Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

QG-TIS: Quality-Guided Temporal Instance Selection for Resource-Efficient Multimodal Emotion Recognition

Juecheng Cai
Minnan Normal University
Zhoufu Chen
Zhangzhou Public Security Bureau
Jingmin Yang
yjm1758@mnnu.edu.cn

Minnan Normal University

Research Article

Keywords: Multimodal emotion recognition, Parameter-efficient fine-tuning, Multiple instance learning, Temporal instance selection, Instance quality gate

Posted Date: November 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9609891/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

QG-TIS: Quality-Guided Temporal Instance Selection for Resource-Efficient Multimodal Emotion Recognition

Juecheng Cai[1,2], Zhoufu Chen[3], Jingmin Yang[1,2*]

[1]School of Computer Science, Minnan Normal University, Zhangzhou 363000, China.
[2]Key Laboratory of Data Science and Intelligence Application, Fujian Province University, Zhangzhou 363000, China.
[3]Second Branch, Traffic Police Division, Zhangzhou Public Security Bureau, Zhangzhou 363005, China.

*Corresponding author(s). E-mail(s): yjm1758@mnnu.edu.cn;

Abstract

Deploying multimodal emotion recognition models on resource-constrained devices requires a careful balance between recognition performance and computational and memory costs. This paper proposes **QG-TIS**, a quality-guided temporal instance selection framework. The proposed method adopts a partial-freezing strategy for the Swin Transformer backbone and introduces a BiGRU-based temporal selector together with an instance quality gate (IQG) to select and reweight video instances, thereby reducing training and inference costs while preserving discriminative capability. Under the subject-dependent DEAP evaluation protocol, results over three independent runs show that QG-TIS achieves **95.79% ± 0.22%** accuracy and **95.69% ± 0.28%** macro-F1. Compared with full fine-tuning, the number of trainable parameters in the Swin backbone is reduced by **48.46%**. These results demonstrate a strong trade-off between recognition accuracy and resource efficiency. With substantially fewer trainable parameters, the framework is well-suited for deployment on consumer-grade hardware.

Keywords: Multimodal emotion recognition, Parameter-efficient fine-tuning, Multiple instance learning, Temporal instance selection, Instance quality gate

1 Introduction

Affective recognition is a central topic in affective computing, with applications ranging from human–computer interaction and mental health monitoring to education and entertainment. Compared with single-modality approaches, multimodal fusion—such as combining facial video and electroencephalography (EEG)—provides complementary information: the visual modality captures behavioral cues, while EEG reflects underlying neurophysiological responses. Joint modeling of these signals generally improves robustness and fine-grained recognition performance [1–4].

However, transferring high-performing models to resource-constrained environments—such as consumer-grade GPUs or edge devices—remains challenging. Vision Transformers like Swin Transformer achieve strong performance but often require substantial memory and computation for full fine-tuning [5, 6]. Parameter-efficient fine-tuning (PEFT), including adapters, LoRA, prompt tuning, and staged freezing, provides a practical way to reduce trainable parameters while

maintaining performance [7–9]. In this work, we adopt staged freezing to reserve computational budget for temporal modeling and robustness.

In video emotion recognition, multiple instance learning (MIL) offers a natural formulation, where a video is treated as a bag of instances [10, 11]. While attention-based MIL can weight instances, it often neglects temporal order, which is important when emotional cues evolve over time. Recent work addresses this by introducing temporal Transformers, recurrent networks, or differentiable selection mechanisms such as Gumbel–Softmax [12, 13]. These methods highlight the importance of temporal dynamics, but typically do not explicitly consider instance quality.

Building on these observations, this paper investigates how temporal modeling and instance quality assessment can be jointly designed under limited computational resources. We propose **QG-TIS (Quality-Guided Temporal Instance Selection)**, which integrates three key components: (i) staged partial freezing for a parameter-efficient visual backbone, (ii) a lightweight BiGRU-based temporal selector to model instance order, and (iii) an instance quality gate (IQG) to adaptively reweight instances based on reliability. This design improves temporal decision-making while keeping the number of trainable parameters small.

On the DEAP benchmark, QG-TIS is evaluated under the *subject-dependent* protocol, with parameter count, GPU memory, and iteration time reported for fair comparison. Based on three independent runs, the method achieves stable performance with modest resource usage. Due to the limited number of runs, conclusions are presented conservatively, and further validation is left for future work.

The main contributions of this paper are summarized as follows:

- QG-TIS is proposed as a quality-guided temporal instance selection framework for multimodal emotion recognition. Built on a partially frozen visual backbone, it combines temporal modeling with instance quality reweighting to enable adaptive selection and fusion of key video instances.
- A BiGRU-based temporal instance selector and an instance quality gate (IQG) are designed to model intra-bag temporal dependencies and suppress low-quality instances, respectively, thereby improving the identification of emotion-relevant visual cues.
- Systematic experiments and ablation studies are conducted under the *subject-dependent* protocol on DEAP. The results show that the proposed method offers a favorable overall trade-off in terms of accuracy, macro-F1 score, parameter efficiency, and memory cost.

2 Related Work

Studies related to this work can be broadly grouped into four directions: parameter-efficient adaptation of visual backbones, multiple instance learning with temporal modeling, multimodal fusion for emotion recognition, and robustness against noisy inputs. We briefly review each and clarify the position of the proposed method.

2.1 Parameter-Efficient Adaptation of Visual Backbones

Adapting large vision models with limited training overhead has become an important research direction. Representative approaches include lightweight adapters, low-rank adaptation (LoRA), and prompt/prefix tuning [7, 8]. These methods reduce trainable parameters while maintaining performance, but their effectiveness depends on task characteristics and data scale [9, 14]. In this work, we adopt staged freezing as a simple and stable strategy, reserving the limited trainable capacity for temporal modeling and quality assessment.

2.2 Multiple Instance Learning and Temporal Modeling

Multiple Instance Learning (MIL) models a video as a set of instances and learns aggregation under bag-level supervision [10, 11]. Attention-based MIL assigns weights to instances but often ignores temporal order. To address this, later studies introduced recurrent networks, temporal Transformers, and differentiable selection mechanisms to capture intra-bag dynamics [12, 13, 15]. Building on this line, our method incorporates a temporal selector together with instance quality modeling to jointly consider order and reliability.

2.3 Multimodal Fusion for Emotion Recognition

EEG and facial video are complementary modalities: EEG reflects neurophysiological responses, while video captures observable behavioral cues [1, 16]. Existing work has explored cross-modal attention, contrastive learning, and alignment strategies to enhance fusion [17, 18]. In contrast, our approach focuses on improving the quality of selected visual instances before fusion, which is particularly beneficial under resource constraints.

2.4 Robustness and Noise Suppression

Real-world EEG and video data are often noisy due to blur, occlusion, motion, and sensor instability. Robust learning methods typically rely on data augmentation or contrastive objectives [19–21]. Alternatively, instance-level reliability estimation provides a lightweight solution [18]. The proposed IQG module follows this idea by reweighting instances according to their quality, thereby suppressing unreliable inputs.

2.5 Position of QG-TIS

QG-TIS combines parameter-efficient adaptation, temporally-aware MIL, and instance quality gating within a unified framework. Staged freezing provides a lightweight backbone, BiGRU captures temporal dependencies, and IQG models instance reliability. Unlike methods that focus solely on temporal modeling or multimodal fusion, QG-TIS jointly considers temporal order, instance quality, and computational efficiency. The following experiments evaluate the contribution of each component through ablation studies and resource profiling.

3 Method

3.1 Overview

To address the demand for both computational efficiency and fine-grained temporal modeling in multimodal emotion recognition, QG-TIS, a quality-guided temporal instance selection framework, is proposed. The framework is designed for EEG–video emotion recognition and aims to preserve recognition performance while reducing the resource overhead introduced by backbone fine-tuning and instance-level modeling.

As illustrated in Figure 1, QG-TIS processes trial data consisting of synchronized EEG signals and facial video clips. The video is uniformly partitioned into N non-overlapping instances. The overall pipeline consists of three components: **(1)** a partially frozen Swin Transformer backbone for visual feature extraction; **(2)** a BiGRU-based selector together with an instance quality gate for quality-guided temporal instance selection; and **(3)** a compact cross-modal attention module for lightweight multimodal fusion. The following subsections describe each component and the implementation details.

3.2 Partially Frozen Swin Transformer Backbone

A Swin Transformer pretrained on large-scale visual data is adopted as the visual feature extractor. To retain representational capacity while reducing the number of trainable parameters and memory usage, a staged partial-freezing strategy is employed: the early stages of the backbone, which capture low-level and mid-level generic features, are frozen, and only the final stage and classification head are fine-tuned.

For an input frame $\mathbf{I}_t$, the backbone outputs a feature map $\mathbf{F}_t \in \mathbb{R}^{H\times W\times C}$, which is then transformed into a D-dimensional vector $\mathbf{f}_t \in \mathbb{R}^D$ via global average pooling. For a video instance containing T frames, a sequence of frame-level features $\{\mathbf{f}_1, \ldots, \mathbf{f}_T\}$ is obtained for subsequent temporal selection.

3.3 BiGRU-Based Temporal Instance Selector

To improve the temporal awareness of conventional static instance scoring, a two-layer bidirectional GRU (BiGRU) is introduced as the instance selector [22]. Given N video instances, temporal average pooling is first applied to each instance to obtain an initial instance representation $\mathbf{x}_i \in \mathbb{R}^D$, where $i = 1, \ldots, N$. Contextual relations among instances are then modeled with a BiGRU:

$$\mathbf{h}_i = \mathrm{BiGRU}(\mathbf{x}_i, \mathbf{h}_{i-1}; \Theta_{\mathrm{GRU}}), \tag{1}$$

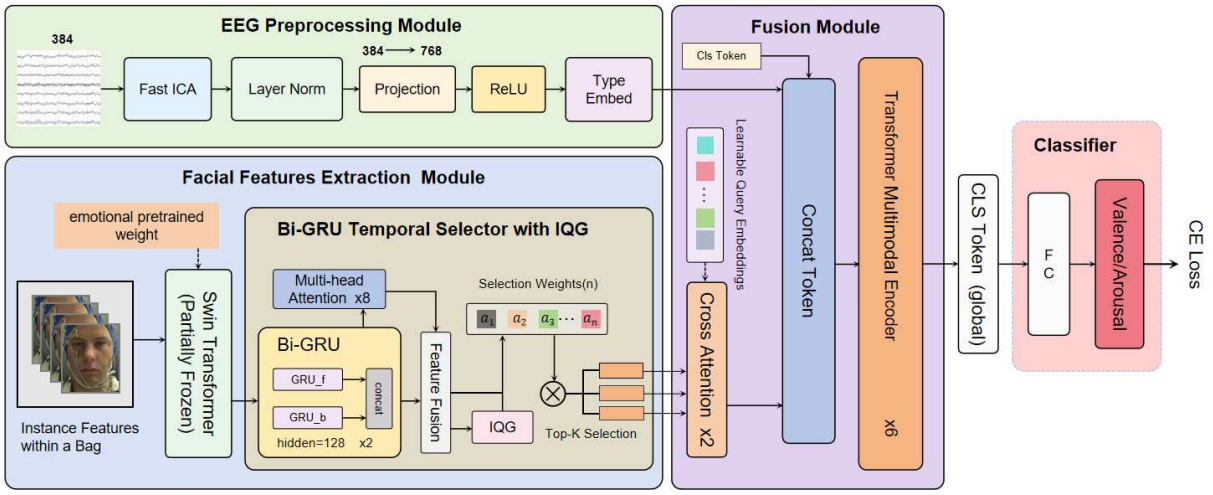

Fig. 1 Overall architecture of QG-TIS.

where $\mathbf{h}_i$ denotes the bidirectional contextual encoding. The initial importance score s_i is computed as

$$s_i = \mathbf{w}^\top \tanh(\mathbf{V}\mathbf{h}_i + \mathbf{b}). \tag{2}$$

During training, a Gumbel–Softmax relaxation with a learnable temperature τ is used to encourage exploration. During inference, the top k instances with the highest s_i are selected. Unless otherwise specified, $N = 10$ and $k = 3$ are used; the remaining hyperparameters are reported in the experimental section or supplementary material.

Compared with Top-k or attention-based selection on flattened sequences, the proposed selector first projects each frame into a low-dimensional space and then aggregates temporal information with BiGRU. This design avoids the extra cost of high-dimensional sequence projection, making it more lightweight in both parameter size and training memory.

3.4 Instance Quality Gate (IQG)

The selected instances may still contain weak noise or blurred segments. To further improve robustness, a lightweight instance quality gate (IQG) is designed. Taking the BiGRU hidden state $\mathbf{h}_i$ as input, IQG predicts a scalar quality score $q_i \in (0, 1)$ through a two-layer MLP:

$$q_i = \sigma(\mathrm{MLP}(\mathbf{h}_i; \Theta_{\mathrm{IQG}})). \tag{3}$$

The final importance score is modulated by

$$\tilde{s}_i = s_i \cdot \big(1 + \alpha \cdot q_i\big), \tag{4}$$

where α is a learnable scaling factor. This mechanism suppresses low-quality instances with minimal parameter overhead while enhancing the contribution of reliable instances.

3.5 EEG Encoding and Multimodal Fusion

EEG signals follow a standard preprocessing pipeline, including band-pass filtering, artifact removal via independent component analysis (ICA), and segmentation into time windows aligned with the video instances. Each EEG segment is represented as $\mathbf{E} \in \mathbb{R}^{C \times L}$ and encoded by a lightweight 1D convolutional encoder:

$$\mathbf{e} = \mathrm{ReLU}(\mathrm{Conv1D}(\mathbf{E})) \in \mathbb{R}^{D_e}. \tag{5}$$

The quality-reweighted visual instances $\{\tilde{\mathbf{h}}_1, \ldots, \tilde{\mathbf{h}}_k\}$, after linear projection into the same dimension, are fused with the EEG embedding $\mathbf{e}$ using a compact cross-modal attention module. The EEG feature is used as the query, and the visual instances are used as keys and values. The

attention output is concatenated with the EEG embedding and fed into a small Transformer encoder, followed by a linear classifier for emotion prediction.

3.6 Preventing Data Leakage

For any transformation based on dataset statistics or learned parameters, such as channel normalization or ICA/PCA fitting, the transformation is fitted only on the training split and then applied to the validation and test splits. Sliding-window or segmentation operations are performed after the data split so that no sample overlap occurs across partitions. These procedures are applied consistently in all cross-validation and leave-one-subject-out settings to avoid exposing test-set statistics during training.

3.7 Training Protocol and Implementation Details

Cross-entropy loss is used to train the differentiable modules end to end, while the frozen backbone is excluded from gradient updates to save memory. The optimizer is AdamW [23]: the initial learning rate for the fine-tuned Swin layers is set to 5×10^{-5}, whereas the remaining modules use an initial learning rate of 1×10^{-4}. The learning rate is decayed with cosine annealing over 100 epochs. Dropout with $p = 0.5$ is applied to the classification head and intermediate MLP layers as regularization [24]. Complete reproduction settings and measurement protocols are provided in the supplementary material.

To improve reproducibility and comparability, each experiment is independently repeated $N = 3$ times under different random seeds, and the results are reported as mean $\pm$ sample standard deviation. The seeds used in this study are 0, 42, and 2025. Because the number of repetitions is limited, comparisons of average performance are presented conservatively, and further validation with more independent runs is recommended.

3.8 Hardware and Measurement Protocol

All training and profiling experiments are conducted on an NVIDIA GeForce RTX 4060 Ti with 8 GB of GPU memory. Due to memory constraints, the maximum batch size during training is 14. Training time and memory measurements are reported under this batch size unless otherwise stated.

To ensure comparability of memory and time measurements, a unified protocol is adopted: for peak training memory, the maximum device memory usage during a complete training iteration is recorded, including forward propagation, loss computation, backward propagation, and optimizer.step(). Each configuration is measured multiple times and reported as the mean with sample standard deviation. Iteration time is obtained by averaging over multiple synchronized iterations. The table captions explicitly indicate whether data-loading overhead is included.

3.9 Parameter Counting and Reproducibility Notes

The "number of trainable parameters" reported in this paper is defined as the total number of parameters participating in gradient computation. In PyTorch, this corresponds to parameters whose `requires_grad` attribute is `True`. If a module-wise parameter breakdown or measurement logs are required during peer review, they can be provided upon request.

4 Experiments and Results

This section provides a comprehensive evaluation of the proposed QG-TIS framework. We first describe the dataset and preprocessing pipeline, and then compare QG-TIS with several representative methods on the DEAP benchmark (DEAP-4). To examine the contribution of each component, we further present ablation studies. In addition, reproducible profiling measurements are reported to quantify resource efficiency and to show that the method is feasible on a consumer-grade GPU. Finally, qualitative visualizations are provided to offer an intuitive understanding of the model's decisions.

4.1 Dataset and Preprocessing

The evaluation is conducted on the widely used DEAP dataset [1]. The dataset contains multimodal recordings from 32 participants who watched 40 one-minute music videos, including synchronized 32-channel EEG signals and facial

videos. Participants provided self-reported ratings of affective dimensions such as valence and arousal on a 9-point Likert scale.

Following standard practice, we focus on valence and arousal and adopt the standard four-class formulation (DEAP-4), in which each dimension is binarized with a threshold of 5 to form four emotional quadrants. EEG preprocessing includes band-pass filtering, artifact removal via independent component analysis (ICA) using MNE-Python [25, 26], and segmentation of each trial into non-overlapping 3-second windows followed by channel-wise normalization. Facial video frames are extracted at 10 fps and resized to 224×224 pixels, with face alignment performed using MediaPipe. Each 3-second EEG window is aligned with 10 consecutive video frames to form a bag of visual instances. All preprocessing steps that require data-driven fitting, such as mean/variance normalization and ICA/PCA decomposition, are fit only on the training split and then applied to the validation and test splits to avoid data leakage (see Section 3).

At the backbone level, after applying the partial-freezing strategy to the Swin Transformer, the total number of parameters is 27,519,354, of which 14,183,856 are trainable (51.54%). Compared with full fine-tuning, this reduces the number of trainable parameters by 48.46%. This setting preserves the representational capacity of the backbone while lowering optimization overhead.

All experiments were conducted on an NVIDIA GeForce RTX 4060 Ti GPU with 8 GB of memory, and the models were implemented in PyTorch. The batch size was set to 14. To report repeatability, each experiment was independently run under $N = 3$ different random seeds (0, 42, and 2025), and results are reported as mean $\pm$ sample standard deviation. Because the number of repetitions is limited, comparisons of average performance are presented conservatively, and further validation with more independent runs is recommended.

4.2 Comparison with State-of-the-Art Methods

Table 1 lists the results of QG-TIS alongside several representative methods on the DEAP-4 task. In the table, reproduced indicates results reproduced under the same subject-dependent protocol and preprocessing as used in this paper, reported as mean $\pm$ std over $N = 3$ runs; paper-reported indicates point estimates taken from the original papers, which may use different protocols or preprocessing and are included for reference only.

4.3 Ablation Study

To verify the contribution of each component, we conducted systematic ablation experiments under the subject-dependent protocol. Each variant was run independently $N = 3$ times (0, 42, and 2025), and the results are reported as mean $\pm$ sample standard deviation. Table 2 summarizes accuracy, macro F1, and the number of trainable parameters.

The parameter count refers to the total number of parameters that require updating in the whole model and is computed using the same trainable-parameter counting criterion throughout. It should be noted that the parameter counts of the baseline variants reflect the actual statistics obtained from our reproduction based on public code and may be higher than those of our method because of implementation details and projection-layer design. The results show that both IQG and BiGRU bring performance gains, and the combination of the two yields the best result.

4.4 Efficiency and Resource Observations

The resource efficiency results indicate that peak memory is determined more by intermediate activations than by the number of trainable parameters. In particular, the BiGRU-based selector compresses the video sequence earlier and reduces the activations retained during training. Under the same hardware, batch size, and input resolution, QG-TIS therefore achieves a better efficiency profile while preserving strong recognition performance.

4.5 Qualitative Analysis and Visualization

To illustrate the model's discriminative behavior and instance selection mechanism, we first present the confusion matrix of QG-TIS on the DEAP-4 test set and then show the selection results for a representative trial.

Table 1 Comparison on the DEAP-4 classification task (mean ± std, $N = 3$).

Method	Modality	Setting	Acc (%)	Macro F1 (%)
Unimodal methods				
Gupta et al. (2018) [27]	EEG	paper-reported	71.43 ± 5.29	–
Marjit et al. (2021) [28]	EEG	paper-reported	81.25 ± 5.47	78.56 ± 8.15
Multimodal methods				
Suhaimi et al. (2020) [29]	EEG (survey)	paper-reported	54.22 ± 2.50	31.00 ± 5.71
Alotaibi and Vuyyuru (2023) [30]	EEG+Face	paper-reported	53.87 ± 2.85	–
Lee et al. (2024) [31]	EEG+Face	paper-reported	83.20 ± 3.52	84.10 ± 2.62
Muhammad et al. (2023) [32]	EEG+Face	paper-reported	85.40 ± 2.15	85.10 ± 2.45
This work				
Partial freeze	EEG+Face	reproduced	94.42 ± 0.38	94.25 ± 0.37
Partial freeze + IQG	EEG+Face	reproduced	95.13 ± 0.46	95.00 ± 0.48
Partial freeze + BiGRU	EEG+Face	reproduced	95.34 ± 0.17	95.20 ± 0.28
QG-TIS	EEG+Face	reproduced	**95.79 ± 0.22**	**95.69 ± 0.28**

Table 2 Ablation results on DEAP-4 ($N = 3$).

Variant	Acc (%)	Macro F1 (%)	Params (M)
Partial freeze	94.42% ± 0.38%	94.25% ± 0.37%	1445.1
Partial freeze + IQG	95.13% ± 0.46%	95.00% ± 0.48%	1445.2
Partial freeze + BiGRU	95.34% ± 0.17%	95.20% ± 0.28%	28.5
QG-TIS (BiGRU + IQG)	**95.79% ± 0.22%**	**95.69% ± 0.28%**	**29.5**

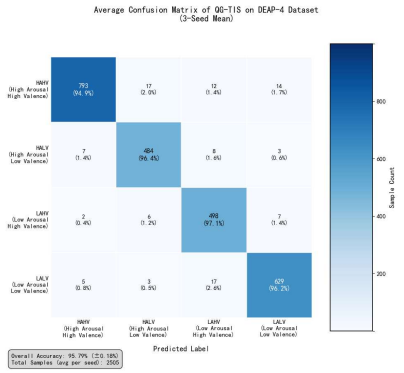

Fig. 2 Confusion matrix of QG-TIS on the DEAP-4 classification task.

The confusion matrix shows that the predictions are concentrated along the main diagonal, indicating that the model can distinguish the four emotion classes in a stable manner. A small number of misclassifications occur mainly in samples with relatively ambiguous emotional boundaries, but the overall effect is limited and does not change the model's overall advantage on the DEAP-4 task.

To further inspect the temporal selection behavior, Figure 3 shows the instance selection process for a representative trial: the top row presents the full video sequence, and the bottom row shows the top three instances selected by BiGRU together with their corresponding IQG scores.

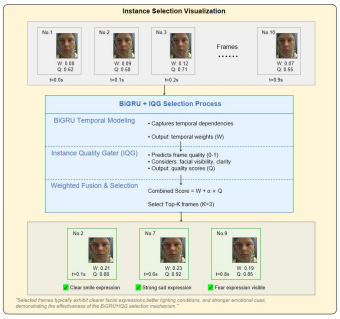

Fig. 3 Visualization of temporal instance selection (example).

As shown in the example, IQG can effectively suppress low-quality instances, such as blurred or heavily occluded segments; these segments receive noticeably lower scores than the selected high-quality ones. This result suggests that instance quality gating helps improve selection reliability and enhances the model's robustness to noisy samples.

5 Conclusion

5.1 Interpretation of Model Behavior

QG-TIS suggests that temporal modeling can partly compensate for the rigidity introduced by freezing a pretrained visual backbone. As shown in Table 2, adding the BiGRU selector improves performance over the partial-freeze baseline, indicating that task adaptation is shifted toward higher-level temporal reasoning. IQG provides an additional robustness mechanism by reweighting instances according to their reliability, which helps suppress blur and occlusion. However, its effectiveness depends on sufficient quality variation in the training data.

5.2 Efficiency Analysis

QG-TIS achieves a better balance between recognition performance and computational cost. The BiGRU-based temporal selector and IQG compress the sequence earlier, which reduces retained activations and improves memory efficiency without sacrificing accuracy.

5.3 Practical Impact and Summary

Under the subject-dependent protocol on DEAP, QG-TIS achieves $\mathbf{95.79\% \pm 0.22\%}$ accuracy and $\mathbf{95.69\% \pm 0.28\%}$ macro-F1, while partial freezing reduces trainable Swin backbone parameters by **48.46%** compared with full fine-tuning. The method is suitable for resource-constrained settings such as edge devices, federated learning, and low-cost experimental platforms. The selected-instance count k provides a simple knob for trading focus and robustness.

Current limitations include the small number of runs ($N = 3$), single-dataset evaluation on DEAP, and unverified cross-dataset generalization. Future work will increase repetitions, evaluate on AMIGOS and SEED, extend IQG to cross-modal reliability estimation, add other physiological modalities, and combine inference-time compression techniques such as quantization and pruning.

Overall, QG-TIS treats efficiency as a design constraint and shifts adaptation capacity from generic feature extraction toward targeted temporal and quality reasoning, achieving competitive performance with fewer trainable parameters.

Statements and Declarations. Funding. This work was supported by Fujian Provincial Natural Science Foundation of China under Grant No. 2024J01818 and the 2024 Annual Regular Project of Fujian Provincial Educational Science Planning Grant No. FJJKBK24-137.

Competing interests. The authors declare that they have no competing interests.

Author Contributions. All authors contributed to the study conception and design. Juecheng Cai and Zhoufu Chen performed material preparation, data collection, and analysis. The first draft was written by Juecheng Cai, and all authors reviewed and approved the final manuscript.

Data Availability. The DEAP dataset is publicly available at http://www.eecs.qmul.ac.uk/mmv/datasets/deap/.

Code Availability. The code will be made available upon reasonable request.

AI Declaration. During preparation of this work, ChatGPT was used for English polishing and wording refinement. The authors reviewed and edited the content and take full responsibility for the final manuscript.

References

[1] Koelstra, S., Mühl, C., Soleymani, M., Lee, J.-S., Yazdani, A., Ebrahimi, T., Pun, T., Nijholt, A., Patras, I.: Deap: A database for emotion analysis using physiological signals. IEEE Transactions on Affective Computing **3**(1), 18–31 (2012) https://doi.org/10.1109/T-AFFC.2011.15

[2] Pan, B., Hirota, K., Jia, Z., Dai, Y.: A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods. Neurocomputing

561, 126866 (2023) https://doi.org/10.1016/j.neucom.2023.126866

[3] Poria, S., Cambria, E., Bajpai, R., Hussain, A.: A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion **37**, 98–125 (2017)

[4] Baltrušaitis, T., Ahuja, C., Morency, L.-P.: Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(2), 423–443 (2019)

[5] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 10000–10012 (2021)

[6] Han, K., Wang, Y., Chen, H., Chen, X., Guo, J., Liu, Z., Tang, Y., Xiao, A., Xu, C., Xu, Y., Yang, Z., Zhang, Y., Tao, D.: A survey on vision transformer. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(1), 87–110 (2023)

[7] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for NLP. In: Proceedings of the 36th International Conference on Machine Learning (ICML). PMLR, vol. 97, pp. 2790–2799 (2019)

[8] Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Chen, W.: LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685 (2021)

[9] Xin, Y., Luo, S., Zhou, H., Du, J., Liu, X., Fan, Y., Li, Q., Du, Y.: Parameter-Efficient Fine-Tuning for Pre-Trained Vision Models: A Survey. arXiv:2402.02242 (2024)

[10] Ilse, M., Tomczak, J.M., Welling, M.: Attention-based deep multiple instance learning. In: Proceedings of the 35th International Conference on Machine Learning (ICML). PMLR, vol. 80, pp. 2132–2141 (2018)

[11] Romeo, L., Cavallo, A., Pepa, L., Bianchi-Berthouze, N., Pontil, M.: Multiple instance learning for emotion recognition using physiological signals. IEEE Transactions on Affective Computing **13**(1), 389–407 (2022)

[12] Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: A video vision transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 6836–6846 (2021)

[13] Jang, E., Gu, S., Poole, B.: Categorical reparameterization with gumbel-softmax. In: International Conference on Learning Representations (ICLR) (2017)

[14] He, H., Cai, J., Zhang, J., Tao, D., Zhuang, B.: Sensitivity-aware visual parameter-efficient fine-tuning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11825–11835 (2023)

[15] Vielzeuf, V., Pateux, S., Jurie, F.: Temporal multimodal fusion for video emotion classification in the wild. In: Proceedings of the 19th ACM International Conference on Multimodal Interaction (ICMI), pp. 569–576 (2017)

[16] Zheng, W.-L., Lu, B.-L.: Investigating critical frequency bands and channels for EEG-based emotion recognition with deep neural networks. IEEE Transactions on Autonomous Mental Development **7**(3), 162–175 (2015)

[17] Huang, Y., Yang, J., Liao, P., Pan, J.: Fusion of facial expressions and eeg for multimodal emotion recognition. Computational Intelligence and Neuroscience **2017**, 2107451 (2017)

[18] Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Advances in Neural Information Processing Systems (NeurIPS) (2020)

[19] Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive

learning of visual representations. In: Proceedings of the 37th International Conference on Machine Learning (ICML). PMLR, vol. 119, pp. 1597–1607 (2020)

[20] He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning (MoCo). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 9729–9738 (2020)

[21] Grill, J.-B., Strub, F., Altché, C., Tallec, C., Richemond, P.H., Buchatskaya, A., Doersch, C., Pires, B., Guo, Z., Azar, M., *et al.*: Bootstrap your own latent: A new approach to self-supervised learning (BYOL). In: Advances in Neural Information Processing Systems (NeurIPS) (2020)

[22] Cho, K., Merriënboer, B., Gülcühre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. arXiv:1406.1078 (2014)

[23] Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. arXiv:1711.05101 (2017)

[24] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: A simple way to prevent neural networks from overfitting. Journal of Machine Learning Research **15**, 1929–1958 (2014)

[25] Gramfort, A., Luessi, M., Larson, E., Engemann, D.A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., *et al.*: MEG and EEG data analysis with MNE-Python. Frontiers in Neuroscience **7**, 267 (2013)

[26] Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. Neural Networks **13**(4-5), 411–430 (2000)

[27] Gupta, V., Chopda, M.D., Pachori, R.B.: Cross-subject emotion recognition using flexible analytic wavelet transform from EEG signals. IEEE Sensors Journal **19**(6), 2266–2274 (2018)

[28] Marjit, S., Talukdar, U., Hazarika, S.M.: EEG-based emotion recognition using genetic algorithm optimized multi-layer perceptron. In: 2021 International Symposium of Asian Control Association on Intelligent Robotics and Industrial Automation (IRIA), pp. 304–309 (2021)

[29] Suhaimi, N.S., Mountstephens, J., Teo, J.: EEG-based emotion recognition: A state-of-the-art review of current trends and opportunities. Computational Intelligence and Neuroscience **2020**, 1–19 (2020)

[30] Alotaibi, Y., Vuyyuru, V.A.: Electroencephalogram based face emotion recognition using multimodal fusion and 1-d convolution neural network (ID-CNN) classifier. AIMS Mathematics **8**(10), 22984–23002 (2023)

[31] Lee, J.-H., Kim, J.-Y., Kim, H.-G.: Contrastive learning for multimodal emotion recognition with EEG and video. IEEE Transactions on Affective Computing **15**(1), 123–135 (2024)

[32] Muhammad, F., Hussain, M., Aboalsamh, H.: A bimodal emotion recognition approach through the fusion of electroencephalography and facial sequences. Diagnostics **13**(5), 977 (2023)

Figures

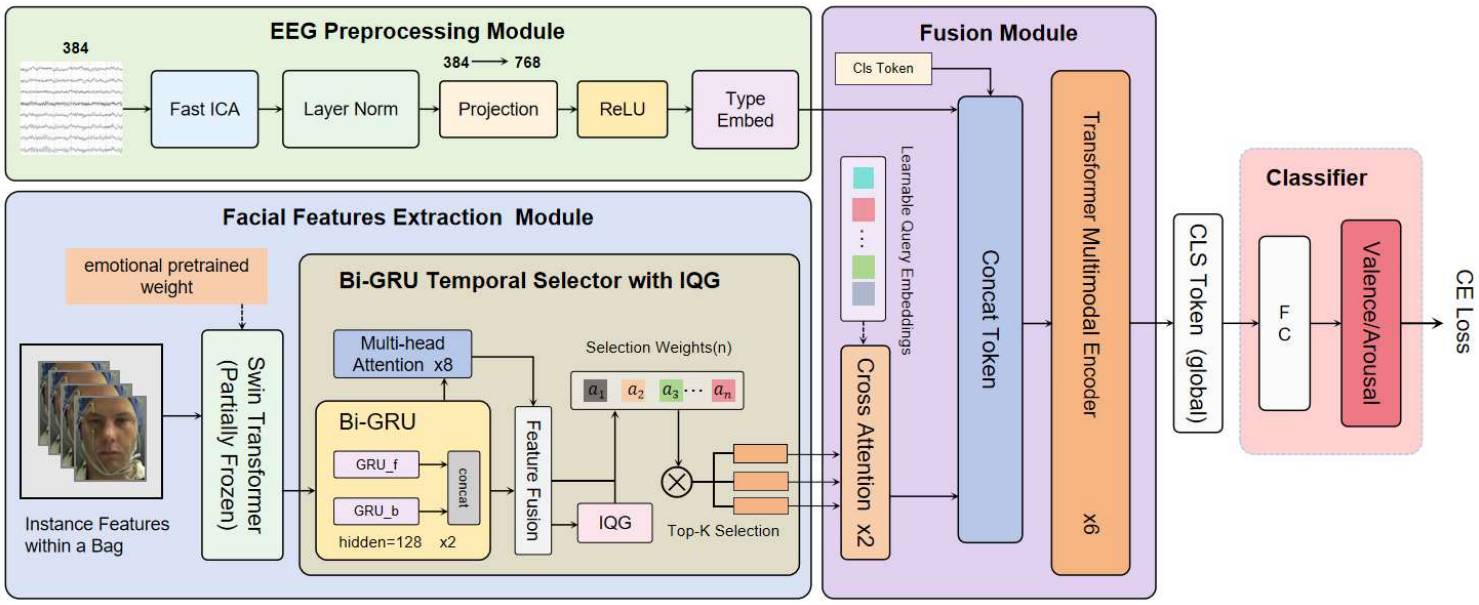

Figure 1

Overall architecture of QG-TIS.

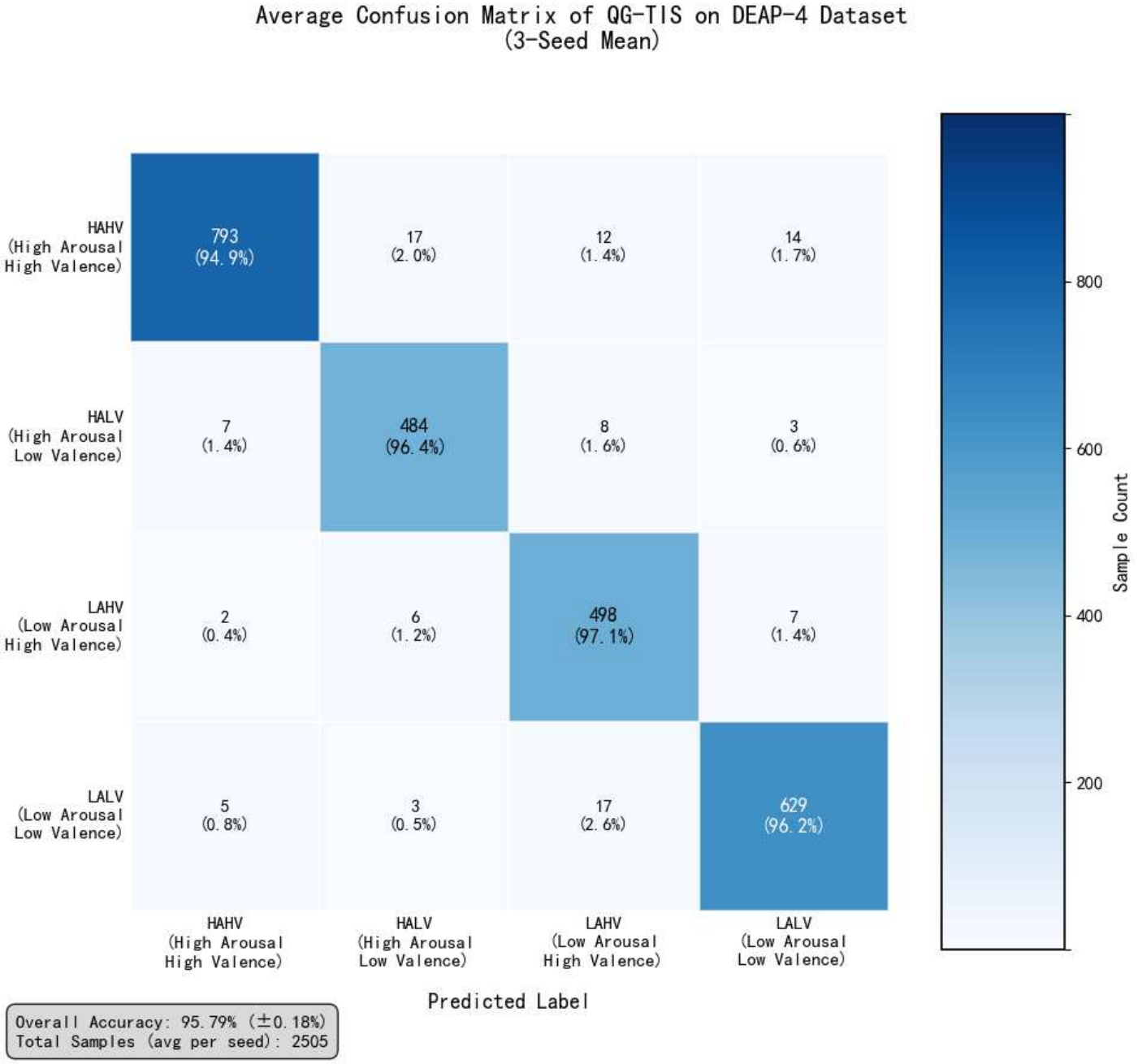

Figure 2

Confusion matrix of QG-TIS on the DEAP-4 classification task.

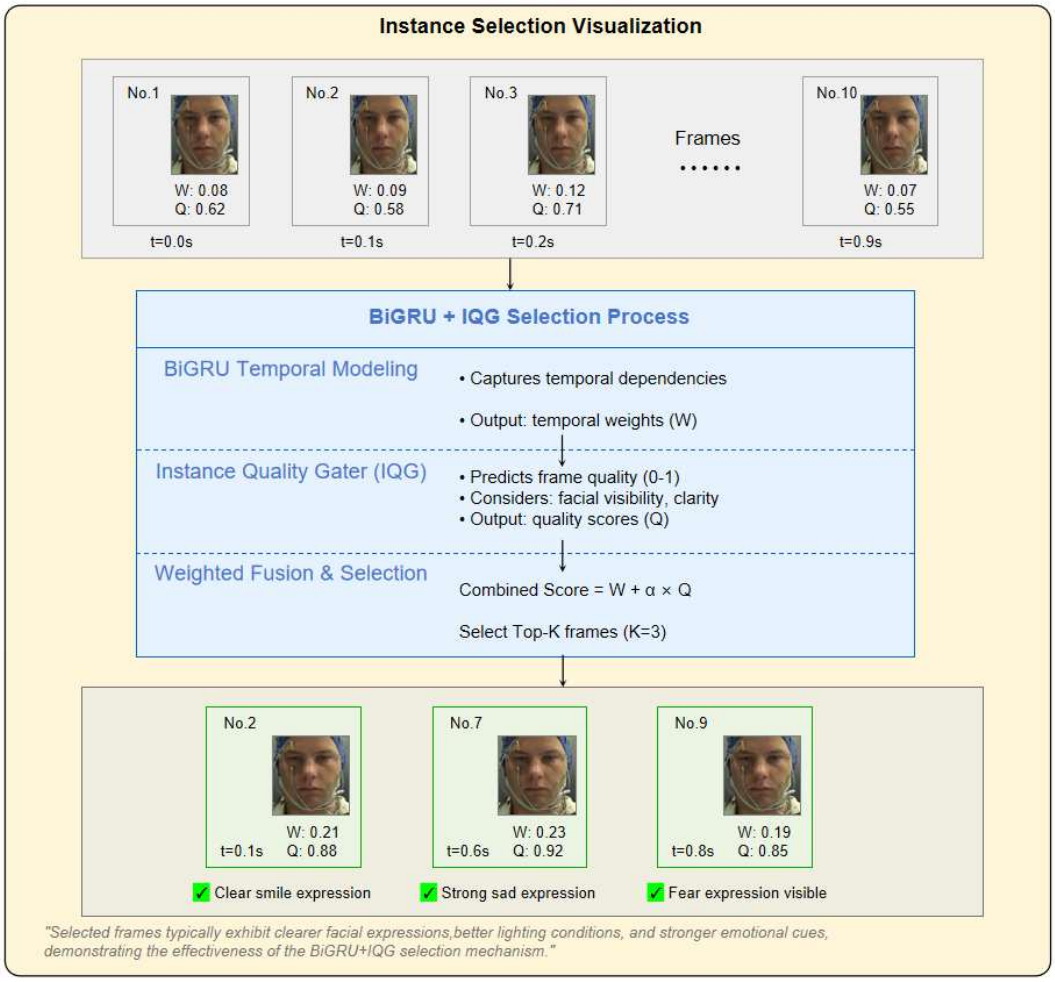

Figure 3

Visualization of temporal instance selection (example).