From Faces to Behavior Analysis: Adaptive Multimodal Fusion for Valence-Arousal Estimation in-the-Wild

Elena Ryumina, Maxim Markitantov, Alexandr Axyonov, Dmitry Ryumin, Mikhail Dolgushin, Denis Dresvyanskiy, Alexey Karpov

ryumina.e@iias.spb.su, markitantov.m@iias.spb.su, axyonov.a@iias.spb.su, ryumin.d@iias.spb.su, dolgushin.m@iias.spb.su, denisdresvyanskiy@gmail.com, karpov@iias.spb.su

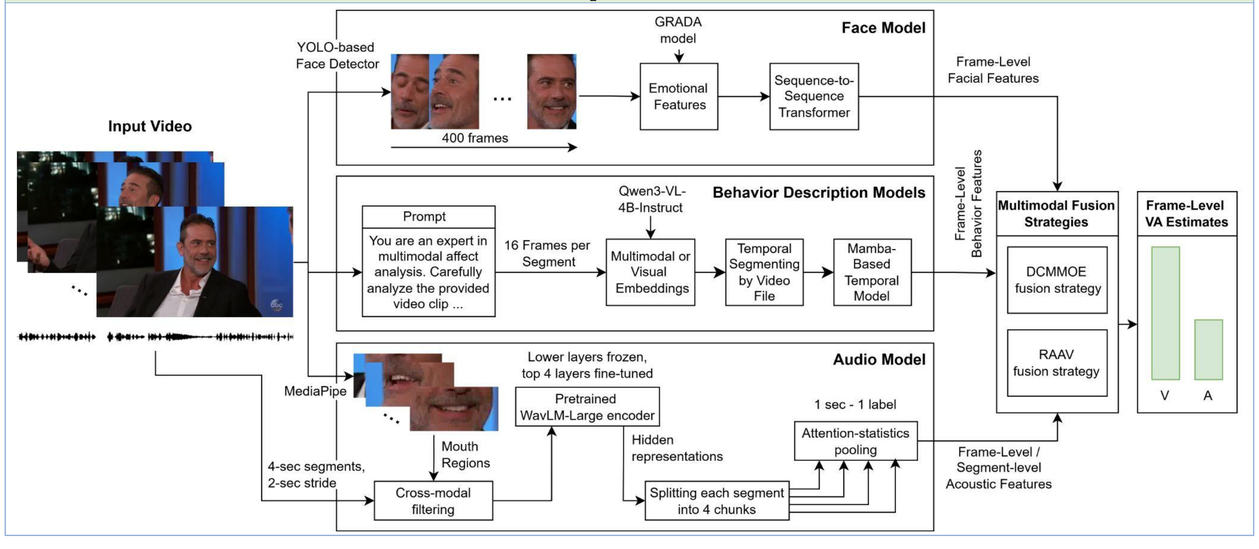
Introduction

- Continuous valence-arousal VA estimation provides flexible emotion representation: valence describes pleasantness, while arousal reflects emotional intensity – however, modeling continuous affect under ITW conditions remains challenging.
- VLM can capture facial expressions, head movements, gestures, posture, and scene context, which may provide complementary information for affect estimation.
- **Problem:** existing VA estimation approaches mainly rely on facial and acoustic features, while the use of multimodal VLM for extracting behavior-related and contextual affective cues remains underexplored.
- **Goal:** improving continuous VA estimation through multimodal fusion of facial dynamics, behavior-related VLM features, and acoustic information.

Proposed Methodology

- Three dedicated unimodal branches are trained independently: a GRADA-based face model, a Qwen3-VL behavior-description model, and a WavLM-based acoustic model.
- Each branch extracts complementary affective representations at a modality-specific temporal resolution: frame-level face features, segment-level behavior embeddings, and 4-second audio features.
- Modality-specific temporal models transform these features into continuous VA estimates or latent affective representations.
- Segment-level behavior and audio predictions are temporally aligned to frame-level resolution using interval assignment and overlap averaging.
- Directed Cross-Modal Mixture-of-Experts (DCMMOE) fusion stratedgy models directed cross-modal interactions between modality pairs with adaptive frame-wise expert gating.
- Reliability-Aware Audio-Visual (RAAV) fusion stratedgy performs reliability-aware frame-level fusion of face and behavior features, using audio as auxiliary context.

Pipeline



Research Corpus and Fusion Setup

Official 10th VA ABAW split [1]:
356 training · 76 development · 162 test videos

Evaluation metric:
Mean CCC of valence and arousal

Fusion search:
Grid search over temporal length, stride, Transformer depth, attention heads, Mamba configuration, and fusion hyperparameters

DCMMOE fusion:
400 frames temporal context · 150 segment stride · 5 cross-attention layers · 16 attention heads

RAAV fusion:
Hidden size 256 · 4 learnable latent audio tokens · 1 Transformer layer · 4 heads · dropout 0.1

RAAV reliability priors:
Face 0.85 · Qwen3 multimodal 0.15

Experimental Results on ABAW'26 VA Challenge

Configuration	Modalities	Public devel.			Private test		
		Valence	Arousal	Avg.	Valence	Arousal	Avg.
1. Face model. GRADA + Transformer	V	0.5869	0.6508	0.6189	0.5155	0.5696	0.5425
2. Visual Behavior Description Model. Qwen3 + Mamba	V	0.2499	0.5515	0.4007	0.3162	0.5221	0.4192
3. Multimodal Behavior Description Model. Qwen3 + Mamba	M	0.4290	0.6480	0.5385	0.5102	0.6075	0.5589
4. Audio model. WavLM + Chunk-Wise Pooling	A	0.3424	0.4644	0.4034	0.2364	0.3416	0.2890
5. DCMMOE (1, 3, 4 frame-level)	V, M, A	**0.6100**	0.6875	0.6487	0.5733	**0.6371**	0.6052
6. RAAV (1, 3, 4 segment-level)	V, M, A	0.6078	**0.7073**	**0.6576**	**0.6084**	0.6351	**0.6217**
7. RAAV (1, 2, 3, 4 segment-level)	V, M, A	0.4527	0.6871	0.5699	0.4794	0.6547	0.5671

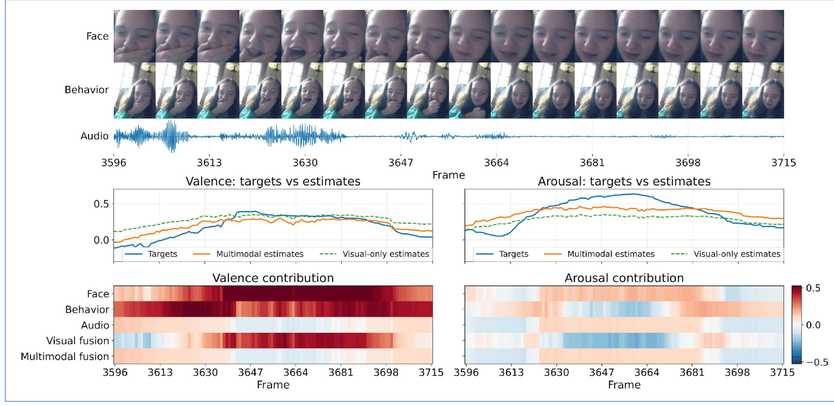
CCC performance.

Approach	Modalities	Private test
Ours [2]	A+V	**0.622**
Lee et al. [3]	A+V	0.580
Gong et al.[4]	A+V	0.530
Savchenko and Tsypliakova [5]	V	0.520

Key findings

- Multimodal Qwen3 behavior features outperform visual-only behavior features on the private test set.
- RAAV provides the best overall fusion performance, surpassing all unimodal branches and DCMMOE.
- The proposed system achieves the best private test performance in the 10th ABAW VA Challenge.

Comparison: Target vs Estimates + Contribution maps



- The multimodal model follows target VA dynamics better than the visual-only model, especially for arousal.
- Valence predictions become more stable and better aligned with the target trajectory.
- Contribution maps show complementary modality roles: face/behavior improve valence, while audio improves arousal.
- Positive fusion gains indicate that improvement comes from multimodal integration, not a single dominant modality.

Conclusions

- We won the ABAW'26 VA Challenge.
- Multimodal fusion consistently outperforms unimodal models, confirming the importance of combining facial, behavior-related, and acoustic cues for continuous VA estimation.
- The Qwen3-based multimodal behavior model outperforms the visual-only behavior model, showing the value of VLM-based behavior-related features.
- Overall, adaptive multimodal fusion with behavior-related VLM features is effective for VA estimation under in-the-wild conditions.

References

[1] Kollias and Zafeiriou. Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and arcface. ArXiv. 2019.
[2] Ryumina et at. Team RAS in 10th ABAW Competition: Multimodal Valence and Arousal Estimation Approach. ArXiv 2026.
[3] Lee et al. Stage-adaptive reliability modeling for continuous valence-arousal estimation. ArXiv 2026.
[4] Gong et al. Conflict-aware multimodal fusion for ambivalence and hesitancy recognition. ArXiv 2026.
[5] Savchenko and Tsypliakova. Hsemotion team at abaw-10 competition: Facial expression recognition, valence-arousal estimation, action unit detection and fine-grained violence classification. ArXiv 2026.