CVPR
SEATTLE, WA JUNE 17-21, 2024

MART: Masked Affective RepresenTation Learning via Masked Temporal Distribution Distillation

Zhicheng Zhang Pancheng Zhao Eunil Park Jufeng Yang

南開大學 Nankai University
成均館大學校 SUNGKYUNKWAN UNIVERSITY

Introduction

Motivation: Inspired by the psychology research & empirical theory, we verify that **degree of emotion** may vary in different segment, thus introducing sentiment complementary and emotion intrinsic on temporal segments

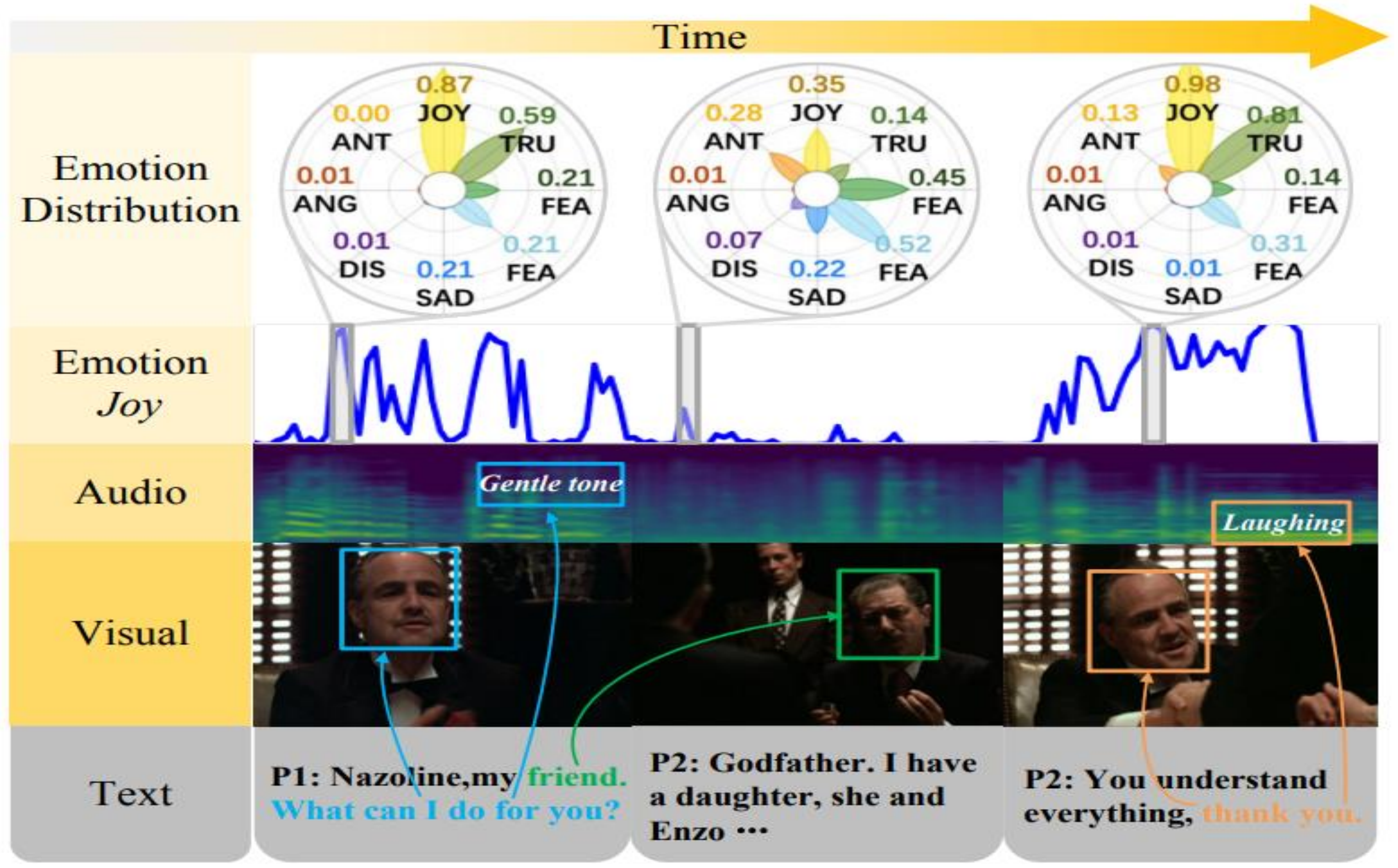
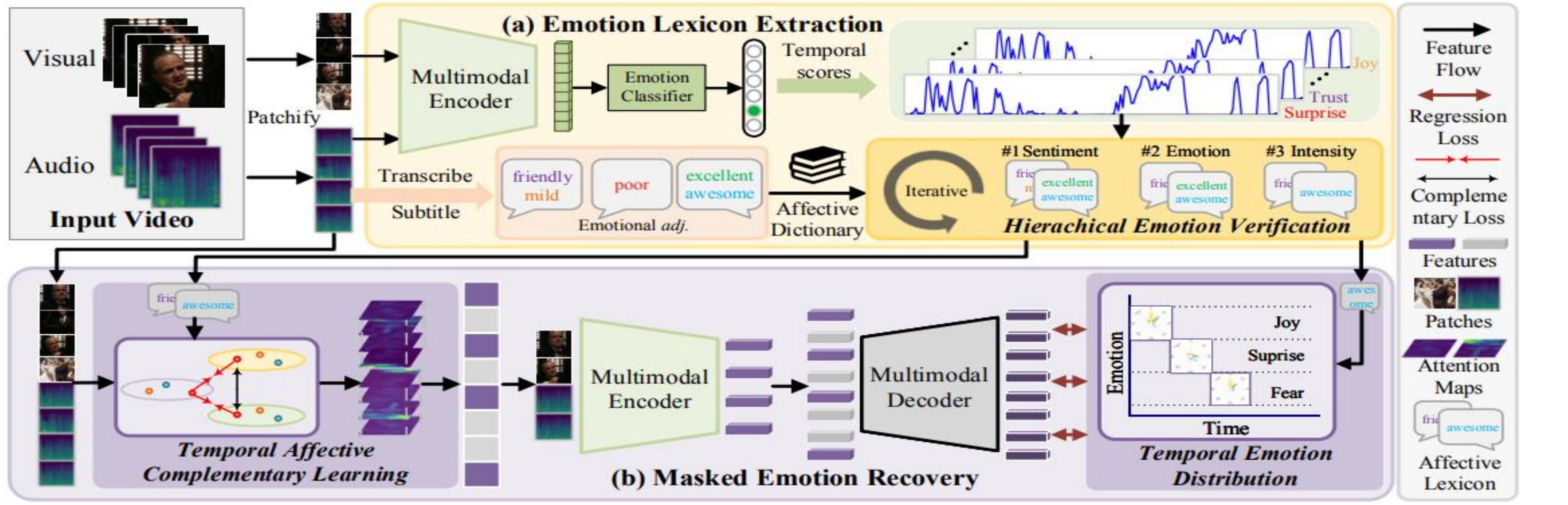

Time
Emotion Distribution
0.87 JOY 0.59 TRU 0.00 ANT 0.21 FEA 0.01 ANG 0.01 DIS 0.21 SAD 0.21 FEA
0.35 JOY 0.14 TRU 0.28 ANT 0.45 FEA 0.01 ANG 0.07 DIS 0.22 SAD 0.52 FEA
0.98 JOY 0.81 TRU 0.13 ANT 0.14 FEA 0.01 ANG 0.01 DIS 0.01 SAD 0.31 FEA
Emotion Joy
Audio
Gentle tone
Laughing
Visual
Text
P1: Nazoline,my friend. What can I do for you?
P2: Godfather. I have a daughter, she and Enzo ···
P2: You understand everything, thank you.

Fig.1. *Godfather* (1972) and its corresponding emotion. The colored texts indicate the emotional words given by affective lexicon and the rectangles show the related content towards emotional words.

Contributions

- **Masked affective modeling** to exploit temporal affective cues among modalities for discriminative representation, which can be integrated into existing video emotion analysis methods as a **plug-in** module.
- Extensive experiments demonstrate effectiveness of MART **covering four downstream areas** on emotion classification and sentiment analysis.

Methodology



Video → Cues

➢ **Emotion Lexicon Extraction**

a) extract emotion words w_e

b) hierarchical Verification

Sentiment $\mathtt{E1:}\ \mathbb{1}(c_s = \overline{c_s}).$

Emotion $\mathtt{E2:}\ \mathbb{1}(c_s = \overline{c_s}) \cdot \mathbb{1}(c_e = \overline{c_e}).$

Degree $\mathtt{E3:}\ \mathbb{1}(c_s = \overline{c_s}) \cdot \mathbb{1}(c_e = \overline{c_e}) \cdot \mathbb{1}(\|i_e - \overline{i_e}\| < \tau).$

Cues → Mask

➢ **Emotion-guided Masking**

Guided by affective semantics, *i.e.,* attention: $A = \frac{1}{H}\sum_{i=1}^{H} a_i \in \mathbb{R}^{K \times d_m},$

Highly hindered masking

$$m_i = \begin{cases} 1, S < idx_i \le \lceil \alpha \cdot M \rceil + S, \\ 0, otherwise. \end{cases},$$

Masked Video → Distribution

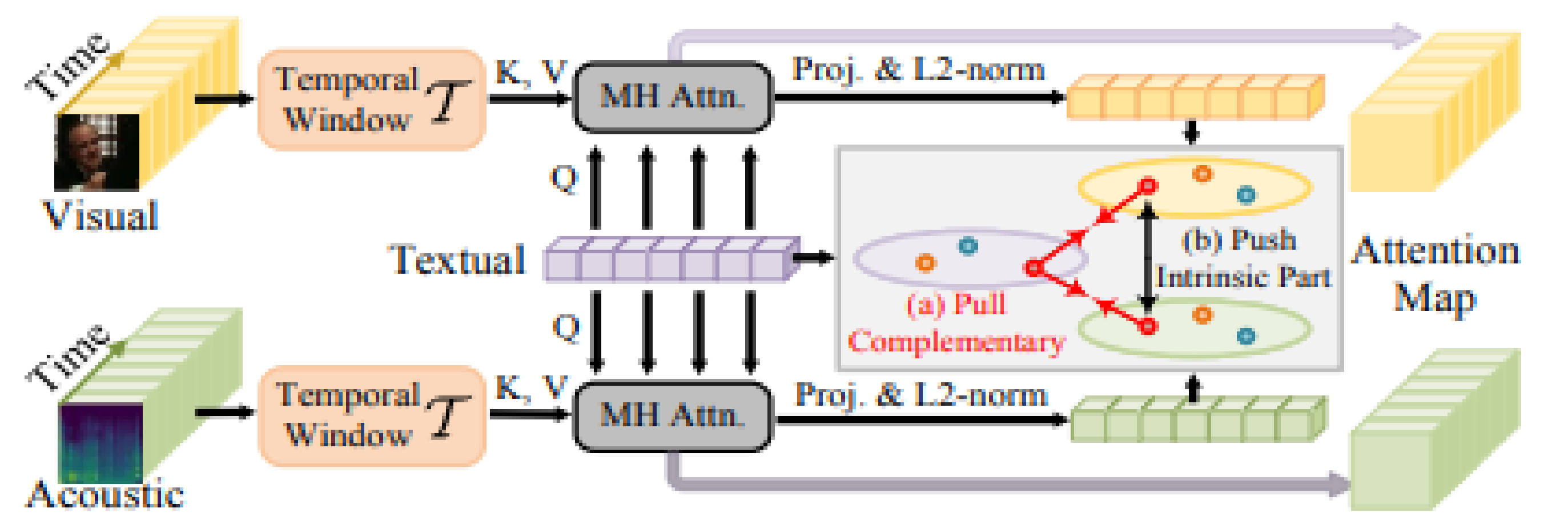
➢ **Temporal Emotion Distribution**

a) recover masked embedding

$$\mathcal{L}_{MAE}(h) = \mathbb{E}_v \mathbb{E}_{v_m, v_s | v}\ \phi(g(f(v_u)), t(v_m)).$$

b1) feature-level Rec: Temporal

b2) patch-level Rec: Spatiotemporal



Affective Mask by TACL

➢ **Temporal Affective Complementary Learning**

a) yielding lexicon-aware features $x^{v \to l}$ by Multi-Head Attn

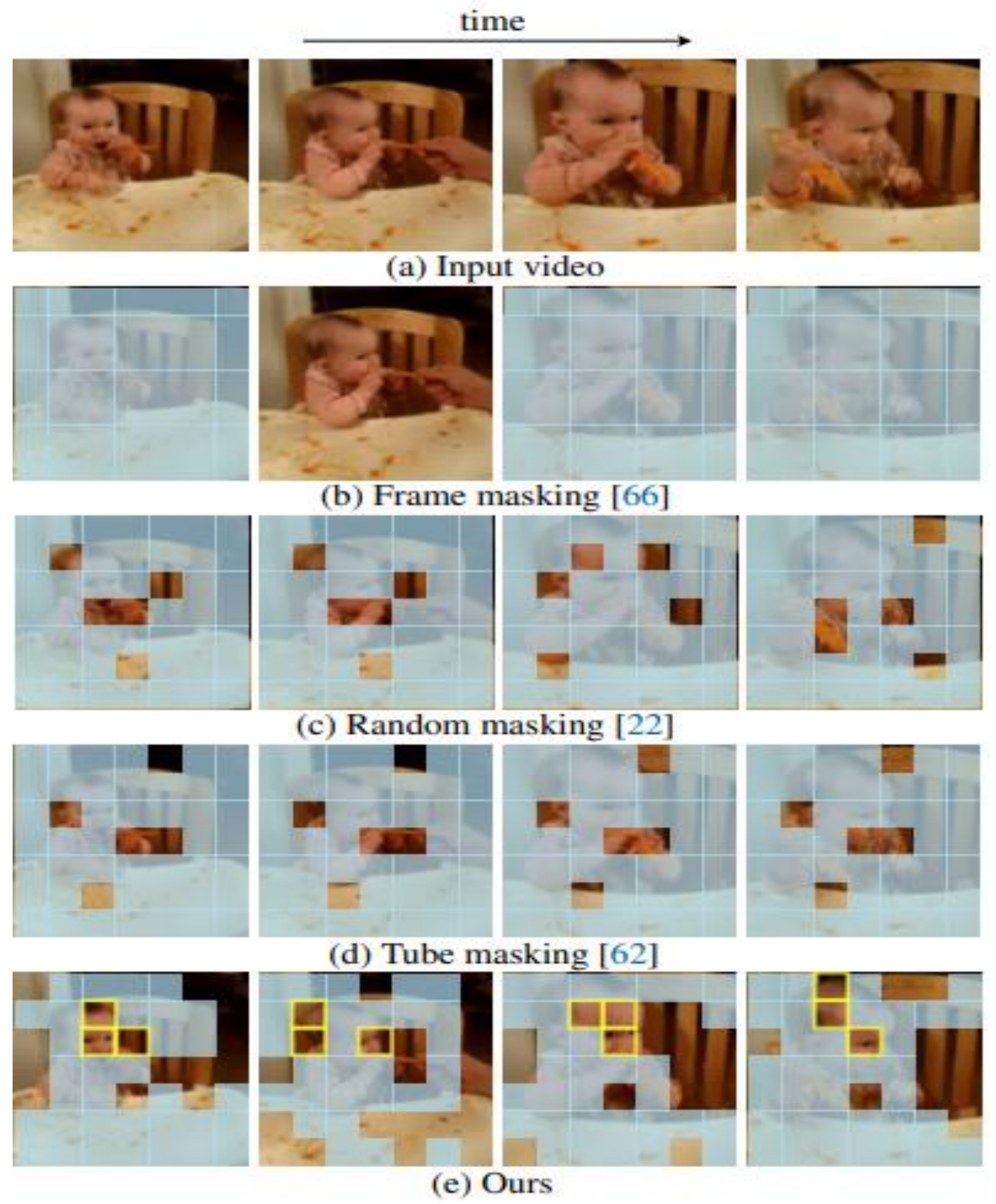
b) align the visual- and acoustic-linguistic to find the corresponding area in multi-modalities

$$\mathcal{L}^v_{tacl} = -\mathbb{E}_{t \in \mathcal{T}^k} \log \frac{D\left(x^{v\to l}_s[t], x^l[t]\right)}{\underbrace{D\left(x^{v\to l}_s[t], x^l[t]\right)}_{pull\ complement} + \sum_{i \notin \mathcal{T}^k} \underbrace{D\left(x^{v\to l}_s[t], x^{a\to l}_s[i]\right)}_{push\ difference}}.$$

c) discussion: serve as multimodal form distance constraint

$$\mathcal{L}_{tacl} = \mathcal{L}^v_{tacl} + \mathcal{L}^a_{tacl}$$
$$\propto 4\tau + \left\|x^{a\to l}_s - x^l\right\|^2 + \left\|x^{v\to l}_s - x^l\right\|^2 - 2\left\|x^{v\to l}_s - x^{a\to l}_s\right\|^2.$$

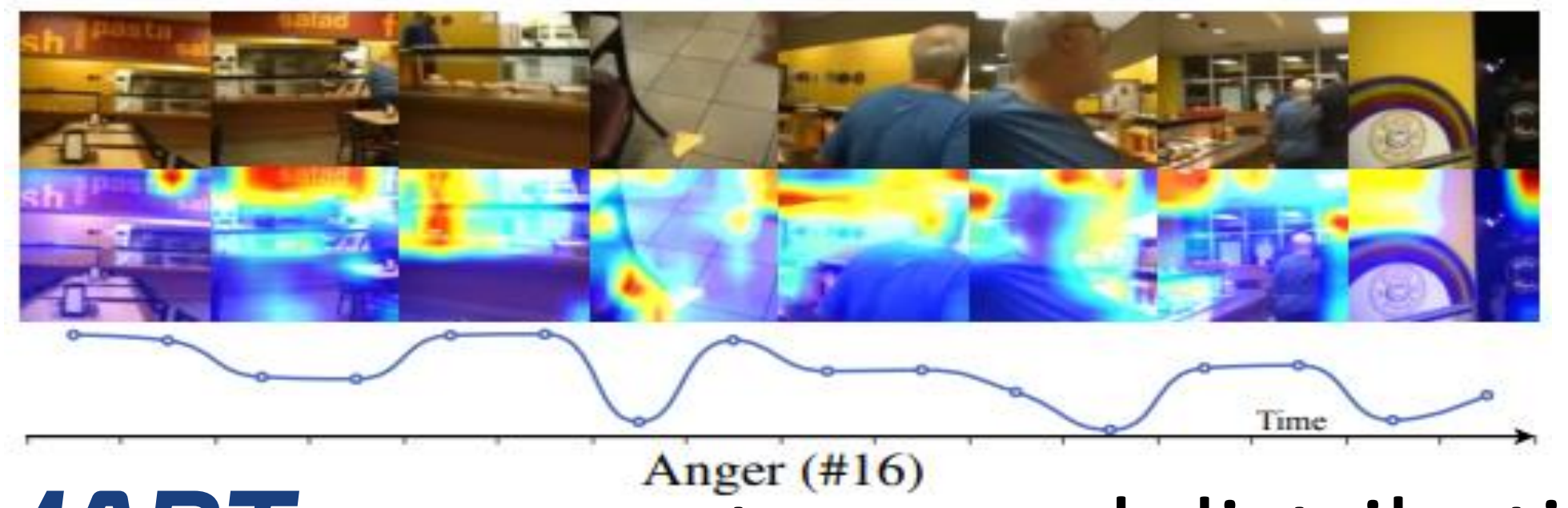
Performance



(a) Input video
(b) Frame masking [66]
(c) Random masking [22]
(d) Tube masking [62]
(e) Ours

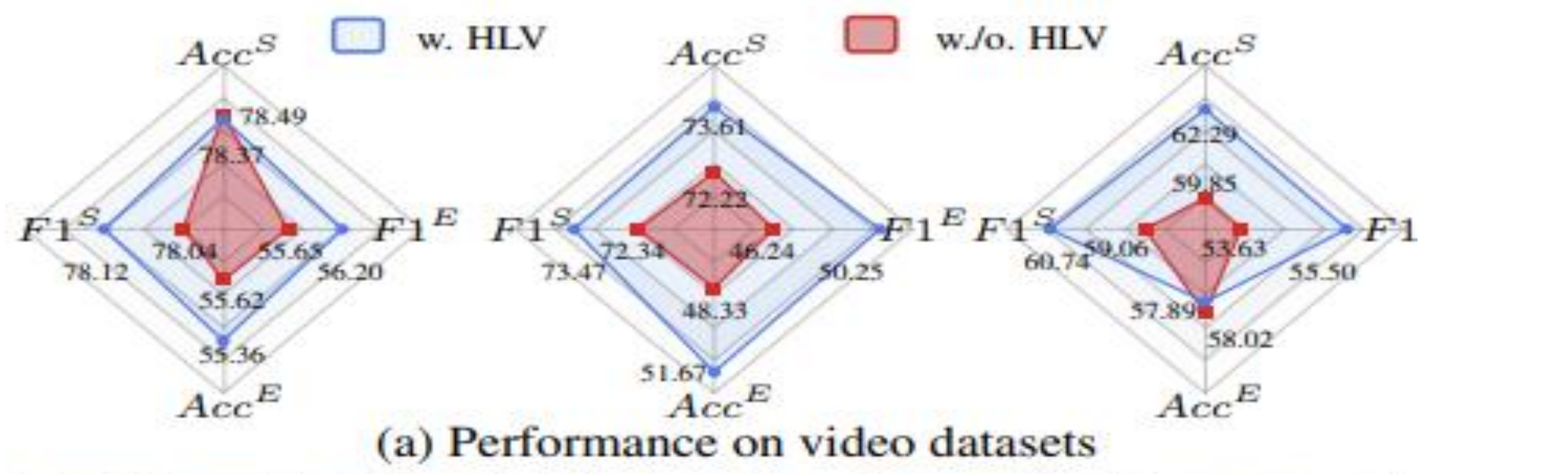
✓ ***MART*** encourages video emotion analysis models to focus on affective area, unlike common mask strategy which randomly select masking area.

✓ ***TACL*** guides the recovery, where highly attentive areas in yellow boxes are reserved for providing affective cues to be recovered.

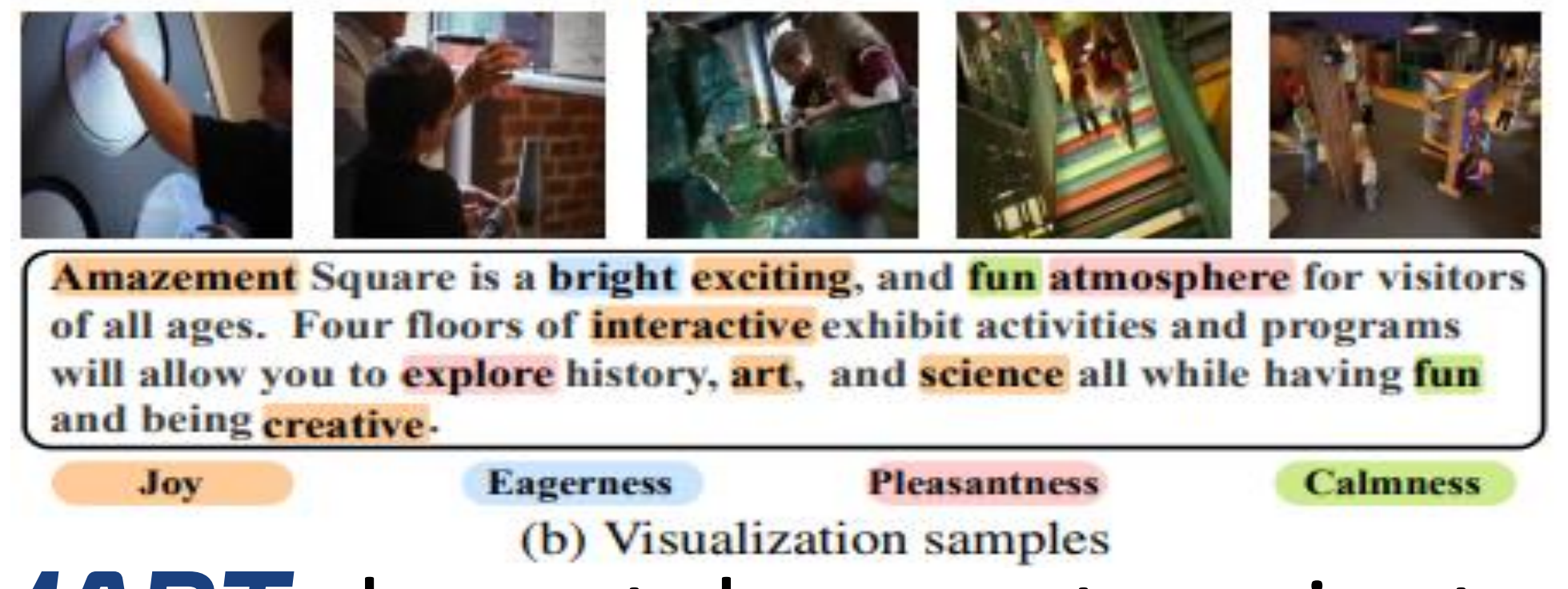


Anger (#16)

✓ ***MART*** recovers temporal distribution.



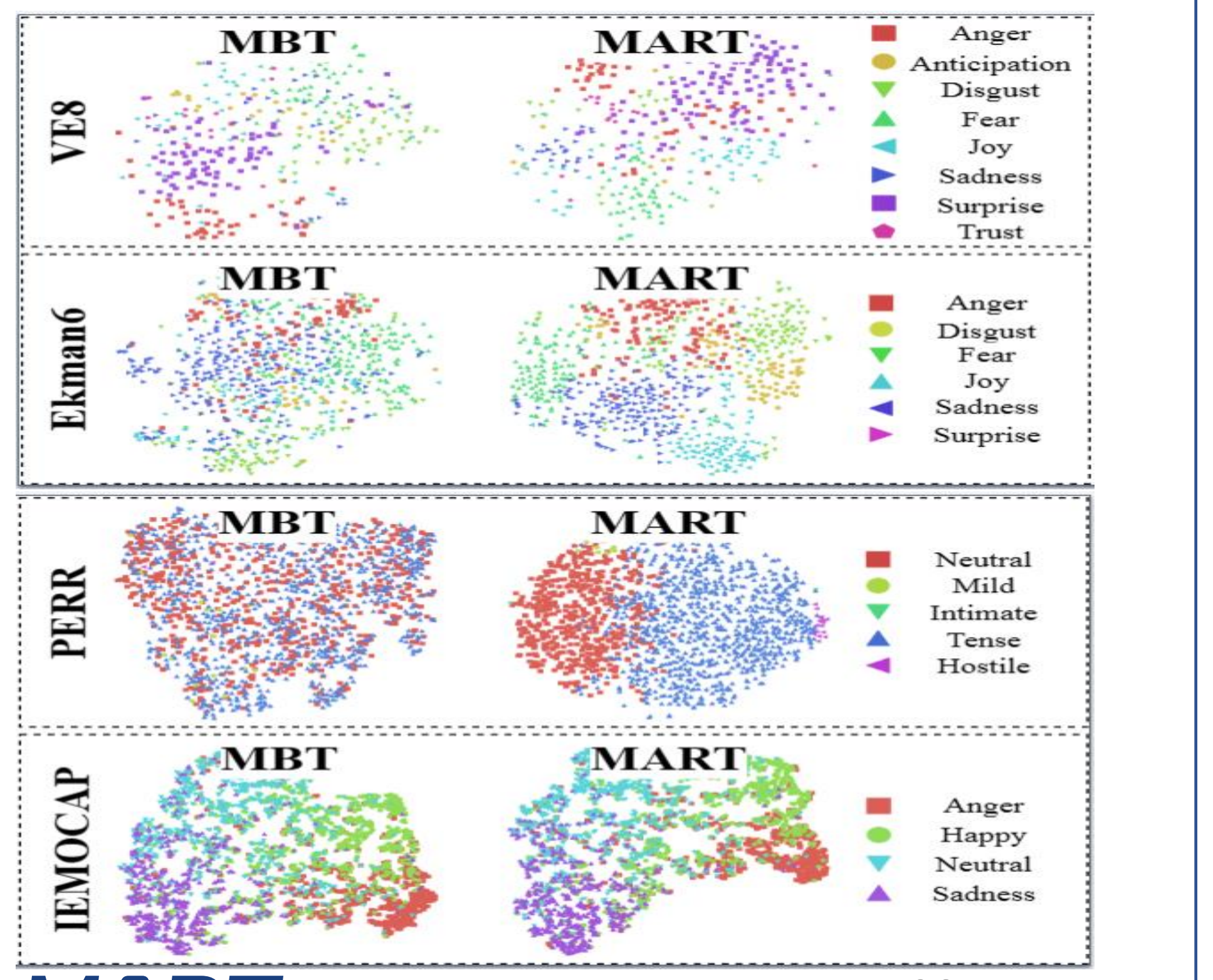
(a) Performance on video datasets



Amazement Square is a bright exciting, and fun atmosphere for visitors of all ages. Four floors of interactive exhibit activities and programs will allow you to explore history, art, and science all while having fun and being creative.

(b) Visualization samples

✓ ***MART*** shows tolerance to noisy text.



✓ ***MART*** extracts robust affective embedding across datasets.

Contact
MART
gloryzzc6@sina.com

Website Project Code