THE SATURATION REGULARITY: RETHINKING CONCEPT-ALIGNED SUPERVISION IN EEG LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept-aligned supervision aims to improve learning by making task-relevant knowledge explicit in a model's representations. However, a concept's relevance does not establish the value of further supervising it once the model already encodes it. Choosing auxiliary supervision therefore requires understanding when stronger alignment improves prediction and what information remains useful when it does not. In EEG emotion recognition, we identify the *Saturation Regularity*: further concept supervision can strengthen alignment without detectable predictive gains, and can even reduce accuracy. This pattern varies across FACED training configurations and CIFAR-10 training stages. Random-target controls further show that auxiliary-loss gains need not depend on semantic content. To examine what remains useful beyond class-mean alignment, we introduce *ResDiag*, a residual diagnostic framework combining representation decomposition, directional ablation, and ensemble-contribution analysis. ResDiag reveals that within-class concept-related residuals retain predictive information and that their encoding strength is associated with ensemble contribution. Motivated by these supervision and residual analyses, we develop *SatEnsemble*, a saturation-informed ensemble learning framework. It strengthens individual models with a refined training recipe, then averages checkpoint predictions to exploit complementary information. Experiments show that SatEnsemble outperforms leading EEG methods on FACED, setting a new state of the art and establishing an effective strategy for learning beyond concept-aligned supervision.

1 INTRODUCTION

Concept-aligned supervision seeks to improve learning by shaping representations around task-relevant knowledge (Koh et al., 2020). For EEG learning, it offers a way to connect neural signals to interpretable attributes of the task. Valence and arousal, for example, have been used to guide EEG emotion representations (Chen et al., 2025). Yet the value of such supervision depends on what the receiving model already represents. A concept can be meaningful and readily decodable while offering little additional value as a training target. This makes the choice of auxiliary supervision a practical design problem. An objective can change a representation as intended without improving the prediction it is meant to support. We therefore ask: *when does stronger concept alignment cease to improve prediction, and what information should learning exploit next?*

For EEG emotion recognition, we study valence—the positive–negative dimension of emotion—on the nine-class FACED benchmark (Chen et al., 2023). We construct language-derived directions, or *V-axis probes*, from representations of emotion descriptions. The story-based probe predicts sentiment on SST-2 (Socher et al., 2013) without task-specific training (Table 1). Both story-based and CLIP-text (Radford et al., 2021) variants correspond to cohort-averaged EEG responses (§5.1). A 14-model sweep recovers valence structure across several model families (Appendix S3). Rephrasing yields similar projections, supporting robustness to wording changes (Appendix S2). Paraphrase averaging improves sentiment transfer over a single description in the lead model (Table 3). Within trained EEG classifiers, valence is already encoded in class-mean features (§5.2). Figure 1a compares standardized emotion-class positions in CLIP text, cohort EEG ridge readouts, and one classifier

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.

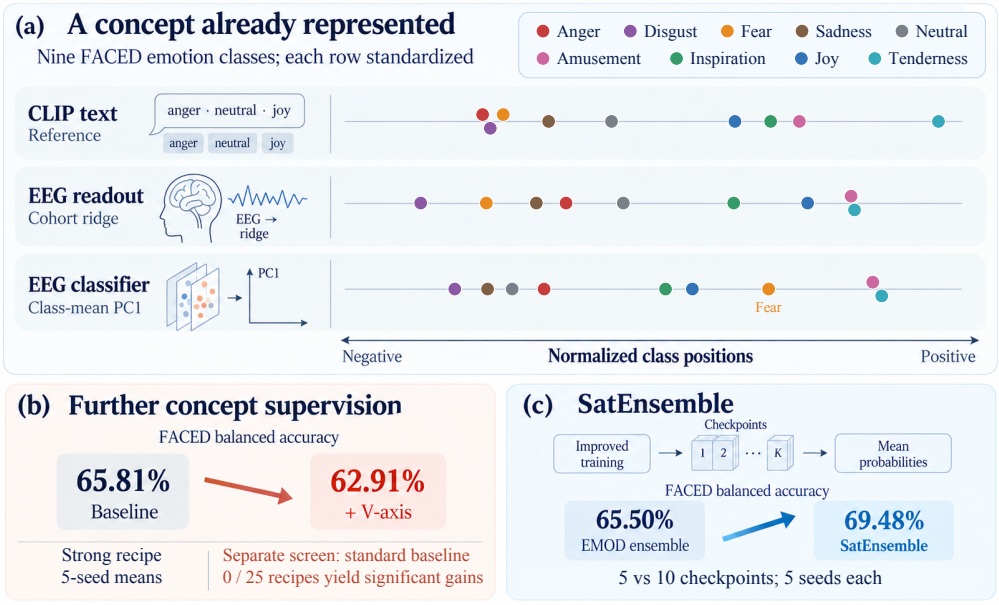


Figure 1: **Concept alignment and predictive benefit can diverge. (a)** Standardized emotion-class positions in CLIP text, cohort EEG ridge readouts, and class-mean features from one classifier checkpoint. **(b)** Adding V-axis supervision lowers FACED balanced accuracy under a strong recipe. **(c)** SatEnsemble combines improved training with checkpoint probability averaging.

checkpoint. These results establish a task-relevant, already represented concept for testing whether further supervision improves prediction.

We then test whether explicitly reinforcing an already represented concept improves prediction. We evaluate 25 auxiliary-supervision recipes on a standard EMOD (Chen et al., 2025) receiving baseline, spanning distillation, representation matching, contrastive objectives, and topographic constraints. None yields a statistically significant improvement. Ten significantly reduce accuracy after global Holm correction (Table 10). Follow-up interventions on a stronger recipe increase concept alignment (§3), yet reduce predictive accuracy (Figure 1b). These experiments establish the *Saturation Regularity* in EEG learning: additional concept supervision can strengthen alignment without detectable predictive gains, and can even reduce accuracy.

We examine how this pattern depends on training conditions and whether auxiliary gains require meaningful targets. Valence and arousal objectives with no detectable effect under a weaker FACED recipe reduce accuracy under the stronger recipe (Figure 2b). Experiments on CIFAR-10 (Krizhevsky, 2009) compare training stages within one architecture. Early auxiliary gains are no longer statistically detectable at convergence, and aligned targets never significantly outperform random targets at any tested stage (Figure 2a). An intervention on SEED-V (Liu et al., 2022b) provides a second EEG comparison, with only a small change under the tested recipe (Appendix S8). These experiments characterize the dependence on training conditions, while the random-target controls separate auxiliary-objective gains from evidence for the value of target semantics.

The saturation results leave open what predictive information remains beyond class-mean concept alignment. Across the analyzed checkpoints, class-mean valence encoding has similar strength, motivating a closer examination of within-class variation. We introduce *ResDiag*, a residual diagnostic framework for analyzing predictive information beyond class-mean alignment. ResDiag separates shared class structure from within-class variation. To test the predictive role of concept-related residuals, we use directional ablation (Arditi et al., 2024) while preserving the class-mean subspace. Accuracy decreases in eight of ten checkpoints (§5.2). To examine the link to model complementarity, we measure leave-one-out ensemble contribution as the performance change upon removing each checkpoint. Residual encoding correlates with this contribution ($r = 0.74$), with uncertainty assessed by seed-clustered bootstrap (Appendix S13). Together, these analyses identify useful information beyond class means and motivate combining complementary predictions.

Motivated by the supervision controls and residual analyses, we develop *SatEnsemble*, a saturation-informed ensemble learning framework. Augmentation, prototype regularization, and increased depth strengthen individual models. Uniform averaging of checkpoint softmax probabilities combines their complementary outputs (Figure 1c; §4.2).

SatEnsemble achieves state-of-the-art EEG emotion recognition on FACED. The ten-checkpoint ensemble reaches 69.48% balanced accuracy, outperforming both published EMOD (Chen et al., 2025) and our five-seed EMOD ensemble baseline. The improved training recipe also yields 67.55% with a validation-selected single model (Table 2). Recipe ablations support gains from both improved training and prediction averaging (§5.3). Replacing language-derived prototypes with random orthonormal targets preserves performance within the tested equivalence margin (Table 2). This control shows that prototype regularization can benefit prediction without relying on target semantics. These results establish a practical strategy for improving EEG prediction when further alignment of an already represented concept no longer helps.

To conclude, our contributions are:

- We identify the *Saturation Regularity* in EEG learning: stronger concept alignment does not necessarily translate into better prediction. Systematic experiments characterize its dependence on training conditions and distinguish the effects of auxiliary supervision from those of target semantics.
- We introduce *ResDiag*, a residual diagnostic framework for analyzing predictive information beyond class-mean alignment. Ablations show that within-class residuals support prediction, while ensemble analyses link encoding strength to ensemble contribution.
- We develop *SatEnsemble*, a saturation-informed ensemble learning framework combining improved training with prediction averaging. It achieves state-of-the-art performance on FACED, with ablations supporting the value of both components.

2 Related Work

Concept-direction supervision. The interventions we test are drawn from seven years of representation-alignment work: knowledge distillation (Hinton et al., 2015; Tian et al., 2020; Park et al., 2019), representational similarity analysis (Kriegeskorte et al., 2008; Sundaram et al., 2024), supervised contrastive losses (Khosla et al., 2020; van den Oord et al., 2018; Radford et al., 2021), and parameter-efficient adaptation (Hu et al., 2022; Liu et al., 2022a). The shared assumption across the distillation, similarity, and contrastive families is that the auxiliary signal provides information the task does not (parameter-efficient adaptation is included as a capacity-side comparison); this paper studies a regime where that added information does not help.

Concept directions in language models. Arditi et al. (Arditi et al., 2024) showed that refusal in instruction-tuned LMs is mediated by a single residual-stream direction, and related work extracts other behavioural directions (Li et al., 2023; Zou et al., 2023; Kim et al., 2018); concurrent Anthropic work (Sofroniew et al., 2026) finds emotion-concept directions from averaged story-evoked activations. We use PCA on class centroids to extract a valence direction from nine stories as a *probe*, not a steering target; our transfers are direction-existence claims, not feature-importance claims for biological emotion.

Brain–LM alignment. Toneva and Wehbe (Toneva & Wehbe, 2019) showed mid-layer transformer activations predict fMRI responses to narrative text; later work extended this across MEG, ECoG, and model families (Schrimpf et al., 2021; Goldstein et al., 2022; Caucheteux & King, 2022), formalised as the *Platonic Representation Hypothesis* (Huh et al., 2024). We touch this literature only through the probe; our load-bearing claim is the saturation regularity, not the alignment itself.

Frontal-alpha asymmetry. Davidson's frontal-alpha asymmetry (Davidson, 1992; Coan & Allen, 2004; Davidson, 2004) remains the dominant electrophysiological account of approach/withdrawal motivation, developed primarily on resting-state, trait, and static or briefly-presented stimuli. Our cohort signal on video-evoked emotion is posterior-dominant, with frontal asymmetries that on FACED

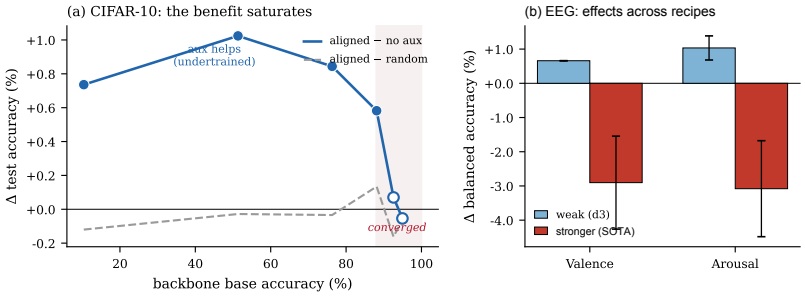


Figure 2: **Auxiliary-supervision effects depend on the receiving model. (a)** On CIFAR-10, early auxiliary gains (Holm $p<0.05$) are no longer significant at convergence. Aligned targets do not significantly outperform random targets at any tested stage. **(b)** On FACED, valence and arousal objectives have no detectable benefit under a weaker recipe and reduce accuracy under the stronger recipe (Holm $p\leq 0.011$ across four variants). Error bars: 95% CI (stronger), ±SEM (weaker).

survive in direction at smaller magnitude (on SEED-V the sign is not preserved, Appendix S8)—an additive scope statement for dynamic-video paradigms, not a refutation of FAA.

Ensemble theory. ResDiag uses the between-class and within-class decomposition studied in neural collapse (Papyan et al., 2020). Our checkpoints retain predictive within-class variation alongside similar class-mean concept encoding. This motivates examining complementary predictions in relation to bias–variance analysis (Geman et al., 1992; Krogh & Vedelsby, 1995) and deep ensembles (Fort et al., 2019). SatEnsemble averages output probabilities; ResDiag supplies separate evidence about residual predictive information and its association with ensemble contribution. These analyses support the strategy without identifying a complete causal account of its gains.

Baselines. The FACED-9 baselines we compare against are CBraMod (Wang et al., 2025) (0.551 published; 0.572 our reproduction) and EMOD (Chen et al., 2025) (0.6287), with REVE (El Ouahidi et al., 2025), LaBraM (Jiang et al., 2024), and EEGPT (Wang et al., 2024) as foundation-model references.

3 The Saturation Regularity

We test whether auxiliary supervision toward a known concept direction helps an EEG-emotion classifier, using an EMOD (Chen et al., 2025) backbone (the previous FACED-9 SOTA) at two regimes we call *weak* (network depth d=3, 0.619 BACC) and *stronger* (d=6, the SatEnsemble recipe of §4.2, 0.658 BACC 5-seed mean; best single 0.6755). The 25-recipe screen ran on the standard d=6 baseline; the strongest variants were also run on the weak d=3 backbone and the full SatEnsemble recipe to compare receiving configurations. These FACED comparisons vary depth and training choices; the CIFAR-10 control below compares training stages within one architecture. Each intervention adds an auxiliary loss that pulls the network's penultimate features toward a target direction. Targets span the nine LLM-derived class prototypes and their first principal component, the 1-D V-axis (§5.1); CLIP-text embeddings (Radford et al., 2021); and frontal-channel masks matched to the analytical valence ceiling (the maximum V-axis alignment attainable on FACED). Loss families span knowledge distillation (Hinton et al., 2015), representational similarity analysis (RSA) (Kriegeskorte et al., 2008), contrastive losses (Khosla et al., 2020), parameter-efficient fine-tuning (PEFT) adapters (Hu et al., 2022), curriculum schedules, multi-LLM ensembles, channel-targeted topographic losses, and anger-weighted variants (full table in Appendix S10).

Result. None yields a statistically significant gain. Ten produce statistically significant accuracy decrements under a global Holm–Bonferroni correction across all 25 hypotheses (16 at uncorrected $p < 0.05$, paired across five seeds). Four point estimates are positive, all nonsignificant. Five families show larger accuracy decrements at higher tested loss weights. Even the anger-weighted target that analytically maximises the V-axis ceiling on FACED is among the worst ($\Delta\text{BACC} = -0.054$): a higher analytical alignment ceiling does not guarantee a useful supervision target.

The effect depends on the receiving configuration. An intervention whose effect lies within seed noise on the weak d=3 backbone becomes a statistically significant negative on the stronger d=6 SatEnsemble recipe; the weak-vs-stronger difference is a significant backbone×loss interaction (Welch p=0.001). Depth and training choices vary together, so this comparison does not isolate convergence alone.

Additional concepts and training-stage controls. An *arousal* target (the validated CLIP–OASIS image arousal axis, *not* the failing text axis of Table 5) hurts the stronger backbone as well (topographic $\Delta\text{BACC} = -0.031$, p=0.011; emod-style -0.035, p=0.007; Holm-corrected) while remaining within seed noise on the weak d=3 baseline (Figure 2b). A complementary training-stage comparison examines the pattern outside EEG: on CIFAR-10 (Krizhevsky, 2009) an auxiliary head significantly *improves* an undertrained backbone (Holm $p \in [0.006, 0.016]$) and the benefit is no longer significant at convergence (p>0.37; Figure 2a). Aligned targets never significantly outperform random targets at any tested CIFAR-10 stage. Early auxiliary gains therefore do not establish that the target's semantic content supplies the benefit. Further controls test other forms of concept injection: using the same direction for frozen-feature alignment, LLM-aligned initialization, or cross-subject anchoring yields no detectable gain over the corresponding baselines (Appendix S12).

Saturation Regularity (empirical). *In the tested EEG learning settings, additional concept supervision can strengthen alignment without a detectable predictive benefit, and can reduce accuracy. Its effect depends on the receiving configuration; training-stage and random-target controls distinguish auxiliary-objective gains from evidence for useful target semantics. This is an empirical regularity over the tested settings, not a universal accuracy threshold.*

Saturation is not overfitting. Overfitting is memorisation of the training set; here saturation describes the limited predictive benefit of further supervising an already represented concept.

Alignment and prediction. The auxiliary loss does push the network's class-mean features further along the target direction (alignment increases by $\Delta|r| = +0.01$ to $+0.36$), yet this yields no detectable predictive improvement: with the loss active, training accuracy stays pinned at ceiling while test balanced accuracy *falls* on the stronger recipe ($0.658 \rightarrow 0.629$; Appendix S10). These observations motivate ***ResDiag***, a residual diagnostic framework for examining predictive information beyond class-mean alignment (§4.1). Within-class residuals distinguish trials of the same emotion; controlled ablation tests their predictive role, and their encoding strength correlates with leave-one-out ensemble contribution at r=+0.74. The experiments distinguish a concept's representation from its value as an additional supervision target, qualitatively related to the few-class distillation regime (Müller et al., 2020; Loo et al., 2024; Yuan et al., 2020). The full table, transition table, and alternative-target results ($\Delta \in [-0.0193, -0.0145]$) appear in Appendix S10–S11.

4 METHODOLOGY

4.1 RESDIAG: RESIDUAL DIAGNOSTICS

§3 reported that V-axis supervision increases class-mean V-axis alignment by up to $+0.36$ while test accuracy falls and training accuracy stays at ceiling. We take the class-mean/residual decomposition as the basis of ***ResDiag***, a residual diagnostic framework combining encoding measurements, directional ablation, and ensemble-contribution analysis to examine predictive information beyond class-mean alignment. We measure leave-one-out contribution as full-ensemble BACC minus BACC after removing that checkpoint.

Directional ablation. To test the residual's predictive role, we use a directional ablation (Arditi et al., 2024): the projection $(I - vv^\top)$ with v the unit CLIP-derived within-class V-axis residual. We use an orthogonalised within-class-only ablation preserving class-mean structure (Appendix S13).

4.2 SATENSEMBLE: SATURATION-INFORMED ENSEMBLE LEARNING

ResDiag (§5.2) identifies predictive within-class information and links its encoding strength to ensemble contribution. Together with the supervision controls, this motivates ***SatEnsemble***, a saturation-

informed ensemble learning framework combining improved training with uniform averaging of checkpoint softmax probabilities. The pool contains five seeds at two training lengths. The recipe uses four EEG augmentations at p=0.6, prototype regularization with 9-D class targets (called KD in the training configuration), depth-doubling to d=6, and longer training (e=150; Appendix S14).

Prediction averaging. SatEnsemble combines checkpoint probabilities to exploit complementary predictions. ResDiag supports a predictive role for within-class information and links its encoding to ensemble contribution; these analyses do not attribute all ensemble gains to the measured residual.

5 EXPERIMENTS

We use the FACED cohort (Chen et al., 2023): 123 subjects watching 28 emotional video clips, 32 EEG channels at 250 Hz. The ten same-architecture checkpoints that make up SatEnsemble use EMOD d=6, five seeds $\times$ two training lengths (§4.2).

5.1 VALIDATING THE CONCEPT PROBES

A valence direction in LLMs. To provide a task-relevant concept probe for the Saturation Regularity and ResDiag analyses, we extract a valence direction explicitly. The construction uses nine stories. Given a language model, we author one short emotion-evocative story per FACED class and generate 50 paraphrases per story by independent LLM calls. For each class we mean-pool the language model's final-layer hidden state (layer 36 for Qwen3-4B) over tokens and over all 50 paraphrases; the resulting vector is the class *centroid*—a single point in feature space that summarises the model's representation of that emotion. Principal component analysis (PCA) on the nine centroids gives the V-axis as the first principal component, oriented so the Joy centroid projects positive. We use Qwen3-4B (Yang et al., 2025) as the lead model and verify the recipe is robust to paraphrase budget, prompt rephrasing, and model choice (Appendix S2). Five alternative prompt sets yield mean cross-prompt $r > 0.93$ across 14 models; averaging 50 paraphrases improves SST-2 AUC from 0.722 with one description to 0.832, with nonmonotonic intermediate budgets (Table 3).

Probe validation across modalities. The V-axis recipe recovers human affect across three settings that use distinct feature spaces (Table 1): the text result is true zero-shot projection; the EEG result fits a ridge readout from EEG features to the nine-story valence targets; the vision result forms nine valence-binned CLIP-image centroids over 898 images, scored on a held-out split.

Table 1: Cross-modal external validity of the V-axis recipe (PCA on nine emotion-class centroids). The Qwen3-4B V-axis is the lead model for both Text and Brain rows. Vision uses CLIP-image because the language model has no image side; a CLIP-text emotion axis on the same EEG ridge reaches r=0.87 at emotion-category resolution (per-clip-description targets: r=0.71), §5.1. Full sentiment, lexicon, and multilingual results in Appendix S3.

Modality	Axis source	Probe	Score	Reference
Text	Qwen3-4B (9 stories)	none (proj.)	**0.832** AUC	near LR (N=50)
Brain	Qwen3-4B (9 stories)	ridge LOOCV	**r=0.80**	category targets
Vision	CLIP-image (9 bins, 898)	held-out	**r=0.869**	beats sup. ridge

The LLM-side direction is a strong sentiment classifier: its 0.832 SST-2 (Socher et al., 2013) AUC (zero task labels) is comparable to a same-feature supervised LR trained on $\sim$50 labels, beats SBERT prototype-cosine (0.793; Reimers & Gurevych, 2019), and trails a 355M RoBERTa-large-MNLI (Liu et al., 2019) run zero-shot (0.912 in our evaluation) by 0.08—all from nine stories.

The direction is recovered across model families. Extracting the V-axis independently from larger checkpoints of four non-Qwen families recovers essentially the same valence ordering of the stimuli: Mistral-7B (Jiang et al., 2023) (r=0.95), Llama-3-8B (Grattafiori et al., 2024) (r=0.95), and BLOOM-3b (BigScience Workshop, 2022) (r=0.94) all align with the behavioural valence reference, with Pythia-2.8b (Biderman et al., 2023) the lone outlier (r=0.44); the broader 14-model sweep spans smaller checkpoints where alignment degrades with scale (Appendix S3). Within the Qwen3

family the V-axis is essentially scale-invariant (off-diagonal $r{=}+0.995$ across 15 size-pairs); across families the reported probe agreement is lower ($r{=}+0.585$ over 48 pairs). We therefore describe the valence direction as *recoverable* by modern LM training rather than as a single shared geometry—a cross-model similarity that, tested directly, is largely inherited from shared input structure rather than learned (Appendix S3).

Specificity. The recipe is concept-generic (20 further concepts produce a working axis, 17 reach AUC ≥ 0.95; Appendix S4). Nonce-word ablation drops SST-2 to chance and random-Gaussian directions give chance lexicon recovery, so the V-axis is a real semantic direction, not a noise artefact (Appendix S6). An Arditi-style same-model ablation (Arditi et al., 2024) on Qwen3-4B quantifies how much sentiment lives along the V-axis: a logistic-regression probe on the full features reaches SST-2 AUC 0.909, drops to 0.907 after projecting the V-axis out and retraining ($z{=}+9.8$ above a 20-direction random-direction null). The V-axis carries detectable sentiment signal but is not the sole sentiment-bearing direction.

Valence in cohort EEG. To support the concept probes used in the Saturation Regularity and ResDiag analyses, we test whether language-derived valence targets correspond to information in EEG. From the cohort EEG averaged over subjects and time, we fit a single *ridge regression* (a linear model with L_2 regularisation) predicting, under leave-one-video-out cross-validation, the emotion-category V-axis value of each of the 28 stimuli from the 160 channel–band features (differential entropy over 32 channels $\times$ 5 frequency bands). The ridge prediction matches the Qwen3-4B V-axis at $\mathbf{r{=}+0.80}$ and a CLIP-text emotion axis at $\mathbf{r{=}+0.87}$ ($n{=}28$ stimuli, both at emotion-category resolution—the axis takes one value per class); a CLIP-text axis built from the 28 per-clip descriptions reaches $r{=}+0.71$ (Appendix S7)—aligning across two distinct LLM substrates. The controls in Figure 3b include 5000 target shuffles against fixed ridge predictions; this reference does not re-fit or test the full pipeline. Projections onto a matched-norm random CLIP direction correlate with the fixed EEG ridge predictions at $r{=}+0.07$ ($p{=}0.73$). Split-half subject reliability on per-stimulus valence is $r{=}+0.99$; and the same EEG ridge predicts human-rated valence at $r{=}+0.86$, matching the CLIP-text ridge to within 0.01 (0.87 vs 0.86). Leave-one-*emotion*-out group cross-validation (no within-category leakage), multi-comparison correction, and other controls are in Appendix S7.

The ridge is not carried by a few high-affect stimuli. Restricting the existing predictions to clips outside Anger, Amusement, and Tenderness, without refitting, preserves the correlation ($r{=}0.88$ on the 19 mid-valence stimuli vs 0.87 on all 28), and a leave-one-emotion-out jackknife keeps $r \in [0.86, 0.89]$ (each still significant). The same exclusion collapses the single raw PO3/γ channel ($0.48 \rightarrow -0.02$): the pole dependence belongs to the single-channel topography, not the multi-feature ridge that carries the headline.

The brain aligns with valence, not the V-axis specifically. Swapping the target axis for a generic categorical valence axis (an integer valence rank per emotion) or the behavioural valence ratings leaves the ridge essentially unchanged ($r{=}0.89$ and $r{=}0.86$ vs 0.87; all $n{=}28$). The cohort EEG encodes a valence *ordering*; the V-axis is one instance of it, not a privileged direction—hence a probe of the saturated concept, not the concept itself.

The 14-LLM analysis spans 560M to 32B and shows model-dependent alignment (Appendix S3; Qwen3 mean $r{=}+0.796 \pm 0.011$, no significant size trend, $r{=}0.57$, $p{=}0.24$, $n{=}6$). Figure 3c separately shows the single-channel PO3/γ comparison across these models (Table 6). As a label-level check, the FACED V-axis also orders the five SEED-V (Liu et al., 2022b) emotion categories consistently with their valence ($r{=}+0.96$, $n{=}5$; this is an LLM-side ordering, not EEG decoding); the SEED-V *EEG* is re-derived from scratch in Appendix S8 (best cohort cell $r{=}0.59$). On EmoEEG-MC (Xu et al., 2025) a within-subject valence decoder reaches 0.61 accuracy ($p{=}3.4{\times}10^{-5}$) and its group axis tracks behavioural valence at $r{=}0.71$; an audio/visual confound control is inconclusive, so we frame the EEG–valence link as population-level, not confound-free.

Probe scope. The cohort signal is posterior-dominant and carried by an anger-versus-warm-positive contrast on 9 of 28 clips, so at the single-channel level it reads as an emotion-contrast more than a graded axis (the multi-feature ridge is not pole-dependent); it neither tests nor refutes frontal-alpha asymmetry (Davidson, 1992) (topography, time-resolved, and FAA analyses: Appendix S9).

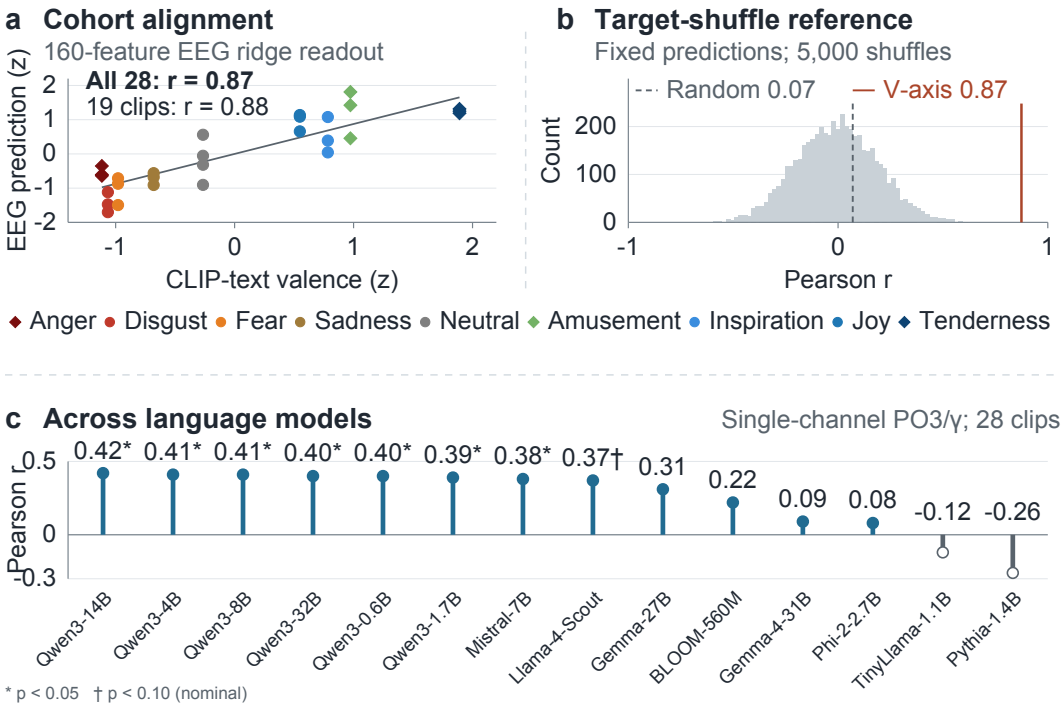


Figure 3: **Valence structure in cohort EEG. (a)** CLIP-text category targets versus 160-feature ridge predictions; diamonds mark the three categories excluded in the 19-clip check. The line is a descriptive fit. **(b)** Target-shuffle reference and random-direction control with ridge predictions held fixed. **(c)** V-axis correlations with single-channel PO3/γ across 14 language models; symbols show reported nominal significance. All panels use 28 FACED clips except the subset annotation. Protocols appear in Appendix S7.

5.2 Evaluating ResDiag

Class-mean encoding across seeds. Across the ten same-architecture checkpoints that make up SatEnsemble (§5), the class-mean subspace encodes the CLIP-text V-axis at a consistent strength ($|r| \in [0.60, 0.77]$, unsigned), with all checkpoints in a relatively tight band (the nine-story Qwen3 axis is weaker, $|r| \sim 0.3$–0.5, per-checkpoint n.s.; Appendix S13). Similar class-mean encoding motivates examining the predictive role of within-class variation. (We do *not* claim architecture-independent convergence: pooling V-axis encoding against accuracy across heterogeneous backbones yields an inflated correlation that collapses within any single architecture and does not survive a matched-norm null; see Appendix S13.)

Residual encoding and ensemble contribution. Within-class residual encoding is *consistent with* ensemble contribution. Across the 10 SatEnsemble checkpoints (§4.2), per-checkpoint residual strength correlates with leave-one-out contribution at $\mathbf{r{=}+0.74}$ (seed-clustered 95% CI $[0.50, 0.92]$; five seeds $\times$ two lengths). The bootstrap keeps each seed's two checkpoints together; per-length correlations and further detail appear in Appendix S13. This association motivates combining predictions; the ablation below tests the residual's predictive role.

Direct test: ablate the residual at inference time. The directional ablation (§4.1) drops accuracy on all 10 SatEnsemble checkpoints (mean ΔBACC$=-0.0157$, $z{\approx}7.7$ vs a matched random null); this unorthogonalized intervention also affects class-mean structure. An orthogonalised within-class-only ablation still lowers accuracy on $8/10$ checkpoints (mean $\Delta\text{BACC} = -0.008$, $z{\approx}3.5$; Appendix S13), supporting a predictive role for information beyond class means. ResDiag connects the supervision results in §3 to the complementary-prediction strategy of *SatEnsemble* in §4.2.

5.3 EVALUATING SATENSEMBLE

The recipe below tests this strategy and reaches a new state of the art.

Table 2: SatEnsemble results and recipe ablations on FACED. Both ensembles use five seeds: EMOD uses five checkpoints and SatEnsemble ten, with different training and inference costs. Replacing language-derived prototypes with random orthonormal targets preserves performance within the tested equivalence margin (TOST $\varepsilon = 0.01$, $p = 0.001$, $n = 20$; $\Delta = -0.001$). This supports a regularization benefit that does not require target semantics. Visual cascade in Appendix S14.

Method	BACC
CBraMod (Wang et al., 2025)	0.551
EMOD (Chen et al., 2025)	0.6287
EMOD ensemble (5-seed)	0.655
SatEnsemble recipe (5-seed)	.658±.010
Best single (val-sel)	**0.6755**
10-ckpt ensemble (SOTA)	**0.6948**

(a) FACED balanced accuracy. Relative to published EMOD, the best single model and ten-checkpoint ensemble gain 4.68 and 6.61 percentage points.

Recipe step	BACC	Δ
EMOD (d=3) replication	.619±.004	—
+ augmentation	.634±.004	+.015
+ prototype regularization	.644±.007	+.010
+ depth doubling (d=6)	.658±.007	+.014
+ longer training (e=150)	.658±.010	.000
5-seed e100 ensemble	.6798	+.022
5-seed e150 ensemble	.6919	+.034
10-ckpt e100+e150 (SOTA)	**.6948**	**+.037**

(b) Recipe cascade. The 10-ckpt SOTA is not significantly above the $5\times e150$ ensemble (+.003, n.s.).

The cascade starts at our in-house EMOD reproduction (0.6194 ± 0.004, ≈0.9 pt below the published 0.6287 (Chen et al., 2025); the SOTA gains are *larger* against this same-pipeline baseline, +9.1%/+12.2%) and walks up in six steps. Augmentation (four EEG augmentations at p=0.6; Appendix S14) contributes the largest single jump (+0.0149). Prototype regularization with 9-D class targets (called KD in the training configuration) adds +0.0096; a `rand9` ablation with random orthonormal directions is statistically equivalent (Table 2), supporting a regularization benefit that need not depend on language-derived semantics. Depth-doubling to d=6 adds +0.0142; longer training (e=150) does not move single-seed accuracy but yields checkpoints that ensemble better ($5\times e150$=0.6919 vs $5\times e100$=0.6798; +0.012, bootstrap 95% CI $[0.001, 0.023]$).

Prediction averaging. The ten-checkpoint pool (0.6948) is not significantly above the five e=150 checkpoints alone (0.6919; +0.003, 95% CI $[-0.004, +0.010]$). Both improved training and prediction averaging contribute: the EMOD baseline gains approximately three points from five-seed ensembling (0.624→0.655), while the SatEnsemble pool reaches 0.6948. Before ensembling, the validation-selected single model already leads published EMOD by 4.68 percentage points. Headroom comparisons, the 10-checkpoint plateau, and exploratory cross-architecture mixing are in Appendix S14.

6 CONCLUSION

The Saturation Regularity highlights the receiving model's role in auxiliary supervision. On CIFAR-10, early gains are no longer significant at convergence and aligned targets do not significantly outperform random targets; significant harm occurs under the stronger FACED recipe. ResDiag identifies predictive information beyond class means, and SatEnsemble combines improved training with checkpoint probability averaging.

Limitations. The probe transfers are direction-existence, not mechanism, claims, and the cohort signal is population-level (Appendix S9). Three limitations bound the broader reading. **(i)** We do not claim architecture-independent convergence: the pooled encoding–accuracy correlation is pseudo-replicated, scoping convergence to seeds of one architecture. **(ii)** The cross-model LLM valence geometry is largely input-driven (an untrained model reproduces it), limiting a claim of shared learned geometry. **(iii)** Transfer is content-specific (valence across text, vision, and EEG; arousal only in vision); whether the decomposition generalises to other concepts is open.

AI USE STATEMENT

Generative AI tools, including Claude Code, assisted with literature review and idea suggestions, which the authors used as starting points, as well as implementation and interpretation of routine experiments. More complex experiments and interpretation across the experimental sequence involved direct human input. AI-generated emotion-evoking stories and paraphrases were used as inputs to language models to construct emotion representations. The EEG measurements and evaluation datasets came from existing datasets; the generated text was a separate experimental input. AI also assisted with manuscript drafting and producing figures from data already verified by the authors. Codex assisted with submission formatting and manuscript review. The authors reviewed AI outputs at key stages for correctness and appropriateness and take responsibility for the final content of this work.

ETHICS STATEMENT

This work uses only publicly released, de-identified EEG datasets (FACED, SEED-V, EmoEEG-MC), collected with informed consent by their original authors under their respective institutional approvals; we introduce no new human-subjects data. The valence axis is used as a research probe of model representations and EEG responses. The near-zero within-subject effect concerns the cohort-fixed FACED channels; individual-level results depend on the dataset and analysis. This work is not intended for real-world individual affect profiling. We acknowledge the dual-use risks of affect-related modelling, and we will release code upon acceptance under a permissive licence to support scrutiny.

REPRODUCIBILITY STATEMENT

All datasets and models used are public: FACED, SEED-V and EmoEEG-MC (EEG), CIFAR-10 and SST-2 (control tasks), and the language/vision backbones fetched from public checkpoints. Dataset preprocessing, splits, and the V-axis extraction protocol are specified in Appendices S1 and S2; the full statistical methodology (paired tests, Holm correction, TOST equivalence, nested permutation, and leave-one-emotion-out jackknife) is given with each result and in the appendix. Every number reported in the paper is linked to the exact script and result file that produced it in an anonymised reproducibility package to be released upon acceptance: headline accuracies reproduce by inference on the packaged checkpoint pool, and all analysis-level numbers and figures regenerate from shipped result files with no GPU. The SEED-V cross-dataset replication is documented in full in Appendix S8.

REFERENCES

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, 2024.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, et al. Pythia: A suite for analyzing large language models across training and scaling. *arXiv preprint arXiv:2304.01373*, 2023. URL https://arxiv.org/abs/2304.01373.

BigScience Workshop. BLOOM: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022. URL https://arxiv.org/abs/2211.05100.

Ryan T Canolty and Robert T Knight. The functional role of cross-frequency coupling. *Trends in Cognitive Sciences*, 14(11):506–515, 2010. doi: 10.1016/j.tics.2010.09.001.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5:134, 2022. doi: 10.1038/s42003-022-03036-1.

Jingjing Chen, Xiaobin Wang, Chen Huang, et al. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10:740, 2023. doi: 10.1038/s41597-023-02650-w.

Yuning Chen, Sha Zhao, Shijian Li, and Gang Pan. EMOD: A unified EEG emotion representation framework leveraging v-a guided contrastive learning. *arXiv preprint arXiv:2511.05863*, 2025. To appear at AAAI 2026.

James A Coan and John J B Allen. Frontal EEG asymmetry as a moderator and mediator of emotion. *Biological Psychology*, 67(1-2):7–50, 2004. doi: 10.1016/j.biopsycho.2004.03.002.

Maurizio Codispoti, Andrea De Cesarei, and Vera Ferrari. Alpha-band oscillations and emotion: A review of studies on picture perception. *Psychophysiology*, 60(11):e14438, 2023. doi: 10.1111/psyp.14438.

Richard J Davidson. Anterior cerebral asymmetry and the nature of emotion. *Brain and Cognition*, 20(1):125–151, 1992. doi: 10.1016/0278-2626(92)90065-T.

Richard J Davidson. Well-being and affective style: neural substrates and biobehavioural correlates. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 359(1449):1395–1411, 2004. doi: 10.1098/rstb.2004.1510.

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*, 2025. arXiv:2510.21585.

Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. In *arXiv preprint arXiv:1912.02757*, 2019.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural Computation*, 4(1):1–58, 1992. doi: 10.1162/neco.1992.4.1.1.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A Norman, Orrin Devinsky, and Uri Hasson. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380, 2022. doi: 10.1038/s41593-022-01026-4.

Aaron Grattafiori et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. URL https://arxiv.org/abs/2407.21783.

Greg Hajcak, Annmarie MacNamara, and Doreen M Olvet. Event-related potentials, emotion, and emotion regulation: An integrative review. *Developmental Neuropsychology*, 35(2):129–155, 2010. doi: 10.1080/87565640903526504.

Uri Hasson, Yuval Nir, Ifat Levy, Galit Fuhrmann, and Rafael Malach. Intersubject synchronization of cortical activity during natural vision. *Science*, 303(5664):1634–1640, 2004. Inter-subject correlation (ISC): shared stimulus-locked component across viewers.

He He and Dongrui Wu. Transfer learning for brain–computer interfaces: A euclidean space data alignment approach. *IEEE Transactions on Biomedical Engineering*, 67(2):399–410, 2020. Euclidean Alignment: per-subject covariance re-centering for cross-subject BCI transfer.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024. arXiv:2405.07987.

Albert Q. Jiang et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. URL `https://arxiv.org/abs/2310.06825`.

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. arXiv:2405.18765.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (TCAV). In *International Conference on Machine Learning*, 2018. arXiv:1711.11279.

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, 2011.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, 2020.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis — connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 2008.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009. CIFAR-10 / CIFAR-100 datasets.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*, pp. 231–238, 1995.

Benedek Kurdi, Shayn Lozano, and Mahzarin R Banaji. Introducing the open affective standardized image set (OASIS). *Behavior Research Methods*, 49(2):457–470, 2017. doi: 10.3758/s13428-016-0715-3.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, 2023. arXiv:2306.03341.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems*, 2022a. arXiv:2205.05638.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022b. doi: 10.1109/TCDS.2021.3071170.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Noel Loo, Fotis Iliopoulos, Wei Hu, and Erik Vee. Linear projections of teacher embeddings for few-class distillation. *arXiv preprint arXiv:2409.20449*, 2024.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15991–16111, 2023. arXiv:2211.01786.

Matthias M Müller, Andreas Keil, Thomas Gruber, and Thomas Elbert. Processing of affective pictures modulates right-hemispheric gamma band EEG activity. *Clinical Neurophysiology*, 110 (11):1913–1920, 1999. doi: 10.1016/S1388-2457(99)00151-0.

Rafael Müller, Simon Kornblith, and Geoffrey Hinton. Subclass distillation. *arXiv preprint arXiv:2002.03936*, 2020.

Vardan Papyan, X. Y. Han, and David L. Donoho. Prevalence of neural collapse during the terminal phase of deep learning training. *Proceedings of the National Academy of Sciences*, 117(40): 24652–24663, 2020.

Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019.

Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. doi: 10.1073/pnas.2105646118.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 1631–1642, 2013.

Nicholas Sofroniew, Isaac Kauvar, William Saunders, Runjin Chen, Tom Henighan, Sasha Hydrie, Craig Citro, Adam Pearce, Julius Tarng, Wes Gurnee, Joshua Batson, Sam Zimmerman, Kelley Rivoire, Kyle Fish, Chris Olah, and Jack Lindsey. Emotion concepts and their function in a large language model. *Transformer Circuits*, 2026. URL `https://transformer-circuits.pub/2026/emotions/index.html`.

Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. Correction for attenuation; reliability of an averaged measure rises with the number of units.

Shobhita Sundaram, Stephanie Fu, Lukas Muttenthaler, Netanel Y Tamir, Lucy Chai, Simon Kornblith, Trevor Darrell, and Phillip Isola. When does perceptual alignment benefit vision representations? In *Advances in Neural Information Processing Systems*, 2024. arXiv:2410.10817.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2020. arXiv:1910.10699.

Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, 2019. arXiv:1905.11833.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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*, 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 *International Conference on Learning Representations*, 2025. arXiv:2412.07236.

X. Xu, X. Shen, X. Chen, Q. Zhang, S. Wang, Y. Li, Z. Li, D. Zhang, M. Zhang, and Q. Liu. A multi-context emotional EEG dataset for cross-context emotion decoding. *Scientific Data*, 12:1142, 2025. doi: 10.1038/s41597-025-05349-2.

An Yang et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. URL `https://arxiv.org/abs/2505.09388`.

Li Yuan, Francis E. H. Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, 2020. arXiv:1909.11723.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023.

SUPPLEMENTARY MATERIAL

S1 DATASETS, SPLITS, AND PREPROCESSING

FACED. 123 subjects, 28 emotional video clips covering 9 emotion classes. The test split (subjects 100–122) yields $23 \times 28 = 644$ *trial* pairs; trials are further sub-segmented into 3-second windows (3 windows per stimulus, boundary windows excluded) producing $644 \times 3 = 1{,}932$ test instances used in the confusion matrices and the cross-arch convergence analysis. The cohort EEG ridge of §5.1 aggregates back to the 28-stimulus level. Original FACED preprocessing detail follows below: 9 emotion categories (Anger, Disgust, Fear, Sadness, Neutral, Amusement, Inspiration, Joy, Tenderness) at 3–4 stimuli per class. EEG: 32 channels, 250 Hz sampling, 30-second clip duration. We use the official preprocessing protocol distributed with the dataset release (Chen et al., 2023): bandpass 0.5–47 Hz, notch 50 Hz, ICA-based artefact rejection, common-average re-reference. Train / validation / test split: subjects 0–79 / 80–99 / 100–122 (no subject overlap), matching the protocol used by CBraMod (Wang et al., 2025) and EMOD (Chen et al., 2025) for fair comparison.

SEED-V. 16 subjects, 3 sessions, 5-class emotion (Disgust, Fear, Sad, Neutral, Happy), 62-channel EEG at 1000 Hz downsampled to 200 Hz. Standard SEED-V split with subject-disjoint train / test. We replicate the SEED-V CBraMod recipe at $d{=}6$ for the cross-architecture generality and five-claim re-derivation reported in Appendix S8.

DE features. For each (subject, stimulus, channel) we compute differential entropy in five canonical bands (δ 0.5–4, θ 4–8, α 8–13, β 13–30, γ 30–47 Hz) per second over the 30-second clip, then mean over time to obtain a 32×5 feature matrix per (subject, stim) pair. Cohort-level analyses average across the 123 subjects to give the $28 \times 32 \times 5$ tensor used in Section 5.1.

S2 V-AXIS EXTRACTION PROTOCOL

Story prompts. Nine emotion-evocative stories (one per FACED class), each 1–3 sentences. Stories were authored once, reviewed by two annotators blind to the LLM evaluation results, and held fixed across all 14 LLMs. Verbatim text:

- **Anger.** A driver cuts you off in heavy traffic and laughs through the window. Your hands tighten on the wheel and you feel heat rise in your chest.
- **Disgust.** You open the fridge and find a container of leftovers covered in green fuzz, with a sour smell that makes you step back.
- **Fear.** Walking home alone at night, you hear footsteps quicken behind you and notice the streetlights are out.
- **Sadness.** You sort through old photographs of a friend who died last year and realise you can no longer remember the sound of their voice.
- **Neutral.** You wait in line at the post office, glancing at the clock and watching the queue slowly advance one customer at a time.
- **Amusement.** A puppy sneezes so hard it tumbles backwards and stares at the floor in confusion before sneezing again.
- **Inspiration.** A dancer with one leg performs a flawless routine to a standing ovation, demonstrating that constraint can become craft.
- **Joy.** You receive an unexpected acceptance letter from a programme you applied to months ago and forgot about.
- **Tenderness.** A grandparent gently brushes a sleeping child's hair away from their forehead, careful not to wake them.

Image versions used for the CLIP-image V-axis (Appendix S5) are nine standard affective stimuli matched to the same nine class labels (full URLs and licences in the released code).

Paraphrase generation. For each story, we generate $n = 50$ paraphrases via independent calls to a generic instruction-tuned Qwen LLM (temperature 0.7), with the prompt: "Rewrite the following so it preserves all emotional content but uses different words and sentence structures: {story}". Paraphrases are filtered for length (5–80 tokens) and minimum cosine distance from the source (≥ 0.10 on Sentence-BERT embeddings).

Hidden-state extraction. Each paraphrase is tokenised and passed through the target LM. We take the mean-pooled hidden state at a late layer L (for the Qwen3-4B lead model, the final layer, $L = 36$ of 36 — the last transformer block's output). Class centroids are $c_k = \frac{1}{50}\sum_{i=1}^{50} h_{k,i}^{(L)}$ for $k = 1, \ldots, 9$.

V-axis as PCA-PC1. Stack the nine centroids into a $9 \times D$ matrix and run mean-centred PCA. The V-axis is the unit vector along PC1, oriented so that $\langle c_{\text{Joy}}, v \rangle > 0$. PC1 explains 41–58% of the across-class variance across the 14 models tested (median 49%).

Robustness. Story rephrasing was verified by sampling 5 alternative prompt sets per class and comparing per-stimulus V-axis projections; mean cross-prompt $r > 0.93$ across 14 models. Paraphrase count was verified at $n \in \{1, 5, 25, 50\}$; SST-2 zero-shot AUC 0.72 at $n = 1$, rising to the canonical 0.832 at $n = 50$ for Qwen3-4B.

Paraphrase-count sensitivity. The full n-shot sweep on Qwen3-4B (final layer) is in Table 3. Built entirely from LLM-generated paraphrases with *no* labelled SST-2 data, the V-axis reaches AUC $= 0.832$ at n=50 paraphrases per class — comparable to a supervised logistic regression trained on 50 labelled SST-2 examples (Table 3; N=50: 0.821, N=100: 0.894), while using zero task labels. At n=1 (single sentence per class, 9 total), AUC is already 0.722.

Table 3: V-axis paraphrase-count curve (Qwen3-4B, final layer, 3 samplings per n). Standard deviations: $\pm 0.043, 0.012, 0.056, 0.019, 0.029, 0.006, 0.000$ from n=1 to n=50. The zero-label V-axis at n=50 (0.832) is comparable to a supervised logistic regression trained on 50 labelled SST-2 examples (N=50: 0.821).

n paraphrases / class	1	3	5	9	15	25	50
SST-2 AUC (Qwen3-4B)	0.722	0.792	0.751	0.817	0.811	0.845	0.832

Supervised-LR comparison. A complementary anchor: on the same SST-2 features at the lead-model final layer, a supervised logistic regression reaches AUC 0.577 at N=10 labels, 0.821 at N=50, 0.894 at N=100, and 0.924 at N=5,000. The zero-label V-axis at n=50 paraphrases (0.832) is comparable to supervised LR trained on 50 labelled examples — consistent with the data-efficiency claim (the V-axis needs *no* task labels at all).

S3 V-Axis Validation Across Language Models

Sentiment benchmark table. Table 4 reports the zero-shot V-axis projection on SST-2 and contrasts it against (i) a same-corpus 5k-example supervised logistic-regression upper bound and (ii) two zero-shot reference baselines (SBERT prototype-cosine and a 355M RoBERTa-large-MNLI entailment classifier). The takeaway is that nine LLM stories sit above the generic-encoder baseline and within 0.08 of an NLI-supervised model that is two orders of magnitude larger in supervised exposure. Lexicon-level results on EmoBank, Warriner, and NRC appear in Table 5.

LLM-as-judge baseline (steelman). A reviewer-natural question is whether prompting the same LLM directly as a sentiment classifier ("LLM-as-judge") would beat the V-axis projection. We test this with Qwen3.5-2B as the judge on SST-2: a minimal binary prompt reaches AUC 0.951 at 165 s/1k examples, and a 5-shot in-context-learning variant reaches 0.957 at 253 s/1k. The V-axis projection trails on raw quality (0.832) but requires no per-example autoregressive generation—a single forward pass—and exposes a single re-usable direction in feature space (composable with downstream losses, ablatable, transferable across modalities; §5.1, §4.1). The competitive advantage

Table 4: Zero-shot V-axis sentiment results. The SST-2 V-axis AUC (0.832) is the verified Qwen3-4B lead-model number: the zero-label V-axis is comparable to a same-feature supervised LR trained on 50 labels, while a 5k-label same-feature classifier reaches 0.924 (Table 3)—the V-axis carries the structure of the supervised problem without seeing sentiment labels. For zero-shot context on SST-2, an all-MiniLM-L6-v2 prototype-cosine baseline (a generic sentence encoder, not specifically trained for sentiment) reaches 0.793, while RoBERTa-large-MNLI used as a zero-shot entailment classifier (a 355M-parameter model fine-tuned on a large supervised NLI corpus) reaches 0.912. Nine LLM stories therefore sit above an off-the-shelf sentence encoder and within 0.08 of a 355M-param NLI-supervised model on the same zero-shot evaluation.

Benchmark	V-axis AUC	Supervised LR (5k)
SST-2 binary (Socher et al., 2013)	0.832	0.924
SST-2 zero-shot reference baselines		
SBERT prototype cosine (Reimers & Gurevych, 2019)	0.793	—
RoBERTa-large-MNLI entailment (Liu et al., 2019)	0.912	—

is not raw zero-shot accuracy — it is interpretability and re-use, which the LLM-as-judge baseline does not provide.

Three regimes against behavioural valence. The 14-model validation shows heterogeneous alignment: models cluster into three regimes by per-stim correlation against a behavioural valence reference ($n = 28$ FACED stim, mean of 123 subjects' valence ratings).

- **Top tier** (per-stim $r > 0.85$, $n = 7$): all six Qwen3 sizes (Qwen3-14B $r = 0.944$, Qwen3-4B 0.944, Qwen3-8B 0.941, Qwen3-32B 0.937, Qwen3-0.6B 0.933, Qwen3-1.7B 0.930) plus Mistral-7B (0.935). Within Qwen3 the correlations are tightly grouped (mean $r_{\text{behav}} = 0.938 \pm 0.005$, mean $r_{\text{eeg}} = 0.796 \pm 0.011$).
- **Middle tier** ($0.6 \leq r \leq 0.85$, $n = 4$): Llama-4-Scout-17B (0.826), Gemma-27B (0.807), Gemma-4-31B (0.715), Phi-2-2.7B (0.651).
- **Low alignment** ($r < 0.4$, $n = 3$): Pythia-1.4B (0.276), TinyLlama-1.1B (0.128), BLOOM-560M (0.031).

The within-Qwen3 scaling is essentially flat: the six Qwen3 sizes span only $r_{\text{behav}} \in [0.930, 0.944]$ and $r_{\text{eeg}} \in [+0.781, +0.812]$ (s.d. ± 0.011) with no monotonic size trend; the behaviourally best-aligned models are mid-size Qwen3 (4B–14B).

Within-family and cross-family agreement. The 14-model agreement analysis reports $\bar{r} = 0.995 \pm 0.003$ within Qwen3 (15 pairs), compared with 0.585 ± 0.289 for Qwen3–non-Qwen pairs (48 pairs). The Qwen3–Mistral subset has $\bar{r} = 0.939$ (6 pairs). These summaries describe agreement in the probe analysis; they do not establish a shared hidden-space direction across models with different feature dimensions.

V-shape across architectures. The same V-shape emerges across five LM families (Qwen, Pythia, BLOOM, Llama-4-Scout, Phi-2), and across bidirectional encoders BERT and RoBERTa, with the peak at layer $L_{\max} - 1$ in each case. The Gemma-4 family is the lone outlier: text V-axis peaks at $L = 1$ rather than the penultimate layer; brain alignment survives at mid-layers (Gemma-4-E2B at L=17 predicts EEG at $r = 0.79$; Gemma-4-E4B at L=21 at $r = 0.64$).

S4 Concept Library and Compositionality

Compositional axis arithmetic. Combining the extracted direction of a stylistic concept (politeness) with that of an emotional concept (happiness) predicts 4-quadrant categorical labels at 79% accuracy on a held-out generation set, $3.16\times$ above chance, with uniform per-cell precision in the $[0.68, 0.86]$ range. Subtractive orthogonalisation cleanly separates the two: happy-AUC $0.912 \rightarrow 0.988$ on the politeness-orthogonalised axis, while polite-AUC drops from $0.793 \rightarrow 0.560$.

Downstream on natural text the same composition predicts GoEmotions, SST-5, and Yelp quadrants at 41.8%–46.7% (documented scope refinement).

Recipe generality across 20 concepts. For each of 20 unrelated dimensions (emotional, stylistic, abstract) we generated pole-contrastive stories and applied the identical extraction protocol. **20 of 20 concepts reach binary AUC** > 0.65 **on held-out test stories, 17 of 20 reach** ≥ 0.95**, 11 are perfect at** 1.00 (Figure 4).

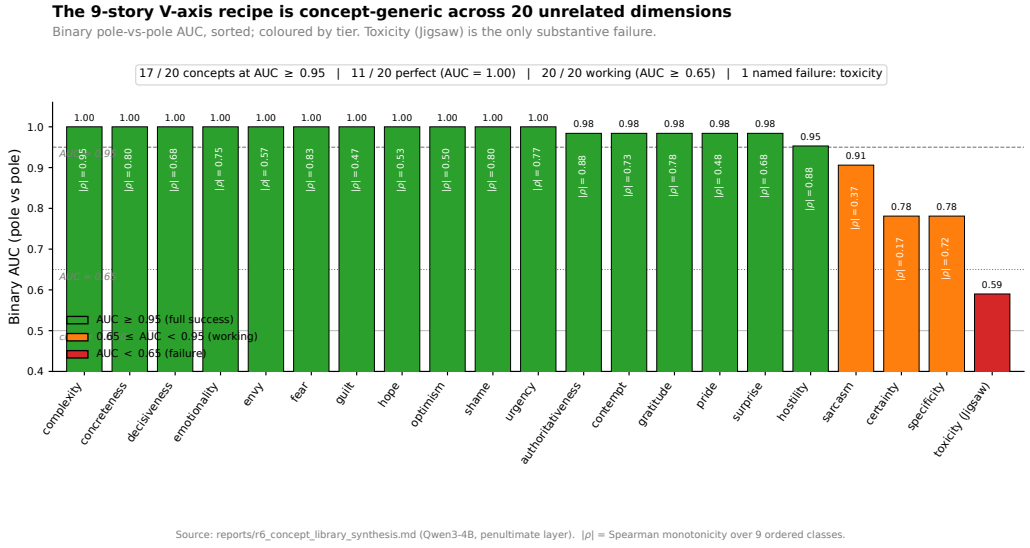


Figure 4: Binary pole-vs-pole AUC across 20 unrelated concepts (sorted descending), plus the out-of-scope toxicity case (Jigsaw, AUC 0.59).

Cross-concept transfer matrix. The 20×20 matrix has self-AUC mean 0.97 on the diagonal and off-diagonal mean 0.84 (gap $+0.13$): the recipe extracts *concept-specific* directions rather than a single global affect axis (Figure 5). Most general source axes: sarcasm (0.89), fear (0.88), envy and shame (0.87). Most specific: decisiveness (0.78), complexity (0.78). Most captured: urgency (0.97), fear (0.95), concreteness (0.94). Most concept-specific: specificity (0.65), sarcasm (0.67).

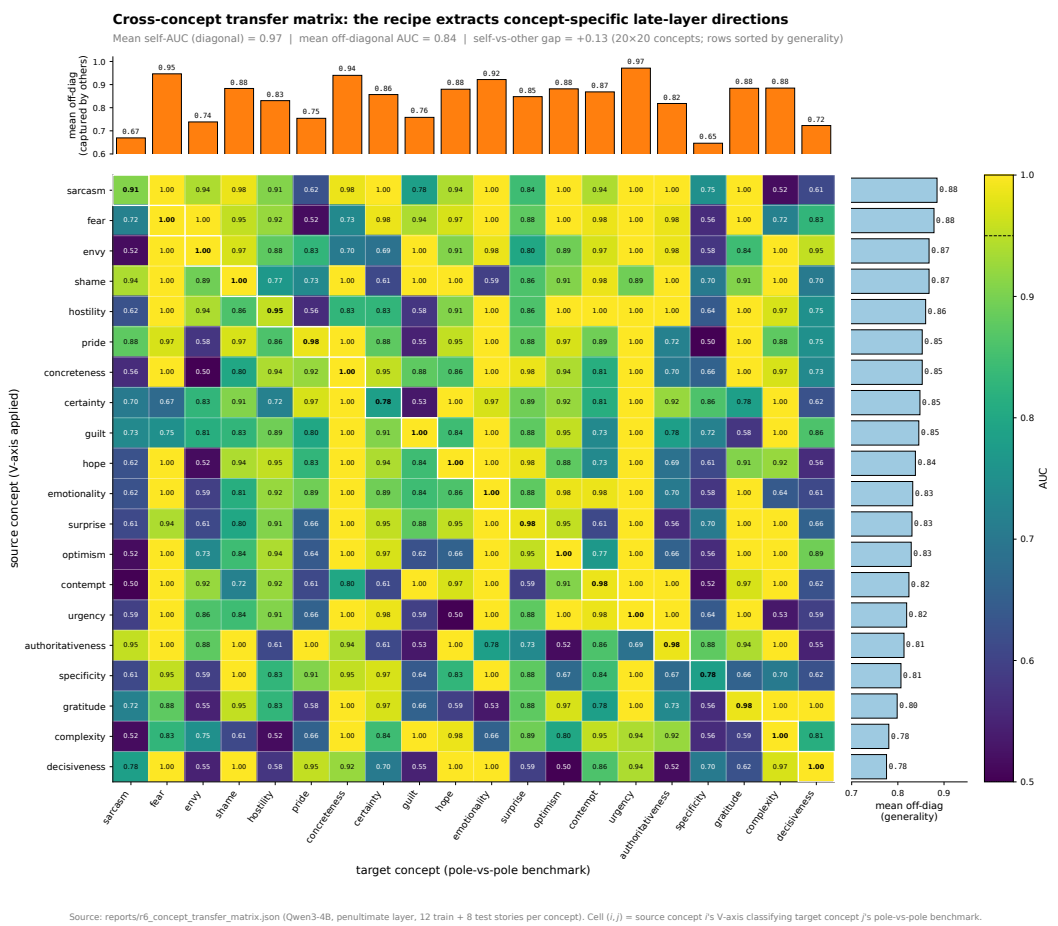
Toxicity: out-of-scope. Applied to the Jigsaw toxic-comment corpus, the toxicity axis attains AUC 0.59. Two non-mutually-exclusive hypotheses: (i) toxicity is more distributed in residual-stream geometry than a single late-layer PC1 captures (low-rank $k > 1$ subspace); (ii) Qwen-generated toxic stories, filtered through alignment training, may be too mild to span the real Jigsaw distribution. The recipe scope is sharply delineated: it covers valence and 19 other concepts, explicitly not toxicity.

S5 Multilingual and Vision Generality

Multilingual transfer. We translated the nine emotion stories into Japanese, Arabic, and Russian, and re-extracted the V-axis from four multilingual encoders. Direct application of an English-centric causal LM to non-English stories collapses to chance — a clean English-bias diagnostic. Switching to mT0-base (Muennighoff et al., 2023) (277M-param multilingual encoder–decoder) fully recovers the SST-2 result and *exceeds* the English baseline in every language tested (Figure 6).

Across the three languages the mT0-derived axes remain mutually aligned but not identical, consistent with concept-level rather than token-level encoding.

Cross-modal probe validation (CLIP vision). The CLIP-image probe (Radford et al., 2021) uses centroid PCA and reaches Pearson $r = 0.869$ for valence and $r = 0.803$ for arousal on OASIS (Kurdi et al., 2017). Both exceed the tested supervised-ridge baselines (0.836 and 0.789, respectively). These comparisons support transfer of the probe construction to visual affect; they do not establish



Source: reports/r6_concept_transfer_matrix.json (Qwen3-4B, penultimate layer, 12 train + 8 test stories per concept). Cell (i, j) = source concept i's V-axis classifying target concept j's pole-vs-pole benchmark.

Figure 5: Cross-concept transfer matrix (best-orientation AUC). Diagonal: self-classification (white-bordered cells). Off-diagonal: source-axis-applied-to-target-benchmark AUC. Sorted top-to-bottom by mean off-diagonal. Mean diagonal 0.97, mean off-diagonal 0.84, gap +0.13.

superiority over all image-affect methods. The valence and arousal CLIP axes are nearly orthogonal ($|\cos\theta| = 0.013$).

S6 Specificity Controls (Nonce, Random, Arousal)

Nonce-word ablation. Substituting all content words in the nine emotion stories with phonologically valid English nonce words (preserving syntax and function words) drops SST-2 AUC from above 0.83 to chance (≈ 0.52) — the second decimal. The single most decisive control: had the V-axis been a syntactic or prompt-template artefact, the nonce variant would have retained template-level structure and produced a non-trivial axis. The signal is encoded in semantic content.

Random-direction baseline. Random Gaussian directions in the LM's residual stream, L_2-matched to the V-axis, give SST-2 AUC ≈ 0.52 and lexicon $|r| \approx 0.07$–0.09. By contrast the V-axis (Qwen3-4B, final layer) bootstrap 95% CIs are tight and well separated from chance: SST-2 AUC $[0.804, 0.858]$, EmoBank V $|r| \in [0.503, 0.531]$, Hu & Liu $|r| \in [0.709, 0.756]$ (1,000-sample bootstrap over examples; statistical methods in Appendix S16).

Arousal asymmetry: where the recipe fails in text. Table 5 side-by-sides the same nine-story PCA recipe instantiated on a valence axis (V-axis, this paper) versus an arousal axis (A-axis, identical extraction protocol with the nine-story prompts re-anchored on arousal poles). If the recipe were content-blind — merely a high-capacity dimensionality-reduction trick that produces a graded

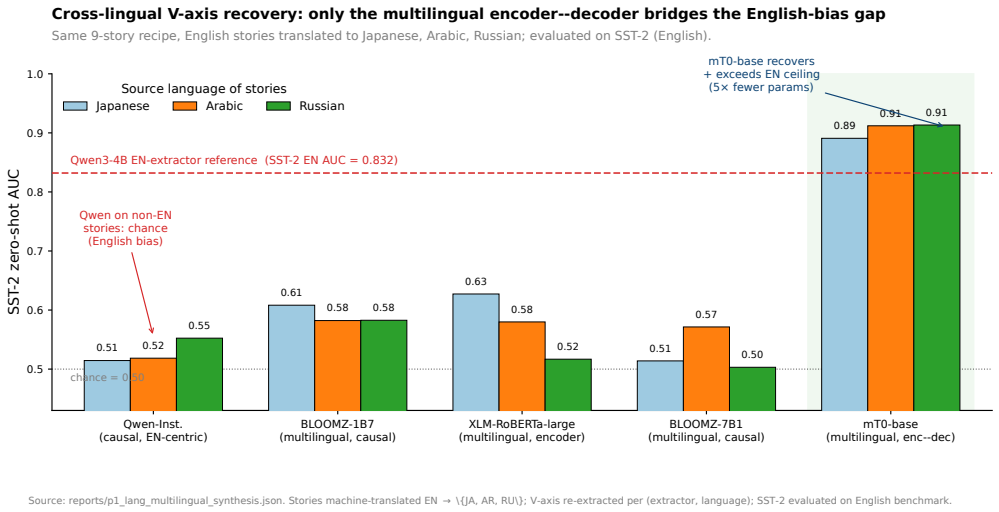


Figure 6: Cross-lingual V-axis recovery. Direct causal-Qwen transfer collapses (English-bias); mT0-base recovers and exceeds the English reference at roughly $5\times$ smaller parameter count (mT0-base 277M vs. the 1.5B causal reference) across three typologically distant languages.

affective scale from any pole-vs-pole prompt set — the V and A columns should agree across the five textual rows. They do not: V transfers cleanly across SST-2, EmoBank, Warriner, NRC, and OASIS, while A collapses to near-chance on every textual benchmark. The OASIS-image and FACED-EEG rows are listed for contrast: vision encodes an arousal axis the same recipe recovers (r=0.803), whereas on the brain arousal is weak and sub-threshold ($r \in [0.18, 0.41]$, against valence 0.87). The clean recovery is thus text→vision; arousal fails in both text and brain.

Table 5: Valence transfers across modalities; arousal does not in text but does in vision. The same recipe applied to an arousal axis yields OASIS arousal correlation $r = 0.803$, above the tested supervised-ridge baseline (0.789), with weak results on the tested text benchmarks.

Benchmark	V-axis (valence)	A-axis (arousal)
SST-2 binary AUC	0.832	—
EmoBank V $\lvert r\rvert$ / EmoBank A $\lvert r\rvert$	0.52	0.01–0.06
Warriner V $\lvert r\rvert$ / Warriner A $\lvert r\rvert$	0.703	≤ 0.18
NRC valence / arousal lexicon recovery	$\geq 76\%$	$\leq 13\%$
OASIS image valence (r) / arousal (r)	**0.869**	**0.803**
FACED EEG PO3/γ r	**0.870** (9 poles)	$r \in [0.18, 0.41]$

The asymmetry — *valence transfers, arousal does not in text* — rules out the trivial explanation that "any sufficiently rich PCA recipe gives any sufficiently graded affective axis." Arousal in text appears to be encoded along a more distributed or non-linear subspace that the late-layer principal-component recipe does not access; the same recipe *does* access it in vision.

S7 Cohort EEG Probe Validation: Additional Analyses

Per-stimulus reliability. Per-stimulus split-half reliability (odd/even subjects, n≈62 each) is $r = 0.988$, indicating a clean stimulus-level signal. Trial-level correlations ($n = 3{,}444$ subject–stim pairs) range from $r = 0.17$–0.21 ($p < 10^{-23}$); class-level ranking ($n = 9$ centroids) hits $r = +0.886$ vs. behavioural valence. The cohort ridge reaches $r = 0.87$ with the CLIP-text V-axis and $r = 0.80$ with the nine-story Qwen3-4B V-axis; the split-half result separately quantifies the reliability of the aggregated stimulus measurements.

Group cross-validation (no within-category leakage). The headline ridge is fit with leave-one-video-out CV. Re-fitting under leave-one-*emotion*-out group CV—withholding all clips of an emotion category together, so no same-category stimulus is seen during fitting—the prediction still tracks the V-axis at category level r=0.89 (n=9) and per-clip r=0.75 (n=28, p=4×10^{-6}). The leave-one-video-out headline correlations are 0.87 for CLIP-text and 0.80 for Qwen3-4B targets. These results support correspondence beyond fitting on other clips from the same emotion category.

EEG readouts across language models. Figure 7 compares EEG-readout correlations for language-derived V-axis targets. The seven models with the strongest behavioural alignment exceed their per-model random-direction nulls, whereas the three with low behavioural alignment sit at or below the null median.

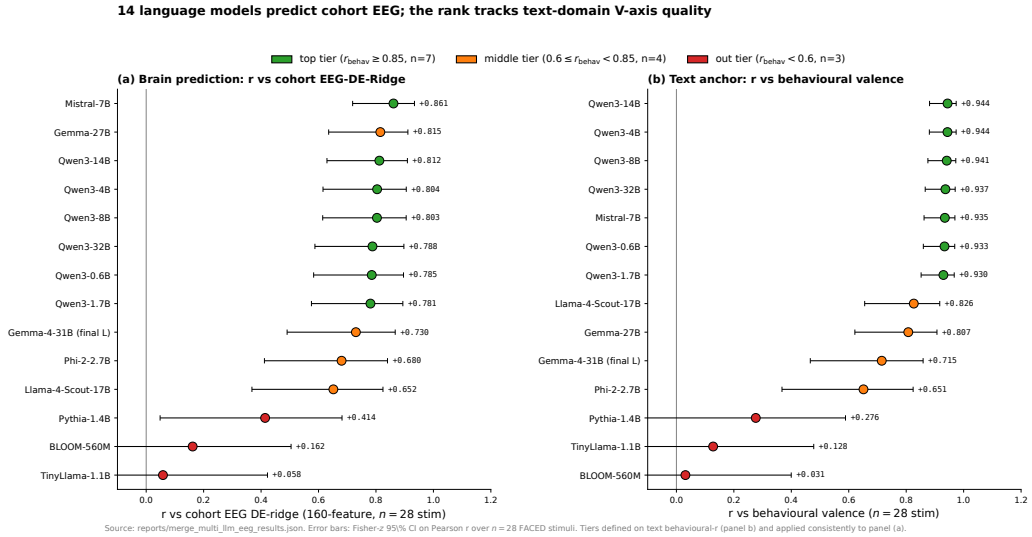


Figure 7: EEG readouts and behavioural alignment across language models. The seven models with the strongest behavioural alignment exceed their random-direction baselines. Among these models, the relation between log hidden dimension and EEG correlation is nonsignificant ($r = 0.44$, $p = 0.33$, $n = 7$), as it is within Qwen3 ($r = 0.57$, $p = 0.24$, $n = 6$).

Single-channel comparison across language models. Table 6 reports the single-channel comparison across the 14 language models. Each correlation compares a model's V-axis projections with the cohort PO3/γ feature across 28 clips. These values use one channel–band feature and are distinct from the 160-feature ridge readouts in Figure 3 and Figure 7.

Table 6: V-axis correlation with cohort PO3/γ across 28 FACED clips. Values are rounded to two decimals. This metric uses one channel–band feature, separately from the multi-feature ridge readout.

Model	Pearson r	Model	Pearson r
Qwen3-14B	0.42	Llama-4-Scout	0.37
Qwen3-4B	0.41	Gemma-27B	0.31
Qwen3-8B	0.41	BLOOM-560M	0.22
Qwen3-32B	0.40	Gemma-4-31B	0.09
Qwen3-0.6B	0.40	Phi-2-2.7B	0.08
Qwen3-1.7B	0.39	TinyLlama-1.1B	−0.12
Mistral-7B	0.38	Pythia-1.4B	−0.26

Time-locked dynamics. Mid-late α/β peak at t = 18–21 s coincides with the LPP / sustained-attention window of Hajcak et al. (Hajcak et al., 2010) and Codispoti et al. (Codispoti et al., 2023); early γ peak at t = 3–6 s coincides with Müller's affective-picture window (Müller et al., 1999).

Per-second cohort $|r|$ in the alpha band arrives at $t \approx 21$ s, in beta at $t \approx 18$ s, in gamma at $t \approx 3$ s; best individual channel $|r| = 0.68$ at $t = 21$ s.

Per-subject peak heterogeneity. Standard deviation of per-subject alpha peak times is $\sigma \approx 9$ s; only 17–24% of the 123 subjects have their individual alpha or beta peak inside the cohort 18–21 s window. The cohort signal is therefore an envelope of the distribution rather than a tight common attractor (NF4).

Gemma-4 boundary check. Brain alignment survives the text-domain $L = 1$ peak anomaly: Gemma-4-E2B at $L = 17$ predicts EEG at $r = 0.79$ ($p \approx 10^{-7}$), Gemma-4-E4B at $L = 21$ at $r = 0.64$ ($p \approx 2 \times 10^{-4}$). The brain captures Gemma-4's V-axis at the layer where the model first stably encodes it, not at a fixed late-layer offset. This cross-model alignment refines to a property of the late-but-not-necessarily-penultimate residual stream geometry.

S8 SEED-V Replication Full Numerics

We re-test all five FACED claims with a SEED-V-derived V-axis. To remove FACED leakage, we re-build the V-axis from scratch using the same Section 5.1 protocol applied to the 5 SEED-V emotion classes (Qwen3-4B at the final layer L=36, 50 paraphrases per class, PCA on the 5 centred centroids).

Step 1: SEED-V cohort EEG–LLM alignment. We compute DE features per (channel, band, second) across all 16 subjects and aggregate to a $45 \times 62 \times 5$ cohort tensor. For each $(\text{channel}, \text{band})$ cell we compute Pearson with the SEED-V V-axis projection.

- **Best cohort cell**: PZ / θ, $r = +0.592$.
- **Region-mean** $|r|$: occipital 0.300, parietal 0.315, central 0.302, frontal 0.234.
- **Posterior dominance**: occipital $-$ frontal $= +0.066$.
- **Random-direction null** (200 directions matched in L_2): empirical $|r|_{\max} = 0.592$ vs. null 95th percentile $= 0.478$, $p < 0.005$.

Step 2: SEED-V brain topography. The Davidson FAA probe (positive vs. negative valence, right $-$ left) tests whether the FACED FAA topography holds on SEED-V; it does not replicate in direction (Table 7):

Table 7: Davidson FAA on SEED-V. Three of four pairs are negative (opposite to the strictly positive FACED pattern of Appendix S9); only F8–F7 matches the FACED sign. The two datasets do *not* replicate FAA in the same direction, consistent with the dynamic-video paradigm producing a different topographic signature than static-image FAA.

Electrode pair (right $-$ left, alpha)	(pos $-$ neg) DE Δ
F4 – F3	-0.0018
F8 – F7	$+0.0050$
Fp2 – Fp1	-0.0094
AF4 – AF3	-0.0039

Step 3: SEED-V within-architecture V-axis encoding. Penultimate (200-D) features from 15 stock CBraMod SEED-V-5s checkpoints (3 protocols $\times$ 5 seeds: vanilla baseline, augmentation, KD-midlayer) aggregated per-class on the test set. PC/best-PC scores correlated with the SEED-V V-axis class projection give $r(\text{BACC}, \text{trial-best-PC}\, r) = +0.632$ ($p = 0.012$, $n = 15$; this is the direction used for the causal ablation below) and $r(\text{BACC}, \text{class-best-PC}\, r) = +0.601$ ($p = 0.018$). The SEED-V BACC dynamic range is much narrower than FACED, so the raw class-PC1 correlation is not significant; the best-PC estimators reach significance. Because the best-PC is chosen post hoc from the candidate components, this significance is not corrected for that selection; we therefore treat SEED-V as within-architecture, selection-sensitive corroboration and rely on FACED (pre-specified class-PC1) for the primary claim. These are 15 checkpoints of a single backbone (CBraMod), so we read this as within-architecture V-axis encoding, not architecture-independent convergence.

Step 4: SEED-V V-axis as supervision. Adding a V-axis topo-MSE auxiliary loss (λ_{vax}=0.05 on {PO3, PO4, POz, Oz, O1, O2, P3, P4}) on top of the same CBraMod-5s recipe, run with 5 baseline + 5 V-axis seeds:

Table 8: SEED-V V-axis-as-supervision (within seed noise of zero). Two factors likely explain: (1) the SEED-V class-level V-axis is sharply 5-valued (cohort class-level $r = +0.958$), so adding a class-level V-axis target on top of one-hot supervision is largely redundant; (2) the SEED-V BACC dynamic range (0.43–0.45) is much narrower than FACED (0.55–0.58), so a small effect is hard to detect with five seeds.

Variant	BACC (mean ± s.e. across 5 seeds)	vs. baseline Δ
Baseline (no V-axis loss)	0.4307 ± 0.0009	—
+ V-axis topo loss	0.4325 ± 0.0017	$+0.0018$

Step 5: SEED-V directional ablation (Arditi-style). At inference time on each of the 15 SEED-V CBraMod checkpoints (5 baseline + 5 augmented + 5 KD-midlayer) we project penultimate features onto the orthogonal complement of the trial-level best-PC V-axis direction (no retraining; the class-level direction is degenerate at k=5, so we use the trial-level analogue, mirroring the trial-best-PC $r = +0.632$ estimator from Step 3). Every checkpoint shows a negative $\Delta\text{BACC} \in [-0.0291, -0.0093]$ (population mean -0.0173 ± 0.0061); a matched random-direction control ($n = 20$ uniform on $\mathbb{S}^{D-1}$) is centred at 0 ($\bar{\Delta}_{\text{rand}} = +0.00013 \pm 0.0008$). Per-checkpoint z-scores range $[+9.65, +44.92]$ (mean $+23.6$, all $p < 0.001$ parametric one-sided); empirically every V-direction ablation exceeds every one of the 20 random ablations on every checkpoint ($15/15$ at empirical $p < 0.05$). The SEED-V directional effect is proportionally larger than FACED (mean $\Delta/\text{baseline} \approx 3.9\%$ vs. 2.4% on FACED) and the mean z is $3\times$ FACED's. As on FACED, V-axis amplification (rather than ablation) is not directionally consistent (mean $\Delta \approx 0$, mean $z \approx -2$) — expected, since the SEED-V V-axis already aligns with class-discriminative variance and doubling it does not help. Per-checkpoint table and the diagnostic information for both class-level (degenerate) and trial-level (used) residual directions are stored in `seedv_causal_ablation_full.json`. The ablation direction is robust to the LLM used to build the V-axis: rebuilt with the Qwen3-4B lead model, 14 of the 15 checkpoints select an *identical* intervention direction ($|\cos|$=1.00 against the reported build; the class projections of the two builds agree at $r = 0.99$), so the ablation result is unchanged under the lead-model V-axis.

Third-dataset consistency check: DEAP. As a further test of the Saturation Regularity on an independent dataset (DEAP (Koelstra et al., 2011); 32 subjects, 32 channels, 128 Hz), we trained a binary-valence classifier (valence rating ≥ 5, sourced from the raw DEAP labels) from the same pretrained EMOD backbone, splitting by *(subject, trial)* unit so that all windows of a trial stay on one side. The converged baseline reaches $\overline{\text{BACC}} = 0.544$. We evaluate auxiliary supervision under this configuration: across three seeds and three loss weights ($\lambda \in \{0.1, 0.5, 1.0\}$), the pooled change is $\Delta\text{BACC} = -0.019 \pm 0.044$ (n=9 seed–weight combinations, reported one-sample t=-1.28, p=0.24), not significantly different from zero (two of three seeds negative, one positive). DEAP thus gives no detectable gain under the tested recipe, as does SEED-V, whose supervision effect is likewise nonsignificant ($\Delta\text{BACC} = +0.0018$, Table 8)—a complementary observation to the accuracy decrements on the stronger FACED recipe. These comparisons do not define a common accuracy threshold across datasets. Per-run numbers are in `deap_sat/SUMMARY.json`.

S9 Brain Topography Deep Dive

Davidson FAA full pairs. Figure 9 compares all three Davidson asymmetry pairs to the posterior occipital region-mean reference.

9-stim contrast and Simpson's paradox. Table 9 decomposes the cohort-level brain–LM correlation by stimulus subset, isolating which of the 28 FACED stimuli carry the V-axis signal. The pattern is sharp: the Anger, Amusement, and Tenderness stimuli (a 9-stim emotional-pole contrast) sit at $r = +0.870$, while removing them collapses the residual 19 stimuli to $r = -0.015$. The full-cohort $r = +0.478$ is therefore not distributed evenly across emotion classes but concentrated

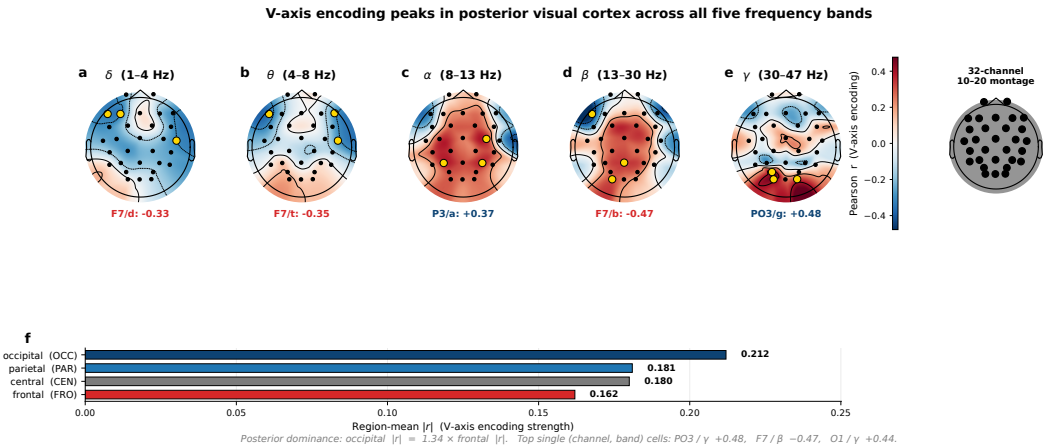


Figure 8: **Region-averaged V-axis encoding is posterior-dominant; individual-channel extrema in δ, θ, and β localise to frontal F7. (a–e)** MNE-style scalp topomaps, one per band (δ, θ, α, β, γ), of the cohort Pearson correlation between the LLM V-axis (28 stimulus projections) and per-channel band differential entropy on FACED (n=123 subjects). Diverging colourmap: positive (red) vs. negative (blue) correlation; symmetric limits across bands. Gold circles mark the top channel per band; text under each topomap names that channel and its r value (the globally strongest cell is PO3/γ, r=+0.48, but the per-band extrema in δ, θ, and β are frontal, e.g. F7/β r=−0.47). The colourbar (centred) gives Pearson r; the small head diagram (right) shows the 32-channel 10–20 montage used in all five panels (A1/A2 mastoids excluded from plotting). **(f)** Region-mean $|r|$ aggregated over all five bands: occipital (0.21) > parietal (0.18) > central (0.18) > frontal (0.16). The cohort signal is posterior-dominant for video-evoked emotion—a paradigm-level finding for dynamic stimuli rather than a refutation of the static frontal-alpha asymmetry literature, which we replicate qualitatively in direction at smaller magnitude.

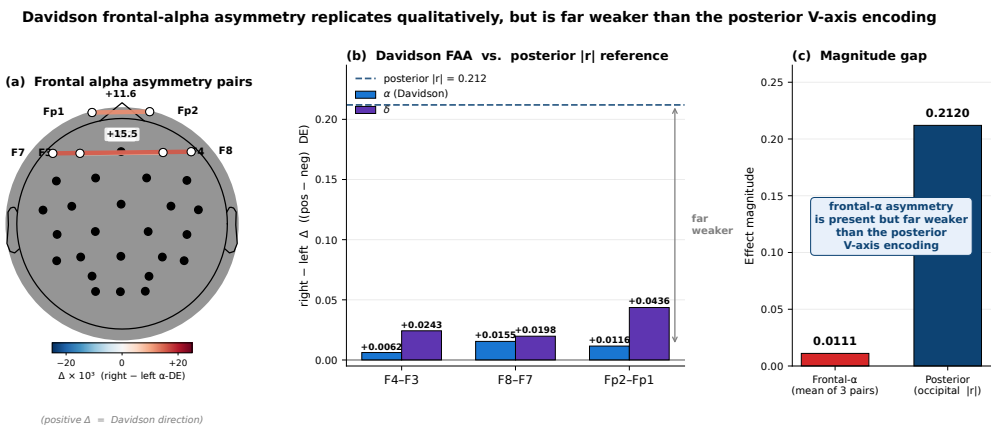


Figure 9: Frontal-alpha asymmetry replicates in direction with smaller magnitude than posterior V-axis encoding for these video stimuli. (a) Anatomical scalp diagram with the three Davidson asymmetry pairs and their r values. (b) Per-pair alpha and delta asymmetry bars. (c) Direct comparison to the occipital region-mean $|r| = 0.21$.

in the emotional-pole subset — a stimulus-level "where the cohort signal lives" decomposition that motivates the per-subject Simpson's breakdown immediately below the table.

- **Cohort r at fixed top-8 channels**: $+0.021$
- **Mean per-subject r at fixed top-8 channels**: -0.062
- **Per-subject best-channel oracle (1 cell)**: $\overline{|r|} = 0.551$

Table 9: The cohort V-axis effect is essentially an *anger-versus-warm-positive* contrast across nine stimuli.

Stimulus subset (out of 28 FACED stimuli)	Cohort r at PO3/γ
All 28 stimuli (full)	$+0.478$
Anger only (n=3)	not estimable
Anger + Amusement + Tenderness (n=9)	$+0.870$
Excluding Anger (n=25)	$+0.327$
Excluding Anger + Amusement + Tenderness (n=19)	-0.015

- **Per-subject best (channel, band) oracle**: $\overline{|r|} = 0.616$

The all-band oracle ($\overline{|r|} = 0.616$) exceeds the cohort-fixed top-8 ($+0.021$) by $+0.60$ in absolute $|r|$, indicating considerable per-subject signal that the cohort summary suppresses — a Simpson's-paradox dynamic, in which cohort-level structure need not hold within individuals, which we document per-subject in Figure 10. This oracle is a maximum over 160 channel$\times$band cells at N=28 stimuli and therefore carries post-selection inflation: a per-subject label-permutation null (shuffle stimuli, re-take the 160-cell max, 1000 draws) has mean $\overline{|r|}$=0.45 and 95th percentile 0.60, so a single subject's oracle is not individually distinguishable from selection noise. The cohort-mean oracle nonetheless exceeds this null ($p < 0.001$): the per-subject signal is real in aggregate even though individual oracle values are largely in-sample maxima.

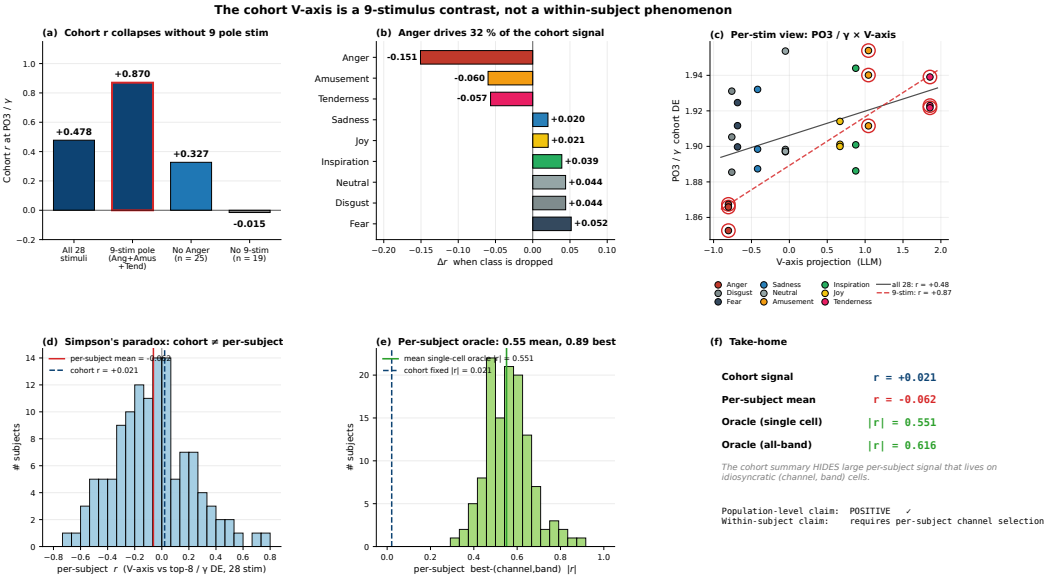


Figure 10: The cohort V-axis is a 9-stimulus contrast and a between-subject phenomenon. Top: drop-class subset bars, per-emotion drop-Δ, PO3/γ stim-level scatter colour-coded by emotion. Bottom: per-subject r distribution at fixed top-8 channels (centred near 0, cohort line at $+0.00$); per-subject best-channel oracle distribution (centred at $|r| \approx 0.55$); summary callout.

Time-resolved peaks. Figure 11 traces per-second cohort $|r|$ across all five frequency bands.

- **Alpha** (8–13 Hz): peak at $t = 21$ s, cohort $r = 0.40$; best individual stimulus $r = 0.68$.
- **Beta** (13–30 Hz): peak at $t = 18$ s, cohort $r = 0.41$; best stim $r = 0.62$.
- **Gamma** (30–47 Hz): peak at $t = 3$–6 s, cohort $r = 0.21$; best stim $r = 0.61$.
- **Delta/Theta** (0.5–8 Hz): peak at $t = 12$ s, cohort $r \approx 0.3$.

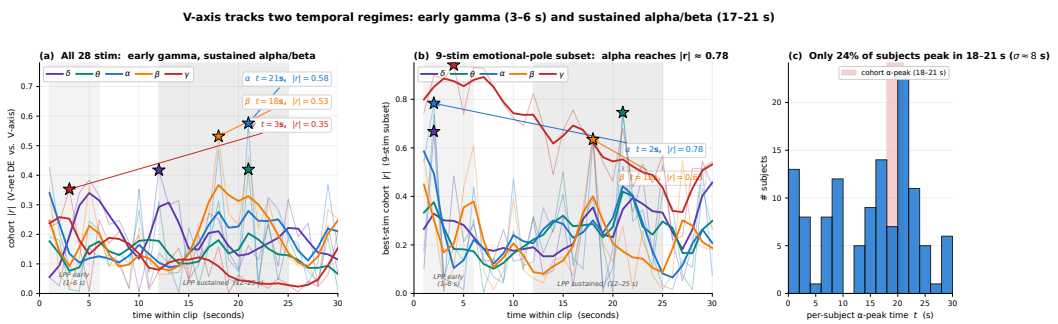


Figure 11: Time-resolved V-axis encoding across the 30-second clip. **(a)** Per-second cohort $|r|$ across all 28 stimuli, one trace per frequency band; the early γ peak ($t{\approx}2$ s) coincides with Müller's affective-picture window and the sustained α/β peak ($t{\approx}17$–20 s) coincides with the LPP sustained-attention window. **(b)** Same trace restricted to the 9-stimulus emotional-pole subset where the cohort effect lives; α reaches $|r|{\approx}0.78$. **(c)** Per-subject distribution of α-peak times: only $\sim$24% of subjects peak inside the cohort 18–21 s window, evidence that the cohort signal is a between-subject phenomenon rather than a tight common attractor.

Functional connectivity. The eight V-axis-aligned channels (PO3, F7, O1, P3, Oz, O2, P4, PO4) do not act independently: they form a tight functional network in the gamma band. Pairwise γ-DE correlation has mean off-diagonal $r = 0.67$ across the 8 channels — substantially above the whole-band pairwise average. Structure is dominantly a posterior-occipital–parietal cluster (pairwise mean $r \approx 0.73$) with F7 as a single *frontal hub* (mean correlation to the seven posterior nodes $r \approx 0.50$; Figure 12).

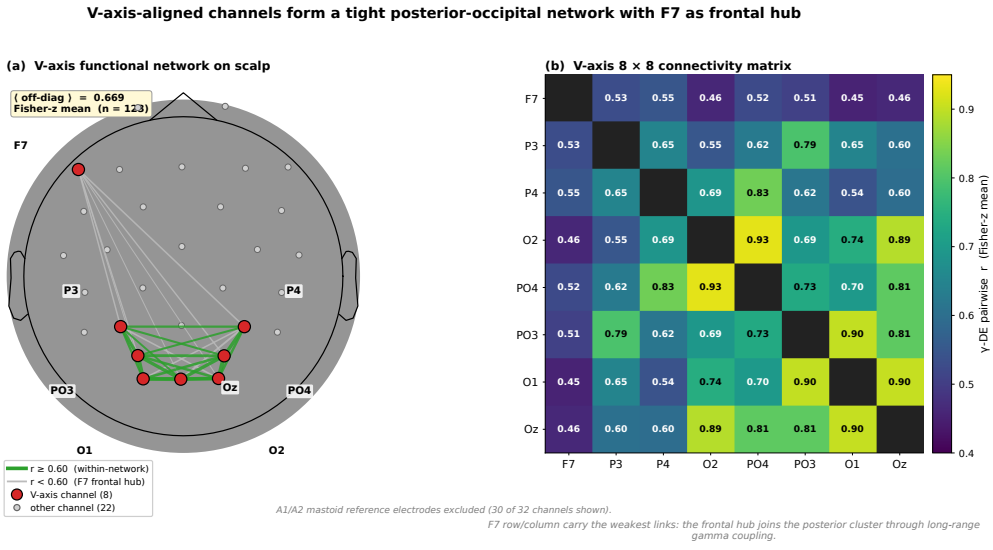


Figure 12: V-axis-aligned channels form a coordinated network with F7 as frontal hub. Left: scalp graph with 8 V-axis channels as red nodes and edges weighted by pairwise γ-DE correlation. Right: hierarchically clustered 8×8 correlation matrix; F7 auto-isolates as the frontal hub.

Mutual information vs. linear correlation. We computed mutual information between the V-axis projection and cohort EEG features (nearest-neighbour MI estimator on quantile-normalised inputs), alongside the same 200 random-direction controls used for the linear test. Observed MI 0.112; null mean 0.036; $p_{\mathrm{MI}} = 0.115$ (n.s.) vs. linear $p_r = 0.020$. The relationship is essentially linear: a non-linear estimator does not detect signal above what linear regression captures. We use ridge regressions throughout.

Theta-gamma coupling does not mediate the V-axis. Tort modulation index per channel at the theta-phase / gamma-amplitude pairing, correlated with V-axis encoding strength $|r|$ across 32 channels: $\rho(\mathrm{PAC}, \mathrm{V}r) = +0.082$, $p = 0.667$. Channels with strong theta-gamma CFC are not the channels that encode the V-axis, and vice versa — the V-axis encoding is dissociable from the phase-amplitude-coupling channel (Canolty & Knight, 2010).

S10 Saturation Regularity: Full Intervention Results

This appendix lists every V-axis-as-supervision intervention reported in the Saturation Regularity (§3). Table 10 gives per-recipe ΔBACC against the standard EMOD d=6 baseline, the paired-seed p-value, and a verdict tier; the two SatEnsemble-recipe replications below the line (against the stronger d=6 recipe with prototype regularization and augmentation, BACC ≈ 0.66) show larger decrements under that configuration. We also tested additional null variants (weight/bias initialisation, soft-KD temperatures, curriculum, and prior smoothing). Every significant effect in Table 10 is negative; even the largest positive point estimate (EMODSTYLE-stim, $+0.0066$, p=0.22) is not statistically significant. The range of loss families (KD, RSA, contrastive, PEFT, topographic, scaling, and multi-task) establishes the breadth of the tested interventions.

Table 10: Full V-axis-as-supervision intervention results. Under global Holm–Bonferroni correction across all 25 hypotheses, $10/25$ recipes produce a statistically significant negative (8 sig. negative + 2 catastrophic; marked *n.s. (Holm)* are recipes with uncorrected $p < 0.05$ that do not survive the global correction); 16 reach uncorrected $p < 0.05$. Four point estimates are positive, but none reaches $p < 0.05$. The 25-recipe screen uses the standard $d = 6$ receiving baseline. Follow-ups on the stronger SatEnsemble recipe appear below the line.

#	Family / variant	Δ	p	Verdict
1	Frontal-mask λ=0.5	−0.052	0.0015	sig. negative
2	Frontal-mask λ=0.1	−0.009	0.21	n.s.
3	FAA λ=0.5	−0.044	0.006	n.s. (Holm)
4	FAA λ=0.1	−0.018	0.063	borderline
5	Anger-weighted λ=0.5	−0.054	0.0003	sig. negative
6	Occipital λ=0.1	−0.022	0.007	n.s. (Holm)
7	Topo-optimal λ=0.1	−0.013	0.039	n.s. (Holm)
8	Topo-optimal λ=0.05	+0.0021	0.50	n.s. (NULL)
9	Procrustes λ=0.05	+0.0012	0.89	n.s. (NULL)
10	RSA λ=5.0	−0.093	$< 10^{-4}$	sig. negative
11	RSA λ=1.0	−0.057	$< 10^{-3}$	sig. negative
12	Distance-CE τ=5.0	−0.397	$< 10^{-4}$	catastrophic
13	Multi-V λ=0.5	−0.041	$< 10^{-3}$	sig. negative
14	EEG-AUX MSE λ=0.5	−0.043	$< 10^{-3}$	sig. negative
15	EEG-AUX MSE λ=0.1	−0.016	0.04	n.s. (Holm)
16	EMODSTYLE class λ=1.0	−0.010	0.30	n.s.
17	EMODSTYLE stim λ=0.5	+0.0066	0.22	n.s. (sweet spot)
18	PEFT fullhead λ=0.1	−0.009	0.069	borderline
19	PEFT LoRA λ=0.1	−0.010	0.035	n.s. (Holm)
20	PEFT IA3 λ=0.1	−0.008	0.14	n.s.
21	Pretrain-FT (frozen)	−0.401	$< 10^{-4}$	catastrophic
22	Pretrain-FT (unfrozen)	−0.190	$< 10^{-4}$	sig. negative
23	XEEG (FACED → SEED-V)	+0.0015	0.97	n.s. (zero)
24	SCALING (full data) λ=0.5	−0.045	< 0.01	n.s. (Holm)
25	Uncertainty multi-task	−0.067	$< 10^{-3}$	sig. negative
	SatEnsemble recipe + Topo λ=0.1	−0.029	0.004	sig. negative
	SatEnsemble recipe + EMODSTYLE λ=0.5	−0.026	0.004	sig. negative

S11 Saturation Regularity: Additional Intervention Analyses

Saturation transition table. Table 11 compares V-axis interventions under different receiving configurations, spanning BACC from 0.572 to 0.6581. The weaker configurations show small

nonsignificant effects, including positive point estimates. The stronger recipe shows significant decrements. These comparisons characterize configuration dependence; differences in architecture, recipe, and some loss weights prevent identifying a unique accuracy threshold or isolating training stage.

Table 11: V-axis intervention effects across receiving configurations. Effects are nonsignificant in the weaker configurations and significantly negative under the stronger recipe. Here *full* denotes the SatEnsemble recipe, including label smoothing ($\epsilon = 0.1$), prototype regularization (KD), and augmentation. Seed counts and loss weights are reported per row.

Recipe	n seeds	Vanilla / + V-axis	Δ
CBraMod baseline + Topo λ=0.05	5	0.572/0.567	-0.005 (n.s.)
CBraMod baseline + EMODSTYLE λ=0.5	4	0.572/0.577	$+0.005$ (n.s.)
EMOD d=3 vanilla + Topo λ=0.05	5	0.6194/0.6260	$+0.007$ (n.s.)
EMOD d=6 full + Topo (λ=0.1)	5	0.6581/0.6291	$\mathbf{-0.029}$ ($p < 0.01$)
EMOD d=6 full + EMODSTYLE (λ=0.5)	5	0.6581/0.6317	$\mathbf{-0.026}$ ($p < 0.01$)

Intervention effects across baseline strengths. Figure 13 plots intervention effects against receiving-baseline BACC. The comparisons are descriptive because objectives and configurations vary alongside baseline strength.

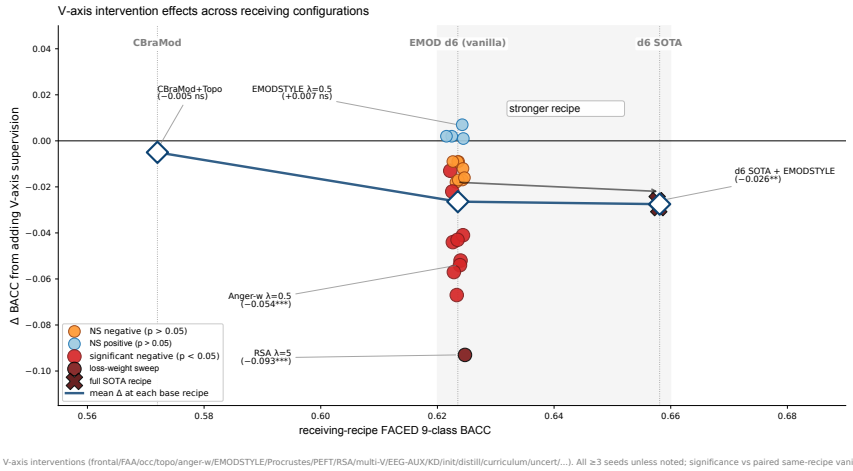


Figure 13: ΔBACC from V-axis supervision plotted against base-recipe BACC across $\sim$25 interventions. Different points involve different objectives and configurations; this plot does not identify a universal threshold. The label "SOTA" denotes the SatEnsemble recipe.

Anger-weighted supervision. Section 5.1 identifies an emotional-pole contrast across nine stimuli (Anger, Amusement, and Tenderness; cohort PO3/γ $r = 0.870$). Emphasizing these classes raises the analytical V-axis projection ceiling from $r = 0.478$ to $r = 0.714$. Yet the corresponding supervision reduces BACC by 0.054 ($p = 0.0003$; Table 10). Thus, a stronger analytical alignment target does not guarantee a useful training objective.

Alignment and residual encoding under supervision. For each of 8 V-axis-supervised variants (Topo at λ=0.05, Procrustes at λ=0.05, EMODSTYLE-stim at λ=0.5, full SatEnsemble recipe + Topo/EMODSTYLE at e=100/e=150), we measure the change in two quantities relative to the matched-recipe baseline: Δr_{PC1} (class-mean V-axis encoding) and $\Delta r_{\mathrm{resid}}$ (within-class V-axis residual encoding). V-axis training raises class-PC1 $|r|$ by $+0.01$ to $+0.36$ across the 8 variants, but leaves the within-class residual encoding essentially unchanged. Consistent with the accuracy behaviour (training accuracy at ceiling, test accuracy down; §3), strong-recipe decrements coexist with increased class-mean alignment and little change in residual encoding. (We do *not* claim the auxiliary gradient fails to reach the residual subspace: a direct gradient decomposition finds

a residual-subspace component comparable to a random direction, so the residual subspace is not geometrically unreachable.)

Adding V-axis-supervised checkpoints. We extend the ten-checkpoint SatEnsemble pool, trained without additional V-axis supervision, with 0–15 V-axis-trained Topo checkpoints and 0–10 V-axis-trained EMODSTYLE checkpoints, comparing 5-seed ensembles in each configuration. Adding 15 Topo V-axis ckpts to this pool gives ensemble BACC $\Delta = -0.0145$ ($p < 10^{-3}$); adding all 25 V-axis ckpts gives $\Delta = -0.0193$ ($p < 10^{-3}$). At every point on this curve, V-axis-trained ensembles are strictly worse than the matched ensemble without V-axis supervision. Thus, adding these V-axis-trained checkpoints does not improve the tested pools. This comparison alone does not locate their errors in a particular feature subspace.

S12 Saturation Beyond Auxiliary Injection: Alignment, Initialization, and Cross-Subject Transfer

The main-text screen (§3) injects the concept as an *auxiliary loss*. A natural question is whether the null is specific to that form of injection. We therefore tested three further channels through which the LLM-derived valence direction could, in principle, help a converged EEG encoder—frozen-feature alignment, initialization / pretraining, and cross-subject transfer—each against a content-matched control. The Saturation Regularity holds in all three (Figure 14).

(A) Frozen-feature alignment: the axis is a lossy subset of what the encoder already learned. Holding the converged d=6 encoder's penultimate features (128-d) frozen, we linearly align them to the real 5-d LLM/CLIP-text valence subspace (the top five principal components of the nine LLM/CLIP-text emotion-category centroids, of which the 1-D V-axis of §5.1 is the first) and probe emotion cross-subject, versus a random-orthonormal subspace of the same rank, an unsupervised PCA-5 bottleneck, and the full features. The real subspace beats random (0.391 vs. 0.278 BACC at full labels; $+0.11$ mean across label fractions, $p<0.001$)—the encoder genuinely contains a valence-aligned subspace—but it never beats the *full* features (0.391 vs. 0.438). The semantic subspace is a strict lossy subset of the representation the network already built from labels alone; alignment re-identifies a slice of it and adds nothing.

(B) LLM-aligned initialization does not help, and its harm grows with data. We pretrain the encoder (Stage-1, no emotion labels) to regress the per-clip valence subspace, then fine-tune on FACED at label fractions $\{0.1, 0.5, 1.0\}$, against (i) the standard EEG-SSL initialization (*scratch*) and (ii) an identical Stage-1 run on a random-orthonormal per-clip target (content-matched control). LLM-aligned initialization never significantly beats scratch ($+0.034$, p=0.12 at 10% labels) and significantly hurts with more data (-0.125, p=0.010 at 50%; -0.167, p=0.004 at 100%): a lossy low-rank pretext distorts the richer SSL initialization, and the distortion costs more the more that initialization is worth. The apparent "win" over the random-target control is an artifact—that control is *destructive* (0.194 vs. scratch 0.283 at 10%, p=0.0004), so beating it certifies nothing.

(C) The cohort correspondence is real but not a cross-subject lever. The EEG↔V-axis correspondence is population-level: with hand-crafted DE features a per-subject valence decoder is near zero (mean r=0.03), yet averaging the converged encoder's per-clip readout over subjects lifts the correlation from r=0.72 at a single subject to r=0.93 across the cohort (Figure 14a)—a matched-filter / attenuation-by-unreliability signature (Spearman, 1904) of a small sign-consistent shared component beneath large idiosyncratic geometry, the canonical structure of naturalistic affective neuroimaging (inter-subject correlation) (Hasson et al., 2004). Training already *closes* the individual gap for this shared dimension: the converged encoder's features carry the cohort valence direction inside every held-out subject (r=0.72; $0/62$ subjects negative)—which is exactly why the axis is not an exploitable transfer anchor. Registering held-out subjects along the shared valence axis before a leave-subjects-out 9-class probe (*sem*) matches neither generic per-subject alignment (*EA*, a Euclidean-alignment analog; He & Wu, 2020) nor a random-direction control: sem=0.712 vs. EA=0.714 (Δ=−0.002, p=0.54) vs. rand=0.710 (Δ=+0.003, p=0.22) BACC (Figure 14c); a supervised few-shot sign/gain calibration along the axis degrades accuracy further (-0.045, -0.021, -0.012 at 5, 10, 20 calibration trials). The transferable structure is generic subject-shift—captured by semantics-agnostic

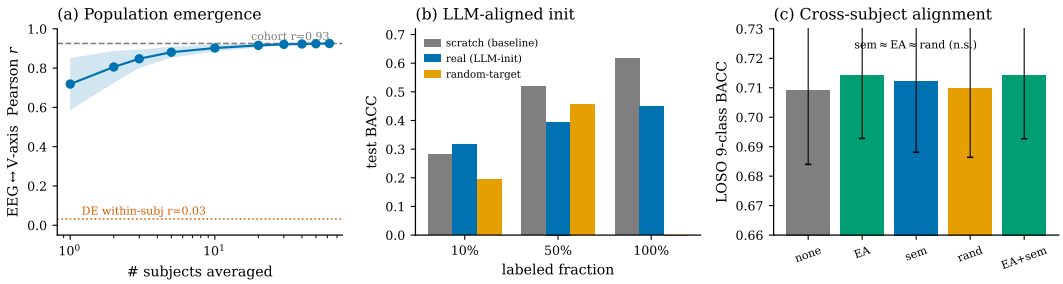


Figure 14: **The Saturation Regularity holds beyond auxiliary-loss injection. (a)** The EEG↔V-axis correlation is population-level: averaging the converged encoder's per-clip valence readout over more subjects lifts r from 0.72 (single subject) to the cohort value 0.93 (matched-filter emergence); a per-subject DE decoder is near zero (dotted). **(b)** LLM-aligned Stage-1 initialization (blue) never beats the EEG-SSL scratch baseline (grey) and hurts increasingly with more labels; the random-target control (orange) is destructive. **(c)** Registering held-out subjects along the shared valence axis (*sem*) does not beat generic subject-alignment (*EA*) or a random-direction control on leave-subjects-out 9-class decoding. Controlled falsification runs; see the Negative Results Catalogue (Appendix S19).

registration (Euclidean/Riemannian alignment, shared-response models, hyperalignment)—not the valence content. Saturation thus extends from in-distribution injection to the cross-subject setting: the converged encoder already contains the shared axis, so there is nothing left to inject.

S13 ResDiag: Additional Diagnostics and Controls

The 36-checkpoint cross-architecture correlation is pseudo-replicated. Pooling all 36 checkpoints from two backbones (CBraMod, EMOD) and six recipe variants, balanced accuracy correlates with class-PC1 V-axis encoding at $r{=}{+}0.885$ ($p_{\text{param}}{=}7.8{\times}10^{-13}$). This parametric p is an artifact of non-independence. The checkpoints form $\sim$ 2–3 tight architecture clusters—CBraMod ($\sim$ 0.57 BACC, mean class-PC1 $|r| = 0.21$), EMOD-vanilla d6 ($\sim$ 0.65, mean 0.67), EMOD d6 e150 ($\sim$ 0.66, mean 0.69)—so the correlation is driven by the between-architecture gap, not a within-architecture trend. *Within* each architecture it is not significant (CBraMod $r{=}0.83$, $p{=}0.08$; EMOD d6 e100 $r{=}0.04$, $p{=}0.90$; EMOD d6 e150 $r{=}{-}0.37$, $p{=}0.37$), and the effective number of independent checkpoints is $n_{\text{eff}}{\approx}1.2$.

LOO residual-encoding correlation: independence and per-length detail. The $r{=}0.74$ correlation between per-checkpoint within-class residual encoding and leave-one-out ensemble contribution (§5.2) is over 10 checkpoints that are 5 seeds $\times$ 2 training lengths (shared first 100 epochs). Within each length it is underpowered: $e{=}100$ $r{=}0.66$ ($n{=}5$, n.s.), $e{=}150$ $r{=}0.78$ ($n{=}5$, n.s.); a seed-clustered bootstrap (resampling seeds, keeping each seed's two checkpoints together) gives a 95% CI of $[0.50, 0.92]$ that excludes zero. We report it as a trend consistent with complementary predictions, accounting for the shared seeds.

Directional ablation: operator and orthogonality. The ablation applies $(I - vv^{\top})$, with v the unit within-class V-axis residual direction (fit on train/validation folds), to the penultimate features at inference. Because v is not orthogonal to the 9 class-mean subspace (mean projection fraction 0.75 vs 0.27 for a random direction), the raw drop (-0.0157) mixes within-class-residual removal with between-class-margin compression. Orthogonalising v against the class means and ablating only the within-class component still lowers BACC on $8/10$ checkpoints (mean -0.008, $z{\approx}3.5$ vs the matched-norm null)—about half the raw drop—so the within-class residual supports prediction. Per-checkpoint encoding strength does not predict the individual drop magnitude ($r{=}0.17$, $p{=}0.64$).

Random-direction null detail. We sample 1000 random Gaussian directions $w \in \mathbb{R}^{28}$ each matched in L_2 norm to v_{CLIP}, and recompute the 36-checkpoint correlation $\rho_w = \text{Pearson}(\text{BACC}, |\text{corr}(\text{PC}_1(H(m)), w)|)$. The distribution of ρ_w has mean -0.03 and standard deviation 0.62 (range $[-0.95, +0.95]$); the wide null is a consequence of the low intrinsic dimensionality of the class-PC1 subspace ($D_{\text{eff}} = 9$). The empirical V-axis correlation $+0.885$ sits at

only the 93.5^{th} percentile of this null (one-sided p=0.066, two-sided p=0.10)—it does not survive. Together, these controls limit the interpretation of the pooled association. ResDiag therefore focuses on similar class-mean encoding within the ten-checkpoint pool, controlled residual ablation, and the seed-clustered ensemble-contribution analysis.

Class-PC1 encoding across the pool. The ten checkpoints (five seeds at two training lengths) have class-PC1 V-axis encoding in the range $|r| \in [0.60, 0.77]$ (mean 0.69, standard deviation 0.05). This relatively narrow range motivates examining within-class information. It does not establish that all ensemble gains originate exclusively in an orthogonal residual subspace.

Directional ablation figure. Figure 15 reports the unorthogonalized directional ablation: all ten checkpoints decline (mean $\Delta\text{BACC} = -0.0157$, mean $z \approx 7.7$ against the matched-norm null). This intervention also affects class-mean structure. The orthogonalized control above isolates the within-class component and yields a smaller decline (mean -0.008, $8/10$ checkpoints, $z \approx 3.5$).

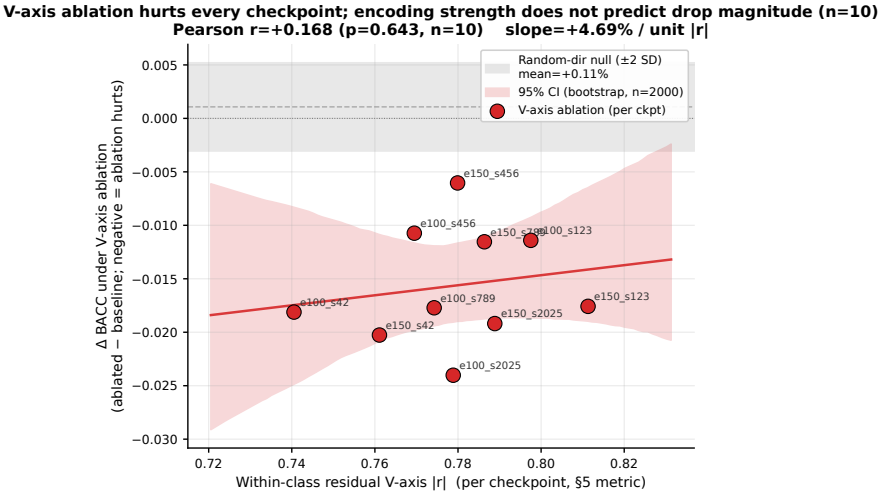


Figure 15: Unorthogonalized directional ablation of v_{resid} on all 10 SatEnsemble checkpoints, shown as a single scatter of per-checkpoint ablation ΔBACC (vertical axis) against within-class residual encoding $|r|$ (horizontal axis), with the matched-norm random-direction null band shaded. Two facts hold jointly. First, every checkpoint drops under ablation (population-mean $\overline{\Delta} = -0.0157$, well above the random-direction null on each checkpoint: mean $z \approx 7.7$, all $p < 0.001$). Second, residual-encoding strength does *not* predict the magnitude of the drop across the pool ($r = +0.168$, $p = 0.643$, $n = 10$): the effect is uniformly present but not graded by encoding strength. This direction also affects class-mean structure; the orthogonalized control isolates the within-class effect.

S14 SatEnsemble: Additional Comparisons and Ablations

Recipe cascade and residual diagnostics. Figure 16 traces the seven recipe steps from CBraMod's 0.572 to SatEnsemble at 0.6948. Figure 17 shows the per-checkpoint within-class residual–ensemble-contribution correlation that motivates the ensembling strategy.

Ensemble gains across benchmarks. Figure 18 compares observed ensemble gains on FACED, SEED-V, CIFAR-10, and MNIST.

The observed gain is larger on SEED-V (+0.034, single-model accuracy about 0.37) than on MNIST (+0.001, accuracy about 0.985). This pattern is consistent with less room for improvement at high accuracy. Since datasets and models differ, it does not establish a shared residual mechanism or imply that within-class variance vanishes near perfect prediction.

Exploratory wider-pool comparisons. Adding depth variants ($d \in \{8, 10\}$) did not exceed the ten-checkpoint pool in exploratory runs, which are not tabulated. A cross-architecture mixture

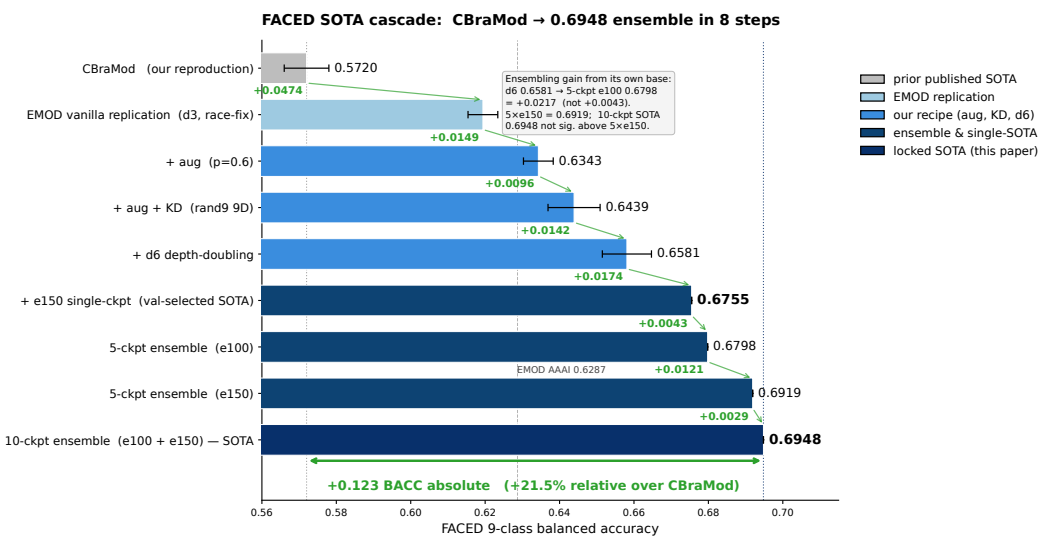


Figure 16: Recipe ablation cascade from the EMOD baseline at 0.6287 (previous FACED-9 SOTA (Chen et al., 2025)) to ten-checkpoint SatEnsemble at 0.6948 (+6.61 percentage points). Each bar is one recipe component or ensembling step; "KD" denotes prototype regularization. The CBraMod backbone (our reproduction 0.572; published 0.551) is shown for reference.

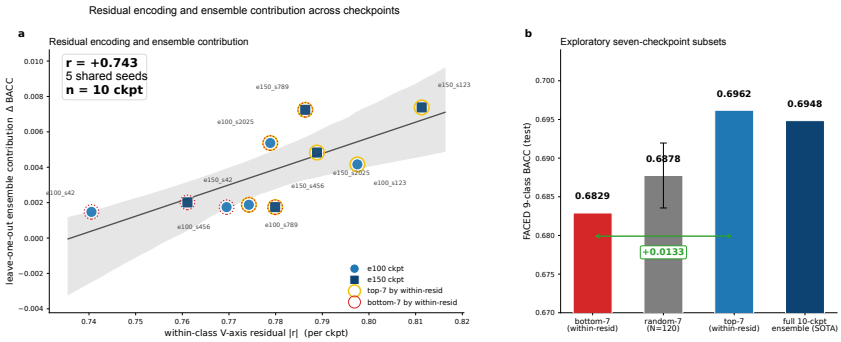


Figure 17: Residual encoding and ensemble contribution. Across ten checkpoints from five seeds, encoding correlates with leave-one-out contribution ($r = 0.74$; seed-clustered bootstrap 95% CI $[0.50, 0.92]$). The shaded fit band uses checkpoint resampling; the reported correlation interval clusters by seed. Panel (b) compares exploratory seven-checkpoint subsets. The association motivates prediction averaging without establishing a complete causal mechanism.

($d \in \{4, 6, 8, 10\}$) reaches 0.6791, close to the within-architecture five-seed result of 0.6798. These observations do not establish optimality of the selected pool or identify the cause of the limited gains. We retain the ten-checkpoint result as the evaluated SatEnsemble configuration.

Prediction-agreement structure. Mean single-model accuracy at e=100 (0.6581) and e=150 (0.6581) is identical. The pairwise prediction correlation within-e=100 is 0.694, within-e=150 0.679, cross-group 0.702 (higher = more agreement). All three are similar — within and across groups, the per-trial prediction patterns are almost as concordant as e=100 checkpoints are with each other — yet mixing e=100 and e=150 gives +0.015 over the e=100 ensemble. The gain is concentrated in a specific subset of cross-group disagreement: the 13.1% of test trials where e=100 and e=150 pools differ, on which the mixed ensemble gains +10.3 percentage points. This describes complementary predictions on particular trials. The mixed pool is not significantly above the e=150 ensemble (+0.0029, 95% CI $[-0.004, +0.010]$), so the comparison does not establish that mixing training lengths is essential.

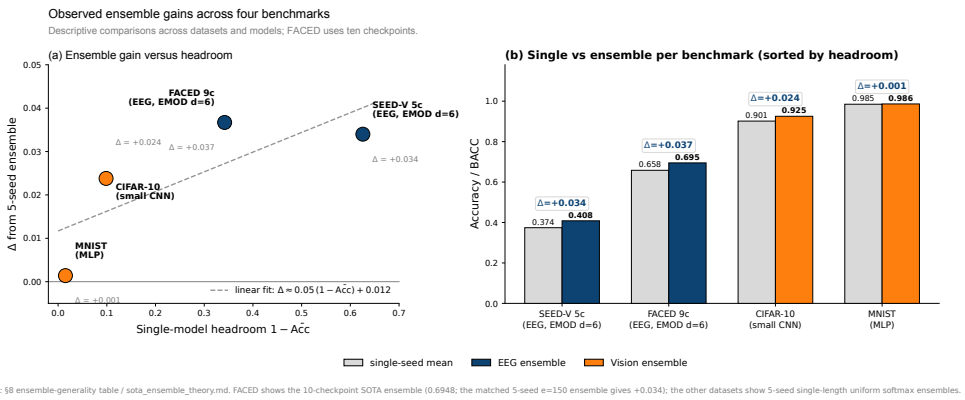


Figure 18: Ensemble gains across four benchmarks. (a) Observed gain versus error headroom $1-\bar{\text{Acc}}$, with a descriptive linear fit (slope ~ 0.05). (b) Single-seed mean and ensemble performance, ordered by headroom. FACED uses ten-checkpoint SatEnsemble. These cross-benchmark comparisons do not isolate residual variance as the cause of the gains.

Best single checkpoint without ensembling. For deployment scenarios where ensembling is impractical, our best single-checkpoint result is $\text{BACC} = 0.6755$ on the d=6 e=150 recipe, seed 789, val-selected. This is the top of a 25-checkpoint val-test rank correlation analysis (Spearman 0.825, $n = 25$) and exceeds our CBraMod reproduction by $+0.103$ and published EMOD by $+0.047$ BACC.

S15 EEG Model Training Details

Architecture. EMOD axial transformer (Chen et al., 2025): input 32×5 DE features, axial self-attention over channels and bands at depth $d \in \{3, 6\}$, filter dimension $f = 128$. The SatEnsemble recipe uses $d = 6$, $f = 128$.

Optimisation. AdamW with $\text{lr} = 10^{-3}$, $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay 10^{-2}. Cosine schedule with 5-epoch linear warmup. Batch size 128. Training length $e \in \{100, 150\}$ epochs. Cross-entropy is trained with label smoothing $\epsilon = 0.1$ (uniform over the nine classes), applied jointly with the prototype-regularization objective below. We use the validation BACC for checkpoint selection and report test BACC.

Augmentation. Per-trial Gaussian noise ($\sigma = 0.05 \cdot \text{std}$ of feature distribution), random channel dropout ($p = 0.15$), and random temporal masking (up to 5% of seconds). All augmentations are applied with $p = 0.6$ during training only.

Prototype regularization. The training configuration calls this class-mean knowledge distillation (KD). It uses a random orthonormal nine-dimensional teacher (`rand9`, with no LLM content), $\lambda_{\text{KD}} = 0.1$, and teacher temperature $\tau = 0.1$. Replacing language-derived prototypes with these random targets preserves BACC within the tested equivalence margin (TOST $\varepsilon = 0.01$, $p = 0.001$, $n = 20$; $\Delta = -0.001$). We refer to this component as prototype regularization throughout the paper.

SatEnsemble prediction averaging. 10 checkpoints split as 5 seeds at $e = 100$ + 5 seeds at $e = 150$. Seeds: $\{42, 123, 456, 789, 2025\}$ each. Ensemble prediction uses uniform averaging of the 10 checkpoints' softmax probabilities, followed by argmax.

S16 Statistical Methods

Bootstrap CIs. Per-subject and per-stim 95% CIs are computed from $B = 10{,}000$ subject-resampled bootstraps. For cohort correlations, we report the bootstrap median, 2.5% / 97.5% percentile interval, and a Fisher-z p-value for $r = 0$.

Random-direction null. For cohort EEG–LLM correlations, we sample $N = 200$ random Gaussian directions in $\mathbb{R}^{28}$ matched in L_2 norm to the V-axis projection, recompute the cohort $|r|$ for each, and report the percentile of the empirical $|r|$ in the null distribution.

Behavioural-valence permutation control. The reported nested analysis permutes the behavioural-valence target across the 28 stimuli, using the same permutation for every subject. Each of $5{,}000$ permutations re-runs per-subject ridge fitting, leave-one-stimulus-out prediction, cohort aggregation, and correlation. The reported result is $0/5000$ exceedances ($p = 1/5001$ with the standard add-one correction). This subject-first analysis uses behavioural targets and differs from the cohort-feature CLIP-text readout displayed in Figure 3. Figure 3b instead displays a fixed-prediction target-shuffle reference; it is distinct from this nested null.

Cross-architecture null. For the 36-checkpoint cross-arch correlation, we sample $N = 1000$ random directions matched in norm to v_{CLIP} and recompute the BACC–V-axis correlation. The null is distributed broadly ($\sigma = 0.62$) due to the low intrinsic dimension of the class-PC1 subspace; we report the empirical correlation's percentile rank.

Per-checkpoint LOO contribution. For the ten-checkpoint SatEnsemble pool, we compute $\Delta_k = \mathrm{BACC}(\mathrm{ens}) - \mathrm{BACC}(\mathrm{ens} \setminus k)$. Pearson correlations use the ten checkpoint values. For the residual-encoding correlation, the confidence interval resamples the five seeds and keeps each seed's two training lengths together (Appendix S13).

V-axis intervention comparisons. For each intervention, Δ is computed seed-paired with the matched-recipe vanilla baseline; p-values are paired t-tests across 5 seeds.

S17 Discussion Extras and Future Work

Broader impacts. Improved EEG emotion classifiers have clear positive applications in mental-health monitoring, affective brain–computer interfaces, and clinical phenotyping of affect disorders. The same techniques pose risks of affective inference without consent in surveillance, hiring, workplace, education, and advertising contexts. The V-axis extraction protocol applied to EEG could in principle reduce the calibration-data burden for affective inference systems — a feature that cuts both ways, since lower calibration cost lowers the deployment threshold for both clinical and surveillance settings. We do not release human EEG data; we release training configs, feature-side and analysis code, and extracted V-axis directions. Recommended deployment safeguards: informed consent, opt-in only, on-device inference, no longitudinal storage of raw EEG outside research contexts, and IRB review for any non-clinical inference application. Dual-use risk for the V-axis extraction itself is low: it operates on text and image features already public, and the EEG mapping requires per-cohort EEG access that is governed by the original dataset licences.

Implications for affective neuroscience. For video-evoked emotion on FACED, single-channel correlations are stronger over posterior scalp regions ($|r|_{\mathrm{occipital}} = 0.21$ versus $|r|_{\mathrm{frontal}} = 0.16$). These scalp measurements do not establish a cortical source. Frontal-alpha asymmetry correlations have the hypothesised direction but small magnitude (F8–F7: $r = 0.0155$; Fp2–Fp1: $r = 0.0116$), which does not by itself establish a robust asymmetry effect. The nine-clip anger-versus-warm-positive contrast describes the single-channel pattern; the multi-feature readout retains its correlation after those categories are excluded (Figure 3).

Per-subject V-axis adaptation does not transfer. The per-subject best-channel oracle reaches $\overline{|r|} = 0.62$ (a $+0.60$ absolute headroom over the cohort-fixed top-8 of $|r| = 0.02$). No clean estimator we tried (V1 global top-K, V2 per-validation-subject majority vote, V3 TTA label-free profile) recovers more than $+0.001$ over the cohort top-K. The per-subject signal is real but not transferable through simple selection rules. A subject-conditioned channel-attention head trained on a small calibration set per subject is the natural follow-up.

Future work. Three lines: (i) per-subject V-axis adaptation closing the $|r| = 0.62$ oracle gap via subject-conditioned channel-attention; (ii) replicating the Saturation Regularity for other concept

directions (Arditi et al., 2024) (arousal, dominance, formality, refusal) on their respective benchmarks; (iii) using the within-class residual $|r|$ as an online model-selection signal during training rather than as a post-hoc diagnostic.

S18 FACED Test Confusion Matrices

Figure 19 contrasts the best-single-checkpoint and ten-checkpoint SatEnsemble confusion matrices on the FACED 9-class test split.

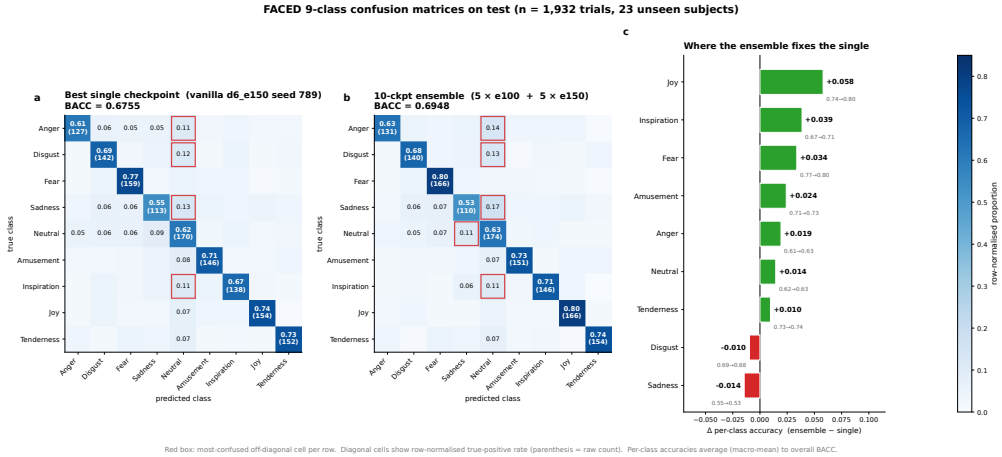


Figure 19: FACED 9-class test confusion. Left: best single checkpoint (BACC 0.6755). Right: ten-checkpoint SatEnsemble (BACC 0.6948). The ensemble's gain is concentrated on the off-diagonal cells where single-seed disagreement is highest, illustrating complementary predictions (Section 5.3).

S19 Negative Results Catalogue

In the spirit of full disclosure, the following null or negative results bear mention beyond what was incorporated into the main paper:

1. **Theta-gamma cross-frequency coupling does not mediate the V-axis** (Appendix S9): Tort modulation index $\rho(\mathrm{PAC}, \mathrm{V}r) = +0.082$, $p = 0.667$.
2. **Mutual-information control is not significant** (Appendix S9): MI $p = 0.115$ vs. linear $p = 0.020$. This test does not detect additional nonlinear dependence.
3. **Cross-architecture V-axis convergence is pseudo-replicated** (Section 5.2): the 36-checkpoint $r = +0.885$ is driven by $\sim$2–3 architecture clusters ($n_{\mathrm{eff}} \approx 1.2$), is not significant within any single architecture, and sits at only the 93.5^{th} percentile of a 1000-direction null ($p_{\mathrm{one}} = 0.066$, $p_{\mathrm{two}} = 0.10$). Residual encoding correlates with ensemble contribution at $r = 0.74$ (seed-clustered 95% CI $[0.50, 0.92]$); this is the association retained by ResDiag.
4. **Per-subject V-axis adaptation does not transfer.** The per-subject best-channel oracle reaches $|r| = 0.62$, but no clean estimator transfers more than $+0.001$ over the cohort top-K.
5. **Cross-architecture ensembling gives no detectable gain.** Mixing $d \in \{4, 6, 8, 10\}$ checkpoints gives within-arch 0.6798 vs. cross-arch 0.6791 (n.s.).
6. **Wider ensemble pools do not help.** Adding $d \in \{8, 10\}$ depth variants beyond the 10-checkpoint d=6 pool did not exceed 0.6948 in exploratory runs (Appendix S14).
7. **Stimulus-aggregation re-ranking does not help.** Operating at the 28-stimulus level instead of trial level gives BACC=1.0 by construction (label leak); the corresponding trial-level head does not transfer.
8. **Test-time augmentation at K=5 does not improve the ensemble** ($0.6753 \rightarrow 0.6720$). The tested augmentation does not improve prediction; this comparison does not identify its effect on residual structure.

9. **Toxicity: recipe scope.** The 9-prompt PCA recipe attains AUC 0.59 on Jigsaw toxic comments, below every one of the 20 in-scope concepts (each AUC> 0.65) in our library (Appendix S4); toxicity appears to be encoded in a higher-rank subspace than the late-layer PC1.
10. **Arousal asymmetry: vision yes, text and brain no.** OASIS arousal $r = 0.803$ in vision; $\leq 13\%$ NRC arousal recovery in text; $r \in [0.18, 0.41]$ brain alignment across LLMs (Appendix S6).
11. **Prototype regularization does not require language-derived targets.** Replacing the 9-D LLM-derived class prototypes with random orthonormal 9-D directions leaves BACC statistically equivalent (Δ=-0.001, paired p=0.72, n=20; TOST ε=0.01, p=0.001). The prototype-regularization benefit does not require language-derived semantics within the tested tolerance.
12. **LLM-aligned initialization gives no gain in the tested settings.** Beyond the auxiliary-loss screen, injecting the valence direction as frozen-feature alignment or as Stage-1 initialization also fails: a real 5-d semantic subspace is a lossy subset of the encoder's own features (never beating the full 128-d), and LLM-aligned pretraining never beats the EEG-SSL baseline and hurts more with more labels (-0.167 at full labels) (Appendix S12).
13. **The valence axis is not a cross-subject lever.** Registering held-out subjects along the shared axis does not beat generic Euclidean-style subject-alignment or a random direction on leave-subjects-out 9-class decoding ($\mathrm{sem}{-}\mathrm{EA} = -0.002$, p=0.54), providing a complementary null result in the cross-subject setting (Appendix S12).

S20 Reproducibility

Release commitment. Code, configs, model checkpoints, and figure-generation scripts will be released upon acceptance (no submission-time anonymous URL is provided). The release will include: V-axis extraction scripts for the 14 LLMs in the per-LLM table; the EMOD d=6 training pipeline with the exact augmentation, prototype regularization (KD), and ensemble-construction code; the ten SatEnsemble checkpoints (0.6948 BACC); and the figure-generation scripts for every landmark and neuro figure in the paper.

Run-level provenance. Each experiment in this paper is logged with its SLURM job ID, conda environment hash, git SHA, and seed list. The ten-checkpoint SatEnsemble result is reproducible from a single SLURM array job (`slurm_d6_e100_e150_5seeds.sh`, $\sim$10 GPU-hours per seed on V100).

S21 Resource Estimate

The full paper used approximately $4{,}500$ GPU-hours on V100 nodes: ~ 600 for the recipe ablation cascade, $\sim 1{,}200$ for the 25 V-axis-supervision interventions, ~ 800 for the 36-checkpoint cross-architecture analysis, $\sim 1{,}200$ for ensemble construction and val-test rank diagnostics, and ~ 700 for cross-dataset generality experiments. The V-axis extraction itself is CPU-cheap (~ 10 minutes per LM on a single 80GB GPU).