BRAIN2FACE: A STUDY ON THE RELATION OF EEG SIGNALS AND SPATIO-TEMPORAL HEAD AVATARS

Anonymous authors
Paper under double-blind review

ABSTRACT

Driving digital head avatars is classically done by hand-crafted artist design or by image-based facial performance capture. In this work, we ask whether spatio-temporal tracking of the humans's head geometry and appearance can be performed solely from brain signals, i.e., Electroencephalography (EEG). Since prior work mostly focused on activity and emotion recognition, this task is relatively unexplored without even having publicly available high-quality datasets and benchmarks. Thus, we contribute a dataset featuring paired EEG and human head geometry as well as appearance annotations. We synchronize 64-channel EEG with per-frame expression latent codes of a personalized Gaussian head avatar. Second, we introduce a baseline to the task of expression recovery of a head avatar from EEG signals. To this end, we propose a dedicated EEG encoder that takes a temporal sequence of EEG signals as input and regresses the latent expression parameters of the 3D Gaussian head avatar. Lastly, we validate our capture and algorithmic design choices through a benchmark protocol that we introduce to this task. We believe this work paves the way towards better understanding of the relation between EEG signals and facial surface as well as appearance changes. Code and data will be made publicly available upon acceptance.

1 INTRODUCTION

Digital human avatars play a central role in virtual reality, telepresence, and human-computer interaction, where the ability to faithfully convey facial expressions is crucial for realism and communication. Existing avatar control paradigms typically rely on manual digital artist design or on visual observations, such as monocular or multi-view images, depth sensors, or facial markers, which inherently require a clear line of sight and explicit user engagement. In contrast, brain-computer interfaces (BCIs) offer a fundamentally different interaction modality by enabling direct control from neural signals. Among these, electroencephalography (EEG) stands out due to its non-invasive nature and practical deployment. Leveraging EEG signals to drive expressive facial avatars would enable hands-free, camera-free avatar control, opening new possibilities for immersive communication, accessibility, and assistive technologies.

Image-based facial performance capture has been studied extensively, from multi-view reconstruction to learned facial models and photorealistic avatars Beeler et al. (2010); Daněček et al. (2022); Qian et al. (2024). EEG-based interfaces have primarily focused on discrete prediction tasks, including affective state recognition Koelstra et al. (2012); Soleymani et al. (2012) and clinical pathology detection Gemein et al. (2020); Schirrmeister et al. (2017), rather than continuous facial motion. Related synthesis approaches generate facial emotion images Esmaeili & Kiani (2024), EEG-conditioned talking faces Park et al. (2024), or intentional facial expressions through brain decoding and generative image synthesis Shimizu et al. (2025). Meanwhile, multi-view capture systems Beeler et al. (2010) and expressive photorealistic head models Qian et al. (2024); Teotia et al. (2025) provide detailed geometry and appearance representations for facial animation. The only work that studied EEG-driven photorealistic avatars is Mind-to-Face Xiong et al. (2025). However, the capture modalities are limited, i.e., EEG channels and views, and data is not publicly available for research. Moreover, there task definition differs from ours (we also refer to related work).

To address these limitations, we introduce a dataset of paired EEG and spatio-temporally coherent and fine-grained facial performance capture data. We captured multiple subjects in a multi-view

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

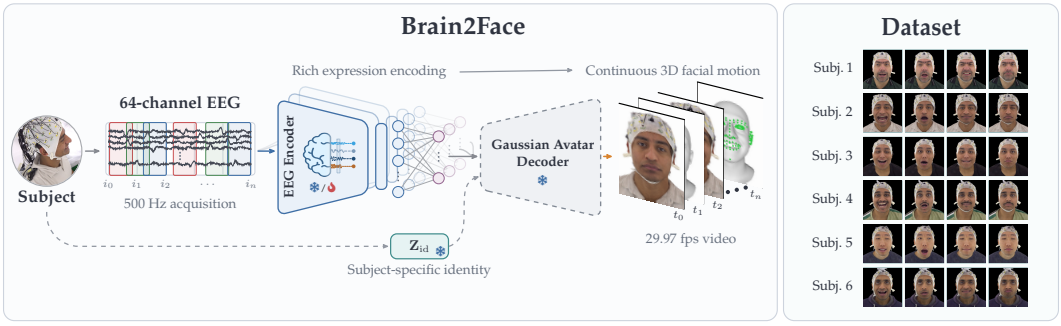

Fig. 1: **Brain2Face.** (*Left*) A 64-channel EEG recording is processed in ordered temporal blocks independently by the shared **EEG-to-Expression Mapper** (decoder) to predict window-level expression-latent chunks. Together with the subject-specific identity latent $\mathbf{Z}_{\text{id}}$, the **Gaussian Avatar Decoder** renders the EEG-driven facial avatar into image space. (*Right*) Fitted targets for six participants (rows) and four expressions (columns) of our newly acquired dataset.

capture setup with up to 33 cameras, paired with 64-channel EEG signals. To recover high-quality face data, we fit a state-of-the-art 3D Gaussian avatar to the data and, thus, encode expressions into low dimensional latent codes. An EEG encoder and prediction head map temporal EEG signals to these expression latents. Supervision is applied in latent and geometry space. We also refer to Fig. 1 for an overview.

Since this is the first study of its kind, we introduce an evaluation protocol for this task and a first baseline as well as ablations with respect to EEG foundation models El Ouahidi et al. (2025); Guetschel et al. (2024), the effect of pretrained weights in general, and the influence of different prediction heads, to only name a few. We believe that our benchmark can serve as a starting point for future works on this task.

To summarize, this work makes several key contributions:

- First, we present a dataset that provides EEG signals paired with high-resolution facial performance capture data as well as dense annotations (encoded as latent expression parameters).
- Second, we propose a model for predicting facial expressions from EEG signals, which allows driving a Gaussian head avatar.
- Third, we ablate several baselines and design factors on our newly introduced benchmark, which serves as a starting point for future studies.

2 Related Work

EEG-based Brain-Computer Interfaces. Electroencephalography (EEG) captures the summation of postsynaptic potentials from cortical pyramidal neurons, providing high-temporal-resolution voltage fluctuations measured at the scalp Gkintoni & Halkiopoulos (2025). Since its discovery by Hans Berger in the 1920s, EEG has evolved from a clinical device into a versatile tool for decoding neural activity in both constrained laboratory and everyday environments Gkintoni & Halkiopoulos (2025); Beyrouthy et al. (2024). The bulk of EEG-based brain-computer interface (BCI) research has focused on *classification* tasks, mapping neural signals to a small, fixed set of categories. A prominent line of work concerns affective state recognition, leveraging multi-modal datasets such as DEAP Koelstra et al. (2012) and MAHNOB Soleymani et al. (2012) to classify valence or arousal labels. Recent methods combine EEG with video and audio streams to improve emotion classification through multimodal fusion Wang et al. (2025b). In the clinical domain, deep convolutional and temporal convolutional networks have demonstrated strong performance for discrete EEG pathology detection Gemein et al. (2020); Schirrmeister et al. (2017), and the Braindecode toolbox Aristimunha et al. (2025) has made such architectures widely accessible. More recently, large-scale EEG foundation models have been proposed to learn general-purpose

representations transferable across subjects and tasks El Ouahidi et al. (2025). A smaller body of work has explored *continuous regression* from EEG, targeting low-dimensional motor outputs. Hosseini and Shalchyan Hosseini & Shalchyan (2022) decode two-dimensional hand-movement trajectories from 63-channel EEG using phase-based connectivity features and linear regression, while Lu et al. Lu et al. (2026) extend continuous decoding to polar-coordinate hand-motion angles. Bridging neural signals and facial motion, Shimizu et al. Shimizu et al. (2025) interface brain decoding with generative image synthesis to restore intentional facial expressions for ALS patients, and Park et al. Park et al. (2024) generate lip-synchronised talking faces conditioned on EEG for dynamic communication. Esmaeili and Kiani Esmaeili & Kiani (2024) synthesise discrete facial emotion images from EEG using conditional adversarial networks. However, none of these works has investigated learning and studying a continuous mapping from EEG signals to facial geometry and appearance parameterized as a latent expression space of a Gaussian head avatar prior.

Facial Performance Capture. Recovering fine-grained facial geometry and appearance from image streams has been studied extensively. Monocular methods Blanz & Vetter (1999); Li et al. (2017); Feng et al. (2021); Daněček et al. (2022) fit parametric face models to single RGB images or video, leveraging shape, expression, and appearance priors to regress 3D morphable model coefficients or neural blendshape parameters. Multi-view systems Beeler et al. (2010) improve reconstruction fidelity by triangulating correspondences across calibrated cameras, enabling denser and more accurate surface recovery. More recent approaches build implicit neural or Gaussian representations directly from video footage, capturing subject-specific geometry and view-dependent appearance with high visual fidelity Gafni et al. (2021); Zheng et al. (2022); Qian et al. (2024). While egocentric setups using head-mounted displays Zhang et al. (2024) have relaxed the requirement for external camera rigs, all of these paradigms share a common dependency: they require a clear line of sight to the subject's face and operate under visual sensing assumptions. EEG-based facial performance capture offers a fundamentally different operational regime: it is inherently camera-free, robust to occlusion and variable illumination, and imposes no constraints on the user's head orientation. Moreover, image-based methods cannot be trivially adapted to EEG inputs, since EEG signals have entirely different statistical properties, spatial structure, and signal-to-noise characteristics compared to pixel data, necessitating dedicated architectures and datasets.

Photoreal Head Avatars. Recent years have seen remarkable advances in the fidelity of animatable head avatars. Neural radiance field (NeRF) based approaches Mildenhall et al. (2020); Gafni et al. (2021) and, more recently, 3D Gaussian splatting representations Kerbl et al. (2023); Qian et al. (2024); Xu et al. (2024) have enabled photorealistic rendering of dynamic faces at interactive frame rates. A range of methods construct person-specific neural avatars conditioned on expression codes derived from 3DMM fits Gafni et al. (2021); Zheng et al. (2022) or tracked blendshape rigs Qian et al. (2024), achieving impressive reconstruction quality. Universal avatar priors that generalize across identities have been proposed to disentangle subject appearance from expression control, allowing a single model to animate novel subjects given only a short capture Teotia et al. (2025). In this work, we adopt such a universal Gaussian head avatar as our rendering backbone and treat its learned expression latent space as the regression target for our EEG encoder. Our goal is to study EEG-to-expression regression using a learned avatar latent space and to compare EEG representations under controlled evaluation protocols.

The most closely related work is Mind-to-Face Xiong et al. (2025), which also reconstructs photo-realistic, dynamic 3D facial expressions from EEG. Their system records synchronized 16-channel EEG and multi-view video during emotion-eliciting stimuli, maps EEG through a CNN-Transformer encoder to dense 3D position maps, and renders the resulting geometry with a Gaussian avatar pipeline. Their study evaluates personalized mappings for two subjects. Brain2Face instead maps 64-channel EEG to the 256-dimensional expression codes of a personalized universal Gaussian head avatar and studies six subjects performing a predefined repertoire of 75 expression types across two sessions. Our emphasis is on comparing EEG foundation models and initialization regimes, with held-out sessions and expression types used to distinguish reconstruction of seen and unseen expressions. Thus, our dataset offers richer annotations and also the task and study significantly differs from theirs.



Fig. 2: Overview of the **Brain2Face** dataset. Our dataset comprises paired EEG and multi-view video data covering multiple facial expressions, views, and subjects.

3 BRAIN2FACE DATASET

The Brain2Face dataset pairs 64-channel EEG with multi-view recordings of controlled facial expressions. Personalized Gaussian head avatars Teotia et al. (2025) provide continuous expression targets disentangled from subject identity. In this section, we describe the capture setup and avatar fitting. Tab. 1 summarizes the statistics of the collected dataset and Fig. 2 shows qualitative examples.

Tab. 1: **Brain2Face dataset statistics.** Paired EEG and multi-view face recordings across all subjects and sessions.

Property	Value
Subjects	6
Sessions per subject	2
Total capture time	4 h 15 min
Expression types per session	75
Expression data	74 min 15 s
EEG channels	64
EEG sampling rate	500 Hz
Cameras per session	26–33
Video frame rate	29.97 fps

3.1 CAPTURE SETUP

Face Capturing. Subjects are recorded by a multi-view setup of 26 to 33 cameras recording video at $f_v = 29.97$ fps. The multi-view images are used to fit a personalized Gaussian head avatar Teotia et al. (2025) to each subject (Sec. 3.2), yielding continuous expression codes and a fixed identity code. The fitted expression codes provide continuous regression targets. Decoding them yields the reference meshes, 3D landmarks, and avatar renders used for supervision and evaluation (Sec. 3.2).

EEG Capturing. Electroencephalography (EEG) measures scalp voltage fluctuations arising from summed postsynaptic activity with high temporal resolution Gkintoni & Halkiopoulos (2025). We acquire 64-channel EEG at 500 Hz using an EasyCap montage, with conductive gel controlling electrode impedance.

Recording Sessions. Every subject follows a standardized experimental protocol with a fixed set of expressions (Secs. A.2 and B in the supplementary material). A baseline EEG recording quantifies the subject's resting neural activity. Then, the subject performs a set of pre-defined facial expressions in two sessions, following video demonstrations of each target expression, separated by a short break.

Each session starts and ends with three jaw clenches that serve as a synchronization signal (Sec. A.1 in the supplementary material).

3.2 Gaussian Avatar Fitting

We build on the Universal Head Avatar Prior (UHAP) introduced by UniGAHA (Teotia et al., 2025), which disentangles identity and expression for photorealistic head rendering using 3D Gaussian splatting (Kerbl et al., 2023). Fitting UHAP to each subject's multi-view recordings yields an identity code $\mathbf{Z}_{\mathrm{id}} \in \mathbb{R}^{512}$ and per-frame expression codes $\mathbf{Z}^t_{\mathrm{exp}} \in \mathbb{R}^{256}$. The personalized decoder $\mathscr{D}_{\mathrm{UHAP}}$ maps the identity and expression codes to N_g Gaussian primitives and a guide mesh, $(\{\mathbf{g}_{k,t}\}_{k=1}^{N_g}, \mathbf{V}_t) = \mathscr{D}_{\mathrm{UHAP}}(\mathbf{Z}_{\mathrm{id}}, \mathbf{Z}^t_{\mathrm{exp}})$. A differentiable Gaussian-splatting renderer R produces the reference image $\mathbf{I}_t = R(\{\mathbf{g}_{k,t}\}_{k=1}^{N_g})$, and a fixed vertex-selection operator Π extracts the 3D landmarks $\mathbf{L}_t = \Pi(\mathbf{V}_t)$. Camera arguments are suppressed for readability; predicted and reference images use the same camera. The images $\mathbf{I}_t$ are fitted-avatar renders, while the captured multi-view images are used to fit the avatar. The personalized decoder and identity code remain frozen during EEG training. Sec. A.3 in the supplementary material describes fitting, inversion, and the construction of the reference geometry.

4 Method

The objective of *Brain2Face* is continuous control of a personalized 3D Gaussian head avatar from EEG signals. Given a preprocessed EEG window $\tilde{\mathbf{X}} \in \mathbb{R}^{C \times N_e}$, a predictor F estimates the corresponding expression-code sequence $\hat{\mathbf{Z}}_{\mathrm{exp}} = F(\tilde{\mathbf{X}}) \in \mathbb{R}^{d \times N_z}$, where $C = 64$, $d = 256$, and N_e and N_z denote the numbers of EEG samples and target frames. Each predicted code drives the frozen personalized avatar decoder defined in Sec. 3.2.

4.1 EEG-to-Expression Encoder

Data preprocessing. We follow the pretraining preprocessing of REVE (El Ouahidi et al., 2025, Sec. 3.1.1): resampling to $f_e = 200$ Hz, 0.5–99.5 Hz band-pass filtering, and clipping at ± 15 standard deviations per session.

Model architecture. We factor the predictor as $F = \mathscr{M}_\phi \circ \mathscr{E}_\theta$, where $\mathscr{E}_\theta$ encodes the EEG window and $\mathscr{M}_\phi$ maps its features to the expression-code sequence. The encoder can be initialized from an EEG foundation model and either frozen or fine-tuned. For the EEG encoder, we use REVE-base El Ouahidi et al. (2025); the mapper flattens the encoder's channel and patch dimensions, projects the features with a 1×1 convolution, and applies L residual convolution blocks along the resulting channel-patch token axis. A learned resampler and a framewise output projection produce N_z expression codes with $d = 256$ coordinates each. We use $L = 2$ for the principal model and study the decoder depth, i.e., the mapper depth L, in Sec. 5.3. During training, an auxiliary head pools the intermediate residual-block features and maps them to a second expression-code sequence $\hat{\mathbf{Z}}^{\mathrm{aux}}_{\mathrm{exp}}$ of the same shape, which supervises the mapper's intermediate representations and is discarded at inference. The foundation-model comparison instead uses a mapper with two fully connected hidden layers to probe each encoder's frozen features (Sec. 5.2). Sec. A.5 provides further details.

Replacing $\mathbf{Z}^t_{\mathrm{exp}}$ with $\hat{\mathbf{Z}}^t_{\mathrm{exp}}$ in the same decoder and renderer gives the predicted mesh $\hat{\mathbf{V}}_t$, landmarks $\hat{\mathbf{L}}_t$, and image $\hat{\mathbf{I}}_t$.

4.2 Supervision

We supervise the predicted expression codes and decoded geometry with the corresponding fitted-avatar targets. Writing each term with its prediction and target arguments, the training objective is

$$\mathscr{L}_{\mathrm{train}} = \lambda_p \mathscr{L}_p(\hat{\mathbf{Z}}_{\mathrm{exp}}, \mathbf{Z}_{\mathrm{exp}}) + \lambda_v \mathscr{L}_v(\hat{\mathbf{V}}, \mathbf{V}) + \lambda_\ell \mathscr{L}_\ell(\hat{\mathbf{L}}, \mathbf{L}) + \lambda_{\mathrm{aux}} \mathscr{L}_{\mathrm{aux}}(\hat{\mathbf{Z}}^{\mathrm{aux}}_{\mathrm{exp}}, \mathbf{Z}_{\mathrm{exp}}). \quad (1)$$

Here, the individual losses sum the predicted and target quantities over the evaluated frames in the training batch and we omit the time index for readability. The first three terms measure latent

Tab. 2: **Frozen EEG foundation model comparison.** Results on test sequences with seen expressions for frozen pretrained EEG encoders with a nonlinear MLP head, evaluated on Subject-03. REVE [scratch] trains the same encoder and head from random initialization, with the encoder unfrozen, with the same data and compute.

Model	LPIPS ↓	Blendshape MAE ↓	Latent MSE ↓	Mesh MAE [mm] ↓	Lmk Wing [mm] ↓
BENDR Kostas et al. (2021)	0.0607	0.0259	0.242	0.906	2.737
BIOT Yang et al. (2023)	0.0671	0.0570	0.247	1.106	2.768
CBraMod Wang et al. (2025a)	0.0624	0.0266	0.253	0.914	2.791
CodeBrain Ma et al. (2026)	0.0587	0.0249	0.231	0.748	2.563
EEGPT Wang et al. (2024)	0.0659	0.0567	0.247	0.877	2.754
LaBraM Jiang et al. (2024)	0.0817	0.0476	0.236	1.308	2.751
LUNA Döner et al. (2025)	0.0704	0.0558	0.254	0.990	2.883
Signal-JEPA Guetschel et al. (2024)	0.0669	0.0566	0.247	1.003	2.742
REVE El Ouahidi et al. (2025)	**0.0575**	**0.0238**	**0.209**	**0.725**	**2.499**
REVE [scratch]	0.0603	0.0258	0.254	1.713	2.737

mean squared error, mesh mean absolute error, and the landmark Wing loss (Feng et al., 2018), which combines a logarithmic penalty on small landmark errors with an ℓ_1 penalty on large ones. The auxiliary term is the mean absolute error between the target codes and the auxiliary sequence $\hat{\mathbf{Z}}_{\text{exp}}^{\text{aux}}$ predicted from intermediate mapper features. It regularizes those features by requiring them to already decode the full expression trajectory. The MLP mapper carries no auxiliary head and therefore omits this term. Geometry gradients pass through the frozen avatar decoder to the trainable predictor parameters. Exact loss definitions and weights are given in Sec. A.4, with optimization and inference details in the supplementary material.

5 Experiments

Next, we study the feasibility of predicting facial expressions from EEG signals. We introduce the evaluation protocol (Sec. 5.1), report our experimental findings (Sec.5.2-5.5), and, finally, show qualitative results of the overall pipeline (Sec. 5.6).

5.1 Evaluation Protocol

Metrics. To quantitatively evaluate the performance of the face predictions, we use the following evaluation metrics: Mesh MAE measures the mean absolute coordinate error over the decoded mesh vertices. Landmark wing measures a robust coordinate-wise deviation of the selected facial landmarks. Blendshape MAE measures the mean absolute error across 52 expression coefficients. LPIPS Zhang et al. (2018) measures perceptual distance between predicted and fitted-avatar ground-truth renders. Additional details about the metrics are provided in the supplementary. All metrics are averaged over all evaluated frames within each recording-expression sequence, then over sequences.

Test splits. We train and evaluate a separate model for each subject using the cross-session split described in Sec. A.2. Both test settings contain sequences excluded from supervised EEG-model training. For *seen expressions* (S), the expression type occurs in training in the opposite recording session; for *unseen expressions* (U), it is absent from training in both sessions. Both sessions contribute training sequences, so this protocol does not evaluate transfer to a completely unseen recording session or participant. The encoder and decoder-depth comparisons use Subject-03, the initialization comparison reports the mean across six subjects, and the channel-masking comparison averages all six subjects.

5.2 On the effectiveness of EEG foundation models for face decoding

We compare nine pretrained EEG encoders $\mathscr{E}_\theta$ by freezing their parameters and training an MLP mapper $\mathscr{M}_\phi$ with two fully connected hidden layers to predict expression-code trajectories. Each predictor maps the full EEG input window to the corresponding target trajectory. Tab. 2 reports results on Subject-03 test sequences with seen expression types.

Tab. 3: **REVE initialization and decoder-depth comparison.** REVE with $L = 1$–4 decoder blocks on seen expression types held out in the test session, evaluated on Subject-03, comparing frozen and unfrozen EEG encoders.

Decoder depth	LPIPS ↓	Blendshape MAE ↓	Latent MSE ↓	Mesh MAE [mm] ↓	Lmk Wing [mm] ↓
Pretrained frozen					
$L = 1$	0.0650	0.0470	0.178	1.034	2.445
$L = 2$	0.0643	0.0434	0.188	1.074	2.447
$L = 3$	0.0624	0.0449	**0.168**	1.003	2.395
$L = 4$	**0.0608**	**0.0424**	**0.168**	**0.989**	**2.386**
Pretrained unfrozen					
$L = 1$	0.0644	0.0465	0.197	1.054	2.525
$L = 2$	0.0568	**0.0345**	**0.164**	**0.943**	**2.195**
$L = 3$	0.0559	0.0385	0.188	1.033	2.436
$L = 4$	**0.0554**	0.0389	0.206	1.068	2.424

Tab. 4: **Effect of foundation-model initialization.** Average results over all six subjects, evaluating held-out test sequences with expressions seen or unseen during training.

Initialization	LPIPS ↓	Blendshape MAE ↓	Latent MSE ↓	Mesh MAE [mm] ↓	Lmk Wing [mm] ↓
Seen expressions (S; held-out session for each expression)					
Scratch	0.0736	0.0463	0.270	1.237	2.820
Pretrained frozen	0.0701	0.0453	0.235	1.163	2.611
Pretrained unfrozen	**0.0651**	**0.0412**	**0.214**	**1.046**	**2.430**
Unseen expressions (U; absent from training)					
Scratch	0.0791	0.0581	0.305	1.305	2.897
Pretrained frozen	0.0735	0.0532	0.262	1.182	2.652
Pretrained unfrozen	**0.0687**	**0.0487**	**0.246**	**1.134**	**2.508**

Among the frozen pretrained encoders, REVE achieves the lowest error on all five reported metrics. Training the same REVE encoder and MLP architecture with randomly initialized weights yields higher errors under the same data and compute budget. In this scratch baseline, both the encoder and mapper are trainable. We therefore use the pretrained REVE for the subsequent experiments.

5.3 Effect of decoder depth

We vary the number L of residual blocks in the mapper $\mathscr{M}_\phi$ defined in Sec. 4.1, comparing frozen and fine-tuned REVE encoders in Tab. 3. All runs use one-second EEG and target windows on Subject-03. With the encoder frozen, $L = 4$ gives the lowest errors, tying $L = 3$ on latent MSE at the reported precision. With fine-tuning, $L = 2$ gives the lowest blendshape MAE, latent MSE, mesh MAE, and landmark Wing loss, while $L = 4$ has the lowest LPIPS. We use $L = 2$ for the subsequent experiments.

5.4 On the effectiveness of EEG pre-training

In this experiment, we more systematically study the effects of unfreezing a pre-trained EEG encoder and show the results in Tab. 4. First, as indicated above, we find that pre-training is beneficial, as from scratch training is substantially worse than both pre-trained variants. Generally, unfreezing the encoder is effective for both, seen and unseen expressions, as the collected EEG data is still different to the pre-training data of REVE. However, it can be observed that the performance gaps between the frozen and the unfrozen settings decrease for unseen expressions. Hence, fine-tuning the encoder is less beneficial for out-of-distribution expressions. This provides an interesting observation for future studies on the generalization capabilities of EEG foundational models.

Tab. 5: **Channel-masking ablations.** Average subject results using pretrained-unfrozen REVE-base and the $L = 2$ decoder.

Variant (zeroed)	LPIPS ↓	Blendshape MAE ↓	Latent MSE ↓	Mesh MAE [mm] ↓	Lmk Wing [mm] ↓
Seen expressions (S; held-out session for each expression)					
All channels (0)	0.0651	0.0412	**0.214**	1.046	2.430
Bad channels (4–19)	0.0668	**0.0398**	0.219	**1.039**	**2.426**
Motor region (20)	**0.0649**	0.0408	0.216	1.051	2.439
Frontal (15)	0.0676	0.0433	0.221	1.082	2.537
Occipital (10)	0.0669	0.0427	0.221	1.077	2.486
Temporal (10)	0.0674	0.0439	0.234	1.100	2.519
Unseen expressions (U; absent from training)					
All channels (0)	0.0687	0.0487	0.246	1.134	2.508
Bad channels (4–19)	0.0694	0.0485	0.249	1.139	2.522
Motor region (20)	0.0701	0.0492	0.252	1.145	2.536
Frontal (15)	**0.0682**	**0.0482**	**0.241**	**1.122**	**2.505**
Occipital (10)	0.0703	0.0489	0.252	1.160	2.547
Temporal (10)	0.0694	0.0492	0.255	1.153	2.550

5.5 On the Informativeness of Different Brain Areals for Face Decoding

To analyze what the model has learned, we study the effect of removing electrodes corresponding to various brain areas, similar to region-wise electrode ablations in prior EEG decoding work (Song et al., 2024; Hu et al., 2026). For this, we retrain models where the respective channels are masked out, i.e., set to zero. Masking motor or prefrontal channels degrades facial expressions reconstructed from EEG (Hu et al., 2026). Furthermore, the frontal and temporal groups also lie closest to the frontalis and temporalis muscles, whose EMG is strongest at these sites and attenuates toward central and posterior sites (Goncharova et al., 2003), whereas the occipital group lies over the visual cortex (Song et al., 2024). Further, we also study the effect of masking out bad channels, which Bigdely-Shamlo et al. (2015) identify by extreme amplitudes, high-frequency noise, or low correlation with, and poor predictability from, the other channels. Bad channels are detected per subject on training expressions with PyPREP Appelhoff et al. (2025), a Python implementation of the PREP pipeline Bigdely-Shamlo et al. (2015). The results are presented in Tab. 5. The findings reveal that for seen in-distribution expressions, a substantial amount of the performance can be attributed to the temporal electrodes, as removing the influence of the those electrodes results in a substantial performance drop. Performance drops are observed when masking other brain areas, including the motor cortex. In contrast, removing bad channels reduces the effect of artifacts in the EEG, eventually even improving the performance compared to experiment where access to all electrodes is given. Interestingly, for unseen expressions, the effect of removing electrodes capturing the frontal brain areal is contrary compared to the seen expression: Results substantially improve when their effect is removed. We hypothesize that this could be attributed to the fact that the motion artifacts in the frontal brain area observed in the EEG signal provide less consistent information for unseen expressions than for seen expressions. In contrast, the signals obtained from the motor cortex electrodes are more consistent, resulting in a performance drops that are comparable to masking temporal brain areas.

5.6 Qualitative Results

Fig. 3 compares EEG-driven predictions with fitted-avatar ground truth for three subjects using the pretrained-unfrozen model. Seen expression types occur in training in the opposite session, whereas unseen types are excluded from training in both sessions. All displayed sequences are held out. The examples show recognizable expression structure in both groups: the seen mouth-and-nose asymmetry of Subject-01 and the unseen open-mouth expressions of Subjects 04 and 05 are reflected in the predictions. Fine details remain imperfect, including lip closure in Subject-05's seen example and the mouth shape and eye closure in Subject-01's unseen example. As each pair shares the same personalized avatar, the comparison shows expression reconstruction rather than identity synthesis.

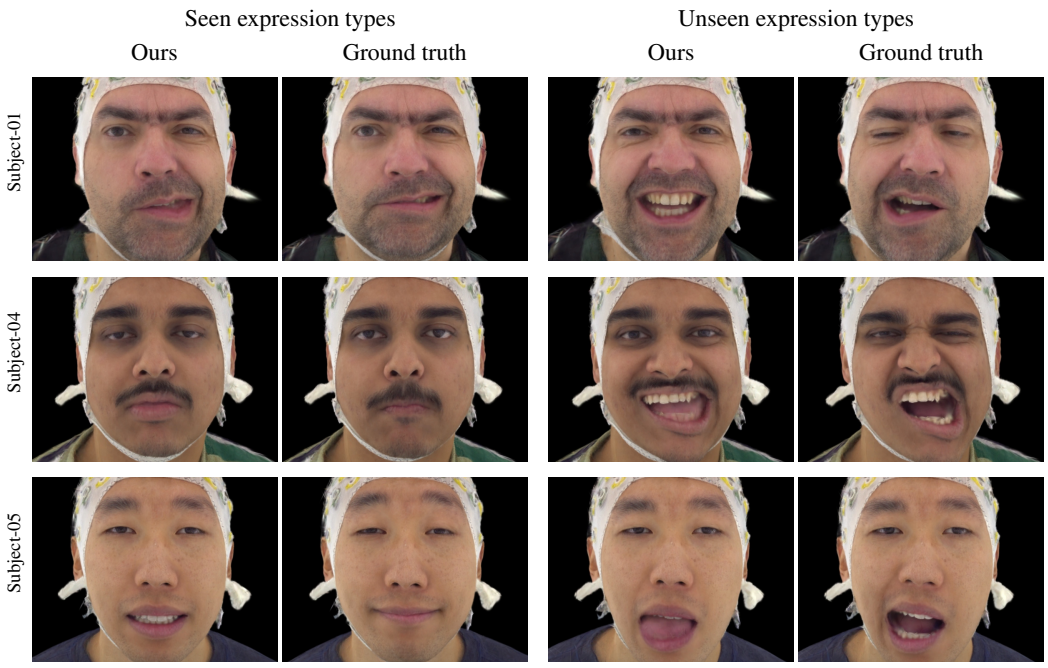

Fig. 3: **Qualitative reconstruction of seen and unseen expression types.** Each row shows one subject, with Ours and Ground truth paired for each expression group. Ground truth denotes the fitted Gaussian-avatar render at the matching frame. Seen/unseen examples are mouth and nose left/jaw right with teeth (Subject-01), left cheek puffed/jaw left with teeth (Subject-04), and stretched smile/jaw right with teeth (Subject-05).

6 Limitations and Future Work

Our initial dataset contains only six participants recorded in two sessions each, motivating larger studies with more participants and recording sessions to assess generalization. As recorded EEG signal are substantially noisier than classical image-driven signals, accurate facial tracking is more challenging with current reconstructions not capturing the level of detail available in image-driven approaches. Future work could explore complementary sensing modalities, such as functional near-infrared spectroscopy (fNIRS), to investigate whether they provide additional information for facial expression decoding. Finally, our framework models only facial expressions and does not account for movements of other body parts - movements that could also impact the signal quality. Future work can also analyze further which EEG signal components drive the prediction of which facial articulations.

7 Conclusion

In this work, we introduced the task of mapping EEG signals to face articulations. For this, we presented a systematically captured dataset that allows studying the relation between the two modalities. Subjects were recorded with an EEG and captured in a multi-camera rig that allows highly accurate reconstructions. We further presented a first baseline to predict face articulations and we studied various design dimensions including the effect of using pre-trained EEG foundational models, the importance of using expressive decoders, and the importance of different brain areals for the prediction of facial expressions. We hope that the dataset and the first analysis serves the community and encourages further analysis on the informativeness of EEG signals to infer facial geometry and articulations.

AI DISCLOSURE

We used AI tools to refine text formatting, grammar, and language, assist with table preparation, and support code development and figure generation through agentic coding. All AI-assisted outputs were reviewed and verified by the authors.

ETHICS STATEMENT

Facial avatars and EEG data were handled with respect for participant privacy, consent, and data security. Their use was limited to the agreed research purposes. Facial avatars were not used to impersonate individuals or misrepresent their identity.

REFERENCES

Stefan Appelhoff, Austin J. Hurst, Aamna Lawrence, Adam Li, Yorguin José Mantilla Ramos, Christian O'Reilly, Liang Xiang, Jonte Dancker, Mathieu Scheltienne, Ole Bialas, Nabil Alibou, Ayush Agarwal, and John Veillette. PyPREP: A Python implementation of the preprocessing pipeline (PREP) for EEG data. Zenodo, version 0.5.0, 2025. URL `https://github.com/sappelhoff/pyprep`. doi: 10.5281/zenodo.16039994.

Bruno Aristimunha, Pierre Guetschel, Martin Wimpff, Lukas Gemein, Cedric Rommel, Hubert Banville, Maciej Sliwowski, Daniel Wilson, Simon Brandt, Théo Gnassounou, Joseph Paillard, Bruna Junqueira Lopes, Sara Sedlar, Thomas Moreau, Sylvain Chevallier, Alexandre Gramfort, and Robin Tibor Schirrmeister. Braindecode: toolbox for decoding raw electrophysiological brain data with deep learning models, 2025. URL `https://github.com/braindecode/braindecode`.

Thabo Beeler, Bernd Bickel, Paul Beardsley, Bob Sumner, and Markus Gross. High-quality single-shot capture of facial geometry. *ACM TOG*, 29(3):40:1–40:9, 2010. doi: 10.1145/1778765.1778777.

Taha Beyrouthy, Nour Mostafa, Ahmed Roshdy, Abdullah S Karar, and Samer Alkork. Review of eeg-based biometrics in 5g-iot: Current trends and future prospects. *Applied sciences*, 14(2):534, 2024.

Nima Bigdely-Shamlo, Tim Mullen, Christian Kothe, Kyung-Min Su, and Kay A. Robbins. The PREP pipeline: standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, 9:16, 2015. doi: 10.3389/fninf.2015.00016.

Volker Blanz and Thomas Vetter. A morphable model for the synthesis of 3D faces. In *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '99, pp. 187–194. ACM Press/Addison-Wesley Publishing Co., 1999. doi: 10.1145/311535.311556.

Radek Daněček, Michael J. Black, and Timo Bolkart. EMOCA: Emotion driven monocular face capture and animation. In *CVPR*, pp. 20311–20322, 2022.

Berkay Döner, Thorir Mar Ingolfsson, Luca Benini, and Yawei Li. LUNA: Efficient and topology-agnostic foundation model for EEG signal analysis. *arXiv preprint arXiv:2510.22257*, 2025.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0760.

Masoumeh Esmaeili and Kourosh Kiani. Generating personalized facial emotions using emotional eeg signals and conditional generative adversarial networks. *Multimedia Tools and Applications*, 83(12):36013–36038, 2024.

Yao Feng, Haiwen Feng, Michael J. Black, and Timo Bolkart. Learning an animatable detailed 3D face model from in-the-wild images. *ACM TOG*, 40(4):88:1–88:13, 2021. doi: 10.1145/3450626.3459936.

Zhen-Hua Feng, Josef Kittler, Muhammad Awais, Patrik Huber, and Xiao-Jun Wu. Wing loss for robust facial landmark localisation with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. URL https://openaccess.thecvf.com/content_cvpr_2018/html/Feng_Wing_Loss_for_CVPR_2018_paper.html.

Guy Gafni, Justus Thies, Michael Zollhöfer, and Matthias Nießner. Dynamic neural radiance fields for monocular 4D facial avatar reconstruction. In *CVPR*, pp. 8649–8658, 2021.

Lukas A.W. Gemein, Robin T. Schirrmeister, Patryk Chrabąszcz, Daniel Wilson, Joschka Boedecker, Andreas Schulze-Bonhage, Frank Hutter, and Tonio Ball. Machine-learning-based diagnostics of eeg pathology. *NeuroImage*, 220:117021, 2020. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2020.117021. URL https://www.sciencedirect.com/science/article/pii/S1053811920305073.

Evgenia Gkintoni and Constantinos Halkiopoulos. Mapping eeg metrics to human affective and cognitive models: An interdisciplinary scoping review from a cognitive neuroscience perspective. *Biomimetics*, 10(11):730, 2025.

I. I. Goncharova, D. J. McFarland, T. M. Vaughan, and J. R. Wolpaw. EMG contamination of EEG: spectral and topographical characteristics. *Clinical Neurophysiology*, 114(9):1580–1593, 2003. doi: 10.1016/S1388-2457(03)00093-2.

Pierre Guetschel, Thomas Moreau, and Michael Tangermann. S-JEPA: Towards seamless cross-dataset transfer through dynamic spatial attention. In *Proceedings of the 9th Graz Brain-Computer Interface Conference*, pp. 11–16, 2024. doi: 10.3217/978-3-99161-014-4-003.

Seyyed Moosa Hosseini and Vahid Shalchyan. Continuous decoding of hand movement from eeg signals using phase-based connectivity features. *Frontiers in Human Neuroscience*, Volume 16 - 2022, 2022. ISSN 1662-5161. doi: 10.3389/fnhum.2022.901285. URL https://www.frontiersin.org/journals/human-neuroscience/articles/10.3389/fnhum.2022.901285.

Jingjing Hu, Dan Guo, Haofan Cheng, Ying Zeng, Zhan Si, Jinxing Zhou, and Meng Wang. See the emotion: A facial emoji proxy modeling for EEG emotion recognition. In *International Conference on Machine Learning*, 2026. URL https://arxiv.org/abs/2607.02912.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D Gaussian splatting for real-time radiance field rendering. *ACM TOG*, 42(4), 2023. doi: 10.1145/3592433.

Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012. doi: 10.1109/T-AFFC.2011.15.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021. doi: 10.3389/fnhum.2021.653659.

Tianye Li, Timo Bolkart, Michael J. Black, Hao Li, and Javier Romero. Learning a model of facial shape and expression from 4D scans. *ACM TOG*, 36(6):194:1–194:17, 2017. doi: 10.1145/3130800.3130813.

Xiaohan Lu, Yifeng Chen, Zhiying Li, Jinqiu Zhao, Yijie Zhou, Dongrui Wu, and Mingming Zhang. Electroencephalography enables continuous decoding of hand motion angles in polar coordinates. *Cyborg and Bionic Systems*, 7:0469, 2026.

Jingying Ma, Feng Wu, Qika Lin, Yucheng Xing, Chenyu Liu, Ziyu Jia, and Mengling Feng. CodeBrain: Bridging decoupled tokenizer and multi-scale architecture for EEG foundation model. In *International Conference on Learning Representations*, 2026.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. doi: 10.1007/978-3-030-58452-8_24.

Ji-Ha Park, Seo-Hyun Lee, and Seong-Whan Lee. Towards eeg-based talking-face generation for brain signal-driven dynamic communication. In *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–5. IEEE, 2024.

Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. GaussianAvatars: Photorealistic head avatars with rigged 3D Gaussians. In *CVPR*, pp. 20299–20309, 2024.

Robin Tibor Schirrmeister, Jost Tobias Springenberg, Lukas Dominique Josef Fiederer, Martin Glasstetter, Katharina Eggensperger, Michael Tangermann, Frank Hutter, Wolfram Burgard, and Tonio Ball. Deep learning with convolutional neural networks for eeg decoding and visualization. *Human Brain Mapping*, aug 2017. ISSN 1097-0193. doi: 10.1002/hbm.23730. URL http://dx.doi.org/10.1002/hbm.23730.

Shinya Shimizu, Airi Ota, and Ai Nakane. Reviving intentional facial expressions: An interface for als patients using brain decoding and image-generative ai. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2025.

Mohammad Soleymani, Jeroen Lichtenauer, Thierry Pun, and Maja Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Transactions on Affective Computing*, 3(1):42–55, 2012. doi: 10.1109/T-AFFC.2011.25.

Yonghao Song, Bingchuan Liu, Xiang Li, Nanlin Shi, Yijun Wang, and Xiaorong Gao. Decoding natural images from EEG for object recognition. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d0865cbe51d35ec322f9af9db7806fc7-Abstract-Conference.html.

Kartik Teotia, Helge Rhodin, Mohit Mendiratta, Hyeongwoo Kim, Marc Habermann, and Christian Theobalt. Audio-driven universal gaussian head avatars, 2025. URL https://arxiv.org/abs/2509.18924.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1239.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025a. URL https://openreview.net/forum?id=NPNUHgHF2w.

Zaitian Wang, Jian He, Yu Liang, Xiyuan Hu, Tianhao Peng, Kaixin Wang, Jiakai Wang, Chenlong Zhang, Weili Zhang, Shuang Niu, and Xiaoyang Xie. Milmer: a framework for multiple instance learning based multimodal emotion recognition, 2025b. URL https://arxiv.org/abs/2502.00547.

Haolin Xiong, Tianwen Fu, Pratusha Bhuvana Prasad, Yunxuan Cai, Haiwei Chen, Wenbin Teng, Hanyuan Xiao, and Yajie Zhao. Mind-to-Face: Neural-driven photorealistic avatar synthesis via EEG decoding. *arXiv preprint arXiv:2512.04313*, 2025. doi: 10.48550/arXiv.2512.04313. URL https://arxiv.org/abs/2512.04313.

Yuelang Xu, Benwang Chen, Zhe Li, Hongwen Zhang, Lizhen Wang, Zerong Zheng, and Yebin Liu. Gaussian head avatar: Ultra high-fidelity head avatar via dynamic Gaussians. In *CVPR*, 2024.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://papers.neurips.cc/paper_files/paper/2023/hash/f6b30f3e2dd9cb53bbf2024402d02295-Abstract-Conference.html.

Qiang Zhang, Tong Xiao, Haroun Habeeb, Larissa Laich, Sofien Bouaziz, Patrick Snape, Wenjing Zhang, Matthew Cioffi, Peizhao Zhang, Pavel Pidlypenskyi, et al. REFA: Real-time egocentric facial animations for virtual reality. In *CVPRW*, 2024.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018. doi: 10.1109/CVPR.2018.00068.

Yufeng Zheng, Victoria Fernández Abrevaya, Marcel C. Bühler, Xu Chen, Michael J. Black, and Otmar Hilliges. I M avatar: Implicit morphable head avatars from videos. In *CVPR*, 2022.

Supplementary Material

A Supplementary Dataset and Method Details

This supplement describes synchronization (Sec. A.1), dataset statistics and splits (Sec. A.2), avatar fitting and inversion (Sec. A.3), training objectives (Sec. A.4), and implementation details (Sec. A.5). The expression repertoire is listed in Sec. B.

A.1 System Synchronization

To synchronize the EEG signals with the facial video recordings, a synchronization event is manually triggered for both the face and EEG recording devices. Additionally, jaw clenching serves as a complementary synchronization signal, as it produces large-amplitude artifacts that are easily detectable both visually in the video and in the EEG traces. These events are used to compute a time-lag offset, enabling precise temporal alignment of the two modalities.

A.2 Dataset statistics and splits

Brain2Face contains recordings from six adult male subjects aged 24 to 47, each recorded in two synchronized EEG and video sessions. The 12 sessions total approximately 4 h 15 min of capture time, excluding the resting EEG baselines. Cameras recorded at 3840×2160 resolution; we downsample the images to 1920×1080 for avatar fitting and inversion. Fig. 2 illustrates expression variation within a subject, simultaneous views of the same expression, and appearance variation across subjects.

The protocol covers 75 predefined expression types per session. For the REVE experiments, we extract five-second paired segments after the annotated expression onsets. The processed dataset retains 891 such segments across the 12 sessions, totaling 74 min 15 s of expression data. Each segment pairs 64-channel EEG at 200 Hz with 256-dimensional avatar expression codes at 29.97 fps.

We train a separate EEG model for each subject and evaluate it on that same subject; these experiments include no held-out-subject split. Four fixed, disjoint expression groups contain 45 types in b_1 and 10 each in b_2, b_3, and b_4. Sec. B lists all expression names and their group membership. All S/U sequences are excluded from supervised EEG-model training. "Seen" means that the expression type occurs in training in the other session; "unseen" means that it is excluded from training in both sessions. Tab. 6 sums the retained segments across the six subject datasets using these categories.

In the cross-session split, b_2 in session 2 and b_3 in session 1 contain seen expression types (S), while b_4 in both sessions contains unseen types (U). Both sessions contribute training sequences, but each S expression type is trained only in the opposite session.

Tab. 6: **Cross-session split.** Cells list the expression groups assigned to each role in each session. Counts give retained five-second segments and their percentage of the 891 segments across six subjects. Each subject has a separate model.

		Test sequences	
	Training (T)	Seen expressions (S)	Unseen expressions (U)
Session 1	b_1, b_2	b_3	b_4
Session 2	b_1, b_3	b_2	b_4
Segments (%)	655 (73.5)	119 (13.4)	117 (13.1)

A.3 Gaussian avatar fitting and inversion

Fig. 4 summarizes how calibrated multi-view expression frames provide RGB images, foreground masks, and triangulated 3D guide points. A frozen gaze-conditioned image encoder from UHAP Teotia et al. (2025) produces the per-frame expression codes $\mathbf{Z}^t_{\text{exp}}$. Fitting adapts the participant's identity

code and decoder weights, together with pose and color calibration, using the multi-view observations and 3D guides. Masked RGB and mouth losses compare rendered and observed colors, while a guide loss matches 200 selected 3D points. Structural similarity and VGG-based perceptual losses compare image appearance. Regularization constrains the identity code and Gaussian scales, offsets, and opacities; a Laplacian loss encourages guide-mesh smoothness, and a locality loss penalizes deviations from the pretrained decoder. After fitting and guide-mesh refinement, the identity code and personalized decoder are frozen, and the saved expression trajectory serves as the EEG regression target.

The decoder supplies a guide mesh with $V = 7306$ vertices in millimetres. The fixed selection operator Π extracts $K_\ell = 200$ landmark-list positions, including repeated indices (193 unique vertices), from both the reference and predicted meshes. Thus, the landmark loss below compares 3D coordinates, rather than image-space projections. The same frozen decoder and camera settings produce the predicted and fitted-avatar reference renders; the 52-dimensional blendshape vectors are used only for evaluation.

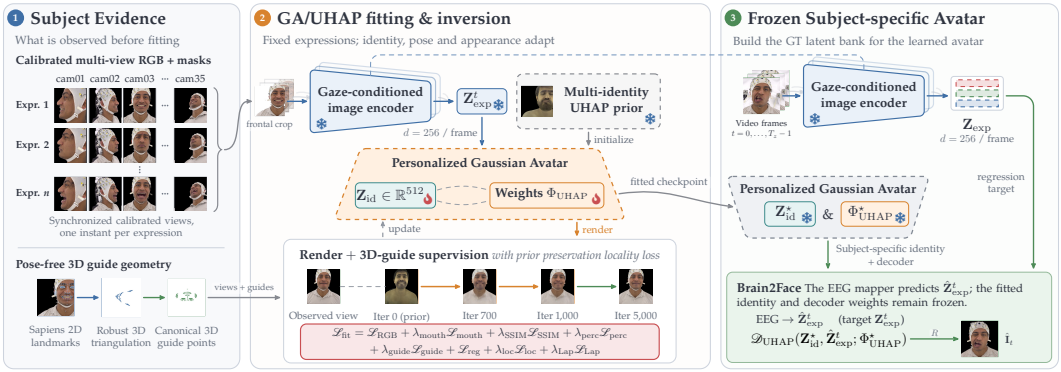

Fig. 4: **Per-subject Gaussian-avatar fitting and inversion.** Calibrated multi-view expression frames provide RGB images and foreground masks. Sapiens landmarks are triangulated and mapped to canonical coordinates to form pose-free 3D guide points. Teotia et al. introduced both the UHAP prior and the gaze-conditioned image encoder used here Teotia et al. (2025). For each frame, the frozen encoder produces an expression code $\mathbf{Z}^t_{\text{exp}} \in \mathbb{R}^d$, where $d = 256$. Fitting adapts the participant's identity code $\mathbf{Z}_{\text{id}} \in \mathbb{R}^{512}$ and the avatar decoder weights Φ_{UHAP}, together with pose and color calibration, using the multi-view images and 3D guides. The shared identity satisfies $\|\mathbf{Z}_{\text{id}}\|_2 \le 5$. The fitting loss includes mouth and regularization terms; Sec. A.3 describes these terms. After the final guide-mesh refinement, the fitted decoder weights $\Phi^\star_{\text{UHAP}}$ and identity $\mathbf{Z}^\star_{\text{id}}$ are frozen. The saved expression trajectory $\mathbf{Z}_{\text{exp}}$ provides the target for EEG-to-expression regression. Predicted expression codes then drive the frozen personalized avatar. The five thumbnails use one camera and frame from participant 2, session 1. They show the observation and initial render, followed by iterations 700, 1,000, and 5,000.

A.4 Training Objective

The combined objective as introduced in Eq. (1) in the main paper uses latent, mesh, landmark, and auxiliary trajectory losses with $(\lambda_p, \lambda_v, \lambda_\ell, \lambda_{\text{aux}}) = (1, 5, 10, 0.5)$. The MLP-head runs use the same main loss types and weights, with $\lambda_{\text{aux}} = 0$.

For a batch of B windows, $\mathbf{Z}^t_{\text{exp},b}$ denotes the length-d vector at frame t of window b. The expression-latent loss averages squared error over the batch:

$$\mathscr{L}_p(\hat{\mathbf{Z}}_{\text{exp}}, \mathbf{Z}_{\text{exp}}) = \frac{1}{BdN_z} \sum_{b=1}^{B} \sum_{t=0}^{N_z-1} \left\| \hat{\mathbf{Z}}^t_{\text{exp},b} - \mathbf{Z}^t_{\text{exp},b} \right\|_2^2 . \tag{2}$$

The mesh loss is the mean absolute coordinate error:

$$\mathscr{L}_v(\hat{\mathbf{V}}, \mathbf{V}) = \frac{1}{3V|\Omega_v|} \sum_{(b,t)\in\Omega_v} \sum_{v=1}^{V} \sum_{a=1}^{3} \left| \hat{V}_{b,t,v,a} - V_{b,t,v,a} \right| . \tag{3}$$

The landmark loss applies a coordinate-wise Wing penalty:

$$\mathscr{L}_\ell(\hat{\mathbf{L}}, \mathbf{L}) = \frac{1}{3K_\ell|\Omega_\ell|} \sum_{(b,t)\in\Omega_\ell} \sum_{k=1}^{K_\ell} \sum_{a=1}^{3} \rho_{\mathrm{W}}(\hat{L}_{b,t,k,a} - L_{b,t,k,a}), \tag{4}$$

where, for $u = |e|$,

$$\rho_{\mathrm{W}}(e) = \begin{cases} w_{\mathrm{W}} \log(1 + u/\epsilon_{\mathrm{W}}), & u < w_{\mathrm{W}}, \\ u - C_{\mathrm{W}}, & u \geq w_{\mathrm{W}}, \end{cases} \quad C_{\mathrm{W}} = w_{\mathrm{W}} - w_{\mathrm{W}} \log(1 + w_{\mathrm{W}}/\epsilon_{\mathrm{W}}).$$

We use the Wing form Feng et al. (2018) with $w_{\mathrm{W}} = 10$ mm and $\epsilon_{\mathrm{W}} = 2$ mm. Both $\mathscr{L}_v$ and $\mathscr{L}_\ell$ have units of mm, but they measure different coordinate penalties. Repeated landmark indices retain their repeated weight.

Each Ω contains the frames evaluated for that loss. The losses average over coordinates and these frames. The auxiliary head reconstructs the complete target trajectory:

$$\mathscr{L}_{\mathrm{aux}}(\hat{\mathbf{Z}}^{\mathrm{aux}}_{\mathrm{exp}}, \mathbf{Z}_{\mathrm{exp}}) = \frac{1}{BdN_z} \sum_{b=1}^{B} \sum_{t=0}^{N_z-1} \sum_{a=1}^{d} |\hat{Z}^{\mathrm{aux},t}_{\mathrm{exp},b,a} - Z^{t}_{\mathrm{exp},b,a}|.$$

Only the trainable EEG encoder and mapper parameters are updated; a frozen encoder remains fixed. Geometry gradients pass through the frozen avatar decoder to the predicted latents; the auxiliary head is part of the mapper.

A.5 Implementation Details

The REVE mapper uses kernel-size-3 residual convolutions with increasing dilation along the flattened channel-patch token axis, which is not a single physical time axis. The multi-scale stem is disabled in the reported residual-mapper runs. The learned resampler uses convolution and transposed-convolution layers; a framewise projection produces 256 coordinates, with linear interpolation to exactly N_z frames when needed. The auxiliary head pools features from the residual blocks and reconstructs a second full trajectory during training. The MLP mapper used for encoder probing flattens the encoder features and applies two 512-unit hidden layers with GEGLU, RMSNorm, and dropout probability 0.75 before projecting to $256N_z$ values. This mapper has no auxiliary reconstruction branch.

EEG models use AdamW under three initialization regimes: pretrained-frozen, pretrained-unfrozen, and scratch. The first keeps the pretrained encoder fixed, the second fine-tunes it, and the third trains it from random initialization. The UHAP identity and decoder remain frozen in all three.

Sec. A.2 defines the expression groups and the cross-session split. Training samples fresh crops from the assigned training sequences and evaluates on fixed test windows; the normalization scope is specified in Sec. 4.1.

The qualitative clips concatenate predictions from non-overlapping EEG blocks. An optional sliding-window mode aligns predictions in time and averages their overlaps. The reported qualitative results use block concatenation; quantitative metrics are evaluated on individual windows.

B List of Expressions

Tab. 7 lists all 75 expression names where expressions from the group b_1 are training expression in session 1 and 2, b_2 are training expressions for session 1, b_3 are training expressions for session 2, and b_4 are only used for evaluation from both sessions.

Tab. 7: **Full expression repertoire.** The 75 expression labels and their fixed group assignments. Indices follow this group order and do not indicate recording order.

Index	Group	Expression name
1	b_1	lips_puffed
2	b_1	raise_left_eyebrow
3	b_1	relaxed_mouth_open
4	b_1	left_side_sneer
5	b_1	show_all_teeth
6	b_1	jaw_open_pull_lips_in
7	b_1	kiss_lips_look_down
8	b_1	stick_lower_lip_out
9	b_1	mouth_open_tongue_out
10	b_1	smile_mouth_closed
11	b_1	jaw_open_sharp_corner_lip_stretch
12	b_1	suck_cheeks_in_brows_up
13	b_1	neck_stretch_brows_up
14	b_1	neutral_eyes_open
15	b_1	nostrils_sucked_in
16	b_1	eyes_wide_open
17	b_1	bite_lower_lip
18	b_1	lips_together_pushed_forward
19	b_1	tongue_out_flat
20	b_1	open_lips_mouth_stretch_nose_wrinkled
21	b_1	tongue_out_left_teeth_showing
22	b_1	tongue_over_upper_lip
23	b_1	jaw_back
24	b_1	suck_cheeks_in
25	b_1	jaw_open_lips_together
26	b_1	blow_cheeks_full_of_air
27	b_1	tongue_out_lips_closed
28	b_1	lower_eyebrows
29	b_1	hide_lips_look_up
30	b_1	jaw_open_upper_lip_raised
31	b_1	raise_left_eyebrow_pucker_lips
32	b_1	tongue_out_rolled
33	b_1	frown_stick_lower_lip_out
34	b_1	wink_left_eye_smile
35	b_1	frown
36	b_1	cheeks_puffed
37	b_1	jaw_open_huge_smile
38	b_1	raise_right_eyebrow
39	b_1	shh
40	b_1	jaw_forward
41	b_1	pressed_lips_brows_down
42	b_1	raise_both_eyebrows_pucker_lips
43	b_1	open_mouth_stretch_nose_wrinkled
44	b_1	bite_tongue
45	b_1	tongue_out_right_teeth_showing
46	b_2	wink_right_eye_smile
47	b_2	right_side_sneer
48	b_2	right_cheek_puffed
49	b_2	left_cheek_puffed
50	b_2	mouth_nose_left
51	b_2	jaw_open_lips_pushed_out
52	b_2	swallow
53	b_2	scrunch_face_squeeze_eyes
54	b_2	smile_stretched
55	b_2	mouth_nose_right
56	b_3	nostrils_dilated

Continued on the next page

Tab. 7 continued

Index	Group	Expression name
57	b_3	open_mouth_wide_tongue_up_and_back
58	b_3	raise_inner_eyebrows
59	b_3	suck_cheeks_in_brows_frowned
60	b_3	oooo
61	b_3	jaw_clench
62	b_3	raise_upper_lip_scrunch_nose
63	b_3	raise_right_eyebrow_pucker_lips
64	b_3	scream_eyebrows_up
65	b_3	tongue_out_thick
66	b_4	neutral_eyes_closed
67	b_4	lips_open_right
68	b_4	jaw_drop_brows_up
69	b_4	bite_upper_lip
70	b_4	jaw_open_mouth_corners_down_nose_wrinkled
71	b_4	raise_cheeks
72	b_4	mouth_open_jaw_right_show_teeth
73	b_4	lips_open_left
74	b_4	smile_mouth_open
75	b_4	mouth_open_jaw_left_show_teeth