M2E: LIGHTWEIGHT MOTION-TO-EMOTION INFERENCE FOR EDGE-ORIENTED MICRO-EXPRESSION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Micro-expression recognition (MER) requires capturing subtle facial movements with low computational cost and providing interpretable decision evidence for edge deployment, yet existing deep models typically learn latent representations from facial images, incurring substantial resource overhead and offering limited process-level interpretability. In this paper, we propose M2E, a lightweight MER framework that encodes compensated facial dynamics as a compact, semantically explicit 12-dimensional regional representation. M2E suppresses head-motion interference, transforms compensated onset–apex optical flow into FACS-guided regional features, and employs an RBF-SVM with class weighting for efficient and interpretable classification. Under the MEGC2019 Composite Database Evaluation protocol, the lightweight anchor-based configuration achieves **76.94%** UF1 and **79.99%** UAR on the 3DB dataset, outperforming LAENet, the most lightweight existing baseline, by **1.26** and **5.94** percentage points, respectively, with lower computational overhead. When equipped with MEFlowNet, a more computationally intensive learning-based optical-flow compensation network, M2E further achieves **79.80%** UF1 and **83.20%** UAR on the 3DB dataset and matches or surpasses high-capacity expert models on several benchmark metrics. Raspberry Pi experiments, ablation studies, and SHAP analyses further demonstrate its edge applicability, component effectiveness, and process-level interpretability, while evaluations on MMEW and CK+ show its cross-dataset and cross-domain generalization. Code and data are available at `https://anonymous.4open.science/r/M2E-F27`.

1 INTRODUCTION

Micro-expressions are brief, involuntary, and low-intensity facial movements lasting only $1/25$–$1/5$ seconds. Unlike macro-expressions, which are pronounced and can be consciously controlled, micro-expressions are associated with affective states that a person attempts to suppress or conceal (He et al., 2025). Their subtle and localized facial dynamics have motivated applications in clinical analysis, human-computer interaction, and public safety (Ekman & Friesen, 1978; Li & Deng, 2020; Lin et al., 2024). In privacy-sensitive or latency-constrained settings, micro-expression recognition (MER) may need to be performed locally, making computational and memory costs important deployment considerations. At the same time, an emotion label alone provides little information about the facial motion underlying a prediction. Edge-oriented MER therefore calls for models that are computationally efficient while retaining motion evidence that can be inspected at the level of facial regions and movement directions.

Early MER methods relied primarily on handcrafted descriptors. LBP-TOP (Zhao & Pietikainen, 2007) characterizes local spatiotemporal appearance, while optical-flow-based methods such as HOOF (Chaudhry et al., 2009) and Bi-WOOF (Liong et al., 2018) summarize inter-frame motion. These representations are computationally efficient, but fixed motion statistics can retain redundant information and remain sensitive to alignment errors or rigid head displacement. Recent neural approaches learn task-specific representations directly from facial frames or dense motion maps. CNN-based methods such as OFF-ApexNet (Gan et al., 2019) and CNN–LSTM (Wang et al., 2018) capture spatial or spatiotemporal facial patterns, while Transformer-based models such as

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.

SLSTT (Zhang et al., 2022) and HTNet (Wang et al., 2024b) model regional and temporal interactions. This additional representation capacity can come with substantial model size; for example, the CNN–LSTM architecture (Wang et al., 2018) contains approximately 80 million trainable parameters and HTNet (Wang et al., 2024b) approximately 149.57 million, limiting their applicability to resource-constrained edge platforms. Lightweight architectures reduce this cost substantially, but struggle to preserve subtle motion cues, making it difficult to balance efficiency and performance. More importantly, image-based networks do not natively expose region-level motion evidence, and their decisions are commonly interpreted through post-hoc spatial attribution rather than explicit motion variables. Methods based on Action Units (AUs) provide more explicit semantics grounded in the Facial Action Coding System (FACS) (Ekman & Friesen, 1978), but typically require reliable AU annotations and additional AU-learning or feature-fusion modules (Xie et al., 2020).

To address these limitations, we propose M2E, a lightweight MER framework guided by the FACS (Ekman & Friesen, 1978) that represents facial dynamics using compact and semantically grounded regional motion evidence. Specifically, M2E computes optical flow between the onset and apex frames, applies motion compensation to suppress dominant head-motion interference, and aggregates the compensated flow over semantically meaningful facial regions, yielding a compact region-level numerical motion representation. The resulting representation is further processed using sample-wise maximum-absolute normalization and directly classified by a lightweight classifier. By integrating motion compensation and anatomical priors into input construction, M2E preserves discriminative non-rigid motion with explicit region-direction semantics, supporting both interpretable prediction and efficient inference on resource-constrained edge devices.

Experiments on standard MER benchmarks show that M2E achieves competitive recognition performance under two complementary compensation configurations tailored to different efficiency–accuracy requirements. With the computationally efficient anchor-based compensation strategy, M2E achieves the best UF1 and UAR among the compared representative handcrafted and lightweight MER methods. These results demonstrate that encoding dense optical-flow maps as compact regional motion evidence can retain discriminative facial-motion cues while maintaining low computational complexity. On-device evaluation on a Raspberry Pi 5 further shows that M2E maintains competitive recognition performance with modest time and memory consumption, supporting its practical applicability to resource-constrained edge platforms. When the specialized learning-based optical-flow compensation network, MEFlowNet (Zhao et al., 2025), is employed to extract cleaner non-rigid motion, M2E matches or surpasses high-capacity expert models on several benchmark metrics. Furthermore, results on MMEW demonstrate cross-dataset generalization under variations in subjects, acquisition conditions, and data distributions, while results on CK+ show that the learned decision boundary can also transfer from subtle micro-expressions to more pronounced macro-expression dynamics. Our main contributions are summarized as follows:

- We encode facial motion as a compact 12-dimensional regional motion representation with explicit anatomical, directional, and inter-region semantics. Constructed from compensated onset–apex optical flow under FACS guidance, the representation preserves discriminative and interpretable motion evidence without requiring task-specific image feature learning or explicit AU supervision.
- We develop an efficient anchor-based compensation strategy for resource-constrained deployment and adopt the existing MEFlowNet module as an accuracy-oriented configuration for obtaining cleaner non-rigid motion. These complementary configurations provide a practical trade-off between computational efficiency and motion-estimation quality.
- Extensive experiments on standard MER benchmarks demonstrate competitive performance against handcrafted, lightweight, and high-capacity models, while Raspberry Pi deployment confirms the feasibility of edge inference. Ablation studies, generalization experiments on MMEW and CK+, and SHAP analyses further validate the component effectiveness, cross-dataset and cross-domain robustness, and process-level interpretability of M2E.

2 Related Work

2.1 Facial-Motion Representation for MER

Early MER studies primarily relied on handcrafted texture and motion descriptors. LBP-TOP (Zhao & Pietikainen, 2007) and LBP-SIP (Wang et al., 2015) encode appearance variations across spa-

tial and temporal planes, whereas optical-flow-based methods characterize facial deformation more explicitly. OSF (Liong et al., 2016) describes local flow deformation, Bi-WOOF (Liong et al., 2018) constructs magnitude- and strain-weighted orientation histograms, and MDMO (Liu et al., 2016) represents each of 36 predefined, AU-related facial regions using the dominant flow direction and corresponding mean magnitude. Although these descriptors enable efficient classification, their predefined partitions and fixed aggregation rules may limit the preservation of fine-grained motion evidence. In particular, MDMO retains only the dominant directional component within each region and temporally averages the resulting regional descriptors before classification. This compression may suppress weaker or co-occurring movements and does not explicitly preserve the signed horizontal and vertical motion responses that characterize regional displacement. Moreover, fine-grained partitioning can introduce redundant statistics, while residual global displacement caused by head motion or imperfect alignment may still contaminate subtle expression-related motion.

Task-specific neural networks learn discriminative representations from facial images or dense motion maps. CNN-based methods (e.g., OFF-ApexNet, Res-CapsNet) extract spatial or multi-scale features from optical flow (Gan et al., 2019; Shu et al., 2023), while attention-based architectures (e.g., C3DBed, FRL-DGT, HTNet) further enhance contextual modeling via Transformer fusion and hierarchical self-attention (Pan et al., 2023; Zhai et al., 2023; Wang et al., 2024b). Although these approaches improve representation capacity, their high parameter counts hinder deployment on resource-constrained edge devices. Lightweight networks mitigate this burden through architectural simplification. Models such as STSTNet (Liong et al., 2019) and LAENet (Gan et al., 2024) reduce their parameter counts to 1,670 and 843, respectively, by streamlining motion streams and adopting shallow architectures; LAENet additionally employs eye-region masking and macro-expression pretraining to address sample scarcity. However, such simplifications risk discarding subtle facial cues, and dense optical-flow inputs remain vulnerable to rigid-motion interference. Crucially, these models typically lack explicit regional motion evidence, outputting primarily emotion labels. Although AU-based methods introduce action-unit semantics for interpretability, they depend on reliable AU annotations and non-transferable learning modules (Xie et al., 2020; Lei et al., 2021).

In contrast, M2E encodes compensated onset–apex optical flow as a compact 12-dimensional, FACS-guided representation with explicit anatomical, directional, and inter-region semantics. This semantically grounded design suppresses redundant pixel-level information while preserving motion evidence that can be directly associated with named facial regions and deformation patterns, thereby supporting lightweight SVM classification and process-level interpretation without explicit AU supervision.

3 METHOD

3.1 PROBLEM DEFINITION

Given a micro-expression video $\mathcal{V} = \{I_t\}_{t=1}^{T}$, where I_t denotes the facial frame at time t, conventional MER methods typically learn a task-specific mapping f_θ from visual inputs to an emotion label:

$$\widehat{y} = f_\theta(\mathcal{V}), \qquad \widehat{y} \in \mathcal{Y}, \tag{1}$$

where $\mathcal{Y}$ denotes the predefined emotion-label space. In contrast, M2E performs recognition from compact regional motion evidence. Specifically, the optical flow between the onset frame I^o and apex frame I^a is compensated and transformed into a numerical representation according to the FACS-guided facial-region partition $\mathcal{R}$:

$$\mathbf{x} = \Psi(\mathcal{M}(I^o, I^a); \mathcal{R}) \in \mathbb{R}^d, \qquad \mathbf{z} = \mathcal{P}(\mathbf{x}), \qquad \widehat{y} = g_\vartheta(\mathbf{z}), \tag{2}$$

where $\mathcal{M}(\cdot)$ produces the compensated optical-flow field from the onset–apex frame pair, $\Psi(\cdot; \mathcal{R})$ constructs the FACS-guided regional motion representation, d denotes its dimensionality, $\mathcal{P}(\cdot)$ denotes feature processing, and g_ϑ is a lightweight classifier. MER is therefore formulated as a regional motion-to-emotion classification problem, enabling efficient recognition from semantically interpretable motion evidence.

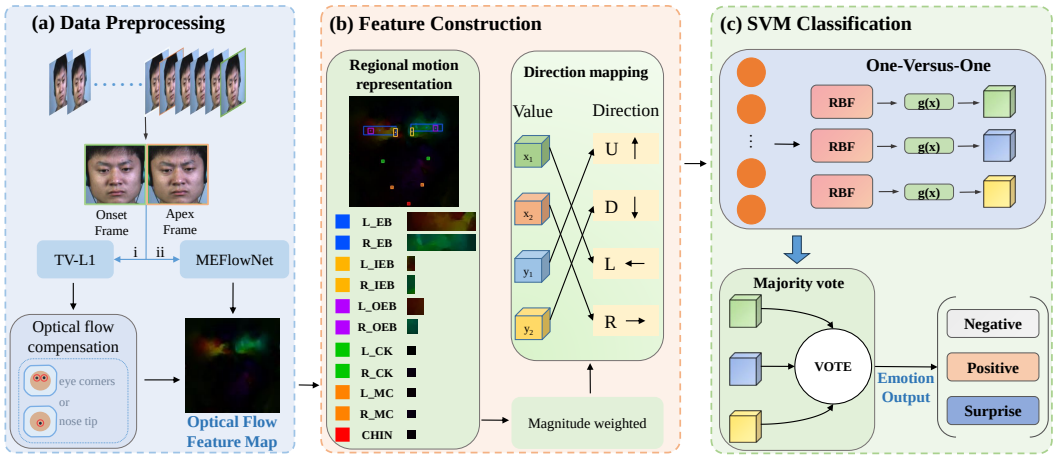


Figure 1: Overview of M2E. The framework consists of three stages: (1) Data preprocessing, which computes onset–apex optical flow by mitigating interference from head motion; (2) Feature construction, which maps compensated flow to facial ROIs and characterizes regional motion through numerical descriptors; and (3) SVM classification, which employs class-weighted one-versus-one RBF classifiers and majority voting to predict the emotion category.

3.2 M2E Overview

Figure 1(a)–(c) presents the overall framework of M2E. The core goal of M2E is to represent subtle facial dynamics as compact, semantically grounded regional motion evidence that can be directly processed by a lightweight classifier for efficient and interpretable recognition. As shown in Figure 1(a), two alternative motion-estimation and compensation strategies are provided. The anchor-based strategy first computes onset–apex TV-L1 optical flow and then removes the dominant head-motion interference estimated from stable facial landmarks. In contrast, the MEFlowNet-based strategy directly processes the onset–apex frame pair and disentangles facial and head-motion flows to obtain a cleaner expression-related motion field. The resulting compensated flow is subsequently mapped to FACS-guided facial regions and aggregated into a normalized numerical representation with explicit regional and directional semantics, as illustrated in Figure 1(b). Finally, the class-weighted RBF-SVM in Figure 1(c) performs one-versus-one classification and determines the emotion category through majority voting.

3.3 Data Preprocessing

To reduce temporal redundancy and computational cost, the onset–apex frame pair is used to summarize the dominant facial motion in each micro-expression sequence. Following the maximum-window strategy adopted in Micro-NesT (He et al., 2025), a shared facial bounding box is determined and applied to both frames, ensuring that the onset and apex faces are represented in a consistent spatial coordinate system. Nevertheless, minor head movement or imperfect alignment may still produce dominant global displacement that interferes with subtle expression-related motion. To address this issue, we introduce two alternative compensation strategies with different trade-offs between computational efficiency and motion estimation quality.

The first strategy performs anchor-based compensation on TV-L1 optical flow (Zach et al., 2007), using flow vectors at stable facial anchors to approximate the dominant global displacement. Specifically, the dense optical-flow field between the cropped onset frame I^o and apex frame I^a is computed as

$$\mathbf{F}^{o\to a} = \text{TV-L}_1\left(I^o, I^a\right) \tag{3}$$

To prevent eye-closure-related motion from contaminating the reference-motion estimate, we use the eye aspect ratio (EAR) (Soukupova & Cech, 2016) to determine the eye state of the onset and apex frames. When eye closure is detected in either frame, the optical-flow vector at the nose tip is used as the reference because it is less affected by periocular deformation. Otherwise, the flow

vectors at the left and right inner eye corners are averaged to reduce local estimation noise. The resulting dominant global displacement is denoted by $\mathbf{r}^{\mathrm{anc}}$, and the compensated flow is computed as

$$\widetilde{\mathbf{F}}^{\mathrm{anc}}(p) = \mathbf{F}^{o\to a}(p) - \mathbf{r}^{\mathrm{anc}}, \qquad \forall p, \tag{4}$$

where p denotes a pixel location.

The second strategy employs the MEFlowNet-based compensation module introduced in MELLM (Zhao et al., 2025) to obtain cleaner expression-related facial motion. Unlike anchor-based compensation, MEFlowNet does not require a precomputed TV-L1 flow field; instead, it directly processes the onset–apex frame pair to estimate the holistic facial-flow and head-motion components:

$$\left(\mathbf{F}^{\mathrm{facial}}, \mathbf{F}^{\mathrm{head}}\right) = \mathrm{MEFlowNet}\left(I^o, I^a\right). \tag{5}$$

The expression-related non-rigid motion is subsequently obtained by subtracting the estimated head-motion component:

$$\widetilde{\mathbf{F}}^{\mathrm{net}}(p) = \mathbf{F}^{\mathrm{facial}}(p) - \mathbf{F}^{\mathrm{head}}(p). \tag{6}$$

3.4 FACS-guided Regional Motion Representation

Dense optical flow retains detailed pixel-level motion but introduces considerable spatial redundancy relative to the localized nature of micro-expressions. Since expression-related deformation is typically concentrated within a small number of anatomically meaningful regions, we use FACS (Ekman & Friesen, 1978) as a spatial prior to construct a compact region-aware motion representation. FACS is employed solely to assist in constructing motion-relevant facial regions and does not require AU-category or AU-intensity annotations.

The compensated optical-flow field is represented as

$$\widetilde{\mathbf{F}}(p) = \left[\widetilde{u}(p), \widetilde{v}(p)\right]^{\top}, \tag{7}$$

where p denotes a pixel location, and $\widetilde{u}(p)$ and $\widetilde{v}(p)$ are its horizontal and vertical components, respectively. Here, $\widetilde{\mathbf{F}}$ denotes either $\widetilde{\mathbf{F}}^{\mathrm{anc}}$ or $\widetilde{\mathbf{F}}^{\mathrm{net}}$, according to the selected compensation strategy.

Because the optical-flow field describes onset-to-apex motion on the same spatial grid as the onset frame, regional motion can be localized directly from onset landmarks. MediaPipe FaceMesh (Lugaresi et al., 2019; Kartynnik et al., 2019) is used to identify the predefined landmark subset $\mathcal{Q}_R$ associated with each anatomical region R. The normalized landmark coordinates are mapped to the optical-flow grid, and an expanded rectangular region of interest (ROI) is constructed as

$$R = \mathrm{Expand}\left(\mathrm{BBox}(\mathcal{Q}_R), \delta\right), \tag{8}$$

where $\mathrm{BBox}(\cdot)$ denotes the minimum enclosing rectangle, and δ is the expansion margin.

Since expanded ROIs exhibit substantial heterogeneity in spatial extent and may contain weak or non-informative motion, we employ magnitude-weighted aggregation without renormalization to achieve area-decoupled motion magnitude estimation. This mechanism ensures that a pronounced regional response arises only when strong and directionally consistent motion is supported by a sufficient proportion of pixels, thereby suppressing spurious high-magnitude artifacts caused by isolated pixel noise. For each $p \in R$, the flow magnitude and regional maximum are computed as

$$m(p) = \sqrt{\widetilde{u}^2(p) + \widetilde{v}^2(p)}, \qquad M(R) = \max_{p\in R} m(p). \tag{9}$$

The magnitude-aware coefficient is then defined as

$$\alpha(p; R) = \frac{m(p)}{M(R) + \epsilon}, \tag{10}$$

where ϵ is a small constant for numerical stability. The regional motion response is subsequently obtained as

$$\boldsymbol{\mu}(R) = \begin{cases} \dfrac{1}{|R|}\displaystyle\sum_{p\in R} \alpha(p; R)\widetilde{\mathbf{F}}(p), & M(R) > \epsilon, \\ \dfrac{1}{|R|}\displaystyle\sum_{p\in R} \widetilde{\mathbf{F}}(p), & M(R) \le \epsilon, \end{cases} \tag{11}$$

where $\boldsymbol{\mu}(R) = [\mu_x(R), \mu_y(R)]^\top$ denotes the horizontal and vertical responses of region R.

To preserve the spatial and directional semantics of facial motion while reducing feature dimensionality, regional responses are encoded according to anatomical and kinematic priors. Leveraging the inherent bilateral symmetry of horizontal facial movements, such as those of the eyebrows, cheeks, and mouth corners, we compute the signed difference between corresponding left-right region pairs. This operation compresses six regional responses into three dimensions; the sign distinguishes outward from inward motion under a fixed coordinate convention, the magnitude quantifies movement intensity, and the bilateral difference captures asymmetry. Since the inner and outer parts of the eyebrows may move with different vertical directions or magnitudes, two dimensions encode their inner-outer vertical-motion differences. The remaining seven dimensions retain the vertical-motion responses of the bilateral eyebrows, cheeks, and mouth corners, together with the chin. Concatenating these components yields the regional motion vector $\mathbf{x} \in \mathbb{R}^{12}$. A detailed specification of each dimension is provided in Table 4 (Appendix A).

To address the variation in absolute scale across samples caused by differences in expression intensity, face size, and optical-flow estimation, we apply sample-wise max-absolute normalization to the compact representation. The normalization factor is defined as

$$\rho = \|\mathbf{x}\|_\infty = \max_{1 \le j \le 12} |x_j| \,, \tag{12}$$

and the normalized representation is obtained by

$$\widehat{\mathbf{x}} = \begin{cases} \dfrac{\mathbf{x}}{\rho}, & \rho > 0, \\ \mathbf{0}, & \rho = 0. \end{cases} \tag{13}$$

3.5 Class-Weighted RBF-SVM Classification

To address the challenges of class imbalance and nonlinear separability in the compact regional motion space, we employ a class-weighted SVM with an RBF kernel (Cortes & Vapnik, 1995). Each feature dimension is standardized using statistics estimated exclusively from the training partition of the corresponding leave-one-subject-out cross-validation (LOSOCV) fold, thereby preventing information leakage. The weight assigned to class c is defined according to its inverse frequency:

$$\omega_c = \frac{N_{\mathrm{tr}}}{K N_{\mathrm{tr},c}}, \tag{14}$$

where N_{tr} is the total number of training samples, K is the number of emotion categories, and $N_{\mathrm{tr},c}$ is the number of training samples belonging to class c. Multiclass recognition follows a one-versus-one decomposition, with the final emotion label determined by majority voting among the pairwise classifiers.

Appendix B provides detailed algorithms for compensated regional motion feature construction and LOSOCV-based lightweight SVM classification.

4 Experiments

4.1 Setup

Datasets and metrics. We evaluate M2E on SMIC (Li et al., 2013), CASME II (Yan et al., 2014), and SAMM (Davison et al., 2016), along with the composite 3DB benchmark, which follows the MEGC2019 (See et al., 2019) Composite Database Evaluation (CDE) protocol. Following the standard MER evaluation practice, we report unweighted F1-score (UF1) and unweighted average recall (UAR) to account for class imbalance. We further evaluate cross-dataset generalization on MMEW (Ben et al., 2022) and cross-domain transferability on the CK+ macro-expression dataset (Lucey et al., 2010), with detailed results provided in Appendix E.8.

Baselines. We compare M2E with three groups of representative MER approaches. The first group includes the handcrafted descriptors LBP-TOP (Zhao & Pietikainen, 2007) and Bi-WOOF (Liong et al., 2018), which characterize facial appearance or motion variations without end-to-end representation learning. The second group consists of the lightweight expert models STSTNet (Liong et al.,

Table 1: Comparison with representative handcrafted and lightweight MER methods on the composite 3DB benchmark and its constituent SMIC, CASME II, and SAMM datasets. "Raspberry Pi" denotes the on-device implementation of the proposed pipeline. The best and second-best results are highlighted in **bold** and underlined, respectively. All results are reported as percentages (%).

Model	3DB		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
Handcrafted methods								
LBP-TOP (Zhao & Pietikainen, 2007)	58.82	57.85	51.23	52.80	70.26	74.26	39.54	41.02
Bi-WOOF (Liong et al., 2018)	62.96	62.27	57.27	58.29	78.05	80.26	52.11	51.39
Lightweight models								
STSTNet (Liong et al., 2019) (1,670 params)	73.53	76.05	68.01	70.13	83.82	86.86	65.88	68.10
LAENet (Gan et al., 2024) (843 params)	75.68	74.05	66.20	65.23	91.01	91.19	68.14	66.20
Proposed method								
Ours (Raspberry Pi)	76.01	78.67	68.23	69.34	91.07	**93.18**	68.53	74.36
Ours (12-D + SVM)	**76.94**	**79.99**	**69.19**	**70.71**	**91.16**	**93.18**	**71.00**	**77.89**

2019) and LAENet (Gan et al., 2024). The third group comprises high-capacity supervised models, including GoogLeNet (Szegedy et al., 2015), VGG16 (Simonyan & Zisserman, 2015), OFF-ApexNet (Gan et al., 2019), SLSTT (Zhang et al., 2022), Res-CapsNet (Shu et al., 2023), FRL-DGT (Zhai et al., 2023), TFT (Wang et al., 2024a), and EMRNet (Liu et al., 2025). These methods provide task-specific references based on image-centric feature learning for evaluating recognition performance.

Implementation details. Unless otherwise specified, the RBF-SVM uses $C_{\mathrm{SVM}} = 0.5$, $\gamma =$ `scale`, a tolerance of 10^{-5}, and `balanced` class weights. Performance is evaluated using LOSOCV, with one subject held out for testing and the remaining subjects used for training. To alleviate the limited sample size of existing micro-expression datasets, label-compatible CK+ samples are additionally incorporated into the training pool, while the subject-independent test folds remain unchanged. Experiments use PyTorch 2.0.1 on an Intel Core i5-10400F CPU and an NVIDIA GeForce RTX 2060 GPU. Edge-device evaluation is conducted on a Raspberry Pi 5 (Raspberry Pi Ltd, 2023) with 2 GB of memory. Further details are provided in Appendix D.

4.2 Main Results

Comparison with handcrafted and lightweight MER methods. As shown in Table 1, M2E achieves the best overall UF1 and UAR using only compact regional motion evidence and a lightweight SVM classifier, validating that the proposed evidence-extraction process preserves discriminative non-rigid facial dynamics without relying on task-specific image-based feature learning. Furthermore, the implementation on a Raspberry Pi 5 maintains competitive performance, with slight deviations from desktop results likely attributable to platform-dependent differences in landmark localization and optical-flow estimation. These results confirm that the extracted motion evidence remains effective on resource-constrained edge devices, supporting practical local inference deployment of M2E. Additional results obtained without CK+ training-pool augmentation are reported in Appendix E.5. The MER-only configuration remains broadly comparable, indicating that the competitiveness of M2E is not solely attributable to supplementary macro-expression data.

Comparison with high-capacity models. As shown in Table 2, the specialized optical-flow compensation network MEFlowNet (Zhao et al., 2025) extracts more discriminative and structured non-rigid motion evidence, thereby enabling M2E to match or surpass high-capacity expert models on several benchmark metrics even with a conventional SVM classifier. Moreover, whereas high-capacity networks generally encode facial dynamics into latent representations that provide limited evidence for their predictions, each feature used by M2E retains explicit facial-region and motion-direction correspondence. This structure enables attribution of model decisions to specific regional movements, providing process-level interpretability alongside competitive recognition performance without relying on a high-capacity image-based recognition network.

Table 2: Comparison with representative high-capacity MER models on the composite 3DB benchmark, SMIC, CASME II, and SAMM. * denotes the use of the MEFlowNet-based optical-flow compensation module from MELLM (Zhao et al., 2025). Best and second-best results are shown in **bold** and underlined, respectively. All results are reported as percentages (%).

Model	3DB		SMIC		CASME II		SAMM	
	UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
GoogLeNet (Szegedy et al., 2015)	55.73	60.49	51.23	55.11	59.89	64.14	51.24	59.92
VGG16 (Simonyan & Zisserman, 2015)	64.25	65.16	58.00	59.64	81.66	82.02	48.70	47.93
OFF-ApexNet (Gan et al., 2019)	71.96	70.96	68.17	66.95	87.64	86.86	54.09	53.92
SLSTT (Zhang et al., 2022)	**81.60**	79.00	74.00	72.00	90.10	88.50	71.50	64.30
Res-CapsNet (Shu et al., 2023)	74.19	74.58	69.00	68.54	81.22	81.31	68.03	68.63
FRL-DGT (Zhai et al., 2023)	81.20	81.10	**74.30**	**74.90**	91.90	90.30	**77.20**	75.80
TFT (Wang et al., 2024a)	81.40	80.10	74.10	71.80	90.70	90.90	70.90	65.60
EMRNet (Liu et al., 2025)	74.68	75.46	65.09	65.96	90.74	89.95	67.82	68.97
Ours (SVM)*	79.80	**83.20**	70.79	72.92	**95.12**	**95.74**	73.90	**81.54**

4.3 Process-level Interpretability

The proposed regional motion representation is semantically structured, with each feature corresponding to an explicit facial region, motion direction, or bilateral deformation. To quantify the class-specific contributions of these features, we employ model-agnostic Kernel SHAP (Lundberg & Lee, 2017), decomposing the difference between a model output and its background expectation into feature-level contributions. Details of the SHAP configuration and the conversion of one-versus-one (OvO) decision margins into class-wise attribution targets are provided in Appendix C. As shown in Figure 2, positive and negative SHAP values indicate that a feature increases or decreases the corresponding class score relative to the background expectation, respectively. For samples predicted as *surprise*, the classifier places the greatest emphasis on the vertical motion of both eyebrows, with larger upward responses generally increasing the class score. Weaker chin responses and mouth-corner deformation provide additional supporting evidence. For the *negative* class, the most influential evidence includes the bilateral horizontal contrasts of the eyebrows and cheeks, together with the vertical responses of the chin, eyebrows, and mouth corners. Their contribution distributions suggest a greater emphasis on inward contraction and downward eyebrow or mouth-corner motion. In contrast, predictions of the *positive* class are primarily supported by upward motion of both mouth corners, while outward cheek and mouth-corner motion contributes complementary evidence.

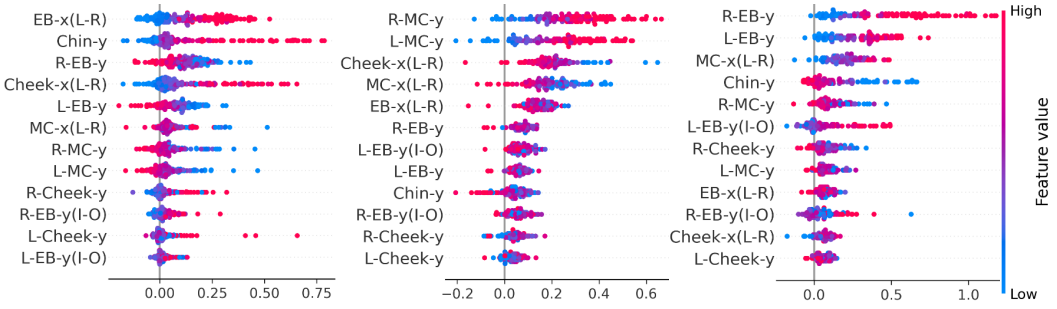


Figure 2: Class-wise SHAP analysis of the regional motion representation on the 3DB dataset. The left, middle, and right panels correspond to samples predicted as *negative*, *positive*, and *surprise*, respectively. Features are ranked by their mean absolute SHAP values. The horizontal axis indicates their contributions to the corresponding class score, while color represents the normalized feature value from low (blue) to high (red). Feature abbreviations and motion-direction conventions are provided in Appendix A.

These class-dependent contribution patterns show that the classifier does not rely on an undifferentiated numerical representation. Instead, its predictions can be traced to explicitly named facial regions, motion directions, and bilateral deformations. Compared with visual attribution methods

that primarily highlight influential image regions, the proposed representation provides semantically grounded motion evidence specifying both where and how the face moves, thereby making the decision process more directly interpretable.

4.4 Ablation Study

We conduct ablation studies on SMIC, CASME II, and SAMM to examine the contributions of optical-flow compensation, magnitude-weighted regional aggregation, and maximum-absolute normalization. The results are reported in Table 3. First, removing anchor-based compensation degrades both UF1 and UAR across all three datasets. The reduction is particularly pronounced on SAMM, where UF1 and UAR decrease by 8.94 and 10.90 percentage points, respectively. This result confirms that suppressing rigid head motion is essential for preserving expression-related non-rigid dynamics. Second, replacing magnitude-weighted aggregation with arithmetic averaging produces slightly higher results on CASME II but lower performance on SMIC and SAMM. This comparison indicates that assigning greater emphasis to stronger regional motion responses tends to improve the discriminability of the resulting evidence, despite dataset-specific variations. Third, removing normalization causes substantial performance degradation across all datasets, demonstrating the necessity of controlling inter-sample scale variation. Although ℓ_2 and absolute-sum normalization remain competitive on SMIC and CASME II, maximum-absolute normalization provides more balanced performance across the three datasets.

Table 3: Ablation studies on SMIC, CASME II, and SAMM. "Mag. Weighting" denotes magnitude-weighted regional aggregation. A check mark indicates that the corresponding proposed component is retained, whereas a cross indicates that it is removed or replaced by the method specified in the first column. All results are reported as percentages (%).

Method Variant	Comp.	Mag. Weighting	Max-Abs Norm.	SMIC		CASME II		SAMM	
				UF1	UAR	UF1	UAR	UF1	UAR
M2E	✓	✓	✓	69.19	70.71	91.16	93.18	71.00	77.89
w/o Compensation	✗	✓	✓	68.70	70.00	90.30	91.85	62.06	66.99
Arithmetic Mean	✓	✗	✓	68.11	69.52	91.81	93.56	69.89	77.52
w/o Normalization	✓	✓	✗	51.83	55.74	81.53	84.72	39.51	47.50
ℓ_2 Normalization	✓	✓	✗	69.94	71.43	91.81	93.56	69.52	76.24
Absolute-Sum Normalization	✓	✓	✗	69.32	70.83	91.81	93.56	69.88	77.16

Appendix E provides additional results and analyses, including confusion matrices, regional motion representation analysis, downstream classifiers, parameter sensitivity, computational efficiency, and SHAP-based interpretations of both correct and misclassified instances.

5 Conclusion

We propose M2E, a lightweight MER framework that represents onset–apex facial dynamics as compact 12-dimensional, FACS-guided regional motion evidence rather than relying on task-specific latent image-feature learning. The anchor-based and MEFlowNet-based configurations provide a practical trade-off between computational efficiency and motion-estimation quality. Experimental results show that the proposed representation retains sufficient discriminative information to outperform representative handcrafted and lightweight methods and achieve competitiveness with high-capacity expert models. Classifier comparisons and feature ablations further indicate that this performance primarily arises from the anatomical and directional structure of the motion evidence, while SHAP analysis enables attributing individual decisions to explicitly named facial regions and motion directions. Raspberry Pi experiments demonstrate the feasibility of resource-constrained local inference, and source-domain evaluations on MMEW and CK+ support robustness to dataset shift and transferability from micro- to macro-expression dynamics. Collectively, these findings support compact structured motion evidence as an efficient, interpretable, and transferable alternative for MER. **Future work** will improve the efficiency of optical-flow estimation and robustness to residual motion noise in unconstrained scenarios.

AI USE STATEMENT

Large language models were used only to assist with language editing, LaTeX formatting, and stylistic refinement of author-provided technical descriptions. They were not used for experimental design, experiment execution, data processing, or the generation of empirical results. All AI-assisted content was manually reviewed and verified by the authors, who take full responsibility for the accuracy and integrity of the manuscript.

ETHICS STATEMENT

This work follows the Code of Ethics. All experiments use publicly available facial-expression benchmarks, including SMIC, CASME II, SAMM, MMEW, and CK+, in accordance with their intended research purposes and applicable terms of use. No new participants were recruited, no additional facial data were collected, and no attempt was made to identify individuals or infer personal attributes beyond the provided expression labels. Results are reported only in aggregate, and no facial images are redistributed. We acknowledge that facial-expression recognition may be affected by demographic bias, annotation subjectivity, and variations in acquisition conditions; therefore, the proposed method should not be used for high-stakes assessment or surveillance without appropriate validation and human oversight.

REPRODUCIBILITY STATEMENT

All datasets used in this work are publicly available benchmarks. The preprocessing procedure, regional motion representation, compensation strategies, SVM configuration, and subject-independent evaluation protocol are described in the main paper. Detailed implementation settings and pseudocode are provided in Appendices D and B; the definitions of the twelve regional motion features are given in Appendix A; and the SHAP configuration and multiclass attribution procedure are documented in Appendix C. Additional ablations, generalization evaluations, and runtime analyses are reported in Appendix E. Code and reproduction instructions are available at `https://anonymous.4open.science/r/M2E-F27`.

REFERENCES

Xianye Ben, Yi Ren, Junping Zhang, Su-Jing Wang, Kidiyo Kpalma, Weixiao Meng, and Yong-Jin Liu. Video-based facial micro-expression analysis: A survey of datasets, features and algorithms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:5826–5846, 2022.

Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

Rizwan Chaudhry, Avinash Ravichandran, Gregory Hager, and Rene Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1932–1939, 2009.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

Corinna Cortes and Vladimir Naumovich Vapnik. Support-vector networks. *Machine Learning*, 20: 273–297, 1995.

Adrian K Davison, Cliff Lansley, Nicholas Costen, Kevin Tan, and Moi Hoon Yap. Samm: A spontaneous micro-facial movement dataset. *IEEE transactions on affective computing*, 9:116–129, 2016.

P. Ekman and W.V. Friesen. *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press, 1978.

Yee Siang Gan, Sze-Teng Liong, Wei-Chuen Yau, Yen-Chang Huang, and Lit-Ken Tan. Off-apexnet on micro-expression recognition system. *Signal Processing: Image Communication*, 74:129–139, 2019.

Yee Siang Gan, Sung-En Lien, Yi-Chen Chiang, and Sze-Teng Liong. Laenet for micro-expression recognition. *The Visual Computer*, 40:585–599, 2024.

Jing He, Yuanhui Xiao, Han Zhang, Jingwen Cai, Li Cai, and Renyang Liu. Micro_nest: multi-scale attention enhanced micro-expression recognition framework. *Expert Syst. Appl.*, 290:128372, 2025.

Yury Kartynnik, Artsiom Ablavatski, Ivan Grishchenko, and Matthias Grundmann. Real-time facial surface geometry from monocular video on mobile gpus, 2019.

Ling Lei, Tong Chen, Shigang Li, and Jianfeng Li. Micro-expression recognition based on facial graph representation learning and facial action unit fusion. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1571–1580, 2021.

Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE transactions on affective computing*, 13:1195–1215, 2020.

Xiaobai Li, Tomas Pfister, Xiaohua Huang, Guoying Zhao, and Matti Pietikäinen. A spontaneous micro-expression database: Inducement, collection and baseline. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6, 2013.

Liang Lin, Song Wang, Jucheng Yang, and Feng Wei. A multi-aware graph convolutional network for driver drowsiness detection. *Knowledge-Based Systems*, 305:112643, 2024.

Sze-Teng Liong, John See, Raphael C.-W. Phan, Yee-Hui Oh, Anh Cat Le Ngo, KokSheik Wong, and Su-Wei Tan. Spontaneous subtle expression detection and recognition based on facial strain. *Signal Processing: Image Communication*, 47:170–182, 2016.

Sze-Teng Liong, John See, KokSheik Wong, and Raphael C-W Phan. Less is more: Micro-expression recognition from video using apex frame. *Signal Processing: Image Communication*, 62:82–92, 2018.

Sze-Teng Liong, Y. S. Gan, John See, Huai-Qian Khor, and Yen-Chang Huang. Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–5, 2019.

Gaqiong Liu, Shucheng Huang, Gang Wang, and Mingxing Li. Emrnet: enhanced micro-expression recognition network with attention and distance correlation. *Artif. Intell. Rev.*, 58:176, 2025.

Yong-Jin Liu, Jin-Kai Zhang, Wen-Jing Yan, Su-Jing Wang, Guoying Zhao, and Xiaolan Fu. A main directional mean optical flow feature for spontaneous micro-expression recognition. *IEEE Transactions on Affective Computing*, 7:299–310, 2016.

Patrick Lucey, Jeffrey F. Cohn, Takeo Kanade, Jason Saragih, Zara Ambadar, and Iain Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pp. 94–101, 2010.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. Mediapipe: A framework for building perception pipelines, 2019.

Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 4768–4777, 2017.

Hang Pan, Lun Xie, and Zhiliang Wang. C3dbed: Facial micro-expression recognition with three-dimensional convolutional neural network embedding in transformer model. *Engineering Applications of Artificial Intelligence*, 123:106258, 2023.

Raspberry Pi Ltd. *Raspberry Pi 5 Product Brief*. Raspberry Pi Ltd, 2023. URL https://pip-assets.raspberrypi.com/categories/892-raspberry-pi-5/documents/RP-008348-DS-6-raspberry-pi-5-product-brief.pdf. Accessed: September 12, 2026.

John See, Moi Hoon Yap, Jingting Li, Xiaopeng Hong, and Su-Jing Wang. Megc 2019 – the second facial micro-expressions grand challenge. In *2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019)*, pp. 1–5, 2019.

Xin Shu, Jia Li, Liang Shi, and Shucheng Huang. Res-capsnet: an improved capsule network for micro-expression recognition. *Multim. Syst.*, 29:1593–1601, 2023.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Tereza Soukupova and Jan Cech. Eye blink detection using facial landmarks. In *21st computer vision winter workshop, Rimske Toplice, Slovenia*, volume 2, pp. 4, 2016.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1–9, 2015.

Su-Jing Wang, Bing-Jun Li, Yong-Jin Liu, Wen-Jing Yan, Xinyu Ou, Xiaohua Huang, Feng Xu, and Xiaolan Fu. Micro-expression recognition with small sample size by transferring long-term convolutional neural network. *Neurocomputing*, 312:251–262, 2018.

Yandan Wang, John See, Raphael C.-W. Phan, and Yee-Hui Oh. Lbp with six intersection points: Reducing redundant information in lbp-top for micro-expression recognition. In *Computer Vision – ACCV 2014*, pp. 525–537, 2015.

Zebiao Wang, Mingyu Yang, Qingbin Jiao, Liang Xu, Bing Han, Yuhang Li, and Xin Tan. Two-level spatio-temporal feature fused two-stream network for micro-expression recognition. *Sensors*, 24 (5), 2024a.

Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayana. Htnet for micro-expression recognition. *Neurocomputing*, 602:128196, 2024b.

Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng. Au-assisted graph attention convolutional network for micro-expression recognition. In *Proceedings of the 28th ACM International Conference on Multimedia*, pp. 2871–2880, 2020.

Wen-Jing Yan, Xiaobai Li, Su-Jing Wang, Guoying Zhao, Yong-Jin Liu, Yu-Hsin Chen, and Xiaolan Fu. Casme ii: An improved spontaneous micro-expression database and the baseline evaluation. *PloS one*, 9:e86041, 2014.

Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime tv-l 1 optical flow. In *DAGM*, pp. 214–223, 2007.

Zhijun Zhai, Jianhui Zhao, Chengjiang Long, Wenju Xu, Shuangjiang He, and Huijuan Zhao. Feature representation learning with adaptive displacement generation and transformer fusion for micro-expression recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pp. 22086–22095. IEEE, 2023.

Liangfei Zhang, Xiaopeng Hong, Ognjen Arandjelović, and Guoying Zhao. Short and long range relation based spatio-temporal transformer for micro-expression recognition. *IEEE Transactions on Affective Computing*, 13:1973–1985, 2022.

Guoying Zhao and Matti Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE transactions on pattern analysis and machine intelligence*, 29:915–928, 2007.

Sirui Zhao, Zhengye Zhang, Shifeng Liu, Xinglong Mao, Shukang Yin, Chaoyou Fu, Tong Xu, and Enhong Chen. Mellm: A flow-guided large language model for micro-expression understanding, 2025.

APPENDIX OVERVIEW

This appendix provides supplementary material that could not be included in the main paper due to space limitations. The appendix is organized as follows:

- Appendix A summarizes the abbreviations, anatomical correspondence, and directional semantics of the regional motion features used in M2E, including the observer-centered coordinate convention and the interpretation of bilateral horizontal-motion differences.
- Appendix B presents the algorithmic procedures for regional motion feature construction and LOSOCV-based lightweight SVM recognition in the proposed M2E.
- Appendix C details the out-of-fold Kernel SHAP configuration, including background construction, coalition sampling, fold-specific prediction pipelines, OvO-to-class score conversion, multiclass attribution, and SHAP visualization.
- Appendix D details the experimental setup, including dataset distributions, evaluation metrics, and the implementation and hyperparameter settings of the Random Forest and XGBoost classifiers.
- Appendix E presents class-wise confusion-matrix analysis, controlled ablations of the regional motion representation, comparisons across downstream classifiers, and sensitivity analysis of the SVM penalty parameter. It further examines CK+ training-pool augmentation, robustness to cross-dataset domain variations, cross-domain applicability on CK+, and strict source-only generalization to MMEW and CK+. Runtime and memory consumption on desktop and Raspberry Pi platforms are also reported, together with instance-level SHAP analyses of representative correct and incorrect predictions to support instance-level interpretability.

A REGIONAL MOTION FEATURE NOTATION

This section defines the abbreviations and coordinate conventions used in the regional motion representation. These definitions establish the anatomical and directional semantics required for interpreting the class-wise and instance-level SHAP results (Lundberg & Lee, 2017). The complete feature definitions are summarized in Table 4. In the main text, the aggregated motion response of region R is denoted by $\boldsymbol{\mu}(R) = [\mu_x(R), \mu_y(R)]^\top$, where $\mu_x(R)$ and $\mu_y(R)$ represent its horizontal and vertical components, respectively. For conciseness, the feature definitions in this appendix use d_x and d_y to denote the same directional responses. Here, *L* and *R* denote the left and right sides from the observer's perspective, while *x* and *y* indicate horizontal and vertical motion. Under the adopted coordinate convention, $d_x > 0$ and $d_x < 0$ represent rightward and leftward motion, whereas $d_y > 0$ and $d_y < 0$ represent upward and downward motion. For a bilateral horizontal feature defined as $\Delta d_x = d_{x,L} - d_{x,R}$, positive values generally indicate inward or convergent motion, while negative values indicate outward or divergent motion. The abbreviations *I* and *O* denote the inner and outer eyebrow regions, and *EB* and *MC* denote the eyebrow and mouth corner, respectively.

Table 4: Abbreviations and semantic definitions of the regional motion features.

Abbreviation	**Description**
EB-x(L-R)	Horizontal-motion difference between the left and right eyebrows
Cheek-x(L-R)	Horizontal-motion difference between the left and right cheeks
MC-x(L-R)	Horizontal-motion difference between the left and right mouth corners
R-EB-y(I-O)	Vertical-motion difference between the inner and outer right eyebrow regions
L-EB-y(I-O)	Vertical-motion difference between the inner and outer left eyebrow regions
R-EB-y	Vertical motion of the right eyebrow
L-EB-y	Vertical motion of the left eyebrow
R-Cheek-y	Vertical motion of the right cheek
L-Cheek-y	Vertical motion of the left cheek
R-MC-y	Vertical motion of the right mouth corner
L-MC-y	Vertical motion of the left mouth corner
Chin-y	Vertical motion of the chin

B ALGORITHM

This section presents the two core algorithms of M2E. Algorithm 1 constructs the regional motion representation by computing onset–apex optical flow, suppressing rigid-motion interference, and aggregating the compensated flow within FACS-guided facial regions (Ekman & Friesen, 1978). It supports two alternative compensation strategies: anchor-based compensation for computationally efficient inference and MEFlowNet-based (Zhao et al., 2025) compensation for extracting cleaner non-rigid facial motion. The resulting regional responses are encoded into a compact motion vector and processed using sample-wise maximum-absolute normalization. Algorithm 2 subsequently performs LOSOCV-based recognition through training-fold standardization and class-weighted RBF-SVM (Cortes & Vapnik, 1995) classification. Multiclass predictions are obtained using the one-versus-one decomposition and majority voting.

Algorithm 1 Regional Motion Feature Construction

Require: Onset frame I^o, apex frame I^a, compensation mode $\kappa \in \{\text{Anchor}, \text{MEFlowNet}\}$
Ensure: Normalized regional motion vector $\widehat{\mathbf{x}} \in \mathbb{R}^{12}$

$\mathcal{Q}^o \leftarrow \text{FaceMesh}(I^o)$
if $\kappa = \text{Anchor}$ **then**
  $\mathbf{F} \leftarrow \text{TV-L}_1(I^o, I^a)$
  $\mathcal{Q}^a \leftarrow \text{FaceMesh}(I^a)$
  $e^o \leftarrow \text{EAR}(\mathcal{Q}^o)$, $e^a \leftarrow \text{EAR}(\mathcal{Q}^a)$
  if $e^o < \tau_{\text{EAR}}$ **or** $e^a < \tau_{\text{EAR}}$ **then**
    $\mathbf{r}^{\text{anc}} \leftarrow \mathbf{F}(\boldsymbol{\ell}^o_{\text{nose}})$
  else
    $\mathbf{r}^{\text{anc}} \leftarrow \dfrac{\mathbf{F}(\boldsymbol{\ell}^o_{\text{li}}) + \mathbf{F}(\boldsymbol{\ell}^o_{\text{ri}})}{2}$
  end if
  $\widetilde{\mathbf{F}}(p) \leftarrow \mathbf{F}(p) - \mathbf{r}^{\text{anc}}, \quad \forall p$
else if $\kappa = \text{MEFlowNet}$ **then**
  $(\mathbf{F}^{\text{facial}}, \mathbf{F}^{\text{head}}) \leftarrow \text{MEFlowNet}(I^o, I^a)$
  $\widetilde{\mathbf{F}} \leftarrow \mathbf{F}^{\text{facial}} - \mathbf{F}^{\text{head}}$
end if
$\mathcal{R} \leftarrow \text{ConstructExpandedROIs}(\mathcal{Q}^o)$
for all $R \in \mathcal{R}$ **do**
  $\boldsymbol{\mu}(R) \leftarrow \text{MagnitudeWeightedAggregate}(\widetilde{\mathbf{F}}, R)$
end for
$\mathbf{x} \leftarrow \text{Encode}\left(\{\boldsymbol{\mu}(R)\}_{R \in \mathcal{R}}\right)$
$\rho \leftarrow \|\mathbf{x}\|_\infty$
if $\rho > 0$ **then**
  $\widehat{\mathbf{x}} \leftarrow \mathbf{x}/\rho$
else
  $\widehat{\mathbf{x}} \leftarrow \mathbf{0}$
end if
return $\widehat{\mathbf{x}}$

Algorithm 2 LOSOCV-Based Lightweight SVM Classification

Input: Labeled feature set $\widehat{\mathcal{X}} = \{(\widehat{\mathbf{x}}, y, u)\}$ and subject set $\mathcal{U}$
Output: Predicted label set $\widehat{\mathcal{Y}}$

$\widehat{\mathcal{Y}} \leftarrow \varnothing$
for $u \in \mathcal{U}$ **do**
  $\left(\mathbf{X}^{\text{tr}}_u, \mathbf{y}^{\text{tr}}_u\right), \left(\mathbf{X}^{\text{te}}_u, \mathbf{y}^{\text{te}}_u\right) \leftarrow \text{LOSO-Split}(\widehat{\mathcal{X}}, u)$
  $(\mathbf{Z}^{\text{tr}}_u, \mathbf{Z}^{\text{te}}_u) \leftarrow \text{Standardize}_{\text{tr}}(\mathbf{X}^{\text{tr}}_u, \mathbf{X}^{\text{te}}_u)$
  $g_u \leftarrow \text{OvO-SVM}(\mathbf{Z}^{\text{tr}}_u, \mathbf{y}^{\text{tr}}_u; \text{RBF}, \text{balanced})$
  $\widehat{\mathbf{y}}_u \leftarrow \text{MajorityVote}(g_u, \mathbf{Z}^{\text{te}}_u)$
  $\widehat{\mathcal{Y}} \leftarrow \widehat{\mathcal{Y}} \cup \widehat{\mathbf{y}}_u$
end for
return $\widehat{\mathcal{Y}}$

C SHAP CONFIGURATION AND MULTICLASS ATTRIBUTION

This section describes the Kernel SHAP (Lundberg & Lee, 2017) configuration used for process-level interpretation, including out-of-fold background construction, multiclass SVM (Cortes & Vapnik, 1995) output formulation, and instance-level attribution.

We employ model-agnostic Kernel SHAP to quantify the contribution of each regional motion feature without modifying the trained classifier. The input to Kernel SHAP is the sample-wise maximum-absolute normalized regional motion vector $\widehat{\mathbf{x}} \in \mathbb{R}^{12}$. For each LOSOCV fold, the training-fold standardizer and the corresponding RBF-SVM are jointly treated as the prediction pipeline to be explained. Specifically, for fold r, the standardized feature vector is obtained as

$$\mathbf{z}^{(r)} = \mathcal{S}^{(r)}(\widehat{\mathbf{x}}), \tag{15}$$

where $\mathcal{S}^{(r)}$ denotes the standardization transformation fitted exclusively on the training partition of fold r. Thus, Kernel SHAP receives $\widehat{\mathbf{x}}$ as its input, while training-fold standardization is performed internally before the feature vector is passed to the SVM.

Explanations are computed in an out-of-fold manner: each test sample is explained using the prediction pipeline fitted in its corresponding LOSOCV fold. Let $\widehat{\mathcal{X}}_{\mathrm{tr}}^{(r)}$ denote the set of normalized 12-dimensional training features in fold r. The corresponding background set satisfies

$$\mathcal{B}^{(r)} \subseteq \widehat{\mathcal{X}}_{\mathrm{tr}}^{(r)}$$

and is uniformly sampled without replacement, with

$$\left|\mathcal{B}^{(r)}\right| = \min\left(50, N_{\mathrm{tr}}^{(r)}\right), \tag{16}$$

where

$$N_{\mathrm{tr}}^{(r)} = \left|\widehat{\mathcal{X}}_{\mathrm{tr}}^{(r)}\right| \tag{17}$$

denotes the number of training samples in fold r. Kernel SHAP uses 200 coalition samples per test instance with $\texttt{l1_reg} = 0$, such that no SHAP-specific feature selection is applied.

For the three-class experiments considered here, the SVM internally learns three one-versus-one (OvO) classifiers. However, class-wise SHAP analysis requires one real-valued output for each emotion category. We therefore apply the OvO-to-OvR transformation of the SVM decision function to convert the pairwise votes and margins into class-aligned decision scores. This transformation changes only the output representation and does not train additional one-versus-rest classifiers.

Let the emotion-label set be

$$\mathcal{C} = \{\mathsf{neg}, \mathsf{pos}, \mathsf{sur}\}, \tag{18}$$

where neg, pos, and sur denote the *negative*, *positive*, and *surprise* classes, respectively. For fold r, the three pairwise margins evaluated by the fold-specific prediction pipeline are

$$h_{\mathsf{neg},\mathsf{pos}}^{(r)}(\widehat{\mathbf{x}}), \qquad h_{\mathsf{neg},\mathsf{sur}}^{(r)}(\widehat{\mathbf{x}}), \qquad h_{\mathsf{pos},\mathsf{sur}}^{(r)}(\widehat{\mathbf{x}}). \tag{19}$$

A non-negative margin supports the first class in the corresponding pair, whereas a negative margin supports the second class. The resulting vote counts are

$$\begin{aligned} V_{\mathsf{neg}}^{(r)}(\widehat{\mathbf{x}}) &= \mathbb{I}\left[h_{\mathsf{neg},\mathsf{pos}}^{(r)}(\widehat{\mathbf{x}}) \geq 0\right] + \mathbb{I}\left[h_{\mathsf{neg},\mathsf{sur}}^{(r)}(\widehat{\mathbf{x}}) \geq 0\right], \\ V_{\mathsf{pos}}^{(r)}(\widehat{\mathbf{x}}) &= \mathbb{I}\left[h_{\mathsf{neg},\mathsf{pos}}^{(r)}(\widehat{\mathbf{x}}) < 0\right] + \mathbb{I}\left[h_{\mathsf{pos},\mathsf{sur}}^{(r)}(\widehat{\mathbf{x}}) \geq 0\right], \\ V_{\mathsf{sur}}^{(r)}(\widehat{\mathbf{x}}) &= \mathbb{I}\left[h_{\mathsf{neg},\mathsf{sur}}^{(r)}(\widehat{\mathbf{x}}) < 0\right] + \mathbb{I}\left[h_{\mathsf{pos},\mathsf{sur}}^{(r)}(\widehat{\mathbf{x}}) < 0\right], \end{aligned} \tag{20}$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

The final emotion label is determined through OvO majority voting:

$$\widehat{y}^{(r)}(\widehat{\mathbf{x}}) = \arg\max_{c \in \mathcal{C}} V_c^{(r)}(\widehat{\mathbf{x}}), \tag{21}$$

with voting ties handled according to the default rule of the SVM implementation.

For class-wise attribution, the accumulated signed pairwise confidences are defined as

$$\begin{aligned} A^{(r)}_{\mathsf{neg}}(\widehat{\mathbf{x}}) &= h^{(r)}_{\mathsf{neg},\mathsf{pos}}(\widehat{\mathbf{x}}) + h^{(r)}_{\mathsf{neg},\mathsf{sur}}(\widehat{\mathbf{x}}), \\ A^{(r)}_{\mathsf{pos}}(\widehat{\mathbf{x}}) &= -h^{(r)}_{\mathsf{neg},\mathsf{pos}}(\widehat{\mathbf{x}}) + h^{(r)}_{\mathsf{pos},\mathsf{sur}}(\widehat{\mathbf{x}}), \\ A^{(r)}_{\mathsf{sur}}(\widehat{\mathbf{x}}) &= -h^{(r)}_{\mathsf{neg},\mathsf{sur}}(\widehat{\mathbf{x}}) - h^{(r)}_{\mathsf{pos},\mathsf{sur}}(\widehat{\mathbf{x}}). \end{aligned} \tag{22}$$

The class-aligned scalar output explained by Kernel SHAP is defined as

$$f^{(r)}_c(\widehat{\mathbf{x}}) = V^{(r)}_c(\widehat{\mathbf{x}}) + \frac{A^{(r)}_c(\widehat{\mathbf{x}})}{3\left(\left|A^{(r)}_c(\widehat{\mathbf{x}})\right| + 1\right)}, \qquad c \in \mathcal{C}. \tag{23}$$

The first term represents the number of pairwise votes received by class c, whereas the second term provides a bounded confidence-based adjustment to the vote count.

Kernel SHAP evaluates the fold-specific vector-valued decision function

$$\mathbf{f}^{(r)}(\widehat{\mathbf{x}}) = \left[f^{(r)}_{\mathsf{neg}}(\widehat{\mathbf{x}}), f^{(r)}_{\mathsf{pos}}(\widehat{\mathbf{x}}), f^{(r)}_{\mathsf{sur}}(\widehat{\mathbf{x}})\right]^\top, \tag{24}$$

and consequently produces a separate set of feature attributions for each emotion category.

For an individual test sample, the waterfall visualization for class c starts from the fold-specific expected score

$$\phi^{(r)}_{0,c} = \mathbb{E}_{\widehat{\mathbf{b}}\sim\mathcal{B}^{(r)}}\left[f^{(r)}_c(\widehat{\mathbf{b}})\right], \qquad c \in \mathcal{C}, \tag{25}$$

where $\widehat{\mathbf{b}}$ denotes a normalized 12-dimensional background feature vector. Because the background set is finite, this expected score can equivalently be written as

$$\phi^{(r)}_{0,c} = \frac{1}{|\mathcal{B}^{(r)}|}\sum_{\widehat{\mathbf{b}}\in\mathcal{B}^{(r)}} f^{(r)}_c(\widehat{\mathbf{b}}). \tag{26}$$

Positive SHAP values increase the corresponding class score relative to its background expectation, whereas negative SHAP values decrease it. After accumulating the contributions of all 12 regional motion features, the class score is approximated by

$$f^{(r)}_c(\widehat{\mathbf{x}}) \approx \phi^{(r)}_{0,c} + \sum_{j=1}^{12}\phi^{(r)}_{j,c}(\widehat{\mathbf{x}}), \qquad c \in \mathcal{C}. \tag{27}$$

In particular, for the predicted class $\widehat{y}^{(r)} = \widehat{y}^{(r)}(\widehat{\mathbf{x}})$, the corresponding additive decomposition is

$$f^{(r)}_{\widehat{y}^{(r)}}(\widehat{\mathbf{x}}) \approx \phi^{(r)}_{0,\widehat{y}^{(r)}} + \sum_{j=1}^{12}\phi^{(r)}_{j,\widehat{y}^{(r)}}(\widehat{\mathbf{x}}). \tag{28}$$

In the waterfall plots, the generic notation $f(x)$ displayed by the SHAP visualization denotes the class-specific score $f^{(r)}_c(\widehat{\mathbf{x}})$ for the class represented by the corresponding panel. Similarly, the displayed baseline $\mathbb{E}[f(X)]$ denotes $\phi^{(r)}_{0,c}$. The generic symbols x and X correspond to the normalized 12-dimensional feature vector $\widehat{\mathbf{x}}$ and its fold-specific background set, respectively. These symbols are distinct from the conventional video-to-label mapping $f_\theta(\mathcal{V})$ introduced in the problem definition.

In an individual waterfall plot, values displayed beside the red and blue bars are SHAP contributions $\phi^{(r)}_{j,c}(\widehat{\mathbf{x}})$, whereas the feature values listed below the plot are the normalized feature components $\widehat{x}_j$. Red bars increase the corresponding class score, whereas blue bars decrease it. In the class-wise summary plots, the horizontal coordinate represents the corresponding SHAP value, while point color represents the corresponding normalized feature value, ranging from low (blue) to high (red). The class-specific SHAP values are computed using the appropriate fold-specific model and background set and are subsequently pooled across LOSOCV folds for visualization.

D EXPERIMENTAL SETUP

D.1 DATASETS AND DISTRIBUTIONS

We evaluate M2E using three widely recognized micro-expression datasets: SMIC (Li et al., 2013), CASME II (Yan et al., 2014), and SAMM (Davison et al., 2016). To ensure consistency and comparability, we adhere to the methodology recommended by the MEGC2019 (See et al., 2019). Specifically, we employ the CDE protocol, which integrates SMIC, CASME II, and SAMM into a unified dataset (3DB), standardizing emotion labels across all three datasets for micro-expression recognition. To systematically assess generalization, we design a two-tier evaluation strategy. First, we evaluate on MMEW (Ben et al., 2022) to test intra-domain generalization against variations in subjects, recording conditions, and motion distributions within the micro-expression domain. Second, to examine generalization beyond this domain, we evaluate on the CK+ macro-expression dataset (Lucey et al., 2010) to determine whether representations of subtle facial dynamics can transfer to more pronounced expressions. Experiments on CK+ follow both three- and five-class protocols. The five-class protocol contains 327 samples distributed across *anger* (45), *disgust* (59), *other* (71), *positive* (69), and *surprise* (83). The distributions of all datasets are summarized in Table 5.

Table 5: Sample distributions of the datasets under the three-class protocol.

Dataset	Negative	Positive	Surprise	Total
SMIC	70	51	43	164
CASME II	88	32	25	145
SAMM	92	26	15	133
3DB	250	109	83	442
MMEW	109	36	89	234
CK+	137	69	83	289

D.2 EVALUATION METRICS

Following standard MER evaluation practice, we adopt unweighted F1-score (UF1) and unweighted average recall (UAR) as the primary metrics because they assign equal importance to each emotion category under class imbalance. Accuracy (ACC) and support-weighted F1-score (F1) are additionally reported where appropriate.

For class c, the class-wise precision, recall, and F1-score are defined as

$$\begin{aligned} \mathrm{Prec}_c &= \frac{TP_c}{TP_c + FP_c}, \\ \mathrm{Rec}_c &= \frac{TP_c}{TP_c + FN_c}, \\ \mathrm{F1}_c &= \frac{2\,\mathrm{Prec}_c\,\mathrm{Rec}_c}{\mathrm{Prec}_c + \mathrm{Rec}_c}. \end{aligned} \tag{29}$$

Given K emotion categories and N_{te} test samples, ACC, weighted F1, UF1, and UAR are computed as

$$\begin{aligned} \mathrm{ACC} &= \frac{\sum_{c=1}^{K} TP_c}{N_{\mathrm{te}}}, \\ \mathrm{F1} &= \sum_{c=1}^{K} \frac{N_{\mathrm{te},c}}{N_{\mathrm{te}}}\,\mathrm{F1}_c, \\ \mathrm{UF1} &= \frac{1}{K}\sum_{c=1}^{K} \mathrm{F1}_c, \\ \mathrm{UAR} &= \frac{1}{K}\sum_{c=1}^{K} \mathrm{Rec}_c, \end{aligned} \tag{30}$$

where TP_c, FP_c, and FN_c denote the true-positive, false-positive, and false-negative counts for class c, respectively, and $N_{\mathrm{te},c} = TP_c + FN_c$ denotes the number of test samples belonging to class c. Thus, F1 weights each class according to its test-set support, whereas UF1 assigns equal weight to all emotion categories.

D.3 Implementation Details

To ensure consistent onset–apex construction across benchmarks, the officially annotated apex frames are used directly for all datasets in which such annotations are available. SMIC is the only evaluated dataset without annotated apex frames; therefore, the apex frame of each SMIC sequence is automatically localized using the D&C-ROIs method (Gan et al., 2019). The detected apex frame is then paired with the corresponding onset frame for subsequent motion estimation and regional feature construction.

For the downstream-classifier comparison using readily available implementations with default or near-default hyperparameters, the linear SVM baseline adopts the same feature processing pipeline, but replaces the RBF kernel with a linear one ($C_{\mathrm{SVM}} = 0.5$), while retaining balanced class weights, optimization tolerance, OvO decomposition, and LOSOCV partitions. The kernel coefficient γ is not applicable in this setting. The Random Forest (Breiman, 2001) contains 200 trees and uses the Gini criterion, bootstrap sampling, and $\sqrt{d}$ candidate features at each split. Its tree depth is unrestricted, with minimum split and leaf sizes of 2 and 1, respectively, and class imbalance is handled using balanced class weights. XGBoost (Chen & Guestrin, 2016) uses the `multi:softprob` objective with multinomial log-loss, 300 boosting rounds, a learning rate of 0.1, and a maximum tree depth of 4. The minimum child weight and split-gain threshold are set to 1 and 0, respectively, while both row- and feature-subsampling ratios are set to 0.8. The L_1 and L_2 regularization coefficients are 0 and 1, respectively, and histogram-based tree construction is employed. Balanced sample weights are computed exclusively from each LOSOCV training fold. Feature standardization is disabled for the random forest and XGBoost, and the random seed is fixed at 51.

E Additional Experiments

This section provides additional quantitative and qualitative analyses of M2E. We first examine class-wise recognition behavior, the anatomical and directional design of the regional motion representation, its discriminability across different downstream classifiers, and sensitivity to the SVM penalty parameter. We then investigate the effect of CK+ training-pool augmentation, robustness to cross-dataset domain variations, cross-domain applicability on CK+, and strict source-only generalization to MMEW and CK+. Finally, we report runtime and memory consumption on desktop and Raspberry Pi platforms and provide instance-level SHAP analyses of representative correct and misclassified predictions.

E.1 Confusion Matrix Analysis

To provide a class-wise view of M2E's recognition behavior under different compensation configurations, Figure 3 presents the row-normalized confusion matrices on SMIC (Li et al., 2013), CASME II (Yan et al., 2014), and SAMM (Davison et al., 2016). Both configurations produce pronounced diagonal structures, indicating that the compact regional motion representation preserves coherent class-discriminative patterns across datasets rather than favoring only a particular compensation strategy. CASME II exhibits the clearest class separation, suggesting that its expression-related regional motion patterns are comparatively well aligned with the three-category label space. By contrast, SMIC and SAMM show greater class-dependent ambiguity, although their confusion patterns differ: under anchor-based compensation, negative samples in SMIC are most frequently confused with surprise, whereas SAMM exhibits more pronounced confusion between the negative and positive categories. MEFlowNet-based (Zhao et al., 2025) compensation generally concentrates predictions more strongly along the diagonal, most notably by improving the separation of surprise on SMIC and reducing the confusion of positive samples with negative on SAMM. Nevertheless, the anchor-based configuration remains competitive and performs better for certain class–dataset combinations. Overall, these patterns demonstrate that the proposed representation retains stable class

structure under both lightweight and learning-based compensation, while the remaining off-diagonal errors exhibit dataset- and category-specific distributions.

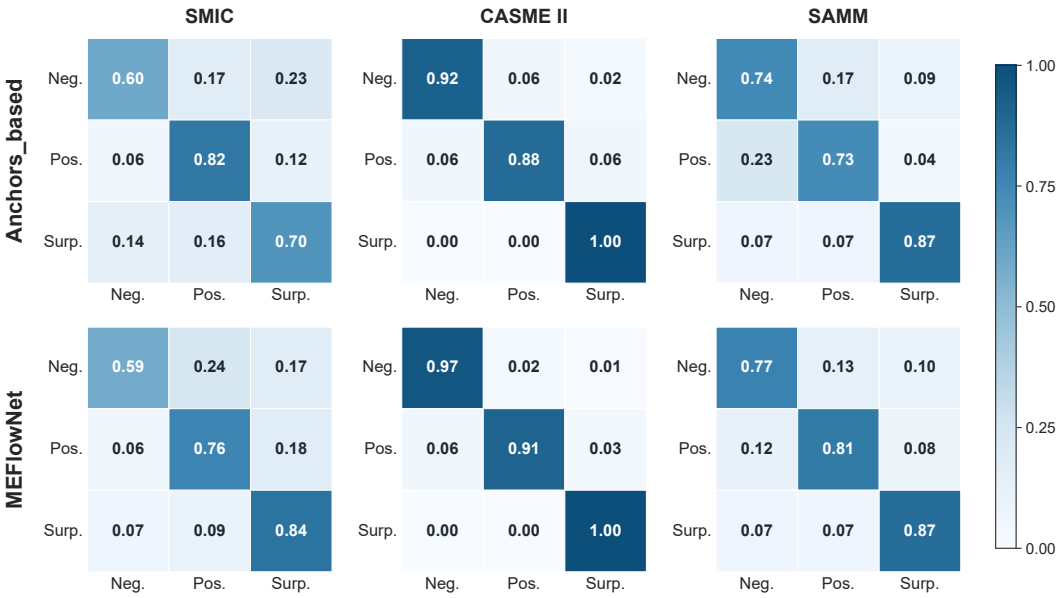


Figure 3: Row-normalized confusion matrices on SMIC, CASME II, and SAMM. The top and bottom rows show the results obtained using anchor-based and MEFlowNet-based compensation, respectively. Rows denote ground-truth classes, while columns denote predicted classes. The emotion labels Neg., Pos., and Surp. denote Negative, Positive, and Surprise.

E.2 Analysis of the Regional Motion Representation

To assess whether the proposed feature design achieves an effective balance between compactness and cross-dataset robustness, we conduct controlled ablations on SMIC (Li et al., 2013), CASME II (Yan et al., 2014), and SAMM (Davison et al., 2016) while retaining identical anchor-based compensation, normalization, RBF-SVM (Cortes & Vapnik, 1995), and LOSOCV settings. As shown in Table 6, the complete 12-dimensional representation does not necessarily maximize every individual dataset–metric pair, but provides the strongest and most consistent overall performance across the three datasets.

The anatomical ablations reveal that the regional components contribute complementary rather than interchangeable information. Eyebrow motion is particularly important for CASME II and SAMM, whereas mouth-corner and chin responses provide complementary lower-face evidence. The effect of cheek removal varies across datasets: it produces small improvements in some settings but substantially degrades performance on SAMM. This variation suggests that the relevance of an individual region depends on the facial characteristics and acquisition conditions of each dataset. Jointly retaining these regions therefore reduces reliance on dataset-specific motion patterns and improves the overall stability of the representation.

The semantic-component ablations further clarify the structure of the encoded evidence. Vertical regional motion constitutes the principal discriminative component, as its removal produces the most pronounced degradation across all datasets. Bilateral horizontal contrasts provide complementary information about relative facial contraction and expansion and are especially important on SAMM. The inner–outer eyebrow contrasts have a smaller effect but preserve localized eyebrow deformation that cannot be fully represented by region-level vertical motion alone. These findings indicate that the representation benefits from combining absolute directional responses with relative bilateral and intra-region deformation.

The two higher-dimensional alternatives show that increasing feature dimensionality does not necessarily improve robustness. Although *Independent L/R* and *Difference + Bilateral Mean* yield

modest gains on SMIC, both are less consistent on CASME II and SAMM. Separately retaining bilateral components or introducing their common-mode responses may preserve redundant, subject-dependent, or residual rigid-motion variation. In contrast, the proposed difference-based encoding emphasizes relative deformation while suppressing shared displacement. Overall, the 12-dimensional design retains the most informative anatomical and directional cues without unnecessary feature expansion, providing a favorable compromise between representational compactness and performance stability across datasets.

Table 6: Ablation analysis of the anatomical and semantic components of the regional motion representation. *Independent L/R* replaces each bilateral horizontal difference with two side-specific horizontal components, whereas *Difference + Bilateral Mean* retains the original differences and appends the bilateral horizontal mean, $(d_{x,L} + d_{x,R})/2$, for the eyebrow, cheek, and mouth-corner regions. All remaining features are unchanged. "Dim." denotes feature dimensionality, and all results are reported as percentages (%).

Representation Variant	Dim.	SMIC		CASME II		SAMM	
		UF1	UAR	UF1	UAR	UF1	UAR
Full representation	12	69.19	70.71	91.16	93.18	71.00	77.89
w/o Eyebrow	7	67.16	68.56	81.50	87.59	60.12	68.54
w/o Cheek	9	70.37	71.43	91.50	91.94	64.87	71.07
w/o Mouth Corner	9	69.40	70.60	86.17	88.92	66.21	75.73
w/o Chin	11	66.73	68.50	88.78	90.14	66.51	70.88
w/o Bilateral Contrast	9	68.95	70.42	87.76	90.05	60.40	71.89
w/o Inner-Outer Contrast	10	68.52	69.93	91.15	93.84	68.24	74.94
w/o Absolute Vertical Motion	5	59.12	60.16	76.19	77.40	58.44	61.26
Independent L/R	15	70.63	71.90	90.00	91.76	70.51	76.03
Difference + Bilateral Mean	15	70.52	71.60	90.65	92.14	69.06	73.80

E.3 Discriminability across Downstream Classifiers

To verify that the recognition performance primarily arises from the discriminative information encoded by the proposed regional motion representation rather than from a particular classifier, we retain the same input features and LOSOCV splits and replace the RBF-SVM (Cortes & Vapnik, 1995) with a linear SVM, Random Forest (Breiman, 2001), and XGBoost (Chen & Guestrin, 2016). As shown in Table 7, the linear SVM achieves the highest UF1 and UAR on SMIC under both compensation configurations and remains competitive with the RBF-SVM on CASME II and SAMM. This result is particularly informative because a linear classifier has limited capacity to construct complex decision boundaries, indicating that the compact regional representation already provides substantial class separability. The RBF-SVM performs better on CASME II (Yan et al., 2014) and SAMM (Davison et al., 2016), suggesting that its nonlinear kernel more effectively captures residual interactions among regional motion components. Nevertheless, the fact that Random Forest and XGBoost achieve viable recognition from the same fixed representation without task-specific image-feature learning confirms that the discriminative information is intrinsic to the regional motion encoding itself.

Overall, these results show that the recognition capability is not an artifact of the selected RBF-SVM. Instead, the proposed representation encodes discriminative anatomical and directional motion information that can be exploited by different classifier families, while linear and RBF SVMs are better suited to the present low-dimensional feature space.

E.4 Analysis of the Penalty Parameter C_{SVM}

To assess the robustness of the classifier configuration, Table 8 reports the recognition performance across different values of the RBF-SVM penalty parameter C_{SVM}, which regulates the balance between margin maximization and empirical error. The influence of C_{SVM} is non-monotonic and varies across datasets and compensation configurations, reflecting differences in their sample dis-

Table 7: Evaluation of the discriminability of the proposed regional motion representation across different downstream classifiers. All classifiers use identical regional motion features and LOSOCV splits; only the downstream decision function is changed. For each dataset within each compensation configuration, the best UF1 and UAR values among the evaluated classifiers are highlighted in bold. All results are reported as percentages (%).

Compensation	Classifier	SMIC		CASME II		SAMM	
		UF1	UAR	UF1	UAR	UF1	UAR
Anchor-based	RBF-SVM	69.19	70.71	**91.16**	**93.18**	**71.00**	**77.89**
	Linear SVM	**70.62**	**71.90**	89.94	92.51	69.43	77.16
	Random Forest	67.66	66.84	86.28	83.86	65.27	67.56
	XGBoost	67.94	68.21	86.04	85.39	67.62	71.04
MEFlowNet	RBF-SVM	70.79	72.92	**95.12**	**95.74**	**73.90**	**81.54**
	Linear SVM	**71.95**	**73.99**	92.35	92.61	71.81	78.59
	Random Forest	65.88	65.23	86.00	83.70	66.99	66.89
	XGBoost	71.37	72.20	87.44	86.73	62.67	64.72

tributions and class boundaries. Importantly, increasing C_{SVM} does not yield systematic improvements, and comparable performance is maintained over a moderate parameter range. This indicates that the discriminative capability of the regional motion representation does not depend on a narrowly optimized penalty value.

Table 8: Sensitivity of recognition performance to the RBF-SVM penalty parameter C_{SVM} under anchor-based and MEFlowNet-based compensation. All remaining feature-construction and classification settings are fixed. The best result for each dataset and compensation configuration is highlighted in bold. All results are reported as percentages (%).

Compensation	C_{SVM}	SMIC		CASME II		SAMM	
		UF1	UAR	UF1	UAR	UF1	UAR
Anchor-based	0.1	**69.99**	**71.43**	88.91	92.05	66.67	76.07
	0.5	69.19	70.71	91.16	93.18	71.00	77.89
	0.8	69.80	71.18	**91.81**	**93.56**	71.89	78.25
	1.2	69.78	71.18	91.03	92.52	71.89	78.25
	1.5	69.78	71.18	91.03	92.52	**72.81**	**78.61**
MEFlowNet-based	0.1	69.01	70.84	92.10	92.90	71.79	78.95
	0.5	70.79	72.92	**95.12**	**95.74**	**73.90**	**81.54**
	0.8	**72.54**	**74.35**	92.90	93.28	73.00	80.25
	1.2	71.94	73.87	92.90	93.28	73.41	80.25
	1.5	71.37	73.40	92.48	93.28	73.41	80.25

E.5 Effect of CK+ Training-Pool Augmentation

The limited scale of existing micro-expression datasets may hinder the estimation of robust classification boundaries. To examine whether heterogeneous facial-motion samples can alleviate this limitation, we map the CK+ (Lucey et al., 2010) macro-expression labels to the same three emotion categories and add the corresponding samples exclusively to the training partition of each LOSOCV fold. The original subject-independent micro-expression test folds and all other evaluation settings remain unchanged.

As shown in Table 9, the effectiveness of CK+ augmentation depends strongly on the consistency of the motion representation across expression domains. Without compensation, incorporating CK+ is generally detrimental, showing that additional samples cannot compensate for mismatched motion distributions. Anchor-based compensation partially reduces this discrepancy but produces dataset-dependent changes, suggesting that a global displacement estimated from fixed facial anchors does not fully reconcile the different motion scales of macro- and micro-expressions. By

contrast, MEFlowNet-based compensation consistently benefits all three datasets, indicating that its learned separation of facial and head motion yields more compatible non-rigid evidence across domains. The results therefore show that the utility of heterogeneous training data is governed primarily by representation compatibility rather than training-set size alone.

The superscript $*$ denotes results obtained on the Raspberry Pi 5 (Raspberry Pi Ltd, 2023). Under the on-device anchor-based configuration, the MER-only training pool remains slightly more effective than its CK+-augmented counterpart. Because the regional features are independently constructed on the edge platform, this behavior may reflect the combined sensitivity to cross-domain motion-scale differences and implementation-level numerical variation. Nevertheless, the competitive MER-only results indicate that the on-device performance of M2E exhibits limited dependence on supplementary macro-expression data, permitting a lower-cost deployment when CK+ feature construction and storage are omitted.

Table 9: Effect of CK+ training-pool augmentation under no compensation, anchor-based compensation, and MEFlowNet-based compensation. "MER + CK+" adds label-compatible CK+ samples to the training partition of each LOSOCV fold, whereas "MER only" uses only micro-expression training samples. The original subject-independent test folds remain unchanged. The superscript $*$ denotes results obtained on the Raspberry Pi 5; all other results are obtained on the desktop platform. All results are reported as percentages (%).

Compensation	Training Pool	SMIC		CASME II		SAMM	
		UF1	UAR	UF1	UAR	UF1	UAR
No compensation	MER + CK+	68.70	70.00	90.30	91.85	62.06	66.99
	MER only	71.15	72.73	90.07	93.47	63.87	70.71
Anchor-based	MER + CK+	69.19	70.71	91.16	93.18	71.00	77.89
	MER only	69.16	70.30	92.48	93.94	70.03	76.03
	MER + CK+*	68.23	69.34	91.07	93.18	68.53	74.36
	MER only*	69.24	70.42	92.10	93.94	69.37	75.66
MEFlowNet-based	MER + CK+	70.79	72.92	95.12	95.74	73.90	81.54
	MER only	68.95	71.49	93.57	93.66	72.19	79.89

E.6 Robustness to Cross-Dataset Domain Variations

To evaluate robustness to cross-dataset domain variations, we compare an MMEW-only (Ben et al., 2022) training pool with a shared training pool that additionally incorporates SMIC (Li et al., 2013), CASME II (Yan et al., 2014), and SAMM (Davison et al., 2016). The MMEW LOSOCV partitions are kept fixed across both settings, ensuring that the observed differences result from introducing heterogeneous samples into the training pool.

As shown in Figure 4, the shared training pool produces contrasting transfer behavior across methods. STSTNet (Liong et al., 2019) exhibits negative transfer, indicating that increasing the amount of training data does not necessarily improve recognition when the learned representation is sensitive to cross-dataset distribution discrepancies. In contrast, both configurations of M2E exhibit positive transfer, showing that the regional motion representation can exploit additional micro-expression samples despite variations in subjects, acquisition conditions, and dataset distributions. The consistency of this trend under both anchor-based and MEFlowNet-based compensation further suggests that the observed robustness is not specific to a single motion-compensation strategy. MEFlowNet-based compensation achieves the strongest overall performance, indicating that cleaner non-rigid motion evidence provides an additional benefit when heterogeneous datasets are combined.

E.7 Cross-Domain Applicability on CK+

To examine whether the proposed regional motion representation remains effective beyond the micro-expression domain, we apply M2E to the CK+ (Lucey et al., 2010) macro-expression dataset under three- and five-class settings. The feature-construction procedure and lightweight SVM architecture remain unchanged, and evaluation follows the LOSOCV protocol. In each fold, all samples

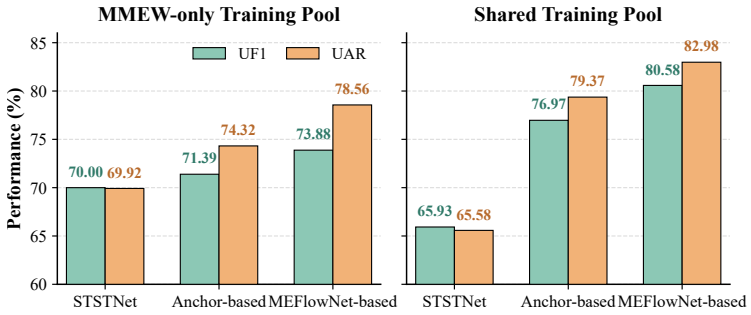


Figure 4: Robustness to cross-dataset domain variations on MMEW. "MMEW only" uses the remaining MMEW subjects for training in each evaluation fold, whereas "Shared" additionally incorporates SMIC, CASME II, and SAMM into the training pool. The held-out MMEW test samples remain unchanged across both settings.

from one CK+ subject are reserved for testing, while those from the remaining subjects are used for training. As shown in Table 10, the compact representation remains highly discriminative for the larger facial deformations of macro-expressions. Under the three-class protocol, MEFlowNet-based (Zhao et al., 2025) compensation surpasses the high-capacity Micro-NesT (He et al., 2025) baseline, while the uncompensated configuration remains competitive. The advantage of motion compensation becomes more pronounced under the five-class protocol, suggesting that cleaner non-rigid motion evidence is increasingly important when distinguishing more fine-grained expression categories. These results indicate that the explicit anatomical, directional, and inter-region semantics of the proposed representation preserve class-relevant deformation patterns across different expression intensities. Consequently, the same compact encoding remains discriminative for the larger and more diverse movements of macro-expressions, allowing a lightweight SVM to perform effective recognition without redesigning the representation.

Table 10: Cross-domain applicability on CK+ under three- and five-class settings using LOSOCV. "None" denotes recognition without optical-flow compensation, whereas "MEFlowNet" denotes MEFlowNet-based compensation. The regional feature construction and SVM architecture remain unchanged across settings. All results are reported as percentages (%).

Setting	Model	Compensation	CK+	
			F1	ACC
Three-class	Micro-NesT (He et al., 2025)	–	96.65	96.64
	Ours (12-D + SVM)	None	95.18	95.16
	Ours (12-D + SVM)	MEFlowNet	98.28	98.27
Five-class	Ours (12-D + SVM)	None	87.74	87.77
	Ours (12-D + SVM)	MEFlowNet	96.01	96.02

E.8 Analysis of Cross-Dataset and Cross-Domain Generalization

To provide a strict source-only evaluation, we train M2E exclusively on the composite 3DB benchmark and directly evaluate the resulting classifier on MMEW (Ben et al., 2022) and CK+ (Lucey et al., 2010). No target-dataset samples are used for model fitting, feature standardization, hyperparameter selection, or adaptation. In particular, the CK+ training-pool augmentation adopted in the main benchmark experiments is disabled, and the feature-processing statistics estimated from 3DB are applied directly to each target dataset. For each target dataset, we evaluate two matched feature-construction configurations. In the anchor-based configuration, both the 3DB training samples and the target samples are processed using fixed-anchor compensation. In the MEFlowNet-based (Zhao et al., 2025) configuration, MEFlowNet compensation is applied consistently to both domains. This protocol avoids introducing differences in feature construction between the source

and target domains, allowing the evaluation to focus on whether the learned class structure transfers across datasets and expression domains.

As shown in Table 11, both configurations retain substantial recognition capability when transferred from 3DB to MMEW without target-specific calibration. MEFlowNet-based compensation provides consistent improvements, indicating that cleaner non-rigid motion evidence enhances robustness to changes in subjects, acquisition conditions, and dataset distributions. Nevertheless, the anchor-based configuration remains competitive, supporting its use when computational efficiency is prioritized. A more pronounced difference is observed on CK+. The comparatively weak anchor-based results suggest that a single global displacement estimated from sparse facial anchors may be insufficient to separate rigid motion from the larger and more spatially heterogeneous deformations of macro-expressions. The strong performance obtained with MEFlowNet-based compensation suggests that separating expression-related deformation from head motion produces a motion representation that transfers effectively from subtle micro-expressions to larger macro-expression movements. Since CK+ is excluded entirely from training and parameter estimation, this result reflects source-only cross-domain transfer rather than adaptation to the target dataset. Collectively, the results support the robustness of the regional motion representation to cross-dataset variation within MER and its applicability beyond the micro-expression domain.

Table 11: Source-only generalization from 3DB to MMEW and CK+. The classifier and feature-processing statistics are learned exclusively from 3DB, without target-dataset training, hyperparameter selection, or adaptation. For each configuration, the same compensation strategy is applied to the source and target samples. All results are reported as percentages (%).

Target Dataset	Compensation	ACC	F1	UF1	UAR
MMEW	Anchor-based	77.78	78.11	75.63	78.38
	MEFlowNet-based	79.06	79.21	76.70	79.57
CK+	Anchor-based	78.55	78.53	77.50	77.43
	MEFlowNet-based	94.12	94.10	94.19	95.71

E.9 Analysis of Runtime and Memory

To evaluate the computational feasibility of M2E on resource-constrained edge devices, we deploy the complete pipeline on a Raspberry Pi 5 (Raspberry Pi Ltd, 2023) with 2 GB of memory. The pipeline comprises facial cropping, onset–apex optical-flow estimation, fixed-anchor compensation, and regional motion encoding. In Table 12, feature-construction time denotes the complete processing interval from loading the onset–apex images to obtaining the 12-dimensional regional motion vector, while memory consumption denotes the corresponding per-sample peak memory increase.

Table 12: Runtime and memory requirements of M2E on the Raspberry Pi 5 and desktop platforms. Face-cropping time is measured over each complete dataset. Feature-construction time covers the processing from onset–apex image loading to generation of the 12-dimensional regional motion vector, including optical-flow estimation, anchor-based compensation, and regional aggregation. Memory is reported as the average per-sample peak increase in process resident memory.

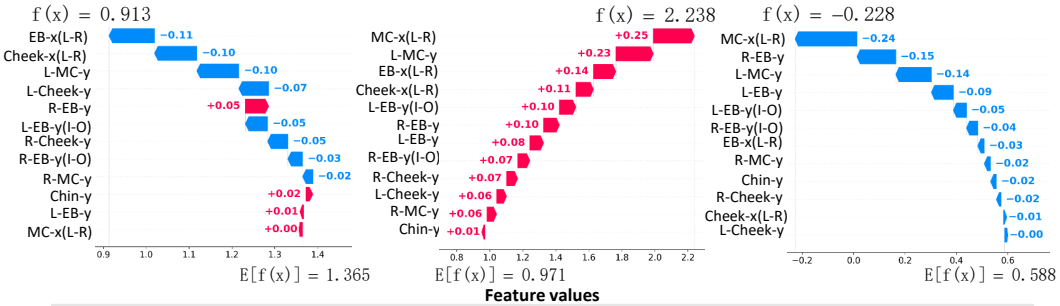
Platform	Dataset	Samples	Face Cropping Total Time (s)	Feature Construction Time/Sample (s)	Peak Memory Increase per Sample (MB)
Raspberry Pi	SMIC	164	14.370	1.13	10.78
	CASME II	145	12.506	1.81	16.43
	SAMM	133	22.918	2.80	27.02
	CK+	289	32.631	5.11	13.06
Desktop	SMIC	164	8.458	0.40	8.72
	CASME II	145	7.705	0.47	11.76
	SAMM	133	12.622	0.77	23.70
	CK+	289	20.169	1.26	8.77

On the Raspberry Pi, feature construction remains within a few seconds per sample, with the peak memory increase remaining below 28 MB across all datasets. CK+ requires the longest processing time, which is consistent with its larger onset–apex deformations and the resulting increase in optical-flow estimation complexity. On the desktop platform, the total processing time across all four datasets, including facial cropping and regional feature construction, is approximately 649 s, which is lower than the approximately 1167 s overall processing time reported for LAENet (Gan et al., 2024). When CK+ training-pool augmentation is disabled, feature construction is restricted to the three constituent datasets of 3DB, eliminating the additional CK+ preprocessing cost and reducing the total training-pool preparation time to approximately 265 s. As reported in Table 9, the resulting MER-only configuration continues to outperform LAENet across the constituent datasets, indicating that the efficiency–accuracy advantage exhibits limited dependence on supplementary macro-expression data. Although cross-study runtime comparisons are affected by differences in hardware and software environments, the combined runtime and recognition results support a favorable efficiency–accuracy trade-off.

Once the regional motion vectors have been constructed, the complete SVM training and testing procedure on the Raspberry Pi requires only 0.0377 s. The runtime measured from the entry of the main function is 2.26 s, including initialization, feature loading, and result processing. Thus, the classification overhead is negligible relative to feature construction. Overall, the modest memory footprint demonstrates that M2E can be executed locally on a low-power, small-form-factor edge computer.

E.10 Instance-level Decision Interpretation

Because each dimension of the regional motion representation retains explicit anatomical and directional semantics, the contribution of individual features can be interpreted in terms of concrete facial movements. Figure 5 presents the Kernel SHAP (Lundberg & Lee, 2017) explanation of a correctly classified sample, for which both the ground-truth and predicted labels are *positive*. Each waterfall plot provides an additive decomposition of a class-aligned decision score from its fold-specific expected value to the output obtained for this sample.



EB-x(L-R)	Cheek-x(L-R)	MC-x(L-R)	R-EB-y(I-O)	L-EB-y(I-O)	R-EB-y	L-EB-y	R-Cheek-y	L-Cheek-y	R-MC-y	L-MC-y	Chin-y
-0.090	-0.208	-0.824	-0.086	-0.039	0.037	-0.029	-0.002	0.187	0.207	1.000	0.114

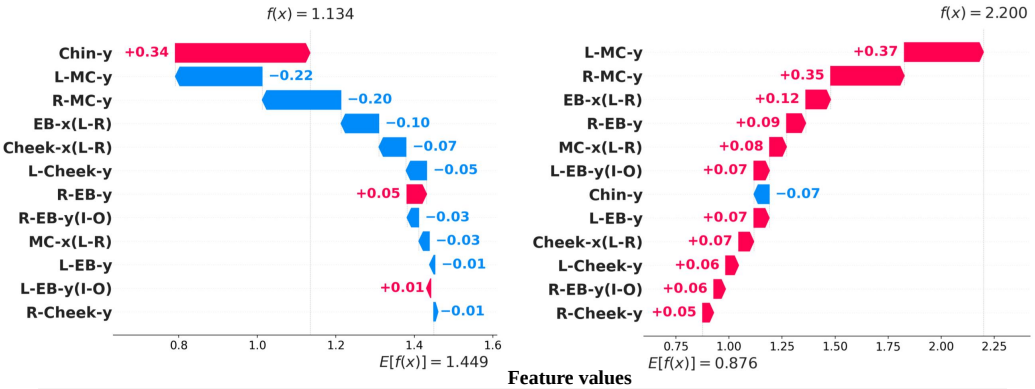
Figure 5: Instance-level Kernel SHAP explanation for a correctly classified *positive* sample. The left, middle, and right panels explain the class-aligned decision scores for *negative*, *positive*, and *surprise*, respectively. Each waterfall plot decomposes $F_c(\mathbf{x})$ from its fold-specific expected value $\mathbb{E}[F_c(\mathbf{X})]$ into additive regional feature contributions. Red bars increase the corresponding class score, whereas blue bars decrease it. The final emotion label is obtained through OvO voting, while the continuous scores shown here are used for SHAP attribution. Feature abbreviations are defined in Appendix A.

For the *positive* class, the decision score increases from 0.971 to 2.238. The largest supporting contribution arises from the upward motion of the left mouth corner (`L-MC-y`), followed by the bilateral horizontal contrasts of the mouth corners and cheeks. Under the adopted coordinate convention, the negative values of these bilateral features indicate outward motion, forming a coherent pattern of mouth-corner elevation and lateral facial expansion. The eyebrow and chin responses provide further, albeit smaller, positive contributions.

The same regional motion pattern provides negative evidence for the competing categories. The *negative* score decreases from 1.365 to 0.913, primarily because of the left mouth-corner motion and the bilateral eyebrow and cheek responses. The *surprise* score decreases more markedly from 0.588 to -0.228, with mouth-corner expansion, left mouth-corner elevation, and eyebrow motion producing the dominant negative contributions. The resulting attribution pattern therefore explains both the evidence supporting the correct *positive* prediction and the evidence suppressing the alternative classes. This example demonstrates how the semantically structured representation enables an individual decision to be traced to explicit facial regions and motion directions.

E.11 Analysis of Instance-level Misclassification

Figure 6 presents the local explanation of a representative misclassified sample whose ground-truth label is *negative*, whereas the predicted label is *positive*. Visual inspection of the corresponding onset–apex frames suggests a pronounced pouting-like deformation, which gives rise to substantial upward displacement of both mouth corners in the extracted motion evidence. Accordingly, the *positive* score increases from its expected value of 0.876 to 2.200. The left and right mouth-corner vertical features, with normalized values of 1.000 and 0.702, contribute $+0.37$ and $+0.35$, respectively, accounting for the dominant positive evidence.



EB-x(L-R)	Cheek-x(L-R)	MC-x(L-R)	R-EB-y(I-O)	L-EB-y(I-O)	R-EB-y	L-EB-y	R-Cheek-y	L-Cheek-y	R-MC-y	L-MC-y	Chin-y
-0.035	-0.055	-0.151	-0.049	-0.036	-0.01	0.012	0.139	0.231	0.702	1.000	0.843

Figure 6: Instance-level Kernel SHAP explanation of a representative misclassified sample, whose ground-truth label is *negative* but whose predicted label is *positive*. The left and right panels decompose the class-aligned scores for *negative* and *positive*, respectively, from their fold-specific expected values into regional feature contributions. Red bars increase the corresponding class score, whereas blue bars decrease it. Feature abbreviations are defined in Appendix A.

By comparison, the lateral deformation of the lower face is considerably weaker. The bilateral horizontal contrasts of the mouth corners and cheeks have normalized values of -0.151 and -0.055, contributing only $+0.08$ and $+0.07$ to the *positive* score. Although their negative signs indicate outward motion under the adopted coordinate convention, their relatively small magnitudes do not constitute strong joint evidence of mouth and cheek expansion. The decision is therefore dominated by bilateral mouth-corner elevation, which renders the pouting-like deformation indistinguishable from the regional motion signature of the *positive* class under the current representation.

For the ground-truth *negative* class, the score decreases from 1.449 to 1.134. Although upward chin motion contributes $+0.34$, the left and right mouth-corner features contribute -0.22 and -0.20, respectively, and are reinforced by several smaller negative contributions from the eyebrow and cheek features. Consequently, the negative-class evidence is insufficient to counteract the strong positive-class support produced by mouth-corner elevation, leading to the final prediction of *positive*.

This case reveals a limitation of the compact regional motion representation. Although regional aggregation preserves anatomical correspondence and motion direction, it compresses fine-grained optical flow into regional summaries and explicitly represents only a limited set of bilateral and intra-region relationships. Facial actions with similar regional displacement patterns may therefore

remain difficult to distinguish. In this example, pouting produces bilateral mouth-corner elevation resembling a positive expression, while the weaker lateral deformation of the mouth corners and cheeks receives insufficient emphasis. This ambiguity may be further amplified by subject-specific deformation, residual rigid motion, or optical-flow estimation errors. Moreover, onset–apex displacement omits temporal evolution and fine-grained intra-region shape changes. Future work will address these limitations by incorporating richer cross-region dependencies, temporal dynamics, fine-grained intra-region shape modeling, and uncertainty-aware weighting to improve robustness in ambiguous cases.