LaScA: Language-Conditioned Scalable Modelling of Affective Dynamics

Supplementary Material

A. LLM Instruction Prompt

Instruction Prompt for Offline Semantic Lexicon Construction

```
You are an expert in affective computing and
    multimodal emotion modelling.

Your task is to convert low-level video (facial
    blendshapes) and audio (acoustic features such
    as MFCCs) feature names into compact semantic
    labels that combine:

1) A brief physical cue
2) A brief affect cue

STRICT RULES:
1. Output must be a valid Python dictionary.
2. No explanations outside the dictionary.
3. Each value 3 to 6 words.
4. No full sentences.
5. No punctuation inside values.
6. Format: "<expression cue> <affect cue>"
7. Video: describe visible movement.
8. Audio: describe acoustic change.
9. Compact affect cue.
10. Avoid deterministic language.
11. Keep vocabulary consistent.
12. Optimise for SentenceTransformer embeddings.
13. Prioritise reproducibility.

Example format:
'feature_name': 'feature meaning and affect
    indication'

Return only the dictionary.
```

B. Description Generation

B.1. Audio Descriptors

Table 8. Affect-aware semantic labels for MFCC acoustic features generated offline.

Feature Name	Semantic Label
mfcc_0	high arousal energy
mfcc_1	dominant low tone
mfcc_2	neutral stability
mfcc_3	tense vocal focus
mfcc_4	clear controlled tone
mfcc_5	urgent brightness
mfcc_6	strained timbre
mfcc_7	assertive pressure
mfcc_8	excited anxiety
mfcc_9	tight emphasis
mfcc_10	breathy nervousness
mfcc_11	withdrawn low energy
mfcc_12	piercing tension

B.2. Facial Descriptors

Table 9. Affect-aware semantic lexicon for facial blendshape features.

Feature Name	Semantic Label
Face_browDownLeft	left brow down focused irritation
Face_browDownRight	right brow down skeptical tension
Face_browInnerUp	inner brows up sad vulnerability
Face_browOuterUpLeft	left outer brow up curious surprise
Face_browOuterUpRight	right outer brow up questioning surprise
Face_cheekPuff	cheeks puff held frustration
Face_cheekSquintLeft	left cheek tight restrained smile
Face_cheekSquintRight	right cheek tight restrained smile
Face_eyeBlinkLeft	left blink stress regulation
Face_eyeBlinkRight	right blink stress regulation
Face_eyeLookDownLeft	left gaze down shame reflection
Face_eyeLookDownRight	right gaze down shame reflection
Face_eyeSquintLeft	left eye squint suspicious focus
Face_eyeSquintRight	right eye squint evaluative doubt
Face_eyeWideLeft	left eye wide alarm arousal
Face_eyeWideRight	right eye wide heightened alertness
Face_jawForward	jaw thrust assertive dominance
Face_jawLeft	jaw left uneasy tension
Face_jawOpen	jaw drop shock surprise
Face_jawRight	jaw right uneasy tension
Face_mouthClose	lips closed emotional restraint
Face_mouthPressLeft	left lip press suppressed anger
Face_mouthPressRight	right lip press controlled tension
Face_mouthRollLower	lower lip roll inhibited emotion
Face_mouthRollUpper	upper lip roll inhibited emotion
Face_mouthFrownLeft	left corner down sad disapproval
Face_mouthFrownRight	right corner down sad disapproval
Face_mouthLowerDownLeft	left lower lip down vulnerable sadness
Face_mouthLowerDownRight	right lower lip down vulnerable sadness
Face_mouthSmileLeft	left smile smirk positivity
Face_mouthSmileRight	right smile genuine positivity
Face_mouthDimpleLeft	left dimple warm engagement
Face_mouthDimpleRight	right dimple warm engagement
Face_mouthStretchLeft	left mouth stretch awkward tension
Face_mouthStretchRight	right mouth stretch awkward tension
Face_mouthFunnel	mouth funnel uncertain anticipation
Face_mouthPucker	lip pucker affection hesitation
Face_mouthShrugLower	lower lip shrug doubt uncertainty
Face_mouthShrugUpper	upper lip shrug skeptical hesitation
Face_mouthUpperUpLeft	left upper lip up disgust contempt
Face_mouthUpperUpRight	right upper lip up disgust contempt
Face_noseSneerLeft	left nose sneer strong disgust
Face_noseSneerRight	right nose sneer strong disgust

B.3. Description Template

For each temporal window, textual descriptions are constructed deterministically from the activated handcrafted features using predefined semantic mappings and structured templates.

Unimodal Facial Template. Facial blendshape features are first mapped to compact affect-aware semantic labels using the fixed lexicon described in Section 3.1. Given the subset of active facial features for a window, their corresponding semantic labels are concate-

nated into a structured textual representation via the `FacialLLMDescriptionGenerator`. The generator produces a consistent phrase ordering to ensure stable downstream embeddings. The resulting facial prompt takes the form:

$$T_t^{\text{face}} = \texttt{concat}(\ell_{i_1}, \ell_{i_2}, \ldots, \ell_{i_k}) \texttt{<|endoftext|>} \tag{7}$$

where ℓ_{i_j} denotes the semantic label associated with the j-th activated facial feature in window t.

Unimodal Audio Template. Audio features (MFCC-derived descriptors) are processed analogously using the `AudioLLMDescriptionGenerator`. Activated acoustic features are converted into their corresponding semantic labels and concatenated into a structured textual prompt:

$$T_t^{\text{audio}} = \texttt{concat}(a_{i_1}, a_{i_2}, \ldots, a_{i_m}) \texttt{<|endoftext|>} \tag{8}$$

where a_{i_j} denotes the semantic label of the j-th active acoustic feature.

Multimodal Template Construction. The multimodal description is formed by concatenating the facial and audio prompts while preserving ordering consistency. The intermediate end-of-text markers are removed to avoid fragmentation, and a single terminal token is appended:

$$T_t^{\text{multi}} = \texttt{strip}(T_t^{\text{face}}) \texttt{ | } \texttt{strip}(T_t^{\text{audio}}) \texttt{<|endoftext|>} \tag{9}$$

This construction ensures: (i) deterministic prompt generation, (ii) stable formatting across samples, (iii) modality-aware separation via the delimiter "|", and (iv) compatibility with sentence-transformer tokenization. All templates are generated offline and remain fixed throughout training and evaluation to ensure full reproducibility.

B.4. Description Examples

To illustrate the structure of the language-conditioned representations used in **LaScA**, we report the five most frequent prompts observed in the dataset. Prompts are constructed by combining active facial descriptors with acoustic cues.

1. `facial: left blink stress regulation, right blink stress regulation, left gaze down shame reflection, right gaze down shame reflection, left eye squint suspicious focus, right eye squint evaluative doubt`
   `audio: Acoustic markers indicate high arousal energy.`

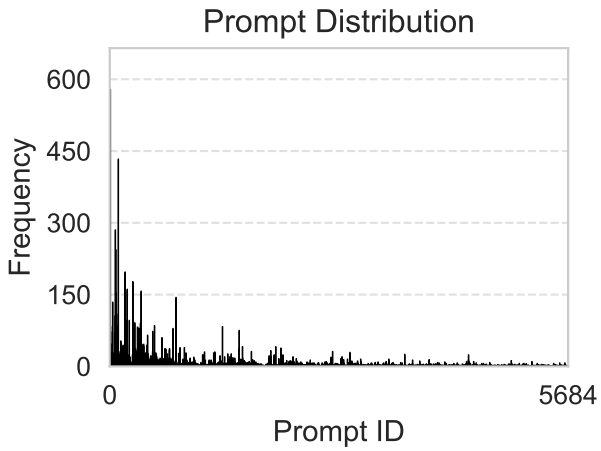

Figure 2. Histogram of unique multimodal prompts

2. `facial: left brow down focused irritation, right brow down skeptical tension, left blink stress regulation, right blink stress regulation, left gaze down shame reflection, right gaze down shame reflection, left eye squint suspicious focus, right eye squint evaluative doubt`
   `audio: Acoustic markers indicate high arousal energy.`
3. `facial: inner brows up sad vulnerability, left outer brow up curious surprise, right outer brow up questioning surprise, left gaze down shame reflection, right gaze down shame reflection`
   `audio: Acoustic markers indicate high arousal energy.`
4. `facial: inner brows up sad vulnerability, left outer brow up curious surprise, right outer brow up questioning surprise`
   `audio: Acoustic markers indicate high arousal energy.`
5. `facial: inner brows up sad vulnerability, left outer brow up curious surprise, right outer brow up questioning surprise, left blink stress regulation, right blink stress regulation, left gaze down shame reflection, right gaze down shame reflection, left eye squint suspicious focus, right eye squint evaluative doubt`
   `audio: Acoustic markers indicate high arousal energy.`