Interpreting Sleep Activity Through Neural Contrastive Learning

Zhongtao Chen[1,2#], Hui Zheng[2,3#], Jianyang Zhou[2,4#], Lin Zheng[2,4#], Peiyang Lin[1], Haiteng Wang[1], Ray Dolan[1,5,6], Tim Behrens[7,8], Yunzhe Liu[1,2*]

1. State Key Laboratory of Cognitive Neuroscience and Learning, IDG/McGovern Institute for Brain Research, Beijing Normal University, Beijing, China
2. Chinese Institute for Brain Research, Beijing, China
3. Academy for Advanced Interdisciplinary Studies, Peking University, Beijing, China
4. Basic Medical Sciences, Capital Medical University, Beijing, China
5. Max Planck UCL Centre for Computational Psychiatry and Ageing Research, London, UK
6. Wellcome Centre for Human Neuroimaging, UCL, London, UK
7. Wellcome Centre for Integrative Neuroimaging, University of Oxford, Oxford, UK
8. Sainsbury Wellcome Centre for Neural Circuits and Behaviour, UCL, London, UK

# These authors contributed equally

* Correspondence: yunzhe.liu@bnu.edu.cn (Y.L.)

Abstract

Memories are spontaneously replayed during sleep, but capturing this process in the human brain has been challenging due to the dominance of slow, rhythmic background activity in sleep, which differs significantly from wakefulness. Each sleep stage, such as NREM and REM, has distinct rhythms, making it even harder for models trained on awake tasks to generalise and decode memory replay during sleep. To overcome this, we developed the Sleep Interpreter (SI), an artificial neural network. We first collected a large EEG dataset from 135 participants, recording brain activity during both awake tasks and overnight sleep. Using a Targeted Memory Reactivation (TMR) technique with 15 pairs of auditory cues and visual images, we tracked when specific memories were reactivated during sleep. The SI model was then trained separately for NREM and REM stages, using contrastive learning to align neural patterns between wakefulness and sleep while filtering out the background rhythms that previously hindered decoding. We also examined how specific sleep rhythms, such as slow oscillations and their coupling with spindles, influenced decoding performance. In a 15-way classification task during sleep, our model achieved a Top-1 accuracy of up to 40.05% on unseen subjects, surpassing all other available decoding models. Finally, we developed a real-time sleep decoding system by integrating an online automatic sleep staging process with the SI model for each sleep stage. This ability to decode brain activity during sleep opens new avenues for exploring the functional roles of sleep. By making our dataset and decoding system publicly available, we provide a valuable resource for advancing research into sleep, memory, and related disorders.

Keywords: sleep, decoding, targeted memory reactivation (TMR), contrastive learning, SO-spindle coupling

Introduction

Sleep is essential for cognition[1,2]. During sleep, the brain spontaneously replays past experiences, a process believed to underlie memory consolidation[3-5]. This replay is particularly active during non-rapid eye movement (NREM) sleep, where the brain shifts into slower, more synchronised activity patterns[6,7]. While memory replay during sleep has been demonstrated in animal models[8,9], capturing and decoding it in humans has remained a significant challenge.

Decoding sleep content is essential for understanding how sleep contributes to memory formation and cognitive function[10]. However, this has been particularly difficult due to the significant differences between brain activity during wakefulness and sleep[11,12]. Sleep is dominated by rhythmic patterns, such as slow oscillations (SOs) and spindles[13,14], which obscure the subtle neural signals linked to memory reactivation. In contrast, brain activity during wakefulness is more varied and less synchronised, making it hard for models trained on awake to generalise to sleep[10].

This challenge is especially pronounced in human neuroscience, where non-invasive methods like EEG, MEG, or fMRI are commonly used. In animal studies, researchers can directly record neuronal spikes, which exhibit similar representations during sleep and wakefulness[15,16]. In contrast, the non-invasive techniques used in human neuroscience, such as EEG, MEG, or fMRI, cannot isolate these spikes from the background activity. The large difference in background activity between wakefulness and sleep creates a 'generalisation gap,' making it difficult for awake-trained models to decode memory reactivation during sleep.

This study aims to solve these challenges by developing a real-time sleep decoding system capable of accurately decoding the content of sleep. It makes three key contributions: 1) the creation of the first publicly available EEG dataset specifically designed for sleep decoding, 2) the development of a novel sleep decoding model—Sleep Interpreter (SI)—which uses a neural contrastive learning approach to align neural representations between wakefulness and sleep, and 3) the introduction of the first real-time sleep content decoding system, combining an online automatic sleep staging process with the SI model for different sleep stages.

The first contribution is the large EEG dataset we collected from 135 participants, recording brain activity during both awake tasks and overnight sleep. This dataset

includes about 1,000 hours of sleep and 400 hours of awake task data, providing numerous instances of the same memories being reactivated during sleep. Using Targeted Memory Reactivation (TMR)[17-20], we paired auditory cues with visual images during wakefulness and replayed these cues during sleep. This allowed us to track which memories were reactivated during sleep. There were 15 pairs of associated memories. Crucially, this dataset enables us to train the SI model not only on awake data but also on memory reactivation during sleep.

The second contribution is the development of the SI model, which uses neural contrastive learning. Unlike typical contrastive learning[21,22], which operates on general features[23], our model works directly with neural signals. This approach aligns neural representations of the same memory content across wakefulness and sleep. Since the same memories are reactivated in both states, this alignment bridges the gap between the distinct rhythms of sleep and the more varied activity of wakefulness. The model is particularly effective because the slow background rhythms that dominate sleep are absent during wakefulness[13,14]. By linking memory reactivation during sleep to its corresponding activity during wakefulness, the SI model can focus on memory-related signals while minimising interference from background rhythms.

Sleep itself is not uniform—it is divided into NREM and rapid eye movement (REM) sleep[24,25]. NREM sleep is characterised by slow oscillations and spindles and is thought to be the primary stage for memory replay[17,26-28]. REM sleep, in contrast, is more desynchronised, similar to wakefulness, and is often associated with dreaming[29,30]. These differences complicate sleep decoding because each stage has distinct neural patterns and plays different roles in cognition[7,31,32], highlighting the need for stage-specific approaches.

To address this, we trained the SI model separately for NREM and REM stages. By training separate models for each stage, we can more accurately decode sleep content and capture the distinct characteristics of both stages. Moreover, we can focus on specific features known to be involved in memory consolidation. For instance, SO-spindle coupling during NREM sleep is closely linked to memory reactivation[3,33-37]. By incorporating these features, our SI model for NREM sleep achieved Top-1 accuracy of up to 40.05% on unseen subjects in a 15-way classification task, outperforming all existing EEG decoding methods[38-42].

The distinction between NREM and REM also highlights the need for an online automatic sleep staging system. This system ensures that the SI model can switch between NREM and REM modes in real-time, allowing for more precise tracking and analysis of memory reactivation during sleep.

By combining the SI model and the online automatic sleep staging, we developed a real-time sleep decoding system capable of tracking and decoding memory reactivation as it happens. This system opens up new possibilities for understanding the functional roles of sleep stages, particularly in how they contribute to memory consolidation, emotional processing, and cognitive function.

By making our dataset and real-time sleep decoding system publicly available, we provide a valuable resource for future research into sleep, memory, and related cognitive processes. This work represents a major advance in sleep research, enabling researchers to explore questions about the content and function of sleep that were previously impossible to study. With real-time decoding capabilities, we can now delve deeper into how different sleep stages support learning, memory, and cognition.

Results

Task design and dataset

Establishing effective real-time sleep decoding first necessitates **ground truth** data that indexes *which* memories get reactivated and *when*. For this we developed a variant of a TMR task[18] linked to simultaneous 64-channel scalp EEG recordings acquired during both awake tasks and overnight sleep (**Figure 1a**). Data, recorded at 500 Hz, were obtained from 135 healthy subjects in all (see Methods for details) such that the dataset encompassed 957.03 hours of sleep, and 362.29 hours of awake, recordings.

During wakefulness, subjects learned 15 pairs of associations between sounds and images, each representing a distinct concept (e.g., 'sheep', 'book'). This was followed by a 'functional localiser' session, wherein subjects were presented with a visual image and an audio clip in random order and tasked to think about their associated semantic content and to then judge whether the semantic content of the picture-sound pair matched. This session was conducted twice: once before sleep (pre-sleep) and once following overnight sleep (post-sleep). In the picture-sound pairing functional localiser sessions test subjects performed well behaviorally (**Figure 1b**), both pre-sleep (image → audio, 95.21±0.30%; audio → image, 96.04±0.29%), and post-sleep (image → audio, 95.17±0.35%; audio → image, 95.85±0.35%), with no significant difference between them (image→audio: $t(134)=0.21$, $P=0.836$; audio→image: $t(134)=0.84$, $P=0.403$, two-sided paired t-test).

Only correctly responded trials with good EEG data quality were epoched (from 0 to 0.8 seconds following the onset of the image or audio stimulus) for later neural decoding analysis (see Methods for details). Across subjects, for both image- and audio-evoked EEG, these signals were segmented into 2153.42±29.08 sample pairs data. Each sample was paired with a corresponding label from the 15 predetermined semantic concepts. For the identical semantic concept the neural data evoked by either the images or audio clips during wakefulness were later used to align with audio-induced neural representations elicited during sleep.

The mean sleep time across subjects was 7.09±0.06 hours (measured from getting into bed to waking up in the morning). During actual sleep we monitored the subjects' sleep stages in real-time including when subjects entered NREM 2/3 or REM sleep (determined by an experienced experimenter). We played an audio clip, randomly

chosen from the 15 pairs, with a 4-6 second gap between successive clips to avoid overlapping representations of different concepts. In offline analysis, two sleep experts independently annotated the sleep stage (Wake, N1, N2, N3, and REM) for each subject in 30-second segments covering the entire night. In cases of inconsistency, the experts engaged in a discussion to reach a consensus. On average, subjects spent 20.53±1.10% of their sleep time in Wake and N1 stages, 43.18±0.72% in N2 stage, 19.61±0.57% in N3 stage, and 16.68±0.50% in REM stage (**Figure 1c**). In relation to auditory cues delivered during the night, 34.88±0.87% were during the W/N1 stage, 54.25±0.84% during NREM 2/3 stage, and 10.87±0.43% during REM stage (**Figure 1d**). After excluding bad segments and auditory cues in the W/N1 stage, the nocturnal TMR-related EEG signals were segmented into 1383.30±40.97 samples for NREM 2/3 stage, and 307.66±15.48 samples for REM stage. Each sample was epoched from 0.3 to 2.3 seconds after the onset of the audio stimulus and paired with the corresponding semantic label.

In total, we obtained 290,712 pairs of image- and audio-evoked EEG samples from the awake tasks (pre- and post-sleep combined), and 186,746 NREM 2/3 and 41,534 REM audio-evoked EEG samples during sleep. These samples were annotated and curated into a dataset for sleep decoding. To the best of our knowledge, this is the first dataset that permits training neural decoders to analyse sleep content. This dataset is accessible from https://osf.io/x7453/ upon publication.

After curating the dataset, we developed the sleep decoding system in three key steps. First, we built an automatic sleep staging model to classify sleep stages without manual intervention. Next, we designed a sleep content decoder using contrastive learning, aligning neural representations between wakefulness and sleep. Finally, we incorporated sleep rhythms associated with memory consolidation, selecting specific features to further improve the model's decoding accuracy. The final zero-shot accuracy on unseen subjects reached up to 40.05% for 15-class classification during NREM sleep, outperforming all available EEG decoding models. We term our model 'Sleep Interpreter', or SI, which in Chinese pronunciation is '思', meaning 'thought'.

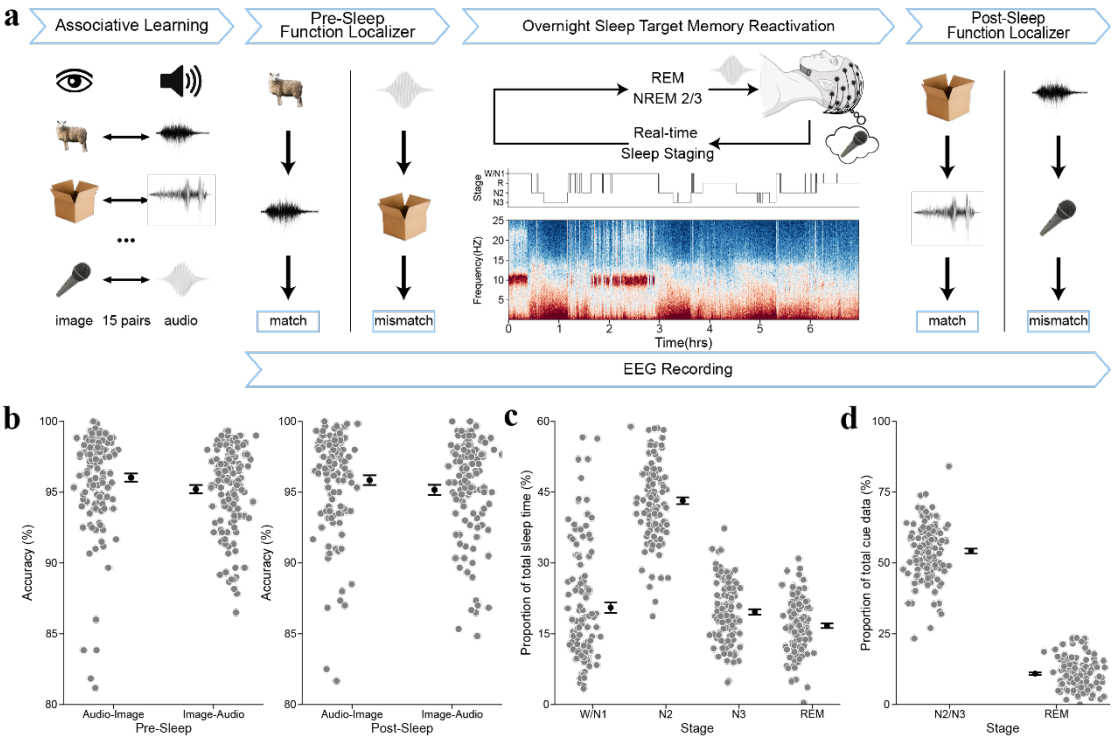


Figure 1: The experimental design for sleep decoding. (**a**) Before the experiment began, subjects were instructed to memorize 15 predefined picture-sound pairs, each sharing the same semantic meaning; for example, a picture labeled "sheep" was paired with the sound of a sheep 'baa'. After this associative learning, subjects undertook a functional localiser task. involving presentation of 1200 trials in total, each containing one picture and one sound selected from the 15 picture-sound sets (the picture and sound in one trial were not required to be paired). These were presented in either image→audio or audio→image order, and subjects were asked to determine whether the pairs corresponded. Subsequently, during overnight EEG recording, online, real-time sleep staging was performed. When subjects entered NREM 2/3 or REM sleep stage, auditory cues—randomly selected from the image-audio pairs—were played every 4-6 seconds. This step provided an approximation to a ground truth regarding which picture memory was likely to be activated and when. Following overnight sleep, subjects were again presented with the 1200 image-audio pairs. Whole-brain 64-channel EEG recordings were collected throughout the experiment, during both sleep and functional localiser tasks. More details can be found in Methods. (**b**) During pre- and post-sleep functional localiser stages, subjects were tasked to determine whether the presented image and audio cues were congruent. The behavioural accuracy is shown separately for image→audio and audio→image, as well as for pre-sleep and post-sleep periods. (**c**) The proportions of different sleep stages within the total collected sleep time, as determined by the offline, expert-based manual sleep staging. (**d**) The proportions of cued sleep data in N2/3 and REM, following offline, expert-based manual sleep staging. The gray dots represent the results for each subject, while the black error bars indicate the mean (M) and standard error of the mean (SE) for the corresponding conditions.

Automatic real-time sleep staging model based on ANN

Sleep staging is typically carried out offline via a human agent using a visual scoring protocol[43,44]. Although automated sleep staging algorithms, such as YASA[45], are available, their real-time capabilities are undetermined. A real-time capability is crucial for controlling timing of cued memory reactivation during sleep, including during distinct sleep states.

As we are interested in zero-shot generalization across subjects, we used expert-labelled sleep staging as a ground truth to train a real-time sleep staging model using an artificial neural network (ANN) (**Figure 2a**, see Methods for details). The model was trained on sleep EEG data from three channels (C3, EOG, EMG) across 123 subjects and tested on sleep data from the remaining 12 subjects (randomly chosen, roughly 10% of the training data). An example of a sleep staging result on testing subjects is shown in **Figure 2b**, alongside its ground truth.

Across 12 test subjects our SI-staging model achieved an average consistency with a ground truth of 90.38±0.96%. This was numerically higher, though not statistically significant, compared to the 90.05±0.76% consistency achieved by YASA[45], an offline staging algorithm ($t(11)=0.40$, $P=0.700$, two-sided paired t-test; **Figure 2c**). When YASA was used in real-time, its consistency with a ground truth was only 63.86±1.19%; our real-time staging model performed significantly better ($t(11)=14.55$, $P<0.001$, two-sided paired t-test).

Our staging model also showed good specificity in classifying each sleep stage (**Figure 2d**). The mean consistency with a ground truth was 88.8±4.05% for the N3 stage, 93.5±1.24% for the N2 stage, and 92.9±1.50% for REM sleep stage. For wakefulness and N1 stage, categorised together, its consistency was 83.8±3.07%. Similarly, the model predicted time for each sleep stage was comparable with the ground truth (expert-labelled) sleep duration (**Figure 2e**). The proportion of N2 or N3 stage in the total sleep time was 43.89±2.61% or 16.19±1.09%, with no significant difference compared to the model prediction ($t(11)=1.24$, $P=0.228$, two-sided paired t-test). The same was true for REM stage (16.83±1.17%, $t(11)=1.33$, $P=0.211$, two-sided paired t-test). These results demonstrate that our SI-staging model achieves both good consistency and specificity in sleep staging, when compared to offline expert-labelled stage results.

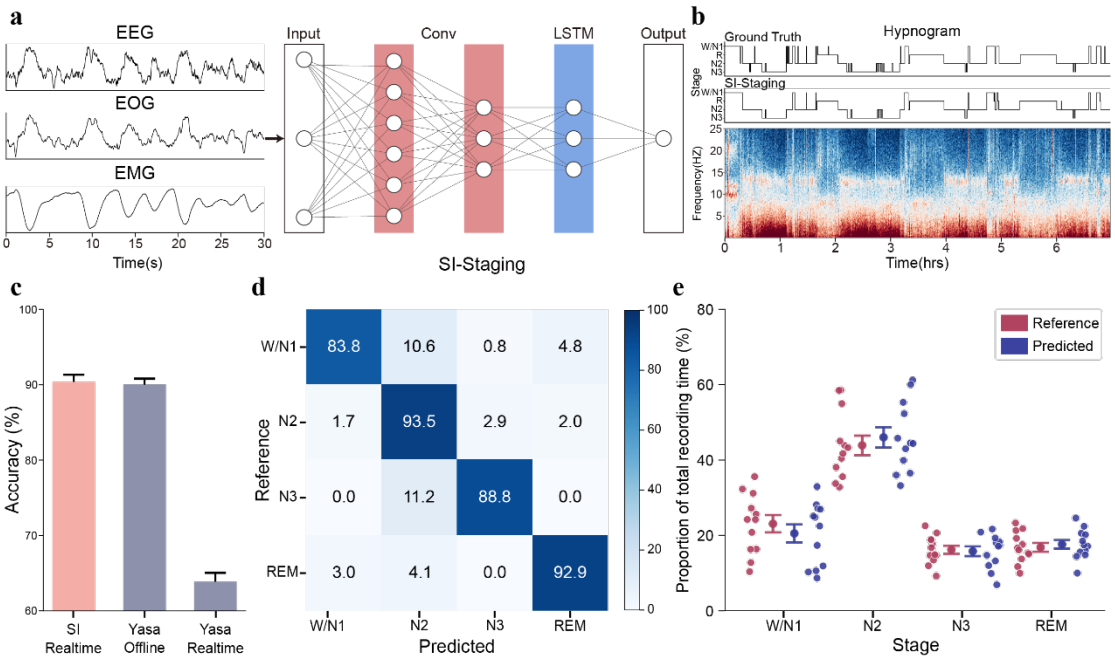


Figure 2: Real-time sleep staging. (**a**) Our Sleep Interpreter (SI) Staging model architecture comprises a CNN block, an LSTM network and a classification head. To maximise staging efficacy, the model utilises 30-second data segments from three channels (EEG-C3, EOG, EMG) as inputs. The model's loss function incorporates both cross-entropy and mean-square error terms, a combination that aims to address an imbalance in the distribution of data volume across each sleep stage, thereby enhancing the accuracy of sleep staging. (**b**) Ground-truth (expert-labeled) and model real-time predicted hypnogram for one test subject. The agreement between the ground-truth and model real-time predicted scoring is 94.41%. Below the hypnogram is the corresponding full-night temporal frequency profile of the central electroencephalogram (EEG) channel – C3, with warmer colours indicating higher power in these frequencies. (**c**) The mean consistency with the ground truth (expert-labeled) for different staging models. Our real-time SI-staging model is represented in light pink, while the two YASA-based models are shown in dark blue. For the YASA real-time model, calculations are based on the past five minutes of data due to the model's limitations. (**d**) Confusion matrix of staging accuracies. The diagonal elements represent the percentage of epochs correctly classified by the SI-staging model, whereas the off-diagonal elements show the percentage of epochs mislabeled by the SI-staging model. (**e**) Proportion of each sleep stage in the total sleep time: ground truth (reference, in red) vs. the SI-staging model predicted (in blue). The dots represent the results for each subject, while the error bars indicate the mean (M) and standard error of the mean (SE) for the corresponding conditions.

Generalisation ability of neural representations across awake and sleep states

To decode reactivation of objects during sleep, we need access to their neural representation, where the latter are typically derived from training during an awake task. In this framework, sleep decoding relies on an assumption of a generalisation ability of object representations across awake and sleep domains. In the current task, it is possible to compare event-related potentials (ERPs) elicited by the same semantic object during awake and sleep states. Qualitatively, there is a striking domain difference between awake and sleep ERPs (**Figure 3a**), with sleep ERPs (audio-evoked) much more synchronised across channels compared to both image- and audio-evoked ERPs acquired during the awake task (for ERP differences between NREM and REM sleep, see also **Supplementary Figure 1**). There is also a discernible difference between audio-evoked and image-evoked ERPs, although to a lesser degree.

With these considerations in mind, we first used a linear model, Lasso GLM, to quantify generalisability between awake and sleep domains (**Figure 3b**) before building the decoder. For each subject, we trained object decoders based on image-evoked (from stimulus onset, -0.2s to 0.8s) or audio-evoked (-0.2s to 0.8s) ERPs in the awake state and tested their performance during sleep. Similarly, we trained neural decoders based on audio-evoked ERPs acquired during sleep (0.3s to 2.3s) and tested these during awake states. Note within a TMR paradigm we know precisely which object should be activated, and when. We implemented this approach separately for NREM and REM sleep stages, where the specific time window was set so as to be consistent with later SI model training, itself selected based on an initial search over the training duration for better sleep decoding accuracy (see details in Methods).

The model was trained and tested on the same subject, using training and testing data from different domains (e.g., sleep vs. awake). The generalisation abilities from awake image- or audio-based neural decoders to NREM 2/3 sleep stage were 8.80±0.14% and 8.82±0.12% respectively, both above chance (chance level, 1/15=6.67%, all $t(134)>14$, all $P<0.001$, one-sample t-test), with no significant differences between decoder types ($t(134)=0.48$, $P=0.630$, two-sided paired t-test). This was also true for decoding an REM sleep stage (image-based: 11.71±0.51%, audio-based: 12.10±0.64%, compared to the chance level: 1/15=6.67%, all $t(131)>8$, all $P<0.001$, one-sample t-test, note three subjects are deleted due to no available data in REM sleep stage was recorded), but where the awake audio-based decoder performed better than awake image-based

decoder ($t(131)=2.14$, $P=0.035$, two-sided paired t-test).

Conversely, the generalisation ability from NREM sleep-trained neural decoders to awake data were 8.18±0.02% for awake-image and 8.08±0.02% for awake-audio, both above chance level (chance level, 1/15=6.67%, all $t(134)>75$, all $P<0.001$, one-sample t-test), with an NREM decoder performing better on awake image data than on awake audio data ($t(134)=8.22$, $P<0.001$, two-sided paired t-test). Similarly, for REM sleep-trained neural decoders to awake data (image-based: 8.16±0.02%, audio-based: 8.06±0.02%, compared to the chance level: 1/15=6.67%, all $t(131)>80$, all $P<0.001$, one-sample t-test), and REM decoder performed better on awake image data than on awake audio data ($t(131)=7.67$, $P<0.001$, two-sided paired t-test).

These results demonstrate a common neural representation of the same concept across both wakefulness and sleep, despite significant differences in their ERP patterns. This suggests that aligning neural representations between these states is possible.

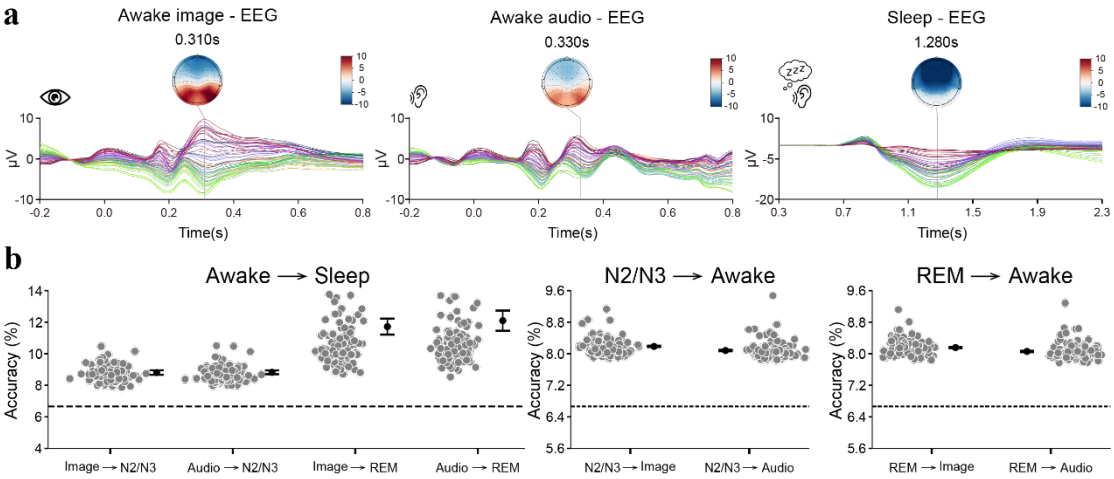


Figure 3: ERPs in awake and sleep domains and their across-domain generalisation ability. (**a**) The averaged ERPs and topographic maps evoked by image stimuli during wakefulness, audio stimuli during wakefulness, and TMR-related audio stimuli during NREM2/3 and REM sleep respectively. The zeros on the horizontal coordinates correspond to the stimulus onset times for different data domains. The averaged ERPs of awake data were computed within a time window from 0.2 seconds before stimulus onset to 0.8 seconds after stimulus onset. The averaged ERPs of sleep data were computed within a window from 0.3 seconds after stimulus onset to 2.3 seconds after stimulus onset. The topographic maps show time points with large neural responses in ERPs, where the same amplitude range (-10 ~ 10 μV) was used to facilitate comparison of activations. (**b**) Performance on sleep decoding based on the decoders trained on awake data, and vice versa. The decoder used here is a simple linear model: the Lasso GLM, designed to probe whether there are common features between the two different data domains. The grey dots represent the results for each subject.

Decoding object representation during sleep with neural contrastive learning

Having identified overlapping representations of object reactivation during awake and sleep states, we went on to develop a cross-domain sleep decoding model, which we term SI-Multi Domain (SI-MD). Using contrastive learning[21], the SI-MD aligns an awake representation of an object with its sleep representation, which we term 'neural contrastive learning' (**Figure 4a**, see details in Methods). The critical computations of the SI-MD model involve two parts: the model architecture and the inclusion of neural contrastive learning loss to align representations across awake and sleep domains.

The model architecture of SI-MD consists of three encoders, one for each data domain: 'awake image', 'awake audio', and 'sleep audio' (**Figure 4a**). For each encoder, the stimuli-evoked whole-brain EEG activity is passed through a one-dimensional (1D) convolution module[46] and a linear module. Additionally, to capture global synchronized brain activation patterns during sleep, the sleep encoder goes through a transformer module[47]. Finally, to decode the identity of an object, each processed data stream goes through a classification layer. Each pair of data streams also share an encoder for calculating neural contrastive loss (see details in Methods).

The typical contrastive learning loss is calculated as:

$$L_{contra} = -\sum_{i\in\{1,\dots,|\mathcal{B}|\}} \frac{1}{|A(i)|} \sum_{k\in A(i)} log \frac{e^{\langle z_i, h_k\rangle}}{\sum_{j\in\{1,\dots,|\mathcal{B}|\}} e^{\langle z_i, h_j\rangle}} \quad (1)$$

where $A(i) = \{k|k \in \{1,\dots,|\mathcal{B}|\}, y_k = y_i\}$, $\langle\cdot,\cdot\rangle$ is the inner product, B is the batch dataset with $|\mathcal{B}|$ the size of batches and y represents the semantic label. Typically, z and h represent data from different domains. In neural decoding[48], z could represent latent embeddings from neural activity, while h could represent processed sound from the wav2vec model[49].

By changing z and h to represent latent embeddings from different kinds of EEG signals (e.g., sleep-TMR-related and image-evoked EEG signals), we derived a neural contrastive learning loss function.

$$L_{contra} = L_{contra}^{tmr2img} + L_{contra}^{tmr2aud} + L_{contra}^{img2aud} \quad (2)$$

In essence, this method aligns the neural representation of the same object across domains, facilitating sleep decoding, and where our expectation was better decoding accuracy compared to models trained on sleep data alone. We then went on to compare

the performance of our SI-MD model with other 'control' models. All sleep decoding models (SI-MD and other control models) were trained on 123 subjects and tested on the remaining 12 subjects. This was implemented separately for NREM 2/3 and REM sleep stage. Note we are particularly interested in the zero-shot, cross-subject, generalisation ability of the sleep decoder, given our aim is to derive a model that can decode sleep in subjects outside the framework of current study.

To evaluate the impact of neural contrastive learning loss, we compared the performance of the SI-MD model with a model that shares the same architecture but lacks this cross-domain contrastive learning, and only trained on the sleep data domain, termed SI-Single Domain (SI-SD). The SI-MD model achieved zero-shot accuracies of 24.12±0.70% and 19.81±1.59% at NREM and REM stages respectively, significantly higher than the SI-SD model's zero-shot accuracies of 21.07±0.43% and 16.05±1.59% (**Figure 4b**, NREM2/3: $t(11)=6.48$, $P<0.001$; REM: $t(11)=3.29$, $P=0.007$, two-sided paired *t*-test). Furthermore, after fine-tuning, the decoding accuracy of the SI-MD model during sleep for the same subjects improved to 30.81±0.65% for NREM 2/3 stage and 31.65±1.96% for REM stage.

To evaluate model architectures, we implemented five different types of sleep decoding models: our SI-MD model, SI-SD model, a linear decoder model (Lasso GLM), and three ANN models (LSTM-based, CNN-based, and Transformer-based). Details on model construction are available in the 'Decoding model comparison and ablation experiment' section in Methods.

We found that the zero-shot decoding results of all models significantly surpassed chance level (**Figure 4b**, NREM2/3: all $t(11) > 11$, all $P < 0.001$; REM: all $t(11) > 6.30$, all $P < 0.001$, two-sided one-sample *t*-test). After fine-tuning on sleep task data, the decoding performance of our SI-MD model, as well as the CNN-based and Transformer-based models, significantly exceeded those trained on sleep data alone (NREM2/3: all $t(11) > 3.30$, all $P < 0.01$; REM: all $t(11) > 3.50$, all $P < 0.01$, two-sided paired *t*-test), whereas the Lasso GLM model performed significantly worse ($t(11) = -7.47$, $P < 0.001$, two-sided paired *t*-test). Moreover, our SI-MD model significantly outperformed the other four models in terms of both zero-shot accuracy (NREM2/3: all $t(11) > 8$, all $P < 0.001$; REM: all $t(11) > 4$, all $P < 0.01$, two-sided paired *t*-test) and fine-tuning accuracy (NREM2/3: all $t(11) > 11$, all $P < 0.001$; REM: all $t(11) > 6.30$, all $P < 0.001$, two-sided paired *t*-test). This indicates wide applicability of our SI-MD

model, including the potential to integrate awake task data for decoding sleep in unseen subjects.

We also compared the performance of SI-MD model to widely used decoding models, such as EEG-Net[38], EEG-Conformer[39], BrainBERT[40], BIOT[41] and LaBram[42]. None of those models achieved zero-shot, cross-subject, decoding accuracy above 20% when applied to NREM 2/3 stage and 15% in REM stage (**Supplementary Table 1**). Further control analysis revealed that sufficient data benefits training the SI-MD model, with decoding accuracy increasing as the training data size increases (**Supplementary Figure 2**). Training on the broadband data signal was more effective than training on specific frequency bands (i.e., delta, theta, alpha, beta, gamma) for both NREM 2/3 and REM sleep stages (**Supplementary Figure 3**).

Together, these results suggest that our SI-MD model, incorporating neural contrastive learning that aligns neural representations across awake and sleep domains, outperforms an SI-SD model trained on sleep data alone. Crucially, it also outperforms other widely used ANN architectures and EEG-based decoding models. This performance advantage was true for both NREM 2/3 and REM sleep stages.

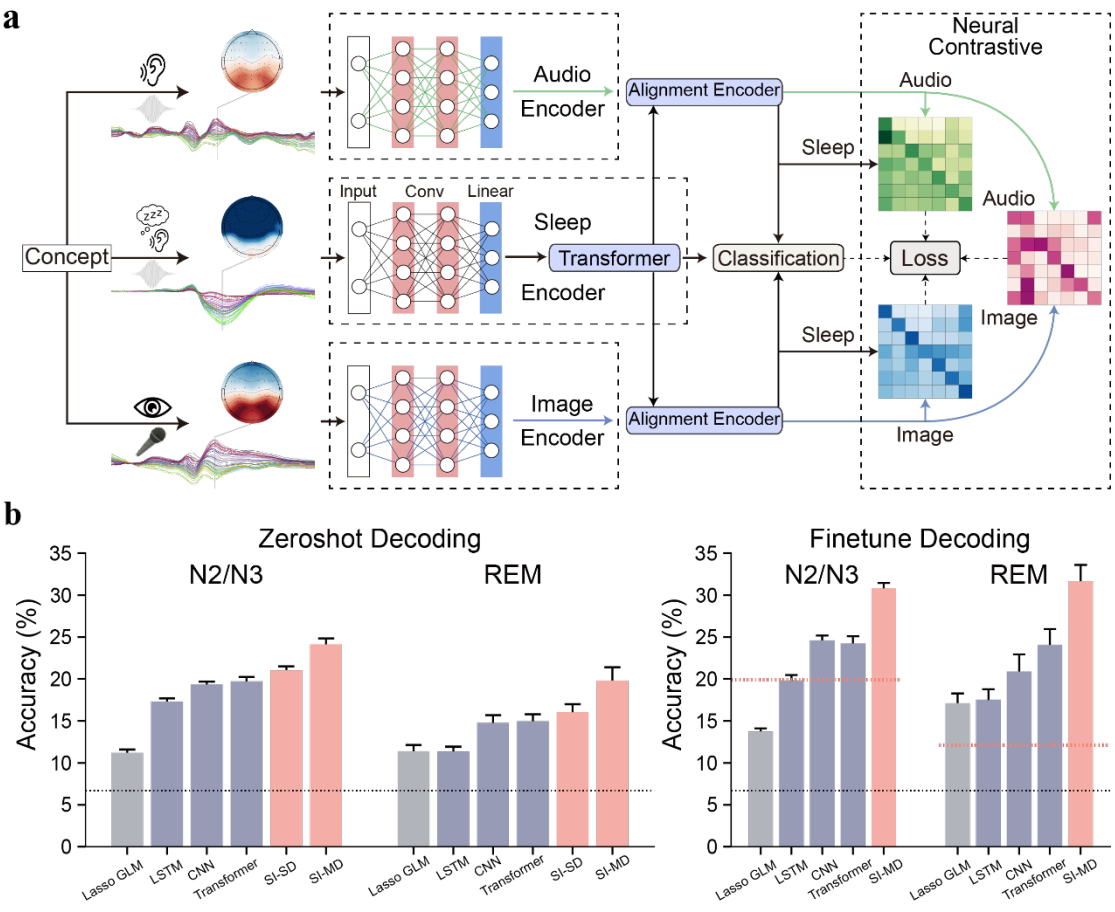


Figure 4: Zero-shot cross-subject sleep decoding with neural contrastive learning. (**a**) Overview of Sleep Interpreter – Multi Domain (SI-MD) model. The SI-MD model architecture consists of neural encoders for both sleep and awake datasets, as well as the alignment encoders. Contrastive loss is used to regularize the latent space, encouraging similarity among features adhering to the same semantic class but coming from different domains (e.g. $D_i^{img}, D_i^{aud}, D_i^{tmr}$) of the same subject, i. The awake EEG dataset and sleep EEG dataset are perceived as disparate domains due to the substantial differences in their neural patterns. By utilizing the alignment encoder approach and integrating neural contrastive learning, the model effectively aligns awake and sleep neural representations across subjects during training. (**b**) Sleep decoding performance. The zero-shot, cross-subject decoding performance of all tested models is shown on the left, and the performance of models fine-tuned was shown on the right, separately for the NREM 2/3 stage and REM sleep stages. The Lasso GLM model is represented by gray bars, widely used ANN models by dark blue bars, and the SI-SD and SI-MD models by light pink bars. The black dashed line represents the chance level accuracy for the decoding task, which is 6.67%. The orange dashed lines represent the average result of the SI-SD model trained and tested on individual subjects across the twelve test subjects, which is 19.87 ± 0.88% for the NREM 2/3 stage and 12.13 ± 1.01% for the REM stage. The bar charts display the mean performance (M) with error bars representing the standard error of the mean (SE) across the testing subjects.

Sleep decoding performance as a function of cue time in SO phases

Previous studies show that, during NREM sleep, the strength of spontaneous memory reactivation varies across different phases of a Slow Oscillation (SO) event[50]. Reactivation is stronger in the "UP" state, defined as the peak of the SO, than in the "DOWN" state, defined as the trough of the SO[6,51]. There is also coupling between spindles and the SO phase, predominantly in UP and transition to UP states, that relates to memory consolidation during sleep[17,33,52,53]. Based on this we explored the influence of cue time in relation to SO phases (and coupling with spindles) on decoding performance during NREM 2/3 sleep. In particular, we were interested in whether a coupling might optimize further model training to improve its ability to learn semantic information during sleep, thereby enhancing decoding performance.

We classified data samples (cued auditory stimuli) in the NREM 2/3 stage into eight non-overlapping categories based on phase of detected SOs (**Figure 5a**), where the latter account for about 40% of the total data samples in the NREM 2/3 stage (see "Slow Oscillation Phases" section in Methods for details). We used the trained SI-SD model to investigate inherent decodability of sleep data, excluding the influence of awake data. We found for the "1st Up" and "2nd Up" categories, decoding results were 23.73±2.06% and 20.68±1.00% respectively, significantly higher than for the "Down" category, which was 16.21±1.66% ("1st Up": $t(11)=3.34$, $P=0.007$; "2nd Up": $t(11)=3.22$, $P=0.008$, two-sided paired t-test). This accords with previous findings that decoding performance is better when the cue presentation is in the Up state of an SO, compared to the Down state[6]. Interestingly, the "Trans 1st UP" category, with a decoding accuracy of 36.17±1.77%, outperformed the Up state categories ("1st Up": $t(11)=4.50$, $P<0.001$; "2nd Up": $t(11)=6.49$, $P<0.001$, two-sided paired t-test). The "Trans 1st Up" category also showed significantly higher decoding accuracy compared to all other categories ("Trans Down", "Down", "Trans 2nd Up", "Post SO", "Other": all $t(11)>4.60$, all $P<0.001$, two-sided paired t-test).

We repeated the generalisation experiments to determine whether the current sleep decoding results are also consistent with cross-domain semantic decoding. We trained Lasso GLM models on different awake domains (awake image-evoked and awake audio-evoked datasets) and evaluated these at different SO phases during NREM 2/3 sleep. Similar sleep decoding results were obtained for classifiers trained on awake image data (**Supplementary Figure 4a**) and awake audio data (**Supplementary**

Figure 4b). Taking the differences in sample sizes in SO phases into consideration, we extracted equal amounts of data from each SO category in the training dataset to train the SI-SD model, then decoded the test dataset. The "Trans 1st Up" was the sole category that showed results consistently above baseline across all nine categories (**Figure 5c**), demonstrating a strong transfer generalisation ability.

We next explored whether SO coupling with sleep spindles influences decoding ability. Temporal frequency analysis of the Fz channel waveforms (from -1s to 4s, based on cue onset) revealed significant coupling of the spindle (12-16 Hz) frequency band with SO phases, especially when the cue was in the "Trans 1st Up" stage of the SO (**Supplementary Figure 5**). Reapplying the trained SI-SD sleep model for decoding, and using the overall decoding accuracy of the corresponding SO phase as baseline for each subcategory (**Figure 5d**; see also **Supplementary Figure 6** for the distribution of spindle coupling frequency and mean magnitude as a function of SO phases), we found that spindle coupling improved decoding accuracy across categories. Thus, the presence of spindle coupling significantly improved decoding accuracy for the "Trans 1st Up" and "Other" categories, as well as for test subjects overall ("Trans 1st Up": $t(11)=2.31$, $P=0.041$; "Other": $t(11)=6.63$, $P<0.001$; "Whole": $t(11)=8.35$, $P<0.001$,two-sided paired *t*-test).

Based on these findings, we retrained the SI-MD model, but now excluding training data from the 'DOWN', 'Post SO' stages, and also 'Other' stage where there were no detectable spindle events. Consequently, the convolutional network module in the SI-MD model architecture was adjusted to account for the reduced size of this 'cleaned' training data (see **Methods** for details). We obtained a retrained SI-MD model that showed a further decoding improvement when confined to the 'Trans 1st Up' part of the decoding, achieving 40.05±1.38% compared to 37.05±1.48% for the original SI-MD model ($t(11)=3.85$, $P=0.003$, two-sided paired *t*-test).

Finally, we built a closed-loop real-time sleep decoding system (**Supplementary Video1**) wherein we combined all the above results and added a diffusion module at the end (see **Methods** for details). This generates an image based on the highest classification probability as the final output and provides real-time sleep staging and decoding of sleep content through generated images derived from neural activity (**Supplementary Figure7**). Note this arrangement allows for closed-loop intervention on specific memory content, opening entirely new directions for modulating learning

and memory processes during sleep.

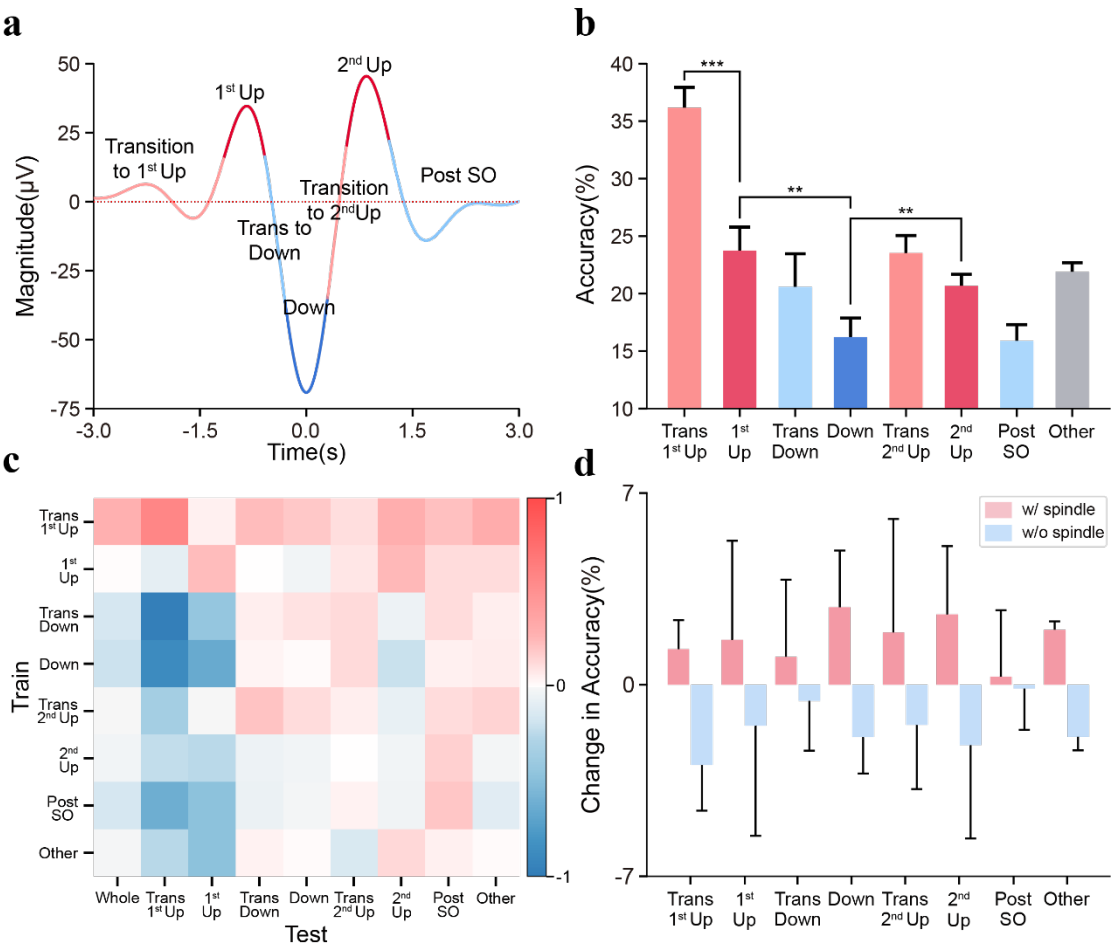


Figure 5: Sleep decoding performance as a function of slow oscillation phases and spindle coupling. (**a**) Slow oscillation (SO) phases. During NREM 2/3 sleep, we divided the phase of a SO event into eight parts (color-coded), based on the onset of audio stimuli. All other cue times were classified as "Other" category. The time represented on the x-axis corresponds to a 6-second data length, with the trough of the SO event as the zero point. (**b**) Sleep decoding performance by SO phase. The decoding performance of different SO phases during NREM 2/3 sleep using the SI-SD model is shown with average performance and standard error across subjects. The decoding results for the "1st Up" and "2nd Up" categories were significantly higher than for the "Down" category, while the result for the "Trans 1st Up" was significantly higher than those for the "1st Up" and "2nd Up" categories. The bar chart colors for all categories correspond to the colors in the diagram of panel (a), with the results for the "Other" category shown in grey. (**c**) Generalisation ability of trained sleep decoders across different SO stages. The same sleep decoder model was re-trained on an equivalent amount of data, separately for each SO stage during NREM 2/3 sleep, and compared to a baseline model, which was re-trained with an equivalent amount of training data randomly sampled from the SO phases, rather than within one of the specific categories. The change in decoding accuracy is color-coded: red indicates accuracy above the baseline model, and blue indicates accuracy below the baseline model. (**d**) Sleep decoding performance with vs. without spindle coupling. Within each SO stage, we further compared the change in decoding performance with and without spindle coupling using the trained sleep decoder. This comparison was based on the presence or absence of a spindle event in the testing trials. The decoding performance for the "with spindle" group is shown in pink, while that for the "without spindle" group in blue.

Discussion

We introduce a real-time, closed-loop sleep decoding system called the Sleep Interpreter (SI), which consists of two key components: a sleep staging model (SI-staging) and a model for decoding sleep content (SI-MD). The SI system is built on a comprehensive dataset collected from 135 participants, combining both awake tasks and overnight sleep using a TMR task with EEG recordings and expert-labelled sleep stages. Upon publication, both the annotated dataset and the code for the real-time system will be made publicly available.

Both the SI-staging and SI-MD involve the use of deep neural networks. SI-staging is developed through supervised learning, with expert-labelled sleep stages serving as an approximation to a ground truth. By contrast, SI-MD is built with the addition of a neural contrastive learning module, which aligns neural representations derived from both sleep and awake states. Critically, we show that in decoding sleep content our model outperforms other ANN models available for EEG-based decoding.

Contrastive learning is typically a self-supervised learning approach that trains models to distinguish between similar and dissimilar data points, pulling semantically similar samples closer to a latent space while pushing dissimilar ones farther apart[23,54,55]. It is widely used in multi-modal models[21,48] and here we use it in a supervised manner[22]—aligning the neural representation of the same concepts across domains (awake and sleep) while separating different concepts. While this allows us to achieve good decoding for the 15 predetermined concepts it also restricts decoding to this 'closed' set of categories. One consequence is that the model would struggle to generalise to concepts beyond these 15 categories. Future work should explore the possibility of using contrastive learning in a self-supervised way[55] or avail of other self-supervised learning approaches, such as masked modelling[40,42], to train a universal sleep decoder for interpreting open-set semantic labels. This may require collecting a new dataset with large numbers of words played to the subjects, similar to the 'THINGS' dataset[56-58] but, in this instance, played to subjects during sleep.

Previous studies found stronger spontaneous as well as cued memory reactivation in the "Up" state than in the "Down" state of SO events, especially during NREM 2/3 sleep[6,51]. Our results align with this literature, but intriguingly, we also found that trials within the "Trans 1st Up" phase of SO events provide for the best decoding performance

compared to other stages of SO, outperforming the “Up” state. This finding is notable because the "Trans 1st Up" phase also has the highest ratio of SO-spindle coupling as well as the highest mean magnitude of neural activity (Supplementary Figure 6). Our observations here mirrors previous sleep findings, underscoring the importance of SO-spindle coupling[17,33,52,53,59]. When re-training the SI-MD model by focusing more on epochs containing SO-spindle coupling (i.e., “Up” and “Transition to Up” states), we obtained additional improvements in sleep decoding accuracy, averaging up to 40.05% across test subjects. It is interesting to speculate that semantic information may be more easily induced during a period of SO-spindle coupling, and if this is the case, we would expect self-supervised learning during this period to extract features likely to be foundational for semantic decoding.

We segregated SI-MD training for NREM and REM sleep on the basis of significant differences in background neural activity, but also on account of their putative distinct cognitive roles[31,60,61]. While SO-spindle coupling, sometimes accompanied by hippocampal ripples (or replay), is characteristic of NREM sleep and thought crucial for memory consolidation[17,26-28], brain activity during REM sleep mirrors elements of preceding awake states, characterized by extensive neural desynchronization and linked with dream occurrence[29,30]. Dreaming is typically studied using subjective report[62,63] and our SI-MD model for REM sleep provides for a possible alternative reporting, although currently constrained to the 15 pre-determined semantic categories. Future work might extend the decoding content, possibly with self-supervised learning in open sets[64,65], potentially investigating the alignment between decoded content acquired during REM sleep and awake-time subjective reports.

In conclusion, we introduce a well-annotated large (135 subjects) EEG dataset for sleep semantic decoding. Leveraging this dataset, we develop a real-time decoding system, comprising both online sleep staging and sleep content decoding. By including neural contrastive learning, we show our decoding model outperforms all baseline and comparison models, establishing it as the state-of-the-art model for decoding content during both NREM and REM sleep. In so doing, we also detail the impact on decoding performance of SO-spindle coupling during NREM 2/3 sleep, validate existing findings, discover new features and incorporated these to further optimise the model for better decoding performance.

Our approach paves the way for developing a universal sleep decoder, enabling more

precise and effective sleep editing and modulation, including decoding of dreams during REM sleep but also targeting of pathological memories that form the content of troubling dreams and nightmares.

Methods

Participants

We tested a total of 149 participants, with 14 participants excluded for a range of reasons: two due to excessive body movements and poor data quality, four because they did not fall asleep, four because of electrode issues (falling off or drifting), two due to snoring, one was interrupted by a wakefulness experiment, and one failed to complete the full wakefulness experiment. All participants were healthy nonsmokers with no color blindness or color processing deficiency, normal hearing, and no allergies to alcohol or scrubs. They maintained normal wake-sleep rhythms, had not permed or colored their hair in the last month, did not suffer from insomnia, problematic snoring, or nighttime sleep disturbances. All participants gave written consent prior to the experiment. The study was approved by the Chinese Institute for Brain Research, Beijing (CIBR) and conducted in accordance with the Declaration of Helsinki.

Experiment Design

Our experiment paradigm comprised three stages: a pre-sleep function localiser stage, an overnight sleep target memory reactivation stage and a post-sleep function localiser stage (**Figure 1a**). First, we selected 15 semantic classes: alarm, apple, ball, book, box, chair, kiwi, microphone, motorcycle, pepper, sheep, shoes, strawberry, tomato, watch. Then, we used corresponding pictures and sounds to form a set of image stimuli (including 15 different pictures) and a set of audio stimuli (including 15 different sounds).

In the pre-sleep function localiser stage, subjects were instructed to perform two tasks: the image-audio task and the audio-image task. In the image-audio task, subjects were presented with 600 trials, with each trial containing one image stimulus and one audio stimulus, in image-audio order. We randomly selected one image stimulus and one audio stimulus from the set of image stimuli and the set of audio stimuli, respectively, where the image stimulus and the audio stimulus could be inconsistent (i.e., derived from different semantic classes). In total, we had 210 inconsistent trials and 390 consistent trials. The purpose of this design is solely to ensure that participants formed veridical and stable stimulus pair memories.

For the image-audio task, subjects were presented with the following study pipeline: a blank screen for 0.6 to 0.9 seconds (inter-trial interval), followed by a cross centered

on the screen for 0.3 seconds, then an image stimulus for 0.8 seconds. After the image, a blank screen was shown for 0.3 seconds, followed by an audio stimulus for 0.5 seconds, and finally, another blank screen for 0.3 seconds. Then, subjects were tasked to indicate whether the presented image stimulus and audio stimulus were matched (i.e., coming from the same semantic class). Finally, we presented behavioral feedback for 0.2s. In total, the image-audio task took about 48 minutes.

The setup of the audio-image task was identical to that of the image-audio task, with the sole difference being that we presented image-audio pairs now in an audio-image order. As in the post-sleep function localiser stage, subjects were required to execute the same tasks in the pre-sleep function localiser stage.

In the overnight sleep target memory reactivation stage, we utilized a near real-time application of the YASA model which refers to use the past five minutes' data in staging. We consider that a subject was in either NREM2/3 or REM stage only when the total probability of NREM2/3 or REM stage in the model output results exceeded 70%. As the staging result subject reached the N2/3 stage of NREM sleep, the subject was presented with audio stimuli alone evert 4-6 seconds (the same as those used during wakefulness, lasting 0.5s), where these were randomly selected from the set of audio stimuli.

Preprocessing

Sleep Staging Data

We followed the standard pipeline of YASA[45]. Accordingly, we provided a central electrode C3 referenced to the mastoid M2 (i.e., C3-M2) as the EEG channel, the right EOG channel REOG referenced to the mastoid M2 (i.e., REOG-M2) as the EOG channel, and the right EMG channel REMG referenced to the reference EMG channel EMGREF (i.e., REMG-EMGREF) as the EMG channel. For each sleep staging dataset, we have expert-labeled manual staging results, which divide the data into 30-second stages. Based on these manual staging results, we segmented the data into 30-second intervals and augmented it with a sliding window of 3-second length. For data segments containing multiple labels, we assigned the label that accounted for the majority of the segment as the corresponding segment label. Finally, we used the 'scale' function of the 'preprocessing' module in the python package 'sklearn' for normalization.

Sleep Decoding Data

We used the python package MNE[66] to epoch the sleep task-related EEG signals from 0.3s to 2.3s for sleep EEG signals according to the onset of the stimuli cue, dropping noisy epochs automatically. Additionally, for sleep EEG signals, only data in the manually labeled NREM2/3 and REM stages were taken out. It should be noted that three subjects had no usable REM data due to the absence of stimuli within their short REM sleep. Each channel of the data segments was then re-referenced by subtracting out the mean signal over all EEG channels (average re-referencing). For data segments in REM stage of sleep, these were filtered within the frequency range of 0.1-40Hz using 'filtfilt' function in python package 'signal' (bandpass, order=4), while line noise at 50Hz (and its harmonics) of data segments in REM stage of sleep were removed using 'iirnotch' function in python package 'signal'. Finally, the preprocessed data segments of NREM2/3 and REM stages were all resampled to 100Hz using the 'decimate' function in python package 'signal'.

The next preprocessing for awake EEG signals was performed using the python package MNE[66]. The awake EEG signals were filtered within the frequency range of 0.1-40Hz (bandpass, order=4). Then, line noise at 50Hz and its harmonics were removed. Each channel was re-referenced by subtracting out the mean signal over all EEG channels (average re-referencing). We then applied Independent Component Analysis (ICA) that removed artifact components automatically[67]. The preprocessed EEG signals were further resampled to 100Hz. Finally, we epoched the EEG signals from -0.2s to 0.8s according to the onset of the stimuli cue (e.g. image, audio), and dropped noisy epochs automatically.

Next, in python we applied the 'robust_scale' function from the 'preprocessing' module of the 'sklearn' package to normalize all the preprocessed awake and sleep data segments, clamping the normalized values greater than 10 and less than -10. For the awake training data, we randomly removed half of the data from the same domain with the same labels for each individual, averaged the remaining data to obtain a new awake data segment, and generated a new awake dataset of the same size as the corresponding sleep dataset for model input.

Real-time Sleep Staging

Following our sleep staging data preprocessing, we formulated the EEG signals as $X \in$

$\mathbb{R}^{C\times T}$, where C is the number of EEG channels and T is the total timestamps. The associated label was denoted as $y \in Y$, where Y represents the set of pre-determined stages. In summary, for each subject, the total dataset was D_s, $s \in S$, where S is the set of subjects and each sub-dataset comprised paired EEG-concept data $(\langle X, y\rangle)$. Therefore, the goal was to decode the corresponding stage label y from the preprocessed EEG signals.

For optimal staging efficacy, the model harnesses data from three channels (EEG, EOG, EMG) as inputs. For each channel, the input data has a sampling rate of 500 Hz and a total length of 30 seconds, which corresponds to $X \in \mathbb{R}^{3\times 15000}$. To meet the requirements of real-time sleep staging, we set the input data with a sliding window applied every 3 seconds for subsequent staging. For the staging results of the real-time sleep model, we considered the initial 30-second staging results as the onset of the initial sleep state. Subsequently, based on this onset, we assigned the staging results after each sliding window to the corresponding sleep state within the 3-second time period. We trained the model in a supervised manner, predicting four sleep stages: W/N1, N2, N3, and REM. The corresponding labeling values for these stages are 0, 1, 2, and 3, respectively.

Our SI-Staging model consists of two main parts: the CNN module and the LSTM module. Through the tandem processing of the CNN and LSTM modules, followed by a final classification layer, we obtained the predicted label y'. The loss term of the model encompasses cross-entropy error and mean square error calculated between the ground truth label y and predicted label y':

$$L_{cls} = L_{cee} + L_{mse} \tag{3}$$

When calculating the cross-entropy error, the ground truth label and predicted label values are converted into one-hot vectors. For the mean square error calculation, they are treated as simple numbers.

The staging model was trained up to 200 epochs by default using Adam with a learning rate of $3 \cdot 10^{-4}$ and a batch size of 256 on one NVIDIA V100 GPU with 32GB of memory.

Generalisation task

For our generalisation task, we formulated the EEG signals after sleep decoding data

preprocessing, as $X \in \mathbb{R}^{T \times C}$, where C is the number of EEG channels and T is the total timestamps. For tasks across data domains (from sleep to awake or from awake to sleep), considering the duration limit of awake EEG signals is 0.8s, we consequently segmented the sleep data into 13 data segments of 0.8s each at 0.1-second intervals (0-0.8 s, 0.1-0.9 s, … ,1.2-2.0 s). This segmentation allowed us to conduct generalization experiments on data with unbalanced timestamp lengths and comprehensively evaluate the final results.

In each time window, we sequentially extracted the sample data of all channels at each single time point in the training data and use a Lasso GLM model for training. For the fifteen classification labels, model was trained for each label and applied to the test data at the corresponding time point. We normalized the test results as probabilities of all categories using the 'softmax' function, applied the 'argmax' function to obtain the model's classification result, and calculated the final accuracy accordingly. We took the maximum value of the 80 test accuracies for each time window and averaged them across all time windows to obtain the final result.

Real-time Sleep Decoding Model

We formulated the awake or sleep EEG signals after sleep decoding data preprocessing as $X \in \mathbb{R}^{T \times C}$, where C is the number of EEG channels and T is the total timestamps. The associated semantic label was denoted as $y \in Y$, where Y represents the set of 15 pre-determined semantic concepts (each paired with both an audio stimulus and an image stimulus), stored in the form of one-hot encoding. In summary, for each subject, the total dataset contained three sub-datasets, i.e., $D_s = \langle D_s^{img}, D_s^{aud}, D_s^{tmr} \rangle, s \in S$, where S is the set of subjects and each sub-dataset comprises paired EEG-concept data ($\langle X, y \rangle$). Therefore, the goal was to decode the corresponding semantic label y from TMR-related raw EEG signals X^{tmr}, or multidomain EEG signals $\langle X^{img}, X^{aud}, X^{tmr} \rangle$.

Our SI-Multi Domain (SI-MD) model contains two parts: (1) Convolutional Encoder (refers to the one-dimensional (1D) convolution module and linear module), and (2) Transformer Encoder.

(1) Convolutional Encoder. As each EEG sample X has multiple channels, it is vital to fuse multiple channels to extract meaningful tokens before token-wise interaction by self-attention. Our convolutional encoder, which comprises a convolution module with

residual convolution blocks and a linear projection module, encodes raw EEG signals into token embeddings. Each residual convolution block contains two 1-D convolution layers and a batch normalization layer[68] and the linear module contains two linear projection layers. We denote the output token embeddings from the convolutional encoder as

$$\mathcal{E}_t = \{e_i^t \in \mathbb{R}^d | i = 1, \ldots, N\} \tag{4}$$

where d is the dimension of the embeddings, and N is the number of token embeddings.

(2) Transformer Encoder. Before feeding the embedding sequence into the transformer encoder, we add the parameter-free position embeddings introduced in previous work[47] to enable the model to be aware of the relative positions among embeddings. Then, the sequence of embeddings will be fed into the transformer encoder to get the final encoded $\mathcal{E} = \{e_i \in \mathbb{R}^d | i = 1, \ldots, N\}$. For the downstream 15-way classification task, we flatten the output embeddings followed by a classification layer.

Since the neural patterns differ greatly among different tasks, we utilize three different convolutional encoders to extract meaningful tokens from raw EEG signals $\langle X^{img}, X^{aud}, X^{tmr} \rangle$. After tokenization, the embeddings are fed into the transformer encoder for sequence modeling. Then, the output embeddings $\mathcal{E}$ are used to predict the semantic label $\hat{y}$ and calculate the classification loss.

To better utilize the template from the awake dataset, we introduce two special model designs: (1) a contrastive objective to align awake and sleep neural latent sequence position-wise, and (2) a alignment encoder to further improve alignment. The overall architecture of model is shown in **Figure 4a**.

Contrastive Objective. we calculate contrastive loss and align the neural representation sequences sharing the same category, originating from different tasks. Our contrastive loss consists of three parts, each aligning representations from two different tasks.

For each batch $\mathcal{B}$, we randomly draw $|\mathcal{B}|$ sample pairs $\langle X^{tmr}, y \rangle$ from the whole TMR-related dataset $D^{tmr} = \{D_s^{tmr}\}_{\{s \in S\}}$. Then, for each sample pair, we further draw one image-evoked sample X^{img} and one audio-evoked sample X^{aud} according to the semantic label y and the subject id s. We have $\mathcal{B} = \{\langle X_i^{img}, X_i^{aud}, X_i^{tmr}, y_i \rangle\}_{i=1}^{|\mathcal{B}|}$,

where different items could come from different subjects. After mapping the raw EEG signals X into encoded embedding sequence $\mathcal{E}$, we utilize the position-wise projection to transform embeddings and flatten them to form the latent embedding z. Then, we calculate the contrastive loss of three combinations based on the Equation-1.

Alignment Encoder. In our experiments, directly aligning the neural representation sequence of TMR-related EEG signals X^{tmr} with that of both X^{img} and X^{aud} often led to a slight degradation, compared to aligning with either of them individually. The initial design required the model to simultaneously find the optimal alignment among them, which posed a challenging task. To further smooth the aligning process, we feed the output embeddings $\mathcal{E}^{tmr}$ from X^{tmr} into the transformer encoders of awake signals, yielding $\mathcal{E}^{tmr2img}$ and $\mathcal{E}^{tmr2aud}$ respectively. Therefore, we utilize the transformed embeddings to calculate contrastive loss, i.e., $\mathcal{E}^{img}$ vs. $\mathcal{E}^{tmr2img}$ and $\mathcal{E}^{aud}$ vs. $\mathcal{E}^{tmr2aud}$.

Besides, all of these embeddings $\langle \mathcal{E}^{tmr}, \mathcal{E}^{tmr2img}, \mathcal{E}^{tmr2aud} \rangle$ are utilized to predict the semantic labels, and averaged to get the final prediction y, which resembles the model ensemble method.

The total loss for training model consists of two parts: (1) classification loss, and (2) contrastive loss. For each data item $\langle X^{img}, X^{aud}, X^{tmr}, y \rangle$, all of these transformed embeddings $\langle \mathcal{E}^{tmr}, \mathcal{E}^{img}, \mathcal{E}^{aud}, \mathcal{E}^{tmr2img}, \mathcal{E}^{tmr2aud} \rangle$ are utilized to calculate the classification loss:

$$L_{cls} = L_{cls}^{tmr} + L_{cls}^{img} + L_{cls}^{aud} + L_{cls}^{tmr2img} + L_{cls}^{tmr2aud} \tag{5}$$

And the contrastive loss is calculated as demonstrated before in Equation-2:

Therefore, the total training objective of model is:

$$L_{total} = L_{cls} + L_{contra} \tag{6}$$

The decoding model was trained up to 100 epochs by default using AdamW with a learning rate of $8 \cdot 10^{-5}$, a weight decay of $2 \cdot 10^{-5}$ and a batch size of 256 on four NVIDIA V100 GPU with 32GB of memory.

Decoding Model comparison and ablation experiment

The five models employed for comparison are as followed: the Logistic GLM model, the LSTM-based model, the CNN-based model, the Transformer-based model, and our

Sleep Interpreter – Single Domain (SI-SD) model. The Logistic GLM follows the standard setup in neuroscience: we train and evaluate a time-specific classifier for each time point, then take the maximum accuracy among these classifiers. The LSTM-based model comprises two basic LSTM layers, while the CNN-based model comprises two basic convolutional layers. The Transformer-based model consists of transformers with a depth of 8 and an attention head count of 6. For the SI-SD model, it includes the same architecture as the SI-MD model but excludes the audio encoder, the image encoder, the convolutional encoders of awake data domains and neural contrastive learning.

Pretrain-finetune pipeline

Our method consists primarily of two stages: the pretraining stage and the finetuning stage. During the pretraining stage, we leave one subject out, e.g. subject i. Then, we use the datasets from the rest subjects to format the training dataset $D_{train} = \{(D_s^{img}, D_s^{aud}, D_s^{tmr})\}_{\{s \in S \setminus \{i\}\}}$. Following the previous training procedure, we get a pretrained model for that subject.

The fine-tuning process comes after the pretraining stage, in which we will finetune the pretrained model using data from the test subjects. For each dataset of test subject i, i.e., $(D_i^{img}, D_i^{aud}, D_i^{tmr})$, we split the dataset into training, validation, testing splits with a size roughly proportional to 80%, 10%, and 10%. We train each model with the training split for 200 epochs, by default using Adam with a learning rate of $2 \cdot 10^{-5}$ and a batch size of 64 on one NVIDIA V100 GPU with 32GB of memory. And then evaluate its performance on validation and testing splits. We take the test accuracy according to the maximum validation accuracy as its performance and compute their average as their performance.

Slow Oscillation Phases

Based on the criteria proposed by existing studies[69], we detected the "Slow Oscillation" (SO) events.

First, raw EEG data was filtered between 0.16–1.25 Hz using ‘sosfiltfilt’ function of python package ‘signal’ (bandpass, order=4). Second, all zero-crossings were determined in the filtered signal, and event duration was determined for SO candidates (Down states followed by Up states) as time between two successive positive-to-negative zero-crossings for Fz channel. Events that met the SO duration criteria

(minimum of 0.8 and maximum of 2 s, 0.5–1.25 Hz) entered the next step. Third, event amplitudes were determined for the remaining SO candidates (trough-to-peak amplitude between two positive-to-negative zero crossing for Fz channel). Events that also met the SO amplitude criteria (≥75% percentile of subject's overall amplitudes, that is, the 25% of events with the largest amplitudes) were considered SO events.

Then, we segmented same raw EEG signals into 6-second samples from -1.5s to 4.5s according to the onset of audio stimulus using python package MNE. We further categorized the data into different SO phases according to the detected SO events in these segmented samples. Since multiple SO events can be detected in each sample, our classification was based on the first detected SO event. Samples were categorized into different SO phases according to the following operational definitions:

1. Trans 1st Up: This phase refers to the onset of the audio stimuli occurring during the transition to the "1st Up" phase of the same SO event. Specifically, the transition corresponds to the onset of audio stimuli located within 1s before the start point of the detected "1st Up" phase.

2. 1st Up: In the time range of one second before the start point of the detected SO event, there could exist a waveform similar to the conventional "Up" state[6,69]. And there is another zero-crossing point in this time range, whereby we can determine a wave peak and take the portion above 50% of the peak value as the "1st Up" phase. If the required zero-crossing cannot be detected in the corresponding one second time period, we choose to use the maximum of the wave peaks in this period as the reference value for division.

3. Trans Down: This phase refers to the onset of the audio stimuli occurs during the transition to the "Down" state of the detected SO event. The transition corresponds to the stage between the end of the "1st Up" phase and the beginning of the "Down" phase.

4. Down: The "Down" phase aligns with the "Down" state in previous studies[6,69]. However, we further refine this by strictly defining the "Down" phase as the portion below 50% of the trough value, and the onset of the audio stimuli occurs during this portion corresponds to the "Down" phase.

5. Trans 2nd Up: This phase refers to the onset of the audio stimuli occurs during the transition to the "2nd Up" phase of the same SO event. This corresponds to the stage between the end of the "Down" phase and the beginning of the "2nd Up" phase.

6. 2nd Up: The "2nd Up" phase aligns with the “Up” state in previous studies[6,69], but we define it more strictly. Only the onset of the audio stimuli occurs during portions of the former Up state waveform that exceed 50% of the peak value corresponds to the "2nd Up" phase.

7. Post SO: This phase refers to the onset of the audio stimuli occurring after the SO event. Instead of using the traditional zero point, we define the end of the SO as the end of the 2nd Up phase of the same SO event.

8. Other: This phase encompasses scenarios where the SO is detected in locations not specified in the above definitions during the detection time period, as well as instances where no SO meeting the defined criteria occurs within the detection time period.

Spindle Coupling

The detection method for spindles and spindles coupling was as follows. First, raw data was filtered between 12–16 Hz using ‘sosfiltfilt’ function of python package ‘signal’ (bandpass, order=4). The mean amplitude was then calculated for the filtered signal using a moving average of 200 ms, with the spindle amplitude criterion defined as the 75th percentile of the subject’s overall mean amplitude values. Finally, whenever the signal exceeded this threshold for more than 0.5s but less than 3s (duration criteria), a spindle event was deemed to have been detected.

For the detected spindles, we used the trough point of the cue-evoked wave in the segmented sample as the starting point for SO-spindle coupling judgment. To simplify the analysis, temporal-frequency analysis showed that the low-frequency signal activation representing the evoked wave appeared 1s after the sound stimulus onset and the trough appeared around 1.3s after the sound stimulus onset. Since our decoding analysis only included data up to 2.3 s after the sound stimulus onset, spindle detection extended to 2.3 s after the sound stimulus onset to ensure that the analysis results match the existing decoding results. Thus, the detection time range for spindle-coupling was defined as 1.3-2.3 s after the sound stimulus onset.

Temporal-Frequency Analysis

Temporal-frequency plots were calculated per event epoch (-1.5s to 4.5s according to the onset of audio stimulus) on Fz channel of raw data, using the ‘STFT’ function of the ‘signal’ package in python. The frequencies ranged from 0.1 Hz to 50 Hz in 13 steps, utilizing a sliding Blackman window with a variable, frequency-dependent length that

always comprised 12 cycles. Time-locked temporal-frequency plots of all epochs were then normalized as percent change from the post-event baseline (4.0 to 4.5 s according to sound stimulus onset) and averaged per participant. The entire temporal-frequency analysis was conducted on Fz channel of the same sleep EEG data used for SO detection.

Frequency Band Analysis

We filtered the raw EEG data using the following ranges: 0.5-4 Hz, 4-8 Hz, 8-12 Hz, 12-30 Hz, and 30-80 Hz, for delta, theta, alpha, beta, and gamma bands, respectively using 'filtfilt' function of python package 'signal' (bandpass, order=4). For the filtered data, we pretrained the SI-MD model separately with the training data filtered by each of the five frequency bands and used the pre-trained models to test on a total of six types of test data (one unfiltered and five filtered).

Magnitude Analysis

We selected the raw sleep EEG data from the Fz channel and calculated the mean magnitude within the range of 0.3-2.3 s relative to the sound stimulus onset. The mean magnitude was extracted using the Hilbert transform and convolved with a time window of 0.2s. The calculation range was adjusted to 0.35-2.25 s to account for the 0.2s sliding window and to avoid interference from out-of-bounds data.

Model Retrain

We first removed the data in "Post SO" phase and then we re-conducted the spindle detection of a loosen amplitude criteria (amplitude higher than 50% of the average amplitude within the data segment [-1.5s, 4.5s]). We next removed the data in "Down" phase without SO-spindle coupling and data with detected SO but without spindle coupling in "Other" phase. Subsequently, we adjusted the depth of the convolutional encoder in SI-MD model structure. Considering the reduced size of the training dataset, we reduced the depth of convolution module from three residual convolution blocks to one block while maintaining the linear module of the convolutional encoder in SI-MD model. Finally, we retrained the adjusted SI-MD model using the cleaned data. The model was trained up to 100 epochs by default using AdamW with a learning rate of $8 \cdot 10^{-5}$, a weight decay of $2 \cdot 10^{-5}$ and a batch size of 256 on four NVIDIA V100 GPU with 32GB of memory.

Closed Loop, Real-Time Sleep Decoding System

The system processes the subject's real-time data in two ways: sleep staging and sleep decoding. First, the device outputs EEG data from channel C3, M2, and REOG, REMG and EMGREF data for sleep staging. Next, the output data is preprocessed according to the preprocessing of sleep staging data in Methods to be the three channel data (EEG, EOG, EMG) and fed into the real-time sleep staging model. The initial 30 seconds of data will be processed by the staging model and directly output as the staging result for this period. After this, the model will perform staging every 3 seconds, updating the original staging results accordingly. If the model detects a stable NREM2/3 or REM stage, a sound stimulus is delivered randomly, with a 4-6 second interval, to the subject. The real-time sleep decoding model is activated whenever it is determined the subject is in NREM2/3 or REM stage. Simultaneously, neural data for sleep decoding (EEG from all channels) is preprocessed according to the preprocessing of sleep decoding data in Methods and input into both the decoding model and the diffusion model. These models process the data to produce the decoding results and generate corresponding images.

Latent Diffusion Model

Based on an open-source latent diffusion model developed for EEG signals in existing studies[70], we replaced the EEG processing model with the sleep data processing component of the SI-MD model and adapted it to our sleep dataset. Since the model requires both EEG data and image data as input, we used the 'Image.open' function from the PIL package to convert the fifteen different image stimuli into array format, matching them to the corresponding sleep data. Next, we input the preprocessed sleep decoding training data along with the matched transformed image data into the latent diffusion model for fine-tuning. The latent diffusion model was trained up to 200 epochs by default using Adam with a learning rate of $2 \cdot 10^{-5}$ and a batch size of 256 on eight NVIDIA V100 GPU with 32GB of memory.

Acknowledgment

Conceptualization, Y.L., Z.C., H.Z., J.Z., L.Z., R.D., and T.B.; Investigation, Z.C., H.Z., J.Z., L.Z., P.L., H.W., Y.L., and T.B.; Writing – Original Draft, Z.C., Y.L.; Writing – Review & Editing, Y.L., Z.C., R.D., and T.B. This study is supported by the National Science and Technology Innovation 2030 Major Program (2022ZD0205500), the National Natural Science Foundation of China (32271093), the Beijing Natural Science Foundation (Z230010, L222033), and the Fundamental Research Funds for the Central Universities. R.D. and Z.C. are supported by UK DRI Key Questions Funding Programme (DRI-KQ2024-S1-1).

Conflict of interest

The authors have indicated they have no potential conflicts of interest to disclose.

Data availability

The benchmark sleep-TMR dataset will be available on https://osf.io/x7453/ upon publication.

Code availability

The code for real-time sleep decoding system will be available on https://gitlab.com/liulab1/sleep-interpreter.git upon publication.

References

1. Diekelmann, S. & Born, J. The memory function of sleep. *Nat. Rev. Neurosci.* **11**, 114-126 (2010).
2. Rasch, B. & Born, J. About sleep's role in memory. *Physiol. Rev.* (2013).
3. Klinzing, J.G., Niethard, N. & Born, J. Mechanisms of systems memory consolidation during sleep. *Nat. Neurosci.* **22**, 1598-1610 (2019).
4. Brodt, S., Inostroza, M., Niethard, N. & Born, J. Sleep—A brain-state serving systems memory consolidation. *Neuron* **111**, 1050-1075 (2023).
5. Gais, S. & Born, J. Declarative memory consolidation: mechanisms acting during human sleep. *Learning & memory* **11**, 679-685 (2004).
6. Ngo, H.-V.V. & Staresina, B.P. Shaping overnight consolidation via slow-oscillation closed-loop targeted memory reactivation. *Proc. Natl. Acad. Sci. USA* **119**, e2123428119. (2022).
7. Girardeau, G. & Lopes-dos-Santos, V. Brain neural patterns and the memory function of sleep. *Science* **374**, 560-564 (2021).
8. Skaggs, W.E. & McNaughton, B.L. Replay of neuronal firing sequences in rat hippocampus during sleep following spatial experience. *Science* **271**, 1870-1873. (1996).
9. Wilson, M.A. & McNaughton, B.L. Reactivation of hippocampal ensemble memories during sleep. *Science* **265**, 676-679. (1994).
10. Liu, Y., Nour, M.M., Schuck, N.W., Behrens, T.E.J. & Dolan, R.J. Decoding cognition from spontaneous neural activity. *Nat. Rev. Neurosci.* **23**, 204-214. (2022).
11. Keefe, F.B., Johnson, L.C. & Hunter, E.J. EEG and autonomic response pattern during waking and sleep stages. *Psychophysiology* **8**, 198-212 (1971).
12. Niedermeyer, E. The normal EEG of the waking adult. *Electroencephalography: Basic principles, clinical applications, and related fields* **167**, 155-164 (2005).
13. Staresina, B.P. Coupled sleep rhythms for memory consolidation. *Trends Cognit. Sci.* (2024).
14. Staresina, B.P., Niediek, J., Borger, V., Surges, R. & Mormann, F. How coupled slow oscillations, spindles and ripples coordinate neuronal processing and communication during human sleep. *Nat. Neurosci.* **26**, 1429-1437 (2023).
15. Storch, C., Höhne, A., Holsboer, F. & Ohl, F. Activity patterns as a correlate for sleep–wake behaviour in mice. *J. Neurosci. Methods* **133**, 173-179 (2004).
16. Sulaman, B.A., Wang, S., Tyan, J. & Eban-Rothschild, A. Neuro-orchestration of sleep and wakefulness. *Nat. Neurosci.* **26**, 196-212 (2023).
17. Schreiner, T., Petzka, M., Staudigl, T. & Staresina, B.P. Endogenous memory reactivation during sleep in humans is clocked by slow oscillation-spindle complexes. *Nat. Commun.* **12**, 3112. (2021).
18. Hu, X., Cheng, L.Y., Chiu, M.H. & Paller, K.A. Promoting memory consolidation during sleep: A meta-analysis of targeted memory reactivation. *Psychol. Bull.* **146**, 218 (2020).
19. Paller, K.A., Creery, J.D. & Schechtman, E. Memory and sleep: how sleep cognition can change the waking mind for the better. *Annu. Rev. Psychol.* **72**, 123-150 (2021).
20. Rasch, B.r., Bu¨chel, C., Gais, S. & Born, J. Odor cues during slow-wave sleep prompt declarative memory consolidation. *Science* **315**, 1426-1429 (2007).
21. Radford, A.*, et al.* Learning transferable visual models from natural language supervision. in *International Conference on Machine Learning (ICML)* 8748-8763 (PMLR, 2021).
22. Khosla, P.*, et al.* Supervised contrastive learning. *Advances in Neural Information Processing*

Systems (NeurIPS) **33**, 18661-18673. (2020).
23. Le-Khac, P.H., Healy, G. & Smeaton, A.F. Contrastive representation learning: A framework and review. *IEEE Access* **8**, 193907-193934 (2020).
24. McCarley, R.W. Neurobiology of REM and NREM sleep. *Sleep medicine* **8**, 302-330 (2007).
25. Nielsen, T.A. A review of mentation in REM and NREM sleep:"covert" REM sleep as a possible reconciliation of two opposing models. *Behav. Brain Sci.* **23**, 851-866 (2000).
26. Chauvette, S., Volgushev, M. & Timofeev, I. Origin of active states in local neocortical networks during slow sleep oscillation. *Cereb. Cortex* **20**, 2660-2674 (2010).
27. Lemieux, M., Chen, J.-Y., Lonjers, P., Bazhenov, M. & Timofeev, I. The impact of cortical deafferentation on the neocortical slow oscillation. *J. Neurosci.* **34**, 5689-5703 (2014).
28. Volgushev, M., Chauvette, S., Mukovski, M. & Timofeev, I. Precise long-range synchronization of activity and silence in neocortical neurons during slow-wave sleep. *J. Neurosci.* **26**, 5665-5672 (2006).
29. Hobson, J.A. REM sleep and dreaming: towards a theory of protoconsciousness. *Nat. Rev. Neurosci.* **10**, 803-813 (2009).
30. Hobson, J.A., Stickgold, R. & Pace-Schott, E.F. The neuropsychology of REM sleep dreaming. *Neuroreport* **9**, R1-R14 (1998).
31. Ackermann, S. & Rasch, B. Differential Effects of Non-REM and REM Sleep on Memory Consolidation? *Curr. Behav. Neurosci. Rep.* **14**, 430 (2014).
32. Vyazovskiy, V.V. & Delogu, A. NREM and REM sleep: complementary roles in recovery after wakefulness. *The Neuroscientist* **20**, 203-219 (2014).
33. Helfrich, R.F., Mander, B.A., Jagust, W.J., Knight, R.T. & Walker, M.P. Old brains come uncoupled in sleep: slow wave-spindle synchrony, brain atrophy, and forgetting. *Neuron* **97**, 221-230. e224 (2018).
34. Latchoumane, C.-F.V., Ngo, H.-V.V., Born, J. & Shin, H.-S. Thalamic spindles promote memory formation during sleep through triple phase-locking of cortical, thalamic, and hippocampal rhythms. *Neuron* **95**, 424-435. e426 (2017).
35. Niethard, N., Ngo, H.-V.V., Ehrlich, I. & Born, J. Cortical circuit activity underlying sleep slow oscillations and spindles. *Proc. Natl. Acad. Sci. USA* **115**, E9220-E9229 (2018).
36. Niknazar, M., Krishnan, G.P., Bazhenov, M. & Mednick, S.C. Coupling of thalamocortical sleep oscillations are important for memory consolidation in humans. *PLoS One* **10**, e0144720 (2015).
37. Tononi, G. & Cirelli, C. Sleep and synaptic down-selection. *Micro-, meso-and macro-dynamics of the brain*, 99-106 (2016).
38. Lawhern, V.J.*, et al.* EEGNet: a compact convolutional neural network for EEG-based brain--computer interfaces. *J. Neural Eng.* **15**, 056013. (2018).
39. Song, Y., Zheng, Q., Liu, B. & Gao, X. EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Trans. Neural Syst. Rehabil. Eng.* **31**, 710-719. (2022).
40. Wang, C.*, et al.* BrainBERT: Self-supervised representation learning for intracranial recordings. *arXiv preprint arXiv:2302.14367* (2023).
41. Yang, C., Westover, M. & Sun, J. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems (NeurIPS)* **36** (2024).
42. Jiang, W.-B., Zhao, L.-M. & Lu, B.-L. Large brain model for learning generic representations with tremendous EEG data in BCI. *arXiv preprint arXiv:2405.18765* (2024).
43. Berry, R.B.*, et al.* Rules for scoring respiratory events in sleep: update of the 2007 AASM

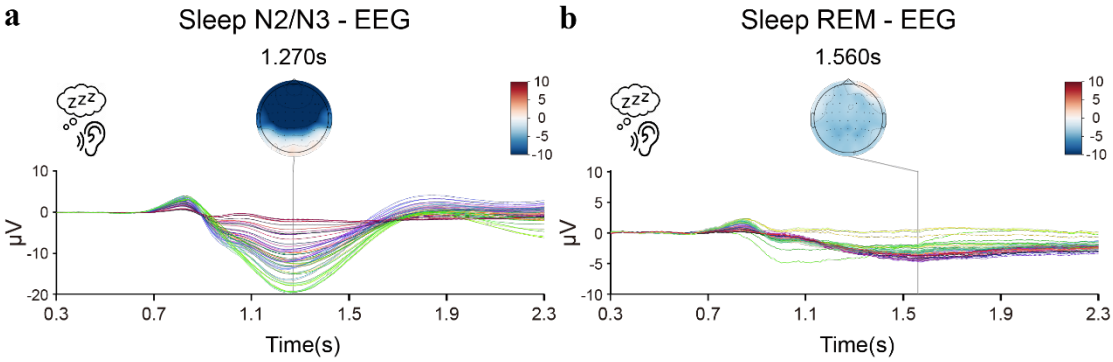
manual for the scoring of sleep and associated events: deliberations of the sleep apnea definitions task force of the American Academy of Sleep Medicine. *J. Clin. Sleep Med.* **8**, 597-619 (2012).
44. Iber, C. The AASM manual for the scoring of sleep and associated events: rules, terminology, and technical specification. (2007).
45. Vallat, R. & Walker, M.P. An open-source, high-performance tool for automated sleep staging. *eLife* **10**, e70092 (2021).
46. Krizhevsky, A., Sutskever, I. & Hinton, G.E. ImageNet classification with deep convolutional neural networks. *Commun. ACM* **60**, 84–90 (2017).
47. Vaswani, A.*, et al.* Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)* **30** (2017).
48. Dfossez, A., Caucheteux, C., Rapin, J., Kabeli, O. & King, J.-R. Decoding speech perception from non-invasive brain recordings. *Nat. Mach. Intell.* **5**, 1097-1107 (2023).
49. Baevski, A., Zhou, Y., Mohamed, A. & Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems (NeurIPS)* **33**, 12449-12460 (2020).
50. Batterink, L.J., Creery, J.D. & Paller, K.A. Phase of spontaneous slow oscillations during sleep influences memory-related processing of auditory cues. *J. Neurosci.* **36**, 1401-1409 (2016).
51. Xia, T., Antony, J.W., Paller, K.A. & Hu, X. Targeted memory reactivation during sleep influences social bias as a function of slow-oscillation phase and delta power. *Psychophysiology* **60**, e14224 (2023).
52. Hahn, M.A., Heib, D., Schabus, M., Hoedlmoser, K. & Helfrich, R.F. Slow oscillation-spindle coupling predicts enhanced memory formation from childhood to adolescence. *elife* **9**, e53730 (2020).
53. Muehlroth, B.E.*, et al.* Precise slow oscillation–spindle coupling promotes memory consolidation in younger and older adults. *Sci. Rep.* **9**, 1940 (2019).
54. Albelwi, S. Survey on self-supervised learning: auxiliary pretext tasks and contrastive learning methods in imaging. *Entropy* **24**, 551 (2022).
55. Jaiswal, A., Babu, A.R., Zadeh, M.Z., Banerjee, D. & Makedon, F. A survey on contrastive self-supervised learning. *Technologies* **9**, 2 (2020).
56. Gifford, A.T., Dwivedi, K., Roig, G. & Cichy, R.M. A large and rich EEG dataset for modeling human visual object recognition. *NeuroImage* **264**, 119754 (2022).
57. Grootswagers, T., Zhou, I., Robinson, A.K., Hebart, M.N. & Carlson, T.A. Human EEG recordings for 1,854 concepts presented in rapid serial visual presentation streams. *Sci. Data* **9**, 3 (2022).
58. Hebart, M.N.*, et al.* THINGS-data, a multimodal collection of large-scale datasets for investigating object representations in human brain and behavior. *eLife* **12**, e82580 (2023).
59. Liu, J.*, et al.* Item-specific neural representations during human sleep support long-term memory. *PLoS Biol.* **21**, e3002399 (2023).
60. Hobson, J.A., Pace-Schott, E.F. & Stickgold, R. Dreaming and the brain: Toward a cognitive neuroscience of conscious states. *Behav. Brain Sci.* **23**, 793-842 (2000).
61. Walker, M.P., Liston, C., Hobson, J.A. & Stickgold, R. Cognitive flexibility across the sleep–wake cycle: REM-sleep enhancement of anagram problem solving. *Cogn. Brain Res.* **14**, 317-324 (2002).
62. Antrobus, J. REM and NREM sleep reports: Comparison of word frequencies by cognitive classes. *Psychophysiology* **20**, 562-568 (1983).
63. Foulkes, D. Dream Research: 1953-1993. *Sleep* **19**, 609-624 (1996).

64. Banville, H., Chehab, O., Hyvärinen, A., Engemann, D.-A. & Gramfort, A. Uncovering the structure of clinical EEG signals with self-supervised learning. *J. Neural Eng.* **18**, 046020 (2021).
65. Thomas, A., Ré, C. & Poldrack, R. Self-supervised learning of brain dynamics from broad neuroimaging data. *Advances in Neural Information Processing Systems (NeurIPS)* **35**, 21255-21269 (2022).
66. Gramfort, A.*, et al.* MEG and EEG data analysis with MNE-Python. *Front. Neurosci.* **7** (2013).
67. Li, A., Feitelberg, J., Saini, A.P., Höchenberger, R. & Scheltienne, M. MNE-ICALabel: Automatically annotating ICA components with ICLabel in Python. *J. Open Source Softw.* **7**, 4484 (2022).
68. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. in *International Conference on Machine Learning (ICML)* 448-456. (PMLR, 2015).
69. Staresina, B.P.*, et al.* Hierarchical nesting of slow oscillations, spindles and ripples in the human hippocampus during sleep. *Nat. Neurosci.* **18**, 1679-1686. (2015).
70. Bai, Y.*, et al.* Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv preprint arXiv:2306.16934* (2023).

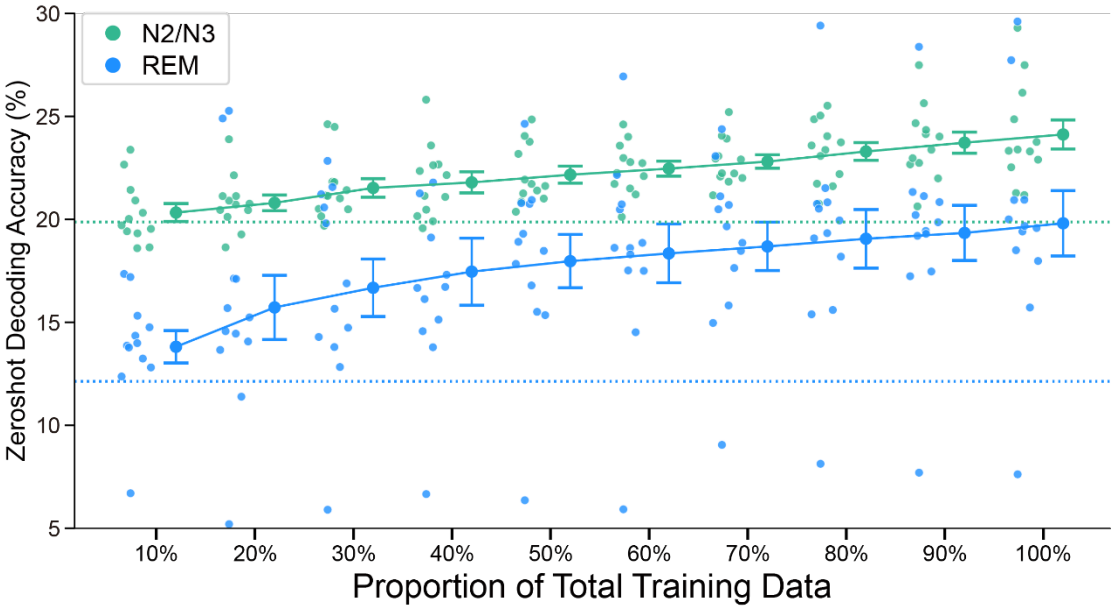
Supplementary Information

Supplementary Table 1: Performance of different EEG-based ANN models in decoding NREM2/3 and REM sleep states. For each model, the 15-way classification accuracy (%) and the standard error (%) across test subjects are reported. The chance level is 1/15 = 6.7%.

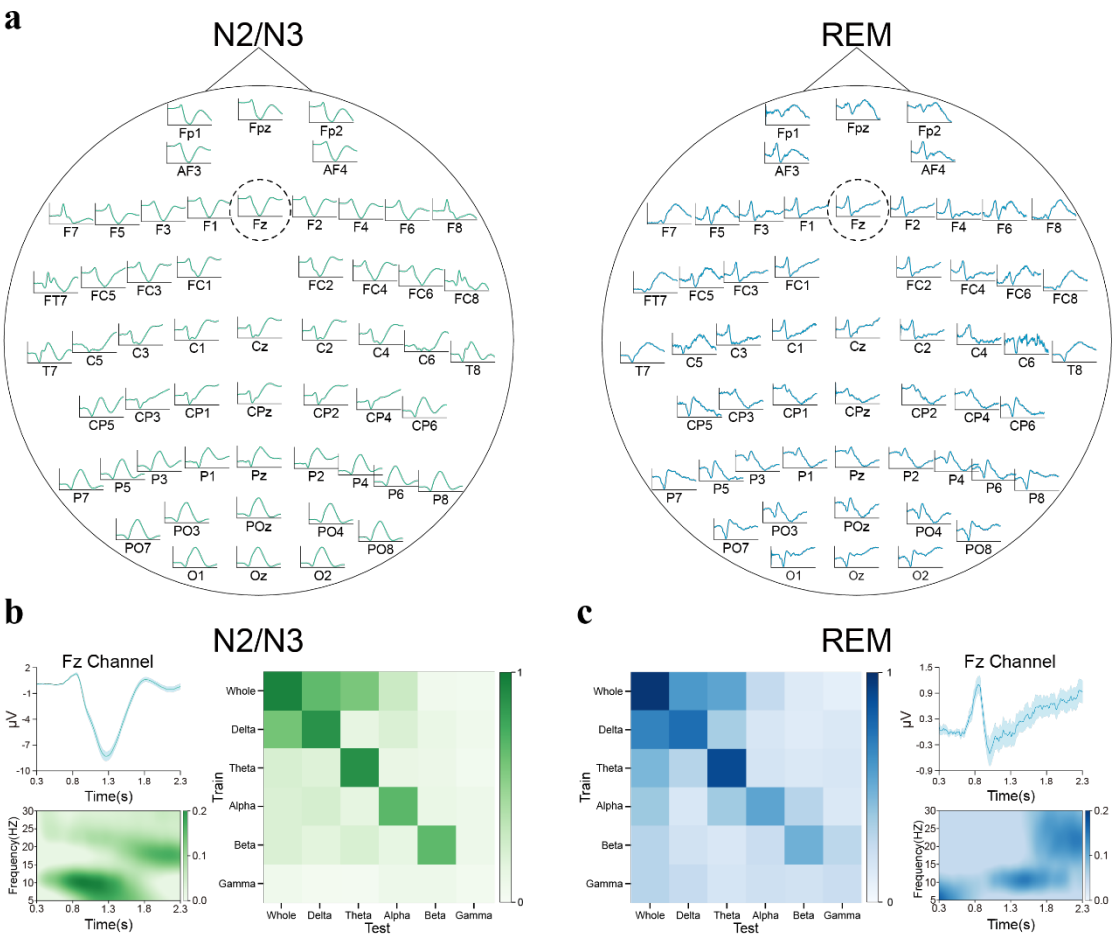
Models	Zero-shot		Finetune	
	NREM 2/3	REM	NREM 2/3	REM
EEG-Net[38]	17.30±0.37%	14.05±1.11%	20.03±0.66%	21.38±1.95%
EEG-Conformer[39]	19.12±0.48%	14.81±0.79%	23.74±0.83%	24.61±1.82%
BrainBERT[40]	12.50±0.78%	10.59±1.31%	17.30±1.44%	18.21±2.02%
BIOT[41]	16.72±0.52%	13.15±1.08%	22.80±1.19%	23.93±2.09%
LaBram[42]	17.91±1.61%	13.64±2.11%	21.99±2.14%	22.27±2.92%
SI-MD	**24.12±0.70%**	**19.81±1.59%**	**30.81±0.65%**	**31.65±1.96%**



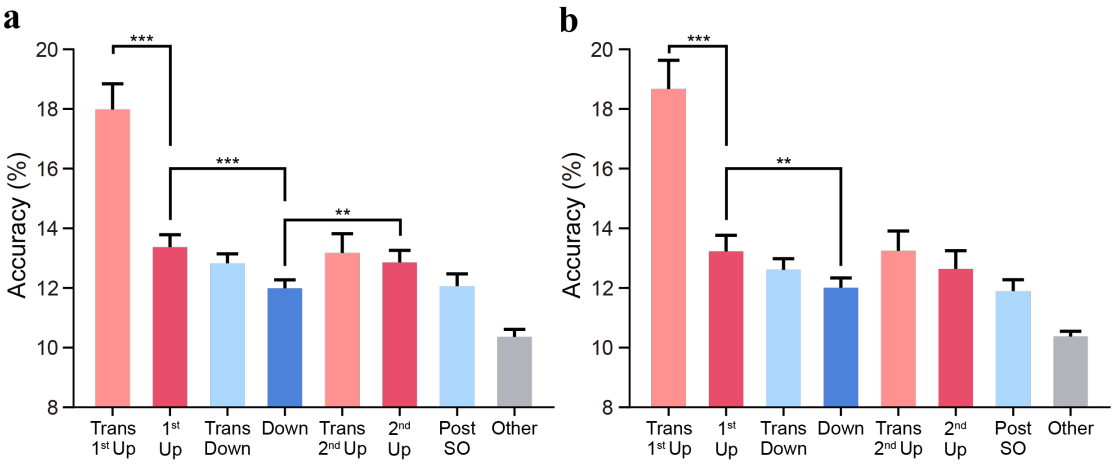
Supplementary Figure 1: The averaged ERPs and the topographic maps evoked by TMR-related audio stimuli during N2/N3 and REM sleep. (**a**) The averaged ERPs and the topographic maps evoked by TMR-related audio stimuli during NREM2/3 sleep. (**b**) The averaged ERPs and the topographic maps evoked by TMR-related audio stimuli during REM sleep. The zeros on the horizontal coordinates correspond to the audio stimulus onset. ERPs were computed within a time window from 0.3 seconds after stimulus onset to 2.3 seconds after stimulus onset. The topographic maps were selected to show the time points with large neural responses in ERPs, where the same amplitude range (-10 to 10 μV) was used to facilitate comparison of activations.



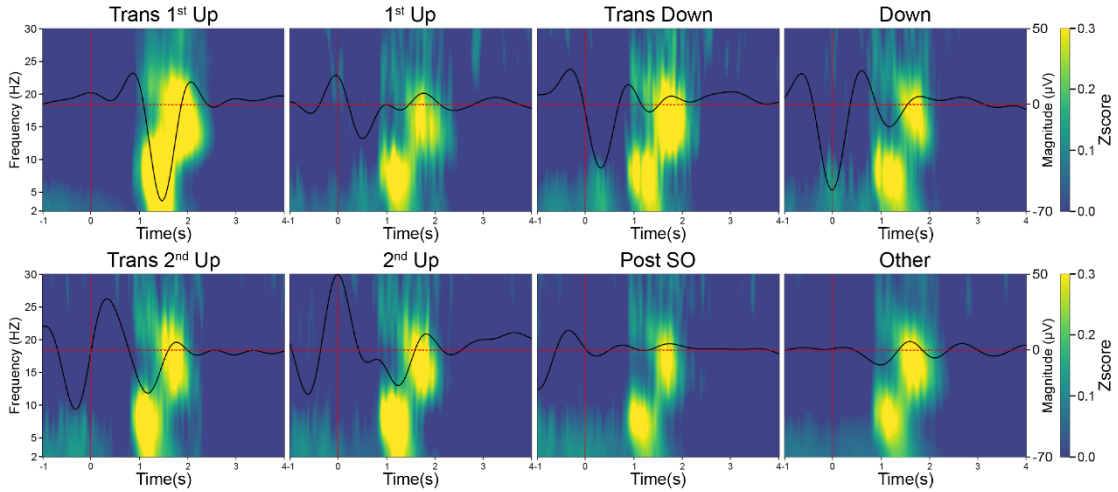
Supplementary Figure 2: Zero-shot sleep decoding accuracy as a function of training data size. We evaluated the changes in zero-shot, cross-subject, sleep decoding performance under different training data volumes. This was done by randomly selecting 10% to 100% of the total training data to retrain the SI-MD model. The results for NREM2/3 stage are shown in green, and the results for REM stage are shown in blue. Each data point represents the result for a test subject.



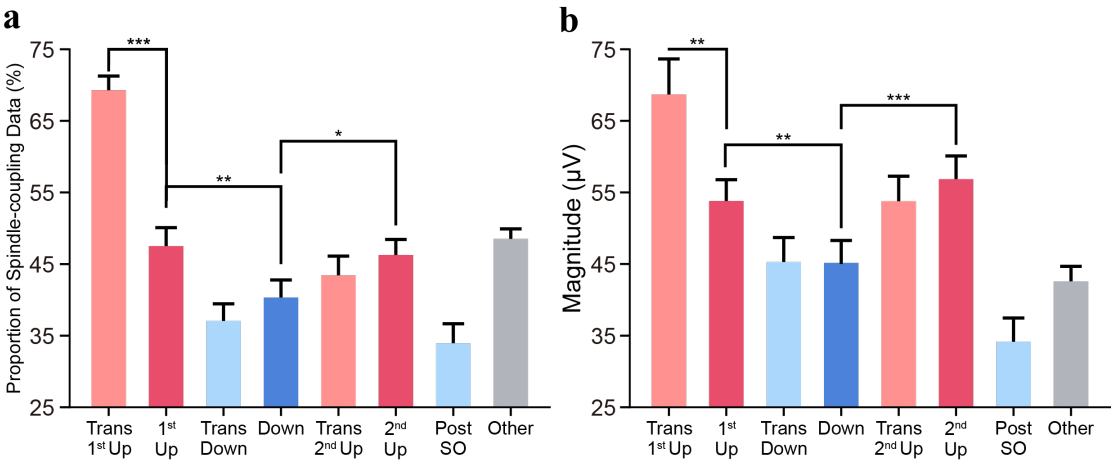
Supplementary Figure 3: ERPs and temporal frequency patterns for auditory-stimuli evoked neural response during sleep. (**a**) Whole brain ERPs during sleep. ERPs of all EEG channels were shown separately for NREM2/3 (left panel) and REM (right panel) sleep stages. The time window spans from 0.3 to 2.3 seconds after stimulus onset. (**b**) Temporal frequency patterns of evoked neural response during NREM2/3 sleep. On the left, taking the Fz channel as an example, the upper left panel illustrates the ERP at the Fz channel during NREM2/3 stage, indicated in green, along with the corresponding standard errors across subjects. The lower left panel presents the time-frequency plots of the EEG signals from the Fz channel during this stage. The right panel shows the trained and tested results of the SI-MD based on specific frequency bands, with sleep decoding accuracy normalized between 0 and 1. (**c**) Temporal frequency patterns of evoked neural response during REM sleep, similar to panel (b), but for REM sleep, shown in blue.



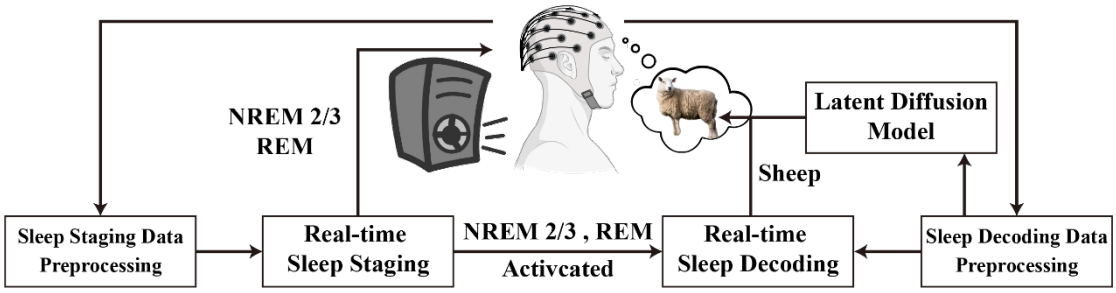
Supplementary Figure 4: Generalisation performance of awake-trained Lasso-GLM on decoding epochs in different Slow Oscillation (SO) stages during NREM 2/3 sleep. (**a**) Lasso-GLM was trained on the awake image-evoked neural responses and tested on epochs in different SO stages during NREM 2/3 sleep of the same subjects. (**b**) Similar to panel (a), but trained on the awake audio-evoked neural responses. The color code is in line with those used in **Figure 5b**. *p<0.05, ** p<0.01, *** p<0.001.



Supplementary Figure 5: Temporal-frequency analysis of all eight SO-groups during NREM2/3 sleep. Temporal frequency analysis of EEG data from the Fz channel was shown separately for each of the eight SO-groups: 'Transition to 1st Up (Trans 1st UP)', '1st Up', 'Transition to Down (Trans Down)', 'DOWN', 'Transition to 2nd Up (Trans 2nd UP)', '2nd Up', 'Post SO', and 'Other'. The y-axis on the left side corresponds to the frequency range in the time-frequency plot, while the y-axis on the right corresponds to the magnitude in the ERP plot. For each SO-group, the time window is [-1s, 4s] relative to the sound stimulus onset. The vertical red lines mark the zero point in time, i.e., the onset of sound stimuli. The horizontal red line corresponds to the zero point of magnitude.



Supplementary Figure 6: Distribution spindle events and mean magnitude across eight SO groups during NREM2/3 sleep. (**a**) The distribution of SO-spindle coupling in different SO groups, with the highest spindle coupling observed in the 'Trans 1st UP' stage. (**b**) The distribution of the mean magnitude of ERPs in different SO groups, also showing the highest magnitude in the 'Trans 1st UP' stage. The color code is in line with those used in **Figure 5b**. *p<0.05, ** p<0.01, *** p<0.001.



Supplementary Figure 7: Data flow of the closed loop, real-time sleep decoding system. The system processes the subject's real-time data in two ways: sleep staging and sleep decoding. First, the device outputs neural data (EEG from channel C3, M2, as well as EOG, EMG) for sleep staging, which is then preprocessed and fed into the real-time sleep staging model. If the model detects a stable NREM2/3 or REM stage, a sound stimulus is delivered randomly, with a 4-6 second interval, to the subject (this can be adapted based on the research goal). The real-time sleep decoding model is activated whenever it is determined the subject is in NREM2/3 or REM stage. Simultaneously, neural data for sleep decoding (EEG from all channels) is preprocessed and input into both the decoding model and the diffusion model. These models process the data to produce the decoding results and generate corresponding images. See also **Supplementary Video 1** for demonstration.

Supplementary Video 1: Closed-loop real-time sleep decoding system. We selected partial real-time data from one of the test subjects to demonstrate the practical application of the system. The results of sleep staging and sleep decoding are displayed in the upper left corner of the screen, along with the TOP-3 results with the highest probabilities to assist experimenters in making additional judgments. The two image frames on the right display the images associated with sound stimuli and the corresponding images generated by the latent diffusion model from sleep data used for decoding. The display box of the images will turn green when the decoding result matches the cue result and red when it doesn't. Below this, the real-time data of the 30-second EEG-C3 channel, re-referenced to the M2 channel and filtered to 0.1-40 Hz, is shown with the onset of all sound stimuli marked by red vertical lines. The bottommost channel displays the sleep stage of the current 30-second data obtained by the model in real-time. Throughout the demonstration, the sleep staging model updates every 3 seconds, and the sleep decoding model updates when sound stimuli occur. The decoding model updates after the stimuli: the first update occurs at 0.3 seconds, the second at 2.3 seconds (corresponding to the time period of decoding results in **Figure 4b**), and the third at 6.0 seconds.