Anchoring Emotions in Text: Robust Multimodal Fusion for Mimicry IntensityEstimation

Lingsi Zhu Yuefeng Zou Yunxiang Zhang Naixiang Zheng Guoyuan Wang Jun Yu1* Jiaen Liang Wei Huang Shengping Liu Ximin Zheng

University of Science and Technology of China 、 Unisound AI Technology Co., Ltd. 、 Pingan Technology Co., Ltd.

Abstract

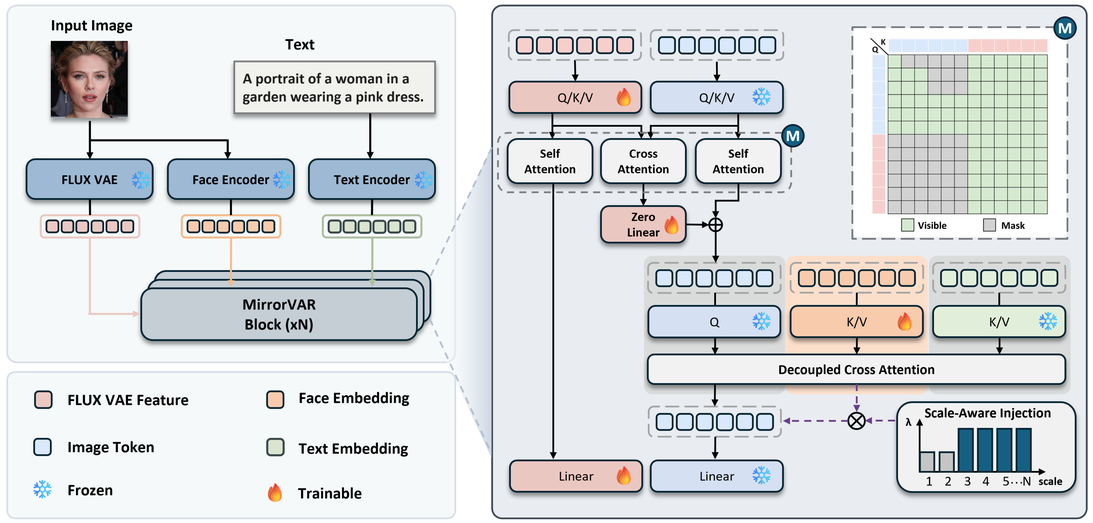
Estimating Emotional Mimicry Intensity (EMI) in naturalistic environments is a critical yet challenging task in affective computing. The primary difficulty lies in effectively modeling the complex, nonlinear temporal dynamics across highly heterogeneous modalities, especially when physical signals are corrupted or missing.

To tackle this, we propose TAEMI (Text-Anchored Emotional Mimicry Intensity estimation), a novel multimodal framework designed for the 10th ABAW Competition. Motivated by the observation that continuous visual and acoustic signals are highly susceptible to transient environmental noise, we break the traditional symmetric fusion paradigm. Instead, we leverage textual transcripts—which inherently encode a stable, time-independent semantic prior—as central anchors.

Specifically, we introduce a Text-Anchored Dual Cross-Attention mechanism that utilizes these robust textual queries to actively filter out frame-level redundancies and align the noisy physical streams. Furthermore, to prevent catastrophic performance degradation caused by inevitably missing data in unconstrained real-world scenarios, we integrate Learnable Missing-Modality Tokens and a Modality Dropout strategy during training. Extensive experiments on the Hume-Vidmimic2 dataset demonstrate that TAEMI effectively captures fine-grained emotional variations and maintains robust predictive resilience under imperfect conditions.

Our framework achieves a state-of-the-art mean Pearson correlation coefficient across six continuous emotional dimensions, significantly outperforming existing baseline methods.

Methodology



Detail

The framework processes temporally aligned audio, visual, and textual modalities.

Following modality-specific feature extraction and linear alignment, the temporal representations are fed into a Text-Anchored Dual Cross-Attention module. Here, the textual features serve as queries to aggregate pertinent audio and visual contexts.

The resulting fused mul-timodal representation is then processed by a Multi-Layer Perceptron (MLP) to predict the continuous emotion intensity vector. Duringtraining, learnable missing-modality tokens and modality dropout are strategically incorporated to enhance the model's robustness.

Results

Table 1. **Statistical overview of the Hume-Vidmimic2 dataset partitions.** The dataset is strictly divided to ensure no data leakage between training, validation, and testing phases.

Partition	Duration	Samples
Train	15:07:03	8072
Validation	9:12:02	4588
Test	9:04:05	4586
$\sum$	33:23:10	17246

Performance comparison of the EMI Estimation Challenge

Category	Model	Modality	Mean ρ_{val}
Baseline	ViT (baseline) [19]	V	0.09
	Wav2Vec2 (baseline) [19]	A	0.24
	ViT + Wav2Vec2 (baseline) [19]	$V+A$	0.25
Previous Solutions	Yu et al. [33]	$V+A$	0.32
	Hallmen et al. [3]	A	0.38
	Savchenko et al. [28]	$V+A+T$	0.44
	Zhang et al. [38]	$V+A+T$	0.46
	Yu et al. [34]	$V+A+T$	0.51
Ours	**TAEMI**	$V+A+T$	**0.55**

Table 3. **Ablation study on key components of TAEMI.** We perform a leave-one-out evaluation to isolate the individual contributions of Text-Anchored Cross-Attention (TA-CA), Learnable Missing Tokens (LMT), and Modality Dropout (MD). The Δ column explicitly highlights the performance drop when a specific module is removed.

TA-CA	LMT	MD	Mean ρ_{val} (↑)	Δ
Simple Feature Concat			0.45	-0.10
×	✓	✓	0.48	-0.07
✓	×	✓	0.52	-0.03
✓	✓	×	0.51	-0.04
✓	✓	✓	**0.55**	-

Table 4. **Impact of different anchor modalities in the Dual Cross-Attention mechanism.** We compare utilizing Vision, Audio, or Text as the central Query to attend to the other two modalities. Text serves as the most effective query anchor due to its high-level semantic stability.

Anchor Modality (Query)	Mean ρ_{val} (↑)
Vision-Anchored	0.39
Audio-Anchored	0.51
Text-Anchored	**0.55**

Conclusion

In this paper, we presented TAEMI, a robust and highlyeffective multimodal framework for continuous EmotionalMimicry Intensity (EMI) estimation. We identified thatsymmetric late-fusion approaches fail to bridge the seman-tic gap between highly heterogeneous affective signals.

To overcome this, we shifted the fusion paradigm by intro-ducing a Text-Anchored Dual Cross-Attention mechanism.By explicitly designating the textual modality as a seman-tic anchor, our architecture successfully guides the fine-grained temporal alignment of visual and acoustic features.

Moreover, we demonstrated that incorporating LearnableMissing-Modality Tokens alongside a Modality Dropouttraining strategy significantly fortifies the network againstthe missing or noisy data frequently encountered in un-constrained environments. Evaluated on the challengingHume-Vidmimic2 dataset for the 10th ABAW Competition,TAEMI established a new state-of-the-art with an overallmean Pearson correlation of 0.55.

Moving forward, we plan to explore the integration of Large Language Models(LLMs) to extract even deeper semantic priors, further ad-vancing the capabilities of multimodal affective computingin the wild.