WORLDMER: RECOGNIZING MICRO-EXPRESSIONS THROUGH TEMPORAL CLOSURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing micro-expression recognition (MER) methods typically aggregate facial motion cues across the onset-to-apex and apex-to-offset phases, overlooking the predictive biomechanical relationship between muscle activation and relaxation. To explicitly model this underlying physical coherence, we draw inspiration from latent world models, which excel at learning transition dynamics in state spaces. We introduce **temporal closure** as a bidirectional predictive constraint between onset-to-apex activation and apex-to-offset relaxation. Given either observed phase, the model predicts the latent representation of its counterpart. Based on this formulation, **WorldMER** learns two direction-specific transitions between onset-to-apex and apex-to-offset latent states. Their predictions encode phase dynamics shared across expressions, while the residuals between predicted and observed states expose sample-specific departures from those dynamics. A reliability-aware gate then regulates predicted and residual cues at the token-channel level while retaining observed motion as the evidence anchor. Under subject-independent evaluation, WorldMER achieves **90.33% UF1** and **90.30% UAR** on the 3DB Composite benchmark, and obtains the best UF1 across the three-, four-, and seven-class CAS(ME)^3 protocols. Ablations show that transition-derived predictions outperform observed motion alone and that bidirectional prediction, residual reasoning, and adaptive fusion provide complementary gains. Our code is available in an anonymous repository at https://anonymous.4open.science/r/WorldMER/.

1 INTRODUCTION

Micro-expression recognition (MER) relies on subtle, localized facial motion under scarce and imbalanced supervision (Davison et al., 2018; Pfister et al., 2011). A micro-expression unfolds through onset-to-apex (OA) activation and apex-to-offset (AO) relaxation. Existing methods improve recognition through motion descriptors (Liong et al., 2018), learned motion modeling (Zheng et al., 2026), or auxiliary objectives and facial priors (Liu et al., 2024; 2025; Cai et al., 2026). CausalNet links the two phases through their near-opposite muscle motion directions and models their relationships using causal attention (Zhang et al., 2025b).

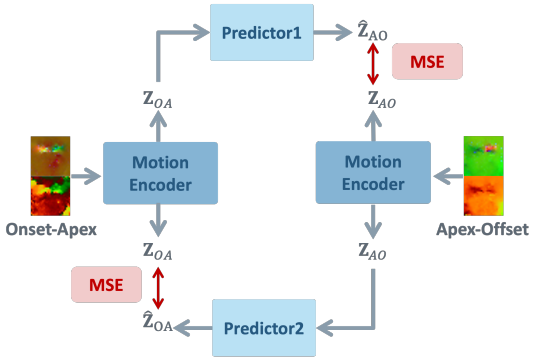

Figure 1: Temporal closure through bidirectional phase prediction. Motion encoders produce observed OA and AO states. Two direction-specific predictors estimate each phase from the other. Mean squared error (MSE) matches predictions to stop-gradient observed targets during training, providing mutual predictive constraints.

Despite these advancements, most existing approaches treat the onset-to-apex (OA) activation and apex-to-offset (AO) relaxation as independent feature sets to be fused. This straightforward aggregation overlooks a fundamental property of facial dynamics: the predictive biomechanical relationship be-

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.

tween muscle activation and relaxation. In a genuine micro-expression, the trajectory of relaxation is heavily conditioned on its preceding activation, and vice versa. We hypothesize that if a model can capture this intrinsic transition rule, the resulting cross-phase predictions—and their discrepancies from the actual observed motion—could provide highly discriminative features for classifying expressions, beyond what the raw observed motion offers alone.

To explicitly model this predictive relationship without reconstructing high-dimensional image pixels, we draw inspiration from latent world models (Ha & Schmidhuber, 2018; Assran et al., 2025). World models excel at learning compact state representations and the intrinsic transition dynamics that govern how a system evolves. By casting the OA and AO phases as latent states within a world model, we introduce **temporal closure** as a bidirectional predictive constraint (Figure 1). Given either observed phase, the model predicts the latent state of its counterpart. Crucially, because facial activation and relaxation are not strictly symmetric or temporally reversible, we learn the forward transition (OA $\rightarrow$ AO) and the backward transition (AO $\rightarrow$ OA) independently. This bidirectional formulation turns the two observed phases into mutual constraints.

Based on this principle, we propose **WorldMER**, a bidirectional latent world model for MER. The model learns direction-specific transitions jointly with expression classification. Rather than merely acting as an auxiliary training objective, temporal closure serves as a core mechanism for feature derivation. Specifically, the model generates predicted phase states (encoding expected shared dynamics) and extracts transition residuals (quantifying sample-specific departures from expectations). An adaptive gating mechanism then evaluates the reliability of these predicted and residual cues, dynamically integrating them with the original observed states to form a comprehensive motion representation. Our empirical analysis confirms that transition-derived predictions inherently outperform raw observed motion, and their combination yields optimal recognition capabilities.

The contributions of this work are as follows:

- We propose temporal closure, modeling the predictive relationship between facial activation and relaxation. By adapting latent world models, we formulate this relationship as a bidirectional constraint between phase states.
- We develop WorldMER, an architecture that explicitly leverages cross-phase predictions and transition residuals as discriminative features. An adaptive fusion mechanism is designed to optimally integrate these transition-derived cues with observed motion anchors.
- Extensive subject-independent evaluations demonstrate that WorldMER achieves state-of-the-art performance on the 3DB Composite (90.33% UF1) and CAS(ME)^3 benchmarks. Comprehensive ablations empirically validate our motivation, proving that bidirectional prediction significantly outperforms unidirectional variants, and that predictions and residuals provide complementary gains over observed features.

2 Related Work

2.1 Micro-Expression Recognition

Micro-expression recognition (MER) relies heavily on motion representation because discriminative facial changes are brief, local, and weak (Ben et al., 2022; Li et al., 2022). Early handcrafted methods summarize regional optical-flow magnitude and direction, while apex-based networks learn compact representations from onset-to-apex flow, optical strain, or shallow spatiotemporal streams (Liu et al., 2016; Liong et al., 2018; Gan et al., 2019; Liong et al., 2019). Recent approaches further strengthen observed motion through flow-guided attention, gradient-motion modeling, local deformation modeling, and continuous weak-trend modeling (Cai et al., 2024; Zhang et al., 2026; 2025a; Zheng et al., 2026). CausalNet also models contraction and relaxation over complete micro-expression intervals to reduce the impact of key-frame errors (Zhang et al., 2025b).

These approaches improve MER through motion representation learning and temporal modeling, including the relationships between activation and relaxation. WorldMER focuses on **learning a predictive relationship between the two phases of a complete micro-expression event**. Direction-specific transitions **predict each phase representation from its counterpart**, with the observed representation serving as the prediction target. The predicted states and their absolute residuals

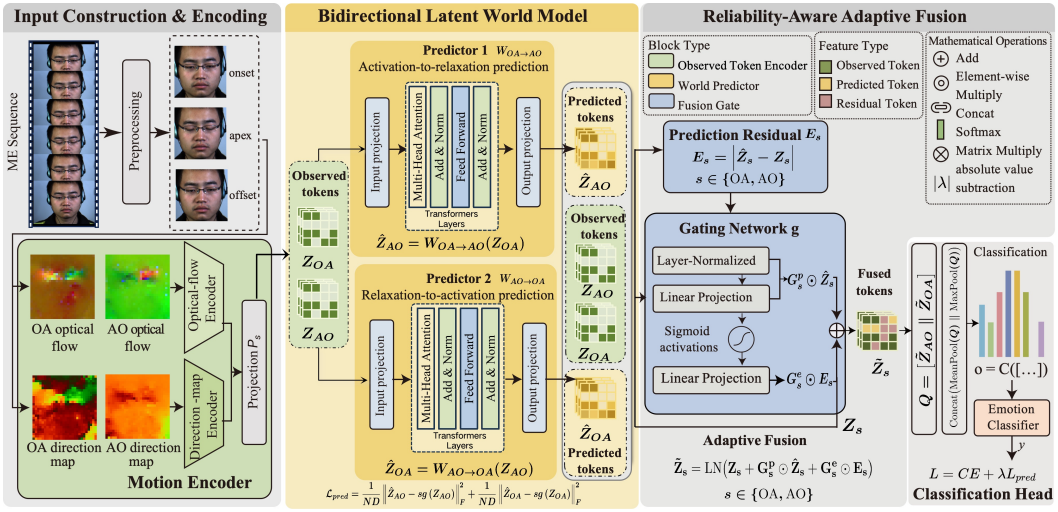

Figure 2: Overview of WorldMER. Observed OA and AO motion states serve as both recognition evidence and sources for bidirectional cross-phase prediction. Two direction-specific latent transitions estimate the complementary-phase expectations, while discrepancies between predicted and observed states produce transition residuals. Prediction and residual evidence are adaptively weighted and introduced as corrections to the observed-state path, preserving observed motion as the evidence anchor. The corrected OA and AO states are finally aggregated for micro-expression recognition.

remain available at inference and are integrated with observed states for classification. This formulation uses cross-phase prediction **both as a training constraint and as a source of recognition features**.

2.2 World Model Technology

World models learn compact states and transition rules that describe how a system evolves. World Models, PlaNet, and Dreamer predict future latent states for planning and decision making (Ha & Schmidhuber, 2018; Hafner et al., 2019; 2020), while recent joint-embedding predictive architectures such as V-JEPA 2 and LeWorldModel move prediction from pixels to learned representations (Assran et al., 2025; Maes et al., 2026). WorldMER adopts this latent-prediction principle to model the relationship between activation and relaxation without reconstructing facial images.

WorldMER applies latent prediction to recognition from complete micro-expression clips, where both activation and relaxation are observed. It learns separate OA-to-AO and AO-to-OA transitions to model the conditional relationship in each direction. The backward transition uses the observed relaxation phase to infer the activation representation, without assuming that facial motion is temporally reversible. Both transitions are trained jointly with recognition, with observed phase representations serving as stop-gradient prediction targets. Their outputs provide cross-phase predictions, while transition residuals quantify absolute discrepancies between predicted and observed states. An adaptive gate integrates these features with observed states during inference. WorldMER thus uses latent prediction within a supervised recognition framework to learn phase relationships useful for expression classification.

3 WorldMER: Recognition by Temporal Closure

WorldMER learns motion representations for micro-expression classification from onset-to-apex (OA) activation and apex-to-offset (AO) relaxation. Both phases are observed inputs, allowing each phase to provide information for predicting the latent representation of its counterpart. The resulting predictions and residuals are used as additional features for expression classification.

Figure 2 provides an overview of the WorldMER framework. WorldMER first represents OA and AO motion as phase states and then learns two direction-specific latent transitions between them. Each transition predicts the representation of one phase from the other, while transition residuals quantify discrepancies between predicted and observed states. An adaptive gate integrates these features with the observed states, and the resulting representations are used for micro-expression recognition.

3.1 Temporal Closure as Cross-Phase Compatibility

Let $Z_{\mathrm{OA}}, Z_{\mathrm{AO}} \in \mathbb{R}^{N\times D}$ denote latent representations of activation and relaxation. For compactness, we use $s \in \{\mathrm{OA}, \mathrm{AO}\}$ to denote one phase and $\bar{s}$ its complementary phase. We introduce a direction-specific latent transition $T_{\bar{s}\to s}$ that predicts the latent representation of phase s from its complementary phase.

We define temporal closure through bidirectional cross-phase compatibility:

$$\mathcal{D}_{\mathrm{TC}} = \sum_{s\in\{\mathrm{OA},\mathrm{AO}\}} d(Z_s,\, T_{\bar{s}\to s}(Z_{\bar{s}}))\,, \tag{1}$$

where $d(\cdot,\cdot)$ measures discrepancy in latent space. A smaller $\mathcal{D}_{\mathrm{TC}}$ indicates closer agreement between observed and cross-phase-predicted representations under the learned transitions.

Temporal closure should not be confused with temporal reversibility. Activation and relaxation are directionally asymmetric facial processes; therefore, WorldMER learns $T_{\mathrm{OA}\to\mathrm{AO}}$ and $T_{\mathrm{AO}\to\mathrm{OA}}$ independently, without parameter sharing or an inverse-transition constraint. The two directions instead describe complementary conditional relationships between the phases.

3.2 Phase-State Representation

Given onset, apex, and offset frames I_o, I_a, and I_f, we construct phase-wise motion representations using Dual TV-L1 optical flow. For each phase $s \in \{\mathrm{OA}, \mathrm{AO}\}$, X_s^f contains horizontal flow, vertical flow, and optical strain, while X_s^d represents motion direction. Both are resized to $28 \times 28 \times 3$.

Flow and direction cues are encoded separately and projected into a common latent space. We write this compactly as $Z_s = P_s([E_f(X_s^f)\|E_d(X_s^d)]) \in \mathbb{R}^{N\times D}$. In our implementation, $N = 4$ corresponds to eye- and mouth-centered regional tokens and $D = 256$.

As shown in the left part of Figure 2, the resulting Z_s has a dual role. It is direct evidence of the facial motion observed in phase s, while simultaneously providing the context from which WorldMER reasons about the complementary phase.

3.3 Cross-Phase Expectation and Transition Residual

The latent transitions are learned by predicting one phase from the other. For a fixed distribution of phase representations with finite second moments, minimizing expected squared prediction error over unrestricted predictors yields

$$T^*_{\bar{s}\to s}(Z_{\bar{s}}) = \mathbb{E}\left[Z_s \mid Z_{\bar{s}}\right]. \tag{2}$$

Equation 2 provides a regression interpretation of cross-phase prediction. In WorldMER, the representations and transitions are learned jointly with recognition. We use *cross-phase expectation* to denote the learned prediction $\hat{Z}_s = T_{\bar{s}\to s}(Z_{\bar{s}})$, motivated by this interpretation. The learned predictor is additionally shaped by the recognition objective.

We quantify the discrepancy between observed and predicted representations using the elementwise absolute residual:

$$R_s = \left|Z_s - \hat{Z}_s\right|, \qquad \mathcal{E}_s = \left\{ \underbrace{Z_s}_{\text{observation}}\,,\ \underbrace{\hat{Z}_s}_{\text{expectation}}\,,\ \underbrace{R_s}_{\text{residual}} \right\}. \tag{3}$$

The observed state Z_s encodes motion in phase s, while $\hat{Z}_s$ is predicted from the complementary phase. The residual R_s records the elementwise absolute difference between these representations.

This discrepancy may provide useful classification cues, but can also reflect errors in motion estimation or latent prediction. WorldMER therefore learns to weight predicted states and residuals when combining them with observed states.

3.4 Transition-Aware Evidence Correction

As illustrated in the right part of Figure 2, WorldMER retains the observed phase state as its evidence anchor and uses cross-phase expectations and transition residuals only to adaptively correct that observation.

To determine how transition-derived evidence should modify the observation, we compute two token-channel weighting fields using a shared adaptive function:

$$(A_s^p, A_s^r) = \mathcal{A}_\phi\Big(Z_s, \hat{Z}_s, R_s\Big). \tag{4}$$

Both A_s^p and A_s^r lie in $[0,1]^{N\times D}$, where $\mathcal{A}_\phi$ denotes the shared adaptive weighting function. The two weighting fields determine, at each regional token and feature channel, how strongly prediction-derived and residual evidence should modify the observation.

The resulting phase representation is written as

$$\widetilde{Z}_s = Z_s + \underbrace{A_s^p \odot \hat{Z}_s + A_s^r \odot R_s}_{\Delta Z_s}, \tag{5}$$

or equivalently $\widetilde{Z}_s = Z_s + \Delta Z_s$.

This formulation retains Z_s through a direct additive path and introduces ΔZ_s as a learned contribution from predictions and residuals. Here, correction denotes an additive update in feature space. The gate controls the contribution of predicted states and residuals at each token and channel.

The corrected OA and AO states are finally summarized by a recognition readout $\mathcal{H}_\psi$ to produce the expression prediction. Standard architectural operations used to instantiate $\mathcal{A}_\phi$ and $\mathcal{H}_\psi$ are described with the implementation details rather than treated as part of the temporal-closure formulation.

3.5 Learning Temporal Closure

The two latent transitions are learned jointly with expression recognition. We write the overall objective as

$$\mathcal{L} = \mathcal{L}_{\text{rec}} + \lambda \cdot \frac{1}{2} \sum_{s\in\{\text{OA},\text{AO}\}} d\big(\hat{Z}_s, \text{sg}(Z_s)\big). \tag{6}$$

where $\mathcal{L}_{\text{rec}}$ is the supervised expression-recognition objective, the summation measures bidirectional latent prediction error, and λ controls its contribution to training.

We instantiate $d(\cdot,\cdot)$ as mean squared error across tokens and channels and use cross-entropy for $\mathcal{L}_{\text{rec}}$. The stop-gradient operator blocks gradients through the target argument of each directional prediction loss. Each observed phase representation remains trainable through expression recognition and its role as the source of the opposite transition.

Importantly, the prediction objective is not used only as an auxiliary regularizer during training. At inference, both transitions remain active: their outputs constitute the cross-phase expectations in Eq. 3, and the corresponding transition residuals participate directly in the adaptive correction in Eq. 5. Temporal closure is therefore both a learning principle and an inference-time source of recognition evidence.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation. We evaluate WorldMER on CASME II (Yan et al., 2014), SMIC (Pfister et al., 2011), SAMM (Davison et al., 2018), and CAS(ME)^3 (Li et al., 2023). Fol-

Table 1: Comparison with state-of-the-art methods on CASME II, SMIC, SAMM, and the three-class Composite benchmark. Values are percentages. Bold and underlined values indicate the best and second-best results, respectively. † denotes video input, and ‡ denotes facial-action priors or auxiliary representation learning.

Method	Pub.-Yr	CASME II		SMIC		SAMM		Composite	
		UF1	UAR	UF1	UAR	UF1	UAR	UF1	UAR
HTNet (Wang et al., 2024)	NC 24	95.32	95.16	80.49	79.05	81.31	81.24	86.03	84.75
SRMCL (Bao et al., 2024)	TAC 24	96.35	96.49	79.46	80.53	84.70	88.66	86.30	88.30
TGMAE† (Liu et al., 2024)	ICME 24	95.97	95.62	82.25	83.44	84.15	78.85	87.90	88.02
EDMDBN (Ma et al., 2025a)	PRL 25	94.84	96.19	79.48	80.85	83.36	86.61	88.21	89.33
CAUT (Jiang et al., 2025)	ICIP 25	96.89	96.88	83.40	83.51	83.06	84.02	88.00	88.33
CausalNet (Zhang et al., 2025b)	ICCV 25	97.48	97.82	84.05	84.33	87.08	85.22	89.30	89.41
MPFNet-C (Ma et al., 2025b)	ICASSP 25	91.10	92.30	80.60	80.90	79.50	83.90	84.00	84.60
CWT-MER† (Zheng et al., 2026)	TAC 26	91.55	90.67	79.34	79.29	83.19	84.47	86.22	86.14
EDR (Zhang et al., 2026)	TAC 26	97.85	97.54	81.97	80.66	85.51	83.89	88.41	86.59
Ra-ICL (Cai et al., 2026)	CVPR 26	96.20	96.20	81.79	82.74	86.68	88.85	88.05	89.11
D-FACE‡ (Deng et al., 2026)	ICLR 26	97.38	97.54	84.22	**84.76**	87.16	85.13	89.43	89.67
UGPT‡ (Liu et al., 2026)	TAC 26	96.89	96.78	**84.79**	84.20	89.43	**89.89**	90.25	89.69
WorldMER	**Proposed**	**98.55**	**97.92**	83.79	84.22	**90.46**	88.40	**90.33**	**90.30**

lowing composite database evaluation (CDE), the first three datasets form the three-class Composite benchmark (3DB), containing 442 samples from 68 subjects. The classes are negative, positive, and surprise. CAS(ME)^3 contains 696 samples from 88 subjects under the three-class protocol and 854 samples from 94 subjects under the four- and seven-class protocols. All experiments use leave-one-subject-out (LOSO) evaluation. We report unweighted F1-score (UF1), unweighted average recall (UAR), and accuracy (ACC).

Implementation. We use the architecture described in Section 3 and train for up to 200 epochs with AdamW. The learning rate is 5×10^{-5}, weight decay is 10^{-4}, batch size is 512, and dropout is 0.1. The prediction-loss weight is $\lambda = 0.50$ unless otherwise specified.

4.2 Comparison with State-of-the-Art Methods

Composite benchmark. WorldMER achieves 90.33% UF1 and 90.30% UAR on 3DB, the highest scores among the compared methods (Table 1). These results exceed UGPT (Liu et al., 2026) by 0.08 and 0.61 percentage points, respectively. WorldMER also leads on CASME II, with 98.55% UF1 and 97.92% UAR, and achieves the highest SAMM UF1 of 90.46%. However, its SAMM UAR remains below UGPT, and neither SMIC metric ranks first. The aggregate improvement therefore does not imply uniform gains across source datasets.

Table 2: Comparison with state-of-the-art methods on CAS(ME)^3 under the three-, four-, and seven-class LOSO protocols (%).

Proto.	Method	Pub.-Yr	UAR	UF1
3-cls	CapsuleNet	FG 19	29.77	23.52
	STSTNet	FG 19	37.61	37.23
	EMR	FG 19	36.56	36.13
	HTNet	NC 24	54.15	57.67
	MER-CLIP	TAFFC 25	76.06	78.32
	FAMNet	IJCNN 25	51.00	43.42
	Lite-Point-GCN	TAFFC 25	74.12	68.19
	Ra-ICL	CVPR 26	74.54	75.85
	WorldMER	**Ours**	**79.37**	**84.02**
4-cls	MER-CLIP	TAFFC 25	62.42	65.44
	Ra-ICL	CVPR 26	61.10	62.30
	WorldMER	**Ours**	**64.51**	**68.02**
7-cls	MER-CLIP	TAFFC 25	**50.14**	48.80
	Ra-ICL	CVPR 26	45.20	46.10
	WorldMER	**Ours**	49.36	**50.72**

CAS(ME)^3 benchmark. Among the compared methods, WorldMER achieves the highest UF1 across all three protocols (Table 2). For three-class recognition, it obtains 84.02% UF1 and 79.37%

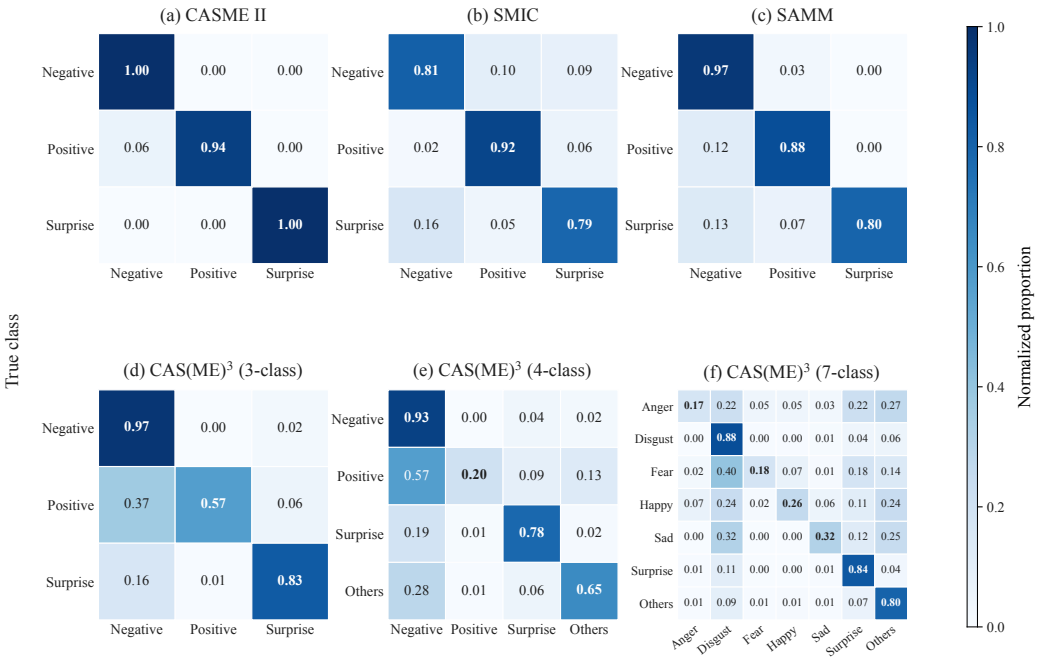

Figure 3: Row-normalized confusion matrices. Top: CASME II, SMIC, and SAMM under CDE. Bottom: three-, four-, and seven-class CAS(ME)^3. Rows indicate true classes and columns indicate predictions; all panels share a common color scale.

UAR, exceeding MER-CLIP (Liu et al., 2025) by 5.70 and 3.31 percentage points. For four-class recognition, it reaches 68.02% UF1 and 64.51% UAR, exceeding MER-CLIP by 2.58 and 2.09 percentage points, respectively. For seven-class recognition, WorldMER achieves the highest UF1 of 50.72%, but its 49.36% UAR trails MER-CLIP by 0.78 points. The lower scores under this protocol highlight the remaining difficulty of fine-grained recognition.

Class-wise performance. Figure 3 shows nearly diagonal predictions on CASME II and greater confusion on SMIC and SAMM. On CAS(ME)^3, positive expressions are frequently classified as negative under both three- and four-class protocols. Under the seven-class protocol, disgust, surprise, and others have higher recall than anger, fear, happy, and sad. These patterns identify class-specific limitations not captured by aggregate scores.

4.3 Ablation Studies

All ablations use the three-class Composite benchmark under LOSO evaluation.

Table 3: Evidence ablation on the three-class Composite benchmark (%). Obs., Pred., and Res. denote observed states, predicted states, and transition residuals. Bold and underlined values indicate the best and second-best results.

Variant	Obs.	Pred.	Res.	Gate	UF1	UAR	ACC
Observed states	✓	–	–	–	88.30	87.14	89.82
Predicted states	–	✓	–	–	89.10	88.84	90.50
Obs. + Pred.	✓	✓	–	–	89.47	89.15	90.72
Obs. + Pred. + Res.	✓	✓	✓	–	89.71	<u>90.01</u>	90.95
Obs. + Gated Pred.	✓	✓	–	✓	<u>89.80</u>	89.68	<u>91.18</u>
WorldMER	✓	✓	✓	✓	**90.33**	**90.30**	**91.63**

Transition-derived evidence. Table 3 compares recognition variants using observed states, predicted states, residuals, and adaptive gating. Predicted states alone exceed observed-state baseline

by 0.80 and 1.70 percentage points in UF1 and UAR, respectively. Combining observed and predicted states yields 89.47% UF1 and 89.15% UAR; including transition residuals gives 89.71% and 90.01%. Gating without residuals yields 89.80% UF1 and 89.68% UAR, while the full model achieves 90.33% and 90.30%. Relative to the observed-state baseline, the full model improves UF1 and UAR by 2.03 and 3.16 percentage points, respectively. These results support the combined use of predictions, transition residuals, and adaptive fusion within the evaluated framework.

Table 4: Transition-direction ablation on the three-class Composite benchmark (%). Source denotes predictor inputs; checkmarks indicate states used for recognition. WorldMER additionally uses the corresponding residuals and gates. Bold denotes the best result within each group.

Variant	Direction	Source	Z_{OA}	Z_{AO}	$\widehat{Z}_{\mathrm{OA}}$	$\widehat{Z}_{\mathrm{AO}}$	UF1	UAR	ACC
Predicted states only	OA→AO	Z_{OA}	–	–	–	✓	87.71	87.10	89.37
	AO→OA	Z_{AO}	–	–	✓	–	78.95	76.13	82.58
	Bidirectional	$Z_{\mathrm{OA}}, Z_{\mathrm{AO}}$	–	–	✓	✓	**89.10**	**88.84**	**90.50**
WorldMER	OA→AO	Z_{OA}	✓	✓	–	✓	89.69	89.46	90.72
	AO→OA	Z_{AO}	✓	✓	✓	–	89.50	89.34	90.72
	Bidirectional	$Z_{\mathrm{OA}}, Z_{\mathrm{AO}}$	✓	✓	✓	✓	**90.33**	**90.30**	**91.63**

Transition direction. Table 4 compares forward, backward, and bidirectional prediction. In the prediction-only setting, OA→AO yields higher scores than AO→OA. The bidirectional variant uses both observed phases as prediction inputs and achieves the highest scores in this group. In the WorldMER variants, both observed phase states are retained for recognition. Bidirectional prediction reaches 90.33% UF1 and 90.30% UAR, compared with 89.69%/89.46% for forward prediction and 89.50%/89.34% for backward prediction. Within this framework, using both prediction directions yields higher recognition scores than using either direction alone.

Table 5: Sensitivity to prediction-loss weight λ on 3DB (%). Bold denotes the best result.

λ	0.01	0.10	0.30	0.50	0.70	0.90	1.00
UF1	88.51	88.56	89.58	**90.33**	89.83	89.17	88.40
UAR	88.00	88.40	90.16	**90.30**	88.99	88.59	88.58
ACC	90.27	90.27	90.95	**91.63**	91.18	90.95	90.05

Prediction-loss weight. Table 5 varies λ in Equation 6 from 0.01 to 1.00. Among the tested values, $\lambda = 0.50$ gives the highest UF1 and UAR. Both metrics remain below 88.60% at $\lambda \leq 0.10$ and decrease to 88.40%/88.58% at $\lambda = 1.00$. These results favor an intermediate prediction-loss weight within the evaluated range.

4.4 Qualitative Analysis

Figure 4 illustrates cross-phase prediction by comparing observed token features with those predicted from the complementary phase over the input optical flow. Arrow directions follow the input flow, colors identify the four tokens, and lengths encode within-panel normalized responses. Predicted columns show latent tokens, not predicted flow. The three OA examples have observed peaks at T3, while predicted responses favor T2, with T0 and T2 nearly tied in (a). In the separate AO examples, predicted responses peak at T2, whereas observed peaks vary between T0 and T3. These are relative responses within independently normalized panels.

Figure 5 visualizes observed and predicted states, transition residuals, and prediction-gate weights for a different selection of samples. It shows the observed–predicted feature residuals and the weights assigned to predictions by the gate. Five of the six residual maps peak at T3; the OA residual in (f) peaks at T0. OA and AO gate responses peak at T1 and T2, respectively. Gaussian markers show scores at four fixed token locations. Independent per-panel normalization supports relative comparisons within each panel; Table 3 reports recognition performance for the corresponding model variants.

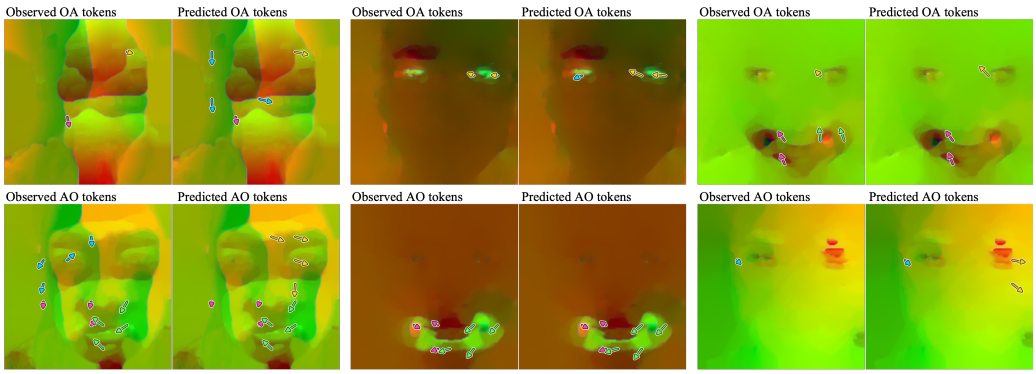

Figure 4: Observed and cross-phase-predicted tokens over input optical flow. Rows show OA/AO examples; paired images show observed/predicted tokens. Cyan, magenta, yellow, and green denote T0–T3, corresponding to the image-left eye, image-left mouth, image-right eye, and image-right mouth, respectively. Arrow directions indicate local motion; lengths represent relative token responses within each panel.

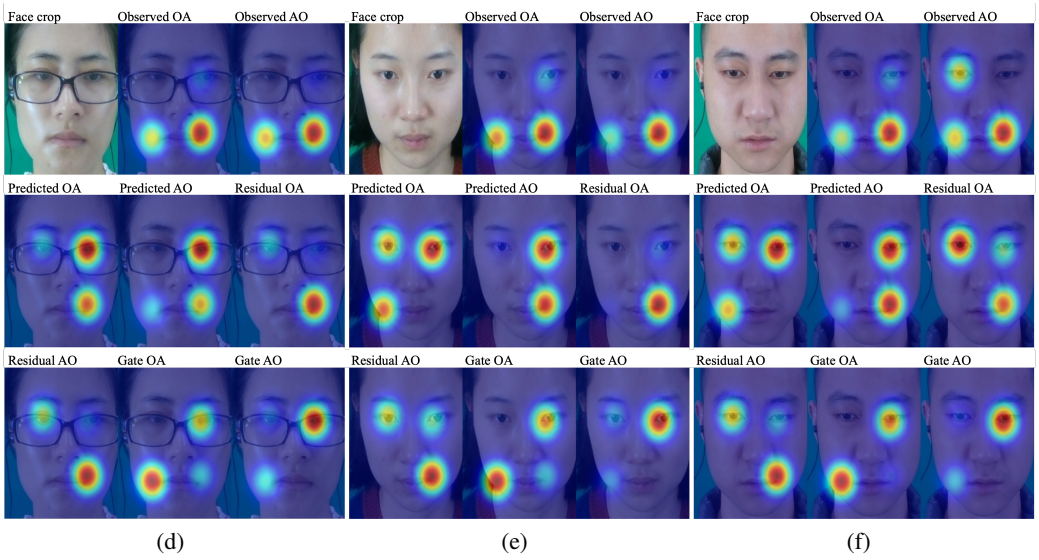

Figure 5: Observed, predicted, residual, and prediction-gate token maps for samples (d)–(f). Feature scores use channel-mean absolute values; gates show the channel mean of A_s^p. Scores are normalized per panel and rendered as Gaussian markers at four fixed eye/mouth tokens.

5 Conclusion and Limitations

WorldMER introduces temporal closure as a bidirectional predictive constraint for micro-expression recognition. Direction-specific latent transitions predict OA and AO representations from their complementary phases. Predicted states and transition residuals are integrated with observed states through adaptive fusion for expression classification. Under subject-independent evaluation, WorldMER achieves 90.33% UF1 and 90.30% UAR on the 3DB Composite benchmark. Component and direction analyses confirm complementary gains from bidirectional prediction, residual evidence, and adaptive fusion, supporting the proposed progression beyond observation-centric recognition. Future work should learn temporal closure under automatically localized, uncertain phase boundaries and evaluate stronger cross-dataset generalization.

AI USE STATEMENT

In this work, we used generative AI tools to help with background research, literature search, LaTeX formatting, table layout, and code development and debugging. We did not use AI to run experiments, generate results, or make final scientific judgments. We also used AI to improve the clarity and structure of some passages. We have reviewed all AI-assisted work. All AI-assisted text was checked and revised by the authors, and all AI-assisted code was inspected, run, and validated before use in experiments. All experimental results, analyses, and conclusions were independently verified by the authors. We take responsibility for the final content of this work, including any text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work uses the previously released CASME II, SMIC, SAMM, and CAS(ME)^3 micro-expression benchmarks, accessed in accordance with the terms and access conditions specified by their original providers. The authors did not collect new data, recruit or interact with human participants, or redistribute dataset files, raw videos, or annotations. Facial-expression analysis can enable sensitive inferences about individuals and may be misused in surveillance, consequential decision-making, or non-consensual settings; the proposed method is evaluated solely for scientific research on benchmark micro-expression recognition, and we do not advocate such uses. Any future deployment should comply with the datasets' access conditions and undergo an application-specific assessment of privacy, informed consent, fairness, potential harms, and applicable laws.

REPRODUCIBILITY STATEMENT

Sections 3 and 4 describe the proposed method, learning objectives, datasets, subject-disjoint LOSO evaluation protocol, evaluation metrics, and principal training settings. The appendix provides further implementation details, including the software and hardware environment, randomization controls, dataset access conditions, label mappings, data preprocessing, training configuration, and metric aggregation procedures. Together, these details enable independent reproduction of the reported experiments, subject to authorized access to the original datasets. We do not redistribute raw videos, image frames, or restricted annotations. An anonymized code package containing the implementation, configuration files, and evaluation scripts is included as supplementary material and will be publicly released upon acceptance.

REFERENCES

Mahmoud Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Yongtang Bao, Chenxi Wu, Peng Zhang, Caifeng Shan, Yue Qi, and Xianye Ben. Boosting micro-expression recognition via self-expression reconstruction and memory contrastive learning. *IEEE Transactions on Affective Computing*, 15(4):2083–2096, 2024. doi: 10.1109/TAFFC.2024.3397701.

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(9):5826–5846, 2022. doi: 10.1109/TPAMI.2021.3067464.

Wenhao Cai, Junli Zhao, Ran Yi, Minjing Yu, Fuqing Duan, Zhenkuan Pan, and Yong-Jin Liu. MFDAN: Multi-level flow-driven attention network for micro-expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(12):12823–12836, 2024. doi: 10.1109/TCSVT.2024.3437481.

Yaomin Cai, C. L. Philip Chen, Shiting Xu, Haiqi Liu, and Tong Zhang. Region-aware instance consistency learning for micro-expression recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 34428–34438, 2026.

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(1): 116–129, 2018. doi: 10.1109/TAFFC.2016.2573832.

Yicheng Deng, Hideaki Hayashi, and Hajime Nagahara. From pixels to semantics: Unified facial action representation learning for micro-expression analysis. In *The Fourteenth International Conference on Learning Representations*, 2026.

Y. S. 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. doi: 10.1016/j.image.2019.02.005.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018. URL https://arxiv.org/abs/1803.10122.

Danijar Hafner, Timothy P. Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning latent dynamics for planning from pixels. In *Proceedings of the International Conference on Machine Learning*, pp. 2555–2565, 2019. URL https://proceedings.mlr.press/v97/hafner19a.html.

Danijar Hafner, Timothy P. Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=H1g5lSkgwH.

Hanyu Jiang, Jiayi Lyu, Xing Lan, and Jian Xue. Continuous action unit intensity modeling for micro-expression recognition. In *Proceedings of the IEEE International Conference on Image Processing*, 2025. doi: 10.1109/ICIP55913.2025.11084479.

Jingting Li, Zizhao Dong, Shaoyuan Lu, Su-Jing Wang, Wen-Jing Yan, Yinhuan Ma, Ye Liu, Changbing Huang, and Xiaolan Fu. CAS(ME)^3: A third generation facial spontaneous micro-expression database with depth information and high ecological validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(3):2782–2800, 2023. doi: 10.1109/TPAMI.2022.3174895.

Yante Li, Jinsheng Wei, Yang Liu, Janne Kauttonen, and Guoying Zhao. Deep learning for micro-expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(4):2028–2046, 2022. doi: 10.1109/TAFFC.2022.3205170.

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. doi: 10.1016/j.image.2017.11.006.

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 *Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition*, pp. 1–5, 2019. doi: 10.1109/FG.2019.8756567.

Jiateng Liu, Hengcan Shi, Feng Chen, Min Liu, Yaonan Wang, and Wenming Zheng. UGPT: Uncertain graph prompt tuning for micro-expression recognition. *IEEE Transactions on Affective Computing*, pp. 1–18, 2026. doi: 10.1109/TAFFC.2026.3708854.

Shifeng Liu, Xinglong Mao, Sirui Zhao, Chaoyou Fu, Ying Yu, Tong Xu, and Enhong Chen. TG-MAE: Self-supervised micro-expression recognition with temporal gaussian masked autoencoder. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pp. 1–6, 2024. doi: 10.1109/ICME57554.2024.10687556.

Shifeng Liu, Xinglong Mao, Sirui Zhao, Peiming Li, Tong Xu, and Enhong Chen. MER-CLIP: AU-guided vision-language alignment for micro-expression recognition. *IEEE Transactions on Affective Computing*, 16(4):3028–3042, 2025. doi: 10.1109/TAFFC.2025.3584918.

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(4):299–310, 2016. doi: 10.1109/TAFFC.2015.2485205.

Bingyang Ma, Lu Wang, Qingfen Wang, Haoran Wang, Ruolin Li, Lisheng Xu, Yongchun Li, and Hongchao Wei. Entire-detail motion dual-branch network for micro-expression recognition. *Pattern Recognition Letters*, 189:166–174, 2025a. doi: 10.1016/j.patrec.2025.01.021.

Chuang Ma, Shaokai Zhao, Yu Pei, Liang Xie, Erwei Yin, and Ye Yan. A multi-prior fusion network for video-based micro-expression recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 2025b. doi: 10.1109/ICASSP49660.2025.10889587.

Lucas Maes, Quentin Le Lidec, Damien Scieur, Yann LeCun, and Randall Balestriero. LeWorldModel: Stable end-to-end joint-embedding predictive architecture from pixels. *arXiv preprint arXiv:2603.19312*, 2026.

Tomas Pfister, Xiaobai Li, Guoying Zhao, and Matti Pietikainen. Recognising spontaneous facial micro-expressions. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1449–1456, 2011. doi: 10.1109/ICCV.2011.6126401.

Zhifeng Wang, Kaihao Zhang, Wenhan Luo, and Ramesh Sankaranarayana. HTNet for micro-expression recognition. *Neurocomputing*, 602:128196, 2024. doi: 10.1016/j.neucom.2024.128196.

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(1):e86041, 2014. doi: 10.1371/journal.pone.0086041.

Christopher Zach, Thomas Pock, and Horst Bischof. A duality based approach for realtime TV-L1 optical flow. In *Proceedings of the Joint Pattern Recognition Symposium*, pp. 214–223, 2007. URL https://doi.org/10.1007/978-3-540-74936-3_22.

Bohao Zhang, Xuejiao Wang, Changbo Wang, and Gaoqi He. Dynamic stereotype theory induced micro-expression recognition with oriented deformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10701–10711, 2025a. doi: 10.1109/CVPR52734.2025.01000.

Tong Zhang, Shiting Xu, and C. L. Philip Chen. Enhanced dynamic representation via gradient-motion modeling for micro-expression recognition. *IEEE Transactions on Affective Computing*, pp. 1–13, 2026. doi: 10.1109/TAFFC.2026.3678926.

Zheyuan Zhang, Weihao Tang, and Hong Chen. Rethinking key-frame-based micro-expression recognition: A robust and accurate framework against key-frame errors. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 12274–12283, 2025b. doi: 10.1109/ICCV51701.2025.01141.

Qiuyan Zheng, Zhiming Zhou, Yue Zhao, Baoqing Zhu, and Qiaoling Han. Modeling continuous weak temporal trends for video-based micro-expression recognition. *IEEE Transactions on Affective Computing*, 17(2):1721–1733, 2026. doi: 10.1109/TAFFC.2026.3676852.

A Reproducibility Conditions

The reference implementation is written in Python using PyTorch and is run on a CUDA-capable device. OpenCV is used to read and crop motion maps; NumPy and Pandas support data handling; scikit-learn computes evaluation metrics; PyYAML loads experiment configurations; and einops implements tensor rearrangements. The dependency list also includes tqdm and openpyxl. Because the inspected dependency specification does not pin package versions, this appendix does not claim unverified version numbers.

All settings needed by the reported 3DB experiment are stored in one YAML configuration. The implementation expects precomputed motion maps, a sample metadata file, and per-subject data directories. Preprocessing is performed once when the program starts; the resulting tensors are then reused while the LOSO folds are executed sequentially. Appendices D– F give the code-matched input construction, model, optimization, and aggregation rules.

B Dataset Access and Usage Conditions

The experiments use CASME II, SMIC, SAMM, and CAS(ME)^3. These datasets are provider-controlled; reproducing the experiments therefore requires applying to the original provider of each dataset and complying with the applicable licence or data-use agreement.

B.1 Dataset-Specific Access Routes

CASME II and CAS(ME)^3. Applications are submitted through the official database portal. The portal provides a blank licence agreement for each dataset; applicants complete and sign the agreement and upload it through the linked application form.

SMIC. Access is requested through the official University of Oulu research-group page. The request should identify the required SMIC version or subset and follow the contact instructions on that page.

SAMM. Applicants use the official SAMM dataset page, complete the licence agreement, and follow the submission instructions provided there.

Access decisions and permitted uses remain subject to the terms specified by the respective dataset providers. The provider pages should be rechecked when the data are requested because forms and access procedures may change.

C Dataset Construction

C.1 Three-Dataset Composite Benchmark

The 3DB Composite benchmark combines CASME II (Yan et al., 2014), SMIC (Pfister et al., 2011), and SAMM (Davison et al., 2018) in a shared three-class label space: label 0 is negative, label 1 is positive, and label 2 is surprise. Table 6 reports the composition read from the experiment metadata.

Table 6: Composition of the 3DB Composite benchmark.

Dataset	Subjects	Negative	Positive	Surprise	Total
CASME II	24	88	32	25	145
SAMM	28	92	26	15	133
SMIC	16	70	51	43	164
Total	68	250	109	83	442

C.2 CAS(ME)^3 PROTOCOLS

CAS(ME)^3 (Li et al., 2023) is evaluated under the three-, four-, and seven-class protocols used in the main paper. The three-class protocol contains 696 samples from 88 subjects, whereas the four- and seven-class protocols each contain 854 samples from 94 subjects. The experiments use the class definitions and annotations supplied with the corresponding benchmark protocols.

D DATA PREPROCESSING

The preprocessing pipeline starts from the videos or image sequences obtained from the official providers of CASME II, SMIC, SAMM, and CAS(ME)^3. Face alignment, motion estimation, and motion-map construction are performed offline. The training program therefore reads the resulting cached maps rather than recomputing them in every LOSO fold.

D.1 FACE ALIGNMENT AND KEY-FRAME PREPARATION

For each micro-expression sample, the onset, apex, and offset frames are identified according to the annotations or key-frame protocol associated with the benchmark. Facial landmarks are then used to register the face to a common coordinate system. The alignment compensates for in-plane rotation, translation, and scale variation, after which a consistent facial region is cropped from the three key frames. All three frames are mapped to the same target coordinate system and cropped consistently so that corresponding pixel locations in the aligned onset, apex, and offset images refer to the same facial regions. This step suppresses motion caused by head displacement or inconsistent image framing before the substantially weaker facial deformation is estimated.

We denote the aligned and cropped frames by $\bar{I}_o$, $\bar{I}_a$, and $\bar{I}_f$.

D.2 PHASE-WISE OPTICAL FLOW

Dual TV-L1 optical flow (Zach et al., 2007) is computed between the aligned onset and apex frames and between the aligned apex and offset frames:

$$F_{\mathrm{OA}} = \mathrm{TVL1}(\bar{I}_o, \bar{I}_a), \qquad F_{\mathrm{AO}} = \mathrm{TVL1}(\bar{I}_a, \bar{I}_f). \tag{7}$$

For phase $s \in \{\mathrm{OA}, \mathrm{AO}\}$, the resulting dense field is $F_s(x, y) = (u_s(x, y), v_s(x, y))$, where u_s and v_s are the horizontal and vertical displacements. Following the optical-flow representation used in micro-expression recognition (Liong et al., 2019), spatial derivatives of the displacement field form the symmetric strain tensor

$$E_s = \frac{1}{2}\left(\nabla F_s + \nabla F_s^{\mathsf{T}}\right), \tag{8}$$

whose magnitude measures local facial deformation. The horizontal flow, vertical flow, and optical-strain magnitude are stacked into the three-channel representation X_s^f. The orientation $\theta_s(x, y) = \mathrm{atan2}(v_s(x, y), u_s(x, y))$ is converted into a three-channel direction map X_s^d.

The four maps X_{OA}^f, X_{AO}^f, X_{OA}^d, and X_{AO}^d are resized to $28 \times 28 \times 3$ and cached as PNG images. This ordering is retained by the loader: streams 0 and 2 form the OA state, whereas streams 1 and 3 form the AO state. The flow and direction maps therefore describe complementary magnitude, deformation, and orientation information for both activation and relaxation.

D.3 REGIONAL CROP CONSTRUCTION

Each stored map is read as a $28 \times 28 \times 3$ PNG image. Five fixed centers are represented in $(\mathrm{row}, \mathrm{column})$ order. After clipping each coordinate to $[7, 21]$ and conversion to integers, they are

$$(9, 11),\ (21, 10),\ (15, 17),\ (10, 21),\ (20, 21).$$

A 14×14 crop is extracted around every center. These centers correspond to the left eye, left mouth corner, central face, right eye, and right mouth corner in the normalized coordinate system. The left eye and left mouth crops are concatenated horizontally, as are the right eye and right mouth crops; the two resulting strips are then concatenated vertically. This yields a $28 \times 28 \times 3$ four-region composite. The central crop is computed by the loader but is not used in the final composite. The same operation is applied to all four motion maps.

D.4 Tensor Assembly

For a sample, the four composites are stacked and converted to floating-point tensors with shape $4 \times 3 \times 28 \times 28$. A batch therefore has shape $B \times 4 \times 3 \times 28 \times 28$. The loader applies no additional rescaling, dataset-statistic normalization, or train/test augmentation to the stored map values.

E Complete Architecture Details

E.1 Phase-State Encoders

Each of the four $28 \times 28 \times 3$ maps is partitioned with a 7×7 patch embedding and processed by a three-level hierarchical Transformer with block repeats $(3, 3, 9)$. OA and AO maps of the same cue type share encoder parameters; flow and direction maps use separate parameter sets. Every stream produces four 256-dimensional tokens. For each phase, the corresponding flow and direction tokens are concatenated, layer-normalized, projected from 512 to 256 dimensions, passed through GELU, and regularized with dropout 0.1. The resulting observed states are $Z_{\mathrm{OA}}, Z_{\mathrm{AO}} \in \mathbb{R}^{4\times 256}$.

E.2 Bidirectional Latent Transition Predictors

The OA→AO and AO→OA predictors have separate parameters. Each projects 256-dimensional tokens to a 512-dimensional hidden space, adds a learned four-token positional embedding, and applies two pre-norm Transformer encoder layers with eight attention heads, a 2048-dimensional feed-forward sublayer, GELU, and dropout 0.1. A final layer normalization and linear projection return four 256-dimensional predicted tokens.

E.3 Transition Residual and Adaptive Weighting

For each phase, the implementation uses the element-wise absolute residual $R_s = |\hat{Z}_s - Z_s|$ without ℓ_2 normalization. The concatenated $[Z_s, \hat{Z}_s, R_s]$ vector is layer-normalized and linearly projected from 768 to 512 dimensions. A sigmoid produces separate 256-dimensional gates for the prediction and residual. The corrected token is $\widetilde{Z}_s = Z_s + A_s^p \odot \hat{Z}_s + A_s^r \odot R_s$, followed by layer normalization and dropout.

E.4 Recognition Readout

Corrected OA and AO tokens are concatenated feature-wise, giving four 512-dimensional tokens. Mean and maximum pooling over the four tokens are concatenated into a 1024-dimensional vector. The classification head applies layer normalization, a $1024 \to 2048$ linear layer, GELU, dropout 0.1, and a final linear layer whose output dimension equals the number of classes in the evaluated protocol.

F Training and Evaluation Protocol

F.1 Optimization Settings

The model is trained for at most 200 epochs using AdamW with learning rate 5×10^{-5}, weight decay 10^{-4}, and batch size 512. Dropout is 0.1. The joint objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}} + \lambda \left[\mathrm{MSE}(\hat{Z}_{\mathrm{AO}}, \mathrm{sg}(Z_{\mathrm{AO}})) + \mathrm{MSE}(\hat{Z}_{\mathrm{OA}}, \mathrm{sg}(Z_{\mathrm{OA}})) \right], \tag{9}$$

where $\lambda = 0.50$, MSE is averaged over all token and feature dimensions, and sg stops gradients through the target states.

F.2 One LOSO Fold

For one LOSO fold, samples from the designated subject are loaded as the held-out set and the remaining samples form the training set. Training batches are shuffled, whereas held-out inference

preserves sample order. A new model and AdamW optimizer are initialized for the fold. After every epoch, the model is evaluated on the held-out subject; the implementation retains the predictions from the epoch with the highest held-out ACC. After all subject folds finish, their retained predictions and ground-truth labels are pooled before computing the reported metrics.

F.3 Metric Definitions and Aggregation

Let C be the number of classes and let TP_c, FP_c, and FN_c denote the class-wise counts after pooling predictions across LOSO folds. The reported metrics are

$$\mathrm{ACC} = \frac{\sum_{c=1}^{C} TP_c}{N}, \tag{10}$$

$$\mathrm{UAR} = \frac{1}{C}\sum_{c=1}^{C} \frac{TP_c}{TP_c + FN_c}, \tag{11}$$

$$\mathrm{UF1} = \frac{1}{C}\sum_{c=1}^{C} \frac{2TP_c}{2TP_c + FP_c + FN_c}. \tag{12}$$

Thus UAR and UF1 are macro-averaged recall and F1, respectively, rather than averages of fold-level scores. A class contribution with a zero denominator is set to zero, matching the evaluation implementation.

G Additional Quantitative Results

G.1 Fixed-Fusion Direction Analysis

Table 7 extends the main fixed-fusion result by separating the two prediction directions. Here, predicted states are added with unit weight; residual evidence and adaptive gates are not used.

Table 7: Direction analysis for fixed observed-plus-predicted fusion on the 3DB Composite benchmark. All metrics are percentages.

Setting	OA representation	AO representation	UF1	UAR	ACC
Observed only	Z_{OA}	Z_{AO}	88.30	87.14	89.82
OA→AO	Z_{OA}	$Z_{\mathrm{AO}} + \hat{Z}_{\mathrm{AO}}$	88.73	87.87	90.50
AO→OA	$Z_{\mathrm{OA}} + \hat{Z}_{\mathrm{OA}}$	Z_{AO}	88.87	88.52	90.05
Bidirectional	$Z_{\mathrm{OA}} + \hat{Z}_{\mathrm{OA}}$	$Z_{\mathrm{AO}} + \hat{Z}_{\mathrm{AO}}$	**89.47**	**89.15**	**90.72**

Either one-way prediction improves UF1 and UAR over observed states alone, and bidirectional fixed fusion is the strongest variant in this family. It remains below full WorldMER (90.33 UF1, 90.30 UAR, and 91.63 ACC), supporting the use of transition residuals and adaptive weighting rather than unregulated addition of predicted features. The qualitative figures already included in the main paper are not repeated here.