# Awe is characterized as an ambivalent experience in the human behavior and cortex:

integrated virtual reality-electroencephalogram study

Jinwoo Yi[1], Danny Dongyeop Han[2], Seung-Yeop Oh[3], Jiook Cha[1,2,4,5]

**Affiliations**

[1] Department of Brain and Cognitive Sciences, Seoul National University, South Korea

[2] Interdisciplinary Program of Artificial Intelligence, Seoul National University, South Korea

[3] AIR: AI Rebirth, South Korea

[4] Department of Psychology, Seoul National University, South Korea

[5] Institute of Psychological Science, Seoul National University, South Korea

**This file includes:**

Abstract

Main text

References

Figure 1-5

Acknowledgement

Author contributions

**Author Note:**

We have no conflicts of interest to disclose. Correspondence concerning this article should be addressed to Jinwoo Yi, BA, and Jiook Cha, PhD, Building 16, Office M512, Gwanak-ro 1, Gwanak-gu, Seoul, South Korea. Email: Jinwoo Yi (adem1997@snu.ac.kr) and Jiook Cha (connectome@snu.ac.kr)

# Abstract

Ambivalent feelings are a defining feature of awe, which has been understood as a possible source of its psychosocial benefits. However, due to the conventional unidimensional model of affective valence, behavior and neural representation of ambivalent feelings during awe remain elusive. To address this gap, we combined awe-inducing virtual reality clips, electroencephalogram, and a deep learning-based dimensionality reduction technique ($N = 43$). Behaviorally, awe ratings were precisely predicted by the duration and intensity of ambivalent feelings, not by single valence-related metrics. In the electrophysiological analysis, we identified latent neural space for each participant sharing valence representation structures across individuals and stimuli. In these spaces, ambivalent feelings during awe were distinctly represented from positive and negative ones, and the variability in their distinctiveness specifically predicted awe ratings. Additionally, frontal delta oscillations mainly engaged in differentiating valence representations. Our findings demonstrate that awe is fundamentally an ambivalent experience reflected in both behavior and electrophysiological activities. This work provides a new framework for understanding complex emotions and their neural underpinnings, with potential implications for affective neuroscience and relevant fields.

# Introduction

Awe, an emotion elicited by vast and enigmatic stimuli, both expands our minds and humbles us to our core. Understanding awe is crucial for unraveling the complexity of human emotional experiences and their impact on cognition, behavior, and well-being. Unlike simpler emotions, awe represents a unique interplay of cognitive expansion and self-diminishment, offering insights into how the human mind processes and integrates seemingly contradictory affective states. Unlike fear, awe is characterized by expanding one's conceptual frameworks in an attempt to comprehend the mysterious. Psychologists define awe through two key dimensions: 'perceived vastness' and 'a need for accommodation' (Keltner & Haidt, 2003). For example, astronauts describe their awe during space flights as being overwhelmed by the vast scale of the universe compared to the smallness of Earth (i.e., perceived vastness) while also feeling a profound sense of beauty and fragility, leading to a realization of the urgent need to preserve this beauty (i.e., need for accommodation)(Yaden et al., 2016).

This multidimensionality of awe shapes its ambivalent nature, and this duality is expected to lead to significant psychiatric, psychosocial, and intellectual benefits, such as stress resilience, empathy, and non-egocentric perspectives (Jiang et al., 2024; Li et al., 2023; Monroy & Keltner, 2023; Rudd et al., 2012; Yin et al., 2024). Hence, as Keltner and Haidt (2003) argued, a comprehensive understanding of awe necessitates explaining its capacity to evoke both intensely positive and negative feelings.

However, despite its profound implications for human well-being, the ambivalent nature of awe remains underexplored in affective science. Current research struggles to capture awe's inherent duality, often attempting to categorize it as simply positive or negative (Gordon et al., 2017; Guan et al., 2019; Piff et al., 2015; Takano & Nomura, 2022). This approach overlooks the key characteristic of awe — its capacity to evoke both positive and negative feelings simultaneously, prompting a fundamental reevaluation of our place in the world.

Two key methodological and theoretical bottlenecks hinder progress in understanding awe's ambivalence. Firstly, traditional valence measurement tools, often relying on unidimensional scales

(Russell, 2003), fail to capture the co-occurrence of positive and negative feelings. Both psychometric (An et al., 2017; Briesemeister et al., 2012) and neurobiological (Berridge, 2019; Lammel et al., 2012; Norman et al., 2011; Reynolds & Berridge, 2008) evidence suggest that a multidimensional model, representing positivity and negativity as distinct dimensions, offers a more accurate depiction of the valence system. For example, reward and aversion circuits, while both involving the ventral tegmental area (VTA), receive input from distinct brain regions: the laterodorsal tegmentum for reward and the lateral habenula for aversion (Lammel et al., 2012). This separation of functional neuroanatomy highlights the distinct pathways associated with positive and negative valence. Moreover, recent studies using negative awe-inducing images have demonstrated a higher co-occurrence of opposing valence than images evoking simple happiness or fear (Chaudhury et al., 2022). This finding suggests that traditional unidimensional valence scales may be insufficient for capturing the complex interplay of positive and negative feelings inherent in awe, underscoring the need for more nuanced measurement approaches.

Secondly, a lack of consensus on the neural representation of ambivalent feelings further complicates the study of awe. While the constructivist perspective posits that mixed feelings arise from rapid fluctuations between opposing valence systems (Barrett & Bliss-Moreau, 2009; Russell, 2017), emerging evidence suggests that ambivalent feelings may have distinct representations in the cortex (Man et al., 2017; Vaccaro et al., 2020). Human fMRI studies indicate that the posterior-anterior axis gradient within the right temporoparietal cortex is associated with valence co-occurrence during movie watching (Lettieri et al., 2019). The ventromedial prefrontal and anterior cingulate cortex also exhibit consistent neural patterns for ambivalent feelings (Vaccaro et al., 2024). Thus, it is required to assess how distinct neural representations mixed feelings during awe show for filling the knowledge gap regarding neural basis of ambivalent feelings.

To address these challenges, in this study, we combine the immersive power of 360° virtual reality (VR) with the high temporal resolution of electroencephalography (EEG) and robust deep learning-based dimensionality reduction analysis. VR offers unprecedented ecological validity crucial for awe research compared with conventional image and movie presentation on a screen (Chirico et al., 2016;

Silvia et al., 2015), allowing participants to experience vast and enigmatic stimuli in a more profound and immersive manner than traditional methods (Chirico et al., 2017; Chirico et al., 2018; Kahn & Cargile, 2021; Quesnel & Riecke, 2018). EEG captures the fine-grained temporal dynamics of brain activity, essential for understanding the rapid interplay of affective processes. For instance, the insula cortex synthesizes bodily signals within a time window of about 125ms to create a global affective representation (Picard & Craig, 2009; Vaccaro et al., 2020; Wittmann, 2013), requiring a high temporal resolution to be measured. This approach enables precise tracking of the dynamic neural processes that underpin the complex affective states experienced during VR immersion.

To examine the distinct neural representation of ambivalent feelings during awe, we employed a contrastive learning-based deep learning approach to dimensionality reduction analysis – CEBRA (Schneider et al., 2023). It learns individual- and stimulus-specific latent neural spaces distinguishing EEG samples in terms of affective valence, efficiently exploring the representation of ambivalent feelings in addition to positive or negative ones. While dimensionality reduction offers a powerful way to reveal latent structures in neural data while preserving statistical power (Cunningham & Yu, 2014), traditional methods have struggled to uncover meaningful affect-related representations from non-invasive recordings. This difficulty stems from the highly nonlinear (Aftanas et al., 1998; Berridge, 2019; Viinikainen et al., 2010) and heterogeneous (Čeko et al., 2022; Lettieri et al., 2024) nature of human affect and low signal-to-noise ratio of non-invasive EEG recordings. By leveraging advanced algorithms based on nonlinear and supervised learning, our approach allows us to identify individual-specific latent neural spaces that capture both idiosyncratic valence representations and their commonalities across participants and stimuli. These individualized spaces, unlike conventional hand-crafted features such as frontal alpha asymmetry, provide a more nuanced and precise representation of the dynamic interplay of valence during awe. This integration of VR with EEG, analyzed through deep learning-based dimensionality reduction, forms a cohesive and innovative approach that bridges the gap between the subjective experience of awe and its neural substrates.

This study investigates the ambivalent nature of awe at both behavioral and neural levels, seeking to answer a fundamental question: How is awe, as a complex emotion characterized by the simultaneous experience of positive and negative feelings, represented in human behavior and the brain? To address this overarching question, we explore three specific inquiries: (1) Are awe ratings better predicted by ambivalence-related behavioral metrics than single valence measures? (2) Does ambivalent feeling during awe have a distinct neural representation, separable from positive and negative feelings? (3) Does the distinctiveness of ambivalent feeling representation in the neural space predict the intensity of awe ratings? By answering these questions, we aim to provide a more nuanced and comprehensive understanding of awe's complex nature, revealing its fundamental connection to the interplay of positive and negative experiences.

# Results

## Overview of experimental protocols and behavioral data

Forty-three participants completed all experimental procedures (see **Fig 1a**). They viewed three awe-inducing clips ('SP' for '*Space*', 'CI' for '*City*', and 'MO' for '*Mountain*') and one control clip ('PA' for '*Park*') in a pseudo-random order, each lasting 120 seconds with sound. These clips were designed to evoke diverse awe experiences based on distinct awe-related factors and sensory features (see 'Methods' and **Supplementary Fig 1**). During watching movies, participants' EEG signals were recorded, and they reported their valence dynamics through real-time keypress, indicating neutral, positive, negative, and ambivalent feelings (see **Fig 1b**). They held the key down for the duration of each affective feelings, and these ratings were synchronized with EEG samples. This approach allowed for a direct association between real-time affective response and neurophysiological data. After each video, they rated overall valence, arousal, awe, and motion sickness. Keypress valence ratings were used to calculate the duration of four valence feelings, while the post-trial valence ratings were applied to quantify the intensity of positive, ambivalent, and negative feelings.

## Awe experiences are consistently associated with longer and stronger ambivalent feelings, but not with single-valence ones

We first explored participants' valence dynamics ratings. Valence dynamics showed idiosyncratic temporal patterns (see **Fig 1c**), but they were also significantly intertwined to visual and acoustic features known to predict perceivers' affective responses - e.g., color hue (Dael et al., 2016; Suk & Irtel, 2010), brightness (Kurt et al., 2017), and loudness (Thao et al., 2019)(see **Supplementary Table 1**). These results validate our real-time valence rating paradigm, showing that individual responses are systematically connected to affect-related sensory inputs, despite their idiosyncrasy.

Next, we examined if awe-inducing clips were associated with higher awe and ambivalence-related metrics. Participants reported significantly higher awe ratings for all awe clips than the control (SP vs. PA: Cohen's $d = 1.837$, $P_{FDR} = 1\times10^{-14}$; CI vs. PA: Cohen's $d = 1.493$, $P_{FDR} = 3\times10^{-12}$; MO vs. PA; Cohen's $d = 1.373$, $P_{FDR} = 2\times10^{-11}$; see **Fig 1d**). The three awe clips also led to significantly longer and stronger ambivalent responses than the control (SP vs. PA: Cohen's $d_{duration} = .562$, $P_{duration/FDR} = .001$, Cohen's $d_{intensity} = .451$, $P_{intensity/FDR} = .005$; CI vs. PA: Cohen's $d_{duration} = .514$, $P_{duration/FDR} = .002$, Cohen's $d_{intensity} = .449$, $P_{intensity/FDR} = .005$; MO vs. PA: Cohen's $d_{duration} = .790$, $P_{duration/FDR} = 2\times10^{-5}$, Cohen's $d_{intensity} = .780$, $P_{intensity/FDR} = 2\times10^{-5}$; see **Fig 1e**). Contrarily, behavioral metrics of positive and negative feelings showed no consistently significant differences between conditions, except for the duration of positive feelings and arousal (see **Supplementary Table 2**). These results show the potential of awe experiences to evoke salient and persistent ambivalent feelings compared to simply positive or negative states.

## Ambivalent feelings predict the awe ratings more precisely than single valence metrics

We tested whether ambivalence-related metrics had greater predictive power for awe ratings than other single-valence features using 10 ratings from each clip and five baseline variables (sex, age, baseline positive and negative mood, and dispositional awe). The univariate linear mixed effects model analysis revealed that only the duration and intensity of ambivalent feelings and the intensity of positive feelings showed significant fixed effects (duration of ambivalent feelings: $\beta = .565$, 95% CI = [.033, 1.097], $P = .039$; intensity of ambivalent feelings: $\beta = .220$, 95% CI = [.090, .351], $P = .001$; intensity of positive feelings: $\beta = .094$, 95% CI = [.007, .180], $P = .035$; see **Fig 2a**), confirming the unique contribution of emotional complexity to awe experiences.

To further explore the relationship between these variables, we employed an AutoML pipeline as multivariate analysis allowing for potential nonlinear interactions. Our best model, selected via 5-fold cross-validation, showed superior predictive performance than a linear regression model (see **Fig 2b**). In

this model, ambivalent feelings' duration and intensity ranked higher in feature importance than other valence metrics (see **Fig 2c**). Shapley values indicated a positive association between ambivalence and awe ratings (see **Fig 2d**). The results from both univariate and multivariate analyses provide compelling evidence that the ambivalent feeling is significantly related to stronger awe experiences.

## Overview of electrophysiological analysis

To examine neural representation of ambivalent feelings during awe and its relationship with awe ratings, we analyzed EEG signals and real-time valence keypress. In this analysis, we used data from 27 participants who showed EEG signals with acceptable qualities and responded ambivalent feelings in their keypress for all awe-inducing clips (see **Fig 1a**). These selection criteria were based on the premise that a consistent reporting of ambivalence across stimuli would provide a clearer signal for assessing its neural correlates. Given our interest of ambivalent feelings' latent neural representation during awe and behavioral findings that the duration of ambivalent feelings was the most salient predictor of awe ratings, such sampling was required and expected not to introduce a sampling bias. Included and excluded participants did not differ significantly in key demographic measures or behavioral awe ratings (all *P*s > .05), suggesting that this selection did not introduce systematic bias (see **Supplementary Table 3**). We applied Short-time Fourier Transformation (STFT) to calculate band power across delta (1-4Hz), theta (4-8Hz), alpha (8-14Hz), beta (14-31Hz), and gamma (31-49Hz) frequency. Afterward, we performed dimensionality reduction to extract latent neural spaces from STFT features, using CEBRA (Schneider et al., 2023), a contrastive learning-based approach. CEBRA can capture complex, non-linear relationships in high-dimensional data while emphasizing distinctions between affective states, allowing for more effective isolation of neural representations specific to each state, including ambivalence. Indeed, CEBRA attracted EEG samples with the same valence state while repelling those with different valence states, revealing cohesive clusters for each valence state in the latent neural space (see **Fig 3a**). We

independently trained CEBRA on STFT features of each participant-clip datum to identify individualized and stimulus-specific latent neural spaces.

To assess if these individualized spaces share a common valence representation structure across participants and stimuli, we conducted two pairwise decoding tasks which predicted valence dynamics of other participants or other clips. Next, we tested if ambivalent feelings formed distinct neural clusters and if distinctiveness of these representations could predict awe ratings. Finally, we analyzed which EEG features mainly engaged in distinguishing valence types in the latent neural spaces.

## Aligned latent neural spaces share valence representation architecture across individuals and stimuli

To evaluate the generalizability and optimal dimensionality of individualized latent neural space constructed with CEBRA, we performed pairwise decoding analysis with two tasks ('across-participants' and 'across-clips') and three conditions ('random', 'not-aligned', and 'aligned'). For the 'across-participants' task, a classifier was trained on participant A's latent space embeddings to decode his valence dynamics and was tested on participant B's embeddings for the same clip. In the 'across-clips' task, a classifier trained on latent space embeddings from clip X was tested on clip Y for the same participant (see **Fig 3b**). To facilitate multi-class classification, only six participants' data for three awe-inducing clips were used, which included all valence categories in the keypress for all awe-inducing clips (see 'Methods' for detailed inclusion criteria). We assessed decoding performance under three conditions: classifiers trained on shuffled valence dynamics labels ('random'), on personalized latent space embeddings without alignment ('not-aligned'), and on latent embeddings aligned with test embeddings through canonical correlation analysis ('aligned'). Alignment helps identify commonalities in individual neural spaces and enhances the interpretability of shared neural representations (Gallego et al., 2020; Safaie et al., 2023). Assuming that the optimal latent space is consistently dimensional across individuals and clips, achieving comparable predictive performance with the simplest dimensionality (see 'Methods'

for rationales), we selected a 7D space for the optimal CEBRA-driven latent neural spaces (see **Fig 3c**). For a comparative baseline model, we performed the same analyses with principal component analysis (PCA)-driven latent spaces, and a 6D space was chosen for the optimal PCA-based embedding space (see **Supplementary Fig 2**).

We then compared decoding performances of the latent neural space across participants and sensory stimuli among CEBRA, PCA, and frontal alpha asymmetry (FAA), a traditional index of affective valence (see **Fig 3d**). In the 'across-participants' task, CEBRA-based classifiers performed significantly better than random chance (aligned vs. random: Cohen's $d = 1.122$, $P_{FDR} = 5\times10^{-17}$; not-aligned vs. random: Cohen's $d = .249$, $P_{FDR} = .020$). Aligned embeddings outperformed not-aligned ones (aligned vs. not-aligned: Cohen's $d = .548$, $P_{FDR} = 2\times10^{-6}$). Also, they provided more accurate predictions than PCA-based (Cohen's $d = 1.093$, $P_{FDR} = 2\times10^{-16}$) and FAA-based latent space embeddings (Cohen's $d = 1.115$, $P_{FDR} = 6\times10^{-17}$) in aligned conditions, but not in random and not-aligned conditions (see **Supplementary Table 4**).

In the 'across-clips' task, only aligned CEBRA embeddings showed significant predictive power for valence dynamics across clips (aligned vs. random: Cohen's $d = .501$, $P_{FDR} = .015$; not-aligned vs. random: Cohen's $d = .103$, $P_{FDR} = .540$). Aligned embeddings outperformed not-aligned ones, although its significance did not reach a threshold (aligned vs. not-aligned: Cohen's $d = .283$, $P_{FDR} = .147$). Aligned CEBRA embeddings more accurately predicted valence dynamics than PCA-based (Cohen's $d = .553$, $P_{FDR} = .006$) and FAA-based embeddings (Cohen's $d = .807$, $P_{FDR} = 8\times10^{-5}$), but not in random and not-aligned conditions (see **Supplementary Table 4**).

These findings suggest that aligned CEBRA-based latent neural spaces capture a shared architecture of valence representations, including ambivalent feelings, across individuals and stimuli, which could not be fully identified by conventional approaches such as PCA or FAA.

**The more distinctively ambivalent feelings are represented in the cortices, the more saliently individuals report awe**

Using not-aligned latent neural spaces, we explored whether ambivalent feelings have distinct neural representations separable from representations of other single valence states. To quantify the cohesiveness of neural representations for each valence state, we calculated the silhouette coefficient of each valence-specific cluster in the latent neural spaces. A higher silhouette coefficient indicates that the corresponding valence type is represented as a more cohesive cluster in the latent space. Since the silhouette coefficient can be overestimated in supervised latent spaces as CEBRA-driven ones, we evaluated its statistical significance using 1,000 times permutation test (see **Fig 4a**).

We observed that the silhouette coefficients of ambivalent clusters varied widely across participants (see **Fig 4b**), while single valence states showed relatively consistent significant silhouette coefficients (see **Supplementary Fig 3**). This individual variability in the ambivalence-related representations suggests that the integration of positive and negative experiences during awe may manifest differently across individuals. Concerned about potential confounding effects of cluster size on silhouette coefficients significance, we tested the correlation between the number of ambivalence-labeled EEG samples in the cluster (i.e., duration of ambivalent feelings) and the *P* values of silhouette coefficients for each clip. No significant correlations were found (SP: Pearson's $R$ = .221, 95% CI = [-.174, .554], Pearson's $P$ = .268; CI: Pearson's $R$ = .198, 95% CI = [-.197, .538], $P$ = .323; MO: Pearson's $R$ = .126, 95% CI = [-.267, .483], $P$ = .531), suggesting that the observed variability in silhouette coefficients for ambivalence is not merely due to differences in the duration of reported ambivalence.

To directly test the relationship between ambivalence-related neural representations and awe experiences, we then examined how spatial properties of valence clusters in the latent neural space relate to awe ratings. We developed a new metric called 'cortical distinctiveness' for cluster $k$, $\phi_k$. It quantifies how distinguishable representations cluster $k$ displays from other valence clusters by averaging cosine

distance with other clusters (see **Fig 4c**). Higher $\phi_k$ indicates that the neural representation of the corresponding valence cluster *k* is more spatially separated from other valence representations within the individual's latent neural space. Using this metric, we constructed four linear mixed-effect models including $\phi$ value of each valence cluster and two random intercepts of participants and clips. In these models, only $\phi_{ambivalent}$ significantly predicted awe intensity ($\beta$ = .817, 95% CI = [.307, 1.326], *P* = .003; see **Fig 4d**). The predictive power of $\phi_{ambivalent}$ persisted even when using medoid cluster distance for calculating this metric ($\beta$ = .803, 95% CI = [.152, 1.455], *P* = .018; see **Fig 4d**), demonstrating the robustness of the findings to the specific method of calculating cortical distinctiveness. As a control analysis, we conducted the same analysis with FAA values and found that FAA did not predict awe ratings significantly ($\beta$ = .000, 95% CI = [-.004, .003], *P* = .789), highlighting the specificity of the relationship between awe and the cortical representation of ambivalence.

In AutoML-based multivariate analysis, adding $\phi_{ambivalent}$ metrics to a model predicting awe ratings with 14 behavioral variables improved the adjusted $R^2$ value by 7.2% (see **Fig 4e**). This supports the significant contribution of ambivalence representation distinctiveness to the subjective experience of awe. Furthermore, $\phi_{ambivalent}$ demonstrated higher predictive power than other behavioral metrics in the predictive model (see **Fig 4f**), suggesting that the distinctiveness of cortical representations for ambivalent feelings may be a key factor driving individual differences in awe experiences.

These results indicate that individual differences in the distinctiveness of cortical representation of ambivalent feelings specifically account for the awe experience.

## The delta oscillation in the frontal areas mainly distinguishes different valence representation in the latent neural space

Lastly, we investigated which EEG features were important for contrasting different valence states when CEBRA constructed latent neural spaces. We utilized explainable AI technique specialized for time-series data, Dynamask (Crabbé & Van Der Schaar, 2021). It estimates a feature's importance by

observing how perturbing it at a specific time point alters the CEBRA model's latent space embedding. Specifically, Dynamask learns a weight matrix by perturbing feature values with nearby values based on a perturbation weight, producing a pseudo-embedding with a high mean-squared error from the original CEBRA embedding (see **Fig 5a**; see 'Methods' for details). If a perturbation weight is close to 1, it indicates that the feature is crucial for the CEBRA model in learning the latent space embedding (i.e., contrasting the valence category at that time point from other valences). We independently learned the perturbation weight matrix for each participant's clip data. By aligning the temporal axis of the weight matrix with participants' valence dynamics ratings, we calculated the mean perturbation weight of each feature for every valence category. To calculate mean perturbation weights for each valence label, we used only data displaying significant silhouette coefficients for the corresponding valence label (e.g., in the case of ambivalent feelings, data colored in **Fig 4b**).

Firstly, we confirmed that all Dynamask models reached plateau in learning performance for learning the weight matrix (see **Supplementary Fig 4**). Delta band power features had higher mean perturbation weights for ambivalent states than other frequency features (see **Fig 5b**), suggesting that delta oscillations may play a crucial role in differentiating ambivalent states from other valences in the brain. Within the delta band, frontal channels had relatively higher weights than channels in other areas (see **Fig 5c**), further implicating frontal delta oscillations in the processing and representation of ambivalent feelings. The same analysis for positive and negative states also showed that delta band power in the frontal channels consistently exhibited higher mean perturbation weights (see **Supplementary Fig 5**), indicating that frontal delta oscillations might be broadly involved in valence processing beyond ambivalent feelings.

To verify the importance of delta oscillation in valence representation, we conducted a post-hoc analysis using a hidden Markov model (HMM). We hypothesized that state changes in delta band power might temporally align with transitions in self-reported valence. We calculated the match rate between event boundaries in delta oscillations and key-pressed valence sequences and assessed its significance through a 1,000 times permutation test – calculating $P_{\text{perm}}$ following the previous literature (Vaccaro et al.,

2024). Neural boundaries from delta-related features significantly matched transitions in participants' valence ratings above random chance (match rate = 53.6%; $P_{perm}$ = .047), providing further support for a temporal link between delta oscillations and the valence dynamics. Among the other frequency band power features, only beta power showed a significant match rate (see **Fig 5d**). These findings suggest that delta oscillation in the frontal channels is crucial for distinguishing different valence states in the latent neural space.

# Discussion

Our study reveals that awe is fundamentally characterized by ambivalence, challenging traditional unidimensional models of valence. By combining immersive virtual reality, EEG, and advanced machine learning, we demonstrate that awe experiences are marked by ambivalent feelings at both behavioral and neural levels. Crucially, the 'cortical distinctiveness' of ambivalence representation, captured through individualized latent neural spaces, significantly predicts the subjective intensity of awe. This novel metric quantifies how distinguishable ambivalent neural representations are from other valence states. Our findings establish a direct link between the neural processing of mixed emotions and the phenomenological experience of awe, offering new insights into how the brain integrates seemingly contradictory affective states. This approach not only illuminates the intricate nature of awe but also provides a novel framework for understanding and quantifying the neural basis of emotional complexity.

The significant association of the duration and intensity of ambivalent feelings with awe intensity further supports our hypothesis that awe is inherently an ambivalent experience. This finding aligns with previous cross-cultural research; for example, Western populations rated threat-awe-inducing images as having stronger ambivalence than stimuli evoking happiness or fear (Chaudhury et al., 2022). This consistency across cultures and stimuli suggests that the ambivalent nature of awe may be a fundamental feature of this emotion.

Our investigation of neural representations of ambivalence during awe revealed that aligned embeddings from individualized latent neural spaces successfully predicted valence dynamics across participants and stimuli, outperforming traditional methods like PCA and FAA. This superiority aligns with known limitations of FAA in valence indexing (Gable & Harmon-Jones, 2010; Harmon-Jones & Gable, 2018; Honk & Schutter, 2006; Wacker et al., 2003) and extends beyond PCA's capabilities in detecting generalizable neural trajectories, such as motor control (Gallego et al., 2020; Safaie et al., 2023). Our deep neural net approach effectively captured idiosyncratic and shared valence representations, reflecting nonlinear brain-valence relationships (Aftanas et al., 1998; Berridge, 2019;

Viinikainen et al., 2010) at the individual level. Notably, our method successfully disentangled affective neural representations from varied sensory inputs, demonstrating its robustness across diverse stimuli.

Our latent neural spaces revealed distinct cortical representations for ambivalent feelings, consistent with studies indicating that ambivalent feelings have unique neural patterns in cortical regions (Lettieri et al., 2019; Man et al., 2017; Vaccaro et al., 2020; Vaccaro et al., 2024). This finding seems to directly challenge the constructivists' view that ambivalent feelings lack specific neural correlates (Barrett & Bliss-Moreau, 2009; Russell, 2017). Nevertheless, our findings of individual variability in these neural representations may integrate these conflicting arguments into a single framework. We observed individual variability in how neural representations of ambivalent feelings are cohesive in the latent space, which paradoxically supports the constructivist perspective that psychological factors (e.g., attentional focus or emotional granularity) influence how mixed emotions are experienced and processed (Hoemann et al., 2017). This dual finding - distinct neural patterns alongside individual variations - offers a novel framework for understanding complex emotions. Our deep neural network (CEBRA) approach, by capturing both aspects, potentially reconciles these divergent views in affective neuroscience, demonstrating the nuanced nature of ambivalence processing in the brain.

A key finding of our study is that the cortical distinctiveness of ambivalent states, quantified by our novel metric, strongly predicts self-reported awe intensity, outperforming traditional behavioral measures. Our approach, which considers the geometrical characteristics of neural representations, revealed relationships between ambivalent feelings and awe intensity that were overlooked by previous studies focused solely on activation levels in specific brain regions (Guan et al., 2019; Hu et al., 2017; Takano & Nomura, 2022). This underscores the value of our analytical method in uncovering the nuanced neural underpinnings of complex emotional experiences like awe.

The positive correlation between cortical distinctiveness of ambivalence and awe intensity aligns with the concept of "holistic meaning-making" in awe experiences (Bonner & Friedman, 2011; Dai et al., 2022; Ihm et al., 2019; Sawada et al., 2024). Awe is often characterized by encounters with stimuli that challenge our existing cognitive schemas, requiring us to accommodate new information and generate

new understandings. This process may involve integrating initially evoked negative feelings, perhaps triggered by the vastness or novelty of the stimuli, with the positive feelings that emerge from successful comprehension and meaning-making. Likewise, awe is often described as a 'self-transcendent' experience, where two conflicting feelings – 'self-diminishment' and 'connectedness' are harmonized (Yaden et al., 2016; Yaden et al., 2019). Awe is thus based on integrating opposing feelings, creating an emotion beyond mere positivity and negativity. The distinctiveness of ambivalence-related neural representations observed in our study may reflect the consequence of this dynamic integration process, suggesting that the brain is actively processing both positive and negative aspects of the experience, ultimately contributing to a more personally meaningful and intense experience of awe.

Our analyses of Dynamask (Crabbé & Van Der Schaar, 2021) and HMM (Kumar et al., 2021) consistently highlighted the significance of frontal delta oscillations in distinguishing valence states. This aligns with previous neuroimaging findings on the role of frontal cortices such as the orbitofrontal cortex and anterior cingulate cortex in processing conflicting affective information (Levens & Phelps, 2010; Rolls & Grabenhorst, 2008; Simmons et al., 2006) and consistent activity patterns when individuals feel ambivalence during movie watching (Vaccaro et al., 2024). The prominence of frontal delta activity suggests its crucial role in integrating positive and negative feelings during awe.

These frequency band, encompassing both delta and low-theta ranges (below 5-6Hz), has been consistently linked to emotional processing and regulation in the literature. Frontal delta power increases during emotional memory retrieval (Brenner et al., 2014; Hutchison & Rathore, 2015; Nishida et al., 2009; Sopp et al., 2017), while the 1-3Hz range specifically decodes valence ratings to images(Shen et al., 2020). Theta burst stimulation, enhancing 5Hz band power in the dorsolateral prefrontal cortex, improves emotion recognition for ambiguous lexical and facial stimuli (Dumitru et al., 2020; Moulier et al., 2021), further supporting the role of low-frequency bands in emotional processing. Moreover, emotion regulation has been linked to 4Hz band power in frontal areas (Cavanagh & Shackman, 2015; Ertl et al., 2013); while individuals with impaired emotion regulation has been linked to decreased power in frequencies below 5Hz (Jiang et al., 2022). Indeed, in our HMM analysis, the synchrony of delta and beta

features, which was aligned with real-time valence dynamics, has also been implicated in efficient emotion and stress regulation (Brooker et al., 2021; Myruski et al., 2022; Phelps et al., 2016; Putman et al., 2012). These findings collectively suggest frontal delta oscillations' critical involvement in constructing coherent emotional responses to the complex, often conflicting feelings elicited by awe-inspiring experiences, supporting our observations on the neural dynamics of ambivalence during awe.

Our findings, coupled with existing literature on frontal delta power and emotion regulation, suggest a potential role for emotion regulation processes in mediating ambivalent feelings during awe experiences. Ambivalent feelings are often thought to stem from the reappraisal process (Vaccaro et al., 2020; Van Tilburg et al., 2018), where initial affective responses are re-evaluated through memory retrieval or knowledge associated with opposite valences. While we did not directly measure emotion regulation strategies in this study, previous studies have shown that reappraisal predicts awe ratings for recalled memories (Chirico et al., 2024; Chirico et al., 2021). Additionally, studies on affect dynamics show emotion regulation types correlate with the duration of negative feelings triggered by stimuli (Van Mechelen et al., 2013; Verduyn et al., 2009; Verduyn et al., 2011), but not the intensity, of negative feelings (Brans & Verduyn, 2014). These findings raise the intriguing possibility that the extended duration of ambivalent feelings we observed during awe experiences might originate from the engagement of implicit reappraisal processes. To further explore this hypothesis, future research could employ cognitive task analysis or think-aloud protocols during VR experiences. Such approaches would allow for direct assessment of reappraisal strategies during awe, illuminating how individual emotion regulation patterns influence ambivalent feelings and awe at both behavioral and neural levels.

Our study has limitations. We acknowledge that our sample comprised only Asian participants, potentially overestimating the role of ambivalent feelings in awe. Western individuals report fewer ambivalent feelings in awe experiences compared to Asians (Nakayama et al., 2020; Stellar et al., 2024). While previous research has found that Western individuals also report higher ambivalence for awe-inducing stimuli than single-valence stimuli (Chaudhury et al., 2022), future studies should investigate cross-cultural differences in the neural representation of ambivalence during awe. Additionally, using

real-time valence ratings through keypresses might have inadvertently influenced participants' emotional experiences (Larsen & Fredrickson, 1999). However, recent naturalistic affective science studies have validated and applied moment-by-moment emotion ratings during watching movies (Belfi et al., 2019; Lettieri et al., 2019; Vaccaro et al., 2024) or listening to music (McClay et al., 2023). Finally, our contrastive learning approach does not directly address how different valence types might share common neural substrates. Investigating these shared representations, particularly in brain regions associated with both pleasure and pain (Lee et al., 2024), could provide a more comprehensive understanding of the neural dynamics underlying ambivalence.

In conclusion, our study provides insights into the ambivalent nature of awe and its neural underpinnings. By demonstrating the link between ambivalent feelings, their distinct cortical representation, and the subjective intensity of awe, our findings advance the field of affective neuroscience and contribute to a more nuanced understanding of this complex and powerful emotion. This work lays the groundwork for further investigations into the cognitive and neural processes that underpin awe, paving the way for a more comprehensive understanding of its potential benefits for mental well-being.

# Methods

## Participants

We recruited 50 healthy young adult Koreans enrolled in psychology courses at Seoul National University for this study. Participants were excluded if they met any of the following criteria: (1) currently taking psychiatric medication, (2) history of psychiatric treatment, (3) left-handedness, (4) vestibular neuritis or balance disorder, (5) visual acuity before correction less than 0.2, (6) non-Korean native speakers, and (7) consumption of alcohol or use of hair rinse 24 hours before the experiment. Following these exclusion criteria, data from 43 participants were completely collected in the analyses (23 females; $M_{age}$ = 20.2 years, $SD_{age}$ = 1.7 years). Seven participants were excluded due to technical issues ($N$ = 3), discontinuance due to motion sickness ($N$ = 2), and lack of fidelity in online valence ratings (i.e., showing accuracy below 75% in the training session for valence keypress; $N$ = 2). The sample size was chosen to be comparable to recent naturalistic affective neuroscience studies – e.g., (Lettieri et al., 2019; Meer et al., 2020; Vaccaro et al., 2024). Participants provided written informed consent before the experiment, and all procedures were approved by the Institutional Review Board of Seoul National University.

## Experimental paradigm

### VR clip design

We collaborated with a professional filmmaker to design four audio-integrated 360° immersive videos using Unreal Engine (version 5.03). Each video lasted 120 seconds. Three of the videos were designed to evoke awe: *Space* (Meer et al.), *City* (CI), and *Mountain* (MO), while the other one, *Park* (PA), served as a control stimulus, designed not to include any awe-related factors. To investigate whether ambivalence is consistently observed in various awe experiences, we varied (1) the clip themes,

(2) the VR cues of awe-related two dimensions – perceived vastness and the need for accommodation, and (3) the perceptual features across the awe-inducing videos.

Firstly, considering that awe is most intensively triggered by massive landscapes (Chirico et al., 2018; Keltner & Haidt, 2003; Shiota et al., 2007; Yaden et al., 2019), we differentiated the semantic theme of the scenery: SP featured supernatural landscapes (i.e., black holes and planets in space), CI showcased urban landscapes (i.e., cityscape viewed from the top of skyscrapers), and MO depicted natural one (i.e., mountain scenery).

Secondly, following the qualitative framework of the ref(Chirico et al., 2018)to design VR videos that effectively elicit awe, we aimed to represent the two key dimensions of awe, perceived vastness and a need for accommodation, through different cues in each video. Each video was designed so that perceivers would first experience vastness during the initial 60 seconds and then feel a need for accommodation in the latter 60 seconds. For example, in SP, participants watched a giant black hole approaching, consuming everything and ultimately drawing them in, followed by the sudden appearance of Earth from space. Different sub-components were applied to realize each dimension across clips. Perceived vastness can be induced through perceptual (e.g., 'width' and 'height') and conceptual cues (e.g., 'complexity')(Chirico et al., 2017; Chirico et al., 2018). MO was designed to evoke vastness through perceptual width, CI through height, and SP through conceptual complexity. For the need for accommodation, we introduced surprise cues in each video around the 60-second mark as a trigger of accommodation (Chirico et al., 2017; Chirico et al., 2018), tailored to the context of each video to ensure that the cause of surprise did not overlap across videos. The design of three awe-inducing clips is summarized in **Supplementary Fig 1a**.

Lastly, to prevent awe from being driven by specific perceptual factors, we intentionally composed the three awe videos with different audiovisual information. We synchronized visual content with ambient sounds using open-source audio samples from Freesound (https://freesound.org) and GarageBand (version 10.4.6). To verify our design, we calculated three perceptual features known to predict perceivers' emotional responses – color hue (Dael et al., 2016; Suk & Irtel, 2010), brightness

(Kurt et al., 2017), and loudness (Thao et al., 2019) – every second for each stimulus and visualized their time-course dynamics. We qualitatively confirmed that each video exhibited very different temporal dynamics for all features (see **Supplementary Fig 1b**).

To validate the awe elicitation, we conducted a preliminary study with 28 independent young adult Koreans (five females; $M_{age}$ = 20.2 years, $SD_{age}$ = 1.9 years), who rated awe intensity using the Awe Experience Scale (AWES)(Yaden et al., 2019) after watching each clip in VR. Participants reported significantly higher awe ratings for three awe clips than the control clip, with large effect sizes (SP vs. PA: Cohen's $d$ = 2.466, $P_{FDR} = 8 \times 10^{-13}$; CI vs. PA: Cohen's $d$ = 2.193, $P_{FDR} = 6 \times 10^{-12}$; MO vs. PA; Cohen's $d$ = 1.52, $P_{FDR} = 1 \times 10^{-6}$).

## Baseline self-report

Before the experiment, participants provided information on their sex, age, baseline mood states, and dispositional traits of feeling awe. Baseline mood states were assessed using the Korean version of the Positive and Negative Affect Schedule (PANAS) validated by (Lim et al., 2010). No participants showed exceptional mood states beyond 1.5 × interquartile range ($M_{positive}$ = 33.581, $SD_{positive}$ = 6.284; $M_{negative}$ = 23.140; $SD_{negative}$ = 6.331). Dispositional awe was measured using the awe-related items from the Dispositional Positive Emotions Scale (DPES)(Shiota et al., 2006) translated into Korean ($M$ = 29.628, $SD$ = 6.626). The Korean-translated DPES items showed acceptable reliability (Cronbach's a = .827) but mixed results in validity (model fit of the original factor structure: CFI = .946, RMSEA = .127).

## Procedures

Participants sat on a sofa in a noise-isolated room and wore an Enobio 20 EEG device (Neuroelectrics) and a Quest 2 VR headset (Oculus). After checking EEG signal quality, the experiment

proceeded as follows: baseline EEG recording, keypress training, VR watching task, post-trial measurement, and a break. Firstly, participants' resting EEG signals were recorded for 120 seconds with their eyes closed (baseline recording). These resting signals were used to normalize signals recorded during the movie-watching trials. Secondly, they practiced real-time valence keypress reporting (keypress training). Participants were explicitly asked to report their valence feelings in real-time keypress with the following auditory instruction: *"While watching the video, please report your affective state by pressing a number pad: '1' for positive, '2' for ambivalent - feeling positive and negative feelings at the same time, and '3' for negative feelings. If you do not feel any affective feelings, please do not press anything. If a specific feeling persists, continue to press and hold the corresponding key. It is important to report your subjective reactions rather than which emotions the video intends to elicit"*. Participants practiced this for 60 seconds using sentences describing affective responses to prevent confusion about which key they should press. Then, participants watched four VR clips in a pseudorandom order, reporting their valence using online keypress (VR watching task). After each trial, they reported awe intensity, overall valence, arousal, and motion sickness using controllers (post-trial measurement). Awe intensity was measured by the Korean-translated AWES (Yaden et al., 2019), valence by Evaluative Space Grid (Larsen et al., 2009), arousal by a conventional 9-point Likert scale (Bradley & Lang, 1994), and motion sickness by a single 7-point Likert scale item. The Korean-translated AWES demonstrated acceptable psychometric properties (Cronbach's a = .928; model fit of the original factor structure: CFI = .881, RMSEA = .079). Participants took a 30-second break with their eyes closed after each trial (break).

## EEG data acquisition

### EEG recording and preprocessing

We recorded EEG signals using 19 dry electrodes: AF3, AF4, F7, F3, Fz, F4, F8, FC5, FC6, C3, Cz, C4, P7, P3, Pz, P4, P8, O1 and O2 with Neuroelectrics Enobio 20 based on the international 10-20

system. Ground and reference electrodes were attached to the right earlobe. The embedded software in the Enobio system assessed signal quality using three levels: good, medium, and bad. We ensured that no electrodes displayed bad signals before starting the signal acquisition. We adopted the automated preprocessing pipeline validated by (Delorme, 2023). EEG signals for each trial were time-locked to the initiation of the video, excluding the last three seconds to avoid end-of-task effects (e.g., loss of attention or emotional confounding). High-pass filtering above 0.5 Hz and Artifact Subspace Reconstruction were performed. Unlike the original pipeline, we used interpolation to maintain consistent recording lengths across participants and trials rather than excluding time windows with poor signal quality. We conducted independent component analysis-based artifact rejection to remove noise components, such as eye or head movements, with over 90% probability. The preprocessed signal for each trial was normalized by subtracting the average resting signal value for each channel. All preprocessing was performed using the "EEGLAB" plugin (Delorme & Makeig, 2004) in MATLAB (version 2021a).

## Short time/Fast Fourier transform

With preprocessed and normalized EEG signals, we performed short time Fourier transform (STFT) and fast Fourier transform (FFT) to calculate the spectral power of five frequency bands for each channel: delta (1-4Hz), theta (4-8Hz), alpha (8-14Hz), beta (14-31Hz), and gamma (31-49Hz). For STFT, a Hanning window with a 500-sample window size (i.e., 1 sec) and a 250-sample hop size was applied. Participants' valence keypresses were embedded as event markers in EEG signals, categorizing EEG samples into one of four valence categories. The valence label of each 500-sample window after STFT was defined as the mode of the corresponding samples' valence labels. FFT was also performed to calculate overall spectral powers marginalized across the whole time series. Using the Welch method, we calculated the power spectral density for each EEG channel and then integrated it over the specific frequency range described above to determine the band power. For relative band power, we normalized

the power within each band by the total power across all frequencies. The "scipy" package (Virtanen et al., 2020) in Python (version 3.8) was used for STFT and FFT.

## Behavioral analysis

### Univariate linear mixed-effects model analysis

Using data from 43 participants, we first assessed the univariate association between AWES ratings and 14 behavioral features measured before, during, and after each trial: sex, age, PANAS positive score, PANAS negative score, DPES awe score (before trial), duration of positive, ambivalent, negative, and neutral feelings (during trial), arousal, motion sickness score, and intensity of positive, ambivalent, and negative feelings (after trial). The duration of each valence type was calculated as the ratio of keypresses for that valence type to the total running time of each clip. Intensity was calculated based on the Evaluative Space Grid responses: positivity (x-axis value), negativity (y-axis value), and ambivalence (minimum value between positivity and negativity, following previous literature (Berrios et al., 2015; Chaudhury et al., 2022; Ersner-Hershfield et al., 2008). Firstly, we performed two-sided paired t-tests to examine statistical differences in AWES scores, duration, and intensity of each valence type, and arousal between the three awe clips and the control clip at $P_{FDR} < .05$. Next, to evaluate the explanatory power of the 14 metrics, we fit linear mixed effect models with each regressor and two random intercepts for participants and clips using the "lmerTest" package (Kuznetsova et al., 2015). Assumptions of normality were examined using the "DHARMa" package (Hartig, 2018). We confirmed that the distribution of residuals did not significantly deviate from a normal distribution using the Kolmogorov-Smirnov test (all $P$s > .05). All variables were standardized before the statistical analysis. All statistical analyses were conducted in R studio (version 2023.03.1+446).

Multivariate machine learning analysis

Next, we conducted machine learning-based predictive modeling with 14 behavioral variables for AWES scores as multivariate analysis, considering potential non-linear interactions among features. Using "h2o" package (LeDell & Poirier, 2020), we split the dataset into training and test sets with a 4:1 ratio and conducted 5-fold cross-validation. A total of 22 models were constructed, and we selected the best model based on the lowest fold-averaged test RMSE value. In the model selection, models that did not provide feature importance information (e.g., stacked ensemble models) were excluded for interpretability. As a result, the gradient boost model (GBM) was chosen as the best model. Its predictive performance was compared to a baseline ridge linear regression model without any interaction terms for metrics: RMSE, MAE, MSE, and $R^2$. To identify the most influential features, we calculated feature importance and shapley values for each variable. All machine learning analyses were performed in R studio (version 2023.03.1+446).

## Electrophysiological analysis

Construction of individualized latent neural space for each clip

Among 43 participants in the behavioral analysis, we excluded 16 individuals due to poor-quality EEG signals even after preprocessing ($N = 6$) and lack of ambivalent keypresses in at least one trial (i.e., who reported ambivalent feelings for less than 5% of the total duration across all videos; $N = 10$). The quality of preprocessed signals was visually inspected. The primary objective of the electrophysiological analysis was to investigate whether ambivalent feelings during awe have distinct cortical representations and to evaluate their relationship with the intensity of awe. Given that our behavioral analysis identified the duration of ambivalent feelings as the most salient predictor of awe ratings, such exclusion is inevitable and expected not to introduce sampling bias regarding the measurement of awe ratings.

Using 27 participants' EEG signals and valence keypress in the three awe clip trials, we constructed a latent neural space for each individual and clip (i.e., a total of 81 latent spaces) using the "CEBRA" package (Schneider et al., 2023). CEBRA employs supervised contrastive learning to extract latent embeddings from the input data (i.e., STFT-processed EEG signals here), maximizing the attraction of EEG samples with the same valence labels and repelling those with different labels. As the dimensionality of latent spaces was elusive, we fitted CEBRA models for each participant-clip pair across dimensions ranging from one to nine. The following hyperparameters were applied: batch_size = length of STFT EEG signals, model_architecture = 'offset-10 model', number_of_hidden_units = 38, learning_rate = .001, the number_of_iterations = 500, and hybrid = False.

## Validation of latent neural spaces with predictive tasks

To test whether the individualized latent spaces hold significant information generalizable across different individuals and clips, we conducted a pairwise prediction task using 2 tasks ´ 3 conditions design. For the 'across participants' task, a classifier trained on Participant A's latent neural space embeddings and valence labels was used to predict the valence keypress of Participant B's embeddings for the same clip. For the 'across clip' task, a classifier trained on clip X's embeddings and labeled valence types was used to predict the valence of clip Y's embeddings within the same participant. Predictions were evaluated under three conditions: (1) a baseline null test with shuffled training valence keypress labels ('random'), (2) prediction using personalized latent neural embeddings without any alignment between train and test embeddings ('not-aligned'), and (3) prediction using aligned embeddings between train and test embeddings ('aligned'). Neural alignment motivates the exploration of commonality among individual latent neural spaces (Gallego et al., 2020; Safaie et al., 2023). Here, canonical correlation analysis (CCA) between train and test embeddings was employed using the "sklearn" package (Pedregosa et al., 2011).

In these prediction tasks, six participants who pressed all four valence labels for at least 5% of the total duration across three clips were selected to facilitate the multi-label classification. A k-nearest neighbors (kNN) classifier with a neighborhood parameter of 15 was used (i.e., the nearest odd number to the square root of the input embedding length following the conventional heuristic). Considering the imbalance in valence keypress labels, prediction performance was evaluated using the weighted F1 scores. A two-sided paired t-test was conducted to compare predictive performances across the three conditions $P_{FDR} < .05$. Construction of latent neural spaces and predictive analyses were performed in Python (version 3.8).

## Dimensionality selection for the latent neural spaces

We selected the optimal dimensionality of latent neural spaces based on two assumptions: (1) The space shares the same dimensionality across individuals and clips. (2) The space displays as good as predictive performances across individuals and clips with other neural spaces based on higher dimensionality, even with fewer dimensions. The second assumption was based on the idea that generalizability, as reflected in predictive performance, should be the criterion for determining the canonical dimensions in dimensionality reduction techniques (Cunningham & Yu, 2014). However, it also acknowledges that such predictive performance may be overestimated as the number of dimensions increases (Cunningham & Yu, 2014; Diaconis & Freedman, 1984). We calculated each dimensionality's mean weighted F1 score n predictive tasks and performed hierarchical clustering analysis to group dimensions with similar prediction performance. The number of clusters was chosen for the highest silhouette coefficients, and the lowest dimension in the highest-performing cluster was selected. Dimensions 6, 7, 8, and 9 formed the high-performance cluster for the across-participant task, and dimensions 7, 8, and 9 for the across-clip task. Thus, the 7D space condition was chosen as the canonical dimensionality. Hierarchical clustering analysis was conducted using the "cluster" package (Maechler et al., 2013) in R studio (version 2023.03.1+446).

## Comparing CEBRA-, PCA-, and FAA-driven latent space embeddings

For a fair comparison of the predictive power of our CEBRA-based latent neural embeddings, we compared its performance with PCA and FAA-driven embeddings. First, to extract PCA embeddings, we input the STFT-processed EEG features and computed latent embeddings with dimensions ranging from one to nine. We identified the 6-dimensional embedding as the optimal latent space due to its modest predictive performance with the fewest dimensions. Second, FAA embeddings were defined as the difference in alpha band power between the F4 and F3 channels for each timepoint in the STFT-featured EEG sequence. These two channels were selected based on previous studies (Brzezicka et al., 2017; Quaedflieg et al., 2016; Van Der Vinne et al., 2017). The prediction tests with the 2 × 3 design were conducted using these three types of embeddings. Test performances under the three conditions – random, not aligned, and aligned – were compared across CEBRA, PCA, and FAA-based embeddings. Additionally, within the CEBRA embeddings, performances were compared across the different conditions. To test the statistical differences in weighted F1 scores, we conducted a two-sided paired t-test at $P_{FDR} < .05$. PCA was performed using the "sklearn" package (Pedregosa et al., 2011) in Python (version 3.8).

## Assessing significance of cortical valence representation

Using the chosen CEBRA-based 7D latent space embeddings, we measured the segregation of ambivalence-related EEG samples from other valence samples using silhouette coefficients. While InfoNCE loss value could quantify the contrast performance of CEBRA as well, it lacks scaling and valence type-specific calculations, so we used silhouette coefficients instead. We computed the average silhouette coefficient for ambivalence-labeled EEG samples for each participant-clip latent space. Since

silhouette coefficients can be overestimated for clusters in the latent space constructed in a supervised manner, we assessed its statistical significance through a permutation test. We randomly shuffled valence keypresses and trained the CEBRA model with the original STFT-featured EEG signals and permuted valence sequence based on the identical hyperparameter set. Average silhouette coefficients of ambivalence-labeled EEG samples were calculated from the trained pseudo-embeddings, and we obtained its null distribution by repeating this 1,000 times. The *P*-value calculated from the permutation test, $P_{\text{perm}}$, is formulated as follows:

$$P_{perm} = \frac{N_{s^*>s}}{N_{s^*}} \tag{1}$$

where $s*$ and $s$ are average silhouette coefficients calculated from permuted and original latent embeddings, and $N_{s*}$ is the number of $s*$ in the null distribution. All $P_{\text{perm}}$ values were FDR corrected. We performed the same analysis with positive and negative valence clusters in the latent neural space.

## Quantifying 'cortical distinctiveness' of each valence in the latent neural space

We developed a metric called 'cortical distinctiveness', $\phi_k$, indicating how distinguishable the reference cluster $k$ is from the other valence clusters in the latent valence-cortical space. $\phi_k$ is defined as:

$$\phi_k = \frac{1}{N}\sum_{i=1}^{N} d(k,c_i) \tag{2}$$

where $N$ is the number of other clusters, $c_i$ is the $i$-th valence cluster, and $d(k, c_i)$ is the cosine distance between the cluster $k$ and $c_i$. We applied cosine distance as a cluster distance metric instead of other conventional metrics (e.g., Euclidean distance), considering that latent CEBRA embeddings are distributed on the hypersphere space. We initially measured $d(k, c_i)$ based on the average cluster distance. The average distance between the cluster $k$ and $c_i$ is calculated as the mean cosine distance between each point in $k$ to every point in $c_i$ as follows:

$$d(k,c_i) = \frac{1}{n_k n_{c_i}} \sum_{p=1}^{n_k} \sum_{q=1}^{n_{c_i}} \left( 1 - \frac{\overrightarrow{x_p} \cdot \overrightarrow{y_q}}{\|\overrightarrow{x_p}\| \, \|\overrightarrow{y_q}\|} \right) \tag{3}$$

where $n$ is the sample size of the corresponding cluster, and $\overrightarrow{x_p}$ and $\overrightarrow{y_q}$ are the vector samples in each cluster. As a sensitivity check for cluster distance metric, we also calculated $\phi_k$ based on the medoid cluster distance. Medoid distance measures the distance between clusters by calculating the cosine distance between 'medoid samples' of each cluster, which shows the closest average cosine distance with samples within each cluster. To test the predictive power of cortical distinctiveness of each valence type for AWES scores, we applied the same univariate and multivariate analysis framework as described in the "Behavioral analysis" section.

Inference of feature importance in CEBRA using Dynamask

Due to the black-box nature of CEBRA, it was challenging to directly evaluate which STFT-processed EEG features were crucial for building latent neural spaces in CEBRA (i.e., contrasting valence types in each time point). To address this issue, we applied "Dynamask" (Crabbé & Van Der Schaar, 2021), a perturbation-based explainable AI techniques, to infer attribution maps from the trained CEBRA models. Dynamask learns perturbation weights, $w$, for each feature at every time point to generate pseudo-embeddings with maximal MSE value compared to the original embeddings with the slightest perturbation. Here, the $i$-th input feature at the timepoint $t$, $x_{i,t}$ is perturbed to $p(x_{i,t})$ as a weighted sum of its own value and the average value in the time window it belongs to, formulated by the following equation:

$$\pi(x_{i,t}) = (1 - \omega_{i,t}) \cdot x_{i,t} + \omega_{i,t} \cdot \frac{1}{2W+1} \sum_{t'=t-W}^{t+W} x_{i,t'} \tag{4}$$

where $W$ is the time window size. $w_{i,t} = 1$ indicates a significant alteration of the CEBRA embeddings upon replacement, implying importance in contrasting valence labels at the given time point. With a receptive field of 10 samples, $W = 5$ was set to align the working behavior of Dynamask with CEBRA's.

We obtained $w$ matrices for each EEG feature from data showing significant silhouette coefficients for ambivalence clusters ($P_{perm} < .05$) using the following parameters: keep_ratio = 0.1, n_epoch = 2,500, initial_mask_coef = 0.5, size_reg_factor_init = 0.5, size_reg_factor_dilation = 100, time_reg_factor = 0, learning_rate = 0.03, and momentum = 0.9. By aligning the time point of STFT-processed EEG features and valence keypress sequence, we calculated $w_{ambivalent}$ of each feature by averaging $w$ values of each feature for time points labeled as ambivalent states in the valence keypress. Consequently, $w_{ambivalent}$ provides an indirect measure of feature importance to contrast ambivalence-labeled EEG samples from other valence-labeled EEG samples to construct latent valence-cortical space. The same analyses were conducted for positive and negative valence states with data displaying significant silhouette coefficient for each state, respectively.

## Post-hoc analysis of feature importance with hidden Markov model

We conducted a post-hoc analysis using the hidden Markov Model (HMM) to confirm Dynamask-driven attribution weights of each feature in distinguishing different valence states within the latent neural space. Particularly, Dynamask revealed that the power of the delta band held greater importance than the power of other frequency bands. Based on this, we hypothesized that hidden states of delta features would be temporally aligned with individual valence dynamics and tested this hypothesis using an HMM. For all HMM analyses, we used the "brainiak" package (Kumar et al., 2021).

We divided the STFT-processed EEG features into frequency bands to generate five input groups – delta, theta, alpha, beta, and gamma – for each participant and clip. These inputs were independently fitted to an HMM, estimating the time points of neural boundaries corresponding to the number of valence transitions reported via keypress. Thus, the event number, a hyperparameter of the HMM, was informed by the participants' reported valence keypress for the clip. Boundaries within ± 3 seconds of the actual valence transition time points were considered a 'match', and the 'match rate' was calculated as the ratio of matched boundaries to the total number of boundaries.

Following the framework of previous literature (Vaccaro et al., 2024), we assessed the statistical significance of the match rates for the five frequency bands across all participants and clips. This framework's advantage is that it accounts for variability in the number of valence transitions reported by each participant for each clip. The approach involves the following steps: First, for each participant and frequency band input, maintain the estimated intervals of neural boundaries but shuffle them randomly, comparing these to the actual valence transition time points to compute a pseudo-match rate. This process is repeated 1,000 times to generate a null distribution of match rates. Second, calculate the difference between the actual match rate and the mean of the null distribution for each frequency band group for each participant-clip data. Averaging these differences across participants and clips yields the average mean difference for feature group $k$, denoted as $M_k$. Third, repeat this process 1,000 times using the null distribution of each participant-clip data to derive a null distribution of 1,000 mean differences between permuted match rates and the mean of the null distribution. Denote the $i$-th permuted mean difference for feature group $k$ as $Q_k^{(i)}$. Last, calculate the $P$ value, $P_{\text{perm}}$, for $M_k$ using the following equation:

$$P_{\text{perm}} = \frac{1}{1000} \sum_{i=1}^{1000} \mathbf{1}\left(Q_k^{(i)} \geq M_k\right) \tag{5}$$

# References

Aftanas, L. I., Lotova, N. V., Koshkarov, V. I., Makhnev, V. P., Mordvintsev, Y. N., & Popov, S. A. (1998). Non-linear dynamic complexity of the human EEG during evoked emotions. *International journal of psychophysiology*, *28*(1), 63-76.

An, S., Ji, L.-J., Marks, M., & Zhang, Z. (2017). Two sides of emotion: Exploring positivity and negativity in six basic emotions across cultures. *Frontiers in psychology*, *8*, 253368.

Barrett, L. F., & Bliss-Moreau, E. (2009). Affect as a psychological primitive. *Advances in experimental social psychology*, *41*, 167-218.

Belfi, A. M., Vessel, E. A., Brielmann, A., Isik, A. I., Chatterjee, A., Leder, H., Pelli, D. G., & Starr, G. G. (2019). Dynamics of aesthetic experience are reflected in the default-mode network. *NeuroImage*, *188*, 584-597.

Berridge, K. C. (2019). Affective valence in the brain: modules or modes? *Nature Reviews Neuroscience*, *20*(4), 225-234.

Berrios, R., Totterdell, P., & Kellett, S. (2015). Eliciting mixed emotions: a meta-analysis comparing models, types, and measures. *Frontiers in psychology*, *6*, 133792.

Bonner, E. T., & Friedman, H. L. (2011). A conceptual clarification of the experience of awe: An interpretative phenomenological analysis. *The humanistic psychologist*, *39*(3), 222-235.

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: the self-assessment manikin and the semantic differential. *Journal of behavior therapy and experimental psychiatry*, *25*(1), 49-59.

Brans, K., & Verduyn, P. (2014). Intensity and duration of negative emotions: Comparing the role of appraisals and regulation strategies. *PloS one*, *9*(3), e92410.

Brenner, C. A., Rumak, S. P., Burns, A. M., & Kieffaber, P. D. (2014). The role of encoding and attention in facial emotion memory: an EEG investigation. *International journal of psychophysiology*, *93*(3), 398-410.

Briesemeister, B. B., Kuchinke, L., & Jacobs, A. M. (2012). Emotional valence: A bipolar continuum or two independent dimensions? *Sage Open*, *2*(4), 2158244012466558.

Brooker, R. J., Mistry-Patel, S., Kling, J. L., & Howe, H. A. (2021). Deriving within-person estimates of delta-beta coupling: A novel measure for identifying individual differences in emotion and neural function in childhood. *Developmental psychobiology*, *63*(6), e22172.

Brzezicka, A., Kamiński, J., Kamińska, O. K., Wołyńczyk-Gmaj, D., & Sedek, G. (2017). Frontal EEG alpha band asymmetry as a predictor of reasoning deficiency in depressed people. *Cognition and emotion*, *31*(5), 868-878.

Cavanagh, J. F., & Shackman, A. J. (2015). Frontal midline theta reflects anxiety and cognitive control: meta-analytic evidence. *Journal of physiology-Paris*, *109*(1-3), 3-15.

Čeko, M., Kragel, P. A., Woo, C.-W., López-Solà, M., & Wager, T. D. (2022). Common and stimulus-type-specific brain representations of negative affect. *Nature neuroscience*, *25*(6), 760-770.

Chaudhury, S. H., Garg, N., & Jiang, Z. (2022). The curious case of threat-awe: A theoretical and empirical reconceptualization. *Emotion*, *22*(7), 1653.

Chirico, A., Borghesi, F., Yaden, D. B., Pizzolante, M., Sarcinella, E. D., Cipresso, P., & Gaggioli, A. (2024). Unveiling the underlying structure of awe in virtual reality and in autobiographical recall: an exploratory study. *Scientific reports*, *14*(1), 12474.

Chirico, A., Cipresso, P., Yaden, D. B., Biassoni, F., Riva, G., & Gaggioli, A. (2017). Effectiveness of immersive videos in inducing awe: an experimental study. *Scientific reports*, *7*(1), 1218.

Chirico, A., Ferrise, F., Cordella, L., & Gaggioli, A. (2018). Designing awe in virtual reality: An experimental study. *Frontiers in psychology*, *8*, 293522.

Chirico, A., Shiota, M. N., & Gaggioli, A. (2021). Positive emotion dispositions and emotion regulation in the Italian population. *PloS one*, *16*(3), e0245545.

Chirico, A., Yaden, D. B., Riva, G., & Gaggioli, A. (2016). The potential of virtual reality for the investigation of awe. *Frontiers in psychology*, *7*, 223153.

Crabbé, J., & Van Der Schaar, M. (2021). Explaining time series predictions with dynamic masks. International Conference on Machine Learning,

Cunningham, J. P., & Yu, B. M. (2014). Dimensionality reduction for large-scale neural recordings. *Nature neuroscience*, *17*(11), 1500-1509.

Dael, N., Perseguers, M.-N., Marchand, C., Antonietti, J.-P., & Mohr, C. (2016). Put on that colour, it fits your emotion: Colour appropriateness as a function of expressed emotion. *Quarterly Journal of Experimental Psychology*, *69*(8), 1619-1630.

Dai, Y., Jiang, T., & Miao, M. (2022). Uncovering the effects of awe on meaning in life. *Journal of Happiness Studies*, *23*(7), 3517-3529.

Delorme, A. (2023). EEG is better left alone. *Scientific reports*, *13*(1), 2372.

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of neuroscience methods*, *134*(1), 9-21.

Diaconis, P., & Freedman, D. (1984). Asymptotics of graphical projection pursuit. *The annals of statistics*, 793-815.

Dumitru, A., Rocchi, L., Saini, F., Rothwell, J. C., Roiser, J. P., David, A. S., Richieri, R. M., Lewis, G., & Lewis, G. (2020). Influence of theta-burst transcranial magnetic stimulation over the

dorsolateral prefrontal cortex on emotion processing in healthy volunteers. *Cognitive, Affective, & Behavioral Neuroscience*, *20*, 1278-1293.

Ersner-Hershfield, H., Mikels, J. A., Sullivan, S. J., & Carstensen, L. L. (2008). Poignancy: mixed emotional experience in the face of meaningful endings. *Journal of personality and social psychology*, *94*(1), 158.

Ertl, M., Hildebrandt, M., Ourina, K., Leicht, G., & Mulert, C. (2013). Emotion regulation by cognitive reappraisal—the role of frontal theta oscillations. *NeuroImage*, *81*, 412-421.

Gable, P., & Harmon-Jones, E. (2010). The motivational dimensional model of affect: Implications for breadth of attention, memory, and cognitive categorisation. *Cognition and emotion*, *24*(2), 322-337.

Gallego, J. A., Perich, M. G., Chowdhury, R. H., Solla, S. A., & Miller, L. E. (2020). Long-term stability of cortical population dynamics underlying consistent behavior. *Nature neuroscience*, *23*(2), 260-270.

Gordon, A. M., Stellar, J. E., Anderson, C. L., McNeil, G. D., Loew, D., & Keltner, D. (2017). The dark side of the sublime: Distinguishing a threat-based variant of awe. *Journal of personality and social psychology*, *113*(2), 310.

Guan, F., Zhao, S., Chen, S., Lu, S., Chen, J., & Xiang, Y. (2019). The neural correlate difference between positive and negative awe. *Frontiers in Human Neuroscience*, *13*, 206.

Harmon-Jones, E., & Gable, P. A. (2018). On the role of asymmetric frontal cortical activity in approach and withdrawal motivation: An updated review of the evidence. *Psychophysiology*, *55*(1), e12879.

Hartig, F. (2018). DHARMa: residual diagnostics for hierarchical (multi-level/mixed) regression models. *R Packag version 020*.

Hoemann, K., Gendron, M., & Barrett, L. F. (2017). Mixed emotions in the predictive brain. *Current Opinion in Behavioral Sciences*, *15*, 51-57.

Honk, J. v., & Schutter, D. J. (2006). From affective valence to motivational direction: the frontal asymmetry of emotion revised. *Psychological science*, *17*(11), 963-965.

Hu, X., Yu, J., Song, M., Yu, C., Wang, F., Sun, P., Wang, D., & Zhang, D. (2017). EEG correlates of ten positive emotions. *Frontiers in Human Neuroscience*, *11*, 26.

Hutchison, I. C., & Rathore, S. (2015). The role of REM sleep theta activity in emotional memory. *Frontiers in psychology*, *6*, 1439.

Ihm, E. D., Paloutzian, R. F., van Elk, M., & Schooler, J. W. (2019). Awe as a meaning-making emotion: On the evolution of awe and the origin of religions. In *The evolution of religion, religiosity and theology* (pp. 138-153). Routledge.

Jiang, H., Ding, X., Zhao, S., Li, Y., Bai, H., Gao, H., & Gao, W. (2022). Abnormal brain oscillations and activation of patients with heroin use disorder during emotion regulation: The role of delta-and theta-band power. *Journal of Affective Disorders*, *315*, 121-129.

Jiang, T., Hicks, J. A., Yuan, W., Yin, Y., Needy, L., & Vess, M. (2024). The unique nature and psychosocial implications of awe. *Nature Reviews Psychology*, 1-14.

Kahn, A. S., & Cargile, A. C. (2021). Immersive and interactive awe: Evoking awe via presence in virtual reality and online videos to prompt prosocial behavior. *Human Communication Research*, *47*(4), 387-417.

Keltner, D., & Haidt, J. (2003). Approaching awe, a moral, spiritual, and aesthetic emotion. *Cognition and emotion*, *17*(2), 297-314.

Kumar, M., Anderson, M. J., Antony, J. W., Baldassano, C., Brooks, P. P., Cai, M. B., Chen, P.-H. C., Ellis, C. T., Henselman-Petrusek, G., & Huberdeau, D. (2021). BrainIAK: the brain imaging analysis kit. *Aperture neuro*, *1*(4).

Kurt, P., Eroğlu, K., Kuzgun, T. B., & Güntekin, B. (2017). The modulation of delta responses in the interaction of brightness and emotion. *International journal of psychophysiology*, *112*, 1-8.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmertest'. *R package version*, *2*(0), 734.

Lammel, S., Lim, B. K., Ran, C., Huang, K. W., Betley, M. J., Tye, K. M., Deisseroth, K., & Malenka, R. C. (2012). Input-specific control of reward and aversion in the ventral tegmental area. *Nature*, *491*(7423), 212-217.

Larsen, J. T., Norris, C. J., McGraw, A. P., Hawkley, L. C., & Cacioppo, J. T. (2009). The evaluative space grid: a single-item measure of positivity and negativity. *Cognition and emotion*, *23*(3), 453-480.

Larsen, R. J., & Fredrickson, B. L. (1999). Measurement issues in emotion research. *Well-being: The foundations of hedonic psychology*, *40*, 60.

LeDell, E., & Poirier, S. (2020). H2o automl: Scalable automatic machine learning. Proceedings of the AutoML Workshop at ICML,

Lee, S. A., Lee, J.-J., Han, J., Choi, M., Wager, T. D., & Woo, C.-W. (2024). Brain representations of affective valence and intensity in sustained pleasure and pain. *Proceedings of the National Academy of Sciences*, *121*(25), e2310433121.

Lettieri, G., Handjaras, G., Cappello, E. M., Setti, F., Bottari, D., Bruno, V., Diano, M., Leo, A., Tinti, C., & Garbarini, F. (2024). Dissecting abstract, modality-specific and experience-dependent coding of affect in the human brain. *Science Advances*, *10*(10), eadk6840.

Lettieri, G., Handjaras, G., Ricciardi, E., Leo, A., Papale, P., Betta, M., Pietrini, P., & Cecchetti, L. (2019). Emotionotopy in the human right temporo-parietal cortex. *Nature communications*, *10*(1), 5568.

Levens, S. M., & Phelps, E. A. (2010). Insula and orbital frontal cortex activity underlying emotion interference resolution in working memory. *Journal of cognitive neuroscience*, *22*(12), 2790-2803.

Li, Y., Mahir, D., & Stellar, J. E. (2023). Exploring the impact of awe on the multifaceted construct of empathy. *Emotion*.

Lim, Y.-J., Yu, B.-H., Kim, D.-K., & Kim, J.-H. (2010). The positive and negative affect schedule: Psychometric properties of the Korean version. *Psychiatry investigation*, *7*(3), 163.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., & Gonzalez, J. (2013). Package 'cluster'. *Dosegljivo na*, *980*.

Man, V., Nohlen, H. U., Melo, H., & Cunningham, W. A. (2017). Hierarchical brain systems support multiple representations of valence and mixed affect. *Emotion Review*, *9*(2), 124-132.

McClay, M., Sachs, M. E., & Clewett, D. (2023). Dynamic emotional states shape the episodic structure of memory. *Nature communications*, *14*(1), 6533.

Meer, J. N. v. d., Breakspear, M., Chang, L. J., Sonkusare, S., & Cocchi, L. (2020). Movie viewing elicits rich and reliable brain state dynamics. *Nature communications*, *11*(1), 5004.

Monroy, M., & Keltner, D. (2023). Awe as a pathway to mental and physical health. *Perspectives on Psychological Science*, *18*(2), 309-320.

Moulier, V., Gaudeau-Bosma, C., Thomas, F., Isaac, C., Thomas, M., Durand, F., Schenin-King Andrianisaina, P., Valabregue, R., Laidi, C., & Benadhira, R. (2021). Effect of intermittent theta burst stimulation on the neural processing of emotional stimuli in healthy volunteers. *Journal of Clinical Medicine*, *10*(11), 2449.

Myruski, S., Bagrodia, R., & Dennis-Tiwary, T. (2022). Delta-beta correlation predicts adaptive child emotion regulation concurrently and two years later. *Biological Psychology*, *167*, 108225.

Nakayama, M., Nozaki, Y., Taylor, P. M., Keltner, D., & Uchida, Y. (2020). Individual and cultural differences in predispositions to feel positive and negative aspects of awe. *Journal of Cross-Cultural Psychology*, *51*(10), 771-793.

Nishida, M., Pearsall, J., Buckner, R. L., & Walker, M. P. (2009). REM sleep, prefrontal theta, and the consolidation of human emotional memory. *Cerebral Cortex*, *19*(5), 1158-1166.

Norman, G. J., Norris, C. J., Gollan, J., Ito, T. A., Hawkley, L. C., Larsen, J. T., Cacioppo, J. T., & Berntson, G. G. (2011). Current emotion research in psychophysiology: The neurobiology of evaluative bivalence. *Emotion Review*, *3*(3), 349-359.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *the Journal of machine Learning research*, *12*, 2825-2830.

Phelps, R. A., Brooker, R. J., & Buss, K. A. (2016). Toddlers' dysregulated fear predicts delta–beta coupling during preschool. *Developmental Cognitive Neuroscience*, *17*, 28-34.

Picard, F., & Craig, A. (2009). Ecstatic epileptic seizures: a potential window on the neural basis for human self-awareness. *Epilepsy & Behavior*, *16*(3), 539-546.

Piff, P. K., Dietze, P., Feinberg, M., Stancato, D. M., & Keltner, D. (2015). Awe, the small self, and prosocial behavior. *Journal of personality and social psychology*, *108*(6), 883.

Putman, P., Arias-Garcia, E., Pantazi, I., & van Schie, C. (2012). Emotional Stroop interference for threatening words is related to reduced EEG delta–beta coupling and low attentional control. *International journal of psychophysiology*, *84*(2), 194-200.

Quaedflieg, C. W., Smulders, F. T., Meyer, T., Peeters, F., Merckelbach, H., & Smeets, T. (2016). The validity of individual frontal alpha asymmetry EEG neurofeedback. *Social cognitive and affective neuroscience*, *11*(1), 33-43.

Quesnel, D., & Riecke, B. E. (2018). Are you awed yet? How virtual reality gives us awe and goose bumps. *Frontiers in psychology*, *9*, 403078.

Reynolds, S. M., & Berridge, K. C. (2008). Emotional environments retune the valence of appetitive versus fearful functions in nucleus accumbens. *Nature neuroscience*, *11*(4), 423-425.

Rolls, E. T., & Grabenhorst, F. (2008). The orbitofrontal cortex and beyond: from affect to decision-making. *Progress in neurobiology*, *86*(3), 216-244.

Rudd, M., Vohs, K. D., & Aaker, J. (2012). Awe expands people's perception of time, alters decision making, and enhances well-being. *Psychological science*, *23*(10), 1130-1136.

Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological review*, *110*(1), 145.

Russell, J. A. (2017). Mixed emotions viewed from the psychological constructionist perspective. *Emotion Review*, *9*(2), 111-117.

Safaie, M., Chang, J. C., Park, J., Miller, L. E., Dudman, J. T., Perich, M. G., & Gallego, J. A. (2023). Preserved neural dynamics across animals performing similar behaviour. *Nature*, *623*(7988), 765-771.

Sawada, K., Koike, H., Murayama, A., Nishida, H., & Nomura, M. (2024). Appreciation processing evoking feelings of being moved and inspiration: Awe and meaning-making. *Journal of Creativity*, *34*(1), 100076.

Schneider, S., Lee, J. H., & Mathis, M. W. (2023). Learnable latent embeddings for joint behavioural and neural analysis. *Nature*, *617*(7960), 360-368.

Shen, X., Hu, X., Liu, S., Song, S., & Zhang, D. (2020). Exploring EEG microstates for affective computing: decoding valence and arousal experiences during video watching. 2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC),

Shiota, M. N., Keltner, D., & John, O. P. (2006). Positive emotion dispositions differentially associated with Big Five personality and attachment style. *The journal of positive psychology*, *1*(2), 61-71.

Shiota, M. N., Keltner, D., & Mossman, A. (2007). The nature of awe: Elicitors, appraisals, and effects on self-concept. *Cognition and emotion*, *21*(5), 944-963.

Silvia, P. J., Fayn, K., Nusbaum, E. C., & Beaty, R. E. (2015). Openness to experience and awe in response to nature and music: personality and profound aesthetic experiences. *Psychology of aesthetics, creativity, and the arts*, *9*(4), 376.

Simmons, A., Stein, M. B., Matthews, S. C., Feinstein, J. S., & Paulus, M. P. (2006). Affective ambiguity for a group recruits ventromedial prefrontal cortex. *NeuroImage*, *29*(2), 655-661.

Sopp, M. R., Michael, T., Weeß, H.-G., & Mecklinger, A. (2017). Remembering specific features of emotional events across time: The role of REM sleep and prefrontal theta oscillations. *Cognitive, Affective, & Behavioral Neuroscience*, *17*, 1186-1209.

Stellar, J. E., Bai, Y., Anderson, C. L., Gordon, A., McNeil, G. D., Peng, K., & Keltner, D. (2024). Culture and Awe: Understanding Awe as a Mixed Emotion. *Affective Science*, 1-11.

Suk, H. J., & Irtel, H. (2010). Emotional response to color across media. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, *35*(1), 64-77.

Takano, R., & Nomura, M. (2022). Neural representations of awe: Distinguishing common and distinct neural mechanisms. *Emotion*, *22*(4), 669.

Thao, H. T. P., Herremans, D., & Roig, G. (2019). Multimodal Deep Models for Predicting Affective Responses Evoked by Movies. ICCV Workshops,

Vaccaro, A. G., Kaplan, J. T., & Damasio, A. (2020). Bittersweet: the neuroscience of ambivalent affect. *Perspectives on Psychological Science*, *15*(5), 1187-1199.

Vaccaro, A. G., Wu, H., Iyer, R., Shakthivel, S., Christie, N. C., Damasio, A., & Kaplan, J. (2024). Neural patterns associated with mixed valence feelings differ in consistency and predictability throughout the brain. *Cerebral Cortex*, *34*(4), bhae122.

Van Der Vinne, N., Vollebregt, M. A., Van Putten, M. J., & Arns, M. (2017). Frontal alpha asymmetry as a diagnostic marker in depression: Fact or fiction? A meta-analysis. *Neuroimage: clinical*, *16*, 79-87.

Van Mechelen, I., Verduyn, P., & Brans, K. (2013). The duration of emotional episodes. In *Changing emotions* (pp. 174-180). Psychology Press.

Van Tilburg, W. A., Wildschut, T., & Sedikides, C. (2018). Nostalgia's place among self-relevant emotions. *Cognition and emotion*, *32*(4), 742-759.

Verduyn, P., Delvaux, E., Van Coillie, H., Tuerlinckx, F., & Van Mechelen, I. (2009). Predicting the duration of emotional experience: two experience sampling studies. *Emotion*, *9*(1), 83.

Verduyn, P., Van Mechelen, I., & Tuerlinckx, F. (2011). The relation between event processing and the duration of emotional experience. *Emotion*, *11*(1), 20.

Viinikainen, M., Jääskeläinen, I. P., Alexandrov, Y., Balk, M. H., Autti, T., & Sams, M. (2010). Nonlinear relationship between emotional valence and brain activity: evidence of separate negative and positive valence dimensions. *Human brain mapping*, *31*(7), 1030-1040.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., & Bright, J. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature methods*, *17*(3), 261-272.

Wacker, J., Heldmann, M., & Stemmler, G. (2003). Separating emotion and motivational direction in fear and anger: effects on frontal asymmetry. *Emotion*, *3*(2), 167.

Wittmann, M. (2013). The inner sense of time: how the brain creates a representation of duration. *Nature Reviews Neuroscience*, *14*(3), 217-223.

Yaden, D. B., Iwry, J., Slack, K. J., Eichstaedt, J. C., Zhao, Y., Vaillant, G. E., & Newberg, A. B. (2016). The overview effect: awe and self-transcendent experience in space flight. *Psychology of Consciousness: Theory, Research, and Practice*, *3*(1), 1.

Yaden, D. B., Kaufman, S. B., Hyde, E., Chirico, A., Gaggioli, A., Zhang, J. W., & Keltner, D. (2019). The development of the Awe Experience Scale (AWE-S): A multifactorial measure for a complex emotion. *The journal of positive psychology*, *14*(4), 474-488.

Yin, Y., Yuan, W., Hao, C., Du, Y., Xu, Z., Hicks, J. A., & Jiang, T. (2024). Awe fosters positive attitudes toward solitude. *Nature Mental Health*, 1-11.

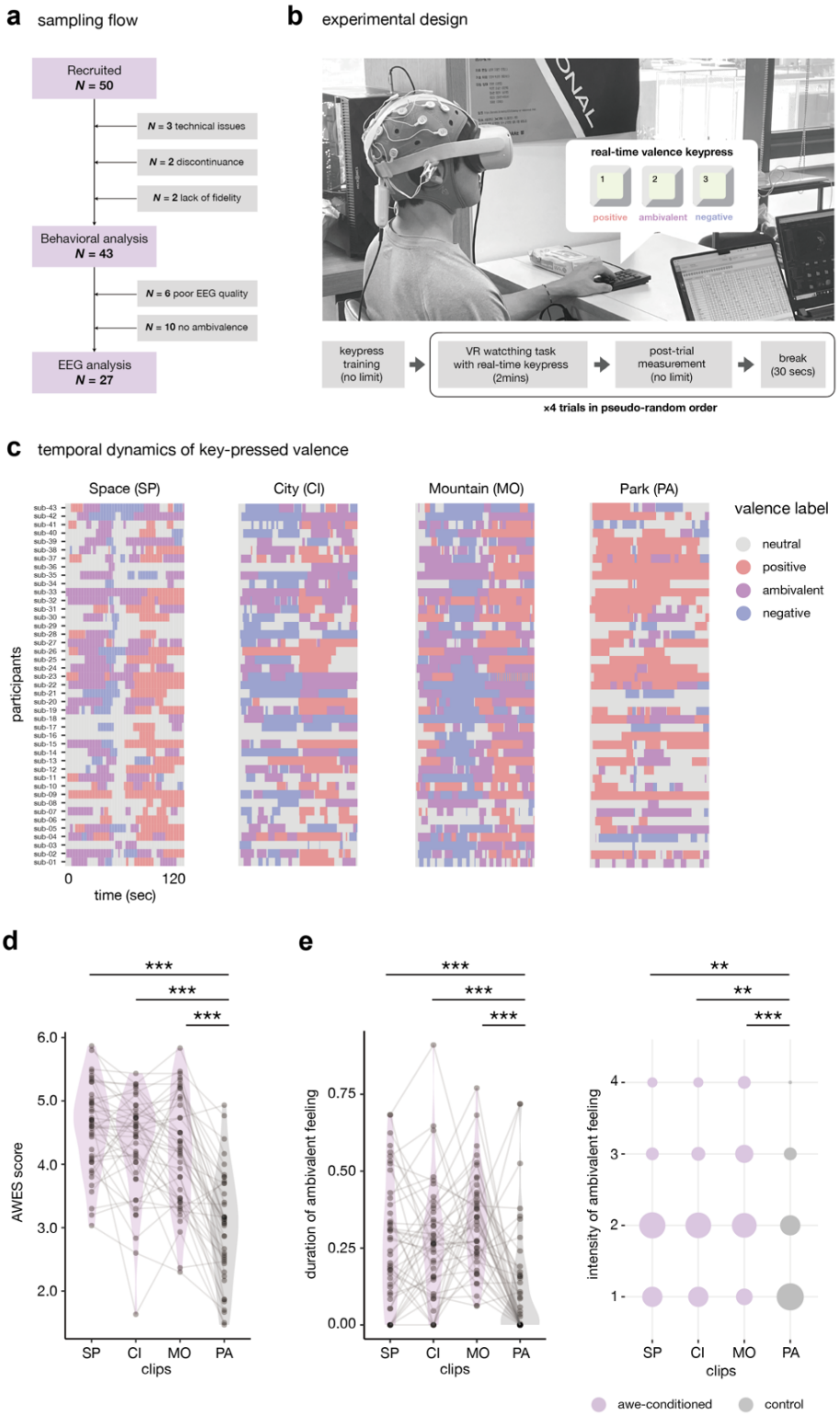



**Fig 1. Valence dynamics, awe, and ambivalent feelings for VR clips. a,** The sampling procedures, summarizing participant recruitment and selection for behavioral and EEG analysis. **b,** The integrated VR-EEG-real time valence

rating protocol, showing the sequence of experimental activities. **c,** Valence dynamics rated by real-time keypress from 43 participants during exposure to each VR clip. **d,** AWES scores evaluated post-trial for each clip, reflecting the intensity of awe experiences. **e**, The duration (left) and intensity (right) of ambivalent feelings for each clip, comparing responses between awe-conditioned and control clips. *$P_{FDR} < .05$; **$P_{FDR} < .01$; ***$P_{FDR} < .001$.

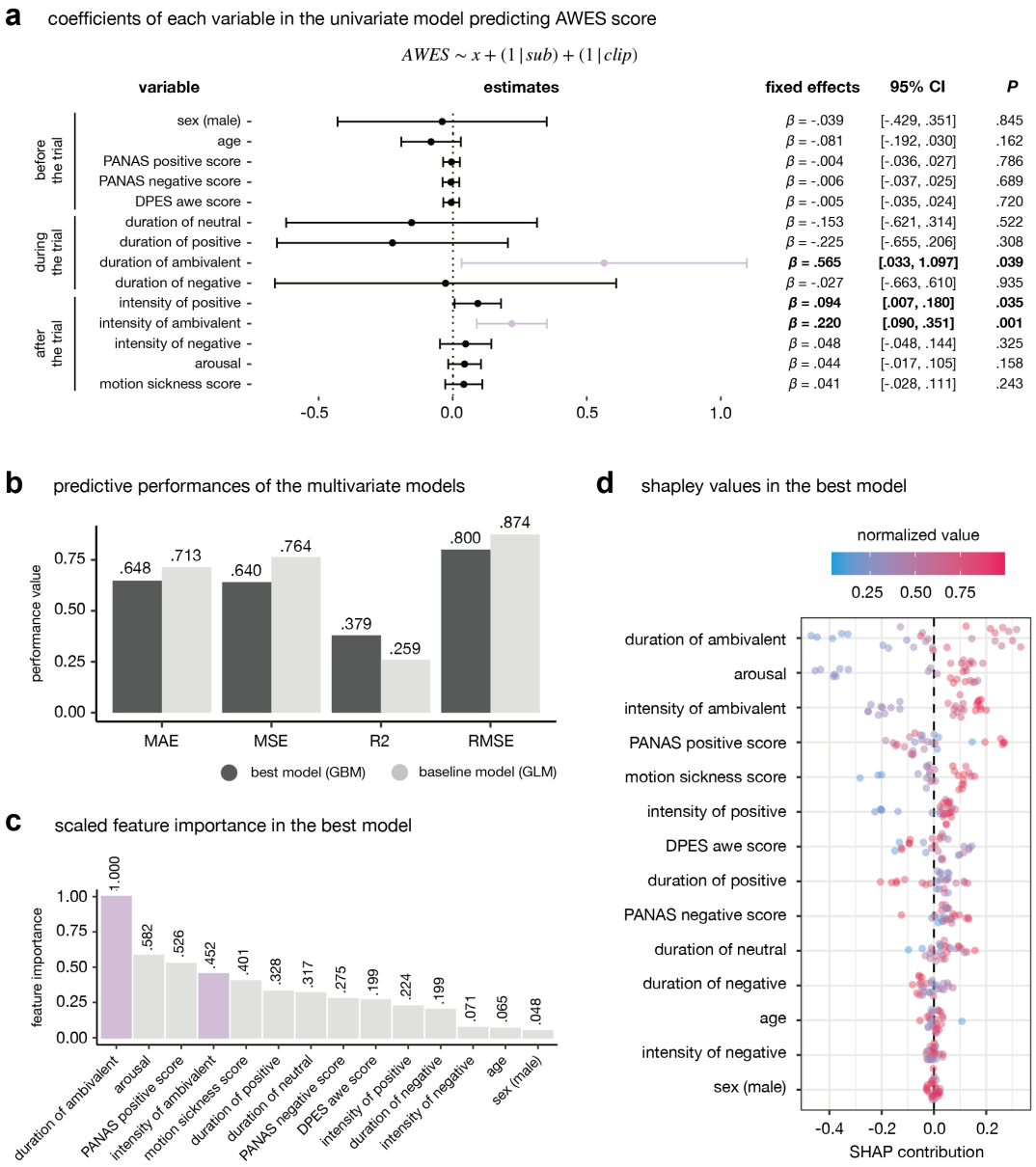



**Fig 2. Association between behavioral metrics and awe ratings. a,** Coefficients of each behavioral metric in the univariate linear mixed effect models. Error bars denote 95% confidence intervals of fixed effects. Purple bars show the estimates of ambivalence-related features. Bolded statistics indicate statistically significant results at $P < .05$. **b,** Comparison of predictive performances between AutoML-based best model (GBM) and linear regression model (GLM). **c,** Feature importance (scale from 0 to 1) of all behavioral features calculated from the best model. Purple bars show the importance of ambivalence-related features. **d,** Shapley values of all features computed from the best model.

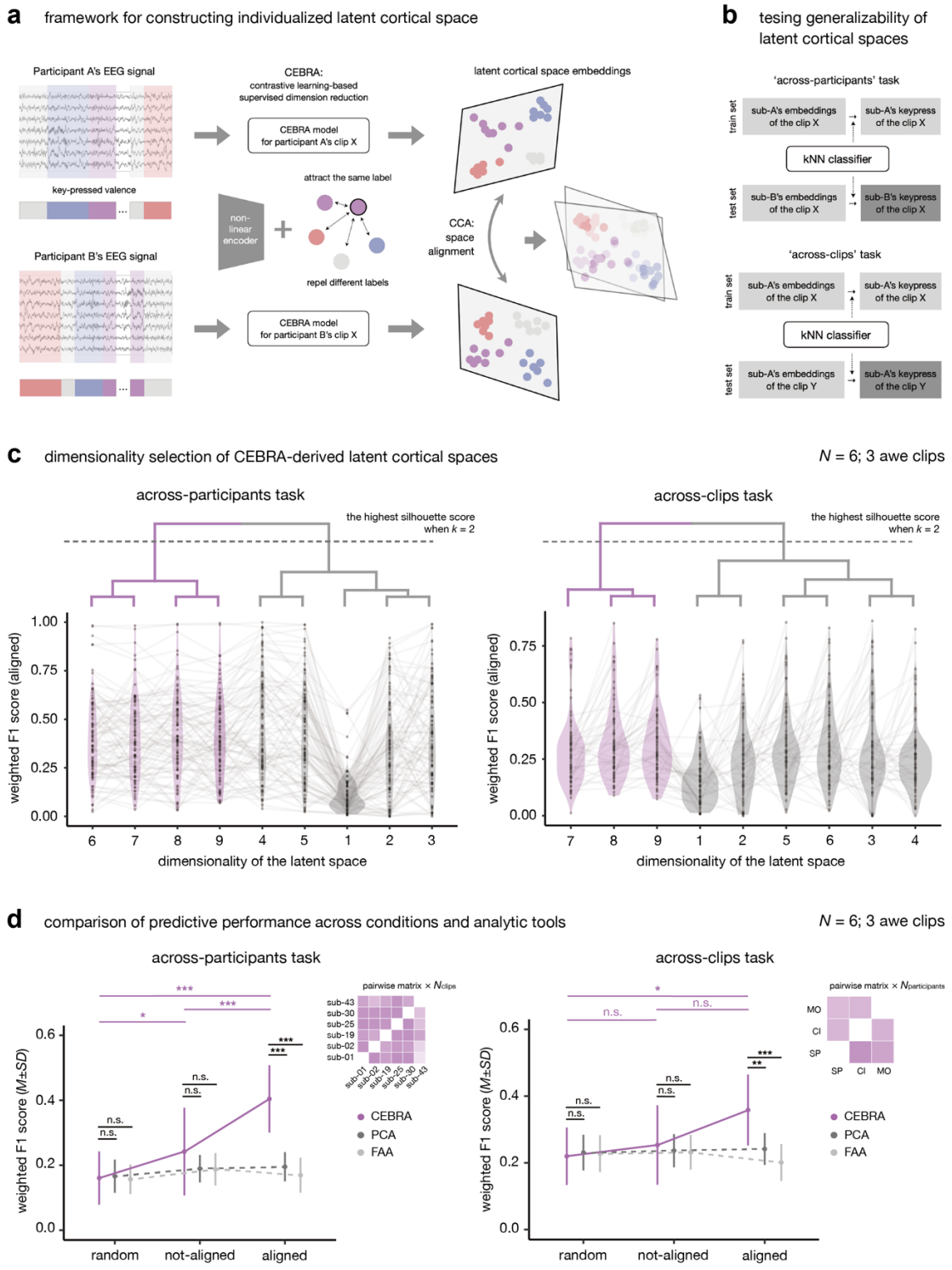



**Fig 3. Identification of individualized latent neural spaces and evaluation of their generalizability. a,** Schematic illustrating the extraction of individualized latent neural spaces using CEBRA and neural alignment. **b,** Schematic depicting the assessment of generalizability through pairwise predictive tasks across participants and

sensory stimuli. **c,** Dimensionality selection for CEBRA-driven latent neural spaces. Hierarchical clustering analysis identified clusters with high performance in predicting individual differences (purple violin plots). **d,** Predictive performance of latent spaces derived from CEBRA, PCA, and FAA across three conditions. CEBRA-based embeddings consistently outperform PCA and FAA-based embeddings in predicting individual differences (black asterisks). Heatmaps visualize conceptual scheme of results from pairwise predictive tasks. Within CEBRA-based predictions, performance varies significantly across conditions (purple asterisks). Results are based on $N = 6$ participants reporting all valence types in all awe-inducing clips. *$P_{FDR} < .05$; **$P_{FDR} < .01$; ***$P_{FDR} < .001$.

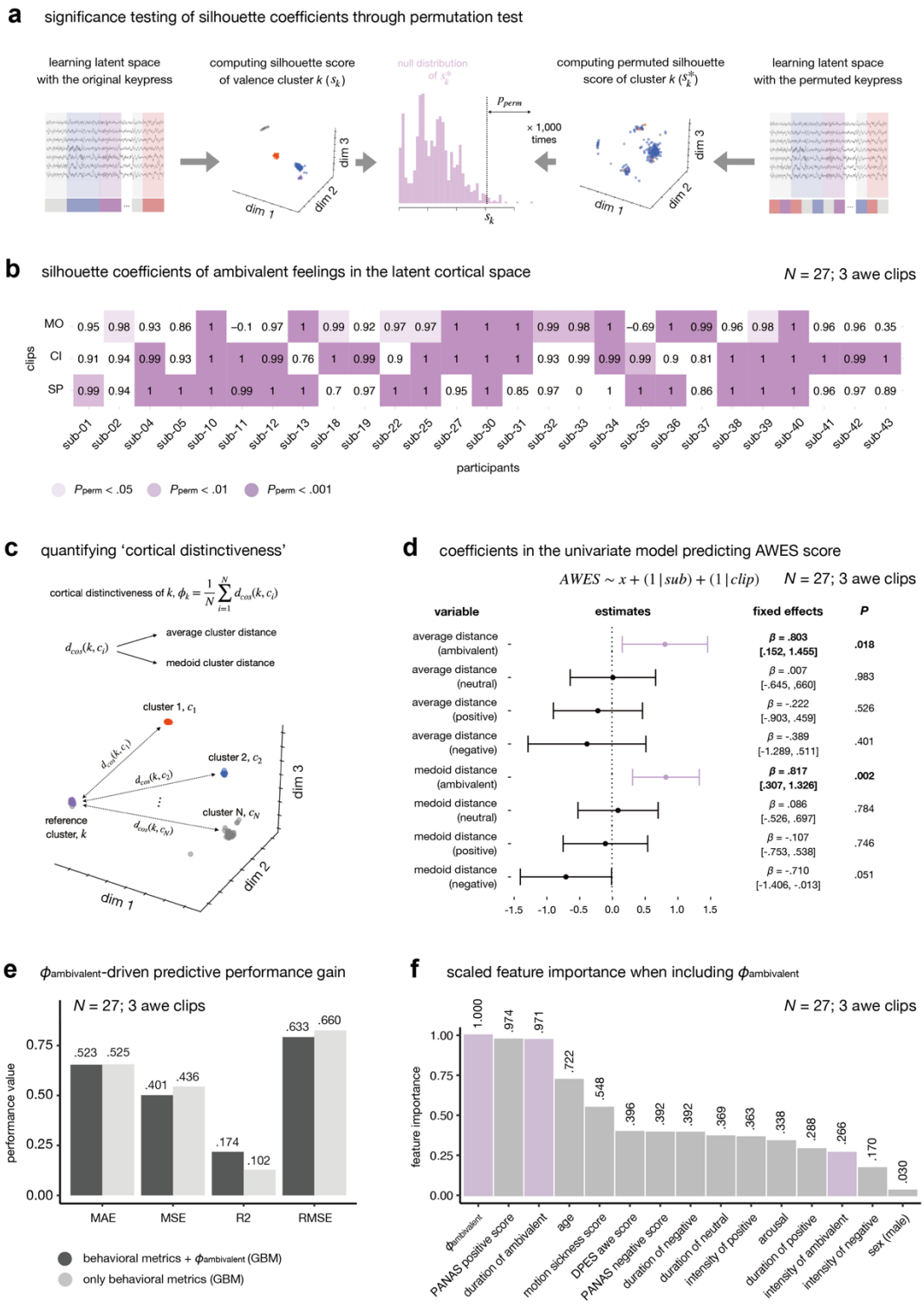



**Fig 4. Individual variability of ambivalence-related latent neural representation and its predictive power on awe ratings. a,** Schematic illustrating the assessment of statistical significance for silhouette coefficients through a

permutation test. **b,** Silhouette coefficients of ambivalent valence clusters within individual latent neural spaces. Color intensity indicates the statistical significance of the silhouette coefficients. **c,** Schematic depicting the calculation of "cortical distinctiveness" for each valence cluster. See Equation (2) and (3) in 'Methods' section for annotations. **d,** Explanatory power of each valence cluster's cortical distinctiveness in the univariate linear mixed-effects models. Error bar denotes 95% confidence interval of fixed effects. Purple bars show the estimates of ambivalence-related features. Bolded statistics indicate statistically significant results at $P < .05$. Notably, the cortical distinctiveness of ambivalent states (purple bars) significantly predicts awe ratings. **e,** Performance gain observed when cortical distinctiveness of ambivalent feelings was added to the behavioral prediction model. **f,** Scaled (0-1) feature importance of cortical distinctiveness of ambivalent feelings and behavioral features. In both, the cortical distinctiveness of ambivalent states (purple bars) emerges as a prominent predictor of awe. Results are based on $N = 27$ participants included in the electrophysiological analysis. *$P < .05$.

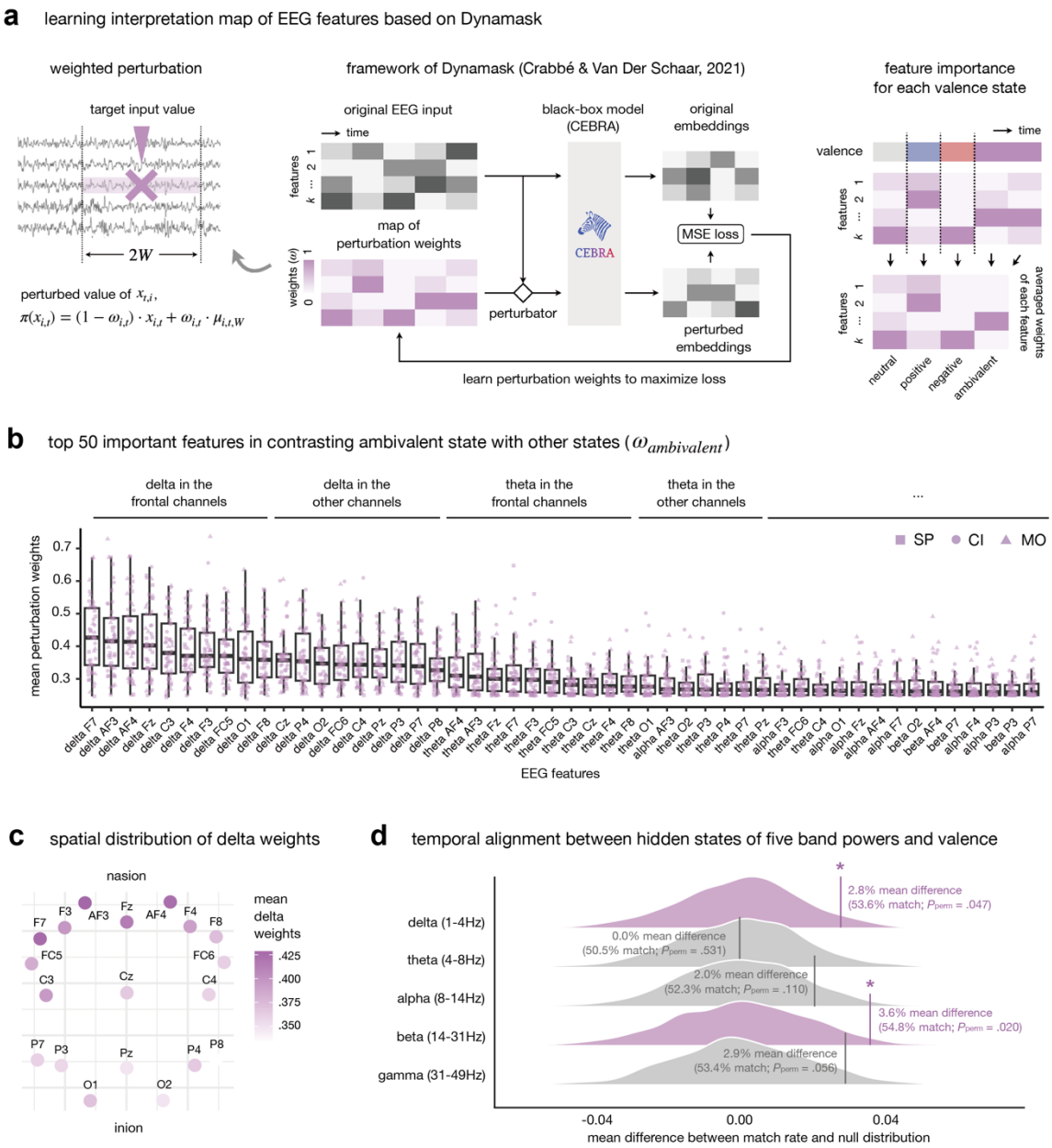



**Fig 5. EEG features distinguishing different valence types. a,** Schematic illustrating the use of Dynamask to learn an attribution map for EEG features. **b,** Top 50 most important features for differentiating ambivalent feelings from other valence categories. **c,** Spatial distribution of feature importance for delta power. **d,** Mean difference between the observed match rate and the null distribution for each frequency band, assessed through a permutation approach. The vertical line indicates the observed mean difference. Purple distributions represent statistically significant differences ($P_{perm} < .05$), while gray distributions are non-significant. Results are based on data from $N = 27$ participants included in the electrophysiological analysis.

# Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2021R1C1C1006503, RS-2023-00266787, RS-2023-00265406, RS-2024-00421268), by Creative-Pioneering Researchers Program through Seoul National University (No. 200-20240057), by Semi-Supervised Learning Research Grant by SAMSUNG (No.A0426-20220118), by Identify the network of brain preparation steps for concentration Research Grant by LooxidLabs (No.339-20230001), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)] by the MSIT (Ministry of Science, ICT), Korea, under the Global Research Support Program in the Digital Field program (RS-2024-00421268) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation), by the National Supercomputing Center with supercomputing resources including technical support (KSC-2023-CRE-0568) and by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2021S1A3A2A02090597), and by Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT (MSIT, Korea) & Gwangju Metropolitan City.

# Author contributions

J.Y. and J.C. conceived the study and designed the behavioral experiment. J.Y. and S.Y.O. designed the VR movie stimuli and experimental protocol. J.Y. and D.D.H. developed the code. J.Y. performed behavioral and EEG data analysis. J.Y., D.D.H., and J.C. interpreted the results and drafted the manuscript.

# Awe is characterized as an ambivalent experience in the human behavior and cortex:

integrated virtual reality-electroencephalogram study

Jinwoo Yi[1], Danny Dongyeop Han[2], Seung-Yeop Oh[3], Jiook Cha[1,2,4,5]

**Affiliations**

[1] Department of Brain and Cognitive Sciences, Seoul National University, South Korea

[2] Interdisciplinary Program of Artificial Intelligence, Seoul National University, South Korea

[3] AIR: AI Rebirth, South Korea

[4] Department of Psychology, Seoul National University, South Korea

[5] Institute of Psychological Science, Seoul National University, South Korea

**This file includes:**

Supplementary Table 1-4

Supplementary Fig 1-5

References

**Author Note:**

We have no conflicts of interest to disclose. Correspondence concerning this article should be addressed to Jinwoo Yi, BA, and Jiook Cha, PhD, Building 16, Office M512, Gwanak-ro 1, Gwanak-gu, Seoul, South Korea. Email: Jinwoo Yi (adem1997@snu.ac.kr) and Jiook Cha (connectome@snu.ac.kr)

**Supplementary Table 1. Associations between affect-related sensory features of clips and participants' valence dynamics**

| | positive vs. neutral | | | ambivalent vs. neutral | | | negative vs. neutral | | |
|---|---|---|---|---|---|---|---|---|---|
| | $\beta$ | SE | $P$ | $\beta$ | SE | $P$ | $\beta$ | SE | $P$ |
| Time | .158 | .063 | .013*** | .052 | .065 | .424 | .063 | .079 | .428 |
| Brightness | 3.758 | .122 | $< 2\times10^{-16}$*** | 1.361 | .115 | $< 2\times10^{-16}$*** | -1.191 | .136 | $< 2\times10^{-16}$*** |
| Hue | 1.581 | .119 | $< 2\times10^{-16}$*** | .256 | .091 | .005** | -.543 | .111 | $1\times10^{-8}$*** |
| Loudness | 1.594 | .093 | $< 2\times10^{-16}$*** | .559 | .093 | $< 2\times10^{-9}$*** | 1.637 | .107 | $< 2\times10^{-16}$*** |

*Note.* We tested multinomial mixed logistic regression model predicting valence label for every time point using sensory features and two random intercepts of participants and clips as follows: $\text{valence}_t = t + \text{brightness}_t + \text{hue}_t + \text{loudness}_t + (1|\text{sub}) + (1|\text{clip})$. *$P < .05$; **$P < .01$; ***$P < .001$.

**Supplementary Table 2. Statistical differences in valence and arousal ratings across clips**

| | SP vs. PA | | | | CI vs. PA | | | | MO vs. PA | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *t* | *df* | *d* | $P_{FDR}$ | *t* | *df* | *d* | $P_{FDR}$ | *t* | *df* | *d* | $P_{FDR}$ |
| **Valence** | | | | | | | | | | | | |
| ambivalent (int) | 2.957 | 42 | .552 | .005** | 2.946 | 42 | .576 | .005** | 5.118 | 42 | .961 | $2\times10^{-5***}$ |
| ambivalent (dur) | 3.685 | 42 | .775 | .001*** | 3.372 | 42 | .770 | .002** | 5.179 | 42 | 1.218 | $2\times10^{-5***}$ |
| positive (int) | 1.427 | 42 | .266 | .352 | -.443 | 42 | .092 | .660 | 1.206 | 42 | .225 | .352 |
| positive (dur) | -3.316 | 42 | .667 | .003** | -3.618 | 42 | .795 | .002** | -3.178 | 42 | .634 | .003** |
| negative (int) | -.295 | 42 | .059 | .769 | 2.562 | 42 | .431 | .021* | 3.024 | 42 | .536 | .013* |
| negative (dur) | .421 | 42 | .082 | .676 | 4.329 | 42 | .839 | $1\times10^{-4***}$ | 5.243 | 42 | 1.085 | $1\times10^{-5***}$ |
| **Arousal** | 5.810 | 42 | 1.151 | $7\times10^{-7***}$ | 7.404 | 42 | 1.433 | $6\times10^{-9***}$ | 9.301 | 42 | 1.777 | $3\times10^{-11***}$ |

*Note.* 'int' and 'dur' denote intensity and duration of the corresponding feelings, respectively. *$P_{FDR}$ < .05; **$P_{FDR}$ < .01; ***$P_{FDR}$ < .001.

**Supplementary Table 3. Statistical differences in key demographics and awe ratings between the participants included and excluded in the electrophysiological analysis**

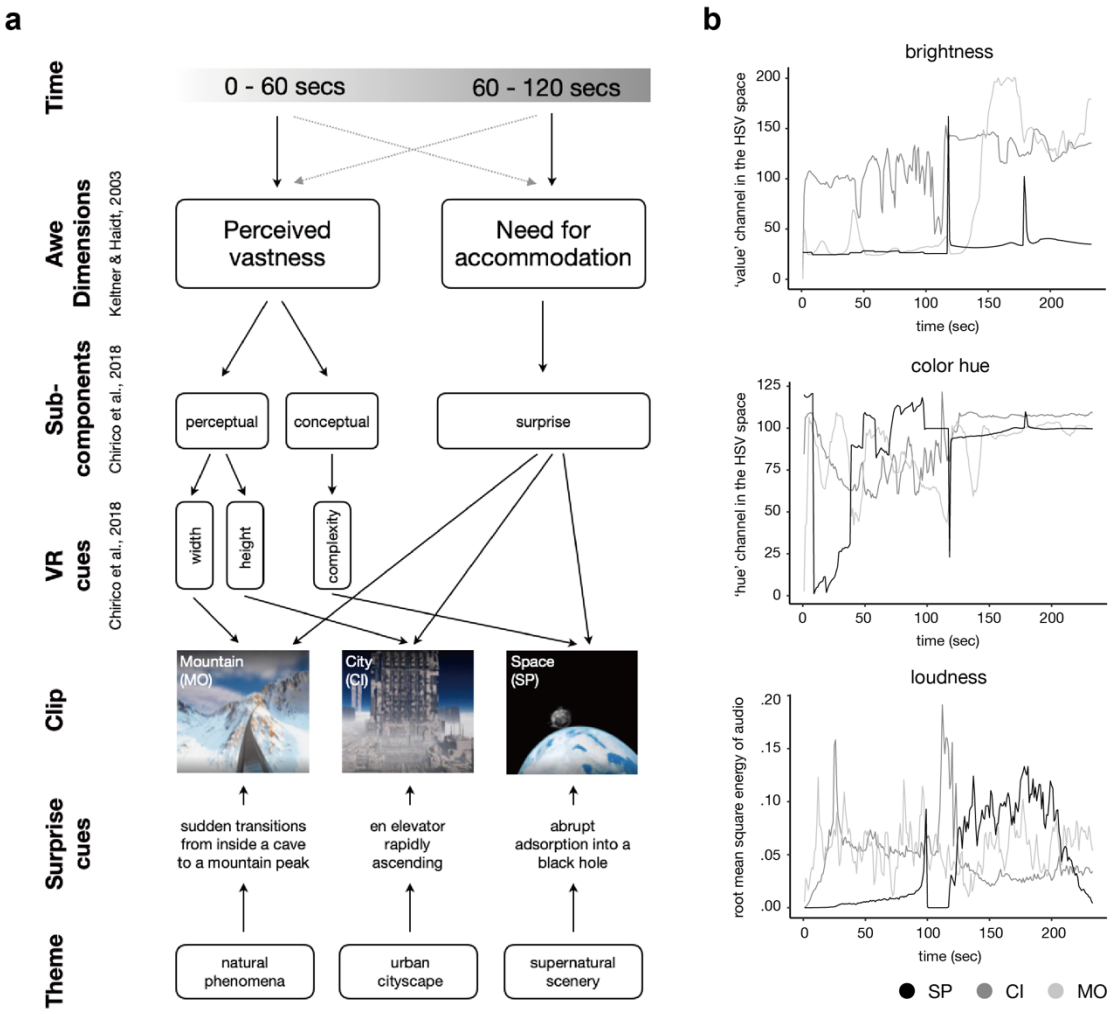
| | **Included (*N* = 27)** | **Excluded (*N* = 16)** | $P_{FDR}$ |
|---|---|---|---|
| **Demographics** | | | |
| male; *N* (%) | 13 (48.1) | 7 (43.8) | .733 |
| age years; *M(SD)* | 20.2 (1.7) | 20.3 (1.7) | .918 |
| **AWES ratings** | | | |
| for SP; *M(SD)* | 4.7 (0.6) | 4.4 (0.7) | .559 |
| for CI; *M(SD)* | 4.6 (0.5) | 3.8 (1.0) | .070 |
| for MO; *M(SD)* | 4.3 (0.8) | 4.1 (1.0) | .733 |

*Note.* Statistical difference of the categorical variable (i.e., male) between groups was assessed by Brunner-Munzel test. For the other variables, we applied two-sided t-test. All *P* values were FDR corrected.

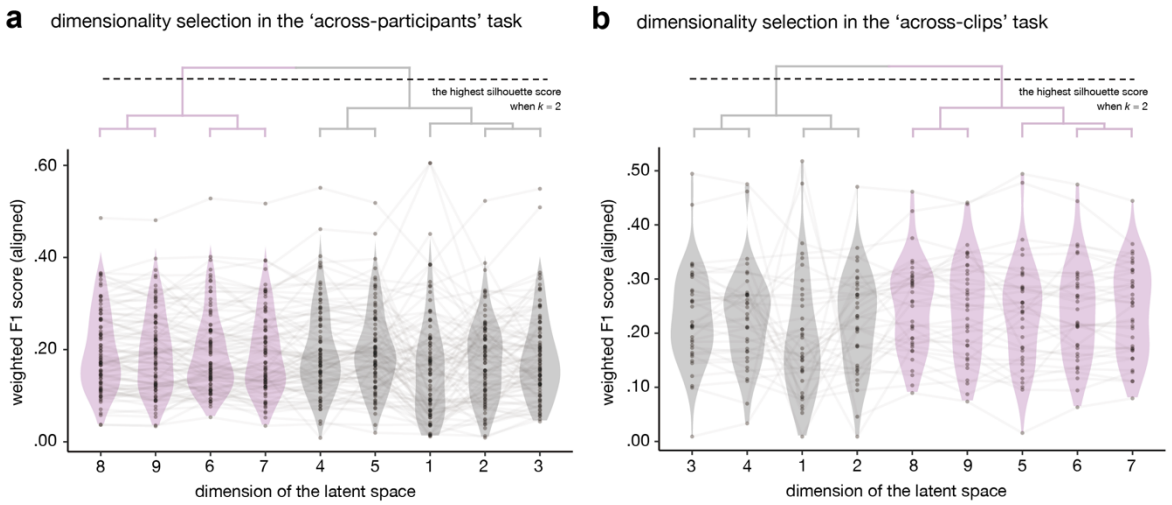
**Supplementary Table 4. Statistical differences in predictive performances of latent neural space embeddings across conditions and analytic approaches**

| | across participants | | | | across clips | | | |
|---|---|---|---|---|---|---|---|---|
| | *t* | *df* | *d* | $P_{FDR}$ | *t* | *df* | *d* | $P_{FDR}$ |
| **within CEBRA embeddings** | | | | | | | | |
| *aligned vs. random* | 10.641 | 89 | 1.122 | $5\times10^{-17***}$ | 3.005 | 35 | .501 | .015* |
| *aligned vs. not-aligned* | 5.200 | 89 | .548 | $2\times10^{-6***}$ | 1.699 | 35 | .283 | .147 |
| *not aligned vs. random* | 2.366 | 89 | .249 | .020* | .620 | 35 | .103 | .540 |
| **between embeddings** | | | | | | | | |
| *aligned* | | | | | | | | |
| CEBRA vs. PCA | 10.371 | 89 | 1.093 | $2\times10^{-16***}$ | 3.320 | 35 | .553 | .006** |
| CEBRA vs. FAA | 10.576 | 89 | 1.115 | $6\times10^{-17***}$ | 4.840 | 35 | .807 | $8\times10^{-5***}$ |
| *not-aligned* | | | | | | | | |
| CEBRA vs. PCA | 1.830 | 89 | .193 | .106 | .360 | 35 | .060 | .721 |
| CEBRA vs. FAA | 1.898 | 89 | .200 | .091 | .452 | 35 | .075 | .782 |
| *random* | | | | | | | | |
| CEBRA vs. PCA | -.379 | 89 | .040 | .706 | -.410 | 35 | .068 | .721 |
| CEBRA vs. FAA | .269 | 89 | .028 | .789 | -.278 | 35 | .046 | .782 |

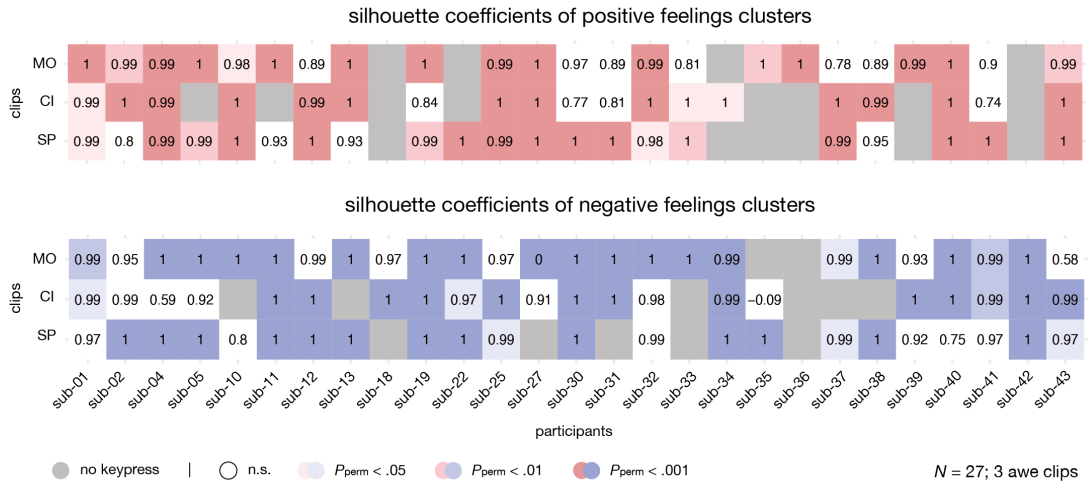
*Note.* *$P_{FDR}$ < .05; **$P_{FDR}$ < .01; ***$P_{FDR}$ < .001.



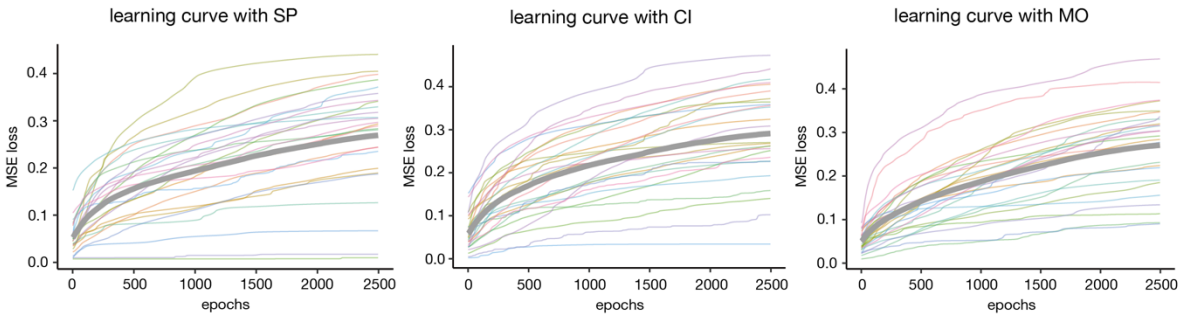

**Supplementary Fig 1. Design of VR clips to induce diverse awe experience using distinct awe-related cues and sensory profiles. a,** We designed the three awe-inducing VR clips to evoke distinct awe experiences while incorporating key dimensions of awe, each implemented through different sub-components in the VR environment. Additionally, each video was intentionally crafted with a unique. The common structure (i.e., 'awe dimensions') was based on the core dimensions of awe, which was formulated by Keltner and Haidt (2003). To ensure that each awe dimension was distinctly represented in the VR environment, we applied the qualitative framework of Chirico et al. (2018), which suggests distinct sub-components of perceived vastness and relevant VR cues. **b,** The heterogeneous temporal patterns of visual and acoustic features in three awe-inducing clips.



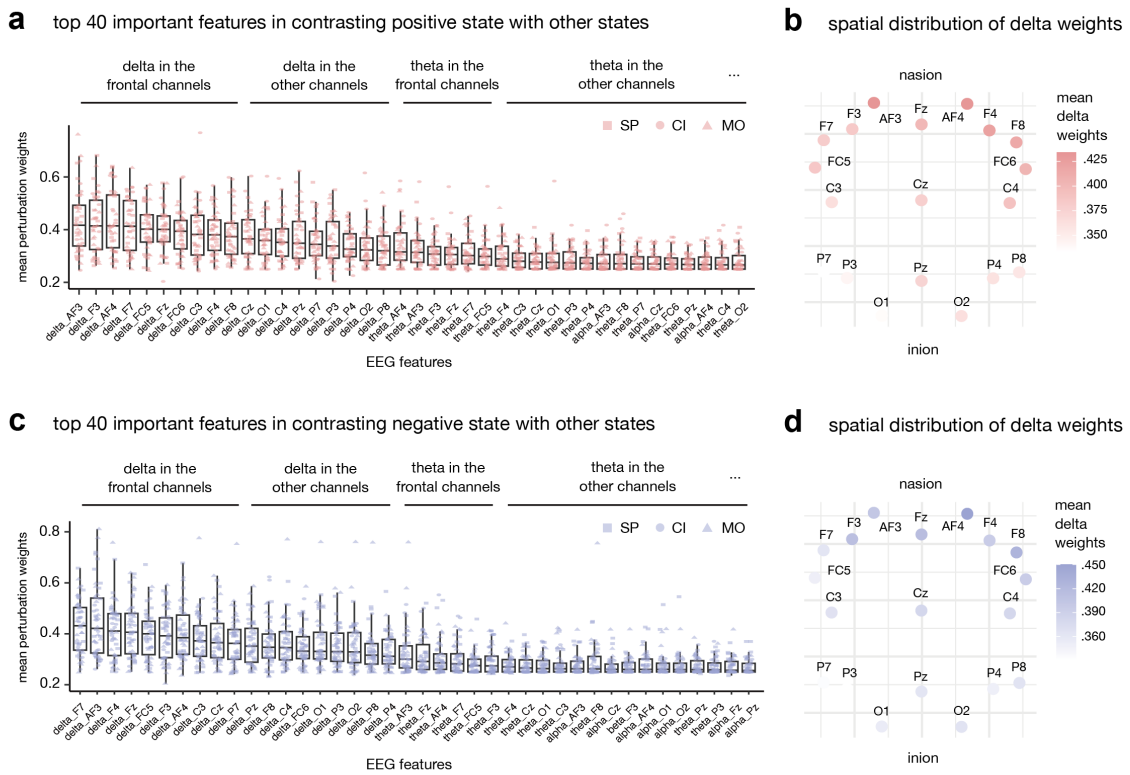

**Supplementary Fig 2. Dimensionality selection for PCA-driven latent neural space. a,** The test performances in the 'across-participants' task. Hierarchical clustering analysis identified clusters with high performance in decoding other participants' valence dynamics (purple violin plots). **b,** The test performances in the 'across-clips' task.




**Supplementary Fig 3. Silhouette coefficients of positive and negative valence clusters.** Silhouette coefficients of positive (red) and negative (blue) feelings clusters within individualized latent cortical spaces. Color intensity indicates the statistical significance of the silhouette coefficients. Some participants have a gray cell (i.e., not respond positive or negative feelings in the keypress) because 27 participants included in this analysis were chosen in terms of whether they responded ambivalent feelings for more than 5% of the total duration across all awe-inducing clips, rather than positive or negative feelings.




**Supplementary Fig 4. Learning curves of Dynamask to learn attribution map.** Dynamask learns perturbation weights of EEG features for every time point in terms of the importance in constructing latent cortical space through CEBRA (i.e., distinguishing the valence at the given timepoint with other ones). Thus, it aims to extract perturbation weights matrix so that resultant pseudo-embeddings from perturbed inputs have large mean-squared error (MSE) with the original latent space embeddings as much as possible. Each colored curve denotes its learning curve of respective participants' data. Thick gray curves visualize the average loss values at every epoch.




**Supplementary Fig 5. Perturbation weights of EEG features for positive and negative feelings. a,** Top 40 most important features for differentiating positive feelings from other valence categories. **b,** Spatial distribution of mean perturbation weights of delta-related features for positive feelings. **c,** Top 40 most important features for differentiating negative feelings from other valence categories. **b,** Spatial distribution of mean perturbation weights of delta-related features for negative feelings.

# References

Chirico, A., et al., *Designing awe in virtual reality: An experimental study.* Frontiers in psychology, 2018. **8**: p. 293522.

Keltner, D. and J. Haidt, *Approaching awe, a moral, spiritual, and aesthetic emotion.* Cognition and emotion, 2003. **17**(2): p. 297-314.