RELATION-AWARE VISUAL–PHYSIOLOGICAL AFFECT REGRESSION WITH AN EVIDENCE-PREFIXED LARGE LANGUAGE MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal affect regression aims to predict continuous emotional states by combining complementary yet imperfect evidence from observable behavior and physiological bodily responses. Facial video captures visible expressions which can be subtle or intentionally regulated, while blood-volume pulse (BVP) measurements are susceptible to motion artifacts, sensor contact noise, and varying individual physiological baselines. We introduce a visual–physiological affect dataset of concurrently-recorded, temporally paired facial-video and BVP streams, with participant-specific, continuous-valued valence–arousal ratings collected for each stimulus presentation. We further propose an evidence-prefixed large-language-model architecture for direct numerical affect regression. Visual and BVP encoders produce modality-specific token sequences, and a continuous learned relation token summarizes their learned-space interactions through modality summaries, absolute difference, element-wise product, and cosine discrepancy. These evidence tokens precede a learnable valence-arousal query in a QLoRA-adapted Qwen model, whose final query hidden state directly predicts valence and arousal. Optimized solely via affect supervision, the relation token requires no manually annotated modality-relation or signal-quality labels. We perform controlled comparisons under a subject-dependent, source-trial-disjoint split and report window-level results for unimodal, non-LLM visual–BVP fusion, and no-relation variants. The relation token represents geometric discrepancy between learned representations rather than verified psychological conflict.

1 INTRODUCTION

Affective computing seeks to recognize human emotional states (Picard, 1997). The valence–arousal (VA) representation models pleasantness and activation on continuous axes (Russell, 1980), but both expressed behavior and physiological response vary across people. Facial video captures appearance and temporal behavior, yet expressions may be subtle, regulated, or affected by pose and illumination (Li & Deng, 2022; Sun et al., 2023). BVP provides a complementary view of cardiovascular dynamics but is also influenced by motion, breathing, sensor contact, and individual baseline (Shaffer & Ginsberg, 2017). Neither modality is universally reliable.

Existing resources combine behavioral and physiological sensing, but differ in participant-facing video, BVP availability, and annotation timescale (Koelstra et al., 2012; Soleymani et al., 2012; Miranda-Correa et al., 2021; Sharma et al., 2019; Ringeval et al., 2013). We contribute high-resolution facial video concurrently recorded and temporally paired with BVP, together with source-trial-level, continuous-valued VA ratings. Because one rating supervises all clips derived from a stimulus presentation, our task is continuous-valued regression under trial-level rather than time-continuous supervision.

The modeling challenge is to preserve both modality-specific token sequences while explicitly representing their interaction. We project the visual and BVP tokens into the embedding space of a pretrained decoder-only Transformer, Qwen2.5-3B-Instruct (Yang et al., 2024). Under its standard causal attention mask, a learned VA query placed after all evidence tokens can attend to the preceding visual, physiological, relation, and instruction tokens. Its final hidden state is then used as

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison by available participant-facing modality and VA annotation unit. Trial denotes one rating per elicitation trial; time denotes ratings that vary within a recording.

Resource	Participant video	BVP	VA unit
DEAP (Koelstra et al., 2012)	subset	yes	trial
AMIGOS (Miranda-Correa et al., 2021)	yes	no (ECG/GSR)	trial
CASE (Sharma et al., 2019)	no	yes	time
RECOLA (Ringeval et al., 2013)	yes	no (ECG/EDA)	time
Aff-Wild (Zafeiriou et al., 2017)	yes	no	time
VP-Affect	yes	yes	source trial

the sole input to the regression head. A continuous relation token summarizes the two modality representations through their pooled summaries, absolute difference, element-wise product, and cosine discrepancy. The token is optimized only through VA supervision and is not interpreted as a categorical relation or a measure of psychological conflict.

Our contributions are: (1) a visual–physiological affect dataset with source-trial-level VA ratings; (2) an evidence-prefixed Qwen architecture for direct numerical VA regression; and (3) a continuous relation token learned without relation-specific annotations, whose predictive utility is assessed through a matched relation-branch ablation alongside unimodal and non-LLM visual–BVP fusion comparisons.

2 Related Work

2.1 Multimodal affect modeling

Video benchmarks such as Aff-Wild evaluate continuous VA prediction from observable behavior (Zafeiriou et al., 2017; Kollias et al., 2023). Physiological affect datasets differ substantially in sensors and supervision. DEAP and MAHNOB-HCI combine peripheral measurements with elicitation and subjective ratings (Koelstra et al., 2012; Soleymani et al., 2012); AMIGOS adds frontal video to EEG, ECG, and GSR (Miranda-Correa et al., 2021). CASE records BVP with time-continuous self-reports (Sharma et al., 2019), while RECOLA combines behavioral recordings, ECG, EDA, and time-continuous VA labels (Ringeval et al., 2013). MVP is a closer video–physiology fusion reference (Strizhkova et al., 2025). Table 1 compares the modalities and annotation units most relevant to our setting.

Multimodal fusion ranges from concatenation to directional cross-modal attention and shared–specific representation learning. MulT models dependencies among unaligned sequences (Tsai et al., 2019); MISA separates modality-invariant and modality-specific spaces (Hazarika et al., 2020); and MMIM seeks to preserve task-relevant information during fusion (Han et al., 2021). These approaches motivate retaining the complete token sequences rather than replacing each stream with one pooled vector. A learned-space distance, however, is not a psychological measurement: it may reflect evidence differences, sensor noise, imperfect alignment, or representation error. We therefore use cosine discrepancy only as one feature alongside both modality summaries, their absolute difference, and their element-wise interaction, and validate the resulting token through held-out prediction.

2.2 Language models for affect

Recent systems use instruction tuning and language generation for emotion recognition and explanation (Lian et al., 2023; Cheng et al., 2024; Lian et al., 2025a;b). In contrast, our model uses a pretrained causal language model as a fusion backbone for direct numerical VA regression. The learned query is placed after all evidence tokens so that it can attend to the complete prefix.

The causal ordering is a functional part of the method. A query inserted before the evidence could not attend to later tokens under a causal mask. Modality-type embeddings mark token provenance, while independent projectors map the heterogeneous sensor representations to the language-model width. Across the Qwen-based ablations, we keep the pretrained backbone, instruction, VA query,

regression head, and training configuration unchanged. The visual-only and BVP-only variants use evidence from a single modality, whereas the multimodal variants use both modalities. The VA query is always placed after all available evidence tokens. The two multimodal variants differ in whether the explicit relation token is included.

Unlike explanation-oriented multimodal systems, our model uses the pretrained language model as a fusion backbone for numerical VA regression. It does not generate emotion descriptions or rationales; instead, the final hidden state of the learned VA query is passed directly to a regression head.

3 Data Acquisition and Preparation

3.1 Elicitation and recording

The collection includes 23 participants and 22 approximately two-minute stimuli. The nominal stimulus sequence followed a neutral–negative–neutral–positive progression, but these tags describe the intended elicitation rather than each participant's experience. Immediately after a presentation, the participant reported the actual experienced valence and arousal on separate 1–9 scales, with 5 as the approximate midpoint. Each target is therefore participant-specific and belongs to one complete stimulus presentation, which we call a source trial.

Before recording, the operator adjusted lighting, camera position, and viewing distance to obtain a clear frontal view. Facial behavior was recorded at 2560×1440 and 30 fps. A CONTEC CMS50E finger pulse oximeter recorded the BVP waveform at 60 Hz. The face and BVP streams were associated by participant and stimulus identifiers.

3.2 Clip Curation and Input Construction

The facial-video and BVP streams were recorded using separate devices and were started and stopped concurrently. Recordings were associated using participant and stimulus identifiers. During manual curation, identical relative start and end boundaries were applied to both modalities to preserve their temporal correspondence.

One or more paired segments of approximately 30 seconds were manually cropped from each participant–stimulus recording, yielding 815 curated facial-video–BVP pairs. These curated input videos, together with their paired BVP segments, constitute the inputs to the training and test pipelines.

Each participant–stimulus recording is treated as a source trial. For dataset partitioning, a source trial is used as an indivisible group: all curated input videos derived from the same source trial are assigned to the same partition. Subject to this grouping constraint, approximately 85% of the curated input videos are assigned to training and 15% to testing. Their paired BVP segments always follow the partition assignments of the corresponding videos.

Window generation is performed after partition assignment within the model's input-processing pipeline. Each curated input pair is divided into five-second windows with 50% overlap, corresponding to a stride of 2.5 seconds. For each window, ten facial frames are sampled uniformly, while the temporally corresponding BVP segment contains 300 samples at 60 Hz. Every window retains its participant, stimulus, source-trial, and parent-video identifiers and inherits both the partition assignment of its parent video and the numerical VA rating of its source trial. Additional details are provided in Appendix A.

4 Method

4.1 Overview

Given paired facial video $\mathcal{V}$ and BVP signal $\mathcal{P}$, the visual and physiological encoders produce modality token sequences X_v and X_p, respectively. A relation encoder further summarizes their interactions in a shared representation space as a continuous relation token z_r. The visual, physiological, and relation tokens are projected into the embedding space of a pretrained large language model and

arranged as an evidence prefix. A learnable valence–arousal query is appended to the end of the prefix, allowing its final hidden state to integrate all preceding multimodal evidence. A regression head then maps this query state to the predicted valence and arousal, $\hat{\mathbf{y}} = [\hat{v}, \hat{a}] \in \mathbb{R}^2$, as illustrated in Figure 1.

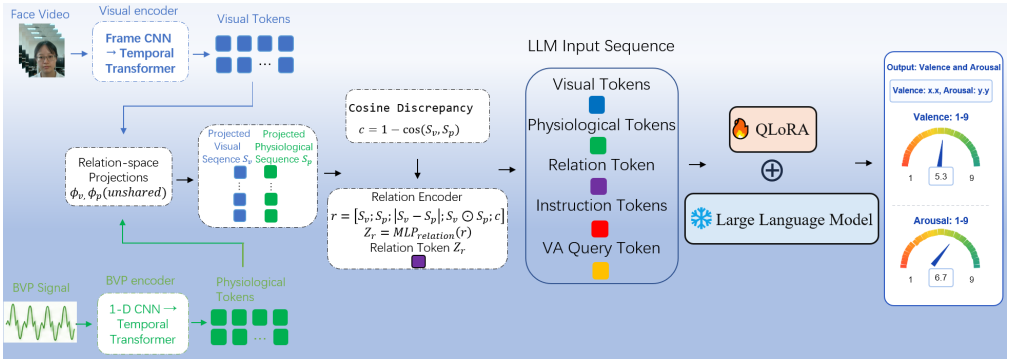

Figure 1: Overview of the proposed relation-aware VA regression framework. The visual and BVP encoders retain their full token sequences, while a separate relation branch constructs a token from pooled modality summaries and their interactions. The evidence sequences and fixed instruction precede a learned VA query in Qwen. The final query state is mapped to valence and arousal by a regression head. The no-relation variant omits the entire relation branch while retaining both modality streams.

4.2 Sensor Encoders and Relation Token

For each five-second window, the visual encoder processes ten uniformly sampled facial frames. Each frame passes through three convolution–group-normalization–GELU blocks, followed by spatial global average pooling and a linear projection. The resulting frame features are processed by a two-layer, four-head temporal Transformer. This branch preserves the ten-frame sequence length, allowing subsequent fusion to access the temporally contextualized visual tokens rather than only a pooled visual vector.

The physiological encoder processes the corresponding 300 BVP samples using a one-dimensional convolutional stack. Adaptive temporal pooling maps the resulting feature sequence to ten positions, followed by a two-layer, four-head temporal Transformer. Thus, the two encoders accommodate different input sampling rates while each producing ten tokens of width $d = 128$:

$$X_v = E_v(\mathcal{V}), \quad X_p = E_p(\mathcal{P}), \qquad X_v, X_p \in \mathbb{R}^{10\times d}. \tag{1}$$

The branches retain their ordered sequences rather than reducing them to one feature vector. For relation modeling only, independently parameterized blocks ϕ_v and ϕ_p map the two sequences into a common-width comparison space:

$$S_m = \phi_m(X_m) = \mathrm{Dropout}\left(\mathrm{GELU}(W_m \mathrm{LN}(X_m) + b_m)\right), \quad m \in \{v, p\}. \tag{2}$$

The two blocks have the same architecture but do not share parameters. "Common" therefore denotes compatible dimensionality, not shared weights. We obtain one summary per stream by temporal mean pooling,

$$s_v = \frac{1}{T_v}\sum_{i=1}^{T_v} S_{v,i}, \qquad s_p = \frac{1}{T_p}\sum_{j=1}^{T_p} S_{p,j}, \qquad s_v, s_p \in \mathbb{R}^d. \tag{3}$$

Their cosine discrepancy is

$$c = 1 - \frac{s_v^\top s_p}{\max(\|s_v\|_2, \varepsilon)\max(\|s_p\|_2, \varepsilon)}, \tag{4}$$

where $\varepsilon = 10^{-8}$ prevents division by zero. The implementation clips cosine similarity to $[-1, 1]$, so $c \in [0, 2]$. A larger value denotes greater directional difference between learned summaries; it is not a measure of psychological conflict or sensor quality.

The discrepancy alone cannot retain the content of either stream. We therefore construct

$$r = [s_v; s_p; |s_v - s_p|; s_v \odot s_p; c] \in \mathbb{R}^{4d+1}, \qquad z_r = \mathrm{MLP}_r(r) \in \mathbb{R}^d. \tag{5}$$

The relation encoder uses layer normalization followed by a two-layer MLP,

$$\mathrm{MLP}_r(r) = W_2 \mathrm{Dropout}\left(\mathrm{GELU}(W_1 \mathrm{LN}(r) + b_1)\right) + b_2, \tag{6}$$

with $W_1 \in \mathbb{R}^{2d \times (4d+1)}$ and $W_2 \in \mathbb{R}^{d \times 2d}$. For $d = 128$, this maps $513 \to 256 \to 128$.

The components of the relation feature play complementary roles. The vectors s_v and s_p retain modality-level evidence; $|s_v - s_p|$ describes dimension-wise difference; $s_v \odot s_p$ represents dimension-wise interaction; and c supplies a global directional discrepancy. A two-layer MLP compresses the resulting $(4d+1)$-dimensional feature to one d-dimensional token. This design does not discretize cross-modal relations or assign pseudo-labels. The utility of the complete relation-token design is evaluated by comparing the full model with a matched multimodal variant that omits the entire relation branch.

4.3 Evidence-Prefix Fusion and Direct Regression

The relation-space projections above are used only to construct z_r; they do not replace either full sensor sequence. Independently parameterized projectors P_v, P_p, and P_r apply layer normalization and a linear map from sensor width d to Qwen width D. Learned type embeddings identify the source of each token:

$$H_v = P_v(X_v) + e_v, \qquad H_p = P_p(X_p) + e_p, \qquad H_r = P_r(z_r) + e_r. \tag{7}$$

Thus $H_v, H_p \in \mathbb{R}^{10 \times D}$ and $H_r \in \mathbb{R}^{1 \times D}$, with $D = 2048$. The causal Transformer input is

$$H_0 = [H_v; H_p; H_r; H_I; H_q], \tag{8}$$

where H_I is a fixed target-free task instruction and H_q is a learned VA query. Neither contains the ground-truth rating. Because the query is last, it can attend to all preceding evidence. The sole prediction path is

$$\widehat{\mathbf{y}} = \boldsymbol{\mu}_{\mathrm{tr}} + \boldsymbol{\sigma}_{\mathrm{tr}} \odot h_{\mathrm{VA}}([F_\theta(H_0)]_{\mathrm{query}}), \tag{9}$$

where F_θ denotes QLoRA-adapted Qwen2.5-3B, and $\boldsymbol{\mu}_{\mathrm{tr}}$ and $\boldsymbol{\sigma}_{\mathrm{tr}}$ are the coordinate-wise mean and standard deviation of the training targets. The no-relation variant omits H_r, and no independent encoder-only residual prediction is added. The causal query state, regression head, and output denormalization are detailed in Appendix B.2.

4.4 Learning Configuration

During final adaptation, the visual and BVP encoders, including their temporal modules, are jointly optimized with the fusion and regression components. The trainable components shared by the two multimodal Qwen variants (with and without the relation branch) include the visual encoder E_v, the BVP encoder E_p, the visual and physiological evidence projectors P_v and P_p, the modality-type embeddings, the learned VA query, the regression head, and the QLoRA adapters.

In the full relation model, the relation-space projectors ϕ_v and ϕ_p, the relation MLP, the relation evidence projector P_r, and the relation-type embedding are additionally trained. The no-relation variant omits the entire relation branch. All trainable components are optimized jointly using the VA regression objective, without relation-specific supervision.

The base Qwen parameters remain frozen and are adapted through QLoRA. The shared components of the with- and without-relation variants use the same initialization and optimization configuration.

4.5 Role of the Explicit Relation Representation

The relation path is auxiliary to, rather than a replacement for, the full modality streams. Mean pooling is used only to construct s_v and s_p for the relation calculation; all ten visual tokens and all ten BVP tokens remain available to the Qwen query. Consequently, the relation token cannot improve prediction merely by restoring sensor content discarded by the main path. It instead supplies an

explicit summary of how the projected streams compare, while the backbone can still recover fine-grained information from the original sequences.

The construction also avoids imposing a categorical ontology on cross-modal behavior. A sample is never labeled as "consistent," "conflicting," or "unreliable." Such categories would require independent human or signal-quality annotation. The MLP is free to transform the summaries, difference, interaction, and discrepancy into whatever continuous representation is useful for VA regression. The ablation can establish predictive utility of this representation, but not a unique causal or psychological interpretation.

To evaluate explicit cross-modal relation modeling, we compare two Qwen-based multimodal variants that share the same visual and BVP evidence, pretrained backbone, VA query, regression head, and training configuration. The full model incorporates the structured relation token H_r into the evidence prefix, whereas the no-relation variant omits the complete relation branch.

5 Experiments

5.1 Input Construction and Optimization

Each paired interval is windowed with a 2.5-second stride. Ten frames are sampled uniformly according to a window's relative start and end times. The visual encoder uses three convolution–normalization–GELU blocks followed by spatial pooling and a two-layer temporal Transformer. Missing BVP samples are interpolated, outliers are clipped, a third-order 0.7–4 Hz Butterworth band-pass filter is applied, and each window is standardized. A 1-D CNN, adaptive temporal pooling, and a two-layer temporal Transformer produce ten physiological tokens.

The visual and BVP encoders, including their temporal modules, are jointly optimized with the evidence projectors, type embeddings, learned VA query, regression head, and QLoRA adapters. In the full model, the relation-space projectors, relation MLP, relation evidence projector, and relation-type embedding are additionally trained. The base Qwen parameters remain frozen. We use Qwen2.5-3B-Instruct with 4-bit NF4 quantization, double quantization, and bfloat16 computation. QLoRA (Dettmers et al., 2023) is applied to the query, key, value, and output projection modules with rank 8, scaling factor 16, and dropout 0.05.

AdamW (Loshchilov & Hutter, 2019) optimization runs for 20 epochs with a mini-batch size of 128 window samples, learning rate 10^{-4}, weight decay 10^{-2}, and gradient-norm clipping at 1.0. We use random seed 42 and apply no learning-rate warm-up or scheduler.

$$\mathcal{L}_{\mathrm{VA}} = \mathrm{SmoothL1}_{\beta=1}(\widehat{\mathbf{y}}, \mathbf{y}) + 0.5 \left[1 - \frac{\mathrm{CCC}(\widehat{v}, v) + \mathrm{CCC}(\widehat{a}, a)}{2} \right] \tag{10}$$

The two loss terms address different aspects of VA prediction. The SmoothL1 term penalizes sample-wise numerical errors, using a quadratic penalty for small residuals and a linear penalty for larger residuals. The CCC term encourages agreement between predictions and targets within each mini-batch by accounting for their covariance and differences in mean and variance. CCC is computed separately for valence and arousal, and the two contributions are averaged. Both loss terms are evaluated using the decoded predictions on the original rating scale. A denominator guard of 10^{-8} is used for numerical stability in the batch-level CCC calculation.

5.2 Protocol and Metrics

We use a subject-dependent, source-trial-disjoint train–test split. Each participant–stimulus recording is treated as an indivisible source-trial group, so all curated input videos derived from the same source trial are assigned to the same partition. Subject to this grouping constraint, approximately 85% of the curated input videos are assigned to training and 15% to testing. The paired BVP segments follow the partition assignments of their corresponding videos. Different source trials from the same participant may appear in both partitions. No separate validation partition is used.

Partition assignment is completed before window generation. Within the model's input-processing pipeline, each curated input pair is divided into five-second windows with a stride of 2.5 seconds. All derived windows inherit the partition assignment of their parent video. Consequently, no source trial, curated input video, or derived window appears in both training and testing.

Each source trial has one numerical VA rating pair collected after the complete stimulus presentation. This rating is inherited by all curated input videos and five-second windows derived from that source trial. Evaluation is performed directly over the labeled windows in the test partition; predictions are not averaged across windows, curated input videos, or source trials.

All models are trained for 20 epochs with a mini-batch size of 128. We compute the training-set MAE after each epoch and select the checkpoint attaining the lowest training MAE over the 20 epochs. No validation data are used for checkpoint selection, and the test partition is used only for the final evaluation. The label mean and standard deviation used in Equation 9 are computed from the training targets only.

We report MAE, RMSE, Pearson correlation coefficient (PCC), and concordance correlation coefficient (CCC) (Lin, 1989) separately for valence and arousal. MAE and RMSE measure prediction errors on the original 1–9 rating scale. PCC measures linear association, whereas CCC additionally accounts for agreement in mean and variance.

5.3 Baselines and Results

We compare five model configurations to examine the roles of input modalities, the fusion architecture, and the explicit relation branch. The Qwen-based visual-only and BVP-only variants receive evidence from one modality while retaining the fixed instruction and final-query regression design. The non-LLM visual–BVP fusion baseline pools and combines the two encoder sequences before direct regression; its complete architecture is given in Appendix B.3. The two multimodal Qwen variants retain the complete visual and BVP token sequences. The no-relation variant excludes the relation branch, whereas the full model additionally inserts the structured relation token into the evidence prefix.

These configurations define three comparisons. First, comparing Qwen without the relation branch against the two Qwen-based unimodal variants assesses the benefit of combining modalities within the same backbone framework. Second, comparing Qwen without the relation branch against visual–BVP fusion without an LLM compares the two implemented fusion architectures. Third, comparing the full Qwen model against the matched no-relation variant evaluates the contribution of the complete relation-token branch. All variants use the same train–test split, 20-epoch training budget, batch size, checkpoint-selection rule, and test evaluation protocol.

Table 2: Window-level results on the subject-dependent, source-trial-disjoint test partition.

	Valence				Arousal			
Method	MAE↓	RMSE↓	PCC↑	CCC↑	MAE↓	RMSE↓	PCC↑	CCC↑
Qwen, Visual only	2.31	2.38	0.142	0.108	2.02	2.15	0.096	0.063
Qwen, BVP only	2.19	2.23	0.176	0.137	1.92	2.05	0.139	0.092
Visual–BVP fusion (no LLM)	2.02	2.15	0.226	0.181	1.75	1.88	0.168	0.105
Qwen, no relation	1.92	2.04	0.274	0.243	1.48	1.56	0.185	0.112
Qwen, full relation	1.61	1.82	0.308	0.276	1.08	1.15	0.215	0.157

Contribution of multimodal evidence. The Qwen-based multimodal model without the relation token outperforms both Qwen-based unimodal variants. Compared with visual-only and BVP-only prediction, valence MAE decreases from 2.31 and 2.19 to 1.92, respectively, while arousal MAE decreases from 2.02 and 1.92 to 1.48. RMSE, PCC, and CCC also improve for both affect dimensions. These comparisons support the benefit of combining visual and physiological evidence within the same Qwen-based architecture.

Comparison of fusion architectures. Compared with visual–BVP fusion without an LLM, the multimodal Qwen model without the relation branch reduces valence MAE from 2.02 to 1.92 and arousal MAE from 1.75 to 1.48. RMSE, PCC, and CCC also improve for both affect dimensions. These results favor the implemented Qwen-based token-fusion architecture over the pooled-MLP fusion baseline under the reported protocol.

Contribution of the relation-token branch. Compared with the matched Qwen-based variant without the relation branch, the full model reduces valence MAE from 1.92 to 1.61 and arousal MAE from 1.48 to 1.08, while also improving RMSE, PCC, and CCC for both dimensions. These improvements support the utility of the proposed relation-token branch, which explicitly represents cross-modal relationships through modality summaries, absolute differences, element-wise products, and cosine discrepancy.

All results in Table 2 are obtained using the same subject-dependent, source-trial-disjoint train–test split and a fixed random seed of 42.

6 DISCUSSION

The results support the benefit of combining facial and physiological evidence for VA regression. Within the Qwen-based architecture, the multimodal variant without the relation token outperforms both the visual-only and BVP-only variants across the reported metrics. Facial video represents observable behavior and temporal expression patterns, whereas BVP provides physiological measurements. Retaining both token sequences allows the fusion model to access these complementary sources of evidence when estimating valence and arousal.

Compared with visual–BVP fusion without an LLM, the Qwen-based multimodal variant without the relation token achieves lower prediction errors and higher PCC and CCC for both affect dimensions. This comparison supports the effectiveness of the implemented Qwen-based fusion architecture under the reported protocol. The learned VA query is positioned after all evidence and instruction tokens, allowing its final hidden state to integrate the complete input prefix through causal attention. A regression head then maps this state directly to numerical VA predictions, making Qwen a multimodal fusion backbone in this framework.

The full model further improves all four metrics for both affect dimensions relative to the matched no-relation variant. The relation branch combines modality summaries, their absolute difference, element-wise product, and cosine discrepancy into a dedicated token. This token provides an explicit representation of cross-modal relationships alongside the original visual and physiological token sequences. The comparison supports the utility of the implemented relation-token design as a whole. Optimized through VA supervision without relation-specific annotations, the token represents learned cross-modal interaction rather than a categorical psychological conflict label or an independently validated signal-quality score.

The supervision and evaluation protocol define the meaning of these predictions. Each participant–stimulus recording provides one source-trial VA rating pair, which is inherited by all curated input videos and five-second windows derived from that recording. Partitioning is performed with source-trial grouping before window generation, so all curated videos and derived windows from one source trial remain in the same partition. Metrics are computed directly over test windows without averaging predictions across windows, curated videos, or source trials. The resulting evaluation measures how effectively short paired observations predict their inherited trial-level VA targets for held-out source trials from participants represented in training.

The reported findings are based on a dataset of 23 participants, one fixed train–test split, and one random seed. Future work can extend the evaluation to participant-disjoint settings and repeated splits and seeds, while examining robustness to changes in BVP quality, temporal alignment, and modality availability. These evaluations would further characterize the applicability of relation-aware fusion across participants and recording conditions.

7 CONCLUSION

We presented an evidence-prefixed Qwen framework for direct VA regression from facial video and BVP. The model preserves the complete visual and physiological token sequences and augments them with a structured, relation-label-free token before a learned VA query. This relation representation explicitly combines modality summaries, absolute difference, element-wise interaction, and cosine discrepancy, while the final query state provides the sole input to the numerical regression head.

Under a subject-dependent, source-trial-disjoint split with window-level evaluation, the full model consistently outperforms the Qwen-based visual-only and BVP-only variants, visual–BVP fusion without an LLM, and the Qwen no-relation variant. It achieves the lowest MAE and RMSE and the highest PCC and CCC for both valence and arousal. These results support the complementary value of facial and physiological evidence, the effectiveness of Qwen-based multimodal fusion, and the contribution of explicit cross-modal relation modeling.

The accompanying dataset contains concurrently recorded and temporally paired facial-video and BVP streams organized into 815 curated input pairs. Each five-second window inherits the numerical VA rating and partition assignment of its source trial. Source-trial grouping ensures that all curated videos and derived windows from the same participant–stimulus recording remain in one partition. Together, the dataset and model provide a unified framework for window-level visual–physiological affect regression.

Reproducibility Statement

The appendix specifies data construction, source-trial-grouped partitioning, window-level evaluation, prediction decoding, the non-LLM baseline architecture, model dimensions, preprocessing, optimization, and the complete loss. The main text reports the split unit, evaluation unit, trainable components, and principal hyperparameters.

AI Use Statement

Generative AI tools were used for language editing, LaTeX restructuring, and literature retrieval for related work discovery. The authors reviewed all generated content and remain fully responsible for the scientific claims, analyses, and final manuscript.

REFERENCES

Z. Cheng, Z.-Q. Cheng, J.-Y. He, J. Sun, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. G. Hauptmann. Emotion-LLaMA: Multimodal emotion recognition and reasoning with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, 2024. doi: 10.52202/079017-3518.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

W. Han, H. Chen, and S. Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In *Proc. Conf. Empirical Methods Natural Language Processing (EMNLP)*, pp. 9180–9192, Online and Punta Cana, Dominican Republic, 2021. doi: 10.18653/v1/2021.emnlp-main.723.

D. Hazarika, R. Zimmermann, and S. Poria. MISA: Modality-invariant and -specific representations for multimodal sentiment analysis. In *Proc. 28th ACM Int. Conf. Multimedia (ACM MM)*, pp. 1122–1131, Seattle, WA, USA, 2020. doi: 10.1145/3394171.3413678.

S. Koelstra, C. Mühl, M. Soleymani, J.-S. Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. DEAP: A database for emotion analysis using physiological signals. *IEEE Trans. Affect. Comput.*, 3(1):18–31, Jan.–Mar. 2012. doi: 10.1109/T-AFFC.2011.15.

Dimitrios Kollias, Panagiotis Tzirakis, Alice Baird, Alan Cowen, and Stefanos Zafeiriou. ABAW: Valence-arousal estimation, expression recognition, action unit detection and emotional reaction intensity estimation challenges. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 5889–5898, 2023.

S. Li and W. Deng. Deep facial expression recognition: A survey. *IEEE Trans. Affect. Comput.*, 13 (3):1195–1215, Jul.–Sep. 2022. doi: 10.1109/TAFFC.2020.2981446.

Z. Lian, H. Sun, L. Sun, H. Gu, Z. Wen, S. Zhang, S. Chen, M. Xu, K. Xu, K. Chen, L. Chen, S. Liang, Y. Li, J. Yi, B. Liu, and J. Tao. Explainable multimodal emotion recognition. *arXiv preprint arXiv:2306.15401*, 2023.

Z. Lian, H. Chen, L. Chen, H. Sun, L. Sun, Y. Ren, Z. Cheng, B. Liu, R. Liu, X. Peng, J. Yi, and J. Tao. AffectGPT: A new dataset, model, and benchmark for emotion understanding with multimodal large language models. In *Proc. 42nd Int. Conf. Mach. Learn. (ICML)*, volume 267, pp. 36993–37014, 2025a.

Z. Lian, H. Sun, L. Sun, H. Chen, L. Chen, H. Gu, Z. Wen, S. Chen, S. Zhang, H. Yao, B. Liu, R. Liu, S. Liang, Y. Li, J. Yi, and J. Tao. OV-MER: Towards open-vocabulary multimodal emotion recognition. In *Proc. 42nd Int. Conf. Mach. Learn. (ICML)*, volume 267, pp. 37015–37050, 2025b.

Lawrence I-Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, 45(1):255–268, 1989. doi: 10.2307/2532051.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Juan Abdon Miranda-Correa, Mojtaba Khomami Abadi, Nicu Sebe, and Ioannis Patras. AMIGOS: A dataset for affect, personality and mood research on individuals and groups. *IEEE Transactions on Affective Computing*, 12(2):479–493, 2021. doi: 10.1109/TAFFC.2018.2884461.

R. W. Picard. *Affective Computing*. MIT Press, Cambridge, MA, USA, 1997.

Fabien Ringeval, Andreas Sonderegger, Juergen Sauer, and Denis Lalanne. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, 2013. doi: 10.1109/FG.2013.6553805.

James A. Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39 (6):1161–1178, 1980. doi: 10.1037/h0077714.

Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:258, 2017. doi: 10.3389/fpubh.2017.00258.

Karan Sharma, Claudio Castellini, Egon L. van den Broek, Alin Albu-Schaeffer, and Friedhelm Schwenker. A dataset of continuous affect annotations and physiological signals for emotion analysis. *Scientific Data*, 6:196, 2019. doi: 10.1038/s41597-019-0209-0.

M. Soleymani, J. Lichtenauer, T. Pun, and M. Pantic. A multimodal database for affect recognition and implicit tagging. *IEEE Trans. Affect. Comput.*, 3(1):42–55, Jan.–Mar. 2012. doi: 10.1109/T-AFFC.2011.25.

Valeriya Strizhkova, Hadi Kachmar, Hava Chaptoukaev, Raphael Kalandadze, Natia Kukhilava, Tatia Tsmindashvili, Nibras Abo-Alzahab, Maria A. Zuluaga, Michal Balazia, Antitza Dantcheva, Francois Bremond, and Laura Ferrari. MVP: Multimodal emotion recognition based on video and physiological signals. *arXiv preprint arXiv:2501.03103*, 2025.

L. Sun, Z. Lian, B. Liu, and J. Tao. MAE-DFER: Efficient masked autoencoder for self-supervised dynamic facial expression recognition. In *Proc. 31st ACM Int. Conf. Multimedia (ACM MM)*, pp. 6110–6121, Ottawa, ON, Canada, 2023. doi: 10.1145/3581783.3612365.

Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics (ACL)*, pp. 6558–6569, Florence, Italy, 2019. doi: 10.18653/v1/P19-1656.

An Yang et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Stefanos Zafeiriou, Dimitrios Kollias, Mihalis A. Nicolaou, Athanasios Papaioannou, Guoying Zhao, and Irene Kotsia. Aff-Wild: Valence and arousal in-the-wild challenge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 34–41, 2017.

A Dataset Construction and Partition Details

The dataset was collected from 23 participants viewing 22 stimulus videos of approximately two minutes each. The nominal presentation sequence followed a neutral–negative–neutral–positive progression. After each stimulus presentation, participants reported their experienced valence and arousal on separate 1–9 scales, with 5 representing the approximate midpoint. Facial video was recorded at 2560×1440 resolution and 30 fps, while BVP was recorded at 60 Hz using a CONTEC CMS50E finger pulse oximeter.

The facial-video and BVP streams were acquired using separate devices and were started and stopped concurrently. Recordings were associated using participant and stimulus identifiers. During manual curation, identical relative start and end boundaries were applied to both streams to preserve temporal correspondence between the facial-video and BVP segments.

One or more paired segments of approximately 30 seconds were manually cropped from each participant–stimulus recording, yielding 815 curated facial-video–BVP input pairs. Each participant–stimulus recording was treated as one source trial and used as an indivisible group for partitioning. All curated input videos derived from the same source trial were assigned to the same partition, together with their paired BVP segments. Subject to this grouping constraint, approximately 85% of the curated input videos were assigned to training and 15% to testing. Thus, the 85:15 ratio refers to the number of curated input videos rather than the number of source trials or derived windows. The split was subject-dependent, so different source trials from the same participant could occur in both partitions. No separate validation partition was used.

Window generation was performed only after partition assignment, within the model's input-processing pipeline. Each curated input pair was divided into five-second windows with 50% overlap, corresponding to a stride of 2.5 seconds. Ten facial frames were sampled uniformly from each window, and the temporally corresponding BVP segment contained 300 samples at 60 Hz. Every window retained its participant, stimulus, source-trial, and parent-video identifiers and inherited the VA rating and partition assignment of its source trial. Consequently, no source trial, curated input video, or derived window appeared in both training and testing.

B Model Implementation Details

B.1 Encoder and Relation Components

The visual encoder comprises three convolution–group-normalization–GELU blocks, spatial global average pooling, a linear projection, and a two-layer, four-head temporal Transformer. It preserves the ten-frame temporal length. The BVP encoder uses a 1-D convolutional stack, adaptive temporal pooling to ten positions, and a two-layer, four-head temporal Transformer. Both outputs have width $d = 128$.

For $m \in \{v, p\}$, the modality-specific relation projection is

$$\phi_m(X_m) = \mathrm{Dropout}(\mathrm{GELU}(W_m \mathrm{LN}(X_m) + b_m)),$$

where $W_m \in \mathbb{R}^{d \times d}$. The two projections share architecture but not parameters. Equation 5 concatenates two summaries, their absolute difference and element-wise product, and the scalar cosine discrepancy, producing $r \in \mathbb{R}^{4d+1}$. The relation encoder is

$$\mathrm{MLP}_r(r) = W_2 \mathrm{Dropout}(\mathrm{GELU}(W_1 \mathrm{LN}(r) + b_1)) + b_2,$$

with $W_1 \in \mathbb{R}^{2d \times (4d+1)}$ and $W_2 \in \mathbb{R}^{d \times 2d}$. Cosine similarity is clipped to $[-1, 1]$ and uses $\varepsilon = 10^{-8}$.

The independent projectors P_v, P_p, and P_r each apply layer normalization followed by a linear map from $d = 128$ to the Qwen width $D = 2048$.

B.2 Prediction Path and Output De-normalization

Equation 9 can be decomposed into three steps:

$$\mathbf{h}_q^{\text{out}} = [F_\theta(H_0)]_{\text{query}} \in \mathbb{R}^D,$$
$$\widehat{\mathbf{y}}_{\text{std}} = h_{\text{VA}}(\mathbf{h}_q^{\text{out}}) \in \mathbb{R}^2,$$
$$\widehat{\mathbf{y}} = \boldsymbol{\mu}_{\text{tr}} + \boldsymbol{\sigma}_{\text{tr}} \odot \widehat{\mathbf{y}}_{\text{std}}.$$

Here, H_0 is the complete sequence of visual, physiological, relation, instruction, and query tokens defined in Equation 8, and F_θ is the QLoRA-adapted Qwen backbone. The operator $[\cdot]_{\text{query}}$ selects the final-layer state at the learned VA-query position. As the final sequence element, the query can causally attend to every preceding evidence and instruction token.

The regression head $h_{\text{VA}} : \mathbb{R}^D \to \mathbb{R}^2$ applies layer normalization, a $2048 \to 1024$ linear layer, GELU, dropout, and a $1024 \to 2$ output layer. Its output $\widehat{\mathbf{y}}_{\text{std}} = [\widehat{v}_{\text{std}}, \widehat{a}_{\text{std}}]$ is expressed in the standardized target coordinates. The vectors $\boldsymbol{\mu}_{\text{tr}}, \boldsymbol{\sigma}_{\text{tr}} \in \mathbb{R}^2$ are the coordinate-wise mean and standard deviation computed only from the training targets, and $\odot$ denotes element-wise multiplication. The final output $\widehat{\mathbf{y}} = [\widehat{v}, \widehat{a}]$ is thereby restored to the original VA rating scale. Only $\mathbf{h}_q^{\text{out}}$ is used for prediction: no other Qwen state is pooled, the target is absent from the prefix, and no independent encoder residual is added.

B.3 Non-LLM Visual–BVP Fusion Baseline

The non-LLM baseline receives the visual and physiological encoder sequences $X_v \in \mathbb{R}^{T_v \times d}$ and $X_p \in \mathbb{R}^{T_p \times d}$. It first applies temporal mean pooling:

$$\overline{\mathbf{x}}_v = \frac{1}{T_v} \sum_{i=1}^{T_v} X_{v,i}, \qquad \overline{\mathbf{x}}_p = \frac{1}{T_p} \sum_{j=1}^{T_p} X_{p,j}.$$

The pooled vectors are concatenated as $u = [\overline{\mathbf{x}}_v; \overline{\mathbf{x}}_p] \in \mathbb{R}^{2d}$ and passed directly to a regression MLP,

$$\widehat{\mathbf{y}}_{\text{base}} = h_{\text{base}}(u) \in \mathbb{R}^2.$$

Here $\widehat{\mathbf{y}}_{\text{base}} = [\widehat{v}_{\text{base}}, \widehat{a}_{\text{base}}]$ is the output scored on the original VA rating scale. The MLP consists of $\text{LayerNorm}(2d)$, $\text{Linear}(2d, d)$, GELU, $\text{Linear}(d, d)$, GELU, and $\text{Linear}(d, 2)$. It contains no Qwen backbone, task instruction, VA query, or relation branch, and otherwise follows the same split, loss, 20-epoch budget, batch size, checkpoint-selection rule, and window-level test protocol.

C Optimization and Reproducibility Details

Missing BVP values are interpolated, outliers are clipped, and a third-order 0.7–4 Hz Butterworth band-pass filter precedes window-wise standardization; no hand-crafted physiological feature tokens are used.

The visual and BVP encoders and their temporal modules are jointly optimized with the evidence projectors, type embeddings, VA query, regression head, and QLoRA adapters. The full model additionally trains the relation-space projectors, relation MLP, relation evidence projector, and relation-type embedding. The frozen Qwen2.5-3B-Instruct backbone uses 4-bit NF4, double quantization, and bfloat16 computation. QLoRA adapts its query, key, value, and output projections with rank 8, scaling factor 16, and dropout 0.05. AdamW training runs for 20 epochs with a mini-batch size of 128, learning rate 10^{-4}, weight decay 10^{-2}, gradient-norm clipping at 1.0, and seed 42, without warm-up or a scheduler.

The decoded VA output is optimized with

$$\mathcal{L}_{\text{VA}} = \text{SmoothL1}_{\beta=1}(\widehat{\mathbf{y}}, \mathbf{y}) + 0.5 \left[1 - \frac{\text{CCC}(\widehat{v}, v) + \text{CCC}(\widehat{a}, a)}{2} \right].$$

Batch-level CCC uses a denominator guard of 10^{-8}. The model uses no relation labels or language-generation loss. Training-set MAE is computed after every epoch, and the checkpoint with the lowest training MAE over the 20 epochs is selected for the reported test evaluation. The test partition is not used for parameter updates, hyperparameter selection, or checkpoint selection.