EIT-1M: One Million EEG-Image-Text Pairs for Human Visual-textual Recognition and More

Xu Zheng[1] *   Ling Wang[1] *   Kanghao Chen[1] †   Yuanhuiyi Lyu[1] †   Jiazhou Zhou[1]   Lin Wang[1,2] ‡

[1]AI Thrust, HKUST(GZ)   [2]Dept. of CSE, HKUST

{yuanhuiyilv, jiazhouzhou}@hkust-gz.edu.cn, zhengxu128@gmail.com, linwang@ust.hk

Abstract

*Recently, electroencephalography (EEG) signals have been actively incorporated to decode brain activity to visual or textual stimuli and achieve object recognition in multimodal AI. Accordingly, endeavors have been focused on building EEG-based datasets from visual or textual single-modal stimuli. However, these datasets offer limited EEG epochs per category, and the complex semantics of stimuli presented to participants compromise their quality and fidelity in capturing precise brain activity. The study in neuroscience unveils that the relationship between visual and textual stimulus in EEG recordings provides valuable insights into the brain's ability to process and integrate multi-modal information simultaneously. Inspired by this, we propose a novel large-scale multi-modal dataset, named **EIT-1M**, with over 1 million EEG-image-text pairs. Our dataset is superior in its capacity of reflecting brain activities in simultaneously processing multi-modal information. To achieve this, we collected data pairs while participants viewed alternating sequences of visual-textual stimuli from 60K natural images and category-specific texts. Common semantic categories are also included to elicit better reactions from participants' brains. Meanwhile, response-based stimulus timing and repetition across blocks and sessions are included to ensure data diversity. To verify the effectiveness of EIT-1M, we provide an in-depth analysis of EEG data captured from multi-modal stimuli across different categories and participants, along with data quality scores for transparency. We demonstrate its validity on two tasks: 1) EEG recognition from visual or textual stimuli or both and 2) EEG-to-visual generation.*

1. Introduction

Electroencephalography (EEG) is a widely applied neuroimaging modality in cognitive neuroscience. It is known for its ability to decipher intricate brain activity patterns during various cognitive processes [29]. In the early days, research focused on constructing EEG datasets for medical purposes, such as detecting and predicting seizures [36]. Recently, EEG signals have been broadly incorporated to decode brain activity to visual or textual stimuli and achieve object recognition in multi-modal artificial intelligence (AI) [3, 7, 28, 31–33]. This enriches the data landscape, allowing for more nuanced and accurate models of brain activity and cognitive processes.

Accordingly, research endeavors have been focused on building EEG-based datasets [10, 11, 14, 18, 34], as summarized in Tab. 1. For instance, ZuCo 1.0 [13] is a pioneering EEG-Text dataset that records neural processes underlying reading and language comprehension during the reading tasks. On the other hand, Brain2Image [17] is a representative EEG-image dataset that includes evoked responses to visual stimuli from 40 classes. However, these datasets have two distinct shortcomings: *1)* They offer limited EEG epochs per category, and the complex semantics of stimuli presented to participants compromise their quality and fidelity in capturing precise brain activity. *2)* They only encompass EEG signals recorded from single-modal stimuli, either visual or textual. This makes them less possible to be used for training high-performance multi-modal AI models.

The study in neural science reveals that EEG recordings reveal a significant relationship between visual and textual stimuli, offering valuable insights into the brain's capacity to integrate multi-modal information simultaneously [29]. This integration is crucial for understanding how the brain processes complex, real-world scenarios where multiple types of sensory input are encountered simultaneously. Inspired by this, we introduce a novel large-scale multi-modal EEG dataset, **EIT-1M**, comprising paired EEG, visual, and textual data for the benefit of research communities. The key insight of our dataset is to record human brain activities while simultaneously processing multi-modal information. To achieve this, data was collected from five participants exposed to random sequences of 60K natural images and their corresponding category descriptions. To date, we have

---

*† equal contribution. ‡ corresponding author.

Dataset	Year	Equipment	Modality	Epochs	CEA	MEA	Purpose
Brain2Image [18]	2017	Brainvision BrainAmp DC	EEG-Image	11,466	×	×	Decoding
EVSR [21]	2018	Emotiv EPOC+	EEG-Image	13,800	✓	×	Recognition
ZuCo 1.0 [13]	2018	EGI Geodesic Hydrocel system	EEG-Text	259,788	×	×	NLP
ZuCo 2.0 [14]	2020	EGI Geodesic Hydrocel system	EEG-Text	272,484	×	×	NLP
THINGS EEG1 [11]	2022	Brainvision actiCHamp	EEG-Image	1,112,400	✓	×	Recognition
THINGS EEG2 [9]	2022	Brainvision actiCHamp	EEG-Image	821,600	×	×	Recognition
Alljoined1 [37]	2024	BioSemi ActiveTwo	EEG-Image	46,080	×	×	Decoding
EIT-1M (Ours)	2024	Brainvision actiCHamp Plus	EEG-Image-Text	1,200,000	✓	✓	Recognition & Decoding

Table 1. EEG datasets. (CEA: Category-level ERP Analysis, MEA: Multi-modal ERP Analysis.)

gathered over **1 million** epochs of brain responses using a 64-channel EEG headset (actiCHamp Plus[1]).

Specifically, we utilize the 10-category dataset CIFAR-10 [20] to construct the visual and textual stimulus. This dataset harnesses an image resolution of 32×32 pixels without excessive details, Empirically, as shown in Fig. 2, we find that low-resolution visual stimuli stimulate more stable neural responses, suggesting they are appropriate and manageable within a brief viewing period. We present visual and textual stimuli sequentially to maintain continuous engagement with the objects and concepts, as shown in Fig. 3. Moreover, our dataset features response-based stimulus timing, repetition across blocks and sessions, and diverse visual and textual classes. To verify the effectiveness, we provide an in-depth analysis of EEG data captured from multi-modal stimuli across different categories and participants. The data analysis includes EEG topographic maps, corresponding signals analysis and ERP analysis. These analysis highlight the distinct ERP characteristics from visual and textual stimuli, providing insights in the multi-modal information processing of brains. For transparency, we include data quality scores (See Tab. 3).

To benchmark our EIT-1M, we demonstrate its validity on two tasks: 1) EEG recognition from visual or textual stimuli or both (See Sec. 5.1) and 2) EEG-to-visual generation (See Sec. 5.2). We expect our dataset to be a benchmark contributor for advancing the research for multi-modal AI [5, 6, 25, 26, 38–41] and potentially for cognitive neuroscience.

2. Related Work

EEG Datasets with Visual Stimuli. They capture EEG waveforms while participants view visual stimuli, facilitating studies of brain activity, as shown in Tab. 1. A representative dataset is Brain2Image [17], which includes evoked responses to visual stimuli from 40 classes, each with 50 images, totaling 2K images. However, this dataset is impeded by its lack of train-test separation during recording, block-specific stimuli patterns, and inconsistency across frequency bands [4, 23]. In contrast, the THINGS-EEG1 [11] and THINGS-EEG2 [9] datasets address these issues by incorporating both main and validation sessions to ensure data quality and consistency. These two datasets contain human EEG responses from 50 subjects to 22,248 images in the THINGS stimulus set.

Regarding the diversity of stimuli, while studies like Brain2Image and [1] involve 40 classes, other studies focus on only 10 different image classes [34]. This limited representation allows for more controlled studies but fails to capture the continuous and diverse nature of naturalistic stimuli due to the limited samples from each category. Other datasets like MindBigData [35] and [1] capture a wide range of images but are derived from a single individual, limiting their potential for training image reconstruction models that generalize to other individuals. Recently, to address these limitations, Alljoined1 [37] includes 10K images per participant from object categories in MS-COCO [24], thereby accounting for the diversity and continuity of real-world images.

EEG Datasets with Textual Stimuli They are primarily developed for brain signal decoding. Notable examples include ZuCo 1.0 [13] and ZuCo 2.0 [14], captured with 128-channel EEG devices. These datasets provide insights into the neural processes underlying reading and language comprehension by recording EEG signals during reading tasks. EEG2Text [2] focuses on translating brain signals into textual descriptions, supporting the development of AI models for decoding and generating text from EEG signals. Despite these advancements, there remains a need for datasets that integrate both visual and textual stimuli to capture the com-
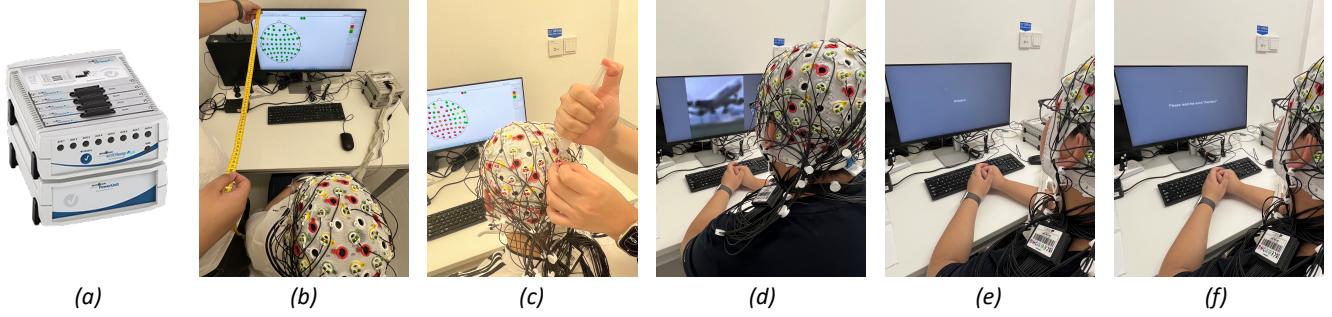

2

Figure 1. (a) actiCHamp Plus device. (b) Experimental setup with monitor 80 cm from participant. (c) Injecting conductive gel. (d) Visual stimuli. (e) Textual stimuli. (f) Speech stimuli.
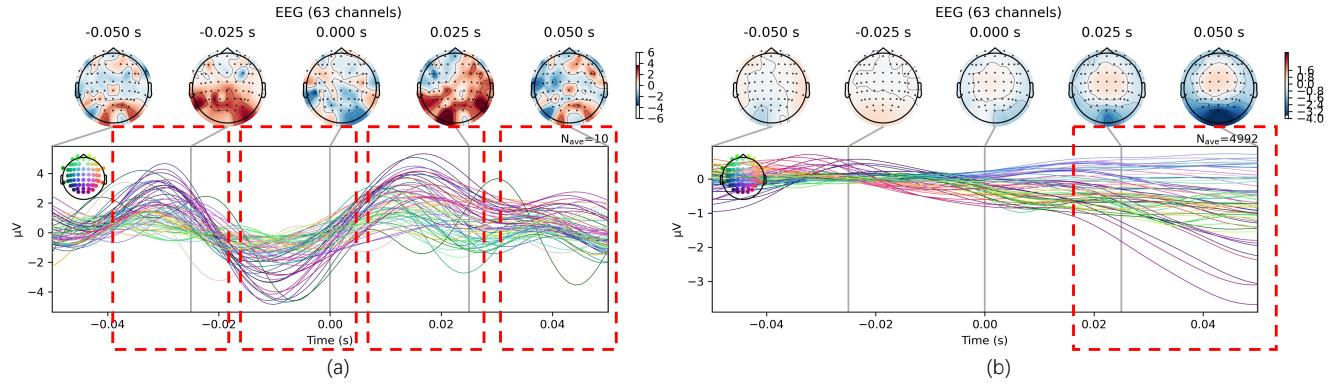

Figure 2. (a) EEG signals from high-resolution visual stimuli.(b) EEG signals from our visual stimuli.

plex interplay between different modalities in the brain. In a nutshell, all these datasets primarily focus on single-modal stimuli, limiting their fidelity for training multi-modal AI models. *Our EIT-1M dataset addresses this gap by providing paired EEG, visual, and textual data, enabling comprehensive multi-modal analysis. Thus our dataset is superior in its capacity of reflecting brain activities in simultaneously processing multi-modal information.*

3. Dataset Collection Methods

Tab. 2 provides an overview of one experiment involving five participants, aged 20-30 years, with a gender distribution of two females and four males. Each participant underwent two 300-minute sessions, during which 1,200,000 events were recorded, including 600K visual and 600K textual stimuli. The stimuli were drawn from ten CIFAR-10 categories for visuals and ten textual categories. EEG recordings were made using a 64-channel headset at a 1000 Hz sampling rate. The dataset ensures high quality with an average signal-to-noise ratio as in Tab. 3, maintaining impedance levels at or below 20 kΩ. Each session featured an average of 10K events, with each event lasting 50 ms and an inter-event interval of 1 second. Preprocessing involved

1-40 Hz band-pass filtering and epoching from -20 to 30 ms relative to stimulus onset, with baseline correction at -20 ms. This dataset aims to support research in EEG analysis and multi-modal recognition.

3.1. Experimental Settings

Participants Five adults (mean age 24.83 years; 1 female, 4 male) participated in this study, all with normal or corrected-to-normal vision, and none of them have suffered or are suffering neurological or psychiatric problems such as ADHD and epilepsy. Each participant provided informed written consent and received monetary reimbursement for their involvement. The study procedures were approved by the ethical committee. It is important to acknowledge the potential limitations of this study, such as the gender imbalance among participants and the low age disparity.

Stimuli. All images used in this study as visual stimuli are sourced from the CIFAR-10 dataset [20]. This dataset is a well-known benchmark in machine learning and computer vision, comprising 60K color images across 10 different classes, with 6K images per class. These classes represent a variety of everyday objects and animals, including airplanes, automobiles, birds, cats, deer, dogs, frogs, horses, ships, and trucks. This dataset is selected for our study due

Item	Description	Details
Participants	Number	5
	Age Range	20-30 years
	Gender Distribution	1 females, 4 males
Sessions	Number per Participant	2
	Duration	4 hours each
Total Events	Total	1,200,000
	Visual Stimuli	600,000
	Text Stimuli	600,000
Stimuli	Categories	10 from CIFAR-10
	Description	Visual: Images from CIFAR-10; Text: category names
Recording Details	Sampling Rate	1000 Hz
	EEG Channels	64
	Equipment	actiCHamp Plus[2]
Data Quality	Impedance Levels	$\leq 20\,\mathrm{k\Omega}$
Event Details	Average Events/Session	120,000
	Event Duration	50 ms
	Inter-event Interval	50 ms
Preprocessing	Filtering	1-40 Hz band-pass
	Epoching	-50 to 50 ms relative to stimulus onset, baseline correction at -50 ms

Table 2. Overview of our proposed EEG-Image-Text Dataset

Category	Airplane	Automobile	Bird	Cat	Deer	Dog	Frog	Horse	Ship	Truck
Average SNR / dB (raw data)	6.14	5.76	6.09	5.69	5.44	4.82	4.47	4.74	4.30	3.67

Table 3. Example raw data quality of participant 04 in our proposed dataset across different blocks (categories) within the first session.

to its diversity and balanced categories, which provide a robust set of stimuli for examining neural responses across different visual contexts. Utilizing this dataset allows for an in-depth exploration of how the brain processes various types of visual information and supports the development of multi-modal models that can generalize across different categories of visual stimuli. Each image in this dataset has a resolution of 32x32 pixels, making it ideal for stimulating brain activities for visual stimuli from participants. The textual stimuli are derived from the category names within the CIFAR-10 dataset: airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

Hardware Setup We recorded data using a 64-electrode actiCHamp Plus system, digitized at a rate of 1024 Hz with 24-bit A/D conversion. The montage was arranged according to the international 10-20 System, and the electrode offset was kept below 40 mV. A 22-inch Dell monitor with a resolution of 1080p at 60 Hz was used to display the visual and textual stimuli. As shown in Fig. 1, the monitor was centrally positioned at a distance of 80 cm from the participant, maintaining a 3.5-degree angle of stimuli. We ensured that the angle remained small to minimize the occurrence of gaze drift.

3.2. Data Collection Procedure

Before viewing the stimuli, conductive gel was injected into each electrode to ensure the resistance was less than 20 ohms, facilitating better signal capture. Participants were then shown images and text over the course of four sessions, each four hours long. Each session comprised multiple blocks, with each block containing images from the same class and the corresponding category name text. The visual and textual stimuli were randomly arranged in a visual-textual-visual-textual order within each block. Different blocks contained stimuli from different classes. Within each block, 1,000 visual stimuli images and 1,000 text stimuli category names from CIFAR-10 were presented.

Within each trial, an image was presented for 50 ms, followed by 50 ms of a black screen. The corresponding category name of the image was also presented for 50 ms, followed by 50 ms of a black screen. A white fixation cross was visible on the screen throughout the entire trial. To ensure focus, participants were prompted to press the space bar after completing two consecutive blocks. Additionally, five to ten-minute breaks were provided between blocks based on participants' needs for better data recording. Fig. 3
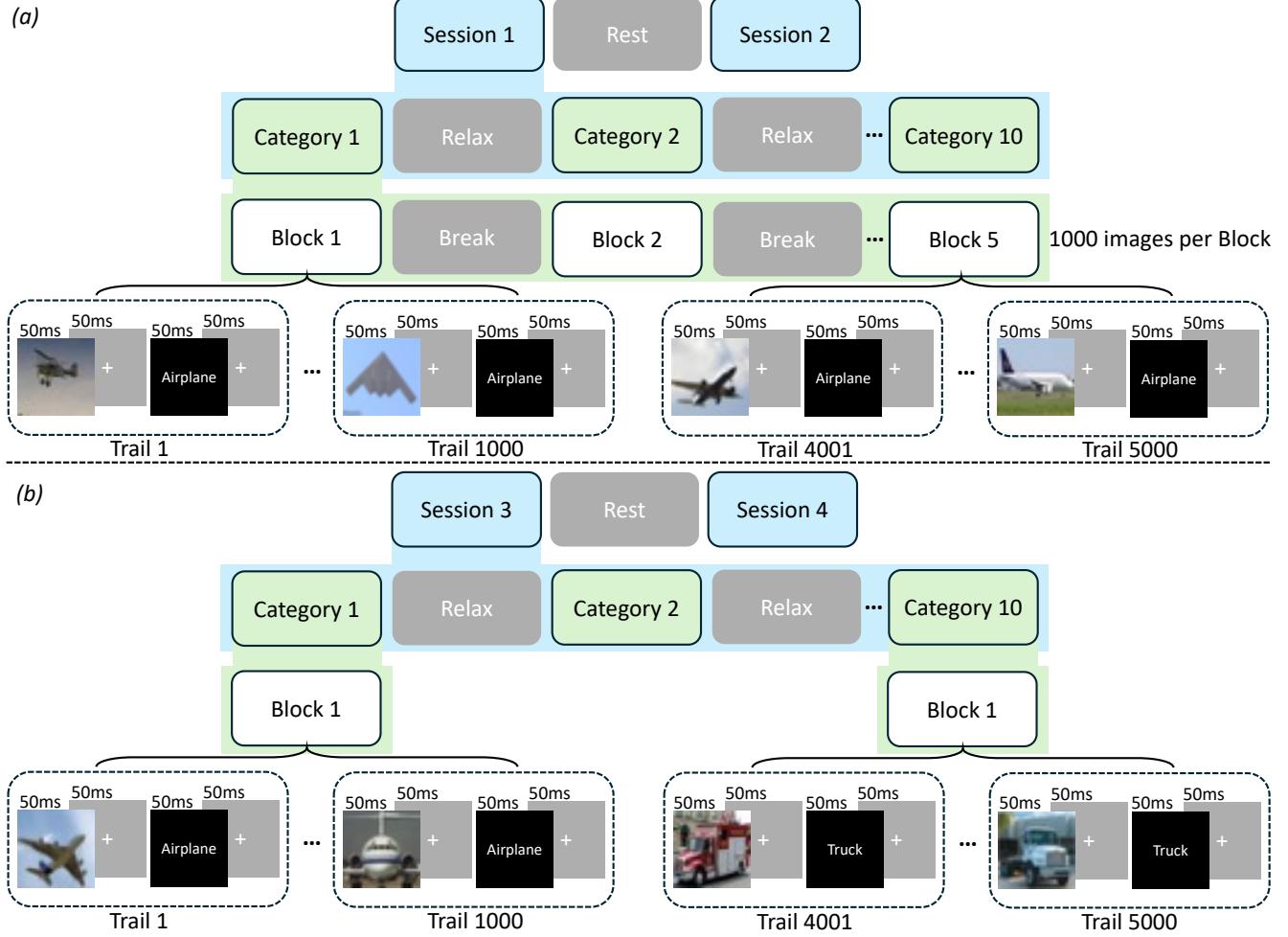
Figure 3. Schematic overview of the structure of trials, blocks, categories and sessions with RSVP paradigm. (a) Training set of CIFAR-10 dataset; (b) Testing set of CIFAR-10 dataset.

(a) shows a schematic overview of the structure of trials, blocks, categories, and sessions, which follows rapid serial visual presentation (RSVP) paradigm [11, 16, 19]. Each of the 5 block-specific CIFAR-10 training images and label text is presented once within each block, and each of the 10 category-specific blocks is presented once within each session. Each participant performed two sessions on different days. Each of the 2 sessions thus consists of 50,000 images and texts within and across blocks. Fig. 3 (b) illustrates that Sessions 3 and 4 consist of 10,000 images and labels from the CIFAR-10 testing set.

4. Data Analysis

4.1. EEG Topographic Maps and Corresponding Signals Analysis

Fig. 4 presents the comparison of EEG signals by showcasing topographic maps and corresponding signals aver-

aged across 63 electrodes (channel FCz as reference) for different stimuli conditions, *i.e.*, visual and textual stimuli with airplane and frog categories. Each column of the figure represents a different stimulus type: visual stimuli (left column) and textual stimuli (right column). The visual stimuli include images from the CIFAR-10 dataset, and the textual stimuli comprise category names from the same dataset. Each row represents different categories, specifically airplane and frog.

Visual Stimuli (Left Column of Fig. 4) The topographic maps show the distribution of brain activity across the scalp at various time points (-0.050s, -0.025s, 0.000s, 0.025s, and 0.050s) after the stimulus onset. The maps reveal distinct patterns of neural activation, indicating how the brain processes visual stimuli over time. For instance, the airplane category (1st row) shows significant activation in the occipital and parietal regions, which are known to be involved in visual processing [27]. The corresponding ERP signals
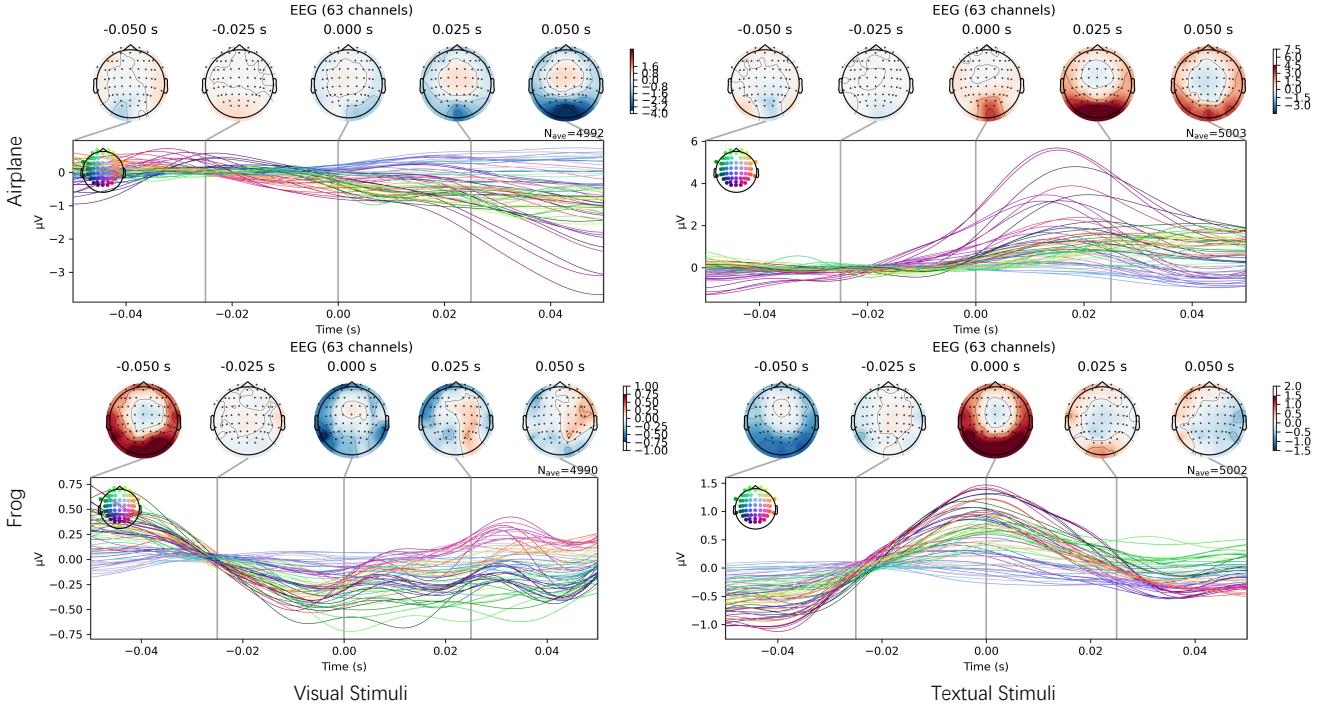
Figure 4. EEG topographic maps and corresponding signals averaged over events for the participant viewing visual stimuli (**left column**) viewing the airplane (1st row) and frog (2nd row) images from the CIFAR-10 dataset, and events for the participant viewing textual stimuli (**right column**) viewing the airplane (1st row) and frog (2nd row) text.

show the average response over time for all electrodes. The signals depict the dynamic changes in brain activity, with notable peaks and troughs corresponding to different cognitive processes. For the visual stimuli, there are clear ERP components around 20ms and 40ms, which might correspond to early visual processing and higher-level cognitive processing, respectively.

Textual Stimuli (Right Column of Fig. 4) Similar to the visual stimuli, the topographic maps for textual stimuli show brain activity at the time points in 50 ms later as visual stimuli. Note that the visual and textual stimuli are presented with a gap of 50 ms. There are noticeable differences in the activation patterns compared to visual stimuli, highlighting the distinct neural processes involved in reading and understanding texts of the participants. For the airplane text (1st row), there is significant activation in the temporal and frontal regions, areas associated with language processing [15]. The ERP signals for textual stimuli also display characteristic peaks, though the patterns differ from those elicited by visual stimuli. The airplane text category shows a strong response between 20 ms to 40 ms, likely reflecting early semantic processing [8].

According to these visualizations, we have the following findings: **(I) Individual and Common Patterns**: Fig. 4 highlights both individual and common brain activity patterns associated with both image and text presentation. This indicates that while there are distinct neural processes for visual and textual stimuli, there are also commonalities in how the brain responds to different types of information. **(II) Temporal Dynamics:** The temporal dynamics of the ERP signals provide insights into the timing of cognitive processes. Early components (within the last 20ms) are typically associated with sensory processing, while earlier components (before 20ms) are linked to cognitive and semantic processing. **(III) Gap Influence**: The 50 ms gap between visual and textual stimuli presentations allowed us to observe the sequential processing of different modalities, showing how the brain transitions between visual and textual information processing. **(IV) ERP Characteristics**: The ERP characteristics, such as the peaks around 20 ms and 40 ms for visual stimuli and between 0 ms to 20 ms for textual stimuli, provide valuable hints for understanding the stages of information processing in the brain.

Unlike previous EEG datasets, such as the THINGS-EEG dataset [11], which use high-resolution images as visual stimuli and introduce a vast number of object concepts (1854), our datasets can address the following limitations of previous ones. Fig. 2 illustrates the differences in EEG signal responses between high-resolution and lower-resolution visual stimuli. The more stable and less variable neural re-
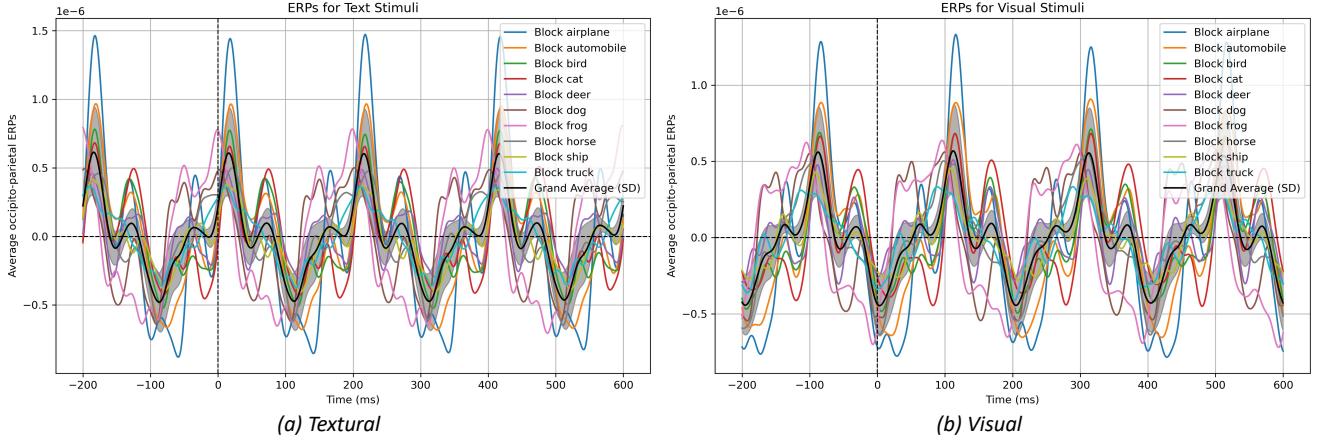
Figure 5. ERPs averaged over occipital and parietal electrodes for the participant viewing stimuli from (a) visual images and (b) the category text. Shaded areas around the grand average ERP represent standard deviations at each time point.

sponses to the lower-resolution images suggest their suitability for creating robust EEG datasets. High-resolution images, on the other hand, require more time for participants to process content and details, making them less suitable for effectively capturing quick neural responses at the millisecond level.

4.2. ERP Analysis

Fig. 5 presents the event-related potentials (ERPs) averaged over occipital and parietal electrodes for a participant viewing visual images (right panel) and category text (left panel). Both plots display ERP data from -200 ms to 600 ms relative to stimulus onset (0 ms), with the average occipito-parietal ERPs fluctuating between approximately -0.5 and 1.5 microvolts for both visual and text stimuli. Each trace, with a distinct color, represents a specific category, including airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.

Regarding the text stimuli (left panel), a significant initial deflection is noticed around 0 ms, showing the brain's quick response to text stimuli. Early components, like peaks and troughs, are seen around 100 ms and 200 ms post-stimulus, typical of early ERP components such as the P1 and N1, which are linked to sensory processing. Additional peaks around 300 ms and beyond likely indicate higher-order cognitive processing. The shaded area around the grand average ERP line signifies the standard deviation, reflecting variability across different trials and categories. This variability is higher at certain peaks, suggesting differences in how the brain processes various text categories.

Concerning the visual stimuli (right panel), a comparable initial deflection is observed around 0 ms. Distinct peaks are evident at approximately 100 ms and 200 ms, corresponding to the P1 and N1 components, which are more pronounced

and consistent across different visual categories compared to textual stimuli. Significant peaks around 300 ms and later may denote the P3 component, indicating cognitive processing associated with visual categorization. The standard deviation shading around the grand average indicates less variability compared to text stimuli, suggesting more consistent brain responses to visual stimuli across various categories.

In comparison, visual stimuli evoke more consistent ERPs across categories than text stimuli, as indicated by the smaller standard deviation areas. Both types of stimuli elicit similar amplitude ranges in the ERP responses, reflecting comparable levels of neural activity. The timing of early and late ERP components is similar for both text and visual stimuli, suggesting that initial sensory processing and subsequent cognitive processing occur within similar time frames for both types of stimuli. In conclusion, the ERPs for both textual and visual stimuli exhibit characteristic early and late components, indicative of sensory and cognitive processing stages. Visual stimuli elicit more consistent responses across categories, whereas text stimuli exhibit greater variability. This analysis provides insights into the sensory and cognitive functions associated with different types of stimuli.

5. Experiments with EIT-1M Dataset

Implementation Details. For preprocessing, a band-pass filter is applied to retain frequencies between 1 and 40 Hz within the raw EEG data. Subsequently, the continuous data is segmented into epochs, each commencing 50 ms prior to the stimulus onset and concluding 50 ms following each event. To train and evaluate the recognition models, the EEG data from one participant (Tab. 4) and two participants (Tab. 5) are divided using an 80/20% split to create training

Models	Image			Text			Image & Text		
	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
EEGNet [22]	25.42	25.63	24.75	25.62	26.30	24.77	20.72	20.96	19.69
MobileNet_v2 [30]	40.84	41.61	40.79	40.17	39.64	39.52	49.76	49.57	49.32
ResNet18 [12]	**56.57**	**56.41**	**56.46**	56.38	56.17	56.17	**63.53**	**63.65**	**63.55**
ResNet34 [12]	56.41	56.15	56.24	**56.47**	**56.34**	**56.39**	58.77	59.22	58.89
ResNet50 [12]	49.45	48.49	48.80	49.93	49.46	49.61	49.34	50.26	49.38

Table 4. Benchmark experiments within one session of one participant.

Models	Image			Text			Image & Text		
	Acc	Recall	F1	Acc	Recall	F1	Acc	Recall	F1
EEGNet [22]	22.08	22.24	21.80	24.15	24.11	23.56	20.73	20.68	19.83
MobileNet_v2 [30]	41.90	42.18	41.66	42.67	43.32	41.59	48.97	49.28	48.86
ResNet18 [12]	53.49	**53.73**	**53.45**	**54.11**	**54.10**	**54.10**	58.60	58.65	58.54
ResNet34 [12]	**54.06**	53.27	52.42	53.57	53.30	52.65	**60.69**	**60.91**	**60.76**
ResNet50 [12]	49.80	49.60	48.53	49.82	48.81	48.17	56.07	54.98	54.98

Table 5. Benchmark experiments across different sessions of two participants.

| *Airplane* | *Automobile* | *Bird* | *Cat* | *Dog* | *Horse* | *Ship* | *Truck* |

Figure 6. Generation results of our dataset using the ThoughtVis Model.

and evaluation sets, respectively. The models are trained using the Adam optimizer, coupled with a step learning rate schedule, across 500 epochs. The default settings for the learning rate, weight decay, and batch size are 1×10^{-3}, 1×10^{-5}, and 2048, respectively. We apply three widely-used metric to evaluate the recognition performance on EIT-1M, including Accuracy, recall, and F1 score.

5.1. Recognition

The results of experiments conducted within one session of a single participant are shown in Tab. 4, illustrating the effectiveness of our dataset in the individual collection procedure. The results in Tab. 4 include the performance across various models with EEG signals captured from visual and textual stimuli. Note that $Image\&Text$ refers to the combined EEG signals from both visual and textual stimuli for recognition. The evaluated models include EEGNet [22], MobileNet-v2 [30], ResNet18 [12], ResNet34 [12], and ResNet50 [12].

Combining EEG signals from image and text stimuli generally enhances performance metrics across all models, suggesting that multi-modal data provides richer informa-

tion, leading to better classification accuracy and robustness. The consistent performance improvements observed from MobileNet-v2 to ResNet architectures indicate that our EIT-1M dataset is well-suited for various deep-learning models. ResNet models, in particular, show significant improvements, highlighting the dataset's capacity to support complex neural networks. Similar performance metrics for image and text stimuli alone indicate that the dataset offers a balanced representation of both modalities. This balance is crucial for training models to generalize well across different types of stimuli. Additionally, the high F1 scores, especially for the ResNet models, reflect good data quality, ensuring that the recorded EEG signals are reliable and effective for training AI models. Tab. 5 summarizes benchmark experiments across different sessions of two participants. The results consistently show that combining EEG signals from both visual and textual stimuli improves performance across all models compared to using either visual or textual stimuli alone. For both visual and textual stimuli, ResNet models maintain consistently high performance, indicating the robustness of ResNet architectures in processing and learning from EEG data.

The analysis of Tab. 4 and Tab. 5 supports the rationality of our EIT-1M dataset. By providing high-quality, balanced, and scalable data, our dataset proves to be an excellent resource for advancing research in multi-modal AI and cognitive neuroscience. The observed improvements in combined image and text stimuli further highlight the importance of multi-modal datasets in capturing the intricate interplay between different types of information.

5.2. Generation

We follow the classic EEG-to-Image generation task proposed by ThoughtVis [34], which obtains images from EEG signals. As shown in the generation results in Fig. 6, our proposed EIT-1M dataset shows the capability to support the EEG-to-Image generation task.

6. Conclusion, Limitations, and Future Work

In this paper, we presented EIT-1M, a large-scale multi-modal dataset comprising 1 million EEG-image-text pairs. We collected the data pairs while participants viewed alternating sequences of visual-textual stimuli from 60K natural images and corresponding label texts. Our EIT-1M is superior in its capacity of recording brain activities in simultaneously processing multi-modal information, *i.e.*, images and text. It features response-based stimulus timing and repetition across blocks and sessions. To verify the effectiveness of EIT-1M, we provided an in-depth analysis of the EEG signals in EIT-1M across different categories and sessions and conducted experiments on two tasks.

Limitations. Despite the robustness of our dataset, there are areas for enhancement. Our current dataset includes data from multiple participants and sessions, but increasing the number of participants and sessions could yield a more comprehensive understanding of neural responses and improve the generalizability of the models trained on this data. Additionally, while we used a well-defined set of visual and textual stimuli, expanding the variety of stimuli, especially for the textual stimuli, could further enhance the dataset's fidelity for studying more diverse and complex neural processes.

Future work. It could be a good direction to integrate additional modalities, such as audio or tactile feedback, to create an even richer multi-modal dataset. This integration could provide deeper insights into the interplay between different sensory inputs and brain activity, advancing research in multi-modal AI and neuroscience. By addressing these limitations and expanding the dataset's scope, we can significantly contribute to the understanding and development of multi-modal AI models.

Broader Impact. EIT-1M advances neuroscience and AI by enabling deeper insights into cognitive processes and sensory integration. It improves brain-computer interfaces and personalized learning. Ethical considerations regarding neural data privacy are crucial for responsible applications.

References

[1] Hamad Ahmed, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. Object classification from randomized EEG trials. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 3845–3854. Computer Vision Foundation / IEEE, 2021. 2

[2] A. Author and B. Author. Eeg2text: Decoding text from brain signals. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pages 12345–12353, 2019. 2

[3] Yohann Benchetrit, Hubert Banville, and Jean-Rémi King. Brain decoding: toward real-time reconstruction of visual perception. *arXiv preprint arXiv:2310.19812*, 2023. 1

[4] Hari M. Bharadwaj, Ronnie B. Wilbur, and Jeffrey Mark Siskind. Still an ineffective method with supertrials/erps - comments on "decoding brain representations by multimodal learning of neural activity and visual features". *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):14052–14054, 2023. 2

[5] Jiahang Cao, Xu Zheng, Yuanhuiyi Lyu, Jiaxu Wang, Renjing Xu, and Lin Wang. Chasing day and night: Towards robust and efficient all-day object detection guided by an event camera. *arXiv preprint arXiv:2309.09297*, 2023. 2

[6] Jialei Chen, Daisuke Deguchi, Chenkai Zhang, Xu Zheng, and Hiroshi Murase. Clip is also a good teacher: A new learning framework for inductive zero-shot semantic segmentation. *arXiv preprint arXiv:2310.02296*, 2023. 2

[7] Michael X Cohen. Where does eeg come from and what does it mean? *Trends in neurosciences*, 40(4):208–218, 2017. 1

[8] Michelle E Costanzo, Joseph J McArdle, Bruce Swett, Vladimir Nechaev, Stefan Kemeny, Jiang Xu, and Allen R Braun. Spatial and temporal features of superordinate semantic processing studied with fmri and eeg. *Frontiers in human neuroscience*, 7:293, 2013. 6

[9] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. 2

[10] Alessandro T Gifford, Kshitij Dwivedi, Gemma Roig, and Radoslaw M Cichy. A large and rich eeg dataset for modeling human visual object recognition. *NeuroImage*, 264:119754, 2022. 1

[11] Tijl Grootswagers, Ivy Zhou, Amanda K Robinson, Martin N Hebart, and Thomas A Carlson. Human eeg recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Scientific Data*, 9(1):3, 2022. 1, 2, 5, 6

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 8

[13] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. Zuco, a simultaneous eeg and eye-tracking resource for natural sentence reading. *Scientific data*, 5(1):1–13, 2018. 1, 2

[14] Nora Hollenstein, Marius Troendle, Ce Zhang, and Nicolas Langer. Zuco 2.0: A dataset of physiological recordings during natural reading and annotation. *arXiv preprint arXiv:1912.00903*, 2019. 1, 2

[15] Nora Hollenstein, Cedric Renggli, Benjamin Glaus, Maria Barrett, Marius Troendle, Nicolas Langer, and Ce Zhang. Decoding eeg brain activity for multi-modal natural language processing. *Frontiers in Human Neuroscience*, 15:659410, 2021. 6

[16] Helene Intraub. Rapid conceptual identification of sequentially presented pictures. *Journal of Experimental Psychology: Human Perception and Performance*, 7(3):604, 1981. 5

[17] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. *Brain2Image*: Converting brain signals into images. In *Proceedings of the 2017 ACM on Multimedia Conference, MM 2017, Mountain View, CA, USA, October 23-27, 2017*, pages 1809–1817. ACM, 2017. 1, 2

[18] Isaak Kavasidis, Simone Palazzo, Concetto Spampinato, Daniela Giordano, and Mubarak Shah. Brain2image: Converting brain signals into images. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1809–1817, 2017. 1, 2

[19] Christian Keysers, D-K Xiao, Peter Földiák, and David I Perrett. The speed of sight. *Journal of cognitive neuroscience*, 13(1):90–101, 2001. 5

[20] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2, 3

[21] Pradeep Kumar, Rajkumar Saini, Partha Pratim Roy, Pawan Kumar Sahu, and Debi Prosad Dogra. Envisioned speech recognition using eeg sensors. *Personal and Ubiquitous Computing*, 22:185–199, 2018. 2

[22] Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. Eegnet: a compact convolutional neural network for eeg-based brain–computer interfaces. *Journal of neural engineering*, 15(5):056013, 2018. 8

[23] Ren Li, Jared S. Johansen, Hamad Ahmed, Thomas V. Ilyevsky, Ronnie B. Wilbur, Hari M. Bharadwaj, and Jeffrey Mark Siskind. The perils and pitfalls of block design for EEG classification experiments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(1):316–333, 2021. 2

[24] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V*, pages 740–755. Springer, 2014. 2

[25] Yuanhuiyi Lyu, Xu Zheng, Dahun Kim, and Lin Wang. Omnibind: Teach to build unequal-scale modality interaction for omni-bind of all. *arXiv preprint arXiv:2405.16108*, 2024. 2

[26] Yuanhuiyi Lyu, Xu Zheng, and Lin Wang. Image anything: Towards reasoning-coherent and training-free multi-modal image generation. *arXiv preprint arXiv:2401.17664*, 2024. 2

[27] Amanda K Robinson, Praveen Venkatesh, Matthew J Boring, Michael J Tarr, Pulkit Grover, and Marlene Behrmann. Very high density eeg elucidates spatiotemporal aspects of early visual processing. *Scientific reports*, 7(1):16248, 2017. 5

[28] Yannick Roy, Hubert Banville, Isabela Albuquerque, Alexandre Gramfort, Tiago H Falk, and Jocelyn Faubert. Deep learning-based electroencephalography analysis: a systematic review. *Journal of neural engineering*, 16(5):051001, 2019. 1

[29] Maham Saeidi, Waldemar Karwowski, Farzad V Farahani, Krzysztof Fiok, Redha Taiar, Peter A Hancock, and Awad Al-Juaid. Neural decoding of eeg signals with machine learning: A systematic review. *Brain Sciences*, 11(11):1525, 2021. 1

[30] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4510–4520, 2018. 8

[31] Prajwal Singh, Pankaj Pandey, Krishna Miyapuram, and Shanmuganathan Raman. 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, 2023. 1

[32] Prajwal Singh, Dwip Dalal, Gautam Vashishtha, Krishna Miyapuram, and Shanmuganathan Raman. 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, 2024.

[33] Michal Teplan et al. Fundamentals of eeg measurement. *Measurement science review*, 2(2):1–11, 2002. 1

[34] Praveen Tirupattur, Yogesh Singh Rawat, Concetto Spampinato, and Mubarak Shah. Thoughtviz: Visualizing human thoughts using generative adversarial network. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 950–958, 2018. 1, 2, 9

[35] David Vivancos and Felix Cuesta. Mindbigdata 2022 A large dataset of brain signals. *CoRR*, abs/2212.14746, 2022. 2

[36] Sheng Wong, Anj Simmons, Jessica Rivera-Villicana, Scott Barnett, Shobi Sivathamboo, Piero Perucca, Zongyuan Ge, Patrick Kwan, Levin Kuhlmann, Rajesh Vasa, et al. Eeg datasets for seizure detection and prediction—a review. *Epilepsia Open*, 8(2):252–267, 2023. 1

[37] Jonathan Xu, Bruno Aristimunha, Max Emanuel Feucht, Emma Qian, Charles Liu, Tazik Shahjahan, Martyna Spyra, Steven Zifan Zhang, Nicholas Short, Jioh Kim, Paula Perdomo, Ricky Renfeng Mao, Yashvir Sabharwal, Michael Ahedor Moaz Shoura, and Adrian Nestor. Alljoined - A dataset for eeg-to-image decoding. *CoRR*, abs/2404.05553, 2024. 2

[38] Xu Zheng and Lin Wang. Eventdance: Unsupervised source-free cross-modal adaptation for event-based object recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17448–17458, 2024. 2

[39] Xu Zheng, Yexin Liu, Yunfan Lu, Tongyan Hua, Tianbo Pan, Weiming Zhang, Dacheng Tao, and Lin Wang. Deep learning

for event-based vision: A comprehensive survey and benchmarks. *arXiv preprint arXiv:2302.08890*, 2023.

[40] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. E-clip: Towards label-efficient event-based open-world understanding by clip. *arXiv preprint arXiv:2308.03135*, 2023.

[41] Jiazhou Zhou, Xu Zheng, Yuanhuiyi Lyu, and Lin Wang. Exact: Language-guided conceptual reasoning and uncertainty estimation for event-based action recognition and more. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18633–18643, 2024. 2