ROCLING 2026
DSA-NIFT SHARED TASK

Source-Aware Lexicon-Augmented Regression for Valence–Arousal Prediction

Affect prediction for new-immigrant reflections with no target labels

Chih-Hsi Chen · **Hang Tan**
NCKU NLP · Dept. of CSIE, National Cheng Kung University

Code & cue lists

Shared task

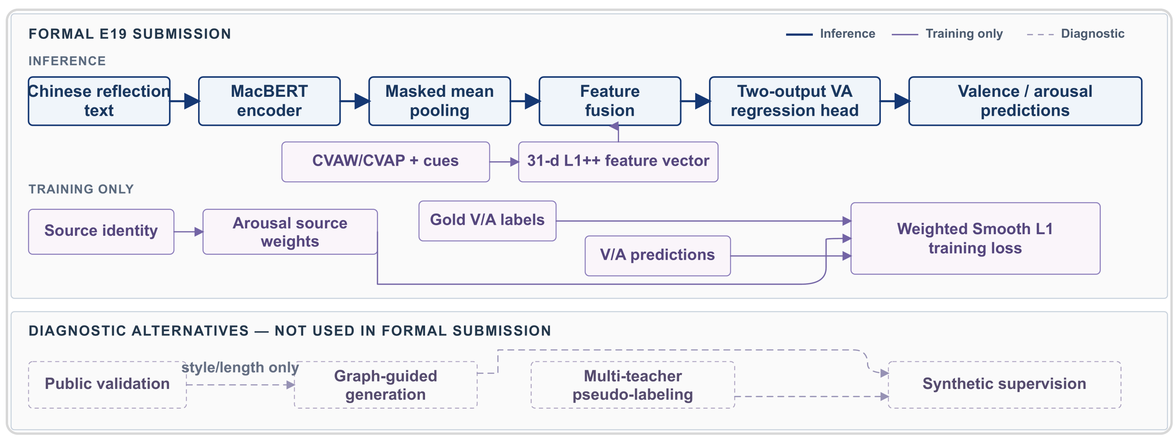
1 Motivation & Setting

Each reflection must be scored for **valence** (pleasantness) and **arousal** (activation) on a 1–9 scale. Systems are ranked by the mean rank of MAE↓ and PCC↑ on both dimensions.

- **No target labels.** The 200 validation and 1,100 test texts come without gold scores; only leaderboard aggregates are visible.
- **Borrowed supervision.** 9,433 labeled texts: EmoBank sentences (2,582) and documents (2,969), medical reflections DSA-MST (2,282), educational reflections ROCLING-2021 (1,600).

Q: which borrowed supervision actually transfers to arousal?

2 System Overview



Top: the submitted E19 system. MacBERT mean-pooled states are fused with a 31-d lexicon vector; source identity changes only the arousal loss weight. **Bottom:** diagnostic branches, not used in the submission.

3 Affective-Lexicon Fusion

L1 (10-d) matches 5,512 CVAW words and 2,998 CVAP phrases, then records coverage, hit count, and the mean, max, min and std of matched V and A.

L1++ (31-d) adds 21 arousal-oriented cues from 14 hand-written lists: punctuation, intensity words, sentence rhythm, negation and contrast, and six migration topics.

4 Source-Aware Arousal Loss ★ KEY

WEIGHTED SMOOTH-L1 OBJECTIVE (EQ. 4)

$$\mathcal{L}_{\text{src}} = \frac{1}{|B|} \sum_{i \in B} \sum_{d \in \{V,A\}} w_d(s_i)\, \ell_{\text{SL1}}\left(\hat{\tilde{y}}_{id},\, \tilde{y}_{id}\right)$$

Valence weights stay at 1.0. Arousal weights follow how closely a source matches **first-person, document-level reflection**:

1.00	0.75	0.50	0.25
DSA-MST medical reflection	ROCLING-2021 short reflection	EmoBank CVAT general documents	EmoBank CVAS general sentences

An ordinal prior, not tuned. Post hoc, non-uniform weights beat uniform by less than the seed SD.

5 Model Selection → Official Test ★ Result

Public validation (n = 200)	V-MAE ↓	V-PCC ↑	A-MAE ↓	A-PCC ↑
EmoBank baseline	0.654	0.867	0.985	0.412
E4: + reflections + L1	**0.600**	0.880	0.882	0.426
E18: L1 → L1++	0.641	0.865	0.909	0.408
E19: + source weights	**0.666**	**0.869**	**0.870**	**0.452**
Raw synthetic labels	0.611	0.870	1.100	**0.461**
Official test, E19 scored once: 0.620 · 0.866 · 0.926 · 0.357				

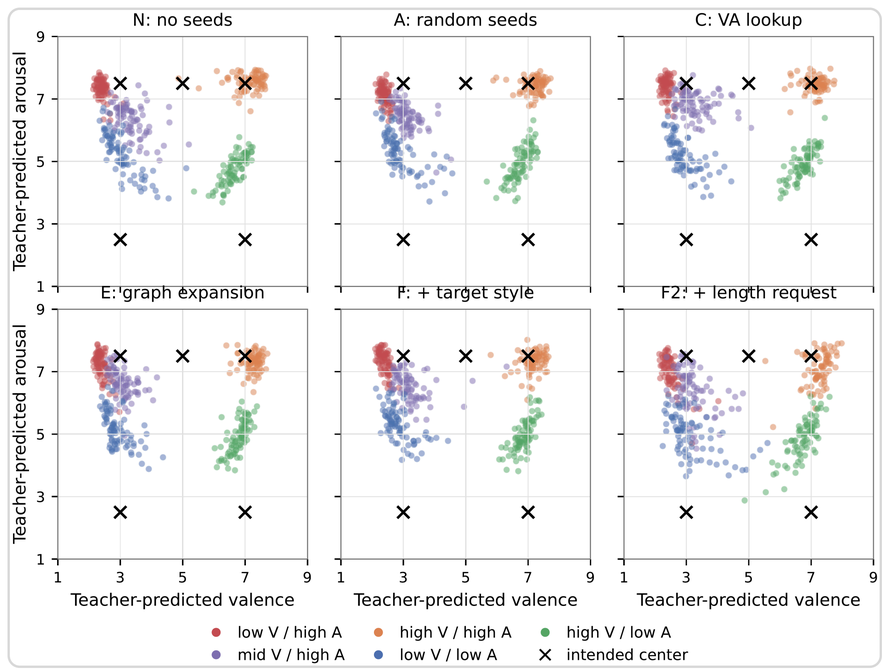
4th / 7	0.870	0.866
official rank (mean rank 4.25)	lowest validation A-MAE	test valence PCC

L1++ alone worsens all four metrics; adding source weights gives the lowest arousal MAE.

6 Graph-Guided LLM Augmentation

An LLM writes 400 reflections for five target VA centers. One prompt component changes per step:

N no seeds → **A** random seeds → **C** VA-filtered → **E** graph kNN → **F** + style → **F2** + length

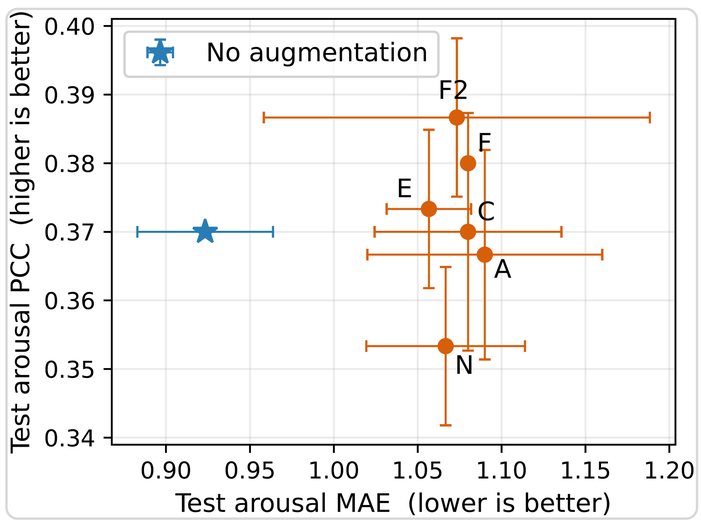


Frozen three-teacher predictions for each condition; × marks the intended centers.

Seed source	Seed coherence ↑	Distance to target ↓	High–low A gap ↑
C: VA lookup	0.0126	**1.631**	**2.173**
E: graph expansion	**0.179**	**1.701**	**2.020**

Graph seeds are **14.2× more coherent**, yet the generated texts are no closer to their targets and mean test A-PCC rises only ≈ **0.003**. **Controllable is not the same as useful.**

7 The Arousal PCC–MAE Trade-off



Matched test ablation, three seeds per condition (F: one seed).

Every synthetic condition moves right: F2 lifts arousal PCC **0.370 → 0.387** but worsens MAE **0.923 → 1.073**. Length matching is only partial: 31.5% of F2 texts are under 50 characters, versus 6.0% of validation texts. Small gaps are also hard to trust:

2.3×	±0.004	< 0.01
wider arousal-PCC 95% interval than valence (bootstrap on dev proxy, n = 200)	test arousal PCC across five training seeds	rounded-score gaps: descriptive only

8 Takeaways ★

- **Supervision.** Source reliability for arousal labels may matter more than extra lexical cues.
- **Generation.** Better prompt control did not translate into better regression labels.
- **Selection.** Arousal PCC and MAE disagree; select and report on both.

Next: stronger target-domain validation, comparative arousal supervision, and trusted target labels to calibrate synthetic data.

In one sentence: without target labels, decide how much to trust each arousal source before adding more features or synthetic data.

Code & cue lists: **doi.org/10.5281/zenodo.22858574** · **chenbrian93@gmail.com**