Distance-aware Soft Prompt Guidance for Multimodal Valence-Arousal Estimation

Byeongjin Jung, Chanyeong Park, Sejoon Lim
Kookmin University, Republic Of Korea

CVPR JUNE 3-7, 2026 DENVER COLORADO

Abstract

Task & Challenge
- Valence-arousal (VA) estimation models human emotions as continuous values to capture subtle emotional changes in naturalistic environments. However, applying CLIP to VA regression is challenging because text prompts are discrete, while VA labels are continuous.

Proposed Method
- We propose **Distance-aware Soft Prompt Guidance**, which bridges semantic text space and continuous VA space. The VA space is divided into multiple emotional regions represented by textual descriptions.

Semantic Soft Labeling
- Instead of hard region labels, Gaussian kernel-based soft labels are generated from the distance between ground-truth VA coordinates and region centers, enabling fine-grained emotional transition learning.

Multimodal Framework & Results
- Visual and audio features are extracted using **CLIP** and **AST**, modeled with **GRUs**, and fused through **cross-modal attention** and **gated fusion**. Experiments on **Aff-Wild2** show improved performance over the official baseline.

Challenge Overview: ABAW 2026 VA Estimation

ABAW Workshop & Competition
The 10th ABAW Workshop and Competition is held with **CVPR 2026** and focuses on multimodal analysis of human affect and behavior in real-world, unconstrained environments.[1]

Valence-Arousal Estimation Challenge
This work participated in the **Valence-Arousal (VA) Estimation Challenge**, which predicts continuous valence and arousal values from in-the-wild audiovisual data.[1]

Dataset
The challenge uses an augmented **Aff-Wild2 database**, containing **594 videos**, approximately **3M frames**, and **584 subjects** with valence-arousal annotations.[1], [2]

Evaluation Metric
Performance is evaluated using the mean Concordance Correlation Coefficient ($\boldsymbol{CCC}$)[3] of valence and arousal. The official baseline reports $\boldsymbol{CCC_V}$ = 0.24, $\boldsymbol{CCC_A}$ = 0.20, and $\boldsymbol{P}$ = 0.22.

Method

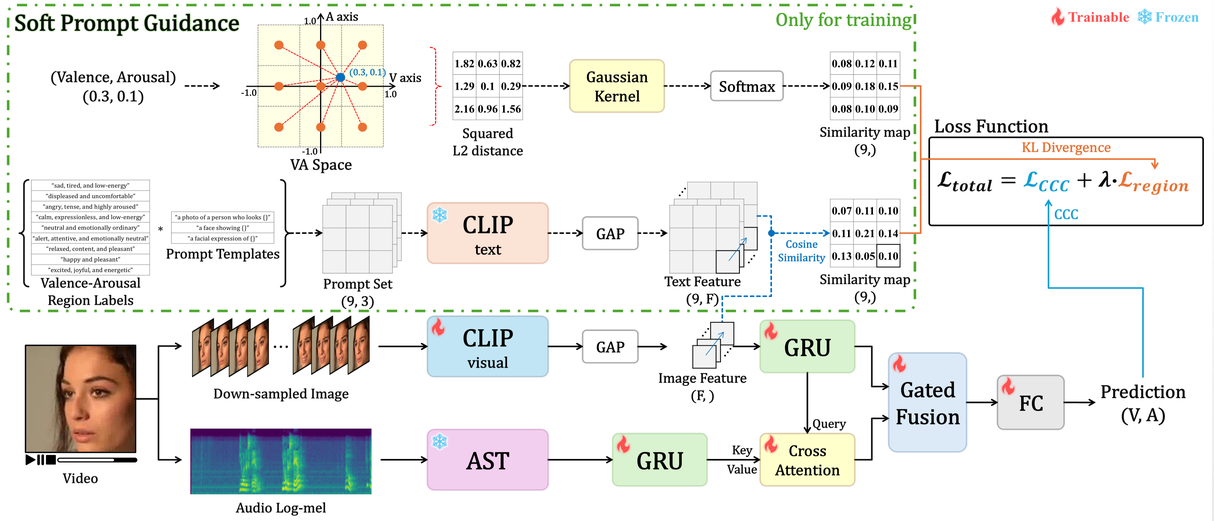

Table 1. **Framework Overall:** The proposed framework consists of four stages: multimodal feature extraction, distance-aware soft prompt guidance, temporal modeling, and hierarchical multimodal fusion for continuous VA prediction.

- **Multimodal Feature Extraction**
  Visual features are extracted using CLIP ViT-B/16[4], while audio features are extracted from Log-Mel spectrograms using AST[5]. Both modalities are projected into a shared 256-dimensional embedding space.
- **Distance-aware Soft Prompt Guidance**
  The continuous VA space is divided into 3 × 3 semantic emotion regions. Each region is represented by text prompts encoded with the CLIP text encoder[4]. Gaussian kernel-based soft labels are generated from the distance between ground-truth VA coordinates and region centers.
- **Temporal Modeling**
  Visual and audio feature sequences are modeled using bidirectional GRUs[6] to capture temporal emotional dynamics across video segments.
- **Hierarchical Multimodal Fusion**
  Cross-modal attention[7] aligns visual and audio features, and gated fusion[8] adaptively integrates complementary audio information while preserving reliable visual cues.
- **Training Objective**
  The model is trained with a joint objective combining $\boldsymbol{CCC}$ loss[3] for VA regression and KL divergence-based semantic region loss for soft prompt guidance.

Results & Discussion

Configuration	λ	CCC_{Mean}
CLIP Frozen	-	0.4438
CLIP Fine-tuned	-	0.4698
CLIP Fine-tuned	0.1	0.4768
Ours(CLIP Fine-tuned)	**0.2**	**0.4897**
CLIP Fine-tuned	0.3	0.4876

Table 1. Effect of semantic region loss weight λ on VA estimation performance

- Applying semantic region loss improves performance over simple CLIP fine-tuning, with the best result achieved at λ = 0.2. This indicates that semantic region supervision effectively complements continuous VA regression.

Teams	CCC_{Mean}
RAS	0.62
EmoDX	0.58
Ours	**0.53**
HSEmotion	0.52
Baseline	0.22

Table 2. Overall Performance

- The proposed method achieved $\boldsymbol{CCC_{Mean}}$ 0.53 on the Aff-Wild2 test set and ranked 3rd place in the ABAW competition, largely outperforming the official baseline.

Conclusions

Contributions
- We propose Distance-aware Soft Prompt Guidance to bridge discrete CLIP text prompts and continuous VA regression.
- We design a multimodal VA estimation framework using CLIP, AST, GRU-based temporal modeling, cross-modal attention, and gated fusion.

Results
- The best validation performance is achieved with GRU + Attention + Gated Fusion, reaching $\boldsymbol{CCC_{Mean}}$ 0.5361.
- Our method achieves $\boldsymbol{CCC_{Mean}}$ 0.53 on the Aff-Wild2 test set and ranks 3rd place in the ABAW competition.

Contact

Name: Byeongjin Jung
Organization: Intelligent Mobility Lab, Kookmin University(Republic Of Korea)
Email: wjdqudwls05@kookmin.ac.kr
Website: lim@kookmin.ac.kr
Phone: +82-10-4326-4396

References

1. 10th Affective & Behavior Analysis in-the-Wild (ABAW) Workshop and Competition, CVPR 2026.
2. D. Kollias and S. Zafeiriou, "Expression, Affect, Action Unit Recognition: Aff-Wild2, Multi-Task Learning and ArcFace," arXiv preprint arXiv:1910.04855, 2019.
3. I. Lawrence and K. Lin, "A Concordance Correlation Coefficient to Evaluate Reproducibility," Biometrics, pp. 255–268, 1989.
4. A. Radford et al., "Learning Transferable Visual Models From Natural Language Supervision," ICML, 2021.
5. Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," arXiv preprint arXiv:2104.01778, 2021.
6. J. Chung et al., "Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling," arXiv preprint arXiv:1412.3555, 2014.
7. A. Vaswani et al., "Attention Is All You Need," NeurIPS, 2017.
8. J. Arevalo et al., "Gated Multimodal Units for Information Fusion," arXiv preprint arXiv:1702.01992, 2017.