CausalAffect: Causally Guided Learning of Psychology-Aligned Facial Affect Relations

Guanyu Hu[1,3,4] **Tangzheng Lian**[5] **Dimitrios Kollias**[1,2*]
Oya Celiktutan[5] **Xinyu Yang**[3*]

[1]Centre for Multimodal AI, Queen Mary University of London, UK
[2]Digital Environment Research Institute, Queen Mary University of London, UK
[3]School of Computer Science and Technology, Xi'an Jiaotong University, China
[4]Department of Computer Science, Aalto University, Finland
[5]King's College London, London, UK

Abstract

Facial affect understanding extends beyond image-to-expression-label prediction. It requires identifying active facial Action Units (AUs), modeling causal facilitation and inhibition among them, and explaining how AU configurations give rise to expression states. Such causal relations have long been studied in cognitive psychology and psychophysiology. The Facial Action Coding System, Component Process Model, and Muscle Synergy Framework provide rich causal priors for facial behavior. In computer vision, however, facial affect models, including structured AU-aware approaches, remain predominantly correlational, with limited grounding in psychological priors. A persistent gap therefore separates biological facial mechanisms from learned visual representations. Without direct biological interventions, existing methods lack a principled way to recover psychology-aligned relations from observational images. To bridge this gap, we propose CausalAffect, a weakly supervised, causally guided framework. It introduces interventions in a semantically grounded AU representation space and requires learned causal relations to explain the resulting downstream responses. This principle guides the learning of directed facilitative and inhibitory relations, capturing shared causal structure while accommodating individual variability. The learned relations recover canonical pathways documented in the psychological literature and reveal plausible inhibitory and sample-specific influences. A blind expert study validates their psychological plausibility. Incorporating these relations into prediction, CausalAffect consistently improves AU detection and facial expression recognition across six benchmarks.

1 Introduction

Facial affect understanding is not merely image-to-label prediction: a psychologically meaningful model should identify active facial Action Units (AUs), model their facilitative or inhibitory interactions, and explain how AU configurations give rise to expression states. These relations have long been studied in cognitive psychology and psychophysiology. The Facial Action Coding System (FACS) [15] provides an anatomical vocabulary for facial muscle actions; the Component Process Model (CPM) [70] views expressions as downstream outcomes of appraisal-driven facial responses; and the Muscle Synergy Framework [11, 2] suggests that facial actions are coordinated through facilitative and antagonistic interactions. For example, AU6 (cheek raiser) and AU12 (lip corner puller) jointly support happiness, while AU24 (lip presser) and AU25 (lips part) illustrate

Email: guanyu.hu@aalto.fi. *Corresponding authors: Dimitrios Kollias (d.kollias@qmul.ac.uk) and Xinyu Yang (yxyphd@mail.xjtu.edu.cn).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

competing movement requirements in compound expressions [14]. These perspectives provide a causal and psychological foundation for modeling both AU→Expression composition and AU→AU coordination, suggesting that facial affect depends not only on which AUs are present, but also on how they interact, reinforce, or suppress one another within coherent AU-level mechanisms.

Despite this causal and psychological foundation, most data-driven facial affect models still learn primarily correlational structures [47, 29, 89]. Their training prioritizes accurate mappings from observed inputs to task outputs [87, 88, 27]. In facial expression recognition (FER), these mappings connect observed cues to emotion labels [31, 19, 39]. AU-aware models incorporate AUs into this predictive mapping, often only as auxiliary labels or intermediate attributes [8, 61]. Structured approaches extend this use of AUs by explicitly modeling their dependencies through attention [59, 86], graph neural networks [75, 85], or co-occurrence statistics [56]. The learning objectives of these relational models, however, typically prioritize *predictive accuracy* over *psychological alignment*. Unlike psychophysiological studies with controlled interventions, these models learn from images and labels alone. They therefore tend to exploit label co-occurrence and dataset-specific shortcuts [7] to improve prediction without capturing facial mechanisms. Without principled constraints, recovering trustworthy relations from such supervision remains difficult. The resulting dependencies can be dense, unsigned, symmetric, or dataset-specific, obscuring whether an AU facilitates or inhibits another AU or expression in a psychologically plausible way.

Recovering psychology-aligned relations is further complicated by fragmented supervision and facial variability. AU datasets annotate muscle actions but usually lack expression labels, whereas expression datasets provide semantic labels without AU annotations [32]. Relation learning must therefore bridge these complementary sources of supervision without fully paired AU and expression annotations. Facial affect also follows no single rigid template: AU combinations, intensities, and co-activation patterns vary across individuals and cultures [34, 35]. These challenges motivate a shared AU interface for integrating disjoint supervision. Causal modeling must also reconcile population-level regularities with individual facial configurations.

A principled solution must therefore connect complementary supervision with a testable causal account of facial behavior. We propose **CausalAffect**, a weakly supervised, causally guided framework for learning psychologically aligned facial affect relations. A shared AU representation connects disjoint AU and expression supervision and grounds relation learning in interpretable facial actions. The framework learns directed facilitative and inhibitory relations, describing how AUs coordinate with one another (AU→AU) and how their configurations support or suppress expressions (AU→Expression). A shared causal structure captures population-level regularities, with instance-specific refinements accommodating individual facial configurations. In the absence of direct biological interventions, CausalAffect constrains these relations through counterfactual interventions in the AU representation space. We formulate an *intervention-consistency principle*: learned relations should explain observed outcomes and target responses to AU interventions. Changes to graph-selected AUs should affect the target, while changes to non-selected AUs should leave it relatively stable. This principle constrains learning beyond label co-occurrence and dataset-specific shortcuts, linking causal interpretation to testable model behavior.

With this intervention-consistency constraint, CausalAffect recovers psychology-aligned AU relations from disjoint supervision, extending canonical AU→Expression pathways with plausible inhibitory and sample-specific dependencies. A blind expert study validates the psychological plausibility of the evaluated excitatory and inhibitory relations. Incorporating these relations into prediction, CausalAffect outperforms prior state-of-the-art methods across six AU detection and FER benchmarks.

Our contributions are summarized as follows:

- **Psychology-aligned causal explanations of facial affect.** CausalAffect explains AU coordination and expression composition through directed facilitative and inhibitory relations beyond symmetric co-occurrence, reconciling population-level regularities with individual facial variation.
- **Intervention-grounded causal learning from observational images.** Disentangled AU variables ground interventions in facial semantics. Target sensitivity to graph-selected AU changes and relative invariance to others constrain causal learning beyond label prediction.
- **Relational evidence beyond canonical AU templates.** Learned inhibitory relations and context-dependent AU roles extend canonical templates with supporting and competing evidence. Literature and expert judgments support their psychological plausibility, while jointly learning relations and predictions improves AU detection and FER across six benchmarks.

2 Problem formulation

We aim to learn *causally guided* and *psychologically plausible* facial affect relations[1] under weak supervision, where AU and expression annotations are disjoint across datasets. Let $\mathcal{D}_{\mathrm{AU}} = \{(I_k^{\mathrm{AU}}, \mathbf{y}_k^{\mathrm{AU}})\}_{k=1}^{K_{\mathrm{AU}}}$ and $\mathcal{D}_{\mathrm{Expr}} = \{(I_k^{\mathrm{Expr}}, y_k^{\mathrm{Expr}})\}_{k=1}^{K_{\mathrm{Expr}}}$ contain K_{AU} AU-labeled and K_{Expr} expression-labeled images. Here k indexes images; $\mathbf{y}_k^{\mathrm{AU}} \in \{0,1\}^{N_{\mathrm{AU}}}$ gives binary labels for N_{AU} AUs where annotated, and $y_k^{\mathrm{Expr}} \in \{1, \ldots, N_{\mathrm{Expr}}\}$ identifies one of N_{Expr} expression classes. No paired AU and expression labels are required. Given an image I, CausalAffect encodes it into AU-specific variables $\mathbf{F}_{\mathrm{AU}} = \{\mathbf{f}_{\mathrm{AU}}^{(i)}\}_{i=1}^{N_{\mathrm{AU}}}$, where $\mathbf{f}_{\mathrm{AU}}^{(i)} \in \mathbb{R}^d$ is the d-dimensional representation of AU i. These variables are shared sources for two directed, signed graph types: **(i) AU→AU graph**, which models facilitative or inhibitory coordination among facial actions; and **(ii) AU→Expression graph**, which models how AU configurations support or suppress expression states. For both types, a **global graph** $\mathcal{G}_{\mathrm{g}}$ captures population-level regularities, while a **sample-adaptive graph** $\mathcal{G}_{\mathrm{s}}(I)$ refines this structure using instance-specific AU evidence.

3 CausalAffect framework

CausalAffect translates the psychobiological organisation of facial actions into a learnable causal framework (Figure 1). Semantically aligned, disentangled AU variables provide shared, interpretable sources for AU→AU coordination and AU→Expression composition. Directed, signed graphs model excitatory and inhibitory influences from these AU sources to AU or expression targets. A global graph captures population-level structure, while a sample-adaptive graph refines it using each face's AU evidence. To constrain the causal meaning of these edges, feature-level counterfactual interventions require target sensitivity to changes in graph-selected AUs and relative invariance to changes in non-selected AUs. Joint optimisation combines these intervention constraints with AU and expression prediction losses under disjoint supervision.

3.1 Disentangled AU bottleneck encoding

The first stage of CausalAffect learns semantically meaningful and explicitly disentangled AU features, which serve as the nodes for causally guided relation learning. As formulated in our problem setup, AU embeddings act as the atomic source nodes for both the biomechanical AU→AU graph and the generative AU→Expression graph. The quality of these nodes is therefore critical. For graph edges to reflect genuine physiological interactions rather than superficial co-occurrences, each node should correspond to a distinct anatomical generator. To this end, we introduce a *Disentangled AU Bottleneck Encoding* scheme that isolates AU semantics from confounders (e.g., identity, demographics), so that message passing operates on purified muscle-activation signals and supports robust causal inference.

Feature Extraction. We first extract global visual features using a backbone-agnostic encoder (ConvNeXt in our setting), producing a D-dimensional image embedding $z_{\mathrm{img}} \in \mathbb{R}^D$. To transition from global visual semantics to fine-grained AU signals, we introduce N_{AU} lightweight projection heads $\{\phi_i\}_{i=1}^{N_{\mathrm{AU}}}$. Each head ϕ_i maps z_{img} into an AU-specific feature $\mathbf{f}_{\mathrm{AU}}^{(i)}$, yielding the feature matrix $\mathbf{F}_{\mathrm{AU}} \in \mathbb{R}^{N_{\mathrm{AU}} \times d}$. To ensure semantic fidelity, each $\mathbf{f}_{\mathrm{AU}}^{(i)}$ is supervised by a binary classifier to predict its activation probability $\hat{y}_{\mathrm{feat}}^{(i)}$ via binary cross-entropy (BCE): $\mathcal{L}_{\mathrm{feat}}^{\mathrm{AU}} = \frac{1}{N_{\mathrm{AU}}} \sum_{i=1}^{N_{\mathrm{AU}}} \mathrm{BCE}(\hat{y}_{\mathrm{feat}}^{(i)}, y_{\mathrm{AU}}^{(i)})$.

Information Bottleneck via HSIC Disentanglement. To enforce disentanglement, we adopt the Deep Information Bottleneck (IB) principle [80] implemented via the Hilbert–Schmidt Independence Criterion (HSIC, see Appendix A.1 for HSIC calculation) [24]. HSIC provides a mathematically grounded measure of statistical dependence without requiring restrictive distributional assumptions, and it captures arbitrary non-linear relationships between variables. This capability allows us to attenuate high-order confounding signals in AU representations, helping each node represent a distinct AU for downstream causal relation learning. Specifically, it imposes two constraints to purify the

[1]The learned edges represent ***causally guided and psychology-aligned*** hypotheses about facial relations: individual AUs facilitate or inhibit other AUs, and AU configurations support or suppress expression states. These hypotheses connect the anatomical description of facial actions [15] with evidence that AU combinations and dynamics convey expression categories [35, 6]. We formulate these relations as directed, signed edges over AU bottleneck representations and constrain their learning through feature-level counterfactual interventions. The resulting relations are evaluated against established literature, expert plausibility judgements, and image-level AU intervention responses.

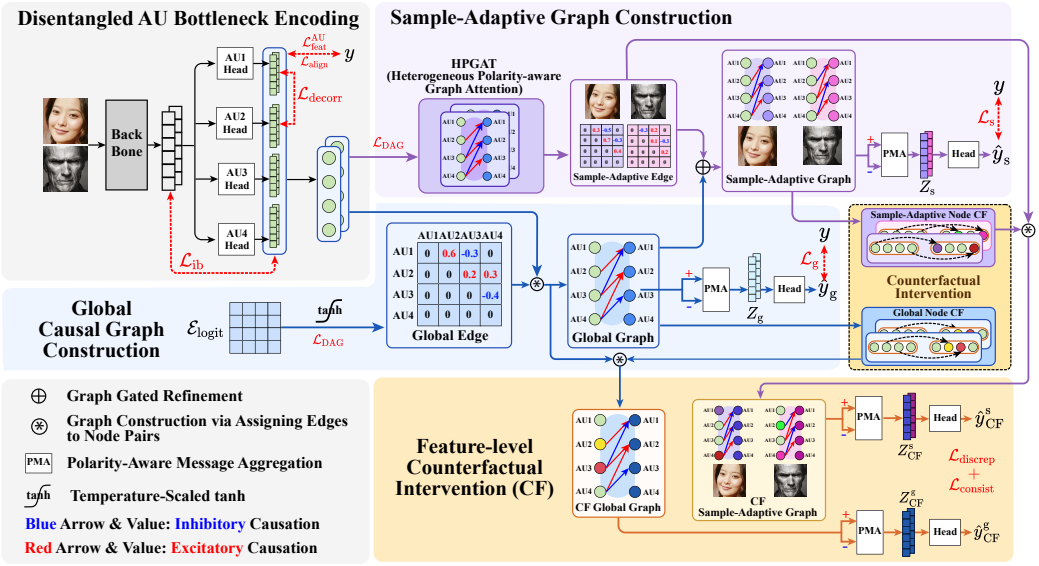

Figure 1: Overview of **CausalAffect**, consisting of four key components: (1) **Disentangled AU Bottleneck Encoding**, which extracts semantically independent AU features; (2) **Global Causal Graph Construction**, which learns a polarity-aware graph capturing population-level dependencies; (3) **Sample-Adaptive Graph Refinement**, which adjusts the global graph for each instance via residual adaptation; and (4) **Feature-Level Counterfactual Intervention**, which enforces causal inductive bias through latent perturbations and factual–counterfactual consistency. The figure illustrates AU→AU branch; AU→Expression branch is analogous and omitted for brevity.

causal nodes: **(i) Removing Confounders ($\mathcal{L}_{\text{ib}}$):** Minimize the dependence between each specific AU feature $\mathbf{f}_{\text{AU}}^{(i)}$ and the global embedding z_{img}. This compresses the representation, forcing the encoder to discard task-irrelevant confounders that are shared globally but causally irrelevant to specific muscle actions. **(ii) Semantic Alignment ($\mathcal{L}_{\text{align}}$):** Maximize the dependence between $\mathbf{f}_{\text{AU}}^{(i)}$ and its label $\tilde{y}_{\text{AU}}^{(i)}$, using ground truth when available and a detached pseudo-label otherwise. This ensures each compressed feature remains strictly aligned with its anatomical definition.

While the standard Information Bottleneck emphasizes the disentanglement of compression and task alignment, we introduce a third, causality-oriented constraint to enforce independence across AU nodes. **(iii) Structural Independence ($\mathcal{L}_{\text{decorr}}$):** We extend the IB objective by minimizing statistical dependence between distinct AU features $\mathbf{f}_{\text{AU}}^{(i)}$ and $\mathbf{f}_{\text{AU}}^{(j)}$ for $i \neq j$. This addresses a key requirement for causal modeling: although AUs may be physiologically coupled through motor synergies, their *latent* node representations should be disentangled to serve as valid source variables. Decorrelation between AU features suppresses feature leakage and encoder redundancy, promoting explicit modelling of causal interactions in the downstream graph.

Objective Formulation. Writing $\mathcal{H}(\cdot,\cdot)$ for HSIC dependence, we instantiate these constraints as:

$$\begin{aligned}
\mathcal{L}_{\text{ib}} &= \frac{1}{N_{\text{AU}}}\sum_{i=1}^{N_{\text{AU}}} \mathcal{H}\left(\mathbf{f}_{\text{AU}}^{(i)},\, z_{\text{img}}\right), \\
\mathcal{L}_{\text{align}} &= -\frac{1}{N_{\text{AU}}}\sum_{i=1}^{N_{\text{AU}}} \mathcal{H}\left(\mathbf{f}_{\text{AU}}^{(i)},\, \tilde{y}_{\text{AU}}^{(i)}\right), \\
\mathcal{L}_{\text{decorr}} &= \frac{1}{N_{\text{AU}}(N_{\text{AU}}-1)}\sum_{i\neq j} \mathcal{H}\left(\mathbf{f}_{\text{AU}}^{(i)},\, \mathbf{f}_{\text{AU}}^{(j)}\right).
\end{aligned} \tag{1}$$

For expression-labeled samples without AU annotations, pseudo-labels are used only for the HSIC-based semantic alignment term and are detached from the computation graph; they are not used as ground-truth AU labels for AU prediction losses.

Total AU Bottleneck Loss. We weight compression by λ_{ib}, decorrelation by λ_{decorr}, and semantic alignment by λ_{align} to obtain

$$\mathcal{L}_{\text{AU}} = \mathcal{L}_{\text{feat}}^{\text{AU}} + \lambda_{\text{ib}} \cdot \mathcal{L}_{\text{ib}} + \lambda_{\text{decorr}} \cdot \mathcal{L}_{\text{decorr}} + \lambda_{\text{align}} \cdot \mathcal{L}_{\text{align}}. \tag{2}$$

3.2 Global causal graph (GC graph)

The second stage of CausalAffect constructs the Global Causal Graph to capture the stable, population-level anatomical backbone of facial behavior. We model both *AU→Expression* and *AU→AU* causal relations using *directed and polarity-aware* edges to capture asymmetric excitatory and inhibitory interactions.

Graph Construction. The global AU→AU graph $\mathcal{G}_{\mathrm{g}}^{\mathrm{AU}}$ and the global AU→Expression graph $\mathcal{G}_{\mathrm{g}}^{\mathrm{Expr}}$ adopt the same structural form. We define the global graph as $\mathcal{G}_{\mathrm{g}} = (\mathcal{N}_{\mathrm{g,\,source}}, \mathcal{N}_{\mathrm{g,\,target}}, \mathcal{E}_{\mathrm{g}})$, where the source nodes $\mathcal{N}_{\mathrm{g,\,source}}$ are the disentangled AU representations $\mathbf{F}_{\mathrm{AU}}$ shared across both graph types. The target nodes $\mathcal{N}_{\mathrm{g,\,target}}$ correspond to representations causally aggregated from $\mathcal{N}_{\mathrm{g,\,source}}$ (expressions in AU→Expression, AUs in AU→AU). The entries of $\mathcal{E}_{\mathrm{g}}$ lie in $(-1, 1)$ and encode the polarity and magnitude of causal influence, thereby specifying the directed topology from source to target. To parameterize these edges, we introduce a learnable matrix $\mathcal{E}_{\text{logit}} \in \mathbb{R}^{N_{\text{target}} \times N_{\text{source}}}$, where each entry $\mathcal{E}_{\text{logit}}^{(j,i)}$ denotes the unbounded edge logit from source node i to target node j, where N_{source} and N_{target} denote the numbers of source and target nodes. To bound the edge weights, we apply temperature-scaled $\tanh$ transformation:

$$\mathcal{E}_{\mathrm{g}}^{(j,i)} = \tanh\left(\tau \cdot \mathcal{E}_{\text{logit}}^{(j,i)}\right), \quad \tau = \exp(T_{\mathrm{g}}), \tag{3}$$

where T_{g} is a learnable log-scale parameter and $\tau > 0$ is the resulting edge temperature scale. This maps the logits into the interval $(-1, 1)$, producing a *signed* edge matrix $\mathcal{E}_{\mathrm{g}} \in \mathbb{R}^{N_{\text{target}} \times N_{\text{source}}}$. The sign of each entry indicates whether the influence is excitatory (> 0) or inhibitory (< 0), while its absolute value measures the strength of the modelled relation.

Polarity-Aware Message Aggregation. To disentangle polarity effects, we decompose the signed edge into excitatory and inhibitory components:

$$\mathcal{E}_{\mathrm{g}}^{+} = \mathbb{I}[\mathcal{E}_{\mathrm{g}} > 0] \odot \mathcal{E}_{\mathrm{g}}, \quad \mathcal{E}_{\mathrm{g}}^{-} = \mathbb{I}[\mathcal{E}_{\mathrm{g}} < 0] \odot (-\mathcal{E}_{\mathrm{g}}), \tag{4}$$

where $\mathbb{I}[\cdot]$ is the element-wise indicator function and $\odot$ denotes the Hadamard product. To aggregate source AU nodes into target nodes (AU or Expression) through graph, we first project the disentangled AU features into two value spaces, i.e., $V_g^{+} = \mathbf{F}_{\mathrm{AU}} W_g^{+}$ and $V_g^{-} = \mathbf{F}_{\mathrm{AU}} W_g^{-}$, where $W_g^{+}, W_g^{-} \in \mathbb{R}^{d \times d_{\mathrm{g}}}$ are the positive and negative projection matrices, d_{g} denotes the dimensionality of the projected representation. The aggregated representation is computed as $Z_{\mathrm{g}} = \mathcal{E}_{\mathrm{g}}^{+} V_g^{+} + \mathcal{E}_{\mathrm{g}}^{-} V_g^{-} \in \mathbb{R}^{N_{\text{target}} \times d_{\mathrm{g}}}$.

AU→AU Homogeneous Soft DAG Constraint. A Directed Acyclic Graph (DAG) provides a structural prior for modelling the direction of muscle interactions. While the heterogeneous AU→Expression pathway inherently aligns with this acyclic generative structure, the homogeneous AU→AU pathway requires additional topological constraints to avoid degenerating into dense, bidirectional loops. The *Muscle Synergy Framework* describes motor behaviour in terms of coordinated muscle activation [11, 2], providing a rationale for modelling AU interactions jointly. To preserve this coordination while clarifying the direction of causal influence, we implement a **soft DAG regularisation**. Its acyclicity penalty discourages cyclic connections during joint optimisation, while allowing reciprocal AU interactions to be learned when their contribution to the other learning objectives outweighs this penalty. We use the trace of a matrix exponential to measure acyclicity [100].

$$\mathcal{L}_{\mathrm{DAG}}^{\mathrm{g}} = \operatorname{tr}\left(\exp\left(\mathcal{E}_{\mathrm{g}}^{\mathrm{AU}} \odot \mathcal{E}_{\mathrm{g}}^{\mathrm{AU}}\right)\right) - N_{\mathrm{AU}}, \tag{5}$$

Global Graph Loss. For AU→AU graph, the prediction is $\hat{y}_{\mathrm{AU}} = \Phi_{\mathrm{AU}}(Z_{\mathrm{g}}^{\mathrm{AU}})$; for AU→Expression graph, the prediction is $\hat{y}_{\mathrm{Expr}} = \Phi_{\mathrm{Expr}}(Z_{\mathrm{g}}^{\mathrm{Expr}})$, where $Z_{\mathrm{g}}^{\mathrm{AU}} \in \mathbb{R}^{N_{\mathrm{AU}} \times d_{\mathrm{g}}}$ and $Z_{\mathrm{g}}^{\mathrm{Expr}} \in \mathbb{R}^{N_{\mathrm{Expr}} \times d_{\mathrm{g}}}$ are the causally aggregated features, and Φ_{AU}, Φ_{Expr} denote task-specific classifier heads. The graph learning is supervised using the corresponding ground-truth labels y^{AU} and y^{Expr} via the following branch-specific losses:

$$\mathcal{L}_{\mathrm{g}}^{\mathrm{AU}} = \mathrm{BCE}\left(\hat{y}_{\mathrm{AU}}, y^{\mathrm{AU}}\right) + \lambda_{\mathrm{DAG}}^{\mathrm{g}} \mathcal{L}_{\mathrm{DAG}}^{\mathrm{g}}, \qquad \mathcal{L}_{\mathrm{g}}^{\mathrm{Expr}} = \mathrm{CE}\left(\hat{y}_{\mathrm{Expr}}, y^{\mathrm{Expr}}\right). \tag{6}$$

Here, $\mathrm{CE}(\cdot)$ denotes categorical cross-entropy, and $\lambda_{\mathrm{DAG}}^{\mathrm{g}}$ controls the strength of the soft DAG regularization for AU→AU graph.

3.3 Sample-adaptive causal graph (SAC graph)

The global causal graph captures stable, population-level causality, but facial causality also exhibits systematic *contextual* variation. Cross-cultural evidence [34, 35] suggests that expressions combine universal AU backbones with culture- and individual-specific modulations (e.g., smiling shares a core AU12+AU6 pattern, while auxiliary AUs and intensity profiles can vary across Western and East Asian groups). A single fixed topology is therefore insufficient, motivating a *sample-adaptive* structure that refines global dependencies per input.

Sample-adaptive Graph as an Edge Residual. For each input $I^{(k)}$, we construct a sample-adaptive causal graph $\mathcal{G}_{\text{s}}^{(k)} = (\mathcal{N}_{\text{s,source}}^{(k)}, \mathcal{N}_{\text{s,target}}^{(k)}, \mathcal{E}_{\text{s}}^{(k)})$. The source nodes are always the disentangled AU variables, $\mathcal{N}_{\text{s,source}}^{(k)} = \mathbf{F}_{\text{AU}}^{(k)}$, and the target nodes depend on the pathway (AUs for AU→AU, expressions for AU→Expression). Instead of relearning a full graph per sample, we treat $\mathcal{E}_{\text{s}}^{(k)}$ as a *structured correction* to the global edges. The model starts from population-level causality and only "edits" the connections that are necessary for the current instance, which stabilizes learning.

HPGAT: Heterogeneous, Polarity-aware, and Residual. To infer the sample-specific residual, we introduce *Heterogeneous Polarity-aware Graph Attention* (HPGAT). The key observation is that standard GAT-style attention (e.g., GAT/GATv2 [4]) is structurally misaligned with causal graph reasoning: it typically assumes a homogeneous node type and yields non-negative attention scores, which precludes modeling inhibitory influences and cross-type aggregation. HPGAT therefore equips attention with three capabilities that match our causal semantics: (i) *heterogeneous* message passing to support AU→Expression aggregation, (ii) *polarity-aware* edge inference with signed weights in $(-1, 1)$ to represent excitatory and inhibitory effects, and (iii) a *gated residual* mechanism that adaptively refines, rather than overwrites, the global causal structure.

Queries/Keys/Values in HPGAT. For image k at layer l, the source AU nodes provide fixed *Keys* and *Values* for both pathways: $K_{\text{s}}^{(k,l)} = V_{\text{s}}^{(k,l)} = \mathbf{F}_{\text{AU}}^{(k)}$. This design reflects the causal semantics, where AUs serve as the underlying generators and provide the evidence to be attended over. The *Queries* specify the target entities whose representations are to be causally explained and aggregated. For AU→AU, we initialize queries with the AU features, i.e., $Q^{(k,1)} = \mathbf{F}_{\text{AU}}^{(k)}$. For AU→Expression, we instead use learnable prototypes $\mathbf{P}_{\text{Expr}}$ as query anchors, which function as semantic receivers that retrieve supporting (or suppressing) evidence from AU generators for the current instance.

Signed Edge Inference (Instance-specific Causality). A core design choice in HPGAT is to infer *signed, bounded* causal edges directly. This removes the non-negativity and simplex constraints, enabling native inhibitory effects while maintaining numerically stable edge magnitudes. Concretely, at layer l we compute $\mathcal{E}_{\text{s}}^{(k,l)} \in \mathbb{R}^{N_{\text{target}} \times N_{\text{source}}}$ as

$$\mathcal{E}_{\text{s}}^{(k,l)}(j,i) = \tanh\Big(\mathbf{a}^{\top} \operatorname{LeakyReLU}\Big(W_Q^{(l)}\mathbf{q}_j^{(k,l)} + W_K^{(l)}\mathbf{k}_i^{(k,l)}\Big)\Big), \tag{7}$$

where $\mathbf{q}_j^{(k,l)}$ is the query of target node j, $\mathbf{k}_i^{(k,l)}$ is the key of source AU node i, $W_Q^{(l)}$ and $W_K^{(l)}$ are projection matrices, and $\mathbf{a}$ is the attention vector. Applying Eq. 7 to all target–source pairs yields $\mathcal{E}_{\text{s}}^{(k,l)} \in \mathbb{R}^{N_{\text{target}} \times N_{\text{source}}}$.

Polarity-Aware Message Aggregation. We apply the polarity decomposition in Eq. 4 to the sample-specific edges, separating excitatory and inhibitory contributions for message aggregation. The target embedding is updated by polarity-aware message passing:

$$Z_{\text{s}}^{(k,l)} = \mathcal{E}_{\text{s}}^{+(k,l)} V_{\text{s}}^{(k,l)} W_V^{+(l)} + \mathcal{E}_{\text{s}}^{-(k,l)} V_{\text{s}}^{(k,l)} W_V^{-(l)}, \tag{8}$$

where d_h is the message dimension and $W_V^{+(l)}, W_V^{-(l)} \in \mathbb{R}^{d \times d_h}$ project excitatory and inhibitory messages.

Gated Residual Refinement. At the final layer L, we fuse the global graph and sample-specific edges using a learned gating function $g_{\text{gate}}(\cdot)$:

$$\widetilde{\mathcal{E}}_{\text{s}}^{(k)} = \mathbf{G}^{(k)} \odot \mathcal{E}_{\text{g}} + \Big(1 - \mathbf{G}^{(k)}\Big) \odot \mathcal{E}_{\text{s}}^{(k,L)}, \qquad \mathbf{G}^{(k)} = \sigma\Big(g_{\text{gate}}\Big(\mathcal{E}_{\text{s}}^{(k,L)}\Big)\Big), \tag{9}$$

where $\mathbf{G}^{(k)} \in (0,1)^{N_{\text{target}} \times N_{\text{source}}}$ is an edge-wise gate and $\sigma(\cdot)$ is the sigmoid function. Applying the same polarity-aware aggregation to the fused edges yields the final sample-adaptive target

embeddings, combining population-level causal structure with instance-specific refinements. The expanded decomposition and final aggregation are provided in Appendix A.3.

Sample-adaptive graph loss. Using the fused sample-adaptive embeddings, we predict AU activations and expression categories via $\hat{y}_{\text{s}}^{\text{AU}} = \Phi_{\text{AU}}(Z_{\text{s}}^{\text{AU}})$ and $\hat{y}_{\text{s}}^{\text{Expr}} = \Phi_{\text{Expr}}(Z_{\text{s}}^{\text{Expr}})$. We supervise SAC learning with the following objectives:

$$\mathcal{L}_{\text{s}}^{\text{AU}} \;=\; \text{BCE}\big(\hat{y}_{\text{s}}^{\text{AU}},\, y^{\text{AU}}\big) \;+\; \lambda_{\text{DAG}}^{\text{s}}\, \mathcal{L}_{\text{DAG}}^{\text{s}}, \qquad \mathcal{L}_{\text{s}}^{\text{Expr}} \;=\; \text{CE}\big(\hat{y}_{\text{s}}^{\text{Expr}},\, y^{\text{Expr}}\big)\,, \tag{10}$$

Here, $\mathcal{L}_{\text{DAG}}^{\text{s}}$ applies the same DAG penalty to sample-adaptive AU→AU edges, weighted by $\lambda_{\text{DAG}}^{\text{s}}$.

3.4 Feature-level counterfactual intervention (CF)

From psychological and motor-control perspectives, both the *Component Process Model* [70] and the *Muscle Synergy Framework* [11, 2] motivate an *interventional* approach to relation learning: if an AU is a true physiological generator or biomechanical coupler, modifying its state *must* alter the downstream outcome. Building on this view, we formulate an *intervention-consistency principle* for learning AU-level causal relations in a semantically grounded representation space. It combines **(i) Causal Sensitivity**, requiring target responses to interventions on graph-selected AU sources, and **(ii) Causal Invariance**, requiring relative target stability under interventions on non-selected sources. This principle guides both AU→AU and AU→Expression learning.

Edge-aware Intervention Masks. We realize this via a thresholded soft mask with edge-aware saliency, which perturbs source node embeddings conditioned on the graph. By contrasting model behaviors under factual and counterfactual conditions, the model receives consistency and discrepancy signals that guide graph refinement, reinforcing informative edges and suppressing spurious ones. A controlled recovery experiment shows that CF corrects perturbed graph assignments and improves predictive performance (Appendix H). Specifically, given a graph edge matrix $\mathcal{E}$ (from either global or sample-adaptive graphs), we compute a soft importance mask $\mathbf{M}_{\text{soft}} \in (0,1)^{N_{\text{target}} \times N_{\text{source}}}$ as:

$$\mathbf{M}_{\text{soft}} = \sigma(\gamma \cdot (|\mathcal{E}| - \boldsymbol{\theta}))\,, \tag{11}$$

where $|\mathcal{E}|$ denotes element-wise edge magnitudes and $\boldsymbol{\theta} \in [0,1]^{N_{\text{target}}}$ is a learnable threshold vector, γ is a sharpness factor, and $\sigma(\cdot)$ is the element-wise sigmoid. Larger γ yields sharper distinctions between relevant and irrelevant source nodes. For each target node j, we extract a row vector $\mathbf{m}^{(j)} \in [0,1]^{N_{\text{source}}}$, and define two complementary intervention masks. The **consistency mask** $\mathbf{m}_{\text{consist}}^{(j)} = 1 - \mathbf{m}^{(j)}$ emphasises weaker incoming edges, while the **discrepancy mask** $\mathbf{m}_{\text{discrep}}^{(j)} = \mathbf{m}^{(j)}$ emphasises stronger ones.

Feature-space Interventions. Counterfactual interventions inject Gaussian perturbations into AU features, with each mask entry applied across its AU's feature dimensions:

$$\mathbf{F}_{\text{AU-CF}}^{(j,\star)} = \mathbf{F}_{\text{AU}} + \boldsymbol{\epsilon}^{(j)} \odot \mathbf{m}_{\star}^{(j)}. \tag{12}$$

Here, $\boldsymbol{\epsilon}^{(j)} \sim \mathcal{N}(0, \tau_{\text{cf}}^2)$ is Gaussian feature noise with standard deviation τ_{cf}, and $\star \in \{\text{consist}, \text{discrep}\}$ selects the mask. Re-running graph aggregation yields counterfactual target representations $Z_{\text{CF}}^{(j,\star)}$ and predictions $\hat{y}_{\text{CF}}^{(j,\star)}$. $Z^{(j)}$ and $\hat{y}^{(j)}$ denote factual target representations and predictions. We define two losses using cosine similarity cos and logit discrepancies $\mathcal{D}_{\star}^{\text{logit}}$ (MSE for AU, KL for expression), with feature/logit weights $\delta_{\text{feat}}, \delta_{\text{logit}}$ for consistency and $\eta_{\text{feat}}, \eta_{\text{logit}}$ for discrepancy:

Consistency loss $\mathcal{L}_{\text{consist}}$: Encourages output invariance under interventions weighted by the consistency mask:

$$\mathcal{L}_{\text{consist}} = \frac{1}{N_{\text{target}}} \sum_{j=1}^{N_{\text{target}}} \Big[\delta_{\text{feat}}\big(1 - \cos(Z^{(j)}, Z_{\text{CF}}^{(j,\text{consist})})\big) + \delta_{\text{logit}} \mathcal{D}_{\text{consist}}^{\text{logit}}\big(\hat{y}^{(j)}, \hat{y}_{\text{CF}}^{(j,\text{consist})}\big)\Big]. \tag{13}$$

Discrepancy loss $\mathcal{L}_{\text{discrep}}$: Encourages a target response to interventions weighted by the discrepancy mask:

$$\mathcal{L}_{\text{discrep}} = \frac{1}{N_{\text{target}}} \sum_{j=1}^{N_{\text{target}}} \Big[\eta_{\text{feat}}\big(1 + \cos(Z^{(j)}, Z_{\text{CF}}^{(j,\text{discrep})})\big) + \eta_{\text{logit}}\Big(1 - \mathcal{D}_{\text{discrep}}^{\text{logit}}\big(\hat{y}^{(j)}, \hat{y}_{\text{CF}}^{(j,\text{discrep})}\big)\Big)\Big]. \tag{14}$$

The final counterfactual objective weighted by λ_{consist} and λ_{discrep} is:

$$\mathcal{L}_{\text{cf}} = \lambda_{\text{consist}} \mathcal{L}_{\text{consist}} + \lambda_{\text{discrep}} \mathcal{L}_{\text{discrep}}. \tag{15}$$

3.5 Unified training objective

CausalAffect is trained end-to-end with disjoint AU/expression labels. AU-labeled samples supervise $\mathcal{L}_{\text{feat}}^{\text{AU}}$, $\mathcal{L}_{\text{g}}^{\text{AU}}$, and $\mathcal{L}_{\text{s}}^{\text{AU}}$, while expression-labeled samples supervise $\mathcal{L}_{\text{g}}^{\text{Expr}}$ and $\mathcal{L}_{\text{s}}^{\text{Expr}}$. $\mathcal{L}_{\text{ib}}$, $\mathcal{L}_{\text{decorr}}$, and $\mathcal{L}_{\text{cf}}$ do not require task labels, and $\mathcal{L}_{\text{align}}$ uses ground-truth or detached pseudo AU labels.

These objectives connect AU representation learning with causal graph learning. $\mathcal{L}_{\text{AU}}$ establishes semantically aligned, disentangled AU variables, while the global and sample-adaptive prediction losses supervise their causal relations at the population and instance levels. The DAG terms guide AU→AU topology, and $\mathcal{L}_{\text{cf}}$ constrains the intervention responses of the learned relations. AU disentanglement gives these interventions a specific semantic target, allowing CF to assess how intervening on an individual AU representation affects its downstream targets.

We evaluate the counterfactual loss separately for each branch, using $\mathcal{L}_{\text{cf}}^{\text{AU}}$ for AU detection and $\mathcal{L}_{\text{cf}}^{\text{Expr}}$ for expression recognition. The per-branch objectives are

$$\mathcal{L}_{\text{total}}^{\text{AU}} = \mathcal{L}_{\text{AU}} + \mathcal{L}_{\text{g}}^{\text{AU}} + \mathcal{L}_{\text{s}}^{\text{AU}} + \mathcal{L}_{\text{cf}}^{\text{AU}}, \qquad \mathcal{L}_{\text{total}}^{\text{Expr}} = \mathcal{L}_{\text{g}}^{\text{Expr}} + \mathcal{L}_{\text{s}}^{\text{Expr}} + \mathcal{L}_{\text{cf}}^{\text{Expr}}. \tag{16}$$

The final objective is

$$\mathcal{L} = \mathcal{L}_{\text{total}}^{\text{AU}} + \lambda_{\text{Expr}} \mathcal{L}_{\text{total}}^{\text{Expr}}, \tag{17}$$

where λ_{Expr} balances the expression branch. See Appendix A for training details.

4 Experiments and analysis

We assess the psychological plausibility of learned AU relations through literature comparisons and expert judgments. Six benchmarks evaluate AU detection and FER under disjoint supervision. Matched comparisons and ablations examine the roles of data composition, graph design, and counterfactual learning. We also assess robustness and efficiency. Appendix B details the setup.

4.1 Psychological plausibility of learned relations

Global causal graphs. Figure 2 shows the two signed graphs that CausalAffect learns under disjoint AU and expression supervision. The AU→Expression graph recovers AU12 as an established marker of Happiness [58, 14] and identifies an inhibitory link from AU6 to Sadness. The AU→AU graph captures both facilitatory links, such as AU6/AU7→AU12, and inhibitory links, such as AU26⊣AU20. The reverse link, AU20→AU26, has a weak positive weight. Symmetric co-occurrence assigns the same score to both directions and therefore leaves the source and target roles indistinguishable. These learned relations serve as **interpretable structural priors** for counterfactual reasoning. Each edge gives a specific hypothesis about which AU or expression prediction should respond to an intervention on the source AU representation. Appendix L compares the learned relations with empirical co-occurrence statistics and GNN-based correlations.

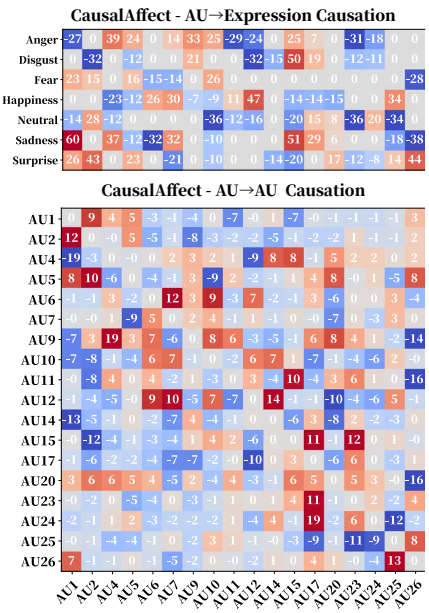

Figure 2: Signed global causal graphs.

Literature alignment. Table 1 compares positive edges in the learned AU→Expression graph with five published profiles from affective-science literature. The graph recovers markers listed in all five profiles, including AU12 for Happiness and AU1, AU4, and AU15 for Sadness. For Surprise, the graph recovers the three AUs shared by all five reference profiles (AU1, AU2, and AU5), as well as AU26, which appears in four of the five. For Disgust, it recovers AU9 and AU17, both listed in all five profiles, and AU15, which appears in two. Together, the Surprise and Disgust results show that agreement with the literature extends beyond the cues shared by all five profiles. For Fear, the five reference profiles agree on AU1, AU4, and AU5. The learned profile recovers AU1 and AU5 from this shared set. Independent studies also support learned links outside the five profiles, including AU7→Sadness [64] and AU10→Fear [3].

Table 1: Comparative AU→Expression relations from prior literature and our global causal graph. **Bold** entries denote AUs reported in the literature and also identified as positive by CausalAffect.

	[16]	[58]	[38]	[14]	[10]	Ours (Global Causal Graph)
Sadness	**1, 4, 15**	**1, 4**, 11, **15, 17**	**1, 4, 15, 17**	**1, 4**, 6, 11, **15, 17**	**1, 4, 15, 17**	**1, 4**, 7, **15, 17**
Surprise	**1, 2, 5, 26**	**1, 2, 5, 25**, 27	**1, 2, 5, 26**, 27	**1, 2, 5, 25, 26**	**1, 2, 5, 26**, 27	**1, 2, 5**, 20, **25, 26**
Happy	**6, 12**	**6, 12, 25**	**6, 12, 25**	**6, 12, 25**	**6, 12**	**6**, 7, 11, **12, 25**
Disgust	**9, 15, 17**	**9**, 10, **15, 17**	**9, 17**	4, **9**, 10, **17**, 24	**9**, 10, **17**	**9, 15, 17**
Fear	**1, 2**, 4, **5**, 20, 26	**1, 2**, 4, **5**	**1**, 4, **5**, 7	**1, 2**, 4, **5**, 20, 25, 26	**1, 2**, 4, **5**, 20, 25	**1, 2, 5**, 10
Anger	**4, 5, 7**, 23	**4, 5, 9, 10, 15, 17**, 23, 24	**4, 5, 7**, 23, 24	**4, 7, 10, 17**, 23, 24	**4, 5, 7, 10, 17**, 23, 25	**4, 5, 7, 9, 10, 15, 17**

Expert validation. To evaluate the psychological plausibility of the discovered relations, particularly the novel inhibitory ones, we conducted a blind study with five domain experts with 5–15 years of experience in affective cognition and facial behavior analysis. As shown in Table 2, CausalAffect attains an Expert True Rate of 0.75, substantially exceeding random controls (0.23) and approaching the literature-supported reference set (0.88). The Majority-True Rate is 0.84 overall, with strong endorsement for both AU→AU inhibition (0.83) and AU→Expression inhibition (0.87). The advantage also holds for high-confidence judgments (confidence scores of 4 or 5), with a True Rate of 0.70 for model-discovered relations versus 0.19 for random controls. The evaluation protocol and metric definitions are detailed in Appendix K.

Table 2: Expert validation statistics. Expert True Rate and Majority-True Rate ($\geq 3/5$); $\pm$ denotes SD across relations.

Relation Set / Subset	#Pairs	Expert True	Maj-True
I. Overall Comparison			
A: Literature-supported	39	0.88 ± 0.21	0.92
B: CausalAffect (Ours)	100	$\mathbf{0.75} \pm 0.24$	**0.84**
C: Random controls	96	0.23 ± 0.20	0.09
II. Breakdown of CausalAffect			
AU→AU (Excitatory)	32	0.72	0.78
AU→AU (Inhibitory)	23	**0.67**	**0.83**
AU→Expr (Excitatory)	22	0.85	0.91
AU→Expr (Inhibitory)	23	**0.77**	**0.87**

4.2 Predictive performance and training data composition

Benchmark performance. CausalAffect outperforms prior SOTA in AU detection and FER (Table 3). Appendix C details per-dataset and per-AU results, and Appendix P analyses AffectNet errors. In the single-dataset (SG) setting, it achieves 66.9 on BP4D, 67.0 on DISFA, and 66.4 on EmotioNet *without external data*. Under weakly supervised joint training, it further reaches 71.5 on DISFA, 62.5 on GFT with DISFA supervision, and the best FER results with BP4D supervision (67.7 on AffectNet and 85.3 on RAF-DB). To assess the relational design under matched supervision, we compare CausalAffect with an adapted AU-ViT baseline [61] and a ME-GraphAU-style relation module [59] using the same training data and splits. CausalAffect achieves higher mean performance in 18 of 19 heterogeneous-training comparisons (Appendix D).

Table 3: SOTA comparison on AU detection and FER. AU results are F1-score (%); FER results on AffectNet and RAF-DB are accuracy (%). CausalAffect values shown with $\pm$ denote mean $\pm$ SD over five runs. **SG**: single-dataset training; **+X**: joint training with dataset X; **+All**: all datasets.

Methods	BP4D	DISFA
JÂA-Net [72]	62.4	63.5
FAUT [36]	64.2	61.5
EmoLA [50]	64.2	65.1
FG-Net [92]	64.3	65.4
KSRL [5]	64.5	64.5
ReCoT [49]	64.8	62.6
MEGraph [59]	65.5	62.4
CLEF [97]	65.9	64.8
MCM [98]	66.0	64.3
MDHRM [86]	66.1	66.2
AUFormer [93]	66.2	66.4
CausalAffect (SG)	*66.9*±0.2	67.0±0.1
CausalAffect (+GFT)	66.6±0.1	*70.6*±0.3
CausalAffect (+EmotioNet)	**67.4**±0.3	71.1±0.2
CausalAffect (+RAF-DB)	67.1±0.1	69.5±0.2
CausalAffect (+AffectNet)	66.6±0.2	68.2±0.2
CausalAffect (+All)	66.7±0.2	**71.5**±0.2

EmotioNet	F1
ME-GraphAU [59]	64.9
JAA-Net [71]	51.8
AUNets [68]	64.6
FBNet [42]	54.0
CTC [101]	64.4
FUXI [96]	*65.4*
SITU [55]	64.2
CausalAffect (SG)	**66.4**±0.1
CausalAffect (+DISFA)	66.0±0.3
CausalAffect (+BP4D)	*65.4*±0.1
CausalAffect (+All)	65.0±0.1

GFT	F1
TCAE [51]	44.2
CDAU [18]	45.3
ARL [73]	50.1
MoCo [25]	52.4
TR [57]	54.7
JÂA-Net [72]	53.7
EmoCo [77]	58.6
EmoLA [50]	*62.1*
CausalAffect (SG)	61.1±0.2
CausalAffect (+DISFA)	**62.5**±0.3
CausalAffect (+BP4D)	60.4±0.4
CausalAffect (+All)	62.4±0.3

AffectNet / RAF-DB	AffN.	RAF.
MVT [46]	64.6	80.4
FBNet [43]	65.0	78.0
DACL [20]	65.2	80.4
ARM [74]	65.2	82.8
Ad-Corre [19]	63.4	79.0
DDAMFN++ [95]	67.4	84.6
FRA [22]	66.1	83.4
CAGE [82]	66.4	83.2
ExpLLM [45]	65.9	-
LOFI [94]	65.7	*84.7*
CausalAffect (+DISFA)	65.6±0.2	83.7±0.1
CausalAffect (+BP4D)	**67.7**±0.1	**85.3**±0.2
CausalAffect (+All)	*66.5*±0.1	84.9±0.2

Mechanism vs. data volume. Pooling datasets in standard multi-task learning (MTL) lowers scores on three AU benchmarks relative to single-task training. With the same backbone and pooled data as MTL, CausalAffect scores higher on all six benchmarks (Appendix D). Table 3 shows how the effect varies with the training mix. On EmotioNet, F1 falls from 66.4 in SG to 65.4 with BP4D and 65.0 with all datasets. Appendix E further shows that AU detection favors the eight most frequent AUs from BP4D, while FER performs best with the full set of 12 BP4D AUs, showing that the preferred AU coverage depends on the prediction task.

4.3 Graph adaptation and intervention analysis

Sample-adaptive graph. Figure 3 shows how SAC refines the global graph using instance-specific evidence. For AU→Expression, the global Sadness graph emphasizes canonical AUs such as AU1, AU15, and AU17, while the adaptive graph retains visually relevant cues such as AU1/AU17 and suppresses conflicting or unsupported AUs including AU5 and AU12. For AU→AU, the adaptive graph focuses on selected active smile-related AUs, including AU6, AU7, AU10, AU12, and AU25, capturing instance-specific excitatory and inhibitory relations among facial actions. Thus, SAC preserves psychology-aligned global structure while adapting to the AU configuration of each sample. Further examples of both relation types appear in Appendix M.

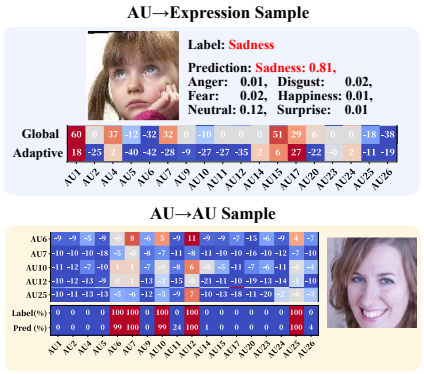

Figure 3: Sample-adaptive relations.

Graph ablations and recovery. GC leads SAC in 11 of 12 matched single-graph comparisons (Rows 3 vs. 7 and 4 vs. 8), and adding SAC to GC and CF improves all six benchmarks (Rows 5→13). Removing the DAG constraint lowers all six scores (Table 15, Rows 14–15). With both graphs, CF raises mean performance by 0.88 points without AU disentanglement (Rows 11→13) and by 3.87 with it (Rows 12→15). After perturbation reduces alignment to zero, CF restores all 13 displaced relations and removes all 13 injected assignments. With matched data order and training budgets, CF aligns 35 of 39 literature-supported AU→Expression relations versus 10 without CF. The same comparison shows gains of 3.8 points in RAF-DB accuracy and 4.1 in EmotioNet F1 (Appendix H).

Image-level and temporal evidence. For 36 source faces, Target and semantic Control edits match AU count and total requested intensity while targeting different expressions. CausalAffect predicts both intended expressions for 75.0% of faces versus 52.8% for Backbone + Dis. CausalAffect retains 34 original predictions under reconstruction without AU edits (Appendix N). The change in route strength favours the edited expression in 31/36 Target and 29/36 Control edits (Table 28). Across nine sampled frames of a smile sequence, excitatory links among AU6, AU12, AU25, and AU26 strengthen towards the smile apex and weaken as the expression fades (Appendix O).

4.4 Robustness and efficiency

Cross-dataset transfer and demographic robustness. With AffectNet-7 expression and BP4D AU supervision, CausalAffect reaches 78.1% accuracy on unseen RAF-DB without target adaptation, exceeding the strongest direct-transfer baseline (77.4%) in Appendix F. Following prior grouping protocols [26, 28], we use demographic annotations only to evaluate disparities alongside predictive utility [30, 53, 54]. On the same test set, accuracy is 77.8%, 77.4%, and 78.6% for the Asian, Black, and White groups, respectively, with a maximum gap of 1.2 percentage points. Across expression classes, the mean statistical-parity (SP) gap is 8.5 percentage points and the mean equal-opportunity (EOP) gap is 9.7 points (Appendix G). Appendix Q discusses scope and ethics.

Graph stability and computational efficiency. Global relations remain consistent across pseudo-label thresholds, AU vocabularies, and random seeds. Source-dataset swaps retain 89.7–94.2% polarity agreement on shared edges (Appendix I). In the largest configuration, training for 80 epochs on ∼800,000 images from six datasets takes ∼5.8 h on one NVIDIA A100, with 28–30 GB of peak GPU memory. Inference reaches ∼200 FPS on pre-cropped 112×112 faces (Appendix J).

5 Conclusion

We introduced **CausalAffect**, a causally guided framework for learning psychology-aligned AU-level relations under disjoint AU/expression supervision. CausalAffect combines AU bottleneck representations, global and sample-adaptive signed graphs, polarity-aware message passing, and feature-level counterfactual intervention to model AU→AU coordination and AU→Expression composition. Experiments on six benchmarks show strong AU detection and FER performance. Literature, expert judgments, and intervention-consistency analyses support the learned relations.

Acknowledgments

This work is supported by Queen Mary University of London. Guanyu Hu is currently a postdoctoral fellow at the Helsinki Institute for Information Technology (HIIT).

References

[1] C. Fabian Benitez-Quiroz, Ramprakash Srinivasan, and Aleix M. Martinez. Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5562–5570. IEEE, June 2016. doi: 10.1109/cvpr.2016.600. URL `http://dx.doi.org/10.1109/CVPR.2016.600`.

[2] Emilio Bizzi and Vincent CK Cheung. The neural origin of muscle synergies. *Frontiers in computational neuroscience*, 7:51, 2013.

[3] Jost U Blasberg, Mathilde Gallistl, Magdalena Degering, Felicitas Baierlein, and Veronika Engert. You look stressed: A pilot study on facial action unit activity in the context of psychosocial stress. *Comprehensive Psychoneuroendocrinology*, 15:100187, 2023.

[4] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.

[5] Yanan Chang and Shangfei Wang. Knowledge-driven self-supervised representation learning for facial action unit recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20417–20426, 2022.

[6] Chaona Chen, Daniel S Messinger, Cheng Chen, Hongmei Yan, Yaocong Duan, Robin AA Ince, Oliver GB Garrod, Philippe G Schyns, and Rachael E Jack. Cultural facial expressions dynamically convey emotion category and intensity information. *Current Biology*, 34(1): 213–223, 2024.

[7] Yuedong Chen, Xu Yang, Tat-Jen Cham, and Jianfei Cai. Towards unbiased visual emotion recognition via causal intervention. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 60–69, 2022.

[8] Yunliang Chen and Jungseock Joo. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991, 2021.

[9] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. Counterfactual fairness for facial expression recognition. In Leonid Karlinsky, Tomer Michaeli, and Ko Nishino, editors, *Computer Vision – ECCV 2022 Workshops*, pages 245–261, Cham, 2023. Springer Nature Switzerland. ISBN 978-3-031-25072-9.

[10] Elizabeth A Clark, J'Nai Kessinger, Susan E Duncan, Martha Ann Bell, Jacob Lahne, Daniel L Gallagher, and Sean F O'Keefe. The facial action coding system for characterization of human affective response to consumer product-based stimuli: a systematic review. *Frontiers in psychology*, 11:920, 2020.

[11] Andrea d'Avella, Philippe Saltiel, and Emilio Bizzi. Combinations of muscle synergies in the construction of a natural motor behavior. *Nature neuroscience*, 6(3):300–308, 2003.

[12] Abhinav Dhall, OV Ramana Murthy, Roland Goecke, Jyoti Joshi, and Tom Gedeon. Video and image based emotion recognition challenges in the wild: Emotiw 2015. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 423–426, 2015.

[13] Iris Dominguez-Catena, Daniel Paternain, and Mikel Galar. Metrics for dataset demographic bias: A case study on facial expression recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5209–5226, 2024.

[14] Shichuan Du, Yong Tao, and Aleix M Martinez. Compound facial expressions of emotion. *Proceedings of the national academy of sciences*, 111(15):E1454–E1462, 2014.

[15] P. Ekman and W.V. Friesen. *Facial Action Coding System*, volume 1. Consulting Psychologists Press, 1978. URL https://books.google.fi/books?id=08l6wgEACAAJ.

[16] Paul Ekman and Wallace V Friesen. Measuring facial movement. *Environmental psychology and nonverbal behavior*, 1(1):56–75, 1976.

[17] Phoebe C Ellsworth and Klaus R Scherer. Appraisal processes in emotion. In Richard J Davidson, Klaus R Scherer, and H Hill Goldsmith, editors, *Handbook of Affective Sciences*. Oxford University Press, 12 2002. ISBN 9780195126013. doi: 10.1093/oso/9780195126013.003.0029. URL https://doi.org/10.1093/oso/9780195126013.003.0029.

[18] Itir Onal Ertugrul, Jeffrey F Cohn, László A Jeni, Zheng Zhang, Lijun Yin, and Qiang Ji. Crossing domains for au coding: Perspectives, approaches, and measures. *IEEE transactions on biometrics, behavior, and identity science*, 2(2):158–171, 2020.

[19] Ali Pourramezan Fard and Mohammad H Mahoor. Ad-corre: Adaptive correlation-based loss for facial expression recognition in the wild. *IEEE Access*, 10:26756–26768, 2022.

[20] Amir Hossein Farzaneh and Xiaojun Qi. Facial expression recognition in the wild via deep attentive center loss. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2402–2411, 2021.

[21] Gege Gao, Huaibo Huang, Chaoyou Fu, and Ran He. Causal representation learning for context-aware face transfer. *arXiv preprint arXiv:2110.01571*, 2021.

[22] Zheng Gao and Ioannis Patras. Self-supervised facial representation learning with facial region awareness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2081–2092, 2024.

[23] Jeffrey M. Girard, Wen-Sheng Chu, Laszlo A. Jeni, Jeffrey F. Cohn, F. De La Torre, and Michael A. Sayette. Sayette group formation task (GFT) spontaneous facial expression database. In *Proceedings of the 12th IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, pages 581–588, Washington, DC, 2017. IEEE. doi: 10.1109/fg.2017.144.

[24] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *International conference on algorithmic learning theory*, pages 63–77. Springer, 2005.

[25] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738, 2020.

[26] Guanyu Hu, Eleni Papadopoulou, Dimitrios Kollias, Paraskevi Tzouveli, Jie Wei, and Xinyu Yang. Bridging the gap: Protocol towards fair and consistent affect analysis. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pages 1–9, 2024. doi: 10.1109/FG59268.2024.10582033.

[27] Guanyu Hu, Jie Wei, Siyang Song, Dimitrios Kollias, Xinyu Yang, Zhonglin Sun, and Odysseus Kaloidas. Robust facial reactions generation: An emotion-aware framework with modality compensation. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2024. doi: 10.1109/IJCB62174.2024.10744499.

[28] Guanyu Hu, Dimitrios Kollias, Eleni Papadopoulou, Paraskevi Tzouveli, Jie Wei, and Xinyu Yang. Rethinking affect analysis: A protocol for ensuring fairness and consistency. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 7(4):914–923, 2025. doi: 10.1109/TBIOM.2025.3550000.

[29] Guanyu Hu, Dimitrios Kollias, and Xinyu Yang. Grounding emotion recognition with visual prototypes: Vega - revisiting clip in merc. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, page 5667–5676, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755340. URL https://doi.org/10.1145/3746027.3755340.

[30] Guanyu Hu, Tangzheng Lian, Na Yan, Dimitrios Kollias, Xinyu Yang, Oya Celiktutan, Siyang Song, and Zeyu Fu. Fairmt: Fairness for heterogeneous multi-task learning. *arXiv preprint arXiv:2512.00469*, 2025.

[31] Guanyu Hu, Dimitrios Kollias, and Xinyu Yang. AURA: Visually interpretable affective understanding via robust archetypes. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 45181–45211. PMLR, 06–11 Jul 2026. URL `https://proceedings.mlr.press/v306/hu26z.html`.

[32] Guanyu Hu, Dimitrios Kollias, and Xinyu Yang. From cognitive priors to instance semantics: A unified framework for multi-task affective computing. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8551–8562, March 2026.

[33] Zi-Yu Huang, Chia-Chin Chiang, Jian-Hao Chen, Yi-Chian Chen, Hsin-Lung Chung, Yu-Ping Cai, and Hsiu-Chuan Hsu. A study on computer vision for facial emotion recognition. *Scientific reports*, 13(1):8425, 2023.

[34] Rachael E Jack, Oliver GB Garrod, Hui Yu, Roberto Caldara, and Philippe G Schyns. Facial expressions of emotion are not culturally universal. *Proceedings of the National Academy of Sciences*, 109(19):7241–7244, 2012.

[35] Rachael E Jack, Oliver GB Garrod, and Philippe G Schyns. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology*, 24(2):187–192, 2014.

[36] Geethu Miriam Jacob and Bjorn Stenger. Facial action unit detection with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7680–7689, 2021.

[37] Jaykumar Kakkad, Jaspal Jannu, Kartik Sharma, Charu Aggarwal, and Sourav Medya. A survey on explainability of graph neural networks. *arXiv preprint arXiv:2306.01958*, 2023.

[38] K Karthick and J Jasmine. Survey of advanced facial feature tracking and facial expression recognition. *International Journal of Advanced Research in Computer and Communication Engineering*, 2(10):3872–3876, 2013.

[39] Pooya Khorrami, Tom Le Paine, and Thomas S. Huang. Do deep neural networks learn facial action units when doing expression recognition? In *2015 IEEE International Conference on Computer Vision Workshop (ICCVW)*, pages 19–27. IEEE, December 2015. doi: 10.1109/iccvw.2015.12. URL `http://dx.doi.org/10.1109/ICCVW.2015.12`.

[40] Mojtaba Kolahdouzi, Alireza Sepas-Moghaddam, and Ali Etemad. Face trees for expression recognition. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 1–5. IEEE, 2021.

[41] Dimitrios Kollias and Stefanos Zafeiriou. Affect analysis in-the-wild: Valence-arousal, expressions, action units and a unified framework. *arXiv preprint arXiv:2103.15792*, 2021.

[42] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Face behavior a la carte: Expressions, affect and action units in a single network. *arXiv preprint arXiv:1910.11111*, 2019.

[43] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for heterogeneous multi-task learning: a large-scale face study. *arXiv preprint arXiv:2105.03790*, 2021.

[44] Dimitrios Kollias, Viktoriia Sharmanska, and Stefanos Zafeiriou. Distribution matching for multi-task learning of classification tasks: a large-scale study on faces & beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2813–2821, 2024.

[45] Xing Lan, Jian Xue, Ji Qi, Dongmei Jiang, Ke Lu, and Tat-Seng Chua. Expllm: Towards chain of thought for facial expression recognition. *IEEE Transactions on Multimedia*, 2025.

[46] Hanting Li, Mingzhe Sui, Feng Zhao, Zhengjun Zha, and Feng Wu. Mvt: mask vision transformer for facial expression recognition in the wild. *arXiv preprint arXiv:2106.04520*, 2021.

[47] Shan Li and Weihong Deng. Deep facial expression recognition: A survey. *IEEE Transactions on Affective Computing*, 13(3):1195–1215, 2022. doi: 10.1109/TAFFC.2020.2981446.

[48] Shan Li, Weihong Deng, and JunPing Du. Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2852–2861, 2017.

[49] Yifan Li, Hu Han, Shiguang Shan, Zhilong Ji, Jinfeng Bai, and Xilin Chen. Recot: Regularized co-training for facial action unit recognition with noisy labels. In *BMVC*, page 102, 2023. URL `http://proceedings.bmvc2023.org/102/`.

[50] Yifan Li, Anh Dao, Wentao Bao, Zhen Tan, Tianlong Chen, Huan Liu, and Yu Kong. Facial affective behavior analysis with instruction tuning. In *European Conference on Computer Vision*, pages 165–186. Springer, 2024.

[51] Yong Li, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Self-supervised representation learning from videos for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer vision and pattern recognition*, pages 10924–10933, 2019.

[52] Yong Li, Menglin Liu, Lingjie Lao, Yuanzhi Wang, and Zhen Cui. Counterfactual discriminative micro-expression recognition. *Visual Intelligence*, 2(1):29, 2024.

[53] Tangzheng Lian, Guanyu Hu, Dimitrios Kollias, Xinyu Yang, and Oya Celiktutan. Fair domain generalization: An information-theoretic view. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(28):23382–23390, Mar. 2026. doi: 10.1609/aaai.v40i28.39508. URL `https://ojs.aaai.org/index.php/AAAI/article/view/39508`.

[54] Tangzheng Lian, Guanyu Hu, Yijing Ren, Dimitrios Kollias, and Oya Celiktutan. A closed-form solution for debiasing vision-language models with utility guarantees across modalities and tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 31672–31682, June 2026.

[55] Chuanhe Liu, Xinjie Zhang, Xiaolong Liu, Tenggan Zhang, Liyu Meng, Yuchen Liu, Yuanyuan Deng, and Wenqiang Jiang. Facial expression recognition based on multi-modal features for videos in the wild. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5872–5879. IEEE, 2023.

[56] Zhilei Liu, Jiahui Dong, Cuicui Zhang, Longbiao Wang, and Jianwu Dang. Relation modeling with graph convolutional networks for facial action unit detection. In *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*, pages 489–501. Springer, 2020.

[57] Liupei Lu, Leili Tavabi, and Mohammad Soleymani. Self-supervised learning for facial action unit recognition through temporal consistency. In *BMVC*, 2020.

[58] 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*, pages 94–101. IEEE, 2010.

[59] Cheng Luo, Siyang Song, Weicheng Xie, Linlin Shen, and Hatice Gunes. Learning multi-dimensional edge feature-based au relation graph for facial action unit recognition. *arXiv preprint arXiv:2205.01782*, 2022.

[60] Jiawei Mao, Rui Xu, Xuesong Yin, Yuanqi Chang, Binling Nie, Aibin Huang, and Yigang Wang. Poster++: A simpler and stronger facial expression recognition network. *Pattern Recognition*, 157:110951, 2025.

[61] Shuyi Mao, Xinpeng Li, Fan Zhang, Xiaojiang Peng, and Yang Yang. Facial action units as a joint dataset training bridge for facial expression recognition. *IEEE Transactions on Multimedia*, 27:3331–3342, 2025. doi: 10.1109/TMM.2025.3535327.

[62] S Mohammad Mavadati, Mohammad H Mahoor, Kevin Bartlett, Philip Trinh, and Jeffrey F Cohn. Disfa: A spontaneous facial action intensity database. *IEEE Transactions on Affective Computing*, 4(2):151–160, 2013.

[63] Thomas Melistas, Nikos Spyrou, Nefeli Gkouti, Pedro Sanchez, Athanasios Vlontzos, Yannis Panagakis, Giorgos Papanastasiou, and Sotirios A Tsaftaris. Benchmarking counterfactual image generation. *Advances in Neural Information Processing Systems*, 37:133207–133230, 2024.

[64] Elizabeth J Miller, Eva G Krumhuber, and Amy Dawel. Observers perceive the duchenne marker as signaling only intensity for sad expressions, not genuine emotion. *Emotion*, 22(5): 907–919, 2022.

[65] Ali Mollahosseini, Behzad Hasani, and Mohammad H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Transactions on Affective Computing*, 10(1):18–31, January 2019. ISSN 2371-9850. doi: 10.1109/taffc.2017.2740923. URL `http://dx.doi.org/10.1109/TAFFC.2017.2740923`.

[66] Yushu Pan and Elias Bareinboim. Counterfactual image editing. *arXiv preprint arXiv:2403.09683*, 2024.

[67] Guruprasad Viswanathan Ramesh, Ashish Hooda, Shimaa Ahmed, Harrison J Rosenberg, Ramya Korlakai Vinayak, and Kassem Fawaz. Counterface: A synthetic face dataset for fine-grained counterfactual evaluation of face recognition systems, 2026. URL `https://arxiv.org/abs/2407.13922`.

[68] Andrés Romero, Juan León, and Pablo Arbeláez. Multi-view dynamic facial action unit detection. *Image and Vision Computing*, 122:103723, 2022.

[69] Harold A Sackeim and Ruben C Gur. Response: Asymmetry in facial expression. *Science*, 209(4458):834–836, 1980.

[70] Klaus R Scherer and Heiner Ellgring. Are facial expressions of emotion produced by categorical affect programs or dynamically driven by appraisal? *Emotion*, 7(1):113–130, 2007.

[71] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Deep adaptive attention for joint facial action unit detection and face alignment. In *Proceedings of the European conference on computer vision (ECCV)*, pages 705–720, 2018.

[72] Zhiwen Shao, Zhilei Liu, Jianfei Cai, and Lizhuang Ma. Jaa-net: Joint facial action unit detection and face alignment via adaptive attention. *International Journal of Computer Vision*, 129:321–340, 2021.

[73] Zhiwen Shao, Zhilei Liu, Jianfei Cai, Yunsheng Wu, and Lizhuang Ma. Facial action unit detection using attention and relation learning. *IEEE Transactions on Affective Computing*, 13(3):1274–1289, July 2022. ISSN 2371-9850. doi: 10.1109/taffc.2019.2948635. URL `http://dx.doi.org/10.1109/TAFFC.2019.2948635`.

[74] Jiawei Shi, Songhao Zhu, and Zhiwei Liang. Learning to amend facial expression representation via de-albino and affinity. *arXiv preprint arXiv:2103.10189*, 2021.

[75] Tengfei Song, Lisha Chen, Wenming Zheng, and Qiang Ji. Uncertain graph neural networks for facial action unit detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(7):5993–6001, May 2021. doi: 10.1609/aaai.v35i7.16748. URL `https://ojs.aaai.org/index.php/AAAI/article/view/16748`.

[76] Tengfei Song, Zijun Cui, Yuru Wang, Wenming Zheng, and Qiang Ji. Dynamic probabilistic graph convolution for facial action unit intensity estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4845–4854, 2021.

[77] Xuran Sun, Jiabei Zeng, and Shiguang Shan. Emotion-aware contrastive learning for facial action unit detection. In *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*, pages 01–08. IEEE, 2021.

[78] Veikko Surakka and Jari K Hietanen. Facial and emotional reactions to duchenne and non-duchenne smiles. *International journal of psychophysiology*, 29(1):23–33, 1998.

[79] Y-I Tian, Takeo Kanade, and Jeffrey F Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 23(2):97–115, 2001.

[80] Naftali Tishby and Noga Zaslavsky. Deep learning and the information bottleneck principle. In *2015 ieee information theory workshop (itw)*, pages 1–5. Ieee, 2015.

[81] John Turnbull, Shivam Grover, Amin Jalali, and Ali Etemad. Evaluating test-time adaptation for facial expression recognition under natural cross-dataset distribution shifts. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2066–2070. IEEE, 2026.

[82] Niklas Wagner, Felix Mätzler, Samed R Vossberg, Helen Schneider, Svetlana Pavlitska, and J Marius Zöllner. Cage: Circumplex affect guided expression inference. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4683–4692. IEEE, 2024.

[83] Tao Wang, Xiangwei Zheng, Lifeng Zhang, Zhen Cui, and Chunyan Xu. A graph-based interpretability method for deep neural networks. *Neurocomputing*, 555:126651, 2023.

[84] Xiaofeng Wang, Tianbo Han, Songling Liu, Muhammad Shahroz Ajmal, Lu Chen, Yongqin Zhang, and Yonghuai Liu. Mhan: Multi-head hybrid attention network for facial expression recognition. *Pattern Recognition*, 170:112015, 2026. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2025.112015. URL `https://www.sciencedirect.com/science/article/pii/S0031320325006752`.

[85] Zihan Wang, Siyang Song, Cheng Luo, Yuzhi Zhou, Shiling Wu, Weicheng Xie, and Linlin Shen. Spatial-temporal graph-based au relationship learning for facial action unit detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5899–5907, June 2023.

[86] Zihan Wang, Siyang Song, Cheng Luo, Songhe Deng, Weicheng Xie, and Linlin Shen. Multi-scale dynamic and hierarchical relationship modeling for facial action units recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1270–1280, 2024.

[87] Jie Wei, Guanyu Hu, Xinyu Yang, Anh Tuan Luu, and Yizhuo Dong. Audio-visual domain adaptation feature fusion for speech emotion recognition. In *INTERSPEECH*, pages 1988–1992, 2022.

[88] Jie Wei, Guanyu Hu, Luu Anh Tuan, Xinyu Yang, and Wenjing Zhu. Multi-scale receptive field graph model for emotion recognition in conversations. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10094596.

[89] Jie Wei, Guanyu Hu, Xinyu Yang, Anh Tuan Luu, and Yizhuo Dong. Learning facial expression and body gesture visual information for video emotion recognition. *Expert Systems with Applications*, 237:121419, 2024. ISSN 0957-4174. doi: https://doi.org/10.1016/j.eswa.2023.121419. URL `https://www.sciencedirect.com/science/article/pii/S0957417423019218`.

[90] Jiannan Yang, Fan Zhang, Bike Chen, and Samee U Khan. Facial expression recognition based on facial action unit. In *2019 tenth international green and sustainable computing conference (IGSC)*, pages 1–6. IEEE, 2019.

[91] Dong Yi, Zhen Lei, Shengcai Liao, and Stan Z Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[92] Yufeng Yin, Di Chang, Guoxian Song, Shen Sang, Tiancheng Zhi, Jing Liu, Linjie Luo, and Mohammad Soleymani. Fg-net: Facial action unit detection with generalizable pyramidal features. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6099–6108, 2024.

[93] Kaishen Yuan, Zitong Yu, Xin Liu, Weicheng Xie, Huanjing Yue, and Jingyu Yang. Auformer: Vision transformers are parameter-efficient facial action unit detectors. In *European Conference on Computer Vision*, pages 427–445. Springer, 2024.

[94] Jinglin Zhang, Qiangchang Wang, Xinxin Zhang, and Yilong Yin. Lofi: Harnessing attention dynamics for facial expression recognition with noisy labels. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[95] Saining Zhang, Yuhang Zhang, Ye Zhang, Yufei Wang, and Zhigang Song. A dual-direction attention mixed feature network for facial expression recognition. *Electronics*, 12(17):3595, 2023.

[96] Wei Zhang, Bowen Ma, Feng Qiu, and Yu Ding. Multi-modal facial affective analysis based on masked autoencoder. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5793–5802. IEEE, 2023.

[97] Xiang Zhang, Taoyue Wang, Xiaotian Li, Huiyuan Yang, and Lijun Yin. Weakly-supervised text-driven contrastive learning for facial behavior understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20751–20762, 2023.

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

[99] Xing Zhang, Lijun Yin, Jeffrey F Cohn, Shaun Canavan, Michael Reale, Andy Horowitz, Peng Liu, and Jeffrey M Girard. Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database. *Image and Vision Computing*, 32(10):692–706, 2014.

[100] Xun Zheng, Bryon Aragam, Pradeep K Ravikumar, and Eric Xing. Dags with no tears: Continuous optimization for structure learning. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL `https://proceedings.neurips.cc/paper_files/paper/2018/file/e347c51419ffb23ca3fd5050202f9c3d-Paper.pdf`.

[101] Weiwei Zhou, Jiada Lu, Zhaolong Xiong, and Weifeng Wang. Leveraging tcn and transformer for effective visual-audio fusion in continuous emotion recognition. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5756–5763. IEEE, 2023.

A Additional method details

A.1 HSIC computation.

We use the empirical Hilbert–Schmidt independence criterion (HSIC) estimator [24] to compute the dependence between any pair of random variables X and Y based on their batch-wise representations. Given n samples $\{(x_k, y_k)\}_{k=1}^{n}$, we first compute the kernel matrices K and L using Gaussian radial basis function (RBF) kernels:

$$K_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_x^2}\right), \quad L_{ij} = \exp\left(-\frac{\|y_i - y_j\|^2}{2\sigma_y^2}\right), \tag{18}$$

where σ_x and σ_y are bandwidth parameters (either fixed or estimated via median heuristic). The empirical HSIC is then computed as:

$$\mathcal{H}(X, Y) = \frac{1}{(n-1)^2}\mathrm{tr}(KHLH), \tag{19}$$

where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ centres the kernel features, with I the $n \times n$ identity and $\mathbf{1}$ the length-n all-ones vector. This formulation enables empirical estimation of the squared Hilbert-Schmidt norm of the cross-covariance operator between X and Y in their respective reproducing kernel Hilbert spaces. In our implementation, we apply the above formulation to all three losses $\mathcal{L}_{\text{ib}}$, $\mathcal{L}_{\text{align}}$, and $\mathcal{L}_{\text{decorr}}$ using mini-batch representations of AU heads, the global image embedding z_{img}, and corresponding (pseudo-)labels. This HSIC-based framework offers a unified and theoretically grounded approach to disentanglement, without relying on adversarial learning or variational estimation.

A.2 Weakly supervised training details

CausalAffect is trained under disjoint AU/expression supervision. AU datasets provide labels of the form $(I, \mathbf{y}^{\mathrm{AU}})$, while expression datasets provide labels of the form (I, y^{Expr}). No image is assumed to have paired AU→Expression annotations. The model connects these two label spaces through the shared AU bottleneck and signed AU-level relation graphs.

A.2.1 Loss masking under disjoint supervision

Let $\mathcal{B}_{\mathrm{AU}}$ and $\mathcal{B}_{\mathrm{Expr}}$ denote mini-batches drawn from AU-labeled and expression-labeled datasets, respectively. Their union is:

$$\mathcal{B} = \mathcal{B}_{\mathrm{AU}} \cup \mathcal{B}_{\mathrm{Expr}}. \tag{20}$$

All samples are passed through the same backbone, AU bottleneck encoder, global graph, and sample-adaptive graph. However, supervised prediction losses are activated only when the corresponding labels are available. For AU-labeled samples, we apply

$$\mathcal{L}_{\mathrm{sup}}^{\mathrm{AU}} = \mathcal{L}_{\mathrm{feat}}^{\mathrm{AU}} + \mathcal{L}_{\mathrm{g}}^{\mathrm{AU}} + \mathcal{L}_{\mathrm{s}}^{\mathrm{AU}}. \tag{21}$$

For expression-labeled samples, we apply

$$\mathcal{L}_{\mathrm{sup}}^{\mathrm{Expr}} = \mathcal{L}_{\mathrm{g}}^{\mathrm{Expr}} + \mathcal{L}_{\mathrm{s}}^{\mathrm{Expr}}. \tag{22}$$

Thus, AU labels are never required for expression datasets, and expression labels are never required for AU datasets.

A.2.2 Pseudo-label construction

For each sample and AU, semantic alignment uses the ground-truth label when available. When the corresponding AU annotation is missing, we construct a confidence-filtered pseudo-label from the AU bottleneck prediction. These pseudo-labels are used only for HSIC-based semantic alignment; supervised AU prediction losses use available ground-truth annotations.

For each image $I_k \in \mathcal{B}$ with a missing annotation for AU i, the AU bottleneck classifier predicts

$$p_k^{(i)} = \sigma\Big(h_i(\mathbf{f}_{\mathrm{AU},k}^{(i)})\Big), \tag{23}$$

where $\mathbf{f}_{\mathrm{AU},k}^{(i)}$ is the feature of AU i for image I_k, $h_i(\cdot)$ is its classifier, $\sigma(\cdot)$ is the sigmoid function, and $p_k^{(i)}$ is the predicted activation probability. Let $r_k^{(i)} = 1$ indicate that image I_k has a ground-truth

Algorithm 1 Weakly Supervised Training of CausalAffect

Require: AU-labeled datasets $\mathcal{D}_{\mathrm{AU}}$, expression-labeled datasets $\mathcal{D}_{\mathrm{Expr}}$; backbone f_θ with parameters θ; AU bottleneck heads $\{\phi_i\}_{i=1}^{N_{\mathrm{AU}}}$; global graph parameters $\mathcal{E}_{\mathrm{logit}}$; heterogeneous polarity-aware graph attention (HPGAT) parameters; classifier heads $\Phi_{\mathrm{AU}}, \Phi_{\mathrm{Expr}}$; T training iterations

for training iteration $t = 1, \ldots, T$ **do**
  Sample $\mathcal{B}_{\mathrm{AU}}^{(t)} \sim \mathcal{D}_{\mathrm{AU}}$ and $\mathcal{B}_{\mathrm{Expr}}^{(t)} \sim \mathcal{D}_{\mathrm{Expr}}$
  Form mixed batch $\mathcal{B}^{(t)} = \mathcal{B}_{\mathrm{AU}}^{(t)} \cup \mathcal{B}_{\mathrm{Expr}}^{(t)}$
  Extract image features $z_{\mathrm{img},k} = f_\theta(I_k)$ for all $I_k \in \mathcal{B}^{(t)}$
  Compute AU bottleneck features $\mathbf{F}_{\mathrm{AU},k} = \{\phi_i(z_{\mathrm{img},k})\}_{i=1}^{N_{\mathrm{AU}}}$
  Construct detached confidence-filtered pseudo-labels for missing AU annotations
  Construct global signed graph $\mathcal{E}_{\mathrm{g}} = \tanh(\tau \mathcal{E}_{\mathrm{logit}})$
  Infer sample-adaptive signed graph $\widetilde{\mathcal{E}}_{\mathrm{s},k}$ using HPGAT and gated refinement
  Perform polarity-aware message aggregation for AU→AU and AU→Expression branches
  Compute supervised AU prediction losses using available ground-truth AU annotations
  Compute expression losses on expression-labeled samples only
  Compute HSIC losses $\mathcal{L}_{\mathrm{ib}}$, $\mathcal{L}_{\mathrm{align}}$, and $\mathcal{L}_{\mathrm{decorr}}$
  Build edge-aware counterfactual masks and compute $\mathcal{L}_{\mathrm{cf}}$
  Aggregate all active losses and update parameters using AdamW
end for

annotation for AU i, and $r_k^{(i)} = 0$ otherwise. The alignment target combines available ground-truth labels with detached pseudo-labels:

$$\tilde{y}_{\mathrm{AU},k}^{(i)} = \begin{cases} y_{\mathrm{AU},k}^{(i)}, & r_k^{(i)} = 1, \\ 1, & r_k^{(i)} = 0,\ \mathrm{sg}(p_k^{(i)}) \geq \rho_+, \\ 0, & r_k^{(i)} = 0,\ \mathrm{sg}(p_k^{(i)}) \leq \rho_-, \\ \varnothing, & \text{otherwise}, \end{cases} \tag{24}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The uncertain case $\varnothing$ is ignored in the semantic alignment term. We set the positive and negative confidence thresholds to $\rho_+ = 0.8$ and $\rho_- = 0.2$, respectively, across all datasets.

The semantic alignment set for AU i is

$$\mathcal{B}_{\mathrm{align}}^{(i)} = \{k \mid I_k \in \mathcal{B},\ \tilde{y}_{\mathrm{AU},k}^{(i)} \neq \varnothing\}. \tag{25}$$

The alignment loss is computed as

$$\mathcal{L}_{\mathrm{align}} = -\frac{1}{N_{\mathrm{AU}}} \sum_{i=1}^{N_{\mathrm{AU}}} \mathcal{H}\left(\{\mathbf{f}_{\mathrm{AU},k}^{(i)}\}_{k \in \mathcal{B}_{\mathrm{align}}^{(i)}}, \{\tilde{y}_{\mathrm{AU},k}^{(i)}\}_{k \in \mathcal{B}_{\mathrm{align}}^{(i)}}\right). \tag{26}$$

If too few confident samples are available for AU i in a mini-batch, this term is skipped for that AU. The stop-gradient operation treats the generated pseudo-labels as fixed targets for learning semantically aligned AU features.

A.2.3 Training algorithm

Algorithm 1 summarizes the end-to-end training procedure. At each iteration, CausalAffect samples AU-labeled and expression-labeled mini-batches, applies supervised losses only to samples with the corresponding labels, computes representation regularizers, and optimizes the AU bottleneck together with the global and sample-adaptive signed graphs.

A.2.4 Inference procedure

At inference time, no pseudo-labels or counterfactual interventions are required. Given a test image, CausalAffect extracts AU bottleneck features, constructs the global and sample-adaptive graphs, and predicts either AU activations or expression probabilities through polarity-aware message aggregation. The sample-adaptive graph can also be retained as an instance-level explanation of which AU relations support or suppress the prediction.

A.3 Sample-adaptive message aggregation

Sample-adaptive message aggregation translates the learned causal edges into target representations while maintaining separate pathways for excitatory and inhibitory contributions. It combines population-level causal structure with the facial evidence in each image, allowing the same AU variables to support both AU→AU and AU→Expression reasoning. The computation proceeds from signed edge decomposition to gated fusion and target aggregation.

Signed edge decomposition. For sample k at layer l, $\mathcal{E}_{\mathrm{s}}^{(k,l)} \in \mathbb{R}^{N_{\text{target}} \times N_{\text{source}}}$ contains the signed edges inferred by Eq. 7. Entry (j, i) describes the causal influence of source AU i on target j; targets are AUs or expressions according to the branch. The polarity decomposition in Eq. 4 is applied element-wise: positive edges enter the excitatory pathway, while the magnitudes of negative edges enter the inhibitory pathway. Explicitly,

$$\mathcal{E}_{\mathrm{s}}^{+(k,l)} = \mathbb{I}\big(\mathcal{E}_{\mathrm{s}}^{(k,l)} > 0\big) \odot \mathcal{E}_{\mathrm{s}}^{(k,l)}, \qquad \mathcal{E}_{\mathrm{s}}^{-(k,l)} = \mathbb{I}\big(\mathcal{E}_{\mathrm{s}}^{(k,l)} < 0\big) \odot \big(-\mathcal{E}_{\mathrm{s}}^{(k,l)}\big). \tag{27}$$

Both component matrices contain non-negative magnitudes, with edge polarity retained by pathway membership. Each non-zero edge contributes to one pathway. The source values $V_{\mathrm{s}}^{(k,l)} = \mathbf{F}_{\mathrm{AU}}^{(k)}$ are the disentangled AU features shared across layers. Separate learnable projections $W_V^{+(l)}$ and $W_V^{-(l)}$ transform these values before aggregation, enabling the two pathways to represent the distinct roles of excitatory and inhibitory causal relations.

Fusion of global and sample-specific relations. At the final layer L, Eq. 9 combines the global edge matrix $\mathcal{E}_{\mathrm{g}}$ with the sample-specific matrix $\mathcal{E}_{\mathrm{s}}^{(k,L)}$. The edge-wise gate $\mathbf{G}^{(k)}$ controls their relative contributions: a larger gate value places more weight on the population-level relation, and a smaller value places more weight on the relation inferred for the current image. This allows individual connections to adapt to sample-specific facial evidence while retaining the global causal structure as a shared reference. We then decompose the fused matrix $\widetilde{\mathcal{E}}_{\mathrm{s}}^{(k)}$ into $\widetilde{\mathcal{E}}_{\mathrm{s}}^{+(k)}$ and $\widetilde{\mathcal{E}}_{\mathrm{s}}^{-(k)}$ using the same rule, so message routing follows the polarity of the final fused edge.

Target aggregation and prediction. The fused edge components weight the separately projected source values to obtain the final target embeddings:

$$Z_{\mathrm{s}}^{(k)} = \widetilde{\mathcal{E}}_{\mathrm{s}}^{+(k)} V_{\mathrm{s}}^{(k,L)} W_V^{+(L)} + \widetilde{\mathcal{E}}_{\mathrm{s}}^{-(k)} V_{\mathrm{s}}^{(k,L)} W_V^{-(L)}. \tag{28}$$

Each row of $Z_{\mathrm{s}}^{(k)}$ aggregates incoming AU messages for one target. The two terms collect excitatory and inhibitory messages through their respective projections and combine them in the target representation. The resulting AU and expression embeddings are passed to the corresponding heads Φ_{AU} and Φ_{Expr} and supervised by the sample-adaptive graph losses defined in Section 3.3. Predictions therefore use AU evidence organised by the fused directed, polarity-aware causal graph, linking instance-specific relation refinement to the downstream recognition tasks.

B Experimental setup

B.1 Datasets

We evaluate CausalAffect on six widely-used facial analysis datasets: BP4D, DISFA, EmotioNet, GFT, RAF-DB, and AffectNet. Among them, BP4D, DISFA and GFT provide frame-level annotations of facial AUs, while EmotioNet, RAF-DB and AffectNet offer image-level labels.

AU datasets. We follow [72, 44] for consistent AU dataset splitting and evaluation. *BP4D*[99] contains 41 subjects with spontaneous expressions annotated over 12 AUs. *DISFA* [62] contains spontaneous facial recordings, for which we use 8 selected AUs. *EmotioNet* [1] consists of over 45K in-the-wild facial images, where we follow the official split and use the 11 most frequent AUs for training and evaluation. *GFT* [23] is a proprietary dataset containing over 130K high-quality face images, for which we use 11 AUs for training. For all AU datasets, we binarize labels using intensity thresholds if available, and compute AU-wise and average F1 scores across the selected AU subsets in each dataset.

Expression datasets. *RAF-DB* [48] contains around 15K images with crowd-sourced annotations of 7 basic expressions. *AffectNet* [65] provides over 250K manually labeled facial images, from which we use the standard 7-class subset (Neutral + 6 basic emotions) for evaluation. We adopt the official train/validation splits provided by both datasets.

B.2 Implementation details

The backbone is ConvNeXt-Base pretrained on WebFace [91]. Input images are resized to 112×112 and globally pooled to obtain $z_{\text{img}} \in \mathbb{R}^{1024}$. AU-specific features $\mathbf{f}_{\text{AU}}^{(i)} \in \mathbb{R}^{64}$ are extracted. HSIC-based disentanglement is applied using RBF kernels with median bandwidth, optimizing $\mathcal{L}_{\text{ib}}$, $\mathcal{L}_{\text{align}}$, and $\mathcal{L}_{\text{decorr}}$ with weights 1.0, 1.0, and 0.2 respectively. We initialize global edge logits in $[-0.01, 0.01]$, use signed 8-head attention, and set both soft DAG weights to $\lambda_{\text{DAG}}^{\text{g}} = \lambda_{\text{DAG}}^{\text{s}} = 0.5$. Sample-adaptive graphs are computed via a 2-layer attention. AU→Expr graphs use learned 64-dimensional expression prototypes as attention queries. Counterfactual intervention uses noise standard deviation $\tau_{\text{cf}} = 0.2$ with gate sharpness $\gamma = 10$ and loss weights $\lambda_{\text{consist}} = \lambda_{\text{discrep}} = 0.5$, $\delta_{\text{feat}} = \delta_{\text{logit}} = \eta_{\text{feat}} = \eta_{\text{logit}} = 1.0$. For CausalAffect (+All), we train for 80 epochs using AdamW (learning rate $1\text{e}{-4}$, weight decay $1\text{e}{-5}$) with a batch size of 360. Table 23 lists the batch sizes and training epochs for representative training configurations.

C Full performance comparison and analysis

This section provides the full benchmark comparison of CausalAffect across six datasets, including four AU detection benchmarks (BP4D, DISFA, EmotioNet, and GFT) and two FER benchmarks (AffectNet and RAF-DB). For AU detection, we use AU-wise and average F1-score (%), while for FER we use classification accuracy (%). **SG** denotes single-dataset training without external data. **+Dataset** denotes weakly supervised joint training with an additional disjoint dataset, and **+All** denotes joint training with all available datasets. For CausalAffect variants, values denote mean and standard deviation over five runs when available. We limit statistical tests to our reproduced comparisons with matched random seeds and training splits, since most published results lack these controls.

The full results serve three purposes. First, they verify whether CausalAffect remains competitive under strict single-dataset training, where performance cannot be attributed to additional data. Second, they examine whether weakly supervised auxiliary data improves performance consistently or only when the auxiliary AU/expression evidence is compatible with the target dataset. Third, the per-AU tables reveal which facial actions benefit most from signed relation modeling and where domain mismatch or AU-frequency imbalance can still limit transfer.

C.1 BP4D AU detection

Table 4: Full per-AU F1-score (%) comparison on BP4D.

Methods	BP4D												
	AU1	AU2	AU4	AU6	AU7	AU10	AU12	AU14	AU15	AU17	AU23	AU24	Avg.
JÂA-Net [72]	53.8	47.8	58.2	78.5	75.8	82.7	88.2	63.7	43.3	61.8	45.6	49.9	62.4
FAUT [36]	51.7	49.3	61.0	77.8	79.5	82.9	86.3	67.6	51.9	63.0	43.7	56.3	64.2
EmoLA [50]	57.4	**52.4**	61.0	78.1	77.8	81.9	**89.5**	60.5	49.3	64.9	46.0	52.4	64.2
FG-Net [92]	-	-	-	-	-	-	-	-	-	-	-	-	64.3
KSRL [5]	53.3	47.4	56.2	79.4	**80.7**	85.1	89.0	67.4	55.9	61.9	48.5	49.0	64.5
ReCoT [49]	51.5	47.8	58.9	79.2	80.2	84.9	88.4	61.5	53.3	64.6	51.8	55.4	64.8
MEGraph [59]	52.7	44.3	60.9	79.9	80.1	85.3	89.2	69.4	55.4	64.4	49.8	55.1	65.5
CLEF [97]	55.8	46.8	63.3	79.5	77.6	83.6	87.8	67.3	55.2	63.5	**53.0**	57.8	65.9
MCM [98]	54.4	48.5	60.6	79.1	77.0	84.0	89.1	61.7	**59.3**	64.7	**53.0**	**60.5**	66.0
MDHRM [86]	54.6	49.7	61.0	79.9	79.4	85.4	88.5	67.8	56.8	63.2	50.9	55.4	66.1
AUFormer [93]	-	-	-	-	-	-	-	-	-	-	-	-	66.2
CausalAffect (SG)	63.4	47.1	56.8	**82.1**	78.8	**85.8**	88.8	71.4	57.5	**67.9**	46.2	56.4	66.9±0.2
CausalAffect (+GFT)	63.7	41.4	60.7	80.0	76.8	85.7	88.5	**71.5**	55.9	67.0	51.4	57.1	66.6±0.1
CausalAffect (+EmotioNet)	**67.1**	43.6	**66.0**	80.1	79.1	84.8	88.9	71.1	55.6	66.6	47.5	58.8	**67.4**±0.3
CausalAffect (+RAF-DB)	63.2	45.7	61.4	81.2	79.1	85.1	88.5	71.0	55.3	67.0	51.3	56.5	67.1±0.1
CausalAffect (+AffectNet)	64.4	44.6	61.2	81.3	80.2	85.5	88.8	**71.5**	52.6	67.4	45.5	56.3	66.6±0.2
CausalAffect (+All)	65.3	42.1	58.0	81.0	78.8	85.0	87.9	71.0	56.3	66.0	48.7	60.3	66.7±0.2

Table 4 gives the full per-AU comparison on BP4D. In the strict single-dataset setting, CausalAffect (SG) achieves 66.9 ± 0.2 average F1, outperforming the strongest prior average result, AUFormer (66.2), by $+0.7$ points without using any auxiliary data. This is important because it shows that the gain is not solely due to data scaling; even when trained only on BP4D, the AU bottleneck and signed AU→AU relation modeling provide useful structural regularization.

The per-AU results show that CausalAffect (SG) is particularly strong on several structurally coupled or expression-relevant AUs. For example, it obtains the best result on AU6 (82.1) and AU17 (67.9), and is also highly competitive on AU10, AU14, and AU15. These AUs often appear in coordinated facial configurations, where modeling dependencies among AUs can help stabilize predictions. AU2 and AU4 remain challenging under single-dataset training. This suggests that the benefit of relation learning is most visible when the target AU has sufficiently reliable visual evidence or useful dependencies with other AUs.

Weakly supervised auxiliary data further changes the performance profile. The best BP4D average result is obtained by CausalAffect (+EmotioNet), reaching 67.4 ± 0.3. Compared with SG, +EmotioNet improves several AUs substantially, including AU1 ($63.4 \rightarrow 67.1$) and AU4 ($56.8 \rightarrow 66.0$), indicating that in-the-wild AU supervision can provide complementary facial-action variation for BP4D. However, the +All setting does not produce the best BP4D result and slightly underperforms SG on the average score. This non-monotonic behavior indicates that adding more datasets can also introduce conflicting AU statistics, domain shift, or annotation-protocol mismatch. Therefore, the BP4D results support a key observation of our framework: auxiliary supervision is beneficial when it enriches compatible AU relations, but more data alone does not guarantee improvement.

C.2 DISFA AU detection

Table 5: Full per-AU F1-score (%) comparison on DISFA.

Methods	DISFA								
	AU1	AU2	AU4	AU6	AU9	AU12	AU25	AU26	Avg.
JÂA-Net [72]	62.4	60.7	67.1	41.1	45.1	73.5	90.9	67.4	63.5
FAUT [36]	46.1	48.6	72.8	56.7	50.0	72.1	90.8	55.4	61.5
EmoLA [50]	50.5	56.9	**83.5**	55.2	43.1	80.1	91.6	60.0	65.1
FG-Net [92]	-	-	-	-	-	-	-	-	65.4
KSRL [5]	60.4	59.2	67.5	52.7	51.5	76.1	91.3	57.7	64.5
ReCoT [49]	51.3	36.2	66.8	50.1	52.4	78.8	**95.3**	69.7	62.6
MEGraph [59]	52.5	45.7	76.1	51.8	46.5	76.1	92.9	57.6	62.4
CLEF [97]	64.3	61.8	68.4	49.0	55.2	72.9	89.9	57.0	64.8
MCM [98]	49.6	44.1	67.2	65.5	49.0	**81.5**	85.9	71.8	64.3
MDHRM [86]	65.4	60.2	75.2	50.2	52.4	74.3	93.7	58.2	66.2
AUFormer [93]	-	-	-	-	-	-	-	-	66.4
CausalAffect (SG)	54.8	61.7	75.4	50.8	62.6	67.5	89.5	73.3	67.0±0.1
CausalAffect (+GFT)	56.7	**66.1**	77.0	**68.3**	63.7	72.2	91.9	68.6	70.6±0.3
CausalAffect (+EmotioNet)	68.1	63.2	77.6	64.1	**74.0**	69.3	83.7	68.7	71.1±0.2
CausalAffect (+RAF-DB)	65.6	65.2	73.4	56.5	61.6	70.5	89.2	74.4	69.5±0.2
CausalAffect (+AffectNet)	59.0	65.7	72.3	62.4	58.3	69.5	83.7	**74.8**	68.2±0.2
CausalAffect (+All)	**72.5**	63.7	80.0	51.9	67.1	75.0	92.6	69.0	**71.5**±0.2

Table 5 gives the full per-AU results on DISFA. DISFA is a challenging spontaneous AU benchmark with subtle activations and strong subject-level variation. CausalAffect (SG) already reaches 67.0 ± 0.1, outperforming the strongest prior average result, AUFormer (66.4), by $+0.6$ points under the same single-dataset setting. This again confirms that the proposed signed AU relation modeling provides benefits even without external supervision.

The gain becomes more pronounced under weakly supervised joint training. CausalAffect (+All) achieves 71.5 ± 0.2, improving over CausalAffect (SG) by $+4.5$ points and over the strongest prior average result by $+5.1$ points. This large improvement suggests that DISFA benefits strongly from auxiliary AU and expression evidence. Because DISFA contains sparse and subtle AU activations, additional datasets help the model learn more stable AU-level relation patterns, while the sample-adaptive graph adjusts these patterns to the current image.

Different auxiliary datasets contribute different AU-specific benefits. For example, CausalAffect (+GFT) achieves the best results on AU2 and AU6, suggesting that GFT provides useful evidence for upper-face and cheek-related activations. CausalAffect (+EmotioNet) strongly improves AU1 and AU9, indicating that in-the-wild AU variation helps certain subtle or mid-face actions. CausalAffect (+All) provides the best overall average by combining complementary sources, with strong performance on AU1, AU4, AU9, AU12, and AU25. Some prior methods remain highly competitive on AU4 and AU25. CausalAffect achieves stronger average performance through structured causal AU dependency learning.

C.3 EmotioNet AU detection

Table 6: Full per-AU F1-score (%) comparison on EmotioNet. The average values for CausalAffect variants denote mean±std over five runs when available.

Methods	EmotioNet											
	AU1	**AU2**	**AU4**	**AU5**	**AU6**	**AU9**	**AU12**	**AU17**	**AU20**	**AU25**	**AU26**	**Avg.**
ME-GraphAU [59]	64.1	56.7	68.9	54.9	74.9	61.9	88.9	56.8	44.9	81.9	59.9	64.9
JAA-Net [71]	50.8	43.8	55.8	41.8	61.8	48.8	75.8	43.8	31.8	68.8	46.8	51.8
AUNets [68]	63.7	56.5	68.6	54.6	74.6	61.6	88.6	56.6	44.6	81.6	59.6	64.6
FBNet [42]	53.0	46.0	58.0	44.0	64.0	51.0	78.0	46.0	34.0	71.0	49.0	54.0
CTC [101]	63.5	56.3	68.4	54.4	74.4	61.4	88.4	56.4	44.4	81.4	59.4	64.4
FUXI [96]	63.7	55.9	69.9	53.1	76.6	62.2	92.9	52.8	41.8	90.9	59.6	65.4
SITU [55]	63.2	56.2	68.2	54.2	74.2	61.2	88.2	56.2	44.2	81.2	59.2	64.2
CausalAffect (SG)	64.3	56.8	**71.0**	54.3	**77.8**	**63.5**	**93.4**	53.7	42.8	**91.6**	61.2	**66.4**±0.1
CausalAffect (+DISFA)	63.9	56.2	70.4	53.8	77.0	62.8	93.1	53.0	42.1	91.1	**62.6**	66.0±0.3
CausalAffect (+BP4D)	**64.4**	**57.4**	69.4	**55.4**	75.4	62.4	89.4	**57.4**	**45.4**	82.4	60.4	65.4±0.1
CausalAffect (+All)	63.4	55.7	69.5	52.4	76.2	61.9	92.8	52.3	41.2	90.8	59.4	65.0±0.1

Table 6 gives the full per-AU results on EmotioNet. Among the configurations in Table 6, strict SG gives CausalAffect its highest EmotioNet average F1 of 66.4 ± 0.1. This improves over the strongest prior average result, FUXI (65.4), by $+1.0$ point. This gain under single-dataset training supports the contribution of the AU bottleneck and signed relation learning.

The per-AU breakdown shows that CausalAffect (SG) performs especially well on several high-frequency and expression-relevant AUs, including AU4, AU6, AU9, AU12, and AU25. For instance, it achieves 93.4 on AU12 and 91.6 on AU25, both of which are important lower-face actions for expression composition. It also improves AU4 and AU6, suggesting that the learned AU$\rightarrow$AU structure helps coordinate upper-face and cheek-related activations in in-the-wild images.

In contrast to DISFA, adding laboratory AU datasets such as BP4D or DISFA does not improve the EmotioNet average score. CausalAffect (+DISFA), (+BP4D), and (+All) obtain 66.0 ± 0.3, 65.4 ± 0.1, and 65.0 ± 0.1, respectively, all below the SG result. The per-AU results reveal why this happens: auxiliary data can improve some individual AUs, such as AU2, AU5, AU17, and AU20 under +BP4D, but it can simultaneously reduce performance on strong in-the-wild cues such as AU12 and AU25. This indicates that laboratory datasets may introduce domain-specific AU statistics that are not fully aligned with EmotioNet's in-the-wild distribution. Thus, EmotioNet provides strong evidence against the "more data is always better" explanation and highlights the importance of dataset compatibility in weakly supervised relation learning.

C.4 GFT AU detection

Table 7: Per-AU F1-score (%) comparison on GFT, evaluated on the same 10 AUs as prior work. Rows with "-" indicate that the corresponding per-AU scores were not reported in the original source.

Methods	GFT										
	AU1	**AU2**	**AU4**	**AU6**	**AU10**	**AU12**	**AU14**	**AU15**	**AU23**	**AU24**	**Avg.**
TCAE [51]	43.9	49.5	6.3	71.0	76.2	79.5	10.7	28.5	34.5	41.7	44.2
CDAU [18]	43.7	44.9	19.8	74.6	76.5	79.8	**50.0**	33.9	16.8	12.9	45.3
ARL [73]	51.9	45.9	13.7	79.2	75.5	82.8	0.1	44.9	59.2	47.5	50.1
MoCo [25]	45.3	48.2	20.3	80.7	78.8	78.1	22.6	46.0	53.9	50.3	52.4
TR [57]	-	-	-	-	-	-	-	-	-	-	54.7
JÂA-Net [72]	46.5	49.3	19.2	79.0	75.0	84.8	44.1	33.5	54.9	50.7	53.7
EmoCo [77]	65.9	55.9	40.7	83.1	75.1	81.4	21.3	48.5	58.0	**56.5**	58.6
EmoLA [50]	74.8	52.8	47.9	84.2	78.9	83.4	40.9	**48.9**	55.8	53.4	62.1
CausalAffect (SG)	73.8	51.6	46.2	83.4	77.2	83.1	38.9	47.7	57.6	51.5	61.1±0.2
CausalAffect (+DISFA)	69.8	**59.1**	**52.8**	**89.3**	73.0	**85.3**	32.3	47.6	**63.1**	52.2	**62.5**±0.3
CausalAffect (+BP4D)	73.2	51.0	45.0	82.7	76.5	82.5	38.2	46.9	56.4	51.6	60.4±0.4
CausalAffect (+All)	**75.1**	52.9	47.5	84.8	**79.0**	84.6	40.7	48.8	58.5	51.9	62.4±0.3

GFT highlights the importance of relational evidence for causal AU learning. CausalAffect gains additional predictive information from the co-activation and contrasting configurations available for each AU pair. Learning a causal edge therefore depends on how often its source and target

AUs appear in informative configurations. Within GFT, this evidence is uneven: AU6, AU10, and AU12 occur in more than 30% of valid frames, whereas AU4 occurs in 3.5%. The per-AU results in Table 7 reflect this variation, with the SG model performing strongly on AU6, AU10, and AU12 and obtaining lower F1 on AU4, AU14, and AU15. This imbalance helps explain the smaller overall gain on GFT: relations involving rare AUs receive fewer informative configurations, limiting the additional evidence available for causal inference.

The AU-composition experiment in Table 12 directly examines how the choice of auxiliary AUs affects this relational supervision. With BP4D fixed as the auxiliary dataset, adding the two most frequent AUs to the six-AU core raises GFT F1 from 60.3 to 61.3; adding the two least frequent AUs yields 58.6. At the same vocabulary size, the eight most frequent AUs achieve 62.7, compared with 58.3 for the eight least frequent AUs. Using all 12 AUs gives 60.4. These controlled comparisons show that AU composition determines how effectively auxiliary supervision strengthens causal relation learning. Frequently observed AUs supply richer configurations for learning the graph, allowing a compact AU vocabulary to outperform a larger one.

Compatible auxiliary supervision translates this advantage into improved GFT performance. CausalAffect (+DISFA) reaches 62.5 ± 0.3, improving over the SG result of 61.1 ± 0.2 by $+1.4$ points and over EmoLA (62.1) by $+0.4$ points. The gains extend across upper- and lower-face actions: +DISFA achieves the best results on AU2, AU4, AU6, AU12, and AU23. The +All configuration reaches 62.4 ± 0.3 and achieves the best results on AU1 and AU10. These complementary per-AU gains show how auxiliary datasets strengthen different parts of the causal graph, explaining why the composition of the added supervision matters for overall performance.

This performance is achieved with a compact ConvNeXt-based causal graph model. EmoLA [50] uses a LLaVA-1.5-7B-based multimodal large language model combining CLIP and landmark features. CausalAffect exceeds its GFT average with DISFA supervision and reaches 62.7 in the frequency-controlled AU-composition experiment. The GFT results demonstrate the predictive value of learning causal AU relations from sufficiently observed configurations and show how targeted auxiliary supervision enhances that value.

C.5 Facial expression recognition on AffectNet and RAF-DB

Table 8: Facial expression recognition comparison on AffectNet and RAF-DB. Accuracy (%) is the metric for both datasets.

Methods	AffectNet	RAF-DB
MVT [46]	64.6	80.4
FBNet [43]	65.0	78.0
DACL [20]	65.2	80.4
ARM [74]	65.2	82.8
Ad-Corre [19]	63.4	79.0
DDAMFN++ [95]	67.4	84.6
FRA [22]	66.1	83.4
CAGE [82]	66.4	83.2
ExpLLM [45]	65.9	-
LOFI [94]	65.7	84.7
CausalAffect (+DISFA)	65.6±0.2	83.7±0.1
CausalAffect (+BP4D)	**67.7**±0.1	**85.3**±0.2
CausalAffect (+All)	66.5±0.1	84.9±0.2

Table 8 compares facial expression recognition results on AffectNet and RAF-DB. Unlike the AU detection setting, FER uses AU-labeled datasets as weak auxiliary supervision to learn AU→Expression relations. This directly tests whether AU-level structure can improve expression recognition even when AU and expression annotations are disjoint.

CausalAffect (+BP4D) achieves the best results on both FER benchmarks, reaching 67.7 ± 0.1 on AffectNet and 85.3 ± 0.2 on RAF-DB. On AffectNet, this improves over the previous best DDAMFN++ result (67.4) by $+0.3$ points. On RAF-DB, it improves over the strongest prior result LOFI (84.7) by $+0.6$ points. These gains demonstrate the value of disjoint AU supervision for

causal AU→Expression learning on established FER benchmarks, without requiring paired AU and expression annotations.

The comparison among CausalAffect variants shows that BP4D is more beneficial for FER than DISFA. CausalAffect (+DISFA) obtains 65.6 ± 0.2 on AffectNet and 83.7 ± 0.1 on RAF-DB, while +BP4D improves them to 67.7 ± 0.1 and 85.3 ± 0.2. This difference is consistent with the broader AU coverage of BP4D, which provides a richer AU basis for composing expression categories. A larger AU vocabulary helps the AU→Expression graph capture more fine-grained expression evidence, especially for ambiguous or compound facial configurations.

Interestingly, +All does not achieve the best FER result, despite using the largest amount of data. It remains strong, with 66.5 ± 0.1 on AffectNet and 84.9 ± 0.2 on RAF-DB, but underperforms +BP4D. This indicates that heterogeneous data can introduce conflicting AU→Expression statistics when multiple AU vocabularies, capture conditions, and annotation protocols are combined. The FER results reinforce the AU detection results. Weakly supervised training benefits most from compatible and semantically useful AU evidence rather than from greater data volume alone.

C.6 Overall analysis

The full benchmark results lead to several observations.

CausalAffect is effective without auxiliary data. In the strict SG setting, CausalAffect already achieves 66.9 ± 0.2 on BP4D, 67.0 ± 0.1 on DISFA, and 66.4 ± 0.1 on EmotioNet. These results outperform or match strong prior methods under the same single-dataset protocol. This is an important control because it rules out the simplest explanation that the method only improves by adding more data. Instead, the SG results show that AU bottleneck learning, directed signed graph reasoning, and polarity-aware message passing are already useful within a single dataset.

Auxiliary supervision is beneficial only when it is compatible. The cross-dataset settings show clear non-monotonic behavior. On DISFA, adding all datasets improves the average F1 from 67.0 ± 0.1 to 71.5 ± 0.2, demonstrating that auxiliary evidence can strongly benefit sparse and subtle AU detection. On GFT, adding DISFA improves the average F1 from 61.1 ± 0.2 to 62.5 ± 0.3, showing that a compatible auxiliary AU dataset can strengthen relation learning. However, on EmotioNet, +All reduces the average F1 to 65.0 ± 0.1, below the SG result. Similarly, for FER, +BP4D outperforms +All on both AffectNet and RAF-DB. These patterns show that dataset composition, AU vocabulary overlap, annotation protocol, and domain shift are more important than raw data volume.

AU detection and FER require different auxiliary AU properties. For AU detection, compact and compatible AU evidence is often more useful than broad but heterogeneous supervision. This is reflected by the strong DISFA and GFT results when appropriate auxiliary AU datasets are used, and by the degradation on EmotioNet when less compatible laboratory data are added. For FER, broader AU coverage is more beneficial because expression recognition requires a richer compositional basis. This explains why BP4D, with its wider AU vocabulary, provides stronger FER gains than DISFA on both AffectNet and RAF-DB.

The gains are not explained by naive multi-task learning. The controlled comparison in Appendix D shows that simply pooling all datasets in a conventional MTL framework can cause negative transfer, especially for AU detection. In contrast, CausalAffect organizes heterogeneous supervision through AU bottlenecks and signed AU-level relation graphs. This turns disjoint AU and expression labels into structured relational supervision rather than treating them as unrelated task objectives. The non-monotonic +Dataset results also support this interpretation: CausalAffect does not simply benefit from more samples, but from whether the added data improves the learned AU relation structure.

Performance is stable across runs. For CausalAffect variants, the standard deviations across runs are generally small, typically between 0.1 and 0.4. This suggests that the model does not rely on fragile initialization or dataset-specific tuning, despite combining several components such as AU bottleneck disentanglement, global signed graphs, sample-adaptive refinement, polarity-aware aggregation, and counterfactual intervention. The stability across datasets supports the reproducibility of the proposed training procedure.

Overall, the full results support the central claim of CausalAffect: directed and signed AU-level relation learning provides both predictive benefits and interpretable structure under disjoint weak

supervision. The results also clarify when auxiliary data helps: improvements arise from compatible AU relational evidence, not from data volume alone.

D CausalAffect mechanism vs. data volume: why CausalAffect works?

A primary concern in large-scale weakly supervised learning is determining whether performance gains stem from the proposed methodology or simply from the increased volume of training data. To disentangle the contribution of the *causal relation learner* from *data scaling*, we conducted a controlled experiment comparing CausalAffect against standard MTL baselines under identical conditions (same backbone, same training data). The results are summarized in Table 9.

Negative transfer in naive multi-task learning. Rows 2 and 3 in Table 9 reveal a critical insight: simply pooling all datasets into a standard MTL framework **does not guarantee performance improvements**. While FER tasks see marginal gains due to increased data diversity, AU Detection tasks suffer from severe *negative transfer*. Comparing the Single-Task baseline (Row 1) with Naive MTL (Rows 2 & 3), performance on DISFA drops drastically from 61.4% to 53.0%, and EmotioNet plummets from 61.7% to 47.5%. FER relies on global semantic features, whereas AU detection requires fine-grained local cues. These competing demands can introduce gradient conflicts in the shared backbone, contributing to the observed negative transfer. These results strongly refute the hypothesis that "more data automatically yields better results".

CausalAffect Resolves Negative Transfer. In contrast, CausalAffect (Row 5) achieves consistent and significant improvements across all benchmarks when trained on the same combined data. Crucially, even in the Single-Dataset setting (Row 4), CausalAffect already outperforms both the Single-Task baseline (Row 1) and the data-rich MTL baselines (Rows 2 & 3). For instance, on DISFA, CausalAffect (SG) achieves 67.0% (vs. MTL's 53.6%), and scaling to all datasets further boosts this to 71.5%. Paired t-tests over five matched runs confirm that the gains over the MTL baselines in Table 9 are statistically significant across all metrics ($p < 0.05$). The proposed causal graph mechanism disentangles task-specific dependencies and uses the AU→Expression hierarchy to turn heterogeneous supervision into performance gains, overcoming the negative transfer observed in the standard MTL baselines.

Table 9: Comparison between naive MTL baselines and the proposed CausalAffect. All models use the same training data sources and backbone. Statistical significance is assessed using paired t-tests over 5 matched runs (same data splits and random seeds), with differences considered significant at $p < 0.05$.

Idx	Model Variant	Training Data	AffectNet (FER)	RAFDB (FER)	DISFA (AU)	BP4D (AU)	GFT (AU)	EmotioNet (AU)
1	SG (single-task)	Single-dataset	58.3	69.5	61.4	64.0	53.6	61.7
2	MTL (Shared Head)	All datasets	58.9	70.0	53.0	57.2	57.9	47.5
3	MTL (Task-Specific Heads)	All datasets	60.3	71.9	53.6	57.6	58.1	49.3
4	CausalAffect (SG)	Single-dataset	/	/	67.0 ± 0.1	**66.9 ± 0.2**	61.1 ± 0.2	**66.4 ± 0.1**
5	**CausalAffect (All)**	**All datasets**	**66.5 ± 0.1**	**84.9 ± 0.2**	**71.5 ± 0.2**	66.7 ± 0.2	**62.4 ± 0.3**	65.0 ± 0.1

Controlled comparisons with AU-aware and graph-based baselines. We compare CausalAffect with an adapted AU-ViT baseline [61] and a ME-GraphAU-style relation module under the same ConvNeXt-Base/WebFace initialisation, dataset combinations, splits, augmentations, and five paired random seeds. For AU-ViT, we replace its original IR50 embedding with the spatial feature map from ConvNeXt-Base and retain its Transformer blocks, auxiliary AU branch, and branch-specific objectives. For the ME-GraphAU-style control [59], we retain the backbone, disentangled AU nodes, task heads, label masks, and training protocol, and replace the relation learner with multidimensional edge generation and graph propagation. This comparison evaluates AU targets, as the ME-GraphAU module has no AU→Expression pathway. Tables 10 and 11 give the complete results.

Advantages under heterogeneous supervision. CausalAffect achieves higher mean performance in **18 of the 19 heterogeneous-training comparisons**. Against AU-ViT, it leads in six of seven joint-training settings, with +All gains of 2.8 points on GFT, 2.1 on EmotioNet, and 1.6 on RAF-DB. Against the ME-GraphAU-style module, it achieves higher means in all 16 settings; the +All gains reach **7.3 points on DISFA**, 3.7 on GFT, 2.9 on BP4D, and 1.7 on EmotioNet. These comparisons

highlight the effectiveness of CausalAffect's relational design in integrating heterogeneous supervision. AU-ViT has higher means in the two AU-only settings and on RAF-DB with EmotioNet supervision. The ME-GraphAU comparisons on GFT+BP4D and EmotioNet+BP4D have $p > 0.05$; the 18/19 count refers to higher means.

Table 10: Matched comparison with the ConvNeXt-Base adaptation of AU-ViT. AU detection is evaluated by average F1 (%), and expression recognition by accuracy (%). Values are mean ± SD over five paired seeds; p-values are from two-sided paired t-tests. Higher means are in bold.

Dataset	**Training**	**AU-ViT (ConvNeXt-B)**	**CausalAffect**	p
GFT (AU)	GFT only	$\mathbf{61.4 \pm 0.3}$	61.1 ± 0.2	0.064
	+RAF-DB	59.7 ± 0.6	$\mathbf{61.7 \pm 0.3}$	0.004
	+All	59.6 ± 0.7	$\mathbf{62.4 \pm 0.3}$	0.002
EmotioNet (AU)	EmotioNet only	$\mathbf{66.7 \pm 0.4}$	66.4 ± 0.1	0.119
	+RAF-DB	64.4 ± 0.6	$\mathbf{65.8 \pm 0.2}$	0.008
	+All	62.9 ± 0.6	$\mathbf{65.0 \pm 0.1}$	0.001
RAF-DB (FER)	+GFT	83.1 ± 0.4	$\mathbf{84.0 \pm 0.2}$	0.011
	+EmotioNet	$\mathbf{84.9 \pm 0.3}$	84.5 ± 0.2	0.029
	+All	83.3 ± 0.6	$\mathbf{84.9 \pm 0.2}$	0.003

Table 11: Controlled comparison with the ME-GraphAU-style relation module. Average AU F1 (%) values are mean ± SD over five paired seeds; p-values are from two-sided paired t-tests. Higher means are in bold.

Dataset	**Training**	**ME-GraphAU module**	**CausalAffect**	p
BP4D	BP4D only	66.3 ± 0.4	$\mathbf{66.9 \pm 0.2}$	0.024
	+GFT	65.5 ± 0.5	$\mathbf{66.6 \pm 0.1}$	0.011
	+EmotioNet	66.7 ± 0.4	$\mathbf{67.4 \pm 0.3}$	0.026
	+All	63.8 ± 0.7	$\mathbf{66.7 \pm 0.2}$	< 0.001
DISFA	DISFA only	66.5 ± 0.4	$\mathbf{67.0 \pm 0.1}$	0.041
	+GFT	66.0 ± 0.5	$\mathbf{70.6 \pm 0.3}$	< 0.001
	+EmotioNet	69.0 ± 0.5	$\mathbf{71.1 \pm 0.2}$	0.002
	+All	64.2 ± 0.7	$\mathbf{71.5 \pm 0.2}$	< 0.001
GFT	GFT only	60.0 ± 0.4	$\mathbf{61.1 \pm 0.2}$	0.006
	+DISFA	60.6 ± 0.5	$\mathbf{62.5 \pm 0.3}$	0.002
	+BP4D	60.3 ± 0.5	$\mathbf{60.4 \pm 0.4}$	0.621
	+All	58.7 ± 0.7	$\mathbf{62.4 \pm 0.3}$	< 0.001
EmotioNet	EmotioNet only	65.9 ± 0.4	$\mathbf{66.4 \pm 0.1}$	0.030
	+DISFA	65.3 ± 0.5	$\mathbf{66.0 \pm 0.3}$	0.028
	+BP4D	65.2 ± 0.4	$\mathbf{65.4 \pm 0.1}$	0.339
	+All	63.3 ± 0.5	$\mathbf{65.0 \pm 0.1}$	0.002

E More data is not always better: sensitivity to AU composition

To investigate how the size and composition of AU supervision affect the learned causal structure, we conducted a systematic analysis across different AU subsets from BP4D. Results are summarized in Table 12, covering both AU detection (EmotioNet, GFT) and expression recognition (RAF-DB, AffectNet).

For AU detection, the results highlight that **frequently occurring AUs** play a dominant role in shaping robust causal dependencies. Configurations relying on the most frequent AUs (Row 5) achieve the highest AU detection performance on both EmotioNet (66.8%) and GFT (62.7%), surpassing both the single-dataset baseline and larger AU sets that include low-frequency units. Incorporating low-frequency AUs (Rows 4 and 6) reduces AU detection performance. Their sparse activations provide fewer informative configurations for learning causal relations, helping explain the advantage

Table 12: Effect of AU Set Size on AU Detection and Expression Recognition. Performance is evaluated on EmotioNet/GFT for AU detection and RAF-DB/AffectNet for expression recognition. Best results for each task are highlighted.

#	Setting / Method	EmotioNet (AU)	GFT (AU)	RAF-DB (Expr)	AffectNet (Expr)
1	CausalAffect (SG baseline)	66.4	61.1	–	–
2	CausalAffect (+BP4D, 6 AUs)	65.6	60.3	80.2	63.7
3	CausalAffect (+BP4D, 8 AUs, Row 2 + 2 Most Frequent AUs)	65.8	61.3	83.5	65.1
4	CausalAffect (+BP4D, 8 AUs, Row 2 + 2 Least Frequent AUs)	64.3	58.6	82.0	64.2
5	CausalAffect (+BP4D, 8 AUs, 8 Most Frequent AUs)	**66.8**	**62.7**	84.2	66.3
6	CausalAffect (+BP4D, 8 AUs, 8 Least Frequent AUs)	63.1	58.3	81.7	64.5
7	CausalAffect (+BP4D, 12 AUs)	65.4	60.4	**85.3**	**67.7**

of the frequent-AU subsets. Interestingly, the 12-AU configuration (Row 7) does not outperform the best 8-AU frequent setting on AU detection, further confirming that **more AUs do not necessarily yield better causal modeling when frequency imbalance is severe**.

For expression recognition, a different trend emerges. Larger AU sets consistently improve performance, with the 12-AU configuration achieving the highest accuracy (85.3% on RAF-DB and 67.7% on AffectNet). This indicates that expression recognition benefits from a **richer and more compositional AU basis**, as the model learns to combine fine-grained AUs into higher-level prototypes for emotion categories. Even though low-frequency AUs hinder AU detection, they still provide complementary information that enriches expression-level inference. The gap between frequent-only subsets and the full 12-AU set demonstrates that expression recognition is more tolerant to sparsity and leverages the additional granularity to form more accurate and psychologically valid AU→Expression mappings.

Overall, these results reveal a clear sensitivity of CausalAffect to AU set size and composition: **(i)** AU detection is optimized when relying on a compact set of frequent AUs, which ensures dense and stable causal relations. **(ii)** Expression recognition, however, requires broader AU coverage, where even low-frequency units contribute to refining causal prototypes of emotions. This divergence underscores the importance of tailoring AU supervision to the specific downstream task—favoring *frequency and stability* for AU detection, while emphasizing *richness and compositionality* for expression recognition.

F Zero-shot cross-dataset evaluation

CausalAffect learns AU→AU and AU→Expression causal relations to support expression recognition across datasets. We evaluate this transferability by training on AffectNet-7 and BP4D and testing directly on the unseen RAF-DB test set. This zero-shot experiment tests whether the AU bottleneck and learned causal graphs provide an effective inference pathway when the input domain changes. Keeping the source-trained model fixed allows us to assess how these representations and relations support expression prediction on new images without target-domain training or adaptation.

Experimental setting. AffectNet-7 provides expression supervision for the AU→Expression branch, while BP4D provides disjoint AU supervision for the AU bottleneck and AU→AU branch. This follows the weakly supervised training setting used throughout the paper and combines expression-level information with fine-grained facial action information from separate datasets. RAF-DB is used only as the target test domain and is excluded from training, validation, hyperparameter tuning, and model selection. During evaluation, all model and graph parameters remain fixed. The sample-adaptive graph is computed from each test image through a forward pass, so the learned causal relations can respond to the facial evidence in that image without updating the source-trained model. The experiment therefore evaluates the complete AU-based inference pathway under a cross-dataset shift, extending the in-domain recognition experiments to an unseen target dataset.

We compare CausalAffect with two groups of methods that address cross-dataset recognition. The zero-shot FER baselines, SE-ResNet, DDAMFN, and POSTER++, are trained on the source domain and directly evaluated on RAF-DB without target-domain supervision. The test-time adaptation (TTA) baselines, TENT, EATA, SAR, T3A, NOTE, and RoTTA, additionally use unlabelled target-domain test samples to update model statistics or parameters during inference. Including both groups evaluates the strength of the source-learned causal representations against direct transfer methods

and methods that exploit the target distribution at test time. Table 13 presents their results alongside CausalAffect, which uses the fixed model for every target prediction.

Cross-dataset transfer. CausalAffect achieves **78.1%** accuracy on RAF-DB, outperforming all three zero-shot FER baselines and all six TTA methods in Table 13. Its advantage over the direct-transfer baselines supports the effectiveness of AU-structured causal learning for cross-dataset recognition. Exceeding the TTA baselines further shows that the source-learned inference pathway remains competitive with methods that adapt to the target distribution, while requiring no such adaptation itself.

Table 13: Cross-dataset zero-shot evaluation on RAF-DB. Models are trained on AffectNet-7 and evaluated on the RAF-DB test set. The DDAMFN result is taken from Table 10 of Wang et al. [84]. Results for the six TTA baselines are taken from Turnbull et al. [81]. CausalAffect additionally uses BP4D as disjoint AU supervision, but does not use RAF-DB training data or target-domain adaptation.

Method	Target Adaptation	RAF-DB Acc. (%)
SE-ResNet [33]	No	77.4
DDAMFN [95]	No	75.6
POSTER++ [60]	No	74.0
TENT	Yes	74.8
EATA	Yes	72.3
SAR	Yes	74.8
T3A	Yes	74.3
NOTE	Yes	72.3
RoTTA	Yes	72.1
CausalAffect	**No**	**78.1**

The transfer advantage can be explained by the role of the AU bottleneck in connecting the two source datasets. BP4D supervision grounds the intermediate representations in facial actions, and AffectNet supervision links these actions to expression categories through signed AU→Expression causal relations. Expression prediction consequently draws on the configuration of facial actions and their excitatory and inhibitory contributions. This structure reduces direct dependence on source-specific appearance statistics, such as background, acquisition conditions, and identity distribution, and encourages the model to use facial action patterns that remain informative across datasets. The result on RAF-DB supports the value of this structured use of auxiliary AU supervision for recognising expressions beyond the source domain.

The global and sample-adaptive causal graphs provide complementary support for this transfer. The global graph captures stable population-level AU dependencies and AU→Expression regularities learned from the source data. The sample-adaptive graph refines these causal relations according to the AU intensities and combinations present in each target image. It thus preserves transferable causal priors while accounting for variation in individual facial configurations through input-conditioned inference. These complementary roles help explain how CausalAffect maintains effective expression recognition on RAF-DB with fixed parameters: the learned causal structure supplies a common relational basis, and the sample-adaptive graph connects that structure to the visual evidence in each unseen image.

G Fairness and demographic robustness evaluation

CausalAffect routes expression prediction through AU-level representations and signed AU→Expression relations. We evaluate whether this relational approach maintains comparable recognition performance across demographic groups on the RAF-DB test set. The model uses the same cross-dataset setting as the zero-shot experiment: AffectNet-7 provides expression supervision and BP4D provides AU supervision. RAF-DB is used only for evaluation, without training, validation, model selection, or test-time adaptation. Its demographic annotations are used after inference to calculate accuracy for the Asian, Black, and White groups; no demographic labels are used during training. The scope and deployment implications of this diagnostic are discussed in Appendix Q.

Evaluation metrics. Alongside group-wise accuracy, we measure Statistical Parity (SP) and Equality of Opportunity (EOP) as percentage-point gaps, with lower values indicating smaller disparities. SP measures differences in prediction distributions across groups, while EOP measures differences in true-positive rates for the same expression class. Let A denote the demographic group, Y the ground-truth expression label, $\hat{Y}$ the predicted label, and $\mathcal{C}$ the set of expression classes. We compute the class-macro gaps as

$$\mathrm{SP} = \frac{100}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left[\max_a P(\hat{Y} = c \mid A = a) - \min_a P(\hat{Y} = c \mid A = a) \right], \tag{29}$$

and

$$\mathrm{EOP} = \frac{100}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \left[\max_a P(\hat{Y} = c \mid Y = c, A = a) - \min_a P(\hat{Y} = c \mid Y = c, A = a) \right]. \tag{30}$$

Demographic robustness. Table 14 shows that CausalAffect maintains comparable accuracy across the three evaluated RAF-DB groups: 77.8% for Asian, 77.4% for Black, and 78.6% for White. The maximum gap is 1.2 percentage points, and all three accuracies are close to the global accuracy of 78.1%; the SP and EOP gaps are 8.5 and 9.7 percentage points, respectively. This pattern is consistent with the model's design motivation: AU-specific bottleneck representations and signed causal relations organise expression prediction around facial action patterns. This relational interface may reduce reliance on appearance shortcuts associated with identity, skin tone, or image acquisition conditions, supporting demographic robustness within the evaluated setting.

Table 14: Diagnostic demographic robustness evaluation on the RAF-DB test set. Lower SP and EOP indicate smaller group disparity.

Asian Acc.	Black Acc.	White Acc.	Global Acc.	SP	EOP
77.8	77.4	78.6	78.1	8.5	9.7

H Ablation study

To better understand the contributions of individual modules and their interactions, we conduct an extensive ablation study on CausalAffect (+All) using 15 model variants (Rows 1–15 in Table 15). The goal of this experiment is to disentangle the roles of the Global Causal Graph (GC), Sample-Adaptive Causal Graph (SAC), Counterfactual Intervention (CF), AU Disentanglement (Dis), and the DAG constraint. Performance is evaluated across six benchmarks (AffectNet, RAF-DB, DISFA, BP4D, GFT, EmotioNet).

Table 15: Ablation Study on CausalAffect (+All Setting), exploring the effect of GC, SAC, CF, and Dis. Best results are highlighted.

Idx	Model Variant	w/o	AffectNet	RAF-DB	DISFA	BP4D	GFT	EmotioNet
1	Backbone	GC + SAC + CF + Dis	58.9	70.0	53.0	57.2	57.9	47.5
2	Backbone + Dis	GC + SAC + CF	57.1	69.3	54.2	55.5	57.8	59.2
3	Backbone + GC	Dis + CF + SAC	62.3	80.2	62.4	61.0	60.9	61.7
4	Backbone + GC + Dis	CF + SAC	61.9	78.2	61.1	59.8	60.5	61.4
5	Backbone + GC + CF	Dis + SAC	62.5	79.8	61.5	62.1	61.7	62.3
6	Backbone + GC + Dis + CF	SAC	64.4	83.3	65.8	66.6	61.0	63.6
7	Backbone + SAC	Dis + CF + GC	62.0	78.1	60.5	61.1	58.9	60.7
8	Backbone + SAC + Dis	CF + GC	60.7	77.9	60.1	59.5	57.9	59.2
9	Backbone + SAC + CF	Dis + GC	61.3	77.5	60.6	60.7	57.5	60.2
10	Backbone + SAC + Dis + CF	GC	62.7	78.5	61.7	62.4	59.1	61.2
11	Backbone + GC + SAC	Dis + CF	63.1	82.9	64.1	61.5	60.4	62.1
12	Backbone + GC + SAC + Dis	CF	62.6	81.4	66.4	61.3	60.1	62.0
13	Backbone + GC + SAC + CF	Dis	64.3	83.4	62.9	62.8	**62.6**	63.4
14	CausalAffect (w/o DAG)	w/o DAG	65.5	84.6	71.3	64.4	62.1	64.7
15	**CausalAffect (GC + SAC + Dis + CF)**	/	**66.5**	**84.9**	**71.5**	**66.7**	62.4	**65.0**

Global vs. Sample-Adaptive Graphs: When trained individually, the Global Graph generally outperforms the Sample-Adaptive Graph (see Rows 3–4 vs. Rows 7–8). This outcome is expected, as

GC captures stable population-level causal structures that are inherently more robust across datasets. Nevertheless, the Sample-Adaptive Graph plays an essential complementary role: adding SAC to GC+CF improves performance on all six benchmarks (Rows 5→13), confirming that personalized inference adds value by tailoring causal reasoning to individual instances.

Effect of AU Disentanglement The effect of AU Disentanglement can be assessed through matched comparisons in Table 15: Rows 1 vs. 2, 3 vs. 4, 5 vs. 6, 7 vs. 8, 9 vs. 10, 11 vs. 12, and 13 vs. 15. Without disentanglement, the model may rely on spurious correlations such as demographic or identity-specific biases, which can be exploited as shortcuts. When AU Dis is present, it enhances both interpretability and robustness, guiding the model toward psychologically meaningful AU activations. Importantly, AU Dis interacts synergistically with CF: in Rows 6, 10, and 15, CF becomes more effective when paired with disentangled features, suppressing irrelevant cues and amplifying informative dependencies.

Role of Counterfactual Intervention With AU disentanglement, adding CF improves performance on all six benchmarks across the three matched comparisons (Rows 4→6, 8→10, and 12→15). Without disentanglement, the gains vary across graph configurations and datasets. CF encourages the model to contrast factual and counterfactual settings. This allows the system to emphasize influential features while suppressing misleading ones. A notable example is observed in the AU6–Disgust pathway: without CF, spurious correlations dominate, but with CF, the model prunes this dependency, yielding more interpretable causal structures.

Recovery from perturbed graph initialisation. We evaluated CF's ability to recover literature-supported relations after deliberately perturbing the initial graph. We trained a common No-CF warm-start with all objectives except $\mathcal{L}_{\text{cf}}$ and evaluated its global AU→Expression graph against the 39 literature-supported excitatory relations in Appendix K. A relation is aligned when its edge exceeds the learned threshold for the corresponding expression and has positive polarity; below-threshold or negative edges are discordant. The warm-start aligns with 13/39 relations, leaving 26/39 discordant.

For each aligned relation, we swapped its edge parameters with those of an unselected source under the same target. This perturbation displaced all 13 supported assignments and reduced alignment to 0/39, preserving each target row's edge values and polarity distribution. The other 26 reference relations remained unchanged. The literature reference set defined the perturbations and the recovery evaluation.

We continued training from both the unmodified and perturbed graphs, each with and without CF, giving four branches: Unmodified + No CF, Unmodified + CF, Contradicted + No CF, and Contradicted + CF. The branches used fresh optimisers, identical data order, and the same training budget, with all parameters trainable and masks recomputed as the graph evolved.

From the perturbed graph, CF restores **all 13 displaced literature-supported relations**, removes all 13 injected assignments, and corrects 22 of the 26 initially discordant relations (Table 16). It reaches **35/39 alignment compared with 10/39 without CF**. Compared with continued training without CF, RAF-DB performance increases from 80.8 to 84.6, and EmotioNet performance from 61.8 to 65.9. From the unmodified graph, CF retains all 13 initially aligned relations and corrects 21/26 prior conflicts, reaching 34/39 alignment compared with 14/39 without CF. These results show that CF corrects erroneous graph assignments and improves predictive performance.

Table 16: Controlled recovery of literature-supported AU→Expression relations. Initial 13 denotes retained or restored relations; Prior 26 denotes initially discordant relations. Contradicted branches start from 0/39 alignment after the parameter swaps. CF-enabled continuations are highlighted.

Condition	Aligned ↑	Initial 13 ↑	Prior 26 corrected ↑	Injected retained ↓	RAF-DB	EmotioNet
No-CF warm-start	13/39	**13/13**	0/26	—	81.4	62.0
Unmodified + No CF	14/39	12/13	2/26	—	81.6	62.2
Unmodified + CF	34/39	**13/13**	21/26	—	84.4	65.5
Contradicted + No CF	10/39	4/13	6/26	8/13	80.8	61.8
Contradicted + CF	**35/39**	**13/13**	**22/26**	**0/13**	**84.6**	**65.9**

Human-Aligned Causal Structure Rows 6, 10, and 15 highlight the complementary contributions of CF and AU Dis. Adding CF to the disentangled models improves performance on all six benchmarks. Adding Dis to the GC+SAC+CF configuration improves five of the six benchmarks, including an 8.6-point gain on DISFA (Rows 13→15). With both components, CausalAffect captures reliable, semantically grounded AU→Expression relations. These complementary effects support the learning of psychologically plausible causal relations that bridge low-level facial activations and high-level emotion inference.

Effect of DAG Constraint Removing the DAG constraint (Row 14) reduces performance on all six benchmarks compared with the full model (Row 15). The constraint discourages cyclic connections during joint optimisation, reducing redundant feedback and guiding the graph towards clearer directional causal pathways while retaining the flexibility to model physiologically grounded feedback loops. These results support the value of complementing predictive supervision with topological guidance for learning causal AU interactions, benefiting both AU detection and expression recognition.

The ablation study highlights the complementary roles of different modules. The Global Graph captures stable population-level dependencies, while the Sample-Adaptive Graph introduces instance-level flexibility. AU Disentanglement prevents shortcut exploitation and provides psychologically meaningful features. Counterfactual regularization provides consistent performance gains across all six benchmarks when combined with AU disentanglement. Finally, the DAG constraint ensures structural clarity and interpretability, yielding the strongest overall performance. These components allow CausalAffect to achieve not only higher accuracy but also more human-aligned and semantically coherent causal graphs.

I Hyperparameter and sensitivity analysis

This section examines the sensitivity of CausalAffect to its hyperparameters, explains the selected values, and evaluates the stability of its learned global graphs. We use a **fixed set of hyperparameters** across datasets and experiments, including multi-dataset training and ablation studies, with each loss weight controlling the contribution of its corresponding objective to joint optimisation. The results below show that the selected weights achieve the highest RAF-DB accuracy and BP4D F1 among the tested settings.

I.1 AU bottleneck hyperparameters

The AU bottleneck promotes a disentangled and compact AU representation by balancing **information compression**, **semantic alignment**, and **inter-AU decorrelation**. The three hyperparameters operate as follows:

- λ_{ib} regulates HSIC-based information compression.
- λ_{align} aligns AU features with corresponding binary ground-truth or confidence-filtered pseudo AU labels.
- λ_{decorr} weights the HSIC-based dependence penalty between distinct AU features to reduce redundancy.

We choose $\lambda_{\text{ib}} = 1$ and $\lambda_{\text{align}} = 1$ so that compression and alignment exert balanced supervision. Smaller values produce under-regularized and noisy AU representations, whereas larger values cause excessive compression or overfitting. The decorrelation term serves as a lightweight regularizer. For this reason, we set $\lambda_{\text{decorr}} = 0.2$.

Table 17 shows that both compression and alignment achieve their best performance at a weight of 1.0. For decorrelation, a small weight (0.2) effectively suppresses redundancy, but higher values overly suppress natural physiological AU co-activation patterns.

I.2 DAG regularization hyperparameters

Both the global and sample-adaptive AU→AU graphs employ the same soft acyclicity constraint. We vary their weights $\lambda^{\text{g}}_{\text{DAG}}$ and $\lambda^{\text{s}}_{\text{DAG}}$ together. When both weights are too small, the constraint becomes ineffective and yields dense, potentially cyclic structures. When both are too large, the constraint becomes over-restrictive. To balance these effects, we set $\lambda^{\text{g}}_{\text{DAG}} = \lambda^{\text{s}}_{\text{DAG}} = 0.5$.

Table 17: Sensitivity analysis for AU bottleneck hyperparameters (λ_{ib}, λ_{align}, λ_{decorr}). Best results are highlighted.

λ Name	Value	RAF-DB Acc.	BP4D F1	Interpretation
λ_{ib}	0.5	84.4	66.3	Too little compression → AU redundancy appears.
λ_{ib}	**1.0 (Best)**	**85.0**	**67.1**	Balanced compression; most faithful AU bottleneck.
λ_{ib}	2.0	84.6	66.5	Excessive compression → information loss.
λ_{align}	0.5	84.2	66.2	Under-alignment → label-feature mismatch.
λ_{align}	**1.0 (Best)**	**85.0**	**67.1**	Optimal alignment with semantic AU labels.
λ_{align}	2.0	84.7	66.8	Over-penalizing alignment yields diminishing returns.
λ_{decorr}	0.1	84.7	66.6	Weak decorrelation → slight redundancy remains.
λ_{decorr}	**0.2 (Best)**	**85.0**	**67.1**	Balanced AU disentanglement.
λ_{decorr}	0.4	84.2	66.4	Too strong → overly sparse AU basis.

As shown in Table 18, setting both weights to 0.5 achieves the optimal balance between structural faithfulness and expressive flexibility.

Table 18: Sensitivity analysis for DAG regularization weights ($\lambda^g_{\text{DAG}}, \lambda^s_{\text{DAG}}$).

Value	RAF-DB Acc.	BP4D F1	Interpretation
0.1	84.8	66.7	Acyclicity too weak → noisy cyclic edges emerge.
0.5 (Best)	**85.0**	**67.1**	Optimal structural balance; best causal fidelity.
1.0	84.7	66.4	Over-regularized → graph becomes under-expressive.

I.3 Counterfactual supervision hyperparameters

The counterfactual intervention module relies on three sets of parameters: the loss weights for consistency/discrepancy ($\lambda_{\text{consist}}, \lambda_{\text{discrep}}$), the mask sharpness ($\gamma$), and the feature/logit-level weights ($\delta_{\text{feat}}, \delta_{\text{logit}}, \eta_{\text{feat}}, \eta_{\text{logit}}$).

Consistency vs. Discrepancy Weights. Counterfactual supervision encourages similar outputs under interventions on unselected AUs (consistency) and divergent outputs under interventions on graph-selected AUs (discrepancy). These forces must be balanced. As shown in Table 19, setting $\lambda_{\text{consist}} = \lambda_{\text{discrep}} = 0.5$ yields the strongest performance. Small weights make the intervention ineffective, while large weights force trivial solutions (e.g., always-consistent outputs).

Table 19: Sensitivity analysis for Counterfactual loss weights ($\lambda_{\text{consist}} = \lambda_{\text{discrep}}$).

Value	RAF-DB Acc.	BP4D F1	Interpretation
0.25	84.0	66.2	Penalty too weak → interventions ineffective.
0.5 (Best)	**85.0**	**67.1**	Balanced consistency–discrepancy.
1.0	84.4	66.4	Penalty too strong → suppresses causal diversity.

Mask Sharpness and Level-wise Regularization. We further analyze the mask sharpness γ (which controls the steepness of the sigmoid gating) and the relative importance of feature-level versus logit-level regularization (Table 20).

- **Effect of γ:** Setting $\gamma = 10$ achieves the best balance, producing a binary-like mask that reliably activates strong edges. Increasing γ to 20 overly sharpens the mask, hindering gradient flow, while decreasing it to 5 blurs causal boundaries.
- **Feature vs. Logit Weights:** Equal weighting (all set to 1.0) achieves the best trade-off. Reducing feature-level weights leads to larger degradation than reducing logit-level weights, indicating that feature-space constraints are critical for reliable causal inference.

Table 20: Sensitivity analysis of mask sharpness (γ) and feature/logit-level causal regularization weights. Best results are highlighted.

Parameter Setting	RAF-DB Acc.	BP4D F1
Mask Sharpness (γ)		
$\gamma = 20$	84.1	66.7
$\gamma = 10$ **(Best)**	**85.0**	**67.1**
$\gamma = 5$	84.3	66.2
Regularization Levels ($\delta_{feat}, \delta_{logit}, \eta_{feat}, \eta_{logit}$)		
All set to 1.0 (Best)	**85.0**	**67.1**
$\eta_{\text{feat}} = \delta_{\text{feat}} = 0.5,\ \eta_{\text{logit}} = \delta_{\text{logit}} = 1.0$	83.2	66.3
$\eta_{\text{feat}} = \delta_{\text{feat}} = 1.0,\ \eta_{\text{logit}} = \delta_{\text{logit}} = 0.5$	84.5	66.6
All set to 0.5	83.0	66.1

Overall, these results highlight that CausalAffect is not overly sensitive to fine-grained hyperparameter tuning, provided that the opposing forces (e.g., consistency vs. discrepancy, compression vs. alignment) are balanced. The selected weights and $\gamma = 10$ maximize performance while maintaining psychologically interpretable causal structures.

I.4 Stability of global causal relations

We evaluated the stability of the global AU→AU and AU→Expression graphs across pseudo-label thresholds, AU vocabularies, random seeds, and source datasets. Signed-weight Spearman correlation (ρ) measures agreement in edge ranking, top-20% directed Jaccard measures overlap between the strongest directed edges, and polarity agreement measures sign consistency among edges selected by both graphs.

Stability across thresholds, AU vocabularies, and random seeds. We retrained the model with pseudo-label thresholds $(\rho_+, \rho_-) \in \{(0.7, 0.3), (0.8, 0.2), (0.9, 0.1)\}$ using five matched seeds, with $(0.8, 0.2)$ as the reference setting. For vocabulary changes, we used the BP4D subsets in Appendix E: the 6-AU core, the core plus two frequent AUs, and the core plus two rare AUs, each compared with the full 12-AU model. AU→AU comparisons use shared source and target AUs; AU→Expression comparisons use shared AU sources and the same expression targets. Threshold and vocabulary comparisons were paired with their same-seed reference and averaged across five seeds. Random-seed stability was evaluated over all ten pairs formed by five independent runs of the default configuration.

Table 21 shows that threshold changes and random seeds retain signed-weight correlations of **0.85–0.93** and polarity agreement of **92.4%–96.5%**. Under AU-vocabulary changes, shared relations retain 88.4%–92.8% polarity agreement. Adding frequent AUs preserves the full graph more closely than adding rare AUs across all three metrics, highlighting the value of well-observed AU configurations for consistent relation learning.

Stability across source domains. We compared global graphs under two source-domain changes: AffectNet versus RAF-DB with BP4D fixed, and BP4D versus DISFA with AffectNet fixed. The expression-domain comparison uses the full BP4D AU vocabulary. The BP4D–DISFA comparison uses their shared induced AU subgraph: both endpoints are shared for AU→AU relations, while AU→Expression relations use shared AU sources and the common seven expression targets.

With BP4D fixed, changing the expression dataset between AffectNet and RAF-DB preserves **92.6%–94.2% polarity agreement** (Table 22). Changing the AU dataset between BP4D and DISFA retains **89.7%–90.8% polarity agreement** and signed-weight correlations of 0.76–0.78 on their shared subgraph. These results show that CausalAffect preserves a consistent core of excitatory and inhibitory relations across source domains.

Table 21: Global relation stability across pseudo-label thresholds, AU vocabularies, and random seeds. Thresholds are compared with the default $(0.8, 0.2)$ setting, and vocabulary subsets with the full 12-AU model on shared variables. Threshold and vocabulary results are averaged over five matched seeds; seed results over all ten pairs of five default runs.

Comparison	Relation	Signed ρ	Top-20% Jaccard	Polarity agreement
Threshold 0.70/0.30	AU→AU	0.90	0.76	95.2%
	AU→Expression	0.88	0.73	94.1%
Threshold 0.90/0.10	AU→AU	0.87	0.71	93.6%
	AU→Expression	0.85	0.69	92.4%
6-core vocabulary	AU→AU	0.75	0.60	89.7%
	AU→Expression	0.73	0.58	88.4%
6-core + 2 frequent AUs	AU→AU	0.84	0.69	92.8%
	AU→Expression	0.82	0.67	91.6%
6-core + 2 rare AUs	AU→AU	0.78	0.63	90.5%
	AU→Expression	0.76	0.61	89.3%
Five random seeds	AU→AU	0.93	0.80	96.5%
	AU→Expression	0.91	0.77	95.4%

Table 22: Global relation stability across source domains. The AffectNet–RAF-DB comparison uses the full BP4D AU vocabulary; the BP4D–DISFA comparison uses their shared induced AU subgraph.

Source-domain change	Relation	Signed ρ	Top-20% Jaccard	Polarity agreement
AffectNet ↔ RAF-DB	AU→AU	0.84	0.71	94.2%
(BP4D fixed)	AU→Expression	0.81	0.68	92.6%
BP4D ↔ DISFA	AU→AU	0.78	0.63	90.8%
(AffectNet fixed)	AU→Expression	0.76	0.61	89.7%

J Training and inference efficiency

We conduct representative training and inference experiments to assess the computational efficiency of CausalAffect under different scales of training data (Table 23). All experiments were conducted on a single NVIDIA A100 GPU with an Intel(R) Xeon(R) Gold 6142 CPU (2.60 GHz), running Rocky Linux. We adopt ConvNeXt-Base as the backbone with an input resolution of 112×112.

Table 23: Training and inference efficiency of CausalAffect across different experimental configurations.

Experiment	+All	AffectNet+BP4D	DISFA+RAF-DB	EmotioNet
Total Training Images	∼800K	∼257K	∼100K	∼24K
Training Batch Size	360	120	120	60
Training Epochs	80	43	21	39
Time per Epoch	∼260s	∼108s	∼98s	∼43s
Total Training Time	∼5.8h	∼1.29h	∼0.57h	∼0.46h
Peak GPU Memory Usage	28–30 GB	18–20 GB	18–19 GB	10–12 GB
Inference Speed	∼200 FPS	∼250 FPS	∼250 FPS	∼270 FPS

Large-scale configuration. CausalAffect (+All) supports joint learning from approximately 800K images across six datasets on a single A100 GPU. Training for 80 epochs takes approximately 5.8 hours with a peak GPU memory usage of 28–30 GB, while inference reaches approximately

200 FPS. These results demonstrate that large-scale learning of AU-level causal relations remains computationally practical while retaining efficient prediction.

Medium- and small-scale configurations. Smaller configurations further reduce training and memory costs while achieving approximately 250–270 FPS inference. AffectNet+BP4D and DISFA+RAF-DB require approximately 1.29 and 0.57 hours of training, with peak memory usage of 18–20 and 18–19 GB, respectively; both reach approximately 250 FPS. EmotioNet requires approximately 0.46 hours and 10–12 GB, while reaching approximately 270 FPS, demonstrating the framework's efficiency in a smaller training setting.

K Expert evaluation of the psychological plausibility of learned causal relations

To assess whether the discovered AU→AU and AU→Expression causal relations reflect *psychologically plausible mechanisms*, we conducted an additional expert validation study. This section provides the full details on the participating experts, evaluation materials, evaluation procedure, metric definitions, and aggregation methods. It then presents the evaluation results.

K.1 Participants

We invited **five domain experts** to independently evaluate the learned causal relations, including three specialists in affective cognition and emotion science, one researcher in cognitive science, and one psychologist with expertise in facial behavior and affect. All experts had **5–15 years of research experience** in their respective fields. Each expert was asked to judge whether the presented AU→AU and AU→Expression relations, which included literature-supported relations, model-discovered relations, and randomly sampled controls, were **psychologically plausible** by providing a `True/False` decision together with a confidence rating on a 1–5 scale.

K.2 Evaluation materials

We constructed **three categories of evaluation pairs**, each designed to test a different aspect of the psychological plausibility of the learned causal relations.

(A) Literature-supported "True" relations as positive controls (39 AU→Expression, excitatory only). This category contains 39 AU→Expression relations drawn from established Facial Action Coding System (FACS) documentation [15] and affective-science literature [16, 58, 38, 14, 10]. All relations in this set are excitatory, meaning the AU is known to *increase* the likelihood or intensity of a specific expression. Typical examples include "AU12 (Lip Corner Puller) increases Happiness", "AU4 (Brow Lowerer) increases Anger", and "AU1+AU2 (Inner and Outer Brow Raisers) increase Surprise". These relations serve as literature-supported positive controls, enabling us to verify that experts consistently recognize well-established psychological dependencies.

(B) CausalAffect-discovered relations as evaluation targets (55 AU→AU / 45 AU→Expression, excitatory and inhibitory). This category contains 100 relations extracted from our learned causal graphs, covering both AU→AU and AU→Expression directions and including both excitatory and inhibitory effects. A subset of these relations overlaps with the literature-supported set A, demonstrating that the model can recover literature-supported patterns. The remaining edges provide new causal hypotheses, including subtle inhibitory relations. Examples include "AU17 (Chin Raiser) increases AU24 (Lip Pressor)", "AU26 (Jaw Drop) inhibits AU20 (Lip Stretcher)", "AU15 (Lip Corner Depressor) increases Disgust", and "AU6 (Cheek Raiser) inhibits Sadness". All overlapping relations between sets A and B were removed before expert evaluation. This category evaluates the psychological plausibility of the causal relations recovered and newly discovered by CausalAffect.

(C) Random relations as negative controls (68 AU→AU / 28 AU→Expression, excitatory and inhibitory). This category consists of 96 randomly sampled relations across AU→AU and AU→Expression pairs that appear in neither set A nor set B. These relations include both excitatory and inhibitory directions but lack any known support from FACS or affective psychology. Example items include "AU4 (Brow Lowerer) increases AU12 (Lip Corner Puller)", "AU20 (Lip Stretcher) inhibits Fear", and "AU9 (Nose Wrinkler) increases Happiness". These pairs serve as negative

controls for assessing experts' rejection of unsupported relations and provide a baseline against which the model-discovered relations in set B can be assessed.

K.3 Evaluation procedure

All evaluation pairs from sets A, B, and C were merged and deduplicated, then randomly shuffled for each expert independently. Experts were blind to both the originating set (A, B, or C) and whether a relation was literature-supported, recovered by the model, or entirely novel. Each expert judged every relation as **True** (psychologically plausible) or **False** (implausible), and additionally provided a **1–5 confidence score**. This procedure yielded **five independent evaluations per relation**, enabling us to assess both overall plausibility and inter-expert reliability.

K.4 Evaluation metrics

We use several quantitative metrics to evaluate both the psychological plausibility of the relations and the reliability of expert judgments. Throughout this section, let N denote the number of relations being evaluated, and K the number of experts (here $K = 5$). For each relation $i \in \{1, \ldots, N\}$ and expert $e \in \{1, \ldots, K\}$, the binary variable $y_{i,e} \in \{0, 1\}$ represents the expert's plausibility judgment, where 1 indicates that the relation was considered psychologically plausible ("True") and 0 indicates "False". The corresponding confidence rating is denoted by $c_{i,e} \in \{1, 2, 3, 4, 5\}$. The indicator function $\mathbf{1}(\cdot)$ returns 1 if its input condition is satisfied and 0 otherwise.

Expert True Rate. The first metric is the overall proportion of "True" decisions across all relations and experts. It quantifies how frequently experts endorse the relations in a given set as psychologically plausible. Formally, this is computed as

$$\text{TrueRate} = \frac{1}{NK} \sum_{i=1}^{N} \sum_{e=1}^{K} \mathbf{1}(y_{i,e} = 1)\,.$$

For each relation, we compute the fraction of its five expert judgments labelled "True". The Expert True Rate is the mean of these relation-level endorsement fractions; the corresponding SD describes their variation across relations within each set.

Majority-True Rate. A relation is considered "majority plausible" if at least three of the five experts judged it as "True". Let $\sum_{e=1}^{K} \mathbf{1}(y_{i,e} = 1)$ denote the number of "True" votes for relation i. The majority-true rate is defined as

$$\text{MajTrue} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\left(\sum_{e=1}^{K} \mathbf{1}(y_{i,e} = 1) \geq 3 \right).$$

High-Confidence True Rate. To assess plausibility in high-confidence judgments, we compute the proportion of "True" votes restricted to evaluations with confidence ratings $c_{i,e} \geq 4$. This metric is given by

$$\text{HighConfTrue} = \frac{\displaystyle\sum_{i=1}^{N} \sum_{e=1}^{K} \mathbf{1}(y_{i,e} = 1 \wedge c_{i,e} \geq 4)}{\displaystyle\sum_{i=1}^{N} \sum_{e=1}^{K} \mathbf{1}(c_{i,e} \geq 4)}.$$

Inter-expert agreement. Inter-expert agreement is assessed using Fleiss' κ:

$$\kappa = \frac{\bar{P} - P_e}{1 - P_e},$$

where $\bar{P}$ is the observed agreement among experts and P_e is the expected chance-level agreement.

K.5 Results and analysis

K.5.1 Overall plausibility comparison

Table 24 compares expert endorsement across the three evaluation sets. Literature-supported relations (set A) receive an Expert True Rate of 0.88, compared with 0.23 for random controls (set C). The relations discovered by *CausalAffect* (set B) achieve a rate of 0.75, which is **more than three times the rate** for random controls and closer to that of the literature-supported relations. The majority-vote results follow the same ordering: 84% of model-discovered relations are endorsed by at least three experts, compared with 92% of literature-supported relations and 9% of random controls.

The advantage over random controls also holds for high-confidence judgments. The High-Confidence True Rate is 0.70 for the model-discovered relations, compared with 0.19 for random controls.

Table 24: Expert validation statistics across the three evaluation sets. Expert True Rate values are mean $\pm$ SD across relations.

Relation Set	#Pairs	Expert True Rate	Maj-True Rate	High-Conf True	Interpretation
A: Literature-supported	39	0.88 ± 0.21	0.92	0.90	Experts strongly endorse well-established psychological relations.
B: CausalAffect	100	0.75 ± 0.24	0.84	0.70	Majority of model-discovered relations judged plausible; performance far above random controls.
C: Random negatives	96	0.23 ± 0.20	0.09	0.19	Experts reliably reject implausible, randomly generated relations.

The endorsement rates show that experts judge the model-discovered relations more plausible than random controls. These results provide external support for the **psychological plausibility** of the CausalAffect-discovered relations. The following analysis examines this support specifically for inhibitory relations.

K.5.2 Psychological plausibility of inhibitory relations

To evaluate the psychological plausibility of the **inhibitory relations** discovered by CausalAffect, we further stratified the CausalAffect-discovered relation set (B) by both *relation type* (AU→AU vs. AU→Expression) and *causal polarity* (excitatory vs. inhibitory), and evaluated expert judgments separately for each subset.

As shown in Table 25, all four subsets receive plausibility scores substantially above the random-negative baseline. In particular, the inhibitory subsets achieve True Rates of 0.67 (AU→AU) and 0.77 (AU→Expression), together with majority endorsement rates of 0.83 and 0.87, respectively. Both inhibitory True Rates are markedly higher than the random control True Rate (0.23), demonstrating that experts judge the inhibitory edges to be **psychologically plausible**. These judgments support the interpretation of the learned inhibitory relations as meaningful **regulatory patterns** in facial musculature and emotional expression.

Table 25: Breakdown for CausalAffect-discovered relations (Set B).

Subset	#Pairs	True Rate	Majority-True Rate	Interpretation
AU→AU, excitatory	32	**0.72**	**0.78**	Positive AU interactions broadly accepted by experts.
AU→AU, inhibitory	23	**0.67**	**0.83**	Inhibitory AU→AU relations judged plausible and non-random.
AU→Expression, excitatory	22	**0.85**	**0.91**	Strong alignment with known affective expression dynamics.
AU→Expression, inhibitory	23	**0.77**	**0.87**	Inhibitory cues viewed as meaningful psychological structure.

These results highlight CausalAffect's ability to discover **novel and subtle inhibitory relations** that external experts judge to be psychologically plausible. Combined with the endorsement patterns in sets A and C, these findings support the inhibitory relations as interpretable causal hypotheses about AU and expression structure.

L Global causal relation analysis

CausalAffect constructs global causal graphs over both AU→Expression and AU→AU spaces, revealing human-aligned interpretable, directional, and semantically grounded dependencies. It captures not only canonical facial expression cues and co-activation patterns but also inhibitory and statistically inaccessible relations—offering structural priors that go beyond co-occurrence statistics. In all CausalAffect relation heatmaps, edge weights are multiplied by 100 for visualisation.

L.1 AU→Expression global causal relation

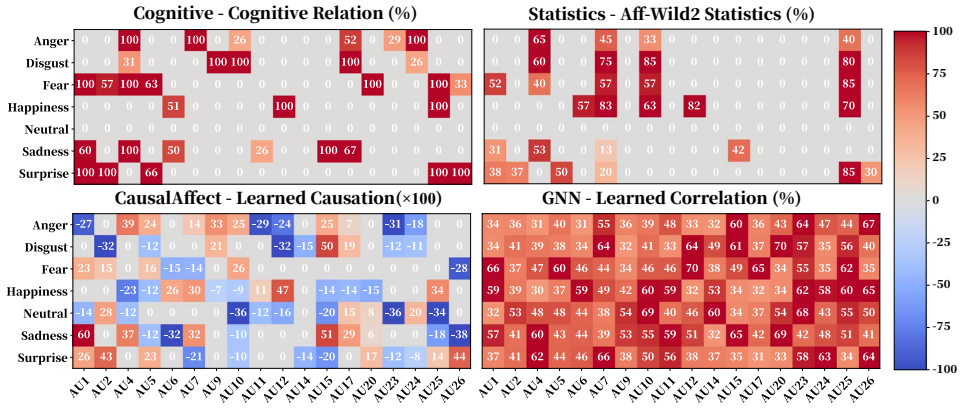

Figure 4: Comparison of AU→Expr Relations across cognitive priors, statistical co-occurrence, GNN-learned correlation (trained in the +All setting), and CausalAffect-learned causal relations (trained in the +All setting).

Alignment with human priors and expression semantics. The AU→Expression causal graph (Figure 4) learned by **CausalAffect** reveals both **causally grounded** and **psychologically plausible** structures. The discovered relations align closely with established findings in facial behavior research [15], as well as with prior **statistical co-occurrence relations** [44] and **cognitive models** [14]. Notably, CausalAffect recovers several canonical expression markers, including AU12 → Happiness, AU1,AU4,AU15 → Sadness, and AU2,AU26 → Surprise, all of which are supported by both *cognitive neuroscience evidence* and empirical patterns observed in datasets such as *Aff-Wild2*. These relations are consistent with well-understood affective mechanisms: AU12 (lip corner puller) is the primary indicator of enjoyment-related expressions such as happiness or amusement; AU1, AU4, and AU15 (inner brow raiser, brow lowerer, and lip corner depressor) are prototypical components of Sadness, reflecting upper-face tension, concern, and downward mouth pull associated with grief or emotional pain; AU2 and AU26 (outer brow raiser and jaw drop) are hallmark components of Surprise, with outer-brow elevation and jaw opening forming a characteristic upper- and lower-face response to sudden or unexpected stimuli. Importantly, these dependencies are learned without any AU→Expression co-annotation. This highlights CausalAffect's ability to infer semantically aligned and interpretable structures in a fully data-driven manner—effectively bridging low-level facial actions and high-level affective understanding.

Modeling inhibitory causal relations. Beyond capturing canonical positive dependencies, CausalAffect also discovers *inhibitory causal relations*—directed negative influences that are largely absent from existing statistical or cognitive structures. These relations do not merely indicate suppression or co-inhibition, but rather reflect a form of **inhibitory precondition**: the absence of a particular AU provides *causal support* for inferring a given expression. In the learned global graph, AU6 (cheek raiser) and AU26 (jaw drop) receive negative weights for Sadness. These edges represent competing cues alongside the positive contributions of AU1, AU4, and AU15. Such negative causal links highlight the model's ability to reason not only about what *is present*, but also about what *is absent* for an emotion to be plausible. This aligns with psychological theories of emotional exclusivity and supports more precise disambiguation in overlapping facial configurations. Overall, these inhibitory relations allow CausalAffect to move beyond symmetric co-activation and toward truly directional understanding of facial expressions.

Causal graph as a diagnostic prior for label auditing. Beyond modeling facial behavior, the learned causal graph also demonstrates strong **diagnostic utility**. While **Neutral** is conventionally assumed to co-occur with the absence of active AUs in psychology, CausalAffect uncovers consistent positive causal links from AU24 (lip pressor), AU2 (outer brow raiser), and AU17 (chin raiser) to Neutral. These findings **challenge conventional assumptions**, suggesting that many Neutral-labeled samples actually exhibit subtle but structured AU activations. These causal patterns provide a structural basis for identifying potential annotation inconsistencies in Neutral-labelled samples and directing label auditing. In this context, the causal graph functions as a **structural prior** that can support **label auditing**, **confidence calibration**, and improved **annotation robustness**. By identifying unexpected or semantically inconsistent activations within annotated Neutral instances, CausalAffect provides a principled mechanism for evaluating annotation quality and guiding data refinement.

L.2 AU→AU global causal relation

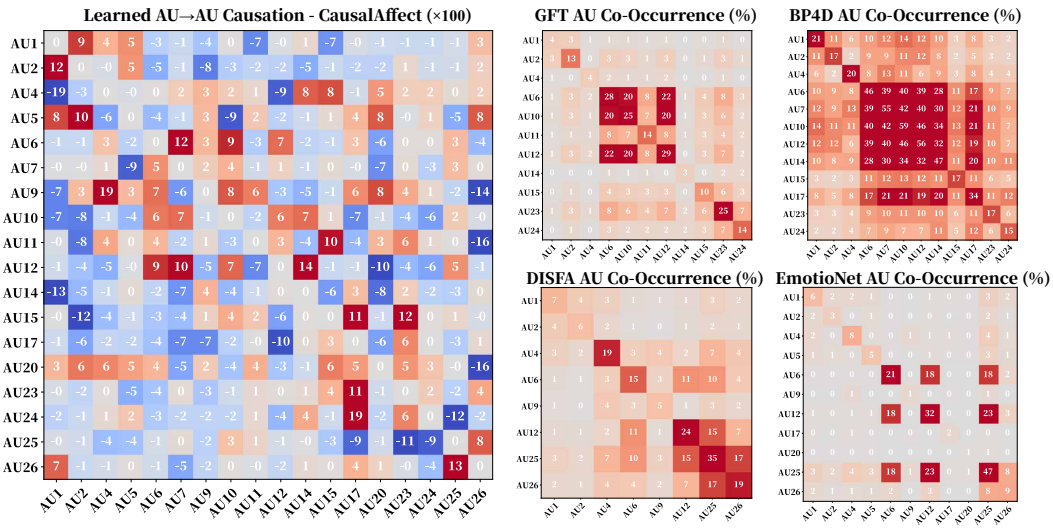

Figure 5: Comparison between the AU→AU causal relations learned by CausalAffect (trained with the +All setting) and AU co-occurrence statistics from four datasets (GFT, DISFA, EmotioNet, BP4D).

To the best of our knowledge, this work presents ***the first*** data-driven framework to learn human-aligned AU→AU causal dependencies from weakly labeled data. Unlike AU→Expression mappings, which have been studied in psychology and affective computing, there exists no established ground truth or cognitive theory that defines directed causal relations between AUs themselves. As a result, we evaluate the plausibility of our learned causal graph by comparing it with AU co-occurrence statistics derived from four widely-used facial expression datasets: GFT, DISFA, EmotioNet, and BP4D.

While co-occurrence statistics provide a simple way to analyze AU correlations, they suffer from several inherent limitations: **(i) Incomplete relational coverage:** Co-occurrence statistics cannot establish pairwise relations across all 18 AUs due to the lack of overlapping AU annotations among the datasets. In contrast, CausalAffect learns a unified causal graph over the full 18-AU vocabulary (see the left panel of Figure 5). **(ii) Symmetry assumption:** Joint co-occurrence probabilities are inherently symmetric, i.e., $P(AU_i, AU_j) = P(AU_j, AU_i)$, which fails to capture the directional nature of inter-AU influences. In contrast, our learned causal graph reveals asymmetric causal dependencies that reflect realistic directional interactions between AU pairs. **(iii) Lack of inhibitory relations:** Joint co-occurrence statistics capture co-activation patterns. They are unable to model inhibitory or suppressive relationships. In contrast, CausalAffect captures both excitatory and inhibitory causal effects, enabling finer-grained interpretation of inter-AU dynamics.

Compared to the AU→Expression causal links (Figure 4), which exhibit strong influence patterns, the AU→AU relations tend to have lower overall magnitudes, with the maximum absolute value reaching approximately 19. AU→AU interactions capture local muscular coordination or antagonism, such as

the co-activation of AU6 and AU7 around the eyes, rather than higher-level expression structure. As a result, these causal effects are more distributed and less dominant than the structured compositional dependencies observed in AU→Expression modeling.

Alignment with human priors and co-occurrence trends. Despite being trained without explicit supervision from domain knowledge or handcrafted priors, CausalAffect successfully recovers human-aligned AU→AU causal patterns that are consistent with known trends in facial actions. For instance, CausalAffect learns strong positive causal relations such as AU7 → AU12 and AU6 → AU12, which correspond to canonical activation pathways in genuine (Duchenne) smiles. These relations are also prominent in dataset-level statistics—e.g., GFT AU6 → AU12 co-occurrence at +22%, while BP4D shows even stronger associations for AU6 → AU12 and AU7 → AU12. Similarly, the model learns AU25 → AU26, capturing the natural progression from lip parting to jaw drop, which aligns with co-occurrence strengths observed in DISFA. Another example is AU10 → AU12, often seen in expressions of contempt or disgust, which is mirrored by co-occurrence patterns in GFT and BP4D.

Capturing inhibitory causal relations. One of the most distinctive advantages of CausalAffect over prior approaches is its ability to model *inhibitory causal relations* between AUs—i.e., directed negative influences that reflect mutual exclusivity or muscular suppression. This capacity is absent from unsigned co-occurrence matrices, which do not explicitly encode negative edges. Such inhibitory relations are especially important in AU detection, where many AUs are known to be semantically incompatible. For example, the global graph assigns AU1 → AU4 a weight of -19 and the reverse edge a weight of $+4$. The negative edge is consistent with antagonism between inner brow raising and brow lowering. The model also learns negative weights from AU26 to AU20 and AU11, linking jaw drop to lip stretching and nasolabial deepening within the learned graph. In addition, we observe semantically suppressive relationships in the upper and lower face. For instance, AU1 ⊣ AU14 reflects the tension between dimple-induced controlled smiles and inner brow raising, which signal conflicting emotional states such as restrained positivity versus concern. The relation AU2 ⊣ AU15 highlights the opposition between lip corner depression (sadness) and outer brow raising (surprise), while AU20 ⊣ AU12 encodes the mismatch between smiling and horizontal lip tension typically associated with fear or anxiety.

Directional relations beyond co-occurrence. The learned graph assigns a positive weight to AU15 → AU20, a pair without joint coverage in the four co-occurrence panels in Figure 5. The AU4–AU9, AU15–AU11, and AU17–AU24 pairs appear in the co-occurrence panels. The learned graph further distinguishes the source and target AUs through the directed edges AU4 → AU9, AU15 → AU11, and AU17 → AU24.

Effect of DAG Constraint: To guide the model toward learning interpretable structures, we incorporate a DAG constraint during training. This constraint serves to prioritize *directed, asymmetric, and semantically meaningful* causal relationships over symmetric statistical associations. In particular, it helps suppress noisy bidirectional correlations and encourages the model to resolve causal directionality. Our implementation adopts the DAG constraint as a *soft regularization*, allowing localized reciprocal connections within the learned causal graph. For example, AU1 → AU2 has a weight of +12 and AU2 → AU1 has +9. In psychological and behavioral literature [15], many AU pairs are known to exhibit *symbiotic or reciprocal* relationships. AUs such as AU1 and AU2 frequently co-activate in expressions like surprise or concern. This design allows the model to retain the flexibility needed to capture biologically plausible reciprocity in AU behavior, while still being biased toward uncovering dominant and interpretable directional dependencies.

M Case study: sample-adaptive causal relation analysis

In this section, we present examples of both AU→Expression and AU→AU sample-adaptive causal graphs to demonstrate how **CausalAffect** dynamically constructs instance-specific causal structures. These case studies reveal how the model adapts its reasoning to each individual input, capturing both prototypical and idiosyncratic facial dynamics beyond what is reflected in global statistical trends.

M.1 AU→Expression sample-adaptive causal relation

Sample-adaptive graph aligns with global priors while preserving expression-specific consistency. Through systematic case-by-case analysis across six primary emotions and neutral states

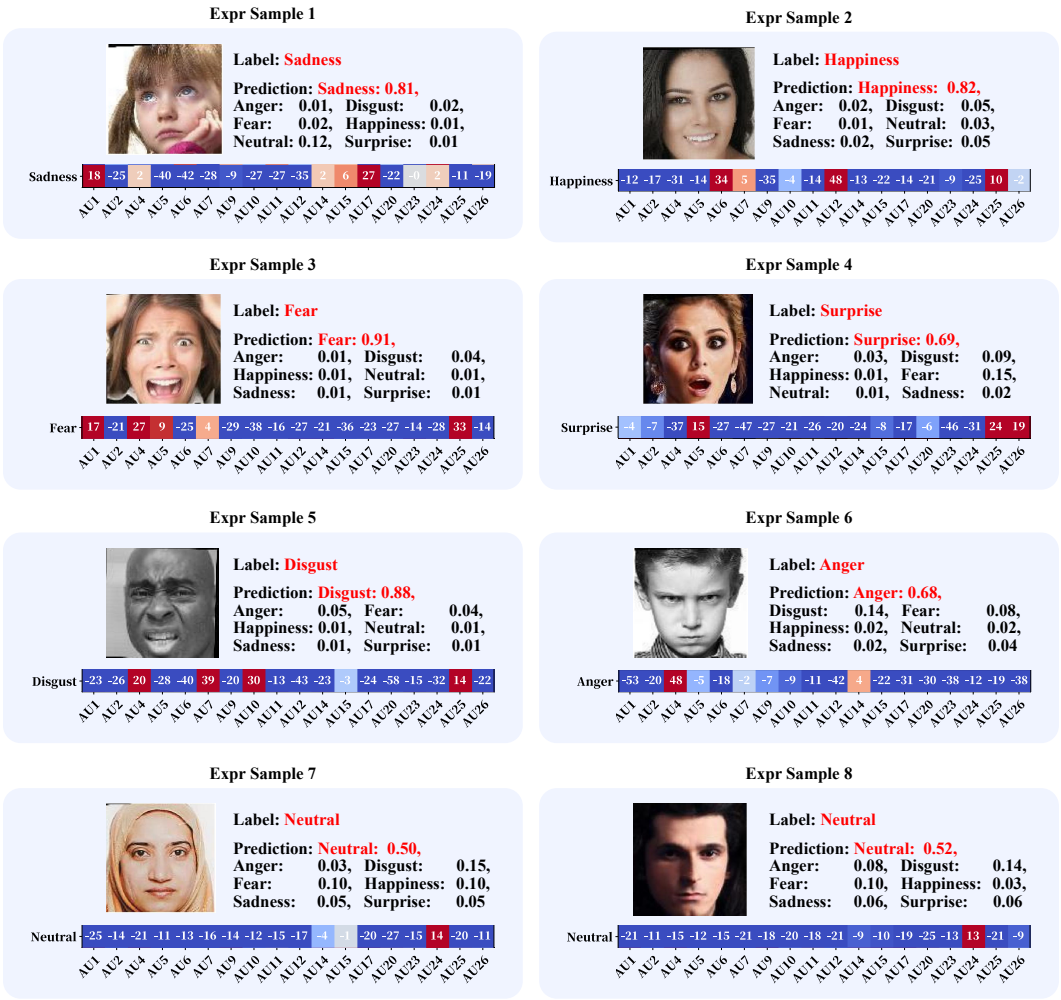

Figure 6: Sample-adaptive AU→Expression causal relations dynamically inferred for individual samples. We randomly sampled one image per basic expression (and two for Neutral) to illustrate how CausalAffect captures instance-specific causal structures. For clarity, only the subgraph corresponding to the predicted expression is visualized.

(Figure 6), we observe that **Sample-adaptive graph** consistently recovers AU→Expression structures that align well with global affective patterns. For example, in **Sample 2 (Happiness)**, Sample-adaptive graph identifies AU12 and AU6 as dominant contributors—closely matching the global graph. Similarly, **Sample 4 (Surprise)** features AU25, AU26, and AU5, which are also prominent in the global graph. In both cases, Sample-adaptive graph additionally suppresses conflicting AUs such as AU4 (in Sample 2) or AU23 (in Sample 4), maintaining semantic consistency with the global graph.

While Sample-adaptive graph accurately replicates global causal trends, it also demonstrates strong adaptability in tailoring inference to the sample-specific AU configuration. For example, in **Sample 3 (Fear)**, the sample-adaptive graph lowers the signed weight of AU10 from $+26$ to -38 and that of AU5 from $+16$ to $+9$ relative to the global graph. A similar pattern appears in **Sample 6 (Anger)**, where Sample-adaptive graph focuses nearly all attribution on AU4, whereas the global graph distributes importance across AU10, AU9, and AU15—none of which are visibly active in the instance. In **Sample 5 (Disgust)**, the sample-adaptive causal graph highlights AU10, AU7, and AU4 as the primary causal drivers—reflecting the visible upper-face tension and mid-face wrinkling characteristic of disgust. Meanwhile, AU6 and AU12 are strongly suppressed. This contrasts with the global AU→Disgust graph, which emphasizes AU15 as the dominant lower-face contributor, along with inhibitory weights on AU12 and AU2. The discrepancy illustrates how CausalAffect dynamically adjusts its causal reasoning based on observed facial features, prioritizing context-relevant AUs.

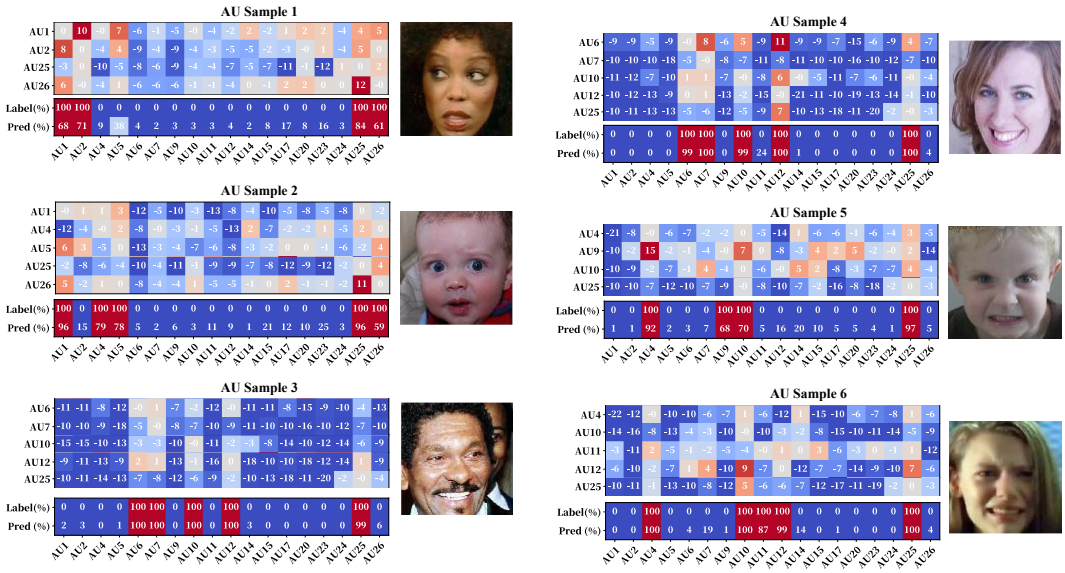

Figure 7: Sample-adaptive AU→AU causal relations for six randomly selected EmotioNet images. Rows show predicted active target AUs and columns show source AUs.

Instance-level causal graphs can guide label auditing. In **Neutral Samples 7 and 8**, the visible facial activity is subtle. The sample-adaptive graph assigns AU24 (lip pressor) positive weights of $+14$ and $+13$. AU2 (outer brow raiser) receives -14 and -11, while AU17 (chin raiser) receives -20 and -19. The global graph assigns positive weights to all three AUs for Neutral, so the polarity changes in AU2 and AU17 show how the model adjusts the evidence it uses for individual faces.

The positive AU24 edge gives annotators a specific feature to revisit. A slight lip press can be difficult to distinguish from a resting mouth posture in a still image. The sample-adaptive graph thus provides a structural cue for reviewing ambiguous annotations and targeting data-quality checks.

M.2 AU→AU sample-adaptive causal relation

Unlike dataset-level co-occurrence, CausalAffect learns fine-grained, sample-specific AU→AU causal graphs (Figure 7). In **Sample 1**, the model assigns positive weights from AU2, AU5, AU26, and AU25 to AU1. The first three edges are also positive in the global graph, while AU25 → AU1 changes from -1 globally to $+4$ in this sample. The AU25 → AU26 edge remains positive ($+13$ globally and $+12$ locally), retaining the learned link from lip parting to jaw drop. The two edges from AU25 illustrate how the graph preserves a global coordination pattern while adjusting a separate relation to the individual face.

In **Sample 1**, AU25 receives mostly negative weights, including -10 from AU4 and -11 from AU17, alongside a small positive weight of $+2$ from AU26. A similar pattern appears in **Sample 5**, where AU25 receives negative weights from AU20 (-8), AU23 (-18), and AU17 (-16). The negative AU20 edge suggests competition between horizontal lip stretching and lip parting in this sample. These patterns support an exclusion-based reading of the model's prediction. Weak activity in competing AUs is consistent with predicted lip parting, even when positive inputs are sparse. Sample 5 also shows two positive inputs to AU9 (nose wrinkler) from AU4 and AU10 ($+15$ and $+7$). With the AU10 → AU4 edge at zero at the displayed precision, these are converging inputs to AU9. The graph thus distinguishes coordination around AU9 from competing influences around AU25 in the same face.

Causal inference of non-root AU activations via inhibitory dependencies. A central strength of CausalAffect is its ability to model **non-root AUs**—those that do not originate from direct causal initiators but instead emerge as residual consequences of **the absence of suppressive or competing AUs**. This is clearly demonstrated in **Sample 3**, where all predicted AUs (e.g., AU6, AU7, AU12, AU25) are inferred primarily via negative causal inputs. For example, AU6 is suppressed by AU20, AU5, AU1 and AU2, while AU12 is inferred through the absence of AU14 and AU20. These patterns

suggest that the facial configuration reflects controlled or ambiguous expression states (e.g., social smiling or emotional masking), where expressive AUs are not initiated directly but emerge under mutual inhibition constraints.

This form of inference is primarily enabled by **soft DAG constraint**, which promotes directional causal structures by discouraging cyclic connections. By discouraging indiscriminate bidirectional correlations, the DAG bias guides the model to identify minimal and interpretable causal pathways—including those where an AU's activation is inferred not through direct excitation, but via the causal absence of its antagonistic counterparts. When dominant upstream AUs are absent—as in **Sample 2**, where AU2 is inactive—the model adaptively reweighs negative contributors to AU1, incorporating AU6, AU11, AU9, and AU15 to compensate for the lack of canonical support. Similarly, in **Sample 6**—a sadness-pain expression—AU12 is inferred not from global graph drivers like AU6, but through modest support from AU10, AU25, and AU7, suggesting indirect lower-face coordination. In contrast, AU4—despite receiving strong inhibition from AU1, AU15, and others—remains active, reflecting how the model integrates competing causal influences in its prediction. Together, these observations demonstrate that CausalAffect goes beyond surface-level co-activation and captures latent causal architecture shaped by antagonism, compensation, and expressive tension—key factors in modeling natural, psychologically grounded facial dynamics.

Psychological interpretation of AU→AU causal differences across expressions. A comparison of samples with similar active AUs but different causal structures further illustrates CausalAffect's nuanced understanding of expressive dynamics. Despite sharing AU6, AU7, AU10, AU12, and AU25, **Samples 3 and 4** have distinct AU→AU graphs. In **Sample 3**, these AUs are inferred primarily via inhibitory input—e.g., AU7 is negatively influenced by AU5, AU20, and AU24, and AU12 is suppressed by AU14, AU11, and AU20. This structure highlights the role of inhibitory causal interactions in shaping the observed AU configuration, providing a causal perspective on expressive modulation [78]. In contrast, **Sample 4** exhibits a coherent and positively coordinated causal graph: AU6 is supported by AU12, AU7, AU10, and AU25, and AU25 is driven by AU12. This feedforward causal configuration captures coordinated reinforcement among smile-related AUs. These differences align with psychological theories of expressive modulation—such as **Duchenne vs. non-Duchenne smiles, emotional masking, or communicative gestures**—highlighting that the same AUs can play different causal roles depending on the expressive context.

N Image-level AU intervention validation

The feature-level counterfactual module in CausalAffect enforces intervention-style consistency and discrepancy in the disentangled AU latent space. The image-level study tests whether the learned relations remain informative when interventions modify visible facial appearance rather than latent AU features. To provide an additional external validation, we conduct an **image-level AU intervention** study. This study tests whether explicit visual edits of target AUs induce prediction changes and graph adaptations that are consistent with the learned AU→Expression and AU→AU structures.

N.1 Protocol

The illustrative examples in Figures 8 and 9 start from source images confidently predicted as *Happiness*. This choice provides a controlled starting point because happiness is characterized by a relatively stable AU configuration, usually involving AU6, AU7, AU12, and AU25. For each source image, we edit the facial appearance by activating a target set of AUs associated with another expression. The edited AU sets are selected according to established FACS-style expression compositions and the positive AU→Expression relations learned by CausalAffect. We then feed both the original and the edited images into the same trained CausalAffect model and visualize three outputs:

- the expression prediction before and after intervention;
- the predicted AU activations after the visual AU edit;
- the sample-adaptive AU→Expression and AU→AU graphs inferred from the edited image.

In Figures 8 and 9, the row labeled "Label" gives the intended intervention target for each edited sample. The row labeled "Pred" shows the model prediction on the edited image. Thus, a successful

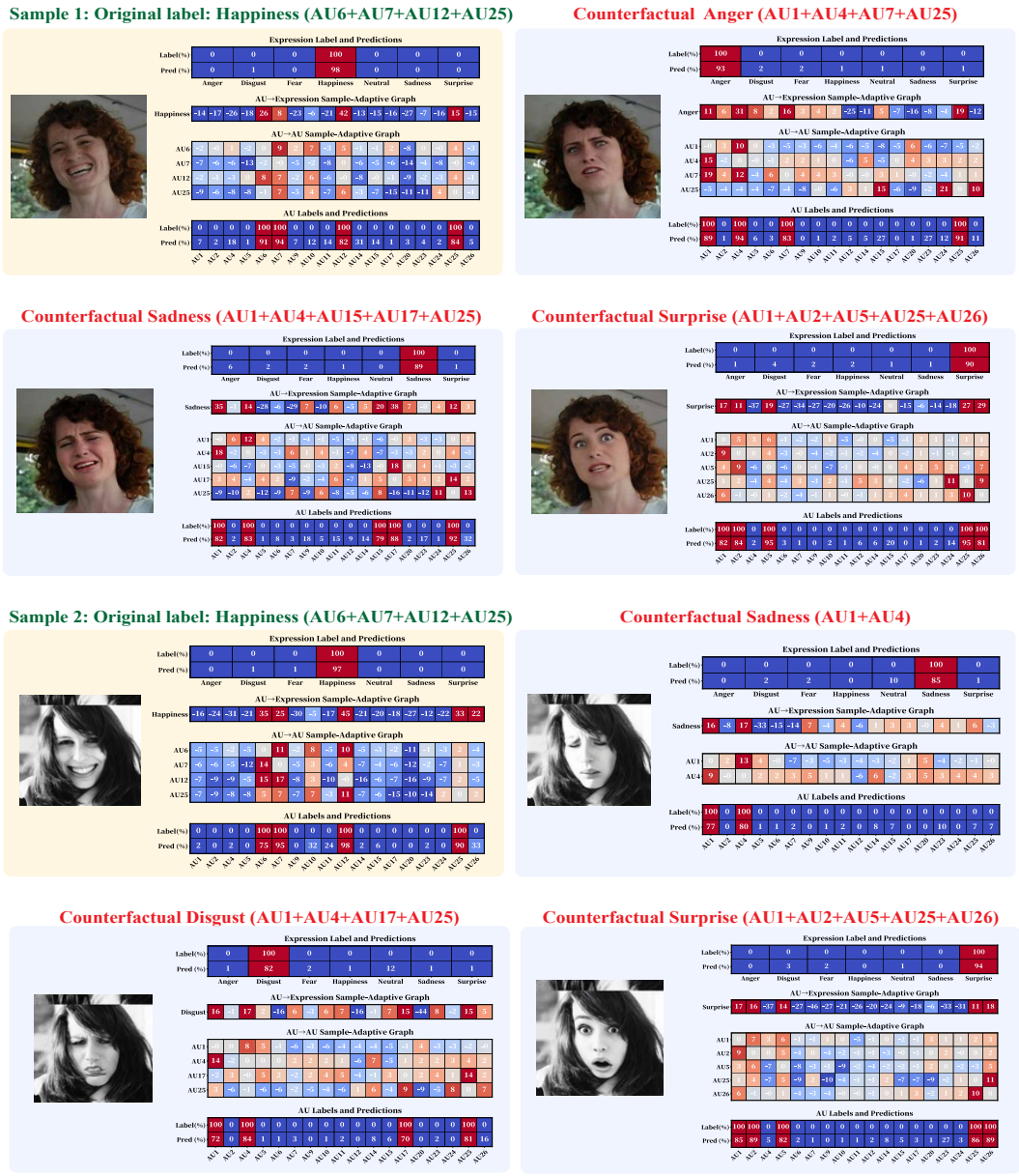

Figure 8: **Image-level AU intervention examples for Samples 1–2.** Starting from happiness images, we edit target AU sets corresponding to anger, sadness, disgust, and surprise. For each original or edited image, we show expression predictions, the sample-adaptive AU→Expression graph, the sample-adaptive AU→AU graph over the active target AUs, and AU predictions. The edited images induce expression changes consistent with the intended AU interventions and trigger corresponding graph reconfiguration.

intervention should satisfy three criteria: **(i)** the predicted expression should shift toward the intended target expression; **(ii)** the edited AUs should be detected as active by the AU branch; and **(iii)** the sample-adaptive graph should reorganize its signed relations around the edited AU evidence rather than preserving the original happiness structure.

We extend this analysis to a paired study of 36 held-out source faces (18 from AffectNet and 18 from RAF-DB), with 12 assigned to each target expression. Each source produces three outputs: a Target (T) edit activating the predefined AU set for expression y_T and its corresponding routes R_T; a semantic Control (C) edit activating a different, coherent AU set for expression y_C and routes R_C;

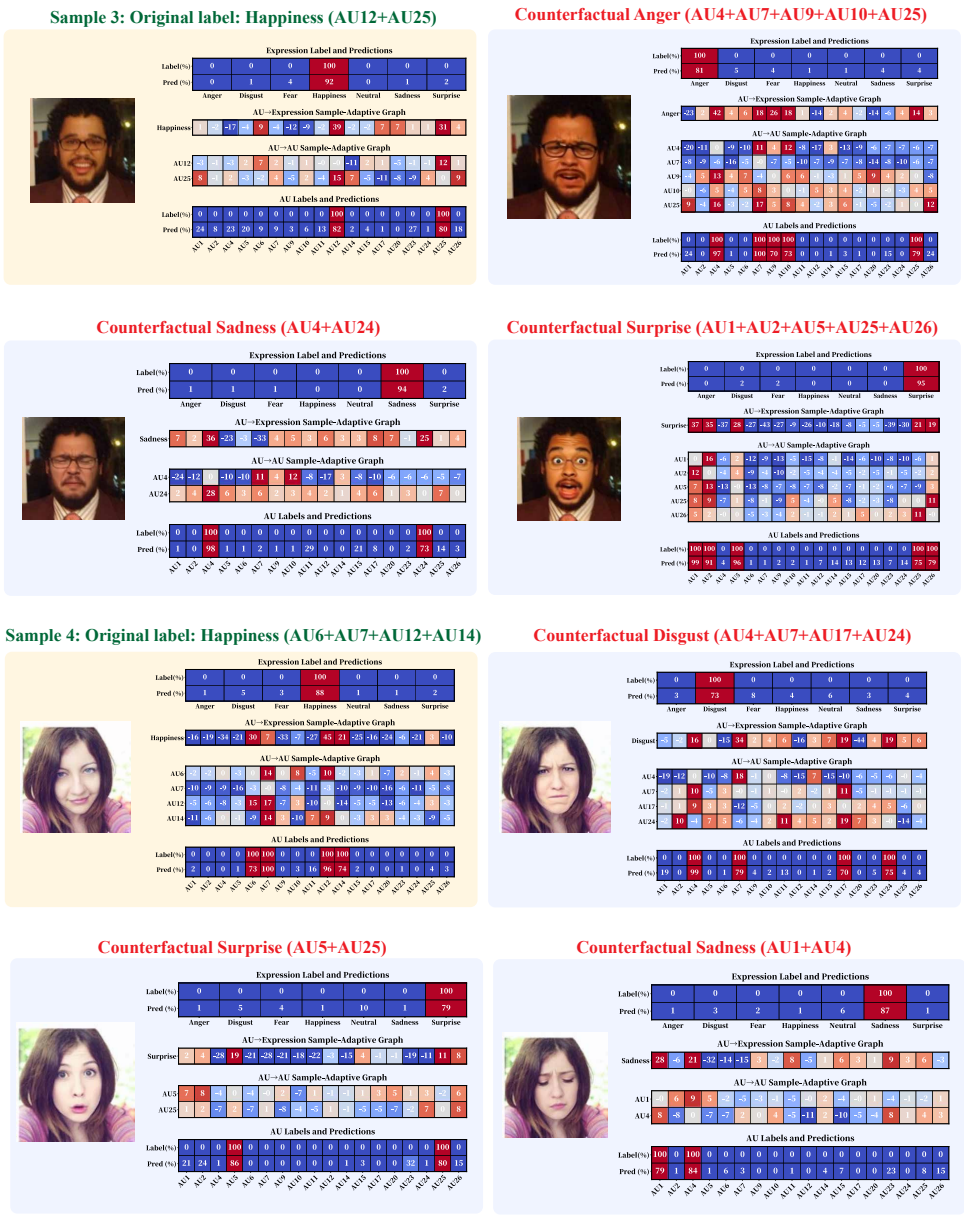

Figure 9: **Image-level AU intervention examples for Samples 3–4.** The same source expression can be transformed into different target expressions by editing different AU sets. CausalAffect responds by shifting the predicted expression and by adapting the sample-specific causal graph to the edited AU evidence.

and a Reconstruction without an AU change. Target and Control are matched in the number of edited AUs and total requested intensity. For example, a Sadness edit using AU1+AU4+AU15 is paired with a Happiness Control using AU6+AU12+AU25. The Control is therefore expected to induce a response towards its own expression and routes. The three outputs share the source image, diffusion seed, prompt template, and sampling parameters, yielding 108 outputs with 36 independent statistical units.

Three evaluators, blinded to the requested AU pattern, assigned expression, and model predictions, independently assessed visible AU changes and identity, pose, and background preservation. Overall, 99/108 outputs (91.7%) satisfy the prespecified verification criteria, with Fleiss' $\kappa = 0.73$. The nine

mutually exclusive failures comprise five insufficient AU changes, three unintended AU co-activations, and one generation artefact. All outputs are retained in the evaluation.

To isolate the contribution of relational modelling, we compare CausalAffect with Backbone + Dis (Table 15, Row 2), which uses the same backbone and disentangled AU representations but removes the global and sample-adaptive causal graphs and graph-dependent counterfactual intervention.

N.2 Expression-level intervention consistency

Let $p(y)$ denote the model-predicted probability of expression label y, and let Δ_T and Δ_C denote edited-minus-source changes under Target and Control edits, respectively. We define the prediction-level crossover interaction as

$$I_p = [\Delta_T p(y_T) - \Delta_T p(y_C)] - [\Delta_C p(y_T) - \Delta_C p(y_C)], \tag{31}$$

This interaction is positive when the Target edit favours y_T while the Control edit reverses the response towards its own y_C.

As shown in Table 26, CausalAffect correctly aligns both active edits with their respective expressions for **27/36 sources (75.0%)**, compared with 19/36 (52.8%) for Backbone + Dis, an improvement of **22.2 percentage points** (95% paired-bootstrap confidence interval (CI): $[5.6, 38.9]$; exact paired McNemar $p = 0.039$). The probability-level crossover interaction increases from 0.34 to 0.58, a paired gain of 0.24 (95% CI: $[0.14, 0.34]$; paired permutation $p < 0.001$). Reconstruction preserves the original prediction in more than 90% of cases for both models. These results show that CausalAffect more effectively aligns its predictions with the distinct AU evidence introduced by each active edit.

Table 26: Prediction-level edit–label alignment over 36 paired sources. Brackets denote 95% CIs: Wilson intervals for rates and source-level bootstrap intervals for the interaction.

Metric	**Backbone + Dis**	**CausalAffect**
Target $\rightarrow y_T$	28/36 (77.8%) $[61.9, 88.3]$	**31/36 (86.1%)** $[71.3, 93.9]$
Control $\rightarrow y_C$	25/36 (69.4%) $[53.1, 82.0]$	**30/36 (83.3%)** $[68.1, 92.1]$
Both aligned	19/36 (52.8%) $[37.0, 68.0]$	**27/36 (75.0%)** $[58.9, 86.2]$
Interaction I_p	0.34 $[0.26, 0.42]$	**0.58** $[0.50, 0.66]$
Reconstruction preserved	33/36 (91.7%) $[78.2, 97.1]$	**34/36 (94.4%)** $[81.9, 98.5]$

In the four illustrative source images in Figures 8 and 9, the original images are consistently recognized as *Happiness*, with prediction confidences of 98%, 97%, 92%, and 88%, respectively. Across these twelve illustrative edits, all edited images are classified as the intended target expression. The target-expression confidences are summarized in Table 27. Across these twelve edits, the average confidence in the edited target expressions is 86.8%, indicating a strong prediction response to visible AU changes.

The results show several consistent patterns. First, activating sadness-related upper-face AUs, such as AU1 and AU4, reliably shifts the prediction from happiness to sadness. For example, editing only AU1+AU4 in Sample 2 changes the prediction to sadness with 85% confidence, and the same AU pair in Sample 4 yields sadness with 87% confidence. This suggests that CausalAffect does not require a complete prototypical AU set to respond to a target expression; instead, it can infer the expression from a compact but causally informative subset when the edited AUs are strong target drivers.

Second, surprise edits involving AU1, AU2, AU5, AU25, and AU26 produce high target confidence in Samples 1–3, reaching 90%, 94%, and 95%, respectively. These edits activate both upper-face surprise cues and lower-face mouth-opening cues, yielding a complete AU configuration for surprise. In contrast, the weaker surprise edit in Sample 4 uses only AU5+AU25 and obtains a lower but still successful confidence of 79%. This difference is informative: the model assigns higher confidence when the edited AU set covers the canonical surprise pathway more completely, while a partial edit

Table 27: Summary of the twelve illustrative image-level AU interventions. Each source image is originally predicted as Happiness. After editing the target AU set, CausalAffect predicts the intended counterfactual expression.

Sample	Source Prediction	Edited Target AUs	Target Expression	Target Pred.
Sample 1	Happiness, 98%	AU1+AU4+AU7+AU25	Anger	93%
Sample 1	Happiness, 98%	AU1+AU4+AU15+AU17+AU25	Sadness	89%
Sample 1	Happiness, 98%	AU1+AU2+AU5+AU25+AU26	Surprise	90%
Sample 2	Happiness, 97%	AU1+AU4	Sadness	85%
Sample 2	Happiness, 97%	AU1+AU4+AU17+AU25	Disgust	82%
Sample 2	Happiness, 97%	AU1+AU2+AU5+AU25+AU26	Surprise	94%
Sample 3	Happiness, 92%	AU4+AU7+AU9+AU10+AU25	Anger	81%
Sample 3	Happiness, 92%	AU4+AU24	Sadness	94%
Sample 3	Happiness, 92%	AU1+AU2+AU5+AU25+AU26	Surprise	95%
Sample 4	Happiness, 88%	AU4+AU7+AU17+AU24	Disgust	73%
Sample 4	Happiness, 88%	AU5+AU25	Surprise	79%
Sample 4	Happiness, 88%	AU1+AU4	Sadness	87%

produces a weaker prediction. Thus, the response is not merely caused by arbitrary visual distortion; it reflects the compositional completeness of the target AU configuration.

Third, anger and disgust edits also induce the intended predictions, with lower average confidence than the full surprise edits among the examples shown. For anger, Sample 1 edited with AU1+AU4+AU7+AU25 is classified as anger with 93% confidence, whereas Sample 3 edited with AU4+AU7+AU9+AU10+AU25 reaches 81%. For disgust, Sample 2 and Sample 4 reach 82% and 73%, respectively. These lower values are expected because anger and disgust often share overlapping facial actions with other negative expressions, making them more ambiguous under image-level editing. The fact that CausalAffect still changes the prediction to the intended class supports the intervention consistency of the learned AU→Expression relations.

N.3 Graph-level reconfiguration under image edits

We further test whether the sample-adaptive graph follows the same edit-dependent crossover in the 36-source study. Let $\Delta e(R)$ denote the edited-minus-source change in the polarity-aligned strength of route set R. The graph-level interaction is

$$I_e = [\Delta_T e(R_T) - \Delta_T e(R_C)] - [\Delta_C e(R_T) - \Delta_C e(R_C)]. \tag{32}$$

Table 28 shows that Target edits produce a larger change in their corresponding routes than in the alternative routes (0.24 versus 0.05). Under Control edits, the pattern reverses: the corresponding R_C routes change more strongly than R_T (0.21 versus 0.04). The corresponding route set is dominant for 31/36 Target edits and 29/36 Control edits, producing a graph-level crossover interaction of $I_e = 0.36$. The graph response thus follows the distinct AU configuration introduced by each edit.

Reconstruction serves as a generator-only control without an AU intervention. Its mean absolute route drift is 0.018 (95% CI: $[0.014, 0.023]$), compared with corresponding-route changes of 0.24 under Target and 0.21 under Control. The Target–Reconstruction and Control–Reconstruction differences are 0.23 and 0.21, with both Holm-adjusted paired permutation tests giving $p < 0.001$.

The illustrative examples provide a complementary view of these graph responses.

Beyond prediction changes, the edited images also induce meaningful changes in the sample-adaptive graphs. This is important because a purely correlation-driven classifier could change its output after an image edit without producing an interpretable relational explanation. CausalAffect, in contrast, adapts both its AU→Expression graph and AU→AU graph according to the edited AU evidence.

For the original happiness images, the sample-adaptive AU→Expression graph assigns strong positive influence to happiness-related AUs such as AU6, AU7, AU12, and AU25, while suppressing AUs that are inconsistent with happiness. After intervention, this structure is no longer preserved blindly.

Table 28: Graph-level edit–route alignment. Δe is the polarity-aligned route-strength change from the source. Brackets denote source-level bootstrap 95% CIs.

Condition	Target-route Δe	Control-route Δe	Aligned margin	Corresponding route dominant
Target	**+0.24** [0.19, 0.29]	+0.05 [0.02, 0.08]	+0.19 [0.13, 0.25]	31/36 (86.1%) [71.3, 93.9]
Control	+0.04 [0.01, 0.07]	**+0.21** [0.17, 0.25]	+0.17 [0.11, 0.22]	29/36 (80.6%) [65.0, 90.2]
Reconstruction	+0.01 [−0.002, 0.022]	0.00 [−0.011, 0.012]	—	—

Instead, the graph reorients toward the target expression. For example, after surprise edits, the AU→Expression graph highlights the edited surprise-related AUs, including AU1, AU2, AU5, AU25, and AU26. Similarly, after sadness edits, the graph increases the contribution of AU1 and AU4, and when additional sadness-related AUs such as AU15 and AU17 are present, these AUs are also incorporated into the target-specific causal explanation. For anger and disgust edits, the graph emphasizes AUs associated with brow tension, lid tightening, nose or upper-lip actions, and lower-face tension, depending on the edited configuration.

The AU→AU graphs provide a complementary view. After editing the image, CausalAffect does not only detect the target AUs independently; it also constructs a sample-specific interaction pattern among the edited AUs. For surprise edits, the model tends to form coherent dependencies among upper-face raising AUs and lower-face opening AUs, such as AU5, AU25, and AU26. For sadness edits, AU1 and AU4 become central to the inferred AU structure. For anger and disgust edits, the graph shifts toward tension-related interactions involving AUs such as AU4, AU7, AU9, AU10, AU17, AU24, and AU25. These graph changes show that the model recomputes its relational explanation for the edited facial configuration.

N.4 Evidence for signed and inhibitory reasoning

The image-level interventions also provide evidence that signed relations are useful for expression disambiguation. When a source happiness image is edited toward a negative expression, the graph not only activates target-compatible AUs, but also suppresses AUs that are incompatible with the target expression. For instance, sadness and disgust edits reduce the explanatory role of happiness-related AUs such as AU6 and AU12. Conversely, in the original happiness images, negative-expression AUs such as AU1, AU4, AU15, AU17, AU20, or AU24 tend to receive low or inhibitory contributions. This behavior is consistent with the signed AU→Expression graph: expression inference depends not only on what facial actions are present, but also on which incompatible actions are absent or suppressed.

This observation is particularly important for distinguishing CausalAffect from unsigned attention or co-occurrence-based AU models. A non-negative attention model can assign importance to visible AUs, but it cannot explicitly represent that the presence of one AU should reduce the plausibility of a competing expression. The image-level edits show that CausalAffect uses both excitatory and inhibitory evidence when adapting to the edited facial configuration.

N.5 Relation to feature-level counterfactual intervention

This image-level study complements the feature-level counterfactual objective used during training. The feature-level module uses graph-derived masks to intervene on AU representations. Interventions on graph-selected AUs should change the target representation or prediction, whereas interventions on unselected AUs should leave it stable. The image-level experiment tests the same principle in a more observable setting. When the visual appearance of target AUs is edited, the prediction changes toward the corresponding expression, and the sample-adaptive graph reorganizes around the edited AU evidence.

Importantly, these image-level results help distinguish the proposed counterfactual mechanism from ordinary noise injection or feature dropout. A generic noise regularizer would only encourage

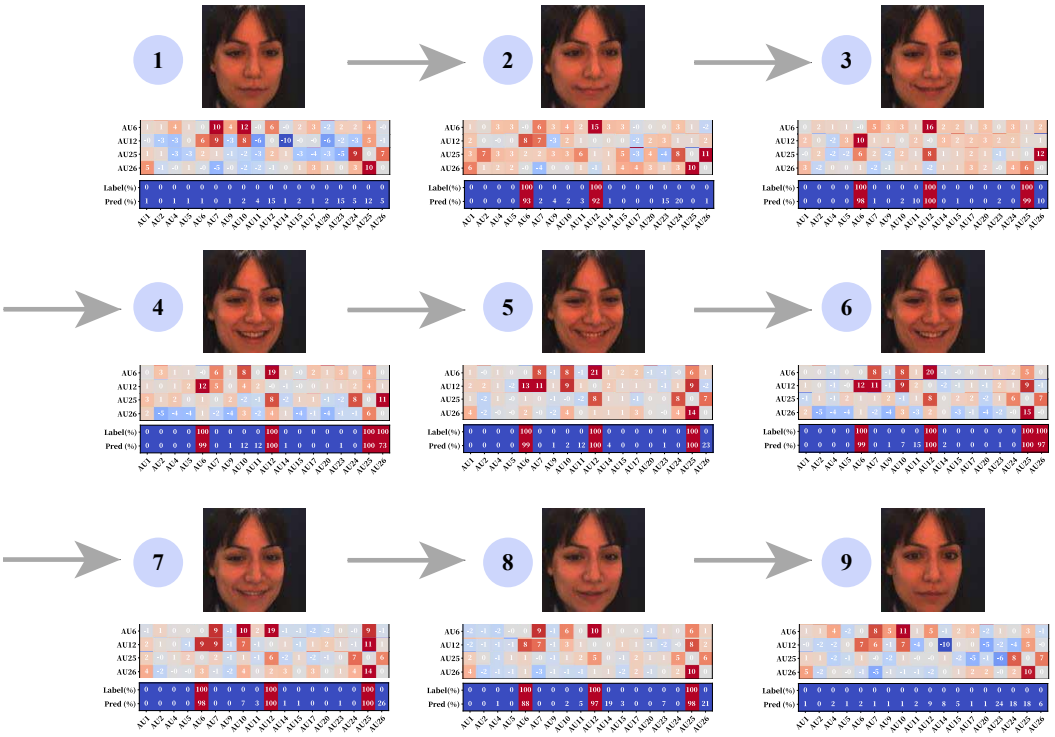

Figure 10: Sample-adaptive AU→AU causal graphs for nine images sampled at five-frame intervals from a smile sequence. Rows represent target AUs and columns represent source AUs. Warm and cool colors denote positive and negative weights. The sequence covers smile onset, apex, and offset. Cross-frame comparisons use the displayed weights.

robustness to perturbations and would not predict which visible AU edits should change the expression, nor how the graph should reconfigure after the edit. CausalAffect uses the learned causal graph to select source AUs for each target, and the image-level edits show that the same AU-level structure remains behaviourally meaningful under visible facial changes.

O Qualitative analysis from a video and temporal perspective

To understand how AUs interact throughout the evolution of an expression, we examine the AU→AU causal graphs extracted directly from a continuous video sequence (Figure 10). From the full video, we sample one frame every 5 frames to obtain nine representative snapshots that jointly cover the onset, apex, and offset phases of a smile. For each selected frame, the model infers a sample-adaptive AU→AU causal graph.

At **smile onset**, movement around the eyes and mouth is subtle, and the weights into AU6 are distributed across several source AUs. As the smile develops, AU12 → AU6 rises from $+6$ in Frame 1 to $+16$ in Frame 3 and $+21$ in Frame 5. By Frame 5, AU12 is the strongest source for AU6, overtaking AU10, which has the strongest incoming weight in Frame 1 ($+12$). This change in the dominant source links the shift in the learned graph to the increasingly coordinated motion of the lip corners and cheeks as the smile develops.

Near the **smile apex**, the stills show tighter eye corners, raised lip corners, and a more open mouth. In Frame 6, AU12 → AU6 has a weight of $+20$ and the reverse edge has a weight of $+12$. The AU25 → AU26 edge has a weight of $+15$, with $+7$ in the reverse direction. The two pairs illustrate reciprocal coordination in the learned graph, linking cheek raising with lip corner pulling and lip parting with jaw opening. The positive AU12 → AU25 edge ($+8$) further connects lip corner movement to mouth opening at this stage.

During the **offset phase**, the cheeks relax and the mouth moves toward its resting posture. The AU12 $\rightarrow$ AU6 weight falls to $+19$, $+10$, and $+5$ in Frames 7–9. AU12 $\rightarrow$ AU25 also weakens, from $+8$ in Frame 6 to zero at the displayed precision in Frame 9. The lip-parting to jaw-opening edge weakens more gradually and retains a weight of $+10$ in Frame 9. The inferred graph therefore captures different rates of change across the facial actions as the smile fades.

Tracking the same directed edges across frames links changes in the learned relations to the visible progression of the smile. The different trajectories of cheek–lip and lip–jaw coordination illustrate how the sample-adaptive graph can distinguish which parts of a facial configuration change over the course of a single expression.

P Failure analysis

We investigated expression-recognition errors through two complementary analyses: an examination of misclassified images and a controlled study of the AU vocabulary used for training. These address different aspects of the evidence available to CausalAffect: the visual AU cues in an individual image and the AU relations that can be learned from the supplied supervision.

For the image analysis, we examined all misclassified samples from the complete AffectNet evaluation set using CausalAffect (+All). Some errors occur when cropping or occlusion hides facial regions containing relevant AUs, or when blur and low resolution make subtle AU activations difficult to discern. The model then receives incomplete visual evidence for AU$\rightarrow$Expression inference and can predict the wrong expression category. Subtle or compound expressions can also provide similar evidence for multiple categories, while a smaller subset of errors involves incorrect AU estimation or relation assignment. These cases show how missing or ambiguous AU cues affect expression judgement even when the full AU vocabulary is available.

To examine the role of training-time AU coverage, we used the experiments in Appendix E, which keep BP4D fixed as the auxiliary dataset and vary the number and frequency composition of its supervised AUs. Expression recognition is evaluated on AffectNet and RAF-DB across these configurations. The full vocabulary achieves the highest accuracy on both datasets; among equally sized subsets, frequent AUs provide better results than infrequent AUs. Restricting the vocabulary excludes AU$\rightarrow$AU and AU$\rightarrow$Expression pathways, while sparse observations provide less evidence for learning the remaining causal relations. These comparisons support broader AU coverage for expression composition and show why the composition of auxiliary supervision matters.

Q Limitations and ethics

Limitations. CausalAffect learns directed and signed causal relation hypotheses through its model architecture and supervision from facial affect benchmarks. This design enables interpretable modeling of AU$\rightarrow$AU coordination and AU$\rightarrow$Expression composition, while relying on a predefined AU vocabulary to maintain comparability with FACS-based psychological knowledge. As a result, the current relational interface mainly covers the selected AU set. Extending it to finer-grained facial primitives, micro-expressions, and open-vocabulary facial actions would broaden the coverage of the framework and support more flexible facial behavior modeling.

Ethics. CausalAffect is intended for consent-based research on interpretable AU and expression analysis, including controlled human–computer interaction and behavioural research. We do not support the use of CausalAffect for covert surveillance or tracking, identity or demographic profiling, deception detection, inference of internal mental states, or high-stakes decisions about individuals, nor as a standalone psychological or clinical assessment system. The fairness analysis provides a dataset-bounded diagnostic, not a comprehensive fairness guarantee across cultural settings, real-world lighting and pose conditions, or unmeasured demographic axes such as age and gender. Broader deployment requires dedicated audits with representative, intersectionally annotated data from the intended operating conditions, together with informed consent, privacy protection, context-aware interpretation, and human oversight.

R Related work

R.1 Modeling structured relations between AUs and expressions

Learning the structured relationships among AUs and between AUs and expressions is crucial for improving facial affect recognition. Psychological and cognitive science studies consistently show that facial expressions emerge through coordinated AU activations [15, 35], and AU→AU as well as AU→Expression dependencies provide strong priors for enhancing both AU detection and expression recognition [14, 79, 34, 59]. A model that accurately captures such dependencies is better positioned to reason about facial behavior, reduce ambiguity among visually similar patterns [12, 40], and provide explanations aligned with human expert interpretation.

Expert-Curated Relations: FACS [15] provides the foundation for describing these relations, and many early approaches incorporate cognitive or psychological priors. Prior-based methods rely on manually defined AU compositions or expert-curated AU groups [15, 14], ensuring a degree of interpretability by enforcing human-understandable structures. However, such approaches assume fixed population-level mappings and cannot adapt to **individual differences** (e.g., idiosyncratic AU combinations or varying AU intensities) as demonstrated in psychological studies of AU variability [17]. Furthermore, hand-crafted rules cannot represent **inhibitory** effects or **asymmetric** AU relations frequently observed in natural facial behavior [69, 34].

Data-Driven Modeling Relations: To address the limitations of fixed psychological priors, recent work employs graph neural networks and attention-based architectures to *learn* AU→AU and AU→Expression relations directly from data [76, 85, 56, 79, 90]. These approaches introduce powerful relational inductive biases and often yield substantial gains in recognition performance. However, models trained primarily for discriminative accuracy typically learn undirected AU affinity graphs rather than psychologically grounded causal relations. As a consequence, the learned graphs tend to be **dense**, highly **entangled**, and **difficult to interpret** [37, 83], lacking the directionality, asymmetry, and polarity-aware dependencies that human experts routinely rely on when decoding facial behavior.

Data-Driven Co-occurrence Relations: Another class of works derives AU or expression dependencies from statistical co-occurrence in jointly annotated datasets [41, 43]. While co-occurrence patterns provide useful weak structural cues, they are known to encode dataset-specific biases, demographic imbalances, and annotation noise [8, 13]. Moreover, because these methods require AU and expression labels to appear on the same images, they cannot take advantage of the many AU-only or expression-only datasets widely used in practice [90, 43]. Finally, frequency-based relations quantify **shared occurrence rates**, which cannot distinguish **excitatory** from **inhibitory** mechanisms.

Overall, these existing approaches struggle to provide **human-aligned interpretability**: they cannot explain ***which*** AU activations lead to a particular expression prediction, ***how*** AUs influence one another, or ***why*** a model arrives at certain decisions. This motivates the need for a causal modeling framework that behaves more like a trained expert— one that learns directed, asymmetric, and polarity-aware relations, generalizes across heterogeneous datasets, and offers coherent, human-aligned explanations for model predictions.

R.2 Causal and counterfactual modeling in facial behavior

Causal reasoning and counterfactual interventions have recently been explored in computer vision, often in the context of generative modeling for fairness, bias mitigation, or semantic editing [9, 7, 52, 66]. In facial affect analysis, several works leverage counterfactuals to debias or interpret expression classifiers, for example by generating counterfactual faces to study fairness in facial expression recognition [9, 7], or by introducing counterfactual attention mechanisms and discriminative counterfactual sequences for macro- and micro-expression recognition [52]. Beyond affect analysis, a broader line of research studies counterfactual image editing and counterfactual image generation using deep generative models, where interventions are performed in pixel space or in a disentangled latent space of semantic attributes [66, 63, 67]. Related work on causal representation learning for faces incorporates causal effects of pose or expression into generative pipelines, with a focus on controllable synthesis [21].

These causal and counterfactual approaches demonstrate the benefits of intervention-based reasoning for bias mitigation, robustness, and explanation, but they do not address the fine-grained FACS structure that is central to AU-based analysis. In generative approaches, these interventions typically

operate on **pixels** or **coarse semantic attributes** (e.g., gender or race). Moreover, most of these methods assume a **fixed, global causal structure** tied to a single dataset or attribute set, and do not provide sample-adaptive refinement of the relational structure.

Our framework differs from prior work in three key aspects: (1) it learns asymmetric, polarity-aware AU→AU and AU→Expression causal relations through *causally guided relation learning*; (2) it supports *weakly supervised multi-dataset learning* without joint annotation; (3) it introduces efficient, feature-level *counterfactual interventions* in the disentangled AU latent space.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's main contributions, supported by the method formulation and experimental results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses limitations and ethics in appendix.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The appendix provides dataset construction, metrics, model architectures, training protocol, and hyperparameters needed to reproduce the main experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper provides detailed experimental and implementation information for reproducibility.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The appendix provides dataset, metrics, architectures, losses, hyperparameters, thresholding rules, and baseline protocols.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports means and standard deviations over repeated runs for key benchmark results and matched baseline comparisons, together with statistical significance tests and confidence intervals where applicable.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses benchmark datasets and standard model training procedures, and the paper discusses fairness, privacy, and deployment considerations.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper discusses potential limitations and deployment impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper cites the original datasets and pretrained models used in the experiments, and the experiments follow their original access terms and usage conditions.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We conducted a voluntary expert validation study with five domain experts. They judged AU→AU and AU→Expression relations using True/False labels and 1–5 confidence ratings. Details are provided in Appendix K; responses were anonymized and reported only in aggregate.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not conduct new human-subject experiments or collect new participant data.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.