

Mind-to-Face: Neural-Driven Photorealistic Avatar Synthesis via EEG Decoding^{*}

Haolin Xiong^{*}, Tianwen Fu^{*}, Pratusha Bhuvana Prasad, Yunxuan Cai, Wenbin Teng, Haiwei Chen, Hanyuan Xiao, and Yajie Zhao[†]

University of Southern California
USC Institute for Creative Technologies
^{*}Equal contribution. [†]Corresponding author.

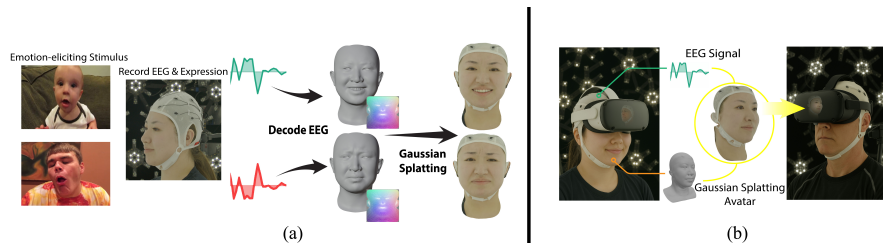


Fig. 1: Overview of Mind-to-Face, our neural-driven avatar framework. (a) We record synchronized EEG and facial expressions while subjects view emotion-eliciting stimuli, then decode the EEG into dense 3D position maps, and render photorealistic avatars using 3D Gaussian Splatting [11]. (b) While exploratory and not intended as a complete VR system, this work suggests an application: decoding EEG to drive expressive avatars when the face is occluded by a head-mounted display. Mind-to-Face points toward future systems that enable expressive avatar control under full facial occlusion.

Abstract. Current expressive avatar systems rely heavily on visual cues and often fail when faces are occluded or emotions remain internal. We present **Mind-to-Face**, the first framework to decode non-invasive electroencephalogram (EEG) signals directly into high-fidelity facial expressions. We build a dual-modality recording setup that captures synchronized EEG and multi-view facial video during emotion-eliciting stimuli, providing precise supervision for neural-to-visual learning. Our model uses a CNN-Transformer encoder to map EEG signals into dense 3D position maps that sample over 65k vertices, capturing fine-scale geometry and subtle emotional dynamics, and renders them through a modified 3D Gaussian Splatting pipeline for photorealistic, view-consistent results. Extensive evaluations show that EEG alone can reliably predict dynamic, subject-specific facial expressions, including subtle emotional responses, demonstrating that neural signals contain far richer affective and geometric information than previously assumed. **Mind-to-Face** establishes a new paradigm for neural-driven avatars, enabling personalized, emotion-aware telepresence and cognitive interaction in immersive environments.

^{*} Project page: https://tianwen-fu.github.io/mind_to_face

Keywords: Neural Decoding · Affective Computing · EEG · BCI · Gaussian Avatar · 3DGS · Light Stage

1 INTRODUCTION

Current avatar and emotion systems rely almost entirely on visual cues, such as cameras, motion tracking, or speech analysis, to infer and render human expressions. These approaches fail when faces are occluded by Head-mounted Displays (HMDs) or when emotions are internally experienced but not outwardly expressed. This limitation overlooks a fundamental fact: human emotion and cognition originate from neural activity that precedes facial movement and behavior. Yet current research treats these modalities separately—vision models reconstruct external geometry, affective models classify EEG into coarse labels (e.g., “happy” or “sad”), and brain-computer interfaces (BCIs) focus on symbolic or motor signals. A critical gap remains in linking neural states to visual expressions. Bridging this gap would enable emotionally transparent communication when visual channels are blocked, EEG-driven avatars for gaming and collaboration, adaptive avatars that reflect personalized neural patterns, and new neuro-visual datasets for decoding attention, emotion, and memory into visual form.

Electroencephalography (EEG) is a non-invasive method for measuring electrical brain activity and is widely used in brain-computer interfaces (BCIs). Prior work demonstrates strong correlations between EEG signals and emotional states [4, 50] or symbolic facial expressions [52], indicating their potential for decoding expressive information directly from brain activity. However, despite increasing interest in EEG-based applications, EEG has rarely been used as a driver for visual synthesis. Recent learning-based approaches classify discrete expression labels from EEG [15, 18], but remain limited to coarse categories and cannot capture the continuous, fine-grained dynamics required for realistic avatar control. Achieving high-fidelity reconstruction from EEG is challenging due to the low dimensionality and noisiness of neural signals.

We present Mind-to-Face, a framework that bridges neural activity and high-fidelity facial expression, as an exploratory method to enable expression decoding from brain activity rather than what cameras capture. Our approach addresses three core challenges. First, we design a dual-modality data acquisition setup that synchronously records EEG signals and multi-view facial video during emotion-eliciting stimuli, solving a key engineering challenge in aligning neural activity with visual expression. Second, we introduce a neural decoding pipeline that translates EEG patterns into photorealistic renderings via 3D Gaussian Splatting.

Unlike images used in traditional face tracking, neural responses are highly individualized: identical stimuli can produce distinct EEG signatures across subjects. Therefore, our work focuses on subject-specific expression control rather than scalable cross-subject modeling. In this sense, our approach follows the spirit of prior BCI research (e.g., [45]), where evaluation on a small number

of participants is common at exploratory stages. Accordingly, we collect a dual-modality dataset from two representative subjects across six trials, each containing multiple stimulus videos. Evaluating two subjects allows a controlled study of neural-to-geometry mapping quality, providing a foundation for future work on cross-subject generalization.

To extract features from EEG responses, we adopt a CNN-Transformer hybrid encoder inspired by EEG-Conformer [37], capturing both local spatial correlations and long-range temporal dependencies in EEG. Each EEG slice is encoded into a latent representation of brain activity, which is then decoded into a dense 3D position map [20] to produce continuous facial geometry in UV space. Unlike strain-gauge systems [14], EEG offers a richer and more holistic measure of cognitive and affective states. For photorealistic rendering, we build on the 3D Gaussian Splatting framework [11] following GaussianAvatars [28], but replace the low-dimensional FLAME model [17] with a high-degree-of-freedom 3D position map. This dense representation captures subtle geometric details and fine expression nuances that are smoothed out in low-dimensional models. We further modify the deformation and skinning system of GaussianAvatars to support this representation and incorporate a robust facial-tracking pipeline during preprocessing, enabling accurate reconstruction even under occlusion and challenging expressions.

In summary, our key contributions include:

1. We introduce *Mind-to-Face*, the first system that decodes non-invasive EEG signals into dense 3D facial geometry and renders photorealistic avatars. Our approach establishes a continuous neural-to-visual mapping, enabling expressive facial motion to be reconstructed directly from brain activity.
2. We design a custom **dual-modality data acquisition system** that synchronously records 16-channel EEG signals and high-speed multi-view facial videos while subjects watch emotion-eliciting film clips. This setup provides precise temporal alignment between neural activity and facial responses, enabling supervised learning for neural-driven 3D facial synthesis. We will publicly release the resulting synchronized EEG-to-3D dataset, which to the best of our knowledge is the first of its kind, supporting future research in neural decoding, telepresence, and affective computing.

2 RELATED WORK

2.1 EEG Decoding and Emotion Recognition

Electroencephalography (EEG) decoding is an active area of research in brain-computer interface (BCI), emotion recognition, and cognitive state estimation [38]. Traditional approaches rely on handcrafted features such as band power [49], Common Spatial Patterns (CSP) [25], or functional connectivity [46] to summarize EEG into low-dimensional descriptors, often limiting the richness of information available for learning expressive behaviors. Recent advances leverage deep learning to directly extract spatial and temporal features from raw EEG.

Convolutional networks (CNNs) are widely used for spatial filtering across channels [13, 29, 37], while RNNs [26] and Transformers [36, 37] have been explored for temporal modeling.

There are two primary theoretical models for representing human emotions. The first is the discrete model, which categorizes emotions into a fixed set of basic classes—originally six, as introduced by Ekman [5], and later expanded to include up to fifteen [24]. In contrast, the dimensional model represents emotions along continuous axes such as arousal, valence, pleasure, displeasure, sleepiness, and distress, as proposed by Russell [31]. These models have been widely adopted in emotion recognition research and dataset design.

Early EEG-based emotion recognition systems [21, 34] visualized categorical emotional states using digital avatars. The SEED dataset [51] introduced a collection of EEG recordings for tasks including sleep state detection, eye blink detection, and emotion classification. The publicly available dataset most similar to our setting is MAHNOB-HCI [33], which includes one RGB and five monochrome facial video streams synchronized with 32-channel EEG recordings. However, due to its low resolution and monochrome format, MAHNOB-HCI cannot be used for 3D facial reconstruction. To the best of our knowledge, our Mind-to-Face data capture setting is the first to collect a high-fidelity multi-view dataset that enables photogrammetric 3D face reconstruction synchronized with EEG signals as well as stimuli videos.

2.2 Head Reconstruction and Neural Avatars

Recent advances in differentiable rendering and implicit scene representations have substantially improved the realism and controllability of digital head avatars. Thies et al. [42] introduced a real-time face tracking and transfer system, establishing the foundation for neural avatar animation. Subsequent progress in face image synthesis [2, 41] has enabled more expressive control over facial expressions, lip synchronization, and head motion [12, 39, 47].

More recently, Neural Radiance Field (NeRF)–like methods [22] have since emerged as a powerful paradigm for realistic head avatar modeling and rendering. Gafni et al. [6] learn a dynamic NeRF conditioned on expression vectors from monocular video input, while Grassal et al. [7] extend the FLAME model [17] by subdividing the mesh and introducing expression-dependent geometric and texture offsets to enhance realism. Other approaches, such as IMavatar [53], employ neural implicit functions to model avatars and solve for deformations between observed and canonical poses via iterative root-finding.

Since 2023, 3D Gaussian Splatting (3DGS) [11] has emerged as a powerful yet efficient alternative to NeRF-like methods. A growing body of work has explored 3DGS for human avatar animation, including controllable full-body avatars [8, 54], motion synthesis [23, 48], and head reconstruction [40]. Notably, GaussianAvatars [28] introduced a framework that combines FLAME-based explicit expression control with 3DGS rendering, achieving high-fidelity geometry and photorealistic appearance through anisotropic splatting. Importantly, while preserving the fine-grained visual fidelity of NeRFs, 3DGS offers an explicit scene

representation that enables precise control over geometry and appearance, making it particularly well-suited for animation tasks.

Our work builds upon the rendering paradigm of GaussianAvatars [28]. Replacing the traditional blendshape and FLAME-based geometric control with a dense, per-pixel 3D position map representation, our method captures fine-grained facial details and subtle expression dynamics with significantly higher fidelity.

3 METHOD

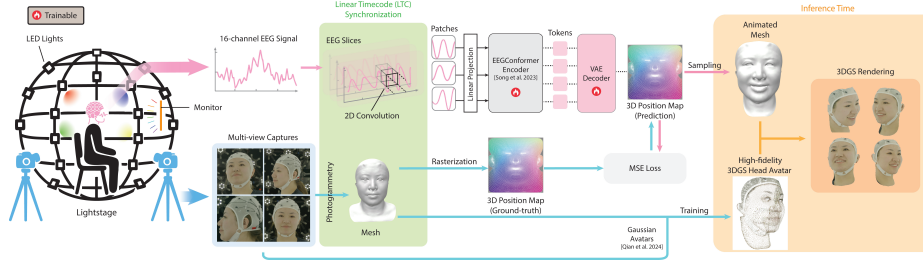


Fig. 2: Overview of Our Mind-to-Face Pipeline. Our system decodes raw EEG signals into dense 3D position maps that are rendered into photorealistic avatars. Data are collected using a custom multi-view capture rig composed of synchronized high-speed RGB cameras and a 16-channel EEG headset, with all modalities temporally aligned via Linear Timecode for frame-accurate correspondence between EEG and facial expressions. Multi-view videos are reconstructed into ground-truth 3D facial meshes through photogrammetry. During training, EEG slices are encoded using a CNN-Transformer encoder (EEG-Conformer [37]) and decoded into 3D position maps using a randomly initialized VAE decoder, supervised with MSE loss against photogrammetric position maps. At inference, the predicted position maps are resampled into meshes and rendered using a modified GaussianAvatars pipeline [28] for high-fidelity avatar synthesis.

Our objective is to decode continuous facial expressions from EEG signals. To achieve this, we first design a multi-modal data capture system (Sec. 3.1) and perform synchronized EEG-video data acquisition (Sec. 3.2) followed by signal and geometry preprocessing (Sec. 3.3). Using the processed data, we develop an EEG encoder framework based on a convolutional transformer architecture [37], coupled with an decoder [30] that learns to map brain signals to facial geometry in the format of positional map (Sec. 3.4). Finally, the reconstructed EEG-driven meshes are rendered photorealistically using a variant of GaussianAvatars pipeline [28], enabling realistic and view-consistent avatar synthesis. Our overall framework is illustrated in Fig. 2.

3.1 Data Capture Setup

Facial Performance Capturing. We record a participant’s responses to stimulus videos from multiple views under a constant lighting condition. Sixteen cinema-grade global-shutter cameras (up to 8K resolution, 120 fps) are evenly distributed around the subject at 15° intervals, providing dense view coverage for photogrammetric reconstruction. A display positioned in front of the participant presents emotionally evocative video stimuli to elicit natural expressions. An illustration of our capture setup is shown in the supplementary materials. All cameras are synchronized using genlock and timecode generators to ensure precise temporal alignment across all views for high-fidelity reconstruction.

EEG Capturing Device. We use the Cyton-Daisy biosensing kit by OpenBCI for EEG data acquisition due to its portability and reliable signal quality. The device records 16-channel EEG signals at a sampling rate of 125Hz. Electrodes are positioned according to the international 10–20 system, as illustrated in the supplementary materials, providing evenly distributed coverage across the scalp. Although the number of channels is fewer than that of clinical-grade EEG systems, this configuration offers sufficient spatial resolution to capture representative neural activity across major cortical regions [1, 10], making it well-suited for our expression decoding task. Conductive gel is applied to each electrode to maintain impedance below 500k Ω , ensuring stable and high-quality signal acquisition.

Synchronization. Synchronization between the camera array and the EEG headset is achieved through a linear signal generated by an Arduino Teensy board, ensuring frame-accurate temporal alignment between EEG measurements and corresponding facial expressions. This setup yields a dual-modality dataset with precisely synchronized EEG and multi-view facial recordings, enabling reliable cross-modal learning.

3.2 Experiments Design and Data Acquisition

Stimulus Videos. The stimulus videos are manually selected from the Open Library for Affective Videos (OpenLAV) [9] and emotion eliciting clips curated by Coan and Allen [3]. The selected clips are grouped into five emotion specific categories—neutral, disgust, funny, angry, and sad—each constituting a single trial in subsequent experiments. In addition, a separate video containing mixed emotional content is compiled from the EMOSTIM dataset [35] and used as a leave-one-trial-out testing sequence.

Experiment Procedure. The subject is seated at the center of the Light Stage while wearing the EEG headset. The stage is uniformly illuminated with a high depth-of-field lighting configuration to minimize motion blur during performance capture. An illustrative example of the capture setup is shown in the supplementary materials. Stimulus playback, EEG recording, and multi-view

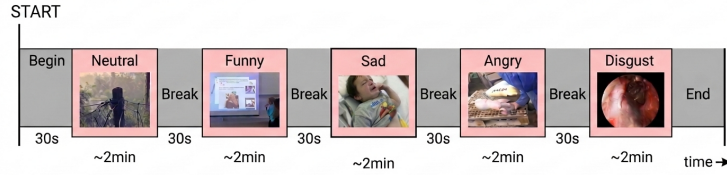


Fig. 3: Trial Design. The experiment consists of five emotion-specific trials, each separated by a 30-second rest period. Within each trial, participants view a sequence of video clips conveying a single, consistent emotion corresponding to that trial’s theme.

video capture are initiated simultaneously for precise temporal alignment. At the beginning of each trial, a prompt appears on the display indicating the target emotion category, preparing the subject’s mental state for the upcoming stimuli. The corresponding stimulus videos are then presented, followed by a 30-second rest period between trials to prevent emotional carryover effects. Each subject completes five stimulus trials, as illustrated in Fig. 3, resulting in approximately 25 minutes of total recording time. The setup yields a synchronized dual-modality dataset comprising high-resolution multi-view videos and EEG signals. In total, around 30 minutes video data per-subject is captured in compressed RAW format from 16 synchronized cameras, producing roughly 7TB of data per subject. Despite the dataset’s large size and preprocessing demands, all recordings were successfully processed and prepared for decoding.

3.3 Preprocessing

EEG Preprocessing. For each trial, the raw EEG data is represented as a time-varying signal with C channels sampled at each time step. A 6th-order Butterworth band-pass filter with a passband of 4-40Hz is applied to suppress high-frequency noise and remove low-frequency drift and DC components introduced by the recording hardware. Per-channel z-score standardization is then performed using statistics computed from the training set, with the corresponding mean and variance applied to the test set for normalization consistency. The filtered and normalized EEG signals are subsequently segmented into overlapping temporal windows, each denoted as $E^{(t)} \in \mathbb{R}^{W \times C}$, where $W = 375$ is the window size, and $C = 16$. Each EEG window $E^{(t)}$ is temporally aligned with a corresponding ground-truth 3D position map $P^{(t)}$ reconstructed via multi-view photogrammetry.

Face Geometry Reconstruction. To enable supervised learning for decoding facial expressions, we reconstruct a 3D facial mesh $M^{(t)}$ for each video frame $I^{(t)}$ captured during the trial sessions. We track the subject’s performance by first solving for the facial identity of a neutral expression frame using the 3D Morphable Model (3DMM) from ICT-FaceModel [16], and then fitting the identity template mesh to the multi-view photogrammetry observations while maintaining consistent topology by Laplacian deformation. Optical flow-based refinement is also employed for temporal stability.

We apply a rigid Procrustes alignment to eliminate the head poses and align each reconstructed mesh to a canonical coordinate frame shared across all timesteps (see details in supplementary materials) to ensure that the decoder focuses on predicting non-rigid, expression-related deformations. For each aligned mesh $M^{(t)}$, we then compute a corresponding 3D position map $P^{(t)} \in \mathbb{R}^{256 \times 256 \times 3}$, where each pixel encodes the (x, y, z) coordinates of the facial surface in UV space. This representation preserves geometric continuity while enabling dense supervision for EEG decoding. Each position map is temporally aligned with its corresponding EEG window described in Sec. 3.3, forming paired samples $(E^{(t)}, P^{(t)})$ for training the EEG-to-face decoder in Sec. 3.4.

Neural Avatar Preprocessing. We convert the predicted 3D position maps $\hat{P}^{(t)}$ into meshes by sampling vertex coordinates $\hat{\mathbf{v}}^{(t)} \in \mathbb{R}^{V \times 3}$ to form meshes $\hat{M}^{(t)}$ according to the template UV layout. While the shared topology provides spatial consistency, the temporal smoothness of the reconstructed sequence primarily came from the overlapping EEG windows used during decoding, which yields gradually varying position maps over time. To prepare training data for GaussianAvatars, each mesh $M^{(t)}$ is projected onto the corresponding multi-view images to generate binary foreground masks, which are used to extract white-background facial crops $\{I_c^{(t)}\}$. The synchronized meshes $\{\mathbf{v}^{(t)}, \mathbf{f}\}$ and masked images $\{I_c^{(t)}\}$ are then used to train our modified GaussianAvatars model.

3.4 Decoding 3D Position Maps from EEG

To map EEG signals to dense 3D position maps, we first encode the preprocessed signals using a hybrid Convolution-Transformer architecture inspired by EEG-Conformer [37]. Each EEG slice $E^{(t)} \in \mathbb{R}^{W \times C}$ is reshaped into a single-channel tensor of size $1 \times C \times W$, which can be interpreted as a spatiotemporal image capturing both temporal dynamics and inter-channel structure. Following prior work [29, 32], we apply two sequential convolutional layers: the first along the temporal axis to model temporal patterns, and the second across channels to capture spatial correlations between electrodes. The resulting features pass through batch normalization, an ELU activation, and an average-pooling layer that reduces temporal redundancy and promotes more stable, emotion-relevant representations.

The resulting feature sequence is passed through a transformer encoder that models long-range temporal dependencies within the EEG window and outputs a compact latent embedding $\mathbf{z}^{(t)}$. The position map provides a UV-parameterized mapping from surface geometry to a regular 2D grid in texture space, enabling efficient processing with standard 2D convolutional networks and image decoders. So, we employ a latent variational autoencoder (VAE) to generate the dense 3D position map $\hat{P}^{(t)}$, bridging the spatiotemporal EEG domain with the spatial structure of the human face.

A binary mask m restricts supervision to the inner-face region, where facial expression dynamics occur, encouraging the decoder to focus on expressive geometry rather than rigid head motion. To further enhance surface smoothness

and continuity, we include a self-supervised Laplacian regularization term (details in the Supplementary Materials). The overall position map loss is defined as:

$$\mathcal{L}_{\text{pm}} = \lambda_{\text{rec}} \mathcal{L}_{\text{rec}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}},$$

where $\mathcal{L}_{\text{rec}}$ denotes the per-pixel mean squared error (MSE), and $\mathcal{L}_{\text{smooth}}$ represents the Laplacian smoothing loss applied to $\hat{P}$ within the masked region.

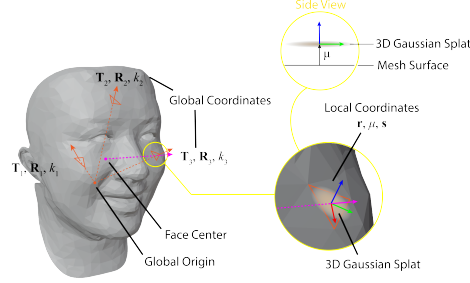


Fig. 4: Illustration of the Gaussian Binding Strategy. Each ellipsoidal Gaussian splat is defined by its local rotation $\mathbf{r}$, position μ , and scale $\mathbf{s}$ relative to the face center with global transform $\mathbf{T}$ and $\mathbf{R}$. During animation, $\mathbf{T}$ and $\mathbf{R}$ are updated over time according to the predicted facial geometry, while $\mathbf{r}$, μ , and $\mathbf{s}$ remain fixed as learnable parameters. The mesh is decimated for clarity in visualization.

3.5 Avatar Rendering with 3D Gaussian Splatting

We adopt 3DGS for avatar head rendering, replacing traditional texture-based methods with a view-dependent representation that captures realistic lighting via spherical harmonics. In this framework, a scene is modeled as a set of ellipsoidal Gaussian splats parameterized by a covariance matrix Σ and center μ :

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mu)^T \Sigma^{-1}(\mathbf{x} - \mu)\right),$$

where Σ is defined by rotation $\mathbf{r} \in \mathbb{R}^{3 \times 3}$, scale $\mathbf{s} \in \mathbb{R}^3$, and position $\mu \in \mathbb{R}^3$. Each Gaussian’s color $\mathbf{c}$ is modeled using 3rd-order spherical harmonics, and the final color is rendered by alpha compositing overlapping splats:

$$\mathbf{C} = \sum_{i=1} \mathbf{c}_i \alpha'_i \prod_{j < i} (1 - \alpha'_j)$$

At initialization, each triangular face of the mesh is assigned one Gaussian with local parameters $(\mathbf{r}, \mu, \mathbf{s})$, transformed into global coordinates during rendering as:

$$\mathbf{r}' = \mathbf{R}\mathbf{r}, \quad \mu' = k\mathbf{R}\mu + \mathbf{T}, \quad \mathbf{s}' = k\mathbf{s}, \quad (1)$$

where $\mathbf{R}$, $\mathbf{T}$, and k denote the face’s rotation, translation, and scale, respectively. This provides a geometry-aware initialization for stable reconstruction (see Fig. 4).

At the training phase, we do not compute the face geometry from the tracked FLAME parameters as in [28]. Instead, we obtain directly the topology-consistent and facial geometry from our face tracking (see details in Sec. 3.3), since it offers greater flexibility and thus fits closer to the actual vibrant facial expressions. Our face tracking pipeline also improves upon the baseline in our scenario with occlusion and extreme expressions, demonstrated in Sec. 6.1. For each iteration, we sample a triplet of (mesh M , camera p , image I) uniformly at random, and render the Gaussian splats, each with the rigid transformation defined in Eq. (1) based on M . The properties are optimized with loss

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_1 + \lambda\mathcal{L}_{\text{D-SSIM}} + \lambda_{\text{pos}}\mathcal{L}_{\text{pos}} + \lambda_{\text{scale}}\mathcal{L}_{\text{scale}}, \quad (2)$$

where $\mathcal{L}_1$ and $\mathcal{L}_{\text{D-SSIM}}$ compare the difference in the rendered images, $\mathcal{L}_{\text{pos}} = \|\max(\mu, \epsilon_{\text{pos}})\|_2$ regularizes the spatial shift from the face center that the splat is bound to, and $\mathcal{L}_{\text{scale}} = \|\max(s, \epsilon_{\text{scale}})\|_2$ regularizes scales of the Gaussians. During rendering, we move and render the optimized splats by the mesh deformed according to the position maps (see Sec. 3.4). For 3DGS training, we use the same hyperparameters as documented in [28].

4 IMPLEMENTATION DETAILS

CNN-Transformer Encoder. Each EEG input window is represented as a tensor of shape $(B, 1, 16, 375)$, where 16 denotes the electrode channels and 375 the temporal samples. The encoder begins with a temporal convolution with 40 output channels and kernel size 40 to capture short-range temporal patterns. This is followed by a spatial convolution with kernel size 16, spanning all electrodes to model spatial correlations between channels. The resulting feature maps are processed by Batch Normalization, an ELU nonlinearity, and an average pooling layer to reduce temporal resolution while preserving salient dynamics. After pooling, the temporal dimension is reduced from 375 to approximately 19 feature patches. A 1×1 convolution then projects the feature maps to an embedding dimension of $E = 40$, forming a token sequence $\mathbf{z}_0^{(t)} \in \mathbb{R}^{B \times N \times E}$ with $N = 19$.

The token sequence $\mathbf{z}_0^{(t)}$ is then processed by a stack of six Transformer encoder layers. Each layer uses pre-normalization with LayerNorm, multi-head self-attention with 10 heads, and a feed-forward network that expands the embedding dimension by a factor of 4, followed by GELU activation and dropout ($p = 0.5$). Residual connections are applied to both attention and feed-forward sub-layers. The final output $\mathbf{z}^{(t)}$ encodes temporally integrated EEG features for time window t , which are passed to the 3D position map decoder.

Position Map Decoder. To reconstruct dense 3D position maps from the EEG feature $\mathbf{z}^{(t)}$, we employ a latent variational autoencoder (VAE) as an image decoder. The decoder receives a flattened $z^{(t)}$ of size $B \times d_{\text{in}}$, which is projected through a sequence of fully connected layers into a latent feature map of dimension $B \times 4 \times 8 \times 8$. The latent tensor is then decoded through the VAE into a

coarse 3-channel output, followed by two transposed convolution layers to produce a dense 3D position map $\hat{P} \in \mathbb{R}^{256 \times 256 \times 3}$, where each pixel encodes the (x, y, z) coordinates of the reconstructed facial surface.

The decoder parameters are randomly initialized and trained jointly with the EEG transformer encoder to predict dense 3D position maps from EEG signals. The predicted face mesh is sampled from the resulting position map and used to drive the trained Gaussian splats attached to the faces to render the final facial image.

5 EXPERIMENTS

As a pioneering work, there is no direct baseline to compare against; we present visual results under different emotional stimuli of two subjects as our own qualitative results to assess system performance.

Data Split. Our data for evaluation consists of two types of trials: the emotion-specific ones (Angry, Disgust, Funny, Neutral, Sad) and the standalone evaluation trial *EMOSTIM*. Each emotion trial consists of several stimulus videos presented sequentially to the subject, during which synchronized EEG and multi-view facial recordings are captured. For the emotion-specific trials, we use responses corresponding to the final 15% of frames of each trial as the test set and the remaining frames for training, following standard signal processing practices to minimize statistical leakage. In addition, the standalone *EMOSTIM* sequence, which contains a distinct set of mixed-emotion stimulus videos, is excluded from training and used solely to evaluate cross-trial generalization.

Qualitative Results. Fig. 5 shows responses of two subjects across three stimulus types. The EEG-driven avatars reproduce the facial expressions while preserving each subject’s identity and adapting to subject-specific dynamics, such as smile intensity, indicating effective personalization rather than broad categorization. For avatar rendering, we further refine the mouth interior using a pretrained GFP-GAN [44] model, following the post-processing strategies adopted in recent neural avatar works [19, 43]. The enhancement is strictly confined to the mouth interior, which is frequently occluded during the capture and thus hard to be modeled by GS-based avatar models, and does not tamper with the decoded emotion, as we show in the third and fourth columns of the figure. More results, including multi-view rendering and videos are provided in Supplementary Materials.



Fig. 5: Qualitative Results of EEG-driven Facial Expression Synthesis. For three trials (Sad, Funny, Disgust), we show the emotion-eliciting video frames from the testing set (first row), the corresponding captured facial expressions from two subjects (second row), the corresponding renders produced by our EEG-driven avatar (third row), and our final result produced with mouth-interior enhancement (fourth row). The reconstructed avatars reflect emotional responses (*e.g.* frown for sadness, smiles for funny clips, brow/nose tightening for disgust) with subject-specific intensities and capture both neutral and strongly expressive frames. The GFP-GAN enhancement is applied only to the mouth interior and does not modify the emotion response (*cf.* the third and fourth row). Quantitative metrics can be found in supplementary materials.

6 ABLATION STUDIES

In this section, we conduct an ablation study to evaluate the impact of key components of our method and justify our design choices.

6.1 Comparison with GaussianAvatars

The GaussianAvatars pipeline is optimal in the common scenario where the expressions are mild and the majority of the face is clear. However, in our setting,

to capture synchronized image and EEG data, it is necessary for the subjects to wear a headcap with sensors, which may confuse common facial landmark decoders, such as the VHAP library [27] used in their pipeline; the extreme expressions, triggered by the emotion-eliciting videos we use, are often beyond the expressivity of FLAME parameters [17]. Therefore, as demonstrated by Fig. 6, our ablation study shows that the standard GaussianAvatar pipeline does not generalize to our data capture setup, and it may predict inaccurate expressions (b) or even fail catastrophically (c), whereas our method maintains robustness in the complex capture settings.

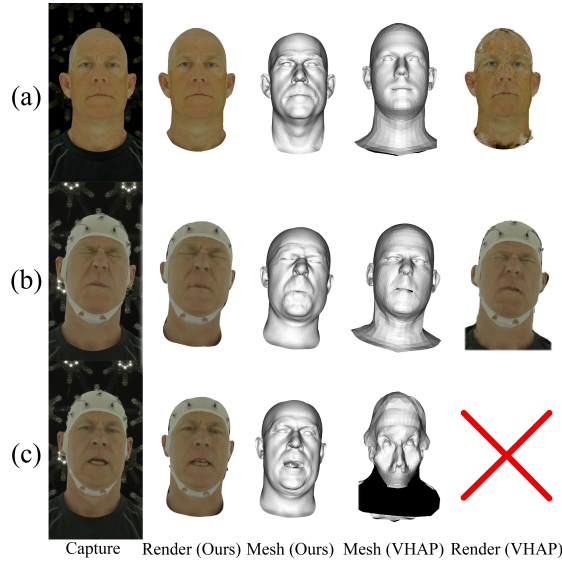


Fig. 6: We demonstrate that the face tracking pipeline of GaussianAvatars does not suit our scenario. In (a), we show that in an occlusion-free scenario, the avatar rendered by our pipeline is similar to that of GaussianAvatars. However, in our complex capture setup with extreme expressions and headcap-induced occlusion, errors in landmark detection of the GaussianAvatars pipeline can lead to incorrect expression tracking (b) or even completely broken geometry (c). The red cross indicates that the model fails to produce a faithful render due to collapsed geometry.

6.2 Why Position Maps Instead of Blendshapes?

Using blendshapes presents two key limitations. First, mapping EEG signals to a fixed set of 52 coefficients produces a sparse and less expressive representation, making it difficult to capture subtle and nuanced facial expressions. Second, standard blendshape models contain relatively few vertices ($\sim 5,000$ in FLAME), which limits their ability to represent mid-frequency geometric details

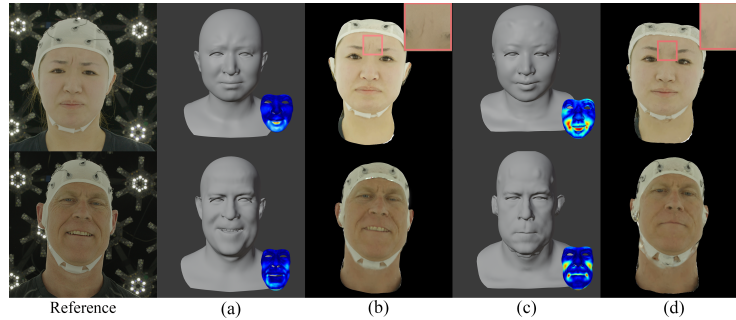


Fig. 7: Qualitative comparison between position map-driven and blendshape-driven avatars. Subject 1 (top) and Subject 2 (bottom). From left to right: Reference image, (a) position map-driven avatar mesh, (b) position map-driven avatar render, (c) blendshape-driven avatar mesh, and (d) blendshape-driven avatar render.

Trial	Position Map-driven Avatar						Blendshape-driven Avatar					
	Subject 1			Subject 2			Subject 1			Subject 2		
	Mean	< 1mm(%)	< 3mm(%)	Mean	< 1mm(%)	< 3mm(%)	Mean	< 1mm(%)	< 3mm(%)	Mean	< 1mm(%)	< 3mm(%)
Angry	0.4982	87.37	99.67	0.5534	84.28	99.64	1.0792	60.76	93.94	1.6492	42.79	83.92
Disgust	1.0238	61.89	95.12	0.6347	80.57	98.78	1.7680	42.03	81.85	1.3410	60.35	85.66
Funny	0.9614	69.12	92.82	0.7432	78.24	96.53	1.3385	56.56	87.08	1.6195	48.93	83.58
Neutral	0.4028	92.85	100.0	0.3245	95.84	99.92	0.3270	94.99	99.95	1.3065	52.91	89.98
Sad	0.3571	94.60	99.98	0.3112	96.60	99.55	0.5391	84.95	99.59	1.3134	56.22	88.15
EMOSTIM	0.5158	86.76	99.16	0.7077	78.88	97.55	0.6838	78.78	97.78	1.5078	52.22	86.28

Table 1: Quantitative evaluation using point-to-surface distance within the masked facial region. Errors are reported in millimeters (mm), including the mean distance and the percentage of vertices below 1 mm and 3 mm. The EMOSTIM trial is used as a leave-one-out trial to evaluate cross-trial generalization.

such as wrinkles. In contrast, our approach directly maps EEG signals to position maps derived from ground-truth multi-view stereo reconstructions. A 256×256 position map ($\sim 65k$ vertices) preserves mid-frequency details that blendshape models cannot capture. By avoiding the two-step dimensionality reduction inherent in blendshape-based approaches, the position-map representation better preserves the underlying geometric fidelity.

To evaluate the practical impact of these representations, we compare avatars driven by predicted blendshapes against those driven by predicted position maps. As shown in Fig. 7, the blendshape-driven reconstructions produce plausible facial motion but appear overly smoothed and exhibit reduced expressive amplitude—a consequence of the limited dimensionality of the blendshape basis. Conversely, the position map representation preserves richer geometric detail, allowing wrinkles and localized facial structures to emerge more naturally in the rendered avatars.

Tab. 1 further quantifies this difference. Across all trials except for Neutral, the position map-driven avatar consistently achieves lower geometric error.

These results suggest that the compact blendshape basis acts as a strong bottleneck when mapping noisy signals, often leading the model toward “averaged-out” predictions. The higher-resolution geometry of the position map yields more accurate surface reconstruction; specifically, the percentage of vertices with error below 1 mm is consistently higher for the position map-driven avatar. Ultimately, these results indicate that the dense position map representation not only produces visually richer animations but also yields more accurate geometric reconstructions compared to the low-dimensional blendshape representation in an EEG-driven context.

7 CONCLUSIONS AND FUTURE WORK

We introduce Mind-to-Face, the first framework that maps non-invasive EEG signals directly to photorealistic 3D facial expressions. By decoding EEG into dense 3D position maps and rendering them with 3D Gaussian Splatting, our approach enables subject-specific facial animation driven entirely by neural activity. A synchronized EEG–video capture setup provides the precise alignment needed for supervised learning, establishing a unified pipeline for neural-driven avatars and new opportunities in affective computing, telepresence, and brain–computer interfaces. In future work, we plan to expand the dataset with more subjects and diverse stimuli to improve robustness and cross-subject generalization, and incorporate stimulus context during decoding to better disentangle perceptual and affective signals for more accurate facial synthesis.

References

1. Ameri, S.K., Kim, M., Kuang, I.A., Perera, W.K., Alshiekh, M., Jeong, H., Topcu, U., Akinwande, D., Lu, N.: Imperceptible electrooculography graphene sensor system for human–robot interface. *npj 2D Materials and Applications* **2**(1), 19 (2018). <https://doi.org/10.1038/s41699-018-0064-4> 6
2. Chan, C., Ginosar, S., Zhou, T., Efros, A.A.: Everybody Dance Now. In: *Proceedings of the IEEE/CVF international conference on computer vision*. pp. 5933–5942 (2019) 4
3. Coan, J.A., Allen, J.J.: *Handbook of emotion elicitation and assessment*. Oxford university press (2007) 6
4. Duan, R.N., Zhu, J.Y., Lu, B.L.: Differential entropy feature for eeg-based emotion classification. In: *International IEEE/EMBS Conference on Neural Engineering*. pp. 81–84 (11 2013). <https://doi.org/10.1109/NER.2013.6695876> 2
5. Ekman, P.: An argument for basic emotions. *Cognition and Emotion* **6**(3-4), 169–200 (1992). <https://doi.org/10.1080/02699939208411068> 4
6. Gafni, G., Thies, J., Zollhöfer, M., Nießner, M.: Dynamic neural radiance fields for monocular 4d facial avatar reconstruction. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. pp. 8649–8658 (June 2021) 4
7. Grassal, P.W., Prinzler, M., Leistner, T., Rother, C., Nießner, M., Thies, J.: Neural head avatars from monocular rgb videos. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 18653–18664 (2022) 4

8. Hu, L., Zhang, H., Zhang, Y., Zhou, B., Liu, B., Zhang, S., Nie, L.: Gaussianavatar: Towards realistic human avatar modeling from a single video via animatable 3d gaussians. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (2024) 4
9. Israel, L., Paukner, P., Schiestel, L., Diepold, K., Schönbrodt, F.: Open library for affective videos (OpenLAV) (2021) 6
10. Jiang, L., Stocco, A., Losey, D.M., Abernethy, J.A., Prat, C.S., Rao, R.P.N.: Brain-Net: A Multi-Person Brain-to-Brain Interface for Direct Collaboration Between Brains. *Scientific Reports* **9**(1), 6115 (2019). <https://doi.org/10.1038/s41598-019-41895-7> 6
11. Kerbl, B., Kopanas, G., Leimkuehler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.* **42**(4) (Jul 2023). <https://doi.org/10.1145/3592433>, <https://doi.org/10.1145/3592433> 1, 3, 4
12. Kim, H., Garrido, P., Tewari, A., Xu, W., Thies, J., Nießner, M., Pérez, P., Richardt, C., Zollhöfer, M., Theobalt, C.: Deep video portraits. *ACM Transactions on Graphics (TOG)* **37**(4), 163 (2018) 4
13. Lawhern, V.J., Solon, A.J., Waytowich, N.R., Gordon, S.M., Hung, C.P., Lance, B.J.: Eegnet: a compact convolutional neural network for eeg-based brain-computer interfaces. *Journal of Neural Engineering* **15**(5), 056013 (Jul 2018). <https://doi.org/10.1088/1741-2552/aace8c>, <http://dx.doi.org/10.1088/1741-2552/aace8c> 4
14. Li, H., Trutoiu, L., Olszewski, K., Wei, L., Trutna, T., Hsieh, P.L., Nicholls, A., Ma, C.: Facial-performance-sensing head-mounted display. *ACM Transactions on Graphics* **34** (08 2015). <https://doi.org/10.1145/2766939> 3
15. Li, R., Liu, D., Li, Z., Liu, J., Zhou, J., Liu, W., Liu, B., Fu, W., Alhassan, A.B.: A novel EEG decoding method for a facial-expression-based BCI system using the combined convolutional neural network and genetic algorithm. *Frontiers in Neuroscience* **16**, 988535 (2022). <https://doi.org/10.3389/fnins.2022.988535> 2
16. Li, R., Bladin, K., Zhao, Y., Chinara, C., Ingraham, O., Xiang, P., Ren, X., Prasad, P., Kishore, B., Xing, J., Li, H.: Learning formation of physically-based face attributes (2020) 7
17. Li, T., Bolkart, T., Black, M.J., Li, H., Romero, J.: Learning a model of facial shape and expression from 4D scans. *ACM Transactions on Graphics, (Proc. SIGGRAPH Asia)* **36**(6), 194:1–194:17 (2017), <https://doi.org/10.1145/3130800.3130813> 3, 4, 13
18. Li, X., Zhang, X., Yang, H., Duan, W., Dai, W., Yin, L.: An eeg-based multi-modal emotion database with both posed and authentic facial actions for emotion analysis. In: 2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020). p. 336–343. IEEE (Nov 2020). <https://doi.org/10.1109/fg47880.2020.00050>, <http://dx.doi.org/10.1109/FG47880.2020.00050> 2
19. Li, X., De Mello, S., Liu, S., Nagano, K., Iqbal, U., Kautz, J.: Generalizable one-shot 3d neural head avatar. *Advances in Neural Information Processing Systems* **36**, 47239–47250 (2023) 11
20. Liu, S., Cai, Y., Chen, H., Zhou, Y., Zhao, Y.: Rapid face asset acquisition with recurrent feature alignment. *ACM Transactions on Graphics (TOG)* **41**(6), 1–17 (2022) 3
21. Liu, Y., Sourina, O., Nguyen, M.K.: Real-time eeg-based human emotion recognition and visualization. In: *Proceedings of the 2010 International Conference on Cyberworlds*. p. 262–269. CW '10, IEEE Computer Society, USA (2010).

- <https://doi.org/10.1109/CW.2010.37>, <https://doi.org/10.1109/CW.2010.37> 4
22. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: representing scenes as neural radiance fields for view synthesis. *Commun. ACM* **65**(1), 99–106 (Dec 2021). <https://doi.org/10.1145/3503250>, <https://doi.org/10.1145/3503250> 4
 23. Moon, G., Shiratori, T., Saito, S.: Expressive whole-body 3d gaussian avatar. In: ECCV (2024) 4
 24. Nadeem, M., Imam, R., Al-Refai, R., Chkir, M., Hoda, M., El Saddik, A.: EVOKE: Emotion Enabled Virtual Avatar Mapping Using Optimized Knowledge Distillation. In: 2024 IEEE International Conference on Consumer Electronics (ICCE). pp. 1–6. IEEE (2024) 4
 25. Panachakel, J.T., H, R., K, S.P., Sidharth, S., Samuel, A.A.: Csp- lstm based emotion recognition from eeg signals. In: 2023 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRaine). pp. 289–294 (2023). <https://doi.org/10.1109/MetroXRaine58569.2023.10405666> 3
 26. Postepski, F., Wojcik, G.M., Wrobel, K., Kawiak, A., Zemla, K., Sedek, G.: Recurrent and convolutional neural networks in classification of eeg signal for guided imagery and mental workload detection (2025) 4
 27. Qian, S.: Vhap: Versatile head alignment with adaptive appearance priors (sep 2024). <https://doi.org/10.5281/zenodo.14988309>, <https://github.com/ShenhanQian/VHAP> 13
 28. Qian, S., Kirschstein, T., Schoneveld, L., Davoli, D., Giebenhain, S., Nießner, M.: Gaussianavatars: Photorealistic head avatars with rigged 3d gaussians. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20299–20309 (2024) 3, 4, 5, 10
 29. Riyad, M., Khalil, M., Adib, A.: Mi-eegnet: A novel convolutional neural network for motor imagery classification. *Journal of Neuroscience Methods* **353**, 109037 (2021). <https://doi.org/https://doi.org/10.1016/j.jneumeth.2020.109037>, <https://www.sciencedirect.com/science/article/pii/S016502702030460X> 4, 8
 30. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022) 5
 31. Russell, J.A.: A circumplex model of affect. *Journal of Personality and Social Psychology* **39**(6), 1161–1178 (1980). <https://doi.org/10.1037/h0077714> 4
 32. Schirrmester, R.T., Springenberg, J.T., Fiederer, L.D.J., Glasstetter, M., Eggensperger, K., Tangermann, M., Hutter, F., Burgard, W., Ball, T.: Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping* **38**(11), 5391–5420 (2017). <https://doi.org/https://doi.org/10.1002/hbm.23730>, <https://onlinelibrary.wiley.com/doi/abs/10.1002/hbm.23730> 8
 33. Soleymani, M., Lichtenauer, J., Pun, T., Pantic, M.: A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing* **3**(1), 42–55 (2012). <https://doi.org/10.1109/T-AFFC.2011.25> 4
 34. Sollfrank, T., Kohnen, O., Hilfiker, P., Kegel, L.C., Jokeit, H., Brugger, P., Lertscher, M.L., Rey, A., Mersch, D., Sternagel, J., Weber, M., Grunwald, T.: The Effects of Dynamic and Static Emotional Facial Expressions of Humans and Their Avatars on the EEG: An ERP and ERD/ERS Study. *Frontiers in Neuroscience* **15**, 651044 (2021). <https://doi.org/10.3389/fnins.2021.651044> 4

35. Somarathna, R., Vuilleumier, P., Mohammadi, G.: Emostim: A database of emotional film clips with discrete and componential assessment. *IEEE Trans. Affect. Comput.* **15**(3), 1202–1212 (Jul 2024). <https://doi.org/10.1109/TAFFC.2023.3328900>, <https://doi.org/10.1109/TAFFC.2023.3328900> 6
36. Song, Y., Jia, X., Yang, L., Xie, L.: Transformer-based spatial-temporal feature learning for EEG decoding. *arXiv preprint arXiv:2106.11170* (2021) 4
37. Song, Y., Zheng, Q., Liu, B., Gao, X.: EEG conformer: Convolutional transformer for EEG decoding and visualization. *IEEE Transactions on Neural Systems and Rehabilitation Engineering* **31**, 710–719 (2023). <https://doi.org/10.1109/tnsre.2022.3230250> 3, 4, 5, 8
38. Sun, C., Mou, C.: Survey on the research direction of EEG-based signal processing. *Frontiers in Neuroscience* **17**, 1203059 (2023). <https://doi.org/10.3389/fnins.2023.1203059> 3
39. Suwajanakorn, S., Seitz, S.M., Kemelmacher-Shlizerman, I.: Synthesizing obama: learning lip sync from audio. *ACM Trans. Graph.* **36**(4) (Jul 2017). <https://doi.org/10.1145/3072959.3073640>, <https://doi.org/10.1145/3072959.3073640> 4
40. Tang, J., Davoli, D., Kirschstein, T., Schoneveld, L., Niessner, M.: Gaf: Gaussian avatar reconstruction from monocular videos via multi-view diffusion. In: *Proceedings of the Computer Vision and Pattern Recognition Conference*. pp. 5546–5558 (2025) 4
41. Thies, J., Zollhöfer, M., Nießner, M.: Deferred Neural Rendering: Image Synthesis using Neural Textures. *Acm Transactions on Graphics (TOG)* **38**(4), 1–12 (2019) 4
42. Thies, J., Zollhofer, M., Stamminger, M., Theobalt, C., Nießner, M.: Face2Face: Real-time Face Capture and Reenactment of RGB Videos. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp. 2387–2395 (2016) 4
43. Tran, P., Zakharov, E., Ho, L.N., Tran, A.T., Hu, L., Li, H.: Voodoo 3d: Volumetric portrait disentanglement for one-shot 3d head reenactment. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. pp. 10336–10348 (2024) 11
44. Wang, X., Li, Y., Zhang, H., Shan, Y.: Towards real-world blind face restoration with generative facial prior. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. pp. 9168–9178 (2021) 11
45. Willett, F.R., Kunz, E.M., Fan, C., Avansino, D.T., Wilson, G.H., Choi, E.Y., Kamdar, F., Hochberg, L.R., Druckmann, S., Shenoy, K.V., Henderson, J.M.: A high-performance speech neuroprosthesis. *bioRxiv* (2023). <https://doi.org/10.1101/2023.01.21.524489>, <https://www.biorxiv.org/content/early/2023/04/25/2023.01.21.524489> 2
46. Wolpaw, J.R., del R. Millán, J., Ramsey, N.F.: Chapter 2 - brain-computer interfaces: Definitions and principles. In: Ramsey, N.F., del R. Millán, J. (eds.) *Brain-Computer Interfaces, Handbook of Clinical Neurology*, vol. 168, pp. 15–23. Elsevier (2020). <https://doi.org/https://doi.org/10.1016/B978-0-444-63934-9.00002-0>, <https://www.sciencedirect.com/science/article/pii/B9780444639349000020> 3
47. Xu, Y., Zhang, H., Wang, L., Zhao, X., Huang, H., Qi, G., Liu, Y.: LatentAvatar: Learning Latent Expression Code for Expressive Neural Head Avatar. In: *ACM SIGGRAPH 2023 Conference Proceedings*. pp. 1–10 (2023) 4
48. Yuan, Y., Li, X., Huang, Y., De Mello, S., Nagano, K., Kautz, J., Iqbal, U.: Gavatar: Animatable 3d gaussian avatars with implicit mesh learning. In: *Proceedings of the*

- IEEE/CVF conference on computer vision and pattern recognition. pp. 896–905 (2024) 4
49. Zhang, Y., Yan, G., Chang, W., Huang, W., Yuan, Y.: Eeg-based multi-frequency band functional connectivity analysis and the application of spatio-temporal features in emotion recognition. *Biomedical Signal Processing and Control* **79**, 104157 (2023). <https://doi.org/https://doi.org/10.1016/j.bspc.2022.104157>, <https://www.sciencedirect.com/science/article/pii/S1746809422006115> 3
 50. Zheng, W.L., Liu, W., Lu, Y., Lu, B.L., Cichocki, A.: Emotionmeter: A multimodal framework for recognizing human emotions. *IEEE Transactions on Cybernetics* **49**(3), 1110–1122 (2018). <https://doi.org/10.1109/TCYB.2018.2797176> 2
 51. Zheng, W.L., Lu, B.L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. *IEEE Transactions on Autonomous Mental Development* **7**(3), 162–175 (2015). <https://doi.org/10.1109/tamd.2015.2431497> 4
 52. Zheng, W.L., Lu, B.L.: A multimodal approach to estimating vigilance using eeg and forehead eog. *Journal of Neural Engineering* **14**(2), 026017 (feb 2017). <https://doi.org/10.1088/1741-2552/aa5a98>, <https://dx.doi.org/10.1088/1741-2552/aa5a98> 2
 53. Zheng, Y., Abrevaya, V.F., Bühler, M.C., Chen, X., Black, M.J., Hilliges, O.: I M Avatar: Implicit morphable head avatars from videos. In: *Computer Vision and Pattern Recognition (CVPR)* (2022) 4
 54. Zielonka, W., Bagautdinov, T., Saito, S., Zollhöfer, M., Thies, J., Romero, J.: Drivable 3d gaussian avatars. In: *I3DV* (March 2025) 4