AFFECTIVE CAUSAL GRAPH LEARNER FOR HUMAN-ALIGNED FACIAL BEHAVIOR UNDERSTANDING

Anonymous authors
Paper under double-blind review

A 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 1). The goal of this experiment is to disentangle the roles of Global Causal Graph (GC), Sample-Adaptive Causal Graph (SAC), Counterfactual (CF), AU Disentanglement (Dis), and the Directed Acyclic Graph (DAG) constraint. Performance is reported across six benchmarks (AffectNet, RAF-DB, DISFA, BP4D, GFT, EmotioNet).

Idx	Model Variant	w/o	AffectNet	RAF-DB	DISFA	BP4D	GFT	EmotioNet
1	Backbone	GC + SAC + CF	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**

Table 1: Ablation Study on CausalAffect (+All Setting), exploring the effect of Global Causal Graph (GC), Sample-Adaptive Causal Graph (SAC), Counterfactual Intervention (CF), and AU Disentanglement (Dis). Best results are highlighted.

Global vs. Sample-Adaptive Graphs: When trained individually, the Global Graph consistently 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: when combined with GC (Rows 11-13), the performance improves further, confirming that personalized inference adds value by tailoring causal reasoning to individual instances.

Effect of AU Disentanglement Removing AU Disentanglement leads to noticeable performance drops (e.g., Row 1 vs. Row 11). 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 13, CF becomes more effective when paired with disentangled features, suppressing irrelevant cues and amplifying informative dependencies.

Role of Counterfactual Intervention Counterfactual interventions show benefits only when supported by disentangled AU features. In the absence of AU Dis (Rows 5, 9 and 12 vs. Rows 3, 7, and 11), CF may even introduce misleading signals, as it lacks structural guidance to filter spurious correlations. However, when AU Dis is enabled, CF provides strong gains by forcing 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.

Human-Aligned Causal Structure Rows 6, 10 and 13 illustrate that both CF and AU Dis are essential for human-aligned causal reasoning. Removing either component leads to reduced performance and degraded interpretability. Only when both are present can CausalAffect capture reliable, semantically grounded AU→Expression relations. This validates the psychological plausibility of our learned causal graphs, bridging low-level facial activations and high-level emotion inference.

Effect of DAG Constraint Finally, Row 14 highlights the necessity of the DAG constraint. Without it, the learned graph may contain redundant or semantically implausible loops due to unconstrained topology. Incorporating DAG (Row 15) introduces a soft acyclicity bias that enforces sparse, directional, and interpretable causal pathways. As a result, irrelevant connections are pruned, semantic clarity is enhanced, and overall performance reaches its peak across all benchmarks.

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 further prunes spurious associations but only becomes effective when disentangled representations are available. Finally, the DAG constraint ensures structural clarity and interpretability, yielding the strongest overall performance. Together, these components allow CausalAffect to achieve not only higher accuracy but also more human-aligned and semantically coherent causal graphs.

B 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 2, covering both AU detection (EmotioNet, GFT) and expression recognition (RAF-DB, AffectNet).

#	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
5	CausalAffect (+BP4D, 8 AUs, 8 Least Frequent AUs)	63.1	58.3	81.7	64.5
6	CausalAffect (+BP4D, 12 AUs)	65.4	60.4	**85.3**	**67.7**

Table 2: Effect of AU Set Size on AU Detection and Expression Recognition. Performance is reported on EmotioNet/GFT for AU detection and RAF-DB/AffectNet for expression recognition. Best results for each task are highlighted.

From the 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. In contrast, incorporating rare AUs (Rows 4 and 5, least frequent) significantly degrades performance, since their sparse activations fail to provide stable co-occurrence cues and instead inject noise into causal inference, leading to fragmented and unstable structures. Interestingly, the 12-AU configuration (Row 6) 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.

C 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 statistics approach.

C.1 AU-Expression Global Causal Relation

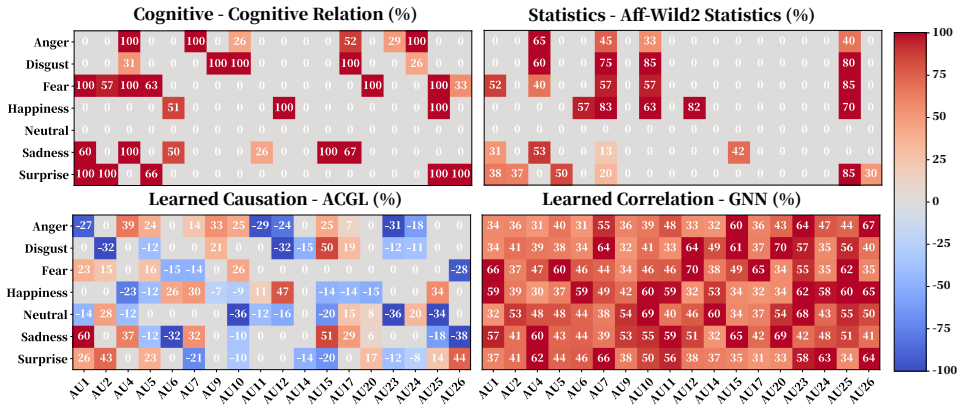

Figure 1: Comparison of AU-Expr Relations across cognitive priors, statistical co-occurrence, GNN-learned correlation(Learned on All-DB), and CausalAffect-learned causation (Learned on All-DB) (%).

Alignment with human priors and expression semantics. The AU→Expression causal graph (Figure 1) learned by **CausalAffect** reveals both **causally grounded** and **psychologically plausible** structures. The discovered relations align closely with established findings in facial behavior researchEkman & Friesen (1978), as well as with prior **statistical co-occurrence relations**Kollias et al. (2024) and **cognitive models**Du et al. (2014). 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, reflecting widened eyes and involuntary jaw relaxation respectively—together forming a classic 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 becomes a *causal prerequisite* for a given expression to be inferred. For example, AU6 ⊣ Sadness and AU26 ⊣ Sadness indicate that the non-activation of smile-related or surprise-related AUs is a necessary condition for confidently inferring Sadness. AU6 (cheek raiser) is a hallmark of joy and amusement, while AU26 (jaw drop) is prominently associated with Surprise. Their presence would contradict the subdued, upper-face tension and downward lip dynamics that define Sadness, including AUs such as AU1, AU4, and AU15. These inhibitory causal links reflect the principle of inhibitory precondition: Sadness becomes a plausible interpretation not only due to the presence of its prototypical AUs, but also because **affectively incompatible** actions like AU6 and AU26 are absent. Such negative causal links highlight the model's ability to reason not only about what *must be present*, but also about what *must be 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. Given the known difficulty in annotating low-intensity or ambiguous AUs, such patterns likely reflect **systematic label noise** rather than genuine neutrality. In this context, the causal graph functions as a **structural prior** that can support **label auditing**, **confidence calibration**, and improved **annotations 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.

C.2 AU-AU Global Causal Relation

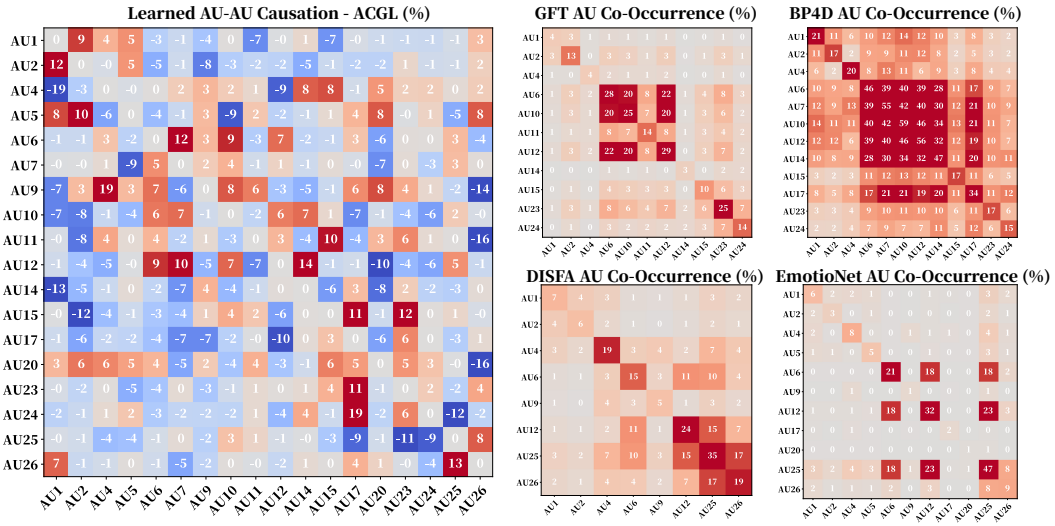

Figure 2: Comparison between the AU-AU causal relations learned by CausalAffect (trained with the +ALL setting) and AU co-occurrence statistics from four datasets (GFT, DISFA, EmotioNet, BP4D).

To our best 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, our CausalAffect framework learns a unified causal graph that covers the complete set of 18 AUs (see left of Figure. 2), without relying on dataset-specific label availability. **(ii) Symmetry assumption:** Co-occurrence measures are inherently symmetric by definition, 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:** Co-occurrence methods can only indicate excitatory association 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 1), which exhibit strong influence patterns, the AU→AU relations tend to have lower overall magnitudes, with the maximum absolute value reaching approximately 19. This observation is consistent with the underlying nature of facial behavior: AU-AU interactions primarily capture low-level muscular coordination or antagonism (e.g., AU6 and AU7 co-activating around the orbital region), rather than reflecting high-level semantic or emotional constructs. 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 knowledges or handcraft prior, 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 largely missing from existing dependency modeling methods, which typically rely on symmetric or co-occurrence-based statistics and thus fail to capture negative interactions. Such inhibitory relations are especially important in AU detection, where many AUs are known to be semantically incompatible. For example, CausalAffect learns a strong negative influence AU4 ⊣ AU1 , reflecting the **physiological antagonism** between brow lowering (corrugator supercilii) and inner brow raising (frontalis). Similarly, AU26 ⊣ AU20 captures the incompatibility between horizontal lip stretching and vertical jaw dropping. The relation AU26 ⊣ AU11 also illustrates this, as AU11 contributes to nasolabial deepening during expressions of effort or sneering, which typically opposes the open-jaw posture characterized by AU26. 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.

Uncovering semantically important but statistically inaccessible relations. Beyond aligning with known co-occurrence trends, CausalAffect also discovers several high-impact causal relations that are statistically inaccessible in existing datasets due to non-overlapping AU annotations. For example, the learned relation AU15 → AU20 captures the dynamic transition from lip corner depression (sadness) to horizontal lip stretch (tension or discomfort), which is rarely annotated together in existing datasets but plays an important role in modeling affective states. In addition, CausalAffect captures several causal dependencies that are **missing from co-occurrence statistics** entirely, yet

are highly meaningful for facial interpretation. For instance, AU4 → AU9 indicates a strong link between brow lowering (anger/focus) and nose wrinkling (disgust), frequently observed in complex expressions such as contempt or intense concentration. The relation AU15 → AU11 reflects a plausible lower-face interaction where lip corner depression activates muscular pathways contributing to nasolabial fold deepening. Similarly, AU17 → AU24 links chin raising with lip pressing—both associated with suppressive, high-tension affective states such as fear or frustration. These examples highlight CausalAffect's ability to infer biologically and semantically grounded causal interactions beyond what is available in co-occurrence statistics, offering richer structural priors that are critical for robust and generalizable AU-based facial analysis.

Effect of Directed Acyclic Graph (DAG) Constraint: To guide the model toward learning interpretable structures, we incorporate a Directed Acyclic Graph (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. Interestingly, we observe that the learned causal graph still contains localized cycles—for example, AU1 → AU2 with a weight of +12 and AU2 → AU1 with +9. While this appears to violate the DAG constraint, it reflects an important characteristic of facial dynamics. In psychological and behavioral literatureEkman & Friesen (1978), 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. Therefore, our implementation adopts the DAG constraint as a *soft regularization* rather than a hard constraint. 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.

D 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.

D.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, 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: AU25, AU26, and AU5. 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)**, sample-adaptive graph prioritizes AU25, AU4, and AU1, which are visually salient in the image, but globally less emphasized compared to AU10 and AU5—both of which are inactive and thus downweighted. 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.

Instance-level causal graphs enable principled diagnosis of label noise. A key strength of **CausalAffect** lies in its ability to detect annotation inconsistencies through fine-grained, instance-

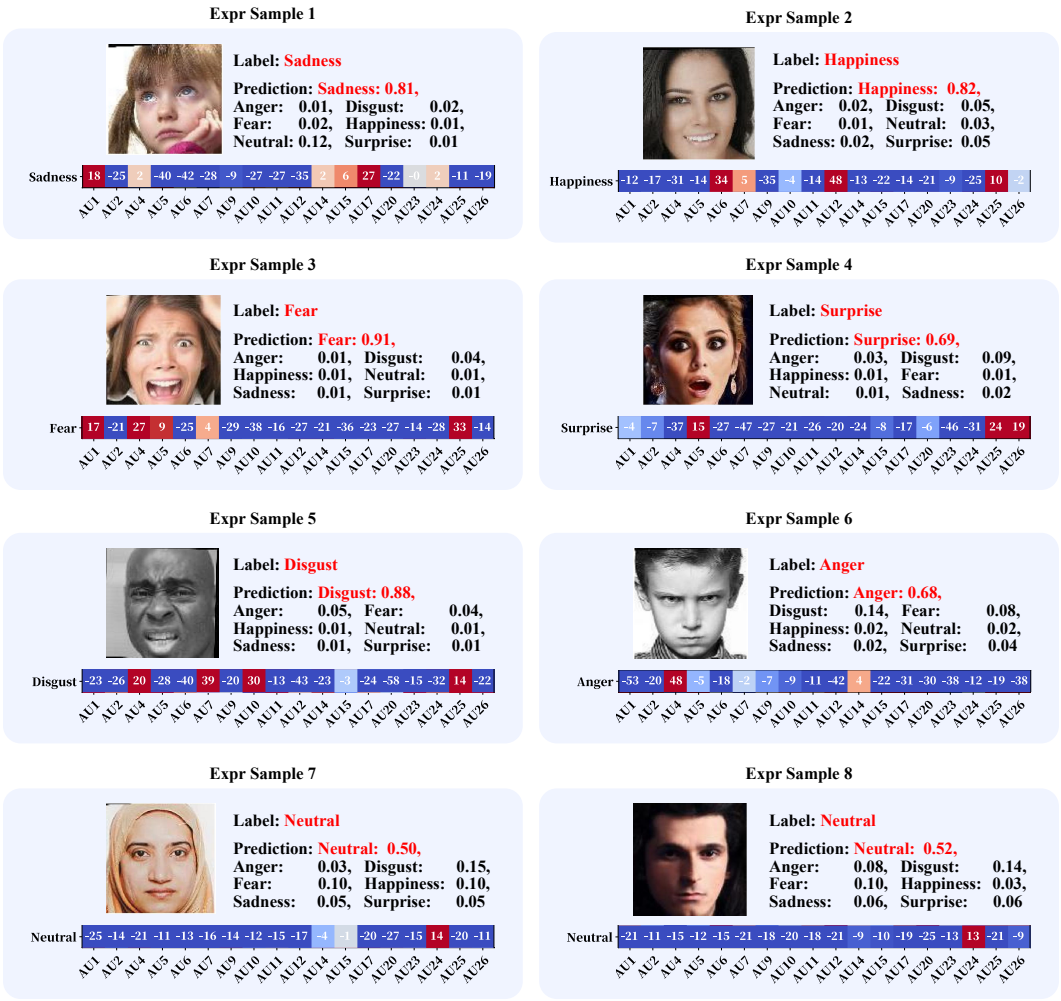

Figure 3: Sample-adaptive AU→Expression causal relations dynamically inferred for individual samples. We random sample one image per basic expression (and two for Neutral) to illustrate how CausalAffect captures instance-specific causal structures. For clarity, only the subgraph corresponding to the predicted expression is visualized.

specific causal reasoning. This capability becomes particularly evident in **Samples 7 and 8 (Neutral)**, both of which display visually static, expressionless faces with minimal muscular activity. Despite the Neutral label, CausalAffect assigns weak positive causal weights to AU24 (lip pressor: $+14$, $+13$), while suppressing all other AUs—including AU2 (outer brow raiser, -35) and AU17 (chin raiser, -29). This discrepancy highlights a critical semantic inconsistency: *by definition*, Neutral should not be causally supported by any strong AU activation. Even the attribution of AU24 as a weak positive contributor appears questionable, particularly given the lack of visible lip pressing in the corresponding images. As a low-saliency AU prone to misinterpretation—often confounded with pre-speech tension or relaxed mouth posture—AU24 is highly susceptible to annotation noise. The fact that CausalAffect assigns marginal support to AU24, while confidently suppressing all other AUs, suggests that such annotations may reflect **systematic over-labeling** rather than true muscular expression. In this sense, the sample-adaptive causal graph provides a valuable structural prior for **label auditing, quality assessment, and data refinement**, enabling the model not only to learn from data but to question its reliability.

D.2 AU-AU Sample-Adaptive Causal Relation

Unlike dataset-level co-occurrence, CausalAffect learns fine-grained, sample-specific AU→AU causal graphs. For instance, in **Sample 1**, AU1 is causally inferred from AU2, AU5, AU26, and AU25, all of which are visibly present in the image. This structure par-

tially aligns with the global AU→AU graph (e.g., AU5 → AU1, AU26 → AU1), yet reflects a clearer subject-specific adjustment. Similarly, AU25 → AU26 recovers canonical part-to-drop progression, demonstrating consistency with global priors.

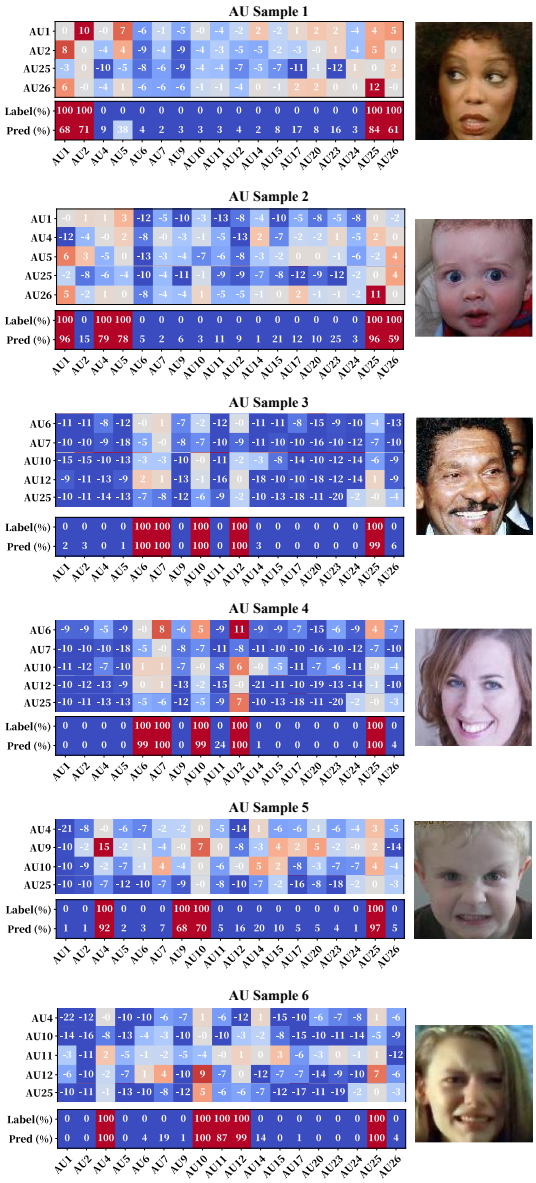

Figure 4: Sample-adaptive AU→AU causal relations dynamically inferred for individual samples. We randomly select 6 images from the EmotioNet dataset to illustrate how CausalAffect captures instance-specific inter-AU causal structures. For clarity, only the subgraphs involving the predicted activated AUs are visualized.

However, in this same example, AU25 is not inferred through direct excitatory causes but via exclusion-based reasoning: it receives strong *negative* causal input from AU4, AU17, and AU15, suggesting that the absence of these mutually exclusive AUs makes AU25 the most plausible outcome. A similar dynamic is observed in **Sample 5**, where AU25 is again inferred not from direct support but from the strong inhibition of AU20—a horizontal lip stretcher biomechanically incompatible with vertical lip separation. These examples illustrate CausalAffect's ability to infer causal dependencies not only through direct excitation but also through *causal exclusion*, wherein the deactivation of specific AUs increases the inferred necessity of others. Additionally, Sample 5 shows that AU9 (nose wrinkler) is causally activated by AU4 and AU10, yet AU4 is itself inhibited by AU10, suggesting muscular antagonism in the upper face. This internal tension reflects CausalAffect's capacity to capture not just co-activation, but also expressive conflict within fine-grained causal paths.

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 sparse and directional structures. 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 emotionally driven override of suppressive muscular input. 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. **Samples 3 and 4** both contain AU6, AU7, AU10, AU12, and AU25, yet their AU→AU graphs diverge sharply. 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 suggests that the system interprets these AUs as **emergent, not volitional—consistent with strained, socially modulated, or ambiguous affect Surakka & Hietanen (1998).** 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 configuration reflects a spontaneous and emotionally consistent smile, where lower-face AUs reinforce each other. 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.

E HSIC COMPUTATION.

We use the empirical HSIC estimator Gretton et al. (2005) 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 RBF (Gaussian) kernels:

$$K_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma_x^2}\right), \quad L_{ij} = \exp\left(-\frac{\|y_i - y_j\|^2}{2\sigma_y^2}\right), \tag{1}$$

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

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

where $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ is the centering matrix that removes the mean from kernel features. This formulation enables unbiased estimation of the squared Hilbert-Schmidt norm of the cross-covariance operator between X and Y in their respective RKHSs. 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.

F EXPERIMENTS

F.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 Shao et al. (2021); Kollias et al. (2024) for consistent AU dataset splitting and evaluation. *BP4D*Zhang et al. (2014) contains 41 subjects with spontaneous expressions annotated over 12 AUs. *DISFA*Mavadati et al. (2013) includes posed and spontaneous videos with 8 selected AUs. *EmotioNet*Fabian Benitez-Quiroz et al. (2016) 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*Girard et al. (2017) is a proprietary dataset containing over 130K high-quality face images annotated with 10 AUs. 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*Li et al. (2017) contains around 15K images with crowd-sourced annotations of 7 basic expressions. *AffectNet*Mollahosseini et al. (2017) 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.

F.2 Implementation Details

The backbone is ConvNeXt-Base pretrained on WebFaceYi et al. (2014). 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 apply a soft DAG loss with $\lambda_{\text{DAG}} = 0.5$. Sample-adaptive graphs are computed via a 2-layer attention. AU→Expr graphs use learned expression prototypes(dim 64) as attention queries. Counterfactual intervention uses $\epsilon \sim \mathcal{N}(0, 0.2^2)$ with gate sharpness $\gamma = 10$ and loss weights $\lambda_{\text{consist}} = \lambda_{\text{discrep}} = 0.5$, $\delta_{\text{feat}} = \delta_{\text{logit}} = \eta_{\text{feat}} = \eta_{\text{logit}} = 1.0$. We train for 80 epochs using AdamW (lr=$1\text{e}{-4}$, wd=$1\text{e}{-5}$), batch size 360

F.3 Sensitivity Analysis of Regularization Parameters

In addition to evaluating the overall effectiveness of CausalAffect, we further investigate the sensitivity of its key regularization parameters, focusing on the mask sharpness parameter γ and the feature/logit-level causal regularization weights. These parameters are not extensively tuned for each dataset but instead are guided by theoretical considerations to ensure robustness and interpretability across heterogeneous settings.

Fixed Robustness Across Settings A key design choice of CausalAffect is to maintain a fixed set of hyperparameters across all datasets and experiments, including multi-dataset training and ablation studies. This consistent configuration ensures robustness and transferability, demonstrating that the model architecture and loss formulation are inherently stable. As reported in main paper Table 1 and Ablation Study Table 1, CausalAffect achieves strong performance without dataset-specific tuning, highlighting its reproducibility and practical usability in real-world scenarios.

CausalAffect (RAF-DB + BP4D)	RAF-DB	BP4D
$\gamma = 20$	84.1	66.7
$\gamma = 10$	**85.0**	**67.1**
$\gamma = 5$	84.3	66.2
$\delta_{\text{feat}} = \delta_{\text{logit}} = \eta_{\text{feat}} = \eta_{\text{logit}} = 1$	**85.0**	**67.1**
$\eta_{\text{feat}} = \delta_{\text{feat}} = 0.5$, $\eta_{\text{logit}} = \delta_{\text{logit}} = 1$	83.2	66.3
$\eta_{\text{feat}} = \delta_{\text{feat}} = 1$, $\eta_{\text{logit}} = \delta_{\text{logit}} = 0.5$	84.5	66.6
$\eta_{\text{feat}} = \delta_{\text{feat}} = \eta_{\text{logit}} = \delta_{\text{logit}} = 0.5$	83.0	66.1

Table 3: Sensitivity analysis of mask sharpness (γ) and causal regularization weights in CausalAffect on RAF-DB and BP4D datasets. Best results are highlighted.

Effect of Mask Sharpness γ The parameter γ controls the steepness of the sigmoid gating mask for counterfactual intervention, thereby determining the sharpness of feature selection. Setting $\gamma = 10$ achieves the best balance, producing a binary-like mask that reliably activates strong edges above the 0.5 threshold while suppressing weak connections. Increasing γ to 20 overly sharpens the mask, reducing generalization and hindering causal learning, while decreasing γ to 5 produces overly soft masks that blur causal boundaries. As shown in Table 3, $\gamma = 10$ consistently yields the strongest results on both RAF-DB and BP4D datasets.

Effect of Feature and Logit Regularization We also examine the relative importance of feature-level ($\delta_{\text{feat}}, \eta_{\text{feat}}$) and logit-level ($\delta_{\text{logit}}, \eta_{\text{logit}}$) causal regularization weights. Equal weighting (all set to 1.0) achieves the best trade-off, balancing causal separability at the feature level and stability at the prediction level. Reducing feature-level weights while keeping logit-level weights high leads to the largest performance degradation, indicating that feature-level constraints are more critical for reliable causal inference. Conversely, reducing logit-level weights produces a smaller drop in performance, but still highlights their relevance. Setting all weights to 0.5 yields the lowest results, confirming that strong and balanced regularization is essential.

Overall, these results highlight that CausalAffect is not overly sensitive to fine-grained hyperparameter tuning. Instead, it benefits from theoretically motivated regularization choices that ensure stability,

interpretability, and transferability across datasets. The optimal configuration corresponds to $\gamma = 10$ and balanced regularization weights at 1.0, which together maximize AU and expression recognition performance while maintaining psychologically interpretable causal structures.

G Training and Inference Efficiency

We report representative training and inference experiments to assess the computational efficiency of CausalAffect under different scales of training data. 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.

Experiment	**+All**	**AffectNet+BP4D**	**DISFA+RAF-DB**	**EmotioNet**
Total Training Images	∼800K	∼257K	∼100K	∼24K
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

Table 4: Training and inference efficiency of CausalAffect across different experimental configurations.

Large-Scale Configuration: The largest-scale training setting corresponds to **CausalAffect (+All)**, which involves approximately 800K images from six datasets. A batch size of 360 (60 images per dataset) was used. Training for 80 epochs required around 260 seconds per epoch, resulting in a total training time of ∼5.8 hours. The peak GPU memory usage was between 28–30 GB, while the inference speed reached ∼200 FPS. This demonstrates that CausalAffect remains computationally tractable even at large scale, making it feasible for deployment on high-performance GPUs.

Medium- and Small-Scale Configurations: For smaller-scale experiments, training time and memory requirements were substantially reduced: **AffectNet+BP4D** (∼257K images): batch size 120, training ∼1.29 hours, memory usage 18–20 GB, inference speed ∼250 FPS. **DISFA+RAF-DB** (∼100K images): batch size 120, training ∼0.57 hours, memory usage 18–19 GB, inference speed ∼250 FPS. **EmotioNet** (∼24K images): batch size 60, training ∼0.46 hours, memory usage 10–12 GB, inference speed ∼270 FPS.

These results highlight that CausalAffect scales gracefully with dataset size: larger datasets increase training time and GPU memory usage but inference remains efficient, consistently above 200 FPS across all settings.

References

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.

Paul Ekman and Wallace V Friesen. Facial action coding system. *Environmental Psychology & Nonverbal Behavior*, 1978.

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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5562–5570, 2016.

Jeffrey M Girard, Wen-Sheng Chu, László A Jeni, and Jeffrey F Cohn. Sayette group formation task (gft) spontaneous facial expression database. In *2017 12th IEEE international conference on automatic face & gesture recognition (FG 2017)*, pp. 581–588. IEEE, 2017.

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

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, pp. 2813–2821, 2024.

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*, pp. 2852–2861, 2017.

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.

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, 2017.

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.

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.

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

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.