SAGE: SUBJECT-ANCHORED AU-GUIDED EXPRESSION FOR PAIN ESTIMATION FROM VIDEO

Anonymous authors
Paper under double-blind review

ABSTRACT

Pain expression is anatomically localized to a few action units (AUs) and varies strongly across individuals. We propose SAGE (Subject-Anchored AU-Guided Expression), which makes both properties explicit on top of a frozen self-supervised video transformer (MARLIN) and trains only a $\sim$2M-parameter ordinal head. *AU-guided token selection* routes patch tokens through the six PSPI AU regions located by facial landmarks, keeping $\sim$12% of the tokens. *Subject calibration* (SubjMean) centers each query on the mean of a few *unlabeled* clips of the same subject. On BioVid leave-one-subject-out, SAGE reaches 5-class QWK 0.399 $[0.35, 0.45]$ (0.342 without calibration) versus 0.199 for a matched frozen baseline (PainFormer) under the same protocol. Calibration accounts for most of the gain and improves all six frozen encoders we test. It is non-parametric, and its gain rises steeply up to about five unlabeled clips per subject and only slowly beyond. Token selection improves over holistic pooling on two face backbones and outperforms both an arbitrary partition and direct AU intensities. Region attention shows that the calibrated model draws on all six AU regions, whereas the uncalibrated one collapses onto a single region. The frozen encoder transfers by linear probe to AI4Pain and zero-shot to a clinical QST pilot, exceeding PainFormer on both.

1 INTRODUCTION

Pain is subjective, yet its assessment is fundamental to clinical care (Merskey & Bogduk, 1994). For patients who cannot self-report—neonates, cognitively impaired or sedated patients—clinicians rely on facial expression (Herr et al., 2011), and their judgments are systematically biased by patient demographics and by the observer's own characteristics (Craig, 1984; Trawalter et al., 2012; Kappesser et al., 2006). Automated, objective estimation from video is therefore a long-standing goal (Werner et al., 2022).

Two properties of pain expression shape the problem. First, it is *anatomically localized*: the Prkachin–Solomon Pain Intensity (PSPI) metric (Prkachin & Solomon, 2008) defines pain expression as a composite of six FACS action units (Ekman & Friesen, 1978)—AU4 (brow lowerer), AU6/7 (cheek raiser / lid tightener), AU9/10 (nose wrinkler / upper-lip raiser), and AU43 (eye closure). Second, it is *subject-specific*: individuals differ substantially in facial expressiveness even under stimuli calibrated to their own pain threshold (Prkachin & Solomon, 2008; Walter et al., 2013), which is why clinical guidelines for non-verbal patients require establishing each patient's baseline before reading deviations as pain (Herr et al., 2011). Video methods for pain estimation nevertheless operate on holistic frame or clip features, train end-to-end on small pain corpora, and apply one population-level decision boundary to all subjects (Werner et al., 2022).

We ask how much of the pain signal in a *frozen* face-video representation can be recovered by making these two properties explicit, without training the backbone. We propose SAGE (Subject-Anchored AU-Guided Expression; Fig. 1), which uses two operations on top of frozen MARLIN (Cai et al., 2023) patch tokens and trains only a $\sim$2M-parameter ordinal head:

1. **Subject calibration from unlabeled clips.** Each query is centered on the mean of K **unlabeled** clips of the same subject (query excluded, no test labels; *SubjMean*). On BioVid LOSO it lifts 5-class QWK from 0.342 to 0.399 and improves every one of six frozen encoders we test; the gain is specific to genuine same-subject reference; the gain rises steeply up to about five unlabeled

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.

clips per subject and only slowly beyond, which we explain through the anchor estimation error; and the plain mean is as good as a learned reference encoder.

2. **AU-guided token selection.** A *zero-parameter*, landmark-driven routing of patch tokens into the six PSPI AU regions. Using $\sim$12% of the tokens, it improves over holistic pooling on two frozen face backbones (a video and an image MAE), outperforms an arbitrary partition and direct AU-intensity inputs, and gives the model a per-AU structure: calibration operates per AU, and a per-AU attention readout (Fig. 2) shows that the calibrated head draws on all six regions, whereas the uncalibrated head collapses onto one.

Together they are competitive with prior end-to-end video methods on BioVid. Against a matched open-source frozen-encoder baseline (PainFormer), SAGE leads on 5-class QWK and on two cross-dataset transfers without backbone fine-tuning, while binary accuracy is comparable to the published PainFormer figure. We report confidence intervals throughout and analyze when calibration helps, across encoders, reference sources, and reference-set sizes. The consistency of that gain suggests that the field's default of training end-to-end with one population-level decision boundary leaves a cheap, reliable improvement unused.

2 Related work

Facial pain estimation. Early systems used hand-crafted geometric and appearance descriptors (Werner et al., 2014; Kaltwang et al., 2012); deep methods moved to CNN–RNN pipelines (Zhou et al., 2016; Rodriguez et al., 2022; Huang et al., 2022b), feature fusion (Yang et al., 2018), non-contact physiological cues from the same video (Yang et al., 2021; Huang et al., 2022a), and transformers over frame features (Xu & Liu, 2021; Gkikas & Tsiknakis, 2023). Closest to our setting is PainFormer (Gkikas et al., 2025), a supervised multi-task face foundation model whose released embedder is kept frozen while a downstream head is trained; we use it as the matched RGB-only baseline. These methods share three traits our work departs from: holistic frame- or clip-level features, end-to-end training on small corpora such as BioVid (Walter et al., 2013) (which transfers poorly across clinical and experimental pain (Prajod et al., 2022; 2024)), and a single population-level decision boundary.

Region-guided AU representation. Landmark- or attention-defined AU regions are well established for AU *detection* on still images (Zhao et al., 2016; Jacob & Stenger, 2021), recently with self-supervised pretraining (Zhang et al., 2024; Wei et al., 2026). We do not predict AUs: we use PSPI AU regions to *select* tokens from a frozen video transformer. This exploits the dense patch grid of masked video autoencoders (Tong et al., 2022), which prior pain work has not used; our backbone MARLIN (Cai et al., 2023) is pretrained on YouTube Faces (Wolf et al., 2011) without pain or affect labels.

Subject normalization and reference-based inference. Per-subject mean normalization is a standard way to remove inter-subject variability in physiological signals (Zhao et al., 2021). In pain estimation it has appeared only as a preprocessing step or inside ordinal metric learning (Ting et al., 2021), not as calibration at inference time. Our anchor is a set function over unlabeled same-subject clips (Zaheer et al., 2017), a subject-level analogue of class prototypes (Snell et al., 2017). Test-time adaptation also adjusts a deployed model from unlabeled test data, via entropy minimization (Wang et al., 2021), feature-statistic alignment (Lin et al., 2023), re-estimated normalization statistics (Schneider et al., 2020), or pseudo-prototypes (Iwasawa & Matsuo, 2021). SAGE differs in three ways: the shift is a per-subject offset in a frozen representation; subject identity is known rather than inferred; and the anchor is simply the mean feature of a few clips, computed after the backbone, so nothing inside the backbone changes. Our experiments show that this calibrated representation is well suited to pain estimation and characterize when it helps, across encoders, reference sources, and reference-set sizes.

3 The SAGE pipeline

SAGE applies AU-guided token selection (Sec. 3.1) and reference-set subject calibration (Sec. 3.2) to frozen MARLIN tokens, and trains a small CORAL ordinal head (Sec. 3.3). Figure 1 summarizes the pipeline.

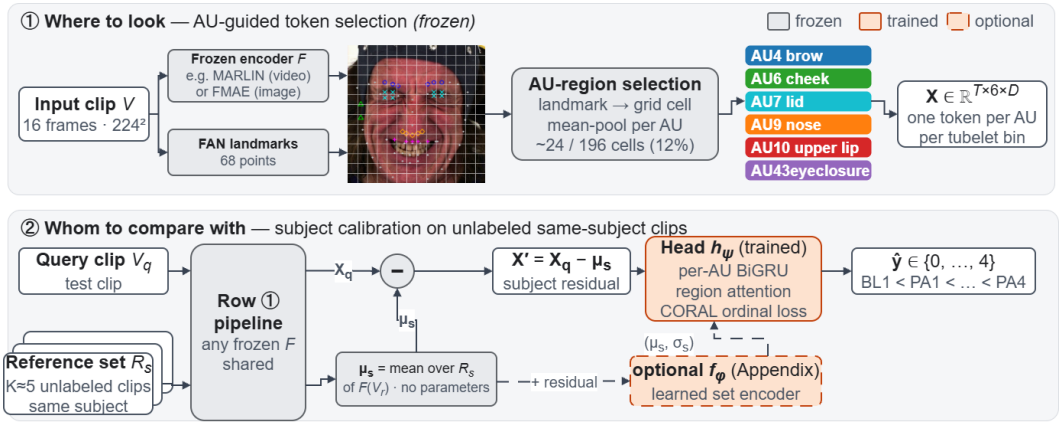

Figure 1: **SAGE.** Row 1 (frozen): patch tokens of a frozen face encoder (MARLIN by default; FMAE in Sec. 4.3) are routed through the six PSPI AU regions located by FAN landmarks and mean-pooled, giving one token per AU per tubelet bin ($\mathbf{X} \in \mathbb{R}^{T\times 6\times D}$). Row 2: the query is centered on the mean of K unlabeled same-subject reference clips (SubjMean, no parameters) and a small ordinal head is trained on the residual. An optional learned encoder f_ϕ (Appendix C) adds a residual to $\boldsymbol{\mu}_s$. Full architecture with tensor shapes: Fig. 5.

3.1 AU-GUIDED TOKEN SELECTION

Pain expression is localized to the six PSPI AUs (Fig. 3, Appendix A). Token selection makes this prior explicit at the representation level by exposing only the anatomically relevant tokens to the head. It requires only a frozen encoder that exposes a spatial patch grid; we use MARLIN (Cai et al., 2023) as the default instance and an image encoder (FMAE (Ning et al., 2024)) in Sec. 4.3. MARLIN maps a 16-frame clip at 224×224 to a patch grid $G \in \mathbb{R}^{T\times H\times W\times D}$ with $(T, H, W, D) = (8, 14, 14, 768)$, so each $G(t, i, j, \cdot)$ embeds one localized spatiotemporal region. For each AU $a \in \mathcal{A} = \{4, 6, 7, 9, 10, 43\}$ we fix a subset $\mathcal{L}_a$ of the 68 FAN (Bulat & Tzimiropoulos, 2017) landmark indices lying in that AU's anatomical extent (Appendix A); with $(y_\ell^{(t)}, x_\ell^{(t)})$ the pixel coordinates of landmark ℓ, the per-AU token is the mean of the grid cells containing its landmarks,

$$\mathbf{X}(t, a, \cdot) \;=\; \frac{1}{|\mathcal{L}_a|} \sum_{\ell\in\mathcal{L}_a} G\Big(t,\ \lfloor y_\ell^{(t)}/16 \rfloor,\ \lfloor x_\ell^{(t)}/16 \rfloor,\ \cdot\Big) \;\in\; \mathbb{R}^D, \tag{1}$$

giving $\mathbf{X} \in \mathbb{R}^{T\times A\times D}$ with A=6. The six regions cover about 24 of the 196 grid cells per frame (∼12% of tokens); the lookup has no learned parameters. For image encoders the same lookup is applied to each frame's grid. Each BioVid video is processed as S=4 non-overlapping sub-clips whose outputs are concatenated in time (T_{v}=32; Appendix B).

3.2 SUBJECT CALIBRATION FROM UNLABELED REFERENCE CLIPS

Identical pain intensities elicit visibly different faces, and each subject's neutral face carries an idiosyncratic bias (Werner et al., 2022) (Fig. 3). Clinical practice reads a patient relative to their own baseline; subject calibration does the same at the representation level.

Protocol. Every query clip is accompanied by a *reference set* R_s: a subset of K unlabeled clips drawn from the same subject's other clips, never including the query itself. K is a free choice: it can be as small as a handful of enrollment clips or as large as all of the subject's remaining clips. Clips in R_s carry no pain labels, so no test label reaches the prediction, and the test subject never appears in training (LOSO). The reference need not be pain-free: what matters is that it samples the subject's expression range, and a neutral-only reference is not better (Appendix G). Unless stated otherwise, tables use the full remaining set (sizes in Appendix B); Sec. 4.3 varies K and shows that the gain saturates around K≈5.

SubjMean (default). For each $r \in R_s$ the token-selection output is mean-pooled over time to $\mathbf{x}_r \in \mathbb{R}^{A \times D}$; the anchor is their mean and the head receives the residual,

$$\boldsymbol{\mu}_s = \frac{1}{|R_s|} \sum_{r \in R_s} \mathbf{x}_r, \qquad \mathbf{X}'(t, a, \cdot) = \mathbf{X}(t, a, \cdot) - \boldsymbol{\mu}_s(a, \cdot). \tag{2}$$

A brow that is furrowed at rest inflates the AU4 channel in every clip of that subject; subtracting $\boldsymbol{\mu}_s$ removes this offset, so only the change from the subject's own baseline reaches the head.

Parameterized variant. A learned permutation-invariant set encoder f_ϕ can replace the mean; it is zero-initialized to reduce to SubjMean at step 0 and adds a scale channel $\boldsymbol{\sigma}_s$ (Appendix C). It is the stronger choice for binary detection (Table 1) and serves as the substrate for the reference-source control (Sec. 4.3); SubjMean is used everywhere else.

3.3 Classifier head and inference modes

Pain levels are ordered but not equally spaced, so the head h_ψ predicts $K-1$ cumulative-logit boundaries with the CORAL rank-consistent loss (Cao et al., 2020) (K=5; a single sigmoid for binary tasks). Each of the A=6 AU branches is encoded in time by a small bidirectional GRU (Cho et al., 2014) and mean-pooled; learned region attention fuses the six branch vectors, and a linear layer emits the logits. Only h_ψ (and f_ϕ when used) is trained; MARLIN receives no gradients. At inference the prediction counts how many logits exceed per-fold recalibrated cutoffs. Details, parameter counts, and pseudocode are in Appendix B.

SAGE has two deployment modes. *Same-subject*: R_s from the test subject's own clips (all main results). *Reference-free*: $\boldsymbol{\mu}_s$=$\mathbf{0}$, a single-clip classifier for deployments without enrollment.

4 Experiments

We evaluate SAGE on BioVid (in-domain LOSO) and on cross-dataset transfer to AI4Pain (linear probe; Appendix H) and QST-Pilot (true zero-shot), a clinical pilot whose protocol mirrors BioVid Part A (Appendix I). The primary external baseline is PainFormer (Gkikas et al., 2025), a matched open-source frozen RGB embedder with a trained pain head; we also report reference-free and direct AU/PSPI controls. Supplementary diagnostics (PSPI linear-probe validation, reference-source controls, composition and label balance) and a full efficiency breakdown are in the appendix.

4.1 Setup

We evaluate on BioVid Heat Pain Part A (Walter et al., 2013), the standard benchmark for automated facial pain estimation: 87 subjects, 5.5 s clips at five intensity levels (BL1 baseline and PA1–PA4, each tuned to the subject's individual pain threshold), 20 clips per level per subject (8,700 clips). All LOSO results use all 87 subjects. Five-level LOSO on BioVid is a hard benchmark: chance accuracy is 20%; adjacent levels differ by small temperature steps inside each subject's individually calibrated range; and a documented subset of subjects shows little or no visible facial response to the stimulus (Werner et al., 2014), which bounds any facial-video method. Published video-only results therefore stay far from ceiling, and under our protocol the matched frozen baseline reaches 5-class QWK 0.199 (Table 1); the extreme-contrast binary task, by comparison, is close to saturation. We extract per-clip AU intensities with OpenFace (Baltrušaitis et al., 2018) and 68-point facial landmarks with FAN.

We use strict LOSO with a deterministic $\sim$10% validation split per fold, used for early stopping and per-fold cutoff recalibration. We report accuracy on all tasks, with macro-F1 and AUC additionally on binary tasks. For the 5-class ordinal task we emphasize Quadratic Weighted Kappa (QWK) (Cohen, 1968), which weights mistakes by their distance on the pain scale, and report fold-mean QWK with percentile bootstrap 95% confidence intervals over LOSO folds. For cross-dataset transfer (Sec. 4.2) we emphasize per-subject Spearman correlation, which is threshold-free and scale-invariant across BioVid (5 levels) and AI4Pain (3 levels). All models are trained on 2$\times$ NVIDIA RTX 4090 (24 GB) with PyTorch 2.5.1.

Table 1: BioVid LOSO comparison against the matched frozen RGB baseline PainFormer (Gkikas et al., 2025). SAGE headline is SubjMean; the parameterized row is an optional large-reference extension. 5-class QWK is fold mean with bootstrap 95% CI over LOSO folds.

	5-class		BL1-vs-PA4		
Method	Acc	QWK [95% CI]	Acc	F1	AUC
Chance (balanced classes)	0.200	0	0.500	0.500	0.500
Without subject calibration:					
PainFormer (Gkikas et al., 2025)[a]	–	0.199 [0.15, 0.25]	**0.7629**	–	–
SAGE, reference-free (ours)	0.286	0.342 [0.29, 0.40]	0.728	0.725	0.781
With subject calibration:					
SAGE (SubjMean)	0.302	**0.399** [0.35, 0.45]	0.697	0.696	0.747
SAGE (parameterized)	**0.304**	0.396 [0.34, 0.46]	0.750	**0.749**	**0.822**

[a]Binary Acc is the published LOSO figure. 5-class QWK is our evaluation of the released frozen embedder + trained Embedding-Mixer under the same LOSO protocol as SAGE. PainFormer is RGB video only (same modality as SAGE).

4.2 Results and comparison with prior work on BioVid and cross-dataset transfer

Table 1 reports in-domain BioVid LOSO results. On 5-class ordinal QWK, SubjMean already exceeds both the reference-free model and our matched PainFormer evaluation (confidence intervals do not overlap the PainFormer band); the parameterized extension does not improve on it for the ordinal task (0.396 vs. 0.399, overlapping CIs), so the non-parametric default is sufficient there. On BL1-vs-PA4, PainFormer's published accuracy is slightly higher than SAGE, while SAGE (parameterized) leads on F1 and AUC. SubjMean itself lowers binary accuracy ($0.728 \rightarrow 0.697$). This is expected: with only the two extreme levels, a population-level threshold already separates them, so subtracting a subject anchor adds estimation noise without adding separation; centering pays off where individual baselines blur adjacent levels, i.e. on the five-level ordinal task. The binary task is also largely saturated across strong frozen encoders, and it is the one setting where the learned encoder helps. Comparing either calibrated row to the reference-free row shows that calibration contributes the dominant share of the 5-class lift.

Per-fold variability and subject difficulty. Fold-level 5-class QWK has a wide spread (std ≈ 0.22 for SAGE parameterized). We attribute this primarily to BioVid subject difficulty rather than training instability. Per-subject SAGE QWK correlates strongly with an independent model (PainFormer; Spearman $\rho = 0.79$, $p < 10^{-14}$): the same subjects are hard for both. Splitting subjects by this *independent* (PainFormer-defined) difficulty into equal halves gives QWK 0.55 ± 0.13 on the easier half versus 0.23 ± 0.17 on the harder half (parameterized variant; mean $\pm$ std over subjects). Within the easier half, SAGE is relatively stable, and about half of the total per-subject variance lies *between* the easy and hard groups. The hard tail aligns with the documented muted-expression subpopulation (little visible change from baseline to pain), which caps what any facial-video method can extract; aggregate 5-class scores are diluted by this group.

Cross-dataset transfer. We test transfer under two target protocols (Table 2). On AI4Pain (Fernandez-Rojas et al., 2024) (Appendix H) the published protocol freezes the encoder and trains only a small head on target clips, with no reference-set inference; our AU-selected features with a 387-parameter linear head match the 2-layer LSTM competitor and exceed the published PainFormer video accuracy, and the positive per-subject Spearman shows that pain ordering transfers, not only class boundaries. This is linear-probe transfer, not zero-shot. On QST-Pilot (Appendix I) no target-domain training is used at all. Here the gap to PainFormer is large even without calibration: reference-free SAGE reaches per-subject Spearman 0.426 against 0.232 for the PainFormer embedder, and SubjMean raises it to 0.488. A plausible reason is that AU-guided selection keeps only the anatomically pain-relevant regions of a self-supervised representation and discards much of the identity-bearing appearance, whereas PainFormer's embedder is trained with identity supervision; separating this from the backbone difference requires a holistic-MARLIN transfer run, which we have not performed.

Table 2: Cross-dataset transfer against PainFormer. **AI4Pain:** linear-probe transfer (frozen BioVid encoder, 387-parameter head trained on the AI4Pain train split; no reference-set inference under the published protocol). **QST-Pilot:** true zero-shot (no target training); SubjMean uses unlabeled same-subject references at test time.

Method	QST-Pilot Spearman	AI4Pain Acc	AI4Pain Spearman
Without subject calibration:			
PainFormer (Gkikas et al., 2025)[a]	0.232	0.5139	–
LSTM (Prajod et al., 2024)	–	0.60	–
SAGE, reference-free (ours)	0.426	0.605	**+0.315**
With subject calibration:			
SAGE (SubjMean)	**0.488**	n/a[b]	n/a[b]

[a]AI4Pain Acc is the published video figure. QST Spearman is our per-frame evaluation of the released embedder (no target training), matched to SAGE's protocol. [b]AI4Pain's protocol freezes the encoder and trains a fresh head without reference-set inference, so calibration is not exercised; the reference-free row is the applicable SAGE transfer setting.

Table 3: Component ablation on 5-class BioVid LOSO. Token selection: AU-guided vs. holistic mean-pool; calibration: SubjMean or parameterized.

		Calibration			
Method	AU selection	SubjMean	parameterized	5-class Acc	5-class QWK
Reference-free	✓	–	–	0.2860	0.3421
Holistic + SubjMean	–	✓	–	0.2952	0.3671
SAGE (SubjMean)	✓	✓	–	0.3020	**0.3990**
SAGE (parameterized)	✓	–	✓	**0.3037**	0.3963

4.3 Ablation studies

That access to a subject's own unlabeled clips helps is to be expected; how it helps is not, and the ablations below show that the effect does not behave like generic extra data. A single reference clip is worse than none, a neutral-only reference is worse than a mixed one, a pool of other subjects' clips hurts, a learned set encoder adds nothing over the plain mean, and the gain is absent on the extreme-contrast binary task. We ablate, in turn, AU-guided token selection and subject calibration, the frozen encoder, the reference-set source, and the reference-set size; a direct OpenFace AU/PSPI input baseline is in Appendix D.

Component ablation. SAGE pairs AU-guided token selection with subject calibration. A natural question is how much of the lift originates from each, and whether either is dispensable. Tab. 3 ablates each component within the full SAGE pipeline on 5-class, toggling one mechanism per row while holding the head, training protocol, and LOSO partition fixed. A separate linear-probe diagnostic validating the PSPI partition against an arbitrary 2×3 block partition is in Appendix E.

Replacing AU selection with a holistic mean-pool costs 0.032 QWK, while removing calibration entirely costs 0.057—subject calibration accounts for the larger share of the SAGE lift, with AU routing contributing a smaller but non-negligible part. The learned reference encoder does not improve on non-parametric SubjMean on this task.

Frozen-encoder sensitivity. Is the BioVid result a property of MARLIN, of AU-guided selection, or of calibration? Table 4 separates the three. The upper block holds encoders that expose a spatial patch grid, so token selection can be applied: MARLIN (face-video MAE) and FMAE (face-image MAE pretrained on Face9M (Ning et al., 2024); per-frame 14×14 grid, temporal axis handled by the head). For each we report holistic mean-pooling versus AU-guided selection, with and without SubjMean. The lower block holds encoders that emit a single vector per frame, on which selection cannot be applied: PainFormer's embedder and three lighter facial models (EmotiEffNet-B0 (Savchenko, 2023; 2022), SwinFace (Qin et al., 2024), IR-18 ArcFace (Deng et al., 2019)). All rows use the same LOSO folds, head, and metric (5-class fold-mean QWK with bootstrap 95% CI); SubjMean is query-exclusive same-subject centering.

Table 4: Frozen-encoder sensitivity on 5-class BioVid LOSO (fold-mean QWK, bootstrap 95% CI). Upper block: patch-grid encoders with AU-guided selection toggled (holistic vs. AU-selected). Lower block: holistic-only encoders, on which selection is not applicable.

Backbone (params)	Input / pretraining	Ref-free	+SubjMean
Patch-grid encoders (AU-guided selection applicable):			
MARLIN ViT-B (∼87M), holistic	video, face SSL (YouTube Faces)	0.301 [0.25, 0.35]	0.367 [0.32, 0.41]
MARLIN ViT-B, AU-selected	video, face SSL	0.342 [0.29, 0.40]	**0.399** [0.35, 0.45]
FMAE ViT-B (∼86M), holistic	image, face SSL (Face9M)	0.281 [0.22, 0.34]	0.335 [0.29, 0.39]
FMAE ViT-B, AU-selected	image, face SSL	0.333 [0.28, 0.39]	0.370 [0.32, 0.42]
Holistic-only encoders (selection not applicable):			
PainFormer (∼33M)	image, supervised multi-task	0.199 [0.15, 0.25]	0.214 [0.17, 0.26]
EmotiEffNet-B0 (∼5M)	image, AffectNet FER	0.222 [0.17, 0.27]	0.254 [0.20, 0.31]
SwinFace (∼30M)	image, multi-task face	0.224 [0.18, 0.27]	0.239 [0.19, 0.29]
IR-18 ArcFace (∼24M)	image, face ID	0.073 [0.04, 0.11]	0.189 [0.15, 0.24]

Table 5: Reference-set source ablation, SAGE (parameterized), BL1-vs-PA4.

Mode	Acc	AUC
No-ref	0.728	0.781
Train-pool ref	0.717	0.762
Same-subj ref	**0.750**	**0.822**

Three readings follow. First, calibration helps every encoder in the table, from $+0.015$ (PainFormer) to $+0.116$ (IR-18), so the SubjMean gain is not a MARLIN artifact. Second, AU-guided selection helps on both patch-grid backbones and more on the image encoder: $+0.041$ QWK on MARLIN and $+0.052$ on FMAE without calibration, $+0.032$ and $+0.035$ with it. Third, the full recipe on FMAE reaches 0.370, within 0.03 of MARLIN and well above every holistic-only encoder, including PainFormer (0.214). Because FMAE shares face-domain self-supervision and scale with MARLIN but sees single frames, this bounds the value of spatiotemporal pretraining at about 0.03 QWK; the much larger gap to the lighter encoders is set by the representation itself, and even holistic MARLIN or FMAE with SubjMean (0.367, 0.335) is far above the best of them (0.254).

Reference-set source. The calibration gain could plausibly arise from two distinct mechanisms: (i) *genuine personalization*, where the test subject's own clips anchor a meaningful per-subject baseline, or (ii) a *generic context regularizer*, where any aggregated reference pool would supply useful normalization. Disentangling these matters for deployment: mechanism (i) requires per-subject enrollment, while mechanism (ii) would not. Tab. 5 varies the reference-set source on BL1-vs-PA4, holding everything else fixed at the parameterized variant: no reference, the test subject's own clips (same-subject), and a *train-pool* control in which R_s is drawn from training subjects, keeping the calibration architecture while removing same-subject information.

Same-subject reference is the only mode that improves over no-ref; train-pool reference is in fact worse, indicating that a non-personalized reference pool actively hurts because the head receives subject-incoherent $(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)$ inputs. The calibration gain therefore depends on genuine same-subject information rather than a generic context signal.

Reference-set size (SubjMean enrollment). A practical concern is whether full same-subject Part A sets (∼100 clips) are required. We vary the number K of *random unlabeled* same-subject clips under the SubjMean default (query-exclusive; no pain labels used), holding the head and LOSO protocol fixed. Fig. 2(d) reports 5-class fold-mean QWK against the anchor estimation error.

Two practical takeaways follow. First, the gain rises steeply up to about five unlabeled clips (recovering 0.030 of the 0.057 QWK full-set gain) and only slowly beyond: twenty clips recover about 70% and the full ∼100-clip set is best. A brief unlabeled enrollment is therefore the operating point, and larger pools refine rather than enable calibration. Second, a single reference clip ($K{=}1$) is *worse* than

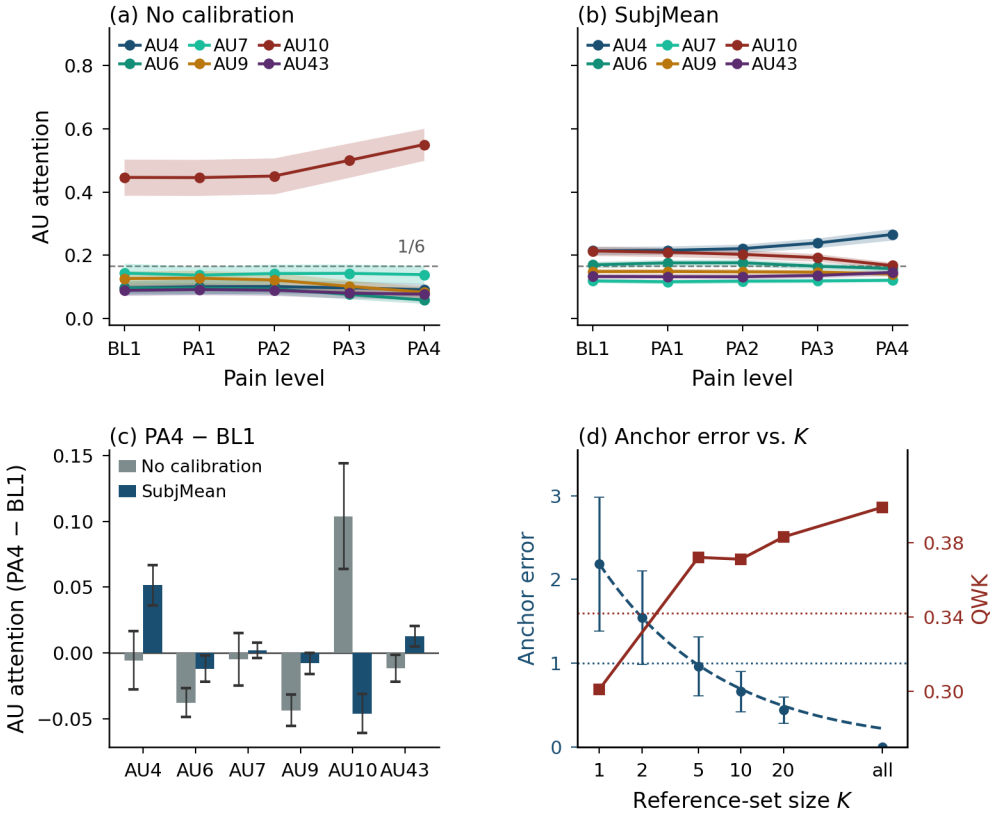

Figure 2: **What calibration changes, and why K matters.** (a–b) Region-attention weight of the held-out head per AU and pain level, without and with SubjMean (mean and 95% CI over 87 subjects; dashed line: uniform $1/6$). (c) Change in attention from BL1 to PA4 for both heads. (d) Anchor estimation error $\|\hat{\boldsymbol{\mu}}_K - \boldsymbol{\mu}_{\text{all}}\|$ relative to the subject's own BL1→PA4 signal $\|\mathbf{m}_{\text{PA4}} - \mathbf{m}_{\text{BL1}}\|$ (mean $\pm$ s.d. over subjects, 50 random draws per K; dashed: $K^{-1/2}$ decay from the $K{=}1$ value) against 5-class fold-mean QWK of SubjMean with K random unlabeled clips ($K{=}0/1/2/5/10/20/$all: $0.342/0.301/0.34/0.372/0.371/0.383/0.399$; dotted: reference-free QWK). A single clip subtracts noise about twice the size of the pain signal; the error drops below 1 by $K{=}5$, where QWK has overtaken the reference-free model.

using no reference at all: pure mean centering then subtracts a high-variance, clip-specific vector rather than a stable subject anchor, so one-shot deployments should fall back to the reference-free mode instead of forcing $K{=}1$. Composition studies in Appendix G (parameterized variant) further show that neutral-only references are not systematically better than mixed-level ones: the anchor is a subject-appearance center, not a pain-free baseline that must be kept clean of painful clips. Learned weighted pooling needs larger sets to help; at small K the non-parametric SubjMean default is the appropriate choice, which is why we headline SubjMean for enrollment realism.

What calibration changes. Fig. 2 offers two observations toward a mechanism. Panel (d) shows that the failure at $K{=}1$ is an estimation effect: the anchor error follows the expected $K^{-1/2}$ decay and crosses the size of the subject's own BL1→PA4 signal between $K{=}2$ and $K{=}5$, exactly where QWK overtakes the reference-free model. The same ratio is largest for muted subjects, whose signal is small, which is where calibration hurts (Sec. 4.2). Panels (a–c) use the per-AU structure that routing provides. Without calibration the head concentrates about half of its region attention on the AU10 mouth region, whose appearance varies most across individuals; with SubjMean the weight redistributes across all six regions, and the channel that grows with pain shifts to AU4 (brow lowerer) and AU43 (eye closure), the cues reported as most consistent across individuals in the PSPI literature (Prkachin & Solomon, 2008). This is consistent with calibration removing static subject

appearance so that the head relies on expression dynamics. Attention weights are an observation rather than a causal test.

4.4 Efficiency and total cost

The trainable footprint is small (SubjMean head ≈2M parameters; 387 for the AI4Pain probe; parameterized variant ≈13.8M), but inference includes the frozen MARLIN ViT-B (≈87.4M). On one RTX 4090 extraction takes 17.6 ms per 16-frame sub-clip (≈70 ms per BioVid video, 608 MiB), about $78\times$ real time; on an 8-thread CPU ≈2.4 s per video. Features are cached once (≈10 min for Part A) and reference anchors once per subject. Total inference cost including the backbone is therefore ≈88M parameters at ≈70 ms per video; PainFormer totals ≈33M including its mixer.

5 Limitations

Enrollment requirement. Calibration needs a few unlabeled clips of each subject; the gain rises steeply up to about five and grows slowly with more. Where no enrollment is possible, the reference-free mode applies (0.342 QWK), and a single clip should not be used as reference, since it is worse than none. Centering targets the fine-grained ordinal task; it does not help the maximal-contrast BL1-vs-PA4 task ($0.728 \rightarrow 0.697$ Acc), where the parameterized variant is used instead.

Representation and subject difficulty. The absolute level depends on the frozen representation: calibration helps every encoder we test, but lighter image encoders remain far below MARLIN, and routing on their feature maps is untested. Per-fold QWK spread is large and tracks an independent baseline's subject difficulty; muted responders sit near chance and calibration does not help them.

Scope. All three datasets contain experimentally induced transient pain in healthy adults; chronic and clinical pain are out of scope, and QST-Pilot is a small unreleased pilot. AU-guided selection needs frontal landmarks and degrades under occlusion or extreme pose. Demographic coverage is discussed in the ethics statement.

6 Conclusion

We showed that two properties of pain expression—anatomical localization and subject specificity—can be made explicit on a frozen face-video transformer with almost no trained parameters. Subject calibration from a few unlabeled clips is the main source of the gain, helps every frozen encoder we tested, requires genuine same-subject reference, and fails at $K{=}1$ for a reason we can measure; a learned reference encoder does not beat the plain mean. AU-guided token selection contributes structure rather than accuracy: it improves over holistic pooling at 12% of the tokens on two backbones and makes the head's behavior readable per AU, which reveals that calibration moves attention from the mouth to brow and eye cues. The resulting model is competitive on BioVid and transfers without backbone fine-tuning. Its absolute level is set by the representation, and clinical generalization remains to be established. Because per-subject calibration is free, backbone-agnostic, and consistently helpful, we suggest it be treated as a standard component and a standard baseline in facial pain estimation: new methods should show what they add beyond a frozen encoder centered on a few unlabeled clips of the same subject.

AI use statement

In this work, we have not used generative AI tools for generating data, developing the conceptual framework, formulating claims, proposing hypotheses, designing the experimental methodology, implementing the proposed methods, processing the datasets, or interpreting the results; these tasks were carried out solely by the authors. We used generative AI tools for the following assistance: identifying and surveying relevant literature; drafting and editing of manuscript text; reviewing, refactoring, and debugging of experiment code written by the authors; and formatting of references and submission files. All AI-assisted text was reviewed, edited, and is endorsed by the authors; AI-assisted code was verified against the logged outputs of our own experimental runs; and no number reported in this paper was produced by generative AI. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

Ethics statement

BioVid Heat Pain and AI4Pain are public datasets and are used under the terms of their original releases. QST-Pilot is an unpublished pilot collected by the authors and is not released. It contains identifiable facial video of healthy adult volunteers, gathered with informed consent under institutional ethics approval, and is used only for the zero-shot evaluation in this paper. Automated pain estimation is a decision-support tool and must not replace clinical assessment. Models fit on laboratory pain in healthy adults may not transfer to clinical care; any such use would require prospective validation and bias auditing. Demographic coverage is limited: BioVid Part A has 87 subjects (44 M/43 F, three age groups) and AI4Pain 65 (23 F, ages 17–52), and neither releases ethnicity labels, so the documented racial bias in human pain assessment cannot be measured for our models here. Clinical facial video is identifiable data and any deployment requires institutional oversight.

Reproducibility statement

Sec. 4 specifies the LOSO protocol, metrics, and the BioVid, AI4Pain, and QST-Pilot evaluation settings. Appendix A gives the exact AU-to-landmark mapping; Appendix B gives the backbone and feature-extraction procedure, reference-set sizes, head architecture, optimization hyperparameters, the deterministic LOSO fold definition, cutoff recalibration, and pseudocode for the forward pass and training step. LOSO folds are fully determined by the subject ordering in Appendix B, so no fold file is needed. Source code and configuration files will be released upon publication. BioVid and AI4Pain are public; QST-Pilot is not released.

References

Tadas Baltrušaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial behavior analysis toolkit. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, pp. 59–66, 2018. doi: 10.1109/FG.2018.00019.

Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2D & 3D face alignment problem? (and a dataset of 230,000 3D facial landmarks). In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1021–1030, 2017. doi: 10.1109/ICCV.2017.116.

Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. MARLIN: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1493–1504, 2023.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331, 2020.

Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220, 1968. doi: 10.1037/h0026256.

Kenneth D. Craig. Observer bias in judging pain in others. *Pain*, 18:S14, 1984. doi: 10.1016/0304-3959(84)90152-0. Conference abstract.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

Paul Ekman and Wallace V. Friesen. *Facial Action Coding System: A Technique for the Measurement of Facial Movement*. Consulting Psychologists Press, Palo Alto, CA, 1978.

Raul Fernandez-Rojas, Calvin Joseph, Niraj Hirachan, Ben Seymour, and Roland Goecke. The AI4Pain grand challenge 2024: Advancing pain assessment with multimodal fNIRS and facial video analysis. In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 55–60, 2024. doi: 10.1109/ACIIW63320.2024.00012.

Stefanos Gkikas and Manolis Tsiknakis. A full transformer-based framework for automatic pain estimation using videos. In *2023 45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pp. 1–6. IEEE, 2023.

Stefanos Gkikas, Raul Fernandez Rojas, and Manolis Tsiknakis. PainFormer: A vision foundation model for automatic pain assessment. *IEEE Transactions on Affective Computing*, 16(4):3369–3386, 2025. doi: 10.1109/TAFFC.2025.3605475.

Keela Herr, Patrick J. Coyne, Margo McCaffery, Renee Manworren, and Sandra Merkel. Pain assessment in the patient unable to self-report: Position statement with clinical practice recommendations. *Pain Management Nursing*, 12(4):230–250, 2011. doi: 10.1016/j.pmn.2011.10.002.

Dong Huang, Xiaoyi Feng, Haixi Zhang, Zitong Yu, Jinye Peng, Guoying Zhao, and Zhaoqiang Xia. Spatio-temporal pain estimation network with measuring pseudo heart rate gain. *IEEE Transactions on Multimedia*, 24:3300–3313, 2022a. doi: 10.1109/TMM.2021.3096080.

Yibo Huang, Linbo Qing, Shengyu Xu, Lu Wang, and Yonghong Peng. HybNet: A hybrid network structure for pain intensity estimation. *The Visual Computer*, 38(3):871–882, 2022b. doi: 10.1007/s00371-021-02056-y.

Yusuke Iwasawa and Yutaka Matsuo. Test-time classifier adjustment module for model-agnostic domain generalization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7680–7689, 2021.

Sebastian Kaltwang, Ognjen Rudovic, and Maja Pantic. Continuous pain intensity estimation from facial expressions. In *International Symposium on Visual Computing (ISVC)*, pp. 368–377, 2012. doi: 10.1007/978-3-642-33191-6_36.

Judith Kappesser, Amanda C. de C. Williams, and Kenneth M. Prkachin. Testing two accounts of pain underestimation. *Pain*, 124(1-2):109–116, 2006. doi: 10.1016/j.pain.2006.04.003.

Juho Lee, Yoonho Lee, Jungtaek Kim, Adam R. Kosiorek, Seungjin Choi, and Yee Whye Teh. Set transformer: A framework for attention-based permutation-invariant neural networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, pp. 3744–3753, 2019.

Wei Lin, Muhammad Jehanzeb Mirza, Mateusz Kozinski, Horst Possegger, Hilde Kuehne, and Horst Bischof. Video test-time adaptation for action recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Harold Merskey and Nikolai Bogduk. *Classification of Chronic Pain: Descriptions of Chronic Pain Syndromes and Definitions of Pain Terms*. IASP Press, Seattle, 2nd edition, 1994.

Mang Ning, Albert Ali Salah, and Itir Onal Ertugrul. Representation learning and identity adversarial training for facial behavior understanding, 2024.

Pooja Prajod, Tobias Huber, and Elisabeth André. Using explainable AI to identify differences between clinical and experimental pain detection models based on facial expressions. In *International Conference on Multimedia Modeling*, pp. 311–322, 2022. doi: 10.1007/978-3-030-98358-1_25.

Pooja Prajod, Dominik Schiller, Daksitha Withanage Don, and Elisabeth André. Faces of experimental pain: Transferability of deep-learned heat pain features to electrical pain. In *2024 12th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, pp. 31–38, 2024. doi: 10.1109/ACIIW63320.2024.00009.

Kenneth M. Prkachin and Patricia E. Solomon. The structure, reliability and validity of pain expression: Evidence from patients with shoulder pain. *Pain*, 139(2):267–274, 2008. doi: 10.1016/j.pain.2008.04.010.

Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. SwinFace: A multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4):2223–2234, 2024. doi: 10.1109/TCSVT.2023.3304724.

Pau Rodriguez, Guillem Cucurull, Jordi González, Josep M. Gonfaus, Kamal Nasrollahi, Thomas B. Moeslund, and F. Xavier Roca. Deep pain: Exploiting long short-term memory networks for facial expression classification. *IEEE Transactions on Cybernetics*, 52(5):3314–3324, 2022. doi: 10.1109/TCYB.2017.2662199.

Andrey V. Savchenko. HSEmotion: High-speed emotion recognition library. *Software Impacts*, 14: 100433, 2022.

Andrey V. Savchenko. EmotiEffNets for facial processing in video-based valence-arousal prediction, expression classification and action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 5716–5724, 2023.

Steffen Schneider, Evgenia Rusak, Luisa Eck, Oliver Bringmann, Wieland Brendel, and Matthias Bethge. Improving robustness against common corruptions by covariate shift adaptation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Jake Snell, Kevin Swersky, and Richard Zemel. Prototypical networks for few-shot learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jie Ting, Yi-Cheng Yang, Li-Chen Fu, Chu-Lin Tsai, and Chien-Hua Huang. Distance ordering: A deep supervised metric learning for pain intensity estimation. In *IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 1083–1088, 2021. doi: 10.1109/ICMLA52953.2021.00177.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

Sophie Trawalter, Kelly M. Hoffman, and Adam Waytz. Racial bias in perceptions of others' pain. *PLoS ONE*, 7(11):e48546, 2012. doi: 10.1371/journal.pone.0048546.

Steffen Walter, Sascha Gruss, Hagen Ehleiter, Junwen Tan, Harald C. Traue, Philipp Werner, Ayoub Al-Hamadi, Stephen Crawcour, Adriano O. Andrade, and Gustavo Moreira Da Silva. The BioVid heat pain database: Data for the advancement and systematic validation of an automated pain recognition system. In *IEEE International Conference on Cybernetics (CYBCO)*, pp. 128–131, 2013. doi: 10.1109/CYBConf.2013.6617456.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. In *International Conference on Learning Representations (ICLR)*, 2021.

Jinsheng Wei, Fengzhou Guo, Yante Li, Haoyu Chen, Guanming Lu, and Guoying Zhao. Micro-AU CLIP: Fine-grained contrastive learning from local independence to global dependency for micro-expression action unit detection. *arXiv preprint arXiv:2603.16302*, 2026.

Philipp Werner, Ayoub Al-Hamadi, Robert Niese, Steffen Walter, Sascha Gruss, and Harald C. Traue. Automatic pain recognition from video and biomedical signals. In *International Conference on Pattern Recognition (ICPR)*, pp. 4582–4587, 2014. doi: 10.1109/ICPR.2014.784.

Philipp Werner, Daniel Lopez-Martinez, Steffen Walter, Ayoub Al-Hamadi, Sascha Gruss, and Rosalind W. Picard. Automatic recognition methods supporting pain assessment: A survey. *IEEE Transactions on Affective Computing*, 13(1):530–552, 2022. doi: 10.1109/TAFFC.2019.2946774.

Lior Wolf, Tal Hassner, and Itay Maoz. Face recognition in unconstrained videos with matched background similarity. In *CVPR 2011*, pp. 529–534. IEEE, 2011.

Haochen Xu and Manhua Liu. A deep attention transformer network for pain estimation with facial expression video. In *Chinese Conference on Biometric Recognition (CCBR)*, pp. 112–119, 2021. doi: 10.1007/978-3-030-86608-2_13.

Ruijing Yang, Xiaopeng Hong, Jinye Peng, Xiaoyi Feng, and Guoying Zhao. Incorporating high-level and low-level cues for pain intensity estimation. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 3495–3500. IEEE, 2018.

Ruijing Yang, Ziyu Guan, Zitong Yu, Xiaoyi Feng, Jinye Peng, and Guoying Zhao. Non-contact pain recognition from video sequences with remote physiological measurements prediction. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 1231–1237, 2021. doi: 10.24963/ijcai.2021/170.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabás Póczos, Ruslan R. Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Xiang Zhang, Huiyuan Yang, Taoyue Wang, Xiaotian Li, and Lijun Yin. Multimodal channel-mixing: Channel and spatial masked autoencoder on facial action unit detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 6077–6086, 2024.

Kaili Zhao, Wen-Sheng Chu, and Honggang Zhang. Deep region and multi-label learning for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3391–3399, 2016. doi: 10.1109/CVPR.2016.369.

Li-Ming Zhao, Xu Yan, and Bao-Liang Lu. Plug-and-play domain adaptation for cross-subject EEG-based emotion recognition. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

Jing Zhou, Xiaopeng Hong, Fei Su, and Guoying Zhao. Recurrent convolutional neural network regression for continuous pain intensity estimation in video. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1535–1543, 2016.

A PSPI AND THE AU-TO-LANDMARK MAPPING

A.1 THE PSPI METRIC

The PSPI (Prkachin & Solomon, 2008) is a FACS (Ekman & Friesen, 1978) composite score that aggregates six pain-relevant AUs into a single per-frame intensity score:

$$\mathrm{PSPI} = \mathrm{AU4} + \max(\mathrm{AU6}, \mathrm{AU7}) + \max(\mathrm{AU9}, \mathrm{AU10}) + \mathrm{AU43}, \tag{3}$$

where each contributing AU intensity is scored on the FACS 0–5 scale (0=absent, A=trace, B=slight, C=marked, D=severe, E=maximum). The two $\max$ operators reflect FACS-level redundancy: AU6 (cheek raiser) and AU7 (lid tightener) co-activate the orbicularis oculi muscle from different angles, and AU9 (nose wrinkler) and AU10 (upper lip raiser) similarly draw on the levator complex; pain expression recruits one or the other but rarely adds anatomically. SAGE uses these six AUs to define the AU-guided token selection: each AU corresponds to a landmark-defined region of interest on the patch grid, and tokens within those regions are pooled into a per-AU feature vector (Sec. 3.1).

A.2 ANATOMY OF THE SIX PSPI AUS

Table 6 lists the anatomical name, dominant muscle, and FACS-defined visible action of each AU. We use these names throughout the paper.

A.3 LANDMARK-TO-TOKEN MAPPING

Each AU's region of interest on the 14×14 patch grid is defined by a small set of FAN-68 landmark indices (Bulat & Tzimiropoulos, 2017). Table 7 lists the exact mapping used in all SAGE experiments. We follow the standard FAN-68 numbering scheme (0–67), where 0–16 trace the jaw, 17–21 and 22–26 the eyebrows, 27–35 the nose, 36–41 and 42–47 the eyes, and 48–67 the mouth.

Table 6: Anatomical interpretation of the six PSPI AUs.

AU	Common name	Visible action / muscle
AU4	Brow lowerer	Inner eyebrows pulled down and together; corrugator supercilii / depressor supercilii.
AU6	Cheek raiser	Cheeks lifted, "crow's feet" wrinkles at outer eye corners; orbicularis oculi (orbital portion).
AU7	Lid tightener	Eyelids tightened, narrowing the eye aperture; orbicularis oculi (palpebral portion).
AU9	Nose wrinkler	Skin on the bridge of the nose pulled upward; levator labii superioris alaeque nasi.
AU10	Upper lip raiser	Upper lip lifted, philtrum deepened; levator labii superioris.
AU43	Eye closure	Eyelids closed; relaxed orbicularis oculi (passive closure, distinguishing AU43 from active blinks).

Table 7: PSPI AU-to-landmark mapping. FAN-68 landmark indices for each AU project onto the 14×14 patch grid; patch tokens at those positions are spatially mean-pooled into a per-AU per-frame $D=768$ vector. The mapping is fixed across subjects and clips; per-frame variation comes only from landmark motion.

AU	FAN-68 landmark indices $\mathcal{L}_a$	Anatomical region covered
AU4	$\{19, 20, 21, 22, 23, 24\}$	Inner brow / glabella (medial portions of both eyebrows).
AU6	$\{1, 2, 36, 39, 42, 45\}$	Cheekbone strip below the eyes (upper-jaw landmarks 1, 2 plus the four eye corners 36, 39, 42, 45).
AU7	$\{38, 41, 44, 47\}$	Eyelid mid-points (upper and lower lid centers of both eyes).
AU9	$\{31, 32, 33, 34, 35\}$	Nose base (nostrils and nose tip).
AU10	$\{49, 50, 51, 52, 53\}$	Upper lip outer contour (cupid's bow / philtrum region).
AU43	$\{37, 38, 40, 41, 43, 44, 46, 47\}$	Full eyelid perimeter (top and bottom of both eyes).

The patch grid positions covered by each $\mathcal{L}_a$ vary per-frame as the subject moves, but the landmark indices themselves are fixed. Within a single temporal slice, the six AUs together cover approximately 24 unique 14×14 grid positions on average (out of 196 spatial positions per slice). Aggregated over the $T=8$ MARLIN temporal slices, this corresponds to roughly 192 of the 1568 spatiotemporal tokens being consulted by token selection ($\sim12\%$); tokens outside the AU regions are discarded. Each AU's selected tokens are then mean-pooled into a single $D=768$-dim vector per temporal slice, giving the $(T=8, A=6, D=768)$ per-sub-clip tensor; concatenating $S=4$ such non-overlapping sub-clips along the time axis (covering the full BioVid clip at temporal stride 2) gives the per-video $(T_{\text{v}}=32, A=6, D=768)$ representation passed to subject calibration and the head.

B Hyperparameters and infrastructure

B.1 Backbone and feature extraction

We use the publicly released MARLIN Vision Transformer-Base (ViT-B) (Dosovitskiy et al., 2021) checkpoint (Cai et al., 2023) (`marlin_vit_base_ytf`, ≈87.4M parameters), pretrained on YouTube Faces with the masked face-video self-supervised objective. Each 5.5-second BioVid clip is divided into $S=4$ consecutive, non-overlapping windows of 32 frames; each window is subsampled at temporal stride 2 to $T_{\text{frame}}=16$ RGB frames at 224×224 and encoded independently, and a window with fewer than 32 frames is mirror-padded. The backbone is loaded once per LOSO fold, weights frozen and dropout disabled, and produces an $8\times14\times14\times768$ patch grid per window. After token selection the four windows are concatenated along the time axis, so the per-video representation has $T_{\text{v}}=4\times8=32$ tubelet bins.



Figure 3: Two BioVid subjects at PA4 (max pain stimulus): strong PSPI activation (left, 8.55) vs. minimal activation (middle, 0.78). The AU-token cosine distance between them is 47% above the within-subject PA4-to-PA4 average.

B.2 Reference-set sizes

The reference set R_s size is task-dependent. For the 5-class ordinal task, R_s contains all $|R_s| = 100$ Part-A clips of the test subject (five pain levels, 20 clips per level), with the query clip excluded. For the binary BL1-vs-PA4 task, R_s is restricted to the two relevant levels and contains $|R_s| = 40$ clips (20 BL1 + 20 PA4) per subject. Reference-set composition and size ablations are reported in Appendix G.

B.3 Parameterized calibration encoder architecture

The DeepSets-style reference-set encoder f_ϕ is a stack of three elements:

1. Per-clip MLP applied to each $\mathbf{x}_r \in \mathbb{R}^{A \times D}$: $\mathrm{Linear}(D = 768 \to 256) \to \mathrm{LayerNorm} \to \mathrm{GELU} \to \mathrm{Dropout}(0.1) \to \mathrm{Linear}(256 \to D)$.
2. Multi-head self-attention over the reference clips ($|R_s|$ tokens, 4 heads, $\mathrm{Linear}(D = 768 \to D)$ key/query/value, residual).
3. Learned weighted pooling (Lee et al., 2019): $\mathrm{Linear}(D \to 1)$ score per clip, softmax, weighted sum into a single $\mathbb{R}^{A \times D}$ vector.

Two zero-initialized residual heads then map the pooled representation to $\boldsymbol{\mu}_s \in \mathbb{R}^{A \times D}$ and $\log \boldsymbol{\sigma}_s \in \mathbb{R}^{A \times D}$. Floor on $\boldsymbol{\sigma}_s$ is the per-AU within-subject empirical standard error of the training reference clips, computed once at the start of each fold.

B.4 Classifier head

The classifier head h_ψ is a region-attention temporal fusion model with one branch per AU:

- Per-region projection: $\mathrm{Linear}(4D = 3072 \to 128) \to \mathrm{LayerNorm} \to \mathrm{GELU} \to \mathrm{Dropout}(0.3)$.
- Per-region BiGRU: hidden 64 per direction (bidirectional, total 128); mean-pooled across the $T = 32$ temporal axis.
- Region attention: tanh score per AU, softmax across the $A = 6$ regions, weighted sum to produce a fused $\mathbb{R}^{128}$ token, concatenated with the per-region max-pool to form a $\mathbb{R}^{256}$ representation, projected back to $\mathbb{R}^{128}$ via $\mathrm{Linear}(256 \to 128)$.
- CORAL ordinal head for 5-class ($K - 1 = 4$ cumulative logits) or sigmoid head for binary tasks.

Total head parameter count: $\sim$2.0M (5-class), $\sim$1.9M (binary). Fig. 4 diagrams the head.

The Holistic baseline used in Sec. 4.3 replaces the per-region branches with a single $D = 768$ branch and otherwise matches this layout.

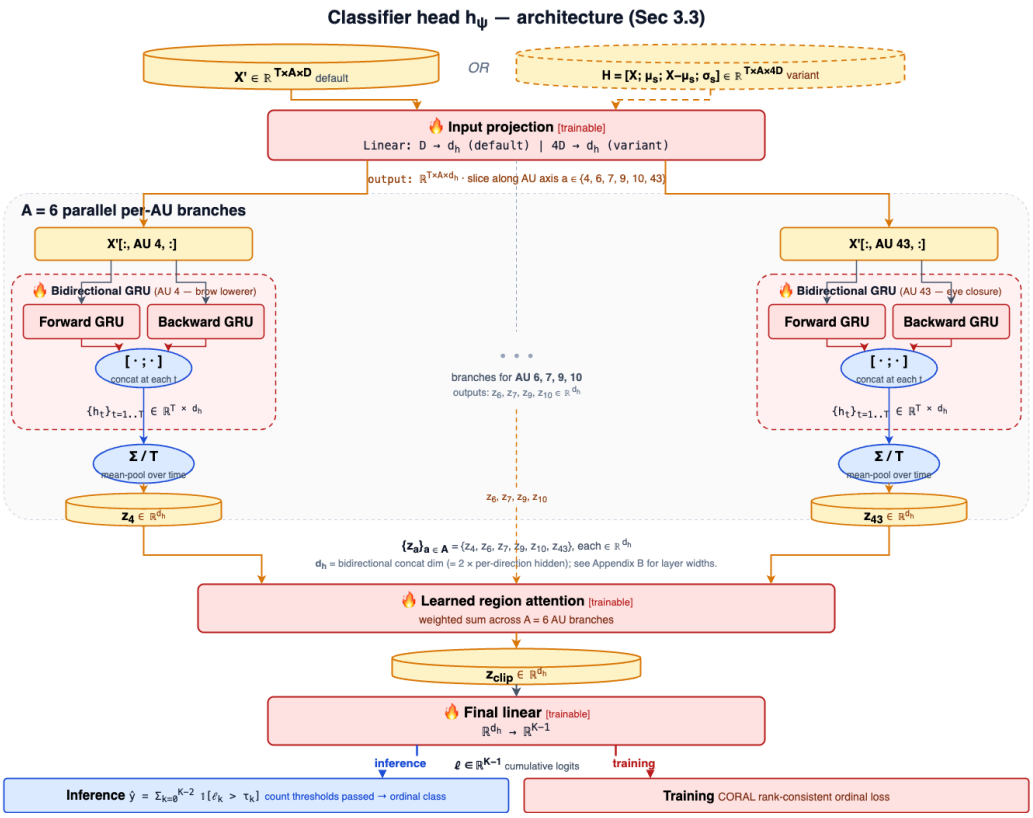

Figure 4: **Classifier head h_ψ.** The head accepts either the default residual input $\mathbf{X}'$ (SubjMean) or the 4-channel concatenation $\mathbf{H} = [\mathbf{X};\, \boldsymbol{\mu}_s;\, \mathbf{X}-\boldsymbol{\mu}_s;\, \boldsymbol{\sigma}_s]$ (parameterized variant); the architecture is identical apart from the input projection's dimension. After projection, $A = 6$ parallel per-AU branches each apply a small bidirectional GRU and mean-pool over time, yielding one per-AU vector $\mathbf{z}_a$. Learned region attention fuses $\{\mathbf{z}_a\}_{a=1}^{6}$ into a single clip-level vector $\mathbf{z}_{\text{attn}}$, which a final linear layer maps to $K-1$ cumulative-logit boundaries $\boldsymbol{\ell}$. Training minimizes the CORAL rank-consistent ordinal loss; at inference, the predicted level counts how many cumulative logits exceed the recalibrated cutoffs $\hat{\boldsymbol{\tau}}$.

Table 8: Optimization hyperparameters. All BioVid LOSO runs share these settings unless noted otherwise.

Setting	Value
Optimizer (head)	AdamW, $\beta_1 = 0.9$, $\beta_2 = 0.999$
Learning rate (head)	1×10^{-3}, cosine decay
Weight decay (head)	1×10^{-4}
Encoder LR multiplier	0.01× head LR (when f_ϕ is used)
Batch size	96 (binary), 48 (5-class)
Max epochs	100
Early stopping	validation QWK (5-class) / F1-macro (binary), patience 20
Mixed precision	bf16
Gradient clipping	$\|g\|_2 \leq 1.0$
Dropout (head)	0.3
Reference set $\|R_s\|$	40 (binary, 20 BL1 + 20 PAk); 100 (5-class, all levels)
Random seed	123

B.5 Optimization

B.6 LOSO fold definition

LOSO folds are defined deterministically from the fold index i:

Algorithm 1 SAGE forward pass (single example).

Require: Query video V_q and query-exclusive reference set $R_s = \{V_1, \ldots, V_n\}$ ($V_q \notin R_s$) from the same subject

// AU-guided token selection on every clip

$\mathbf{X}_q \leftarrow \mathrm{LandmarkPool}(\mathrm{MARLIN}(V_q))$ ▷ $T \times A \times D$ for query

for $i = 1, \ldots, n$ **do**

  $\mathbf{X}_i \leftarrow \mathrm{LandmarkPool}(\mathrm{MARLIN}(V_i))$

end for

// Subject calibration (Sec. 3.2): temporal mean-pool → subject anchor → calibrated head input

for $i = 1, \ldots, n$ **do**

  $\mathbf{x}_i \leftarrow \frac{1}{T} \sum_{t=1}^{T} \mathbf{X}_i(t, \cdot, \cdot)$ ▷ $A \times D$

end for

if parameterized variant **then**

  $(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s) \leftarrow f_\phi\big(\{\mathbf{x}_i\}_{i=1}^{n}\big)$ ▷ learned permutation-invariant aggregation over i

  $\mathbf{H} \leftarrow [\, \mathbf{X}_q \,;\, \boldsymbol{\mu}_s \,;\, \mathbf{X}_q - \boldsymbol{\mu}_s \,;\, \boldsymbol{\sigma}_s \,]$ ▷ $\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s$ broadcast over T axis

  $\boldsymbol{\ell} \leftarrow h_\psi(\mathbf{H})$

else ▷ SubjMean (default)

  $\boldsymbol{\mu}_s \leftarrow \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$ ▷ arithmetic mean over i

  $\boldsymbol{\ell} \leftarrow h_\psi(\mathbf{X}_q - \boldsymbol{\mu}_s)$

end if

return $\boldsymbol{\ell}$

Algorithm 2 SAGE training (one optimizer step).

Require: Batch $\{(V_q^{(i)}, R_s^{(i)}, y^{(i)})\}_{i=1}^{B}$ of (query, reference-set, label) triples

for $i = 1, \ldots, B$ **do**

  $\boldsymbol{\ell}^{(i)} \leftarrow$ Algorithm 1$(V_q^{(i)}, R_s^{(i)})$ ▷ forward pass

end for

$\mathcal{L} \leftarrow \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{\mathrm{CORAL}}(y^{(i)}, \boldsymbol{\ell}^{(i)})$ ▷ or BCE for binary task

Backpropagate $\nabla \mathcal{L}$ to update h_ψ and (parameterized only) f_ϕ ▷ MARLIN frozen, no gradient update

- Test subject: subject i in the deterministic subject ordering (sorted by subject ID).
- Validation subjects: a rotating block of 7 subjects starting from index $i + 1$ in the same ordering, wrapping around as needed.
- Training subjects: the remainder.

The full subject ID list and fold assignments are released with the code.

B.7 Per-fold cutoff recalibration (5-class only)

After each fold's CORAL head is trained, we sweep the four ordinal cutoffs $c_k \in \{-2.0, -1.95, \ldots, +2.0\}$ via three-pass coordinate descent on the same fold's validation subjects' logits, optimizing fold-validation QWK (default) or accuracy. The recalibrated cutoffs are then applied to the held-out test subject's logits. This is a standard ordinal classification hygiene step and is not part of our methodological contribution; binary tasks use a fixed 0.5 threshold.

B.8 Hardware and timing

All BioVid LOSO experiments run on local servers with $2\times$ NVIDIA RTX 4090 (24GB) GPUs. Folds are dispatched two-at-a-time (round-robin by GPU ID). Per-fold wall-clock:

- Binary BL1-vs-PA4: $\sim$15–30 seconds per fold (Holistic baseline) to 2–5 minutes per fold (full anchor pipeline).
- 5-class CORAL: $\sim$30 seconds per fold (Holistic baseline) to 8–12 minutes per fold (full anchor pipeline).

B.9 Reproducibility

LOSO folds are determined by the sorted subject ID ordering above; the AU-to-landmark mapping is Table 7. Source code, configuration files, and per-fold result files will be released upon publication. With seed 123 and the deterministic LOSO partition, all reported numbers should reproduce exactly on the same hardware.

C Parameterized reference-set encoder

Beyond the empirical default, we additionally instantiate subject calibration as a learned reference-set encoder f_ϕ that maps R_s to a subject posterior

$$(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s) = f_\phi\big(\{\mathbf{x}_r\}_{r \in R_s}\big), \tag{4}$$

with $\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s \in \mathbb{R}^{A \times D}$; f_ϕ is a permutation-invariant DeepSets-style (Zaheer et al., 2017) backbone with learned weighted pooling (Lee et al., 2019), and its residual heads are zero-initialized so that at training step 0 the output is identically $\boldsymbol{\mu}_s = \boldsymbol{\mu}_s^{\text{emp}}$, with any deviation from the empirical baseline being data-driven. $\boldsymbol{\sigma}_s$ is floored at the empirical within-subject standard error to prevent variance collapse, and the head consumes a four-channel calibrated input

$$\mathbf{H}(t,k,\cdot) = \big[\,\mathbf{X}(t,k,\cdot)\,;\,\boldsymbol{\mu}_s(k,\cdot)\,;\,\mathbf{X}(t,k,\cdot) - \boldsymbol{\mu}_s(k,\cdot)\,;\,\boldsymbol{\sigma}_s(k,\cdot)\,\big] \in \mathbb{R}^{T \times A \times 4D}, \tag{5}$$

i.e. the query, the anchor, the explicit deviation, and the per-feature uncertainty. It is reported as an optional large-reference extension: it does not improve on SubjMean for the 5-class task (Table 1) and is used only for the binary rows and, as a controlled substrate, in the reference-source diagnostics.

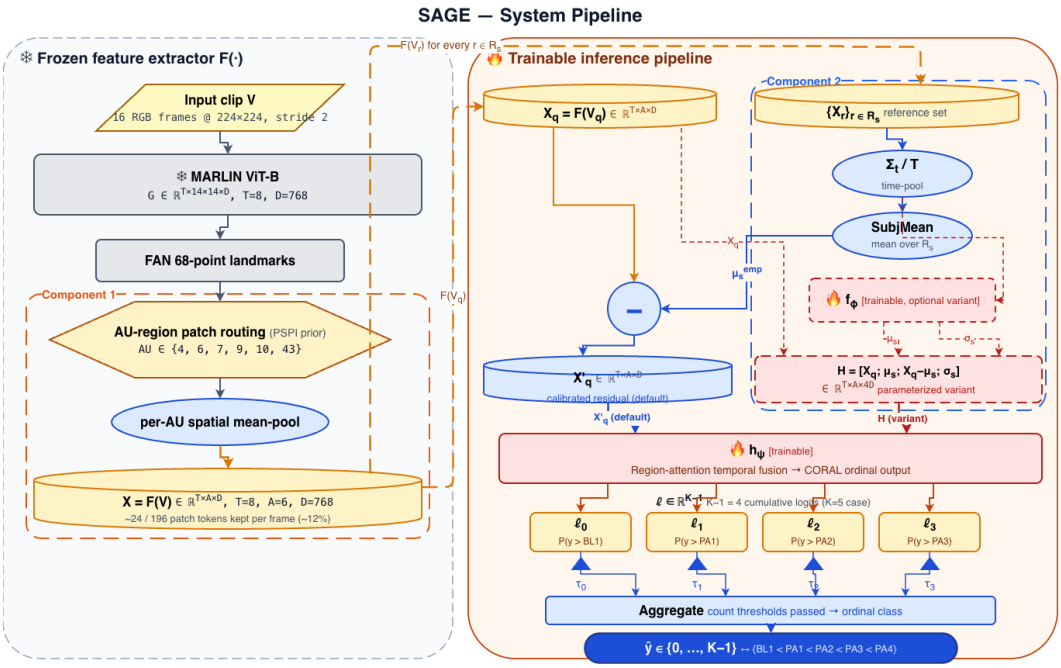

Figure 5: Full SAGE architecture with tensor shapes, including the parameterized variant: both query and reference clips pass through the frozen extractor $F(\cdot)$; f_ϕ maps the reference set to $(\boldsymbol{\mu}_s, \boldsymbol{\sigma}_s)$, which form the four-channel input $\mathbf{H}$ to the head; the CORAL head emits $K-1$ cumulative logits that are thresholded and counted.

D Direct AU/PSPI input baselines

A natural question is whether the SAGE lift reduces to direct AU detection: if MARLIN tokens add nothing beyond what an OpenFace-style AU detector provides, feeding the same head a direct

AU intensity vector should reproduce the SAGE result. Table 9 tests this by feeding the same head (no reference, no calibration) three direct AU inputs in place of MARLIN tokens: scalar PSPI, the pain-relevant OpenFace AU intensities, or the full OpenFace AU set. All three fall well short of the SAGE pipeline—the gain is not reducible to AU detection but requires visual features in MARLIN's tokens that are not summarized by static AU intensity.

Table 9: Direct AU/PSPI baselines on 5-class BioVid (same head, no reference).

Input	QWK	Acc
PSPI scalar	0.2697	0.2490
Pain-relevant AUs	0.3163	0.2682
All AUs	0.3189	0.2661
SAGE (SubjMean)	**0.3990**	0.3020
SAGE (parameterized)	0.3963	**0.3037**

E TOKEN-SELECTION LINEAR-PROBE DIAGNOSTIC

We probe the token-selection representation directly with a BL1-vs-PA4 binary linear-probe diagnostic under strict LOSO, L_2 LogisticRegression with train-only standardization (Table 10). Three variants test two complementary claims. (a) MARLIN holistic (768D, mean over the full 1568-token grid) and AU-guided token selection (4,608D, mean within each PSPI region) reach 63.96% and $\mathbf{64.74}\%$ accuracy respectively ($+0.78$pp absolute): retaining only $\sim 12\%$ of the patch tokens does not lose linear-probe signal, confirming that the $\sim 88\%$ of patches outside the PSPI regions are dispensable. (b) An arbitrary 2×3 block partition of the 14×14 grid, also pooled to 4,608D, reaches only 61.57%—3.17pp *below* PSPI and 2.39pp below the holistic baseline. The PSPI-specific partition is therefore not interchangeable with any 6-region partition: a geometry-only block split that ignores facial anatomy actively hurts the linear probe. We read these together as empirical validation of the PSPI anatomical prior: the prior selects the right tokens *and* the right partition.

Table 10: Token-selection linear-probe diagnostic on BL1-vs-PA4 (strict LOSO, L_2 LogReg; linear-probe diagnostic, not the SAGE head used in Tab. 1). AU selection matches the holistic baseline at $\sim 12\%$ of patch tokens; an arbitrary 2×3 geometric partition at the same dimensionality underperforms PSPI by 3.17pp.

Representation	Dim	Tokens used	BL1-vs-PA4 Acc
MARLIN holistic (full-grid mean)	768	1568 (100%)	63.96%
AU-guided token selection	4,608	$\sim$192 ($\sim$12%)	**64.74**%
Arbitrary 2×3 block partition	4,608	1568 (100%)	61.57%

F 5-CLASS REFERENCE-SOURCE MECHANISM DIAGNOSTIC

The BL1-vs-PA4 train-pool result in Sec. 4.3 shows that a non-personalized reference hurts; we extend the test on the 5-class task with two reference-source replacements that each break a different invariant. We use the parameterized calibration variant trained against same-subject Part A references and, at inference time only, swap the reference set to: (a) Part A clips of a randomly chosen *wrong* training subject (*wrong-subject*, breaks subject identity), and (b) the BL1 clip pool of all training subjects (*train-pop*, breaks subject specificity entirely). Both controls keep the trained encoder fixed; only R_s at inference varies.

Both controls drop T1 QWK by ~ 0.06–0.07, ruling out a "generic prior" explanation: any reference is not enough, the calibration gain depends on genuinely subject-specific information. The encoder's posterior σ_s stays at the same low level (≈ 0.005–0.006) across all three reference modes, meaning the encoder does not internally flag a wrong reference set as uncertain—it confidently produces a $\boldsymbol{\mu}_s$ that simply happens to point at the wrong location, an *anchor mismatch* rather than an uncertainty-driven failure.

Table 11: 5-class reference-source diagnostic (strict LOSO, parameterized calibration; ΔT1 QWK vs. same-subject Part-A anchor). Both controls drop QWK by ~ 0.06–0.07; the encoder's posterior σ_s stays at the anchor level.

Reference source	T1 QWK	Acc	F1	posterior σ_s
Same-subject Part A (anchor)	$\mathbf{0.4151}$	$\mathbf{0.3037}$	$\mathbf{0.3083}$	0.00559
Wrong-subject Part A	0.3565 ($\Delta = -0.0586$)	0.2821	0.2861	0.00521
Train-pop Part A	0.3467 ($\Delta = -0.0684$)	0.2894	0.2921	0.00594

G Reference-set composition, size, and label-balance check

The mechanism test in Sec. 4.3 swaps the reference set's identity; here we ask how much the *quantity* and *label distribution* of same-subject reference clips matter. Table 12 varies the size $K = |R_s|$ and the per-pain-level balance, holding everything else fixed (strict LOSO, parameterized variant, same-subject Part A clips, query-exclusive).

Table 12: Reference-set composition and size on 5-class LOSO (parameterized variant). $K = 100$ same-subject is the strongest; $K = 20$ random loses 0.06 T1 QWK, balancing across levels recovers most of it, BL1-only is worst.

Reference set	$\lvert R_s \rvert$	T1 pooled QWK	Acc	F1
All same-subject Part A (anchor)	100	$\mathbf{0.4151}$	$\mathbf{0.3037}$	$\mathbf{0.3083}$
Balanced across pain levels	20 (4×5)	0.3987	0.3103	0.3128
Random Part A clips	20	0.3549	0.2818	0.2844
BL1 clips only	20	0.3523	0.2858	0.2883

The full $K = 100$ reference is the strongest. Compressing to $K = 20$ random clips loses -0.06 T1 QWK; explicitly balancing the $K = 20$ across pain levels recovers most of the loss (-0.016 vs anchor). BL1-only $K = 20$ is the worst, indicating that the head benefits from seeing the full pain expressivity range of the test subject in the reference set, not just baseline neutral clips.

Label-leakage check. A residual concern is whether the encoder f_ϕ exploits chance label co-occurrence in the same-subject reference. The reference score head consumes only features and a presence mask, never labels, so any label leakage would have to be implicit. We test the hypothesis directly by forcing every reference set to be balanced across pain levels by construction. If the gain were label-leakage-driven, balancing the label distribution should not hurt; if anything, it should help by removing imbalance. Table 13 reports this on an 8-fold sanity batch.

Table 13: Label-balance test on 5-class (8-fold sanity batch, parameterized). Balanced reference does not improve over the unbalanced anchor; BL1-only further hurts—inconsistent with a label-leakage explanation.

Reference set	T1 pooled QWK	T0 pooled QWK	Acc	F1
Same-subject all 100 clips (anchor)	$\mathbf{0.4993}$	$\mathbf{0.4884}$	$\mathbf{0.3325}$	$\mathbf{0.3325}$
Same-subject label-balanced (20×5)	0.4688 ($\Delta = -0.0305$)	0.4588	0.3038	0.3002
Same-subject BL1-only ($K = 20$)	0.4437 ($\Delta = -0.0556$)	0.4495	0.2963	0.2939

Label balancing does not improve performance and slightly hurts; BL1-only further hurts. Neither direction is consistent with a label-leakage explanation, which would predict that flattening the label distribution helps. The gain comes from richer same-subject context, not from label correlation in the reference set.

H AI4Pain dataset details

AI4Pain (Fernandez-Rojas et al., 2024) is an officially-released multimodal pain video dataset organized as a 3-class classification task (Baseline / Pain_LOW / Pain_HIGH), with a fixed split of 41 training and 12 validation subjects. The published competition protocol freezes a pretrained visual encoder and trains a small head on the train split with no reference-set inference; we follow this

protocol unchanged so our results are directly comparable to the published baseline. The no-reference protocol also explains why calibration cannot be exercised on this dataset (see footnote in Table 2).

I QST-PILOT DATASET DETAILS

QST-Pilot is an unpublished pilot collected by the authors. It is not released: the recordings are identifiable facial video and remain with the authors. The pilot records the facial response to thermode-driven thermal pain stimulation under a protocol modeled on the German Research Network on Neuropathic Pain Quantitative Sensory Testing (DFNS-QST) battery. We use it as a cross-dataset zero-shot transfer target because its elicitation, recording, and per-subject calibration choices substantially overlap with BioVid Part A; this overlap is the methodological basis for treating BioVid→QST-Pilot transfer as a meaningful test of the SAGE encoder rather than a domain-bridging exercise.

Participants. 13 healthy adult volunteers ($\geq$18 yr, no current chronic pain or recent acute pain, no condition or medication affecting facial appearance/expression or cognition, no facial cosmetic alterations, no occluding marks, able to follow study instructions) were enrolled. 12 with usable face video are analyzed; one session was excluded for camera failure. Each participant completed a single session with no follow-up.

Recording. A single GoPro HERO-7 action camera mounted on a fixed-arm dolly captured 4K video at 60 fps, framed to keep the participant's full frontal face in view at a fixed angle. The room was conditioned with a uniform single-color background; participants wore noise-cancelling headphones playing neutral imagery on a monitor to reduce observer effects. Hair was tied back and upper-body accessories (eyewear, necklaces) removed.

Stimulation. Thermal stimuli were delivered with the TSA-II Neurosensory Analyzer (Medoc Ltd.), an FDA-approved computer-controlled QST device, via a 3×3 cm Peltier thermode placed on the volar non-dominant forearm. The operating envelope was clipped to $[-2, 50]\,^\circ\mathrm{C}$ around a $32\,^\circ\mathrm{C}$ baseline, with a $1\,^\circ\mathrm{C/s}$ ramp slope. Each session began with DFNS-style threshold testing: warm/cold detection thresholds (WDT, CDT), heat- and cold-pain thresholds (HPT, CPT), and heat- and cold-pain tolerances (HPTol, CPTol), each measured 3 times with a 15 s interstimulus interval; subjects reported sensation onset via a stop button.

Personalized nociceptive input sequence (NIS). Per-subject pain ranges were calibrated from the threshold tests. Heat NIS bounds were (lowest HPT $-2\,^\circ\mathrm{C}$, highest HPTol); cold NIS bounds were (CPTol, CPT $+2\,^\circ\mathrm{C}$), all clipped to the safety envelope. Each subject received three sequence variants—ascending in $1\,^\circ\mathrm{C}$ steps, descending, and randomized—delivered as low-frequency phasic ramp-and-hold pulses (2–4 s hold; default 4 s; 20 s between-stimulus interval; the long inter-pulse interval suppresses temporal summation). After each pulse the participant rated perceived pain on a 0–100 visual analog scale (VAS); subjects were blinded to thermode setpoints.

Labels. For each clip, the recorded VAS score together with the per-subject NIS index yields a 3-class temperature-cut category (low / moderate / high) anchored to that subject's own calibrated pain range rather than to absolute temperatures. The 3-class derivation is what motivates per-subject Spearman as the primary cross-dataset metric in Sec. 4.2: BioVid uses 5 fixed pain levels, so mapping a 5-class CORAL score to 3 VAS-relative bins requires a non-trivial threshold choice that contaminates accuracy and QWK, whereas Spearman is rank-based and threshold-free.

Why a natural transfer target for BioVid-trained models. Both QST-Pilot and BioVid Part A: (i) elicit pain via controlled contact heat on the forearm with a Peltier thermode operated under a DFNS-style QST protocol; (ii) calibrate per-subject pain ranges from detection/threshold/tolerance measurements so each subject is stimulated relative to their own range; (iii) record frontal full-face video during phasic ramp-and-hold heat pulses with rest intervals; and (iv) keep healthy adults in a controlled lab environment. They differ in label space (VAS-derived 3-class vs. 5 fixed pain levels), per-subject clip count (smaller in QST-Pilot), and camera/background details, but the underlying behavioral target (facial response to thermode-calibrated pain) is the same.