FROM PHYSIOLOGICAL COMPOSITION TO AFFECTIVE GEOMETRY IN EEG REPRESENTATIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Emotion is associated with joint activity across multiple brain regions. However, the regional and spectral composition of emotion-related activity remains difficult to characterize explicitly in high-dimensional EEG representations. In this work, we propose a physiological compositional framework that expresses emotion-related EEG activity through a shared vocabulary of region–frequency motifs. For each EEG instance, we estimate motif activations from EEG foundation-model representations, guided by each motif's predefined scalp region and spectral activity within its frequency band. Competition among motifs reduces redundant evidence assignment, and their activations together form an instance-specific physiological composition. The resulting representation provides a shared physiological space for comparing affective geometries across foundation models.

Across all nine dataset–model combinations, physiologically structured motifs outperform randomized mappings, supporting the presence and predictive relevance of physiological organization in the learned composition. And two independently pretrained foundation models also recover highly similar affective geometry, with similarities of 0.976 in emotion distances and 0.935 in centroid angles. Together, these results establish physiological composition as a common, interpretable space for comparing EEG foundation models not only by decoding performance, but also by how they organize affective structure.

1 INTRODUCTION

Emotion is associated with distributed patterns of neural activity (Koide-Majima et al., 2020). In EEG, emotion-related information spans multiple recording sites and frequency bands (Zheng & Lu, 2015; Song et al., 2020). However, the regional and spectral composition of this information remains difficult to characterize explicitly in high-dimensional learned representations. Understanding these representations involves two complementary questions: which physiological components jointly support emotion decoding, and what affective geometry—the relative organization of emotion categories—emerges from their combinations?

EEG foundation models such as LaBraM, CBraMod, and REVE provide reusable representations for downstream decoding (Jiang et al., 2024; Wang et al., 2025; El Ouahidi et al., 2025). Decoding accuracy establishes their predictive utility, but does not describe the physiological components they use or how they organize relations among emotions. Representational similarity analysis compares relational structure across representational spaces (Kriegeskorte et al., 2008), while concept-based methods introduce interpretable intermediate representations, including in settings without concept annotations (Koh et al., 2020; Oikarinen et al., 2023). We bring these perspectives together through a shared physiological vocabulary, enabling predictive composition and affective geometry to be examined within the same representation across EEG foundation models.

We propose a physiological compositional framework built on a fixed vocabulary of 40 region–frequency motifs, defined by eight scalp regions and five canonical frequency bands. For each EEG instance, a trainable motif pathway estimates activations from frozen foundation-model features, guided by each motif's predefined scalp region and spectral evidence from its frequency band. Competition among motifs discourages redundant evidence assignment. Their activations form an instance-specific physiological composition: motif identities are predefined, while their activations are learned without motif-level annotations. The motif pathway contributes additively to the logits

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.

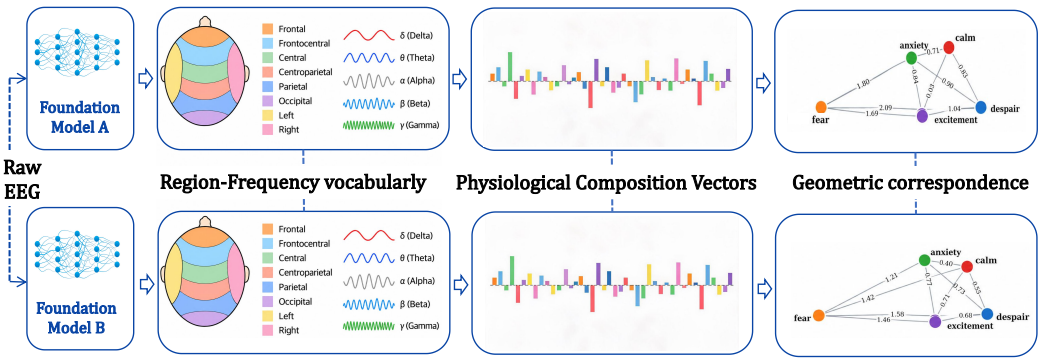

Figure 1: From physiological composition to affective geometry. A shared region–frequency vocabulary expresses EEG foundation-model representations as physiological compositions. Relations among emotion centroids in this space enable cross-model comparison of affective geometry.

of a motif-blind base predictor, allowing us to measure its incremental predictive value and the exact effects of removing individual motif contributions at the logit level.

The same composition supports analysis at the emotion-category level. After subject centering and feature standardization, we compute emotion centroids in the full motif space and characterize their relations through pairwise distances and angles. This defines affective geometry in the learned downstream physiological representation. The shared vocabulary allows us to examine whether model-specific compositions support similar relations among emotions under a common physiological reference and analysis protocol.

We evaluate the framework on SEED, SEED-IV, and a five-class in-house dataset using three frozen EEG foundation models. Adding the motif pathway improves accuracy over the base predictor in all nine dataset–model combinations. Physiologically structured motifs also outperform capacity-matched randomized mappings in all nine settings, supporting the predictive relevance of the intended region–frequency organization. Co-activation analyses reveal dependencies beyond marginal activation rates and subject-specific tendencies in all nine settings, and beyond per-trial activation counts in eight of nine. Together, these results support the predictive value and non-random organization of the learned physiological compositions.

Permutation analyses further reveal emotion-dependent structure in the motif space across all nine settings. On the five-class in-house dataset, LaBraM and REVE exhibit strong cross-model geometric correspondence, with Pearson correlations of 0.976 for pairwise emotion distances and 0.935 for centroid angles. CBraMod does not show comparable agreement, and raw band-power features show weaker correspondence with the LaBraM and REVE geometries. These findings identify shared affective organization between two independently pretrained models while showing that a common motif vocabulary does not guarantee geometric agreement. Physiological composition therefore provides a common basis for examining both how EEG foundation models distinguish emotion categories and how their learned representations organize relations among them.

Our contributions are as follows:

- A physiological compositional framework that expresses emotion-related EEG evidence through a shared region–frequency vocabulary, with additive predictive contributions and explicit physiological references across heterogeneous foundation models.
- An empirical evaluation of predictive utility and compositional structure across three datasets and three frozen backbones, combining comparisons against motif-blind predictors, capacity-matched randomized mappings, and co-activation null models.
- An analysis of affective geometry in the shared motif space, revealing strong correspondence between two independently pretrained models and characterizing its scope through comparisons with another backbone and raw spectral features.

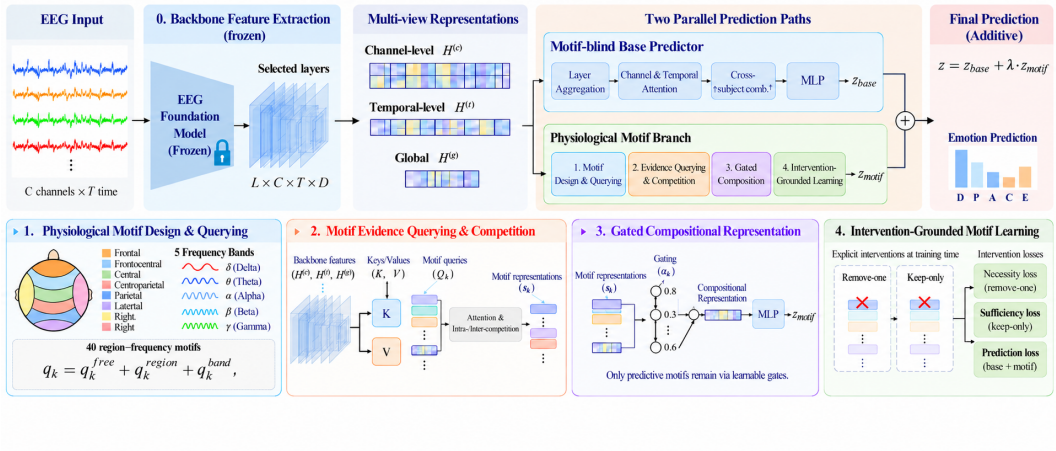

Figure 2: Overview of the proposed framework. Frozen EEG features feed a motif-blind base predictor and a physiological motif branch with structured querying, competition, and gating. The branches contribute additively to prediction, while intervention objectives guide motif learning.

2 METHOD

Figure 2 summarizes the framework. Given an EEG trial x, we extract windowed representations from L selected layers of a frozen EEG foundation model. These features are organized into channel, temporal, and global views:

$$H^{\mathrm{ch}} \in \mathbb{R}^{L\times C\times D}, \qquad H^{\mathrm{tm}} \in \mathbb{R}^{L\times T\times D}, \qquad H^{\mathrm{gl}} \in \mathbb{R}^{L\times D}, \tag{1}$$

where C, T, and D denote the numbers of channels, temporal units, and feature dimensions. The three views preserve complementary aspects of the frozen representation: H^{ch} retains channel-specific information, H^{tm} retains temporal structure, and H^{gl} provides a global summary of the trial. A motif-blind base branch predicts emotion from these features, while a physiological motif branch combines channel-level features with spectral evidence computed directly from the EEG. Only the downstream branches are trained.

Base predictor. Each feature view is aggregated across layers using separate learnable weights, allowing channel, temporal, and global information to draw differently from the selected representation depths. Channel and temporal attention then produce descriptors h_{ch} and h_{tm} by emphasizing informative channels and temporal units, respectively, while the layer-aggregated global features form h_{gl}. The base logits are

$$\mathcal{S}_{\mathrm{base}}(x) = f_{\mathrm{base}}\left([h_{\mathrm{ch}}; h_{\mathrm{tm}}; h_{\mathrm{gl}}; e_s]\right) + b_s, \tag{2}$$

where f_{base} is a multilayer prediction head, and e_s and b_s are a learnable subject embedding and subject-specific class bias. These subject-specific terms account for inter-subject variability in the frozen EEG representations. The base branch therefore provides a motif-blind reference prediction, to which the physiological motif pathway contributes a structured residual.

Physiologically structured motif querying. We define $K = 40$ motifs by crossing eight scalp regions (frontal, frontocentral, central, centroparietal, parietal, occipital, left, and right) with five frequency bands (delta, theta, alpha, beta, and gamma). This fixed vocabulary provides a shared physiological reference in which different foundation-model representations can be expressed using the same region–frequency coordinates. Motif identity is fixed by its region–frequency coordinate, but motif activation is neither predefined nor directly supervised. Instead, the downstream pathway learns how strongly each motif is expressed in a trial from the frozen backbone representation and the corresponding physiological evidence. This separation keeps the semantic coordinates fixed across foundation models while allowing their trial-wise expression to be learned from data.

For motif $k = (r, b)$, a query combines motif-specific parameters with region and band embeddings shared by motifs with the corresponding identities. Given projected key and value vectors $\mathbf{k}_{l,c} =$

$W_K H^{\mathrm{ch}}_{l,c}$ and $\mathbf{v}_{l,c} = W_V H^{\mathrm{ch}}_{l,c}$, we compute

$$q_k = q_k^{\mathrm{free}} + q_r^{\mathrm{region}} + q_b^{\mathrm{band}},$$
$$A_{k,l,c} = \frac{q_k^\top \mathbf{k}_{l,c}}{\sqrt{d}} + \rho M_{k,c} + \eta P_{b,c} + \beta_{k,l}. \tag{3}$$

Here, d is the query dimension, $M_{k,c}$ indicates membership of channel c in motif k's scalp region, $P_{b,c}$ denotes normalized EEG band-power evidence for band b at channel c, and $\beta_{k,l}$ is a learnable layer bias. Fixed coefficients ρ and η control the spatial and spectral guidance. The factorized query separates shared physiological structure from motif-specific flexibility. Motifs associated with the same region or frequency band share the corresponding embedding, while q_k^{free} allows each region–frequency pair to adapt individually to the representation provided by the frozen backbone. Thus, motif semantics are structured without requiring all motifs to use identical latent evidence. The query–key term measures compatibility with the latent foundation-model representation, while the spatial and spectral terms bias the query toward its predefined physiological identity. These terms act as soft rather than hard constraints, allowing attention to distributed latent evidence while preserving a common region–frequency reference.

Competitive evidence assignment. To discourage motifs from repeatedly selecting the same evidence, we combine within-motif attention $\alpha_{k,l,c} = \mathrm{softmax}_{l,c}(A_{k,l,c})$ with cross-motif ownership $o_{k,l,c} = \mathrm{softmax}_k(A_{k,l,c}/\tau)$:

$$\widetilde{\alpha}_{k,l,c} = \frac{\alpha_{k,l,c}\left[(1-\lambda_o) + \lambda_o o_{k,l,c}\right]}{\sum_{l',c'} \alpha_{k,l',c'}\left[(1-\lambda_o) + \lambda_o o_{k,l',c'}\right]}. \tag{4}$$

Temperature τ controls ownership sharpness, and λ_o controls its influence. Thus, within-motif attention identifies evidence relevant to each motif, whereas ownership introduces relative competition for the same layer–channel tokens. These two normalizations therefore answer complementary questions: which layer–channel tokens best support a given motif, and which motif most strongly claims a particular token relative to the alternatives. Their combination preserves motif-specific evidence selection while reducing repeated assignment of the same predictive feature. The resulting attention encourages motifs to specialize rather than repeatedly assigning the same latent evidence to multiple components. Each motif pools latent evidence as $z_k(x) = \sum_{l,c} \widetilde{\alpha}_{k,l,c} \mathbf{v}_{l,c}$. Together with band-power evidence $s_k(x)$ pooled over the attended channels, this produces a scalar activation $a_k(x) = f_{\mathrm{act}}(z_k(x), s_k(x))$. The resulting $a_k(x)$ is therefore a trial-specific scalar describing the evidence assigned to motif k.

Physiological composition and prediction. Each motif has an independent sigmoid gate $g_k = \sigma(\gamma_k)$ shared across trials. The gated activation $c_k(x) = g_k a_k(x)$ defines the composition vector $\mathbf{c}(x) = [c_1(x), \ldots, c_K(x)]$. Thus, gates capture global motif relevance, while activations vary across instances. Because the gates are independent rather than normalized across motifs, multiple motifs can remain simultaneously active and contribute jointly to the prediction. With learned class-weight vectors w_k, the motif and final logits are

$$\mathcal{S}_{\mathrm{motif}}(x) = \sum_{k=1}^{K} c_k(x) w_k + b_{\mathrm{motif}},$$
$$\mathcal{S}(x) = \mathcal{S}_{\mathrm{base}}(x) + \lambda \mathcal{S}_{\mathrm{motif}}(x), \tag{5}$$

where λ scales the motif contribution, and w_k determines how motif k contributes to the different emotion logits. The vector $\mathbf{c}(x)$ provides the shared coordinates used in the compositional and geometric analyses. Each coordinate of $\mathbf{c}(x)$ has a fixed physiological identity but a trial-dependent magnitude. Unlike the original backbone features, whose dimensions need not be semantically aligned across pretrained models, the coordinates of $\mathbf{c}(x)$ refer to the same region–frequency vocabulary by construction. This shared indexing enables subsequent comparison of motif composition and emotion-level geometry across heterogeneous backbones.

Prediction-level interventions and training. Because motif contributions enter the prediction additively, their predictive roles can be evaluated directly at the logit level. Rather than identifying important motifs solely from attention weights or activation magnitudes, we measure how the

prediction changes when a motif contribution is removed or retained in isolation. Additivity allows these interventions to be computed without masking the EEG input or rerunning the frozen foundation model:

$$\mathcal{S}^{(-k)}(x) = \mathcal{S}(x) - \lambda c_k(x) w_k,$$
$$\mathcal{S}^{(k)}(x) = \lambda \left(c_k(x) w_k + \frac{b_{\text{motif}}}{K} \right). \tag{6}$$

The two interventions capture complementary properties. The remove-one operation asks whether motif k supports the prediction in the context of the full model, whereas the keep-only operation asks whether its contribution retains class-relevant evidence on its own. For J emotion classes, define the true-class margin as $m(\mathcal{S}, y) = \mathcal{S}_y - \frac{1}{J-1} \sum_{j \neq y} \mathcal{S}_j$. The removal effect is $E_k = m(\mathcal{S}, y) - m(\mathcal{S}^{(-k)}, y)$; a positive value indicates support for the correct-class margin.

Training combines classification and intervention objectives:

$$\mathcal{L} = \lambda_b \mathcal{L}_{\text{base}} + \lambda_m \mathcal{L}_{\text{motif}} + \lambda_s \mathcal{L}_{\text{keep}} + \lambda_n \mathcal{L}_{\text{nec}}. \tag{7}$$

Here, $\mathcal{L}_{\text{base}}$ and $\mathcal{L}_{\text{motif}}$ supervise the base and motif-only predictions, respectively. The keep-only objective $\mathcal{L}_{\text{keep}}$ applies classification supervision to the single-motif predictions $\mathcal{S}^{(k)}$ in Eq. 6, encouraging each motif to retain class-relevant evidence when considered in isolation. The necessity objective $\mathcal{L}_{\text{nec}}$ encourages positive removal effects, such that removing a useful motif reduces the correct-class margin. The intervention objectives are introduced gradually during training.

3 Experiments

3.1 Experimental Setup and Emotion Decoding

Datasets and foundation models. We evaluate the proposed physiological motif framework on three EEG emotion datasets: SEED, SEED-IV, and an in-house dataset. These datasets span different numbers of emotion categories and experimental settings. The in-house dataset contains 32-channel EEG recordings from 91 participants collected during 35 video-based emotion-elicitation conditions, yielding 3,185 usable trials across five emotion categories: despair, fear, anxiety, calm, and excitement. A dedicated description of this dataset is being prepared for publication, and the dataset is planned to be made publicly available together with the corresponding work. We use three EEG foundation models—CBraMod, LaBraM, and REVE—as frozen feature extractors.

Evaluation protocol. All backbone parameters remain frozen, and only the downstream base and motif pathways are optimized. Data are split at the trial or stimulus level so that held-out samples are never used for model selection. For each outer split, model selection uses only the training partition, after which the model is retrained on the full outer-training set and evaluated once on the held-out data. For the in-house dataset, stimulus identities are explicitly separated across folds to prevent video-level leakage. Out-of-fold predictions, motif activations, and intervention effects are retained for subsequent analyses. The goal is not state-of-the-art emotion classification, but to test whether the physiological motif space preserves useful predictive information. As shown in Table 1, the Full model outperforms the motif-blind Base predictor in all dataset–backbone combinations, while the motif-only pathway retains substantial emotion-discriminative information.

Because the motif representation is learned within the predictive pathway rather than applied post hoc, its physiological structure can be evaluated directly in the subsequent compositional analyses.

3.2 Evidence for Physiological Composition in EEG Representations

Improved decoding alone does not establish that the learned motif representation is genuinely compositional or physiologically grounded. We therefore test whether recurrent motifs form non-random joint configurations and whether their predictive value depends on the intended physiological region–frequency correspondence.

Motifs form non-random combinations. We test whether recurrent motifs exhibit statistical dependence beyond marginal activation rates, subject-specific preferences, and variation in the number

Table 1: Emotion decoding accuracy across datasets and EEG foundation models. Base denotes the motif-blind predictor, Motif denotes the physiologically structured motif-only pathway, Rand. Motif denotes the capacity-matched randomized-mapping control that breaks the intended region–frequency correspondence, and Full combines Base and Motif.

Dataset	Backbone	Base ACC	Motif ACC	Rand. Motif ACC	Full ACC
SEED	CBraMod	.599	.569	.544	.647
SEED	LaBraM	.634	.637	.633	.687
SEED	REVE	.668	.735	.699	.736
SEED-IV	CBraMod	.379	.370	.360	.403
SEED-IV	LaBraM	.374	.381	.356	.402
SEED-IV	REVE	.394	.435	.425	.449
In-house dataset	CBraMod	.273	.235	.223	.298
In-house dataset	LaBraM	.262	.267	.262	.277
In-house dataset	REVE	.278	.279	.269	.307

Table 2: Non-random compositionality of recurrent physiological motifs. Multi. denotes the percentage of held-out trials with at least two active recurrent motifs, and TC denotes total correlation in bits. The independent and subject-conditioned null tests are significant in all nine settings ($p = 10^{-4}$); the exact-count null is significant in eight of nine settings ($p = 10^{-4}$), with LaBraM on SEED-IV as the only exception.

Dataset	Backbone	# Motifs	Multi. (%)	TC
SEED	CBraMod	5	59.3	1.736
SEED	LaBraM	5	68.7	1.042
SEED	REVE	4	58.5	.460
SEED-IV	CBraMod	4	68.1	.228
SEED-IV	LaBraM	2	30.0	.029
SEED-IV	REVE	4	63.0	.171
In-house dataset	CBraMod	3	50.3	.182
In-house dataset	LaBraM	4	57.7	.532
In-house dataset	REVE	3	50.7	.046

of active motifs per trial. For each dataset–backbone pair, recurrent motif activations are binarized using motif-specific median thresholds, and dependence is quantified by total correlation (TC). We compare the observed TC against three null models: an independent null preserving marginal activation rates, a subject-conditioned null additionally preserving subject-specific tendencies, and an exact-count null preserving the number of active motifs per trial while randomizing motif identities. Table 2 shows that the resulting joint configurations cannot be explained solely by activation frequency, subject identity, or the number of active components.

Does the composition depend on physiological structure? Non-random dependence alone does not establish physiological meaning. We therefore construct a capacity-matched randomized-mapping control that preserves the motif architecture and training procedure while disrupting the intended region–frequency correspondence. Specifically, each nominal motif receives spatially mismatched channels and a randomized frequency-band assignment, preserving model capacity while removing its physiological identity.

As shown by the **Motif ACC** and **Rand. Motif ACC** columns in Table 1, the physiologically structured motif pathway outperforms the randomized mapping in all nine dataset–backbone combinations. Thus, its predictive utility depends on the intended spatial–spectral organization rather than merely on introducing an additional 40-dimensional pathway.

Together with the null-model analysis, these results support a motif space that is both non-randomly compositional and dependent on physiological organization.

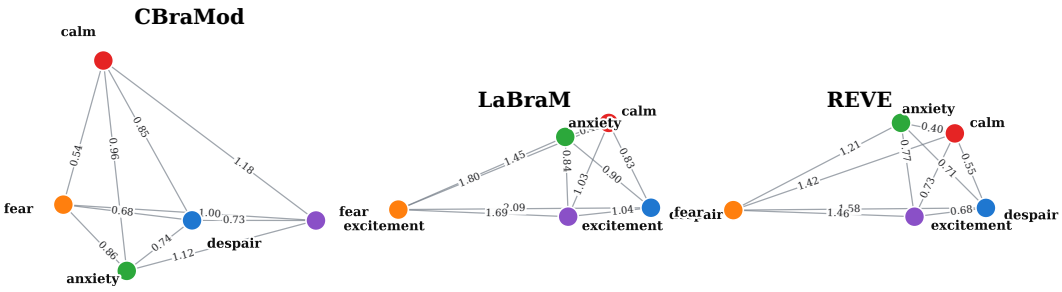

Figure 3: Affective geometry of the in-house dataset recovered from CBraMod (left), LaBraM (middle), and REVE (right). Nodes denote emotion centroids in the subject-centered 40-dimensional physiological motif space, and edge labels indicate pairwise distances. The two-dimensional layouts are optimized to preserve pairwise distances and angular relations of the original geometries.

Table 3: Agreement of affective geometry. Foundation-model comparisons are computed in the shared 40-dimensional physiological motif space; raw band-power provides a control geometry constructed directly from region–frequency spectral features.

Model pair	Dist. Pearson	Dist. Spearman	Angle Pearson	Angle Spearman
Raw band-power–LaBraM	.544	.442	-.108	.139
Raw band-power–REVE	.569	.442	-.071	.236
CBraMod–LaBraM	-.476	-.479	-.196	-.309
CBraMod–REVE	-.386	-.176	-.205	-.442
LaBraM–REVE	**.976**	**.879**	**.935**	**.891**

3.3 From Physiological Composition to Affective Geometry

The preceding analyses establish structured physiological composition at the level of individual EEG instances. We next characterize each emotion by its centroid in the physiological motif space and define affective geometry through the relative distances and directions among these centroids.

Subject-centered affective geometry. For each foundation model, we remove each subject's mean motif representation, standardize the residual motif dimensions, and compute an emotion centroid in the resulting space. Pairwise Euclidean distances and angular relations among these centroids define the affective geometry. Significance is assessed by within-subject label permutations, which preserve subject structure while disrupting emotion labels. Cross-foundation analyses use the complete shared 40-dimensional motif space to ensure a common coordinate system for all comparisons.

For visualization, the 40-dimensional geometry is projected into a two-dimensional layout optimized to preserve pairwise distances and angular relations; all statistical analyses remain in the original motif space. **All nine dataset–backbone settings show significant emotion-dependent organization** under within-subject permutation testing ($p = 2 \times 10^{-4}$ in every case). Because SEED and SEED-IV contain only three and four emotion classes, respectively, we use the five-class in-house dataset for the main cross-foundation analysis. Figure 3 shows the geometries recovered from CBraMod, LaBraM, and REVE.

Cross-foundation correspondence. We next ask whether different pretrained foundation models recover similar higher-order affective organization in the shared physiological coordinate system. LaBraM and REVE recover closely related configurations on the in-house dataset (Figure 3). Table 3 confirms this quantitatively: their pairwise emotion distances show Pearson and Spearman correlations of .976 and .879, while centroid angles correlate at .935 and .891, respectively. Exact emotion-identity permutation tests confirm that all four correspondences remain significant after BH-FDR correction.

The controls in Table 3 indicate that this correspondence is not a trivial consequence of the analysis framework. CBraMod uses the same 40-dimensional motif space and downstream architecture, yet

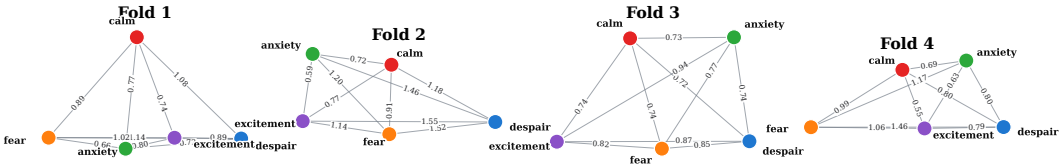

Figure 4: Fold-wise affective geometry of CBraMod on the in-house dataset. The four panels show emotion-centroid configurations obtained independently from each outer fold, illustrating variability of the recovered geometry across data splits.

Table 4: Fold-wise predictive and intervention statistics for CBraMod.

Fold	Necessity mean	Positive fraction	Base AUC	Final AUC
1	0.000446	.4885	.536	.559
2	0.001324	.4950	.613	.624
3	0.001119	.5046	.583	.586
4	**0.001519**	**.5258**	**.638**	**.653**

does not show comparable agreement with either LaBraM or REVE. Likewise, the raw band-power geometry shows only moderate correspondence in pairwise distances and little agreement in centroid angles, with none reaching significance under exact permutation testing. Thus, the LaBraM–REVE correspondence cannot be explained simply by the shared motif bottleneck or coarse spectral structure in the raw EEG.

Geometric stability. The weak global correspondence of CBraMod raises the question of whether it represents a reproducibly different affective organization or less stable recovery across data splits. We therefore examine its four outer folds independently (Figure 4 and Table 4). The recovered configurations vary substantially across folds, with no consistent alternative geometry emerging. Predictive performance also varies, with Fold 4 showing the strongest downstream AUC and the most visually differentiated configuration.

These observations suggest that CBraMod's disagreement with LaBraM and REVE is associated with less stable recovery of the downstream physiological representation rather than a stable but fundamentally different affective geometry. Because only four folds are available, however, we do not infer a direct relationship between decoding performance and geometric similarity.

This result also strengthens the control provided by CBraMod: the shared motif bottleneck does not automatically impose a common affective organization. We use *affective geometry* to denote the relative organization of emotion categories within the EEG-derived physiological representation, without implying a universal psychological geometry or identical geometry in the original pretrained latent spaces.

3.4 Ablation Study

We examine two architectural choices of the motif pathway on the five-class in-house dataset using CBraMod, LaBraM, and REVE as frozen backbones: structured query factorization and ownership competition. The physiological region–frequency correspondence itself is evaluated separately through the randomized-mapping control in the previous section. For each ablation, all remaining training and evaluation procedures are unchanged.

The ablations reveal a distinction between predictive performance and decomposition quality. Removing the region–band factorization does not consistently reduce ACC, but substantially increases attention redundancy across all three backbones, indicating that structured query factorization mainly organizes latent evidence into more differentiated physiological components. Ownership competition shows a similar effect: removing it causes only small, backbone-dependent changes in ACC but consistently increases attention redundancy, suggesting that its main role is to prevent multiple motifs from repeatedly assigning themselves to the same latent evidence.

Table 5: Architecture ablation on the five-class in-house dataset across three frozen EEG foundation models. Attention redundancy is the mean pairwise cosine similarity between motif attention maps, with lower values indicating less redundant evidence assignment.

Backbone	Variant	ACC	Attention Redund.
CBraMod	Full model	.298	**.129**
CBraMod	Free query only	.311	.326
CBraMod	w/o ownership competition	.292	.192
LaBraM	Full model	.277	**.072**
LaBraM	Free query only	.271	.223
LaBraM	w/o ownership competition	.278	.116
REVE	Full model	.307	**.053**
REVE	Free query only	.306	.164
REVE	w/o ownership competition	.294	.083

Overall, structured querying and ownership competition primarily improve the quality of the compositional decomposition while preserving comparable predictive performance.

4 Discussion

Our results suggest that the main value of the proposed framework lies not simply in improving emotion decoding, but in providing a common physiological coordinate system in which heterogeneous EEG foundation models can be compared. The fixed region–frequency vocabulary makes it possible to examine representation structure at two complementary levels: trial-level physiological composition and category-level affective geometry. Across datasets and backbones, the motif space exhibits significant emotion-dependent organization, while the randomized-mapping controls indicate that this structure depends on the intended physiological correspondence rather than on the presence of an arbitrary low-dimensional bottleneck. At the same time, the architectural ablations show that structured querying and ownership competition mainly improve the organization and non-redundancy of the decomposition rather than acting as generic accuracy-enhancing components. This distinction is important because the goal of the framework is not to replace conventional decoding models, but to make their internal affective organization directly comparable and interpretable.

The geometric analysis further shows that such a shared coordinate system can reveal relationships that are not visible from decoding accuracy alone. LaBraM and REVE recover highly similar affective organization, whereas CBraMod does not exhibit the same global correspondence and shows greater variability across data splits. The raw band-power control also indicates that the LaBraM–REVE correspondence cannot be explained simply by coarse spectral structure shared in the EEG input. We therefore view the recovered affective geometry as an empirically defined relational structure in the learned physiological representation. Whether these geometric relations correspond to meaningful psychological distances between emotions, established affective dimensions, or other properties of emotional experience remains an open question that requires more rigorous validation. Future work should therefore examine the recovered geometry against independent behavioral or self-report measures, richer emotion spaces, and additional datasets, and should investigate how the geometry changes as the physiological representation itself is further refined. More broadly, the contribution of this work is methodological as much as empirical: instead of asking only which EEG foundation model predicts emotion better, the proposed perspective asks how different models organize affective information within a common physiological reference space. Although the current evidence is limited by modest absolute decoding performance, the low number of emotion classes in public benchmarks, and the cross-trial evaluation setting, this perspective provides a new direction for EEG representation analysis in which compositional structure and relational geometry become explicit objects of comparison rather than remaining hidden behind downstream accuracy.

REFERENCES

M Catalina Camacho, Ashley N Nielsen, Dori Balser, Emily Furtado, David C Steinberger, Leah Fruchtman, Joseph P Culver, Chad M Sylvester, and Deanna M Barch. Large-scale encoding of emotion concepts becomes increasingly similar between individuals from childhood to adolescence. *Nature neuroscience*, 26(7):1256–1266, 2023.

Pin-Hao A Chen, Eshin Jolly, Jin Hyun Cheong, and Luke J Chang. Intersubject representational similarity analysis reveals individual variations in affective experience when watching erotic movies. *NeuroImage*, 216:116851, 2020.

Yuning Chen, Sha Zhao, Shijian Li, and Gang Pan. Emod: A unified eeg emotion representation framework leveraging va guided contrastive learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 17427–17435, 2026.

Guanglong Du, Jinshao Su, Linlin Zhang, Kang Su, Xueqian Wang, Shaohua Teng, and Peter Xiaoping Liu. A multi-dimensional graph convolution network for eeg emotion recognition. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.

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. *arXiv preprint arXiv:2510.21585*, 2025.

Mateo Espinosa Zarlenga, Katie Collins, Krishnamurthy Dvijotham, Adrian Weller, Zohreh Shams, and Mateja Jamnik. Learning to receive help: Intervention-aware concept embedding models. *Advances in Neural Information Processing Systems*, 36:37849–37875, 2023.

Eloy Geenjaar and Lie Lu. Citrus: Squeezing extra performance out of low-data bio-signal transfer learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 16781–16789, 2025.

Bruno L Giordano, Caroline Whiting, Nikolaus Kriegeskorte, Sonja A Kotz, Joachim Gross, and Pascal Belin. The representational dynamics of perceived voice emotions evolve from categories to dimensions. *Nature human behaviour*, 5(9):1203–1213, 2021.

Antonino Greco, Julia Moser, Hubert Preissl, and Markus Siegel. Predictive learning shapes the representational geometry of the human brain. *Nature communications*, 15(1):9670, 2024.

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

Ming Jin, Changde Du, Huiguang He, Ting Cai, and Jinpeng Li. Pgcn: Pyramidal graph convolutional network for eeg emotion recognition. *IEEE Transactions on Multimedia*, 26:9070–9082, 2024.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348, 2020.

Naoko Koide-Majima, Tomoya Nakai, and Shinji Nishimoto. Distinct dimensions of emotion in the human brain and their representation on the cortical surface. *NeuroImage*, 222:117258, 2020.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008. doi: 10.3389/neuro.06.004.2008.

Kent M Lee and Ajay B Satpute. More than labels: neural representations of emotion words are widely distributed across the brain. *Social Cognitive and Affective Neuroscience*, 19(1):nsae043, 2024.

Han Lu, Xiahai Zhuang, and Qiang Luo. A brain-inspired way of reducing the network complexity via concept-regularized coding for emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 556–564, 2024.

Yumeng Ma and Philip A Kragel. Map-like representations of emotion knowledge in hippocampal-prefrontal systems. *Nature Communications*, 17(1):1518, 2026.

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Glancenets: Interpretable, leak-proof concept-based models. *Advances in neural information processing systems*, 35:21212–21227, 2022.

Tuomas Oikarinen, Subhro Das, Lam M Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. *arXiv preprint arXiv:2304.06129*, 2023.

Pia-Kelsey O'Neill, Lorenzo Posani, Jozsef Meszaros, Phebe Warren, Carl E Schoonover, Andrew JP Fink, Stefano Fusi, and C Daniel Salzman. The representational geometry of emotional states in basolateral amygdala. *Nature Neuroscience*, pp. 1–13, 2026.

Haroon Popal, Yin Wang, and Ingrid R Olson. A guide to representational similarity analysis for social neuroscience. *Social cognitive and affective neuroscience*, 14(11):1243–1253, 2019.

Sungbin Shin, Yohan Jo, Sungsoo Ahn, and Namhoon Lee. A closer look at the intervention procedure of concept bottleneck models. In *International conference on machine learning*, pp. 31504–31520. PMLR, 2023.

Tengfei Song, Suyuan Liu, Wenming Zheng, Yuan Zong, and Zhen Cui. Instance-adaptive graph for eeg emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2701–2708, 2020.

Guagnyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. Eegpt: Pretrained transformer for universal and reliable representation of eeg signals. *Advances in Neural Information Processing Systems*, 37:39249–39280, 2024.

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 *The Thirteenth International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Chaoqi Yang, M Westover, and Jimeng Sun. Biot: Biosignal transformer for cross-data learning in the wild. *Advances in Neural Information Processing Systems*, 36:78240–78260, 2023.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models. *arXiv preprint arXiv:2205.15480*, 2022.

Mateo Espinosa Zarlenga, Pietro Barbiero, Gabriele Ciravegna, Giuseppe Marra, Francesco Giannini, Michelangelo Diligenti, Zohreh Shams, Frederic Precioso, Stefano Melacci, Adrian Weller, Pietro Lio, and Mateja Jamnik. Concept embedding models: Beyond the accuracy-explainability trade-off, 2022. URL https://arxiv.org/abs/2209.09056.

Daoze Zhang, Zhizhang Yuan, Yang Yang, Junru Chen, Jingjing Wang, and Yafeng Li. Brant: Foundation model for intracranial neural signal. *Advances in Neural Information Processing Systems*, 36:26304–26321, 2023.

Tong Zhang, Zhen Cui, Chunyan Xu, Wenming Zheng, and Jian Yang. Variational pathway reasoning for eeg emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2709–2716, 2020.

Zhi Zhang, Shenghua Zhong, and Yan Liu. Beyond mimicking under-represented emotions: Deep data augmentation with emotional subspace constraints for eeg-based emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 10252–10260, 2024.

Li-Ming Zhao, Xu Yan, and Bao-Liang Lu. Plug-and-play domain adaptation for cross-subject eeg-based emotion recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 863–870, 2021.

Wei-Long Zheng and Bao-Liang Lu. 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. doi: 10.1109/TAMD.2015.2431497.

A Related Work

Distributed and compositional representations of emotion. Emotion-related neural activity is widely understood as distributed rather than localized to a single region. In EEG-based affective computing, recent methods explicitly model interactions among electrodes, regions, and functional pathways. Instance-adaptive graph models learn dynamic inter-channel relationships for emotion recognition (Song et al., 2020), while pathway-based approaches identify salient dependencies among spatially and functionally related electrodes (Zhang et al., 2020). More recent graph architectures further integrate local, regional, and global EEG structure, or jointly model spatial and temporal dependencies across the scalp (Jin et al., 2024; Du et al., 2022). Cross-subject approaches further seek representations that preserve emotion-relevant structure while reducing subject-specific variation (Zhao et al., 2021), and recent work has explored emotion-aware latent subspaces for constraining EEG representations (Zhang et al., 2024). These studies support the view that affective information is carried by coordinated activity across multiple physiological components. However, they are primarily designed for discrimination, transfer, or connectivity modeling rather than for explicitly representing each EEG instance as a composition over a fixed physiological vocabulary.

Our work makes this composition explicit. Each EEG instance is expressed through a shared set of region–frequency motifs, allowing joint motif activation itself to become an analyzable object. This enables direct tests of whether physiological components form non-random combinations and whether their predictive value depends on their intended spatial–spectral organization.

Affective geometry and representational similarity. A complementary line of work studies emotion as a structured representational space rather than as a collection of independent class labels. Representational similarity analysis (RSA) characterizes such structure through pairwise relations among neural representations (Kriegeskorte et al., 2008; Popal et al., 2019). In affective neuroscience, multidimensional representational analyses have shown that affective experiences can be captured more precisely by relational structure than by scalar summaries (Chen et al., 2020), and that diverse emotions occupy structured multidimensional representations in the human brain (Koide-Majima et al., 2020). More recent RSA studies further show that emotion-related knowledge is distributed across broad cortical systems (Lee & Satpute, 2024). Related work has characterized the representational geometry of perceived emotions and shown that neural organization can reflect both categorical and dimensional affective structure (Giordano et al., 2021). Other studies have examined emotion representations as structured maps or population-level geometries in neural systems (O'Neill et al., 2026; Ma & Kragel, 2026), while broader neuroscience work has demonstrated that learning and prediction can systematically shape representational geometry (Greco et al., 2024). Developmental studies further suggest that the neural representational structure of emotion concepts can become increasingly shared across individuals (Camacho et al., 2023).

These studies motivate our geometric perspective. We define emotion centroids in a shared physiological motif space and characterize their relative organization through pairwise distances and angular relations. This extends representational analysis from within a single neural system to direct comparison across independently pretrained EEG foundation models.

Foundation models for neural signals. Large-scale pretraining has recently become an important strategy for learning reusable neural-signal representations. LaBraM learns generic EEG representations from heterogeneous large-scale data and transfers them to multiple downstream BCI tasks (Jiang et al., 2024). CBraMod further models structured spatial and temporal dependencies and demonstrates broad transfer across EEG benchmarks (Wang et al., 2025). REVE extends large-scale EEG pretraining across heterogeneous recording configurations and electrode settings (El Ouahidi et al., 2025). Related work has also explored large-scale neural-signal pretraining beyond scalp EEG, such as Brant for intracranial neural recordings (Zhang et al., 2023). More broadly, biosignal pretraining methods such as BIOT seek transferable representations across heterogeneous datasets and recording formats (Yang et al., 2023), while EEGPT develops a general pretrained represen-

tation for diverse EEG tasks (Wang et al., 2024). Recent work has also investigated transfer under limited labeled data through large-scale biosignal pretraining (Geenjaar & Lu, 2025), while emotion-specific pretraining frameworks seek transferable affective representations across multiple EEG datasets (Chen et al., 2026).

Most of these works assess pretrained representations primarily through downstream predictive performance. In contrast, we keep each foundation model frozen and use it as a representation provider. The downstream motif pathway then maps the pretrained features into the same physiological coordinate system, allowing compositional and geometric structure to be compared without rewriting the underlying pretrained representation.

Structured concept and motif extraction. Our motif pathway is also related to concept-based representation learning. Concept bottleneck models introduce predefined intermediate concepts that support interpretable prediction and direct intervention (Koh et al., 2020). Concept Embedding Models enrich this intermediate layer with higher-dimensional concept representations while retaining concept-level interpretability (Zarlenga et al., 2022), and GlanceNets explicitly addresses alignment between learned concept representations and intended semantic concepts (Marconato et al., 2022). Subsequent work has extended concept-based models to settings in which explicit concept annotations are unavailable or incomplete. Post-hoc concept bottleneck models project pretrained representations into interpretable concept spaces (Yuksekgonul et al., 2022), while label-free concept bottleneck models reduce dependence on manually annotated concept labels (Oikarinen et al., 2023). Intervention-aware concept models further study how concept representations can be learned to support reliable human intervention (Espinosa Zarlenga et al., 2023), and analyses of concept intervention procedures highlight that intervention behavior itself is an important property of concept-based representations (Shin et al., 2023). Related emotion-recognition models have also introduced separate concept-oriented pathways to regularize perceptual representations (Lu et al., 2024).

Our setting differs in that no per-trial supervision exists for whether an individual physiological motif is active. We therefore fix motif identity through predefined region–frequency coordinates and infer motif activation from the frozen foundation representation together with spatial and spectral evidence. This provides a structured intermediate representation without requiring motif-level labels, and connects physiological composition with higher-order affective geometry in a shared coordinate system.

AI USE STATEMENT

We used generative AI tools to assist with manuscript drafting and editing, literature search and summarization, and brainstorming and refinement of research ideas. We also used these tools to obtain feedback on methodological descriptions and to discuss the interpretation and presentation of results. Generative AI was not used to implement or conduct the experiments or to generate experimental results. The authors made the final scientific and editorial decisions and take full responsibility for the content of this work, including its claims, results, and references.