforms all tokens

into embeddings. A Masked Autoencoder (MAE) predicts the missing tokens, refining the learned representations. Lan et al. (2023) employs contrastive learning to extract pixel-level semantics from EEG signals while generating a **saliency map of silhouette information** using GANs. It also aligns CLIP embeddings for image captions with an EEG sample-level encoder through a specialized loss function.

3.4 Evaluation Metrics

EEG-to-image generation often begins with object classification to ensure extracted EEG features contain useful class-discriminative information. Metrics like *top-k accuracy* are commonly used (Shimizu and Srinivasan, 2022; Lan et al., 2023; Song et al., 2023), along with qualitative visual analysis and quantitative evaluations. Key quantitative metrics include *Inception Score (IS)* (Salimans et al., 2016), used by (Kavasidis et al., 2017; Li et al., 2020; Bai et al., 2023; Singh et al., 2023) which measures the quality of images, *Frechet Inception Distance (FID)* for measuring realism (Bai et al., 2023; Singh et al., 2024; Ahmadih et al., 2024), and saliency metrics such as *Structural Similarity Index (SSIM)* for assessing perceptual fidelity (Khaleghi et al., 2022; Shimizu and Srinivasan, 2022; Bai et al., 2023; Ahmadih et al., 2024; Sugimoto et al., 2024). Other useful metrics are PixCorr (Pixel-wise Correlation) (Shimizu and Srinivasan, 2022), *Kernel Inception Distance (KID)* (Singh et al., 2024), *LPIPS (Learned Perceptual Image Patch Similarity)* (Bai et al., 2023) and *Diversity Score* (Mishra et al., 2023).

4 EEG-to-Text Generation

This section discusses how AI learns brain signal representations from EEG data and maps them to linguistic representations, with an overview depicted in Figure 3. We survey use cases, techniques, concerns, and EEG feature encoding methods for text generation.

4.1 Use Cases and Addressed Concerns

The studies referenced in this section share a common use case: generating text from EEG signals. Several studies (Biswal et al., 2019; Srivastava and Shinde, 2020; Yang et al., 2023; Rathod et al., 2024) use the closed vocabulary approach, relying on a fixed set of pre-defined words for EEG-based decoding. Among these, Srivastava and Shinde

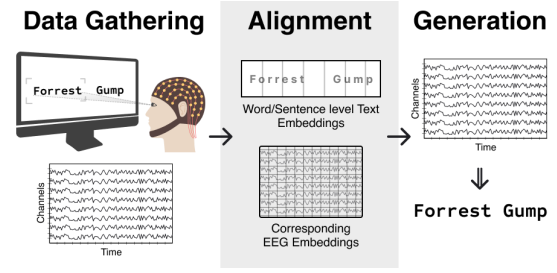


Figure 3: Reconstructing text from EEG, with eye-tracking data used to capture word-level EEG signals

(2020); Yang et al. (2023) investigate text generation using morse code representation of EEG signals, where users’ active intent is captured, mapped to morse codes, and then translated to text format.

Recent studies (Wang and Ji, 2022; Feng et al., 2023; Duan et al., 2023; Liu et al., 2024a; Wang et al., 2024; Amrani et al., 2024; Tao et al., 2024; Mishra et al., 2024; Ikegawa et al., 2024; Chen et al., 2025) overcome closed-vocabulary limitations by exploring open-vocabulary text generation to emulate naturalistic conversations. These studies also address the impact of subjectivity in subject-dependent EEG representation (Feng et al., 2023; Amrani et al., 2024), learn cross-modal representation (Wang et al., 2024; Tao et al., 2024), and capture long-term dependencies in text and also global contextual information from EEG data that transformers might miss (Rathod et al., 2024; Chen et al., 2025).

A significant challenge is the reliance on eye-tracking fixation data as a marker for word-level EEG, which studies like (Duan et al., 2023; Liu et al., 2024a) aim to address. To overcome challenges like defining word-level boundaries in EEG signals and other language processing tasks, some studies have proposed language-agnostic solutions (Mishra et al., 2024; Ikegawa et al., 2024) which capture signals through image modality and leverage advancements in image-text intermodality to generate text from the collected data. Additionally, Yu-Hao Chen et al. (2025) introduce a VAE-based augmentation technique to address the issue of limited EEG-text datasets.

4.2 Techniques Used Across Studies

A noteworthy aspect of these studies is the utilization of **Large Language Models (LLMs)**, particularly BART. Several works (Wang and Ji, 2022; Liu et al., 2024a; Wang et al., 2024; Amrani et al., 2024; Tao et al., 2024; Chen et al., 2025) have used

BART for text generation. In a study by [Mishra et al. \(2024\)](#), LLMs were fine-tuned on EEG embeddings, image and text data in the training stage to generate text from just EEG signals during inference.

Contrastive learning is widely used in studies like ([Feng et al., 2023](#); [Tao et al., 2024](#); [Wang et al., 2024](#)) to identify positive EEG-text pairs (e.g., EEG data from the same sentence across subjects) and negative pairs (e.g., EEG data from different sentences or subjects), improving the model's ability to align EEG signals with corresponding text representations. Another key technique is masked signal modeling, employed by [Liu et al. \(2024a\)](#), where a transformer model is pre-trained to reconstruct randomly masked EEG signals from raw data, enabling the model to learn context, relationships, and semantics within sentence-level EEG signals. An integrated approach by [Tao et al. \(2024\)](#) combines contrastive learning with **masked signal modeling**, where word-level EEG feature sequences are randomly masked and sentence-level sequences deliberately masked, guided by an intra-modality self-reconstruction objective.

In addition to these techniques, bi-directional Gated Recurrent Units (GRUs) are used to dynamically handle the varying lengths of word-level raw EEG signals ([Amrani et al., 2024](#)). Hierarchical GRUs further improve EEG data processing by capturing both long-range dependencies and local contextual information through the organization of hidden layers hierarchically ([Chen et al., 2025](#)). A unique approach by ([Rathod et al., 2024](#)) employs a folded ensemble deep CNN for text suggestion and a folded ensemble Bidirectional LSTM for text generation, effectively addressing class imbalance and significantly enhancing the accuracy of text generation in closed-vocabulary tasks.

4.3 EEG Feature Encoding Techniques

For text generation tasks, EEG signals are encoded into features to capture **temporal patterns and semantic information**. In the study by [Biswal et al. \(2019\)](#), which focuses on generating medical reports, EEG signals are encoded using stacked CNNs to capture **shift-invariant features** and RNNs to capture **temporal patterns**. These features are then used to generate key phenotypes, which hierarchical LSTMs utilize to produce detailed explanations. [Srivastava and Shinde \(2020\)](#) employs an ensemble model to extract EEG embeddings, us-

ing CNNs to capture spatial variations and LSTMs to model temporal sequences and long-range dependencies. Another study by [Yu-Hao Chen et al. \(2025\)](#), proposes two objectives: classification and sequence-to-sequence (seq2seq) text generation, employing residual blocks for feature extraction in both tasks to capture both **spatial and temporal features** of the EEG signals effectively.

Other studies explore the extraction of **spectral and statistical features** alongside temporal or spatial patterns. [Yang et al. \(2023\)](#), aiming to translate active intention into text using Morse code, employed Short-Term Fourier Transform (STFT) to extract spectral features and concatenated these with statistical features (e.g., min, max etc. for each channel), in addition to using 1D CNN for spatial features and RNN for temporal features. [Rathod et al. \(2024\)](#), another closed-vocabulary solution, used features such as Wavelet Transform (WT), Common Spatial Patterns (CSP), and **statistical features** to generate EEG feature vectors for classification.

Various studies have used state-of-the-art transformer architecture for encoding EEG features. ([Wang and Ji, 2022](#)) uses a multi-layer transformer encoder to obtain **EEG mapping from word-level EEG sequences**. [Feng et al. \(2023\)](#) uses a transformer-based pre-encoder to convert word-level EEG features into the Seq2Seq embedding space. Another study by [Tao et al. \(2024\)](#) also uses an encoder to extract EEG embeddings and store them in a cross-modal codebook alongside word embeddings obtained from a transformer-based BART model.

Obtaining word-level EEG signals typically requires markers, often from eye-fixation data like in the Zuco dataset, limiting generalizability. Some studies address this by using **marker-free and sentence-level EEG signals**. [Duan et al. \(2023\)](#) extracts both word-level EEG and raw EEG embeddings. For word-level EEG features with markers, a multi-head transformer layer projects embeddings into feature sequences. For raw EEG waves, a multi-layer transformer encoder is trained for self-reconstruction of waveforms and the transformation of raw EEG signals into sequences of embeddings. In a study by [Liu et al. \(2024a\)](#), a convolutional transformer model is pretrained with sentence-level EEG signals using a masking technique. It uses a multi-view transformer to encode different brain regions with separate convolutional

transformers. Wang et al. (2024) uses both word-level and sentence-level EEG features. It employs a masking technique where word-level sequences are randomly masked and sentence-level features are compulsorily masked.

Chen et al. (2025) uses a stacked Hierarchical GRU-based decoder along with Masked Residual Attention Mechanism to obtain EEG representations that capture both **local and global contextual information**. Amrani et al. (2024) employs a module consisting of bi-directional GRUs to dynamically address varying lengths of word-level raw EEG signals, a subject-specific 1D convolutional layer, and a multi-layer transformer encoder to encode word-level EEG signals.

4.4 Evaluation Metrics

In the surveyed studies, generated text is evaluated against reference text using various established metrics. The commonly used text evaluation metrics are as follows: *METEOR* (Banerjee and Lavie, 2005), employed by (Biswal et al., 2019; Chen et al., 2025); *BLEU* score (Papineni et al., 2002), utilized by (Biswal et al., 2019; Wang and Ji, 2022; Feng et al., 2023); *ROUGE* score (Lin, 2004), adopted by (Wang and Ji, 2022; Feng et al., 2023; Duan et al., 2023; Liu et al., 2024a; Wang et al., 2024); and *BERTScore* (Zhang et al., 2019), used by (Amrani et al., 2024; Mishra et al., 2024). Other metrics include *Word Error Rate (WER)* used by (Feng et al., 2023), *Translation Error Rate (TER)*, and *BLEURT* (Sellam et al., 2020), used by (Chen et al., 2025).

5 EEG-to-Sound/Speech Generation

We review studies focused on EEG-based generation of sound, speech, voice or music and cover use cases, concerns, techniques, and EEG feature encoding methods for generating Sound or Speech from EEG.

5.1 Use Cases and Addressed Concerns

EEG-based generation has been explored in various fields beyond image reconstruction, particularly in audio and speech-related applications. These include speech synthesis (Krishna et al., 2021; Lee et al., 2023a), music decoding and reconstruction (Ramirez-Aristizabal and Kello, 2022; Postolache et al., 2024), emotive music generation (Jiang et al., 2024), voice reconstruction (Lee et al., 2023b), talking-face generation (Park et al., 2024), and

speech recovery (Mizuno et al., 2024). While some studies focus on decoding audio signals for listening tasks in speech or music perception (Krishna et al., 2021; Ramirez-Aristizabal and Kello, 2022; Park et al., 2024; Mizuno et al., 2024; Postolache et al., 2024; Jiang et al., 2024), others also investigate speaking tasks and imagined speech (Krishna et al., 2021; Lee et al., 2023b,a).

For more naturalistic communication, Lee et al. (2023b) converts EEG signals recorded during imagined speech into the user's own voice, aiming for personalized speech synthesis. Similarly, Park et al. (2024) synthesizes speech from EEG along with generating a talking face with lip-sync. Furthermore, these studies tackle issues such as generating fragmented or abstract outputs (Park et al., 2024), challenges of synthesizing complete speech from EEG (Mizuno et al., 2024), being restricted to simpler music with limited timbres (Postolache et al., 2024), and the absence of a standardized vocabulary for aligning EEG and audio data (Jiang et al., 2024).

5.2 Techniques Used Across Studies

Convolutional Neural Network (CNN)-based deep learning models have been used in studies (Krishna et al., 2021; Ramirez-Aristizabal and Kello, 2022) to generate audio waveforms from EEG input. Krishna et al. (2021) explores speech synthesis for both speaking and listening tasks, using a deep learning architecture with temporal convolution layers, 1D layer, and a time-distributed layer to generate audio waveforms directly. Similarly, Ramirez-Aristizabal and Kello (2022) reconstructs music stimuli using sequential CNN regressors.

Lee et al. (2023b) propose NeuroTalk framework for voice reconstruction from imagined speech. The framework uses a generator based on **GRUs** to capture sequential EEG information, which outputs a mel-spectrogram. Mel-spectrogram is then converted into a waveform using a **HiFi-GAN vocoder** (Kong et al., 2020), and the resulting waveform is transcribed into text using an **Automatic Speech Recognition (ASR)** system based on HuBERT (Hsu et al., 2021), a self-supervised speech representation learning method. Park et al. (2024) uses NeuroTalk framework to synthesize audible speech and integrates it with a personalized talking face using **Wave2Lip** (Prajwal et al., 2020) and Apple API-based avatar generator that accurately lip-sync to the synthesized speech.

Transformers and Latent Diffusion Models have been used to reconstruct speech (Mizuno et al., 2024) and music (Postolache et al., 2024; Jiang et al., 2024). Jiang et al. (2024) employs a Transformer model for emotive music generation, while Postolache et al. (2024) decodes naturalistic music from EEG using a ControlNet adapter (Zhang et al., 2023) to guide AudioLDM2 (Liu et al., 2024b), a pre-trained diffusion model, improving control over the generated music.

5.3 EEG Feature Encoding Techniques

For speech, voice, and music decoding or generation from EEG, EEG signals are either transformed into intermediate representations, such as mel-spectrograms, or decoded into acoustic and articulatory features (Krishna et al., 2021), or EEG temporal features (Jiang et al., 2024) are utilized. Mel-spectrograms are especially useful, as they offer a shared representational state for both neural signals and audio, enabling more efficient translation between the two modalities.

Krishna et al. (2021) incorporates an attention model to predict **articulatory features** and another attention-regression model to convert these predicted features into **acoustic features**. Similarly, Jiang et al. (2024) extracts EEG tokens through a multi-step process which includes DBSCAN clustering algorithm to derive **EEG temporal features**. These features are eventually transformed **EEG positional encoding EEG features** using positional encoding, which are used to form EEG tokens.

In studies using **mel-spectrograms** as intermediate representations, Ramirez-Aristizabal and Kello (2022) employs a sequential CNN-based regressor to directly map EEG input to time-aligned music spectra. Lee et al. (2023a) seeks to adapt spoken EEG to the subspace of imagined EEG using Common Spatial Pattern (CSP) filters trained on imagined EEG, aiming to generate a user’s voice from imagined speech. These CSP filters extract temporal oscillation patterns, minimizing distribution differences between spoken and imagined EEG. Similarly, Postolache et al. (2024) applies this technique while temporally aligning users’ voices with brain signals, using triggers to mark onset intervals and clearly distinguish actual utterance intervals in continuous brain signals.

5.4 Evaluation Metrics

EEG-to-speech generation is evaluated using quantitative and qualitative metrics, based on its time-

series structure, which also enables its representation as mel-spectrograms. *Mel Cepstral Distortion (MCD)* and *Root Mean Square Error (RMSE)* measure similarity between reconstructed and original speech signals (Krishna et al., 2021; Park et al., 2024), while *Structural Similarity Index (SSI)* and *Peak Signal-to-Noise Ratio (PSNR)* assess spectrogram quality (Ramirez-Aristizabal and Kello, 2022). Linguistic accuracy is evaluated using Word Error Rate (WER), Character Error Rate (CER), and BERTScore (Mizuno et al., 2024), and perceptual quality is quantified with *Frechet Audio Distance (FAD)* (Postolache et al., 2024). Additional metrics include *Hits@k* for search relevance (Jiang et al., 2024) and *Mean Opinion Score (MOS)* for subjective quality assessment (Lee et al., 2023b).

6 Conclusion and Future Work

With advancements in Generative AI, EEG—once primarily used for classification tasks—is now being harnessed for generation, which marks a significant step toward brain-computer interaction (BCI) applications. Given its portability and non-invasive nature, EEG has strong potential for real-time, widespread applications, particularly in assistive communication by enabling direct thought-to-speech or thought-to-text systems that enhance accessibility and human-computer interaction. However, comparing studies in this field remains challenging due to the lack of standardized benchmarks. Even when studies utilized the same datasets, the subject-dependent nature of EEG data allowed for multiple ways of splitting and processing, either by subject or object category. For a fair and meaningful comparison across the surveyed studies, it is crucial to establish standardized benchmarks that define consistent data partitioning, evaluation metrics, and model validation protocols. This would ensure reproducibility, facilitate progress in the field, and enable a more accurate assessment of various approaches in EEG-based generative AI research. Nevertheless, we remain optimistic about further advancements in EEG processing and its potential for generating different modalities. As research progresses, improved methodologies, larger datasets, and standardized benchmarks will enhance the reliability and effectiveness of EEG-based generative solutions and bring us closer to real-time, practical implementations of EEG-driven generative AI.

Limitations

While this survey provides a comprehensive overview of EEG-based generative AI applications, certain limitations exist due to the focused scope of this work. Firstly, this survey primarily covers EEG-based Brain-Computer Interfaces (BCIs), deliberately excluding other neuroimaging techniques such as fMRI, Magnetoencephalography (MEG), and Near-Infrared Spectroscopy (NIRS). Although these modalities play a significant role in BCI research and offer complementary advantages in terms of spatial resolution and multimodal integration, their detailed discussion is beyond the scope of this work.

Secondly, due to space constraints, in-depth discussions on the cognitive underpinnings of EEG signals – such as their biological origins, neural interpretations, and relationships with brain activity—have been omitted. Similarly, technical details regarding EEG hardware, electrode configurations, and device specifications have been largely excluded for brevity. While these aspects are crucial for practical EEG-based applications, our focus remains on the computational and generative modeling aspects of EEG data processing.

Finally, this survey assumes a general background in EEG signal processing, and generative modeling and expects familiarity with these foundational concepts. While we provide essential explanations, a more in-depth introduction to the fundamentals of EEG and BCI technology is outside the scope of this review.

Ethics Statement

EEG data is inherently sensitive, as it contains neural activity patterns that can potentially reveal cognitive states and sometimes personal information. While the majority of the works covered in this survey adhere to established ethical guidelines and standards, some studies may require additional ethical justifications. We have not conducted an exhaustive review of the ethical compliance of each cited work but emphasize the importance of ethical transparency in EEG research. We do not endorse studies that raise ethical concerns or lack proper ethical oversight. Any research involving EEG data collection and analysis should rigorously follow ethical protocols, including obtaining informed consent, ensuring data anonymity, and minimizing risks to participants.

Additionally, we acknowledge the use of OpenAI's ChatGPT-4 system solely for enhancing writing efficiency, generating LaTeX code, and aiding in error debugging. No content related to the survey's research findings, citations, or factual discussions was autogenerated or retrieved using Generative AI-based search mechanisms. Our work remains grounded in peer-reviewed literature and ethical academic standards.

References

- Hajar Ahmadieh, Farnaz Gassemi, and Mohammad Hasan Moradi. 2024. Visual image reconstruction based on eeg signals using a generative adversarial and deep fuzzy neural network. *Biomedical Signal Processing and Control*, 87:105497.
- Omair Ali, Muhammad Saif-ur Rehman, Marita Metzler, Tobias Glasmachers, Ioannis Iossifidis, and Christian Klaes. 2024. Get: A generative eeg transformer for continuous context-based neural signals. *arXiv preprint arXiv:2406.03115*.
- Hamza Amrani, Daniela Micucci, and Paolo Napoletano. 2024. Deep representation learning for open vocabulary electroencephalography-to-text decoding. *IEEE Journal of Biomedical and Health Informatics*.
- Yunpeng Bai, Xintao Wang, Yan-pei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. 2023. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934*.
- Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.
- Shohini Bhattachali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbbers, and John Hale. 2020. The alice datasets: fmri & eeg observations of natural language comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 120–125.
- Siddharth Biswal, Cao Xiao, M Brandon Westover, and Jimeng Sun. 2019. Eegtotext: learning to write medical reports from eeg recordings. In *Machine Learning for Healthcare Conference*, pages 513–531. PMLR.
- Alessandro Bria, Claudio Marrocco, and Francesco Torella. 2021. Sinc-based convolutional neural networks for eeg-bci-based motor imagery classification. In *International Conference on Pattern Recognition*, pages 526–535. Springer.
- Qiupu Chen, Yimou Wang, Fenmei Wang, Duolin Sun, and Qiankun Li. 2025. Decoding text from electroencephalography signals: A novel hierarchical gated

- recurrent unit with masked residual attention mechanism. *Engineering Applications of Artificial Intelligence*, 139:109615.
- Xun Chen, Chang Li, Aiping Liu, Martin J McKeown, Ruobing Qian, and Z Jane Wang. 2022. Toward open-world electroencephalogram decoding via deep learning: A comprehensive survey. *IEEE Signal Processing Magazine*, 39(2):117–134.
- Yiqun Duan, Jinzhao Zhou, Zhen Wang, Yu-Kai Wang, and Chin-Teng Lin. 2023. Dewave: Discrete eeg waves encoding for brain dynamics to text translation. *arXiv preprint arXiv:2309.14030*.
- Xiachong Feng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. Aligning semantic in brain and language: A curriculum contrastive method for electroencephalography-to-text generation. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*.
- Shu Gong, Kaibo Xing, Andrzej Cichocki, and Junhua Li. 2021. Deep learning in eeg: Advance of the last ten-year critical period. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):348–365.
- Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial networks. In *Advances in Neural Information Processing Systems*.
- Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. 2022. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3.
- Nora Hollenstein, Jonathan Rotsztein, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. 2018. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13.
- Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. 2019. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*.
- Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460.
- Yuya Ikegawa, Ryohei Fukuma, Hidenori Sugano, Satoru Oshino, Naoki Tani, Kentaro Tamura, Yasushi Iimura, Hiroharu Suzuki, Shota Yamamoto, Yuya Fujita, et al. 2024. Text and image generation from intracranial electroencephalography using an embedding space for text and images. *Journal of Neural Engineering*, 21(3):036019.
- Hui Jiang, Yu Chen, Di Wu, and Jinlin Yan. 2024. Eeg-driven automatic generation of emotive music based on transformer. *Frontiers in Neurobotics*, 18:1437737.
- Blair Kaneshiro, Duc T Nguyen, Jacek P Dmochowski, Anthony M Norcia, and Jonathan Berger. 2016. Naturalistic music eeg dataset—hindi (nmed-h). In *Stanford Digital Repository*. Stanford Digit. Repository.
- Blair Kaneshiro, Marcos Perreau Guimaraes, Hyung-Suk Kim, Anthony M Norcia, and Patrick Suppes. 2015. A representational similarity analysis of the dynamics of object processing using single-trial eeg classification. *Plos one*, 10(8):e0135697.
- Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. 2017. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817.
- Nastaran Khaleghi, Tohid Yousefi Rezaii, Soosan Beheshti, Saeed Meshgini, Sobhan Sheykhivand, and Sebelan Danishvar. 2022. Visual saliency and image reconstruction from eeg signals via an effective geometric deep network-based generative adversarial network. *Electronics*, 11(21):3637.
- Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.
- Gautam Krishna, Co Tran, Mason Carnahan, and Ahmed H Tewfik. 2021. Advancing speech synthesis using eeg. In *2021 10th International IEEE/EMBS Conference on Neural Engineering (NER)*, pages 199–204. IEEE.
- Pradeep Kumar, Rajkumar Saini, Partha Pratim Roy, Pawan Kumar Sahu, and Debi Prosad Dogra. 2018. Envisioned speech recognition using eeg sensors. *Personal and Ubiquitous Computing*, 22:185–199.
- Yu-Ting Lan, Kan Ren, Yansen Wang, Wei-Long Zheng, Dongsheng Li, Bao-Liang Lu, and Lili Qiu. 2023. Seeing through the brain: Image reconstruction of visual perception from human brain signals. *arXiv e-prints*, pages arXiv–2308.
- Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. 2018. Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of neural engineering*, 15(5):056013.
- Young-Eun Lee, Sang-Ho Kim, Seo-Hyun Lee, Jung-Sun Lee, Soowon Kim, and Seong-Whan Lee. 2023a. Speech synthesis from brain signals based on generative model. In *2023 11th International Winter Conference on Brain-Computer Interface (BCI)*, pages 1–4. IEEE.

- Young-Eun Lee, Seo-Hyun Lee, Sang-Ho Kim, and Seong-Whan Lee. 2023b. Towards voice reconstruction from eeg during imagined speech. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6030–6038.
- Dan Li, Changde Du, and Huiguang He. 2020. Semi-supervised cross-modal image generation with generative adversarial networks. *Pattern Recognition*, 100:107085.
- Dongyang Li, Chen Wei, Shiyang Li, Jiachen Zou, Haoyang Qin, and Quanying Liu. 2024. Visual decoding and reconstruction via eeg embeddings with guided diffusion. *arXiv preprint arXiv:2403.07721*.
- Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.
- Hanwen Liu, Daniel Hajialigol, Benny Antony, Aiguo Han, and Xuan Wang. 2024a. Eeg2text: Open vocabulary eeg-to-text decoding with eeg pre-training and multi-view transformer. *arXiv preprint arXiv:2405.02165*.
- Haohe Liu, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Qiao Tian, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. 2024b. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.
- Steven Losorelli, Duc T Nguyen, Jacek P Dmochowski, and Blair Kaneshiro. 2017. Nmed-t: A tempo-focused dataset of cortical and behavioral responses to naturalistic music. In *ISMIR*, volume 3, page 5.
- Abhijit Mishra, Shreya Shukla, Jose Torres, Jacek Gwizdka, and Shounak Roychowdhury. 2024. Thought2text: Text generation from eeg signal using large language models (llms). *arXiv preprint arXiv:2410.07507*.
- Rahul Mishra, Krishan Sharma, Ranjeet Ranjan Jha, and Arnav Bhavsar. 2023. Neurogan: image reconstruction from eeg signals via an attention-based gan. *Neural Computing and Applications*, 35(12):9181–9192.
- Tomoaki Mizuno, Takuya Kishida, Natsue Yoshimura, and Toru Nakashika. 2024. An investigation on the speech recovery from eeg signals using transformer. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1–6. IEEE.
- Saydul Akbar Murad and Nick Rahimi. 2024. Unveiling thoughts: A review of advancements in eeg brain signal decoding into text. *arXiv preprint arXiv:2405.00726*.
- D Nemrodov, M Niemeier, A Patel, and A Nestor. 2018. The neural dynamics of facial identity processing: insights from eeg-based pattern analysis and image reconstruction.
- Taiki Orima and Isamu Motoyoshi. 2021. Analysis and synthesis of natural texture perception from visual evoked potentials. *Frontiers in Neuroscience*, 15:698940.
- Simone Palazzo, Concetto Spampinato, Isaak Kavasidis, Daniela Giordano, Joseph Schmidt, and Mubarak Shah. 2020. Decoding brain representations by multimodal learning of neural activity and visual features. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3833–3849.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.
- Ji-Ha Park, Seo-Hyun Lee, and Seong-Whan Lee. 2024. Towards eeg-based talking-face generation for brain signal-driven dynamic communication. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1–5. IEEE.
- Emilian Postolache, Natalia Polouliakh, Hiroaki Kitano, Akima Connelly, Emanuele Rodolà, Luca Cosmo, and Taketo Akama. 2024. Naturalistic music decoding from eeg data via latent diffusion models. *arXiv preprint arXiv:2405.09062*.
- KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. 2020. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pages 484–492.
- Adolfo G Ramirez-Aristizabal and Chris Kello. 2022. Eeg2mel: Reconstructing sound from brain responses to music. *arXiv preprint arXiv:2207.13845*.
- Vasundhara S Rathod, Ashish Tiwari, and Omprakash G Kakde. 2024. Folded ensemble deep learning based text generation on the brain signal. *Multimedia Tools and Applications*, pages 1–29.
- Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. *Advances in neural information processing systems*, 29.
- Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. *arXiv preprint arXiv:2004.04696*.
- H Shimizu and R Srinivasan. 2022. Improving classification and reconstruction of imagined images from eeg signals. *bioRxiv*. retrieved july 5, 2022.
- Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, and Shanmuganathan Raman. 2024. Learning robust deep visual representations from eeg brain recordings. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7553–7562.

- Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. 2023. Eeg2image: image reconstruction from eeg brain signals. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.
- Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. 2023. Decoding natural images from eeg for object recognition. *arXiv preprint arXiv:2308.13234*.
- Yonghao Song, Lie Yang, Xueyu Jia, and Longhan Xie. 2021. Common spatial generative adversarial networks based eeg data augmentation for cross-subject brain-computer interface. *arXiv preprint arXiv:2102.04456*.
- Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. 2017. Deep learning human mind for automated visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6809–6817.
- Aditya Srivastava and Tanvi Shinde. 2020. Think2type: Thoughts to text using eeg waves. *International Journal of Engineering Research & Technology (IJERT)*, 9(06):2278–018.
- Yuma Sugimoto, Goragod Pongthanisor, and Genci Capi. 2024. Image generation using eeg data: A contrastive learning based approach. In *2024 IEEE Canadian Conference on Electrical and Computer Engineering (CCECE)*, pages 794–798. IEEE.
- Congzhong Sun and Chaozhou Mou. 2023. Survey on the research direction of eeg-based signal processing. *Frontiers in Neuroscience*, 17:1203059.
- Yitian Tao, Yan Liang, Luoyu Wang, Yongqing Li, Qing Yang, and Han Zhang. 2024. See: Semantically aligned eeg-to-text translation. *arXiv preprint arXiv:2409.16312*.
- Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. 2018. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 950–958.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*.
- Suguru Wakita, Taiki Orima, and Isamu Motoyoshi. 2021. Photorealistic reconstruction of visual texture from eeg signals. *Frontiers in Computational Neuroscience*, 15:754587.
- Jiaqi Wang, Zhenxi Song, Zhengyu Ma, Xipeng Qiu, Min Zhang, and Zhiguo Zhang. 2024. Enhancing eeg-to-text decoding through transferable representations from pre-trained contrastive eeg-text masked autoencoder. *arXiv preprint arXiv:2402.17433*.
- Zhenhailong Wang and Heng Ji. 2022. Open vocabulary electroencephalography-to-text decoding and zero-shot sentiment classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 5350–5358.
- Weining Weng, Yang Gu, Shuai Guo, Yuan Ma, Zhao-hua Yang, Yuchen Liu, and Yiqiang Chen. 2024. Self-supervised learning for electroencephalogram: A systematic survey. *arXiv preprint arXiv:2401.05446*.
- Jonathan R Wolpaw and Chadwick B Boulay. 2010. Brain signals for brain-computer interfaces. In *Brain-Computer Interfaces: Revolutionizing Human-Computer Interaction*, pages 29–46. Springer.
- Jonathan Xu, Bruno Aristimunha, Max Emanuel Feucht, Emma Qian, Charles Liu, Tazik Shahjahan, Martyna Spyra, Steven Zifan Zhang, Nicholas Short, Jioh Kim, et al. 2024. Alljoined—a dataset for eeg-to-image decoding. *arXiv preprint arXiv:2404.05553*.
- Jing Yang, Muhammad Awais, Md Amzad Hossain, Lip Yee, Ma Haowei, Ibrahim M Mehedi, and AIM Iskanderani. 2023. Thoughts of brain eeg signal-to-text conversion using weighted feature fusion-based multiscale dilated adaptive densenet with attention mechanism. *Biomedical Signal Processing and Control*, 86:105120.
- Terrance Yu-Hao Chen, Yulin Chen, Pontus Soederhaell, Sadrishya Agrawal, and Kateryna Shapovalenko. 2025. Decoding eeg speech perception with transformers and vae-based data augmentation. *arXiv e-prints*, pages arXiv–2501.
- Thorsten O Zander, Christian Kothe, Sabine Jatzev, and Matti Gaertner. 2010. Enhancing human-computer interaction with input from active and passive brain-computer interfaces. *Brain-computer interfaces: Applying our minds to human-computer interaction*, pages 181–199.
- Hong Zeng, Nianzhang Xia, Dongguan Qian, Motonobu Hattori, Chu Wang, and Wanzeng Kong. 2023a. Dmre2i: A framework based on diffusion model for the reconstruction from eeg to image. *Biomedical Signal Processing and Control*, 86:105125.
- Hong Zeng, Nianzhang Xia, Ming Tao, Deng Pan, Hao-hao Zheng, Chu Wang, Feifan Xu, Wael Zakaria, and Guojun Dai. 2023b. Dcae: A dual conditional autoencoder framework for the reconstruction from eeg into image. *Biomedical Signal Processing and Control*, 81:104440.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Shunan Zhao and Frank Rudzicz. 2015. Classifying phonological categories in imagined and articulated speech. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 992–996. IEEE.