NCKU_NLP at ROCLING-2026 Shared Task: Source-Aware Lexicon-Augmented Regression for Valence–Arousal Prediction

Chih-Hsi Chen
Department of Computer Science and Information Engineering
National Cheng Kung University
chenbrian93@gmail.com

Hang Tan
Department of Computer Science and Information Engineering
National Cheng Kung University
harry30701e@gmail.com

Abstract

The ROCLING-2026 DSA-NIFT shared task requires continuous valence and arousal prediction for Chinese new-immigrant reflection texts, but provides no labeled target-domain training data. Our formal submission combines MacBERT, Chinese affective-lexicon features, and source-aware arousal supervision; we additionally evaluate pseudo-labeling, pairwise supervision, and graph-guided LLM augmentation as diagnostic alternatives. The submitted system obtains 0.620/0.866 valence MAE/PCC and 0.926/0.357 arousal MAE/PCC on the official test set, ranking fourth among the seven participating teams. On official public validation, raw synthetic supervision reaches 0.461 arousal PCC but worsens arousal MAE to 1.100, whereas E19 trades slightly lower arousal PCC (0.452) for the lowest arousal MAE (0.870). A controlled augmentation ladder further separates controllability from downstream utility: graph expansion raises intrinsic seed coherence by $14.2\times$ over VA lookup, but yields only an approximately 0.003 increase in mean test arousal PCC based on rounded leaderboard scores. In matched batch-32, three-seed reruns, the synthetic condition with the highest arousal PCC reaches 0.387 versus 0.370 without augmentation, while arousal MAE increases from 0.923 to 1.073. Together, these results highlight the need to balance PCC and MAE when selecting arousal models under weak auxiliary supervision.

Keywords: Chinese Dimensional Sentiment Analysis, Cross-Domain Regression, Affective Lexicon Fusion, Source-Aware Supervision

1 Introduction

Dimensional sentiment analysis represents affect on continuous axes rather than with a single discrete category. In the circumplex account, *valence* describes pleasantness and *arousal* describes activation (Russell, 1980). The ROCLING-2026 shared task on Chinese Dimensional Sentiment Analysis for New Immigrants' Feeling Texts (DSA-NIFT) applies this representation to first-person reflections about migration and settlement (Lin et al., 2026). Systems must predict both dimensions on a 1–9 scale and are evaluated by mean absolute error (MAE) and Pearson correlation (PCC).

The task is difficult because affect must transfer without target-domain labels. Gold labels are not released for either the 200-text public validation set or the 1,100-text test set; participants receive only aggregate leaderboard scores. Consequently, all supervised signals must be borrowed from general Chinese affective resources, medical reflections, or educational reflections. The closest predecessor found that general-domain data and target-like reflections need not help valence and arousal equally (Huang and Shao, 2025). The small public validation set creates a second problem: even a reasonable adaptation may be selected because of finite-sample variation rather than robust transfer.

We study this setting through a common cross-domain VA regressor rather than claiming a new state-of-the-art architecture. Our formal submission combines MacBERT, a 31-dimensional affective-lexicon feature vector, and source-aware arousal supervision. Separately, we evaluate encoder changes, pairwise objectives, multi-teacher pseudo-labels, ensembles, and target-style synthetic reflections as diagnostic alternatives. For the synthetic branch, we explicitly separate arbitrary seed words, VA-filtered seeds, graph-expanded seeds, style anchors, and length matching. This design distinguishes the submitted method from the diagnostic interventions and tests whether each intervention's intended signal survives official evaluation.

Our contributions are threefold:

- We develop a fourth-place cross-domain DSA-

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 486–495
November 6-7, 2026 ©2026 Association for Computational Linguistics

NIFT system combining affective-lexicon fusion with source-aware arousal supervision, and compare it with encoder, pseudo-labeling, pairwise, ensemble, and synthetic-data variants under the official metrics.

- As a diagnostic analysis, we conduct a component-level study of graph-guided LLM augmentation. The graph measurably changes seed coherence, but yields only an approximately 0.003 increase in mean A-PCC over VA lookup based on rounded leaderboard scores; style/length alignment changes more, yet introduces a clear PCC–MAE trade-off.

- We analyze arousal model selection under competing official metrics. Public validation favors different systems under PCC and MAE, while proxy bootstrap and seed analyses show that small A-PCC differences should be interpreted cautiously.

2 Related Work

Chinese DSA resources and shared-task systems. Chinese EmoBank provides word-, phrase-, sentence-, and document-level VA resources (Yu et al., 2016; Lee et al., 2022) and underpins a sequence of Chinese DSA shared tasks on educational (Yu et al., 2021) and medical (Lee et al., 2025) reflections. Pretrained Chinese encoders such as MacBERT (Cui et al., 2020) are natural regression backbones, and affective norms (Köper et al., 2017), word/character fusion (Deng et al., 2023), and dimension-specific attention (Zhu et al., 2019) have each been proposed to complement them; our shared inference path instead uses masked-mean pooling, and only a later stacked configuration examines dimension-specific pooling. Transferring without target labels is the classic unsupervised domain-adaptation setting (Ramponi and Plank, 2020), for which continued pretraining is a standard tool (Gururangan et al., 2020). The ROCLING-2025 systems are the closest precedents: NTULAW used different data-selection policies for valence and arousal (Huang and Shao, 2025), motivating—but not specifying—our source-weighted loss; CYUT-NLP used retrieval-augmented generation and multi-teacher pseudo-labeling (Jian et al., 2025); and TCU combined LLM embeddings with regression and ensembling (Li and Lin, 2025). Because those systems target medical texts, their scores are not comparable to DSA-NIFT; we test the underlying mechanisms on the new task instead.

Weak supervision and model selection. Weak affective supervision can come from LLM-generated psychological indicators (Furniturewala and Jaidka, 2024), from VA structure imposed on categorical errors (Mitsios et al., 2024), or from synthetic training data, whose utility varies with task subjectivity (Li et al., 2023); for continuous intensity, comparative annotations such as pairwise or best–worst judgments are often more reliable than direct rating scales (Bagdon et al., 2024). These findings motivate our intensity features, synthetic pairs, and teacher relabeling, whereas the affective graph is a new seed-control mechanism here rather than an inherited one. Selecting among such variants is itself risky: finite-sample criteria can be overfit when their variance is non-negligible (Cawley and Talbot, 2010). We contribute a task-specific case study of public-validation metric trade-offs and source-domain proxy uncertainty; unlike work with released labels, we cannot run paired error analysis on the official examples.

3 Task and Data

Given a Chinese reflection x_i, the task is to predict $\mathbf{y}_i = (y_i^V, y_i^A) \in [1, 9]^2$. The official scorer returns MAE (↓) and PCC (↑) for each dimension; final task ranking uses the mean rank across the four metrics. We use "absolute error" for MAE and "association" for PCC, and reserve "ranking" for the explicit pairwise objective.

Table 1 lists the labeled source-domain training data: Chinese EmoBank (Lee et al., 2022), DSA-MST (Lee et al., 2025), and ROCLING-2021 (Yu et al., 2021). The 2,282 DSA-MST training documents exclude a 253-document medical-reflection development split. Thus, the training set contains 9,433 labeled examples and the internal development set contains 253 disjoint examples. The development set is useful for checkpoint selection but is an imperfect target proxy: its domain is medical, and its scores are systematically higher than official validation scores.

The public target-domain splits contain 200 validation and 1,100 test texts. Their gold labels are not available locally; only aggregate leaderboard results can be observed. We use the validation texts without labels for DAPT and, in two synthetic conditions, as style and length references. Test texts and test scores were not used to choose the formal

E19 submission. After the evaluation period, the leaderboard remained open and enabled additional diagnostic submissions; Section 5.3 distinguishes these post-hoc analyses from the formal submission.

4 Methodology

4.1 System Overview

Figure 1 separates the shared regression path from optional adaptation branches. Every neural system maps source-labeled text through an encoder and a two-output regressor. Lexicon features enter the representation; source weights alter only the supervised loss; and synthetic texts enter either as direct weak labels, teacher labels, or pairwise constraints. This separation is important because our formal submission uses source weighting but no synthetic examples.

4.2 Base VA Regression

We normalize each gold dimension to $[0, 1]$ as $\tilde{y}_i^d = (y_i^d - 1)/8$, where $d \in \{V, A\}$. A pretrained Chinese encoder produces token states $H_i = [\mathbf{h}_{i1}, \ldots, \mathbf{h}_{iT}]$. We use attention-mask mean pooling,

$$\mathbf{z}_i = \frac{\sum_t m_{it}\mathbf{h}_{it}}{\sum_t m_{it}}, \tag{1}$$

and a sigmoid output head,

$$\hat{\tilde{\mathbf{y}}}_i = \sigma(W[\mathbf{z}_i; \mathbf{l}_i] + \mathbf{b}), \qquad \hat{\mathbf{y}}_i = 1 + 8\hat{\tilde{\mathbf{y}}}_i. \tag{2}$$

Here $\mathbf{l}_i$ is empty in the no-lexicon model and contains the features below otherwise. The base objective sums PyTorch Smooth L1 loss over both dimensions. Unless noted otherwise, the encoder is `hfl/chinese-macbert-base`.

4.3 Affective-Lexicon Fusion

Lexicon fusion supplies a source-independent affect prior when target labels are absent. We combine 5,512 CVAW words and 2,998 CVAP phrases from Chinese EmoBank. After adding multi-character entries to the jieba tokenizer's vocabulary, [1] we compute ten features:

$$\mathbf{l}_i = [r, c, \mu_V, \max_V, \min_V, s_V, \mu_A, \max_A, \min_A, s_A]. \tag{3}$$

r is the fraction of segmented tokens matched by the lexicon, and c is the hit count capped at 20 and divided by 20. VA statistics are normalized to $[0, 1]$; a no-hit document receives zero coverage/count, neutral means/extrema (0.5), and zero standard deviations. E4 concatenates these ten features with the 768-dimensional pooled MacBERT representation. E18 expands this vector to 31 dimensions with the 21 surface and domain features detailed in Table 2. These additions were intended to expose arousal-specific intensity and migration-context signals that the ten aggregate VA statistics do not encode explicitly.

Keyword and domain categories are evaluated independently, so one document can activate multiple groups. Sentence boundaries are punctuation-based; all other added cue features are computed from raw-character substring counts.

Cue lists and availability. The 21 added dimensions are driven by 14 manually written cue lists holding 227 entries. The six intensity classes cover degree adverbs (20 entries; "extremely", "completely", "almost"), bodily reactions (23; "heart pounding", "cannot breathe", "hands shaking"), sleep disturbance (12; "cannot fall asleep", "insomnia", "tossing and turning"), anxiety, fear, and anger (22; "anxious", "panic", "breaking down"), pressure events (20; "pressure", "too late", "scolded"), and low activation (24; "calm", "at ease", "numb"); the negation and contrast features use 8 and 11 markers, respectively. The six migration-domain lists hold 13–15 entries each: language (13; "accent", "cannot understand"), documentation (15; "residence permit", "naturalization"), work (15; "overtime", "wages"), family separation (15; "hometown", "cannot go back"), cultural adaptation (15; "custom", "fit in"), and finances (14; "rent", "cannot afford"). Entries are listed here as English glosses and are matched as raw-character substrings, so the features require no segmentation. The complete lists and the feature-extraction implementation are archived on Zen-

Table 1: Labeled source-domain training data. No row comes from the DSA-NIFT target domain; the 253-document internal development split is excluded.

Source	Documents	Unit/domain
EmoBank CVAS	2,582	sentence; news/reviews
EmoBank CVAT	2,969	document; news/reviews
DSA-MST	2,282	medical reflection
ROCLING-2021	1,600	educational reflection
Total	9,433	

[1] https://github.com/fxsjy/jieba

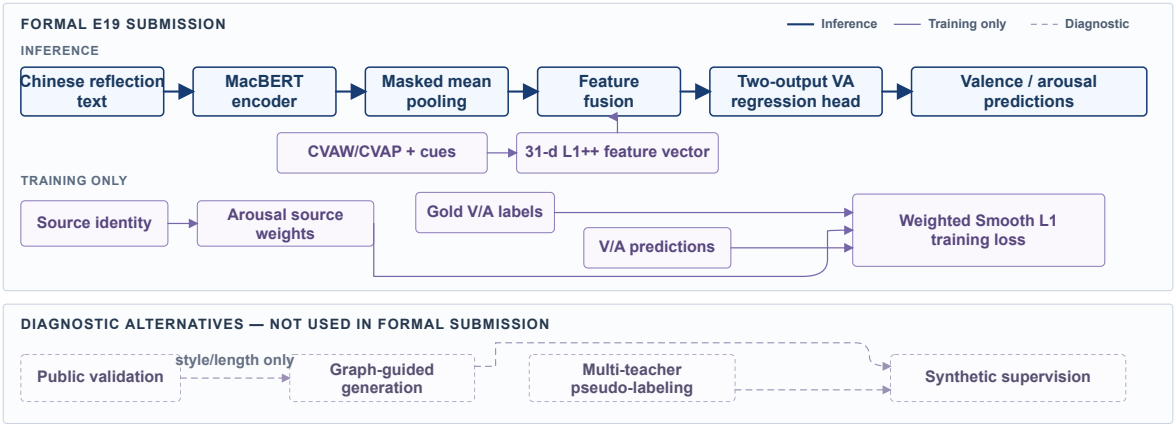

Figure 1: System overview. The upper lane shows the formal E19 system, where MacBERT representations are fused with 31-dimensional L1++ features and source identity affects only the arousal-weighted training loss. The lower lane shows diagnostic alternatives not used in the formal submission; public-validation texts are unlabeled and used only as optional style/length references.

Table 2: Composition of the 31-dimensional L1++ feature vector. Every feature is scaled to $[0, 1]$. Density features count substring matches per 100 characters, cap the density at the listed value, and divide by that cap. The size of each cue list and representative entries are given in the text.

Feature group	Dim.	Computation
Base CVAW/CVAP statistics	10	Token coverage; hit count capped at 20; and V/A mean, maximum, minimum, and standard deviation
Punctuation density	3	Exclamation, question-mark, and ellipsis densities, each capped at 3
Intensity keyword classes	6	Degree adverb, bodily reaction, sleep disturbance, anxiety/fear, pressure event, and low-activation densities, capped at 8/4/3/5/5/4, respectively
Repeated characters	1	Adjacent repeated-character ratio, capped at 0.10
Sentence-length rhythm	3	Mean length divided by 50, length standard deviation divided by 25 (both capped at 1), and fraction of sentences no longer than 10 characters
Negation and contrast	2	Negation and contrast-marker densities, capped at 10 and 5, respectively
Migration-domain cues	6	Language, documentation, work, family separation, cultural adaptation, and financial densities, each capped at 4
Total	31	Concatenated with the pooled encoder representation

odo.[2]

4.4 Source-Aware Supervision

Source-aware training tests whether arousal supervision should depend on corpus proximity. For source s_i, we minimize

$$\mathcal{L}_{\text{src}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{d \in \{V,A\}} w^d(s_i)\, \ell_{\text{SL1}}(\hat{\tilde{y}}_i^d, \tilde{y}_i^d). \quad (4)$$

Valence weights are 1.0 for every real source. The arousal weights reflect two qualitative factors: reflection-style similarity and document granularity. DSA-MST receives the 1.00 reference weight—meaning that its loss is not down-weighted—because it contains multi-sentence, first-person, document-level reflections under the same 1–9 VA setup; its mean raw length is 76.8 characters, versus 19.8 for the educational reflections. CVAS receives 0.25 because it consists primarily of isolated general-domain sentences and therefore mismatches both factors. CVAT retains document granularity but remains general-domain, motivating the intermediate 0.50, whereas the educational corpus shares reflective style but contains substantially shorter texts, motivating 0.75. The quarter-step spacing is a simple ordinal encoding of these priors, not a function of document length or a calibrated domain-similarity score, and it was not selected by a hyperparameter search. E19, our submitted system, combines this loss with the 31-dimensional L1++ feature vector because it achieved the lowest validation A-MAE (0.870) among the systems considered at submission time while retaining an A-PCC of 0.452.

[2] https://doi.org/10.5281/zenodo.22858574

4.5 Graph-Guided Synthetic Supervision

Synthetic augmentation targets VA regions that are scarce in the source data. We train 100-dimensional FastText skip-gram embeddings with character n-grams (Bojanowski et al., 2017) on jieba-segmented source training texts, appending each CVAW entry as a singleton sequence. We then build a cosine k-nearest-neighbor graph ($k = 8$) over the 8,510 CVAW/CVAP entries that receive a FastText vector; the separate embedding-to-VA SVRs in our lexical infrastructure do not determine these graph edges. FastText uses window 5, 20 epochs, and character n-grams of length 1–3. Counting all 8,510 entries as vertices, the undirected graph contains 46,536 edges and has mean degree $2|E|/|V| = 2 \times 46{,}536/8{,}510 \approx 10.94$. An Anthropic-hosted Claude-family generator (recorded model string `claude-opus-4-8`) produces first-person reflections for five intended centers: $(3, 7.5)$, $(5, 7.5)$, $(7, 7.5)$, $(3, 2.5)$, and $(7, 2.5)$. Each condition contains 80 texts per center (400 total), with center labels plus uniform jitter of at most 0.4.

The controlled ladder changes one prompt component at a time when possible: N uses no seed words; A uses random lexicon seeds; C filters seeds by the intended VA range; E expands those seeds through graph neighbors; F adds style examples from the unlabeled 200-text validation set; and F2 further requests lengths drawn from that set. The final E→F2 comparison jointly changes style and length, because F has only one scored test seed; it cannot identify their individual effects.

We also test two alternatives to direct bin-center supervision. E13 averages three teacher predictions (two MacBERT and one RoBERTa), removes examples whose teacher disagreement exceeds 0.60 for valence or 0.75 for arousal, and removes per-bin outliers beyond 1.5 standard deviations, retaining 318 examples. E20 retains E19's L1++, source weights, and real-source regression loss, but uses synthetic labels only to form pairs. For a pair ordered so that intended label i exceeds j by at least 0.1875 (1.5 points on the original scale), its hinge loss on normalized predictions is

$$\mathcal{L}^d_{\text{rank}} = \max\Big(0, 0.125 - (\hat{\tilde{y}}^d_i - \hat{\tilde{y}}^d_j)\Big). \quad (5)$$

The total objective adds $0.10\mathcal{L}^A_{\text{rank}}$ and $0.05\mathcal{L}^V_{\text{rank}}$ to the real-source regression loss.

5 Experimental Setup

5.1 Training Configuration

Core MacBERT systems are trained for up to four epochs with maximum length 256, learning rate 2×10^{-5}, AdamW weight decay 0.01, 10% linear warmup, batch size 32, dropout 0.1, and gradient clipping at 1.0. We retain the epoch maximizing mean(PCC)−mean(MAE) on the 253-document internal development set; this is a checkpoint criterion, not an official metric. Formal training ran on a Google Colab Pro A100 with PyTorch and Transformers ≥ 4.40. The default random seed is 42. The recorded RoBERTa-large configuration replaces the encoder with `hfl/chinese-roberta-wwm-ext-large` and uses batch size 16 and learning rate 10^{-5}. The DAPT checkpoint uses masked-language modeling on the 200 unlabeled target validation texts for 30 epochs with batch size 16, learning rate 5×10^{-5}, and masking probability 0.15 before supervised fine-tuning.

Several historical post-hoc families have documented exceptions. The five E4-family seeds (42, 1, 2, 3, 4) and E13 pseudo-label seeds (42, 1, 2) use batch size 64, while the N/A/C/E/F/F2 augmentation ladder uses batch size 32. For the lexicon and no-augmentation comparisons in Sections 6.5 and 6.6, we rerun No L1, L1, E, and F2 with batch size 32 and seeds 42, 1, and 2; encoder, learning rate, epochs, maximum length, and checkpoint selection are otherwise matched.

5.2 Compared Configurations

Table 3 provides a compact index of the experiment IDs referenced throughout the paper; the text below records the remaining training and supervision details.

E1 trains on the 5,551 EmoBank examples alone, E2 applies DAPT before the same supervised setup, and E3 extends training to all 9,433 labeled sources. The two synthetic-label rows of Table 4 are not indexed above: the raw-synthetic row appends 400 graph-guided texts to E4 with their intended labels, and its shrinkage variant replaces each synthetic label y^d by $\mu_d + 0.6(y^d - \mu_d)$, where μ_d is the real-source mean. E13, E18, and E19 are defined in Section 4. E12 fits its per-dimension fusion weights on the internal development set, and E21b pools concatenated mean and CLS states with separate additive-attention modules and heads per dimension; because no attention-only configuration was

Table 3: Index of frequently referenced experiment configurations. L1 and L1++ denote the 10- and 31-dimensional lexicon feature sets, respectively.

ID	Encoder	Lexicon	Source wt.	Synthetic supervision	Aggregation / pooling
E1	MacBERT	None	No	None	Mean pooling
E4	MacBERT	L1 (10-d)	No	None	Mean pooling
E10	MacBERT	L1 (10-d)	No	None	Five-seed mean
E12	MacBERT + RoBERTa	L1 (10-d)	No	None	Dim.-weighted ensemble
E13	MacBERT	L1 (10-d)	No	Three-teacher pseudo-labels	Mean pooling
E18	MacBERT	L1++ (31-d)	No	None	Mean pooling
E19	MacBERT	L1++ (31-d)	Yes	None	Mean pooling
E20	MacBERT	L1++ (31-d)	Yes	Pairwise constraints	Mean pooling
E21b	MacBERT	L1++ (31-d)	Yes	Pairwise constraints	V/A-specific attention
E22	MacBERT	L1 (10-d)	Yes	None	Mean pooling

scored officially, its attention and pairwise components remain confounded.

5.3 Evaluation Protocol

Our prospective result is the formal E19 submission, selected from public validation and evaluated once on the official test set. The organizers kept the leaderboard open after evaluation; we then submitted fixed checkpoints for method-family, seed, lexicon, and augmentation diagnostics. These post-hoc test scores reveal failure modes but are no longer an untouched estimate for model selection. We report them explicitly to avoid presenting retrospective analysis as prospective performance.

We bootstrap the labeled source-domain development set 5,000 times at $n = 200$ to compare valence and arousal PCC uncertainty on that proxy. This analysis must not be read as an official-validation confidence interval.

6 Results and Analysis

6.1 What Public Validation Suggested

Table 4 reconstructs the official public-validation development path. E4 records the lowest valence MAE (0.600) and a valence PCC of 0.880. Expanding its 10-dimensional feature vector to the 31-dimensional L1++ representation in E18 degrades all four validation metrics, so the expanded representation is not independently beneficial. Holding L1++ fixed, source-aware weighting in E19 reduces arousal MAE from 0.909 to 0.870 and raises arousal PCC from 0.408 to 0.452. We selected E19 because it had the lowest validation A-MAE, and because the only systems with a higher A-PCC were the two synthetic-label conditions (0.461 and 0.460), both of which cost more than 0.18 A-MAE. Thus, the submission decision does not imply that the 31-dimensional representation alone was superior, nor was it determined by PCC alone. Completing the lexicon $\times$ source-weight design, E22 applies the same source weights to the 10-dimensional vector and reaches 0.431 A-PCC and 0.904 A-MAE, against E4's 0.426 and 0.882. Source weighting therefore buys almost no arousal association on top of L1 ($+0.005$ A-PCC) while costing 0.022 A-MAE, whereas the same loss applied to L1++ gains 0.044 A-PCC and 0.039 A-MAE (E18 $\rightarrow$ E19). On public validation, source-aware weighting does not improve both arousal metrics when paired with L1; only with L1++ does it improve both A-PCC and A-MAE, although this comparison rests on single runs.

6.2 Formal Submission Result

On the official test set, the prospective E19 submission obtains 0.620 V-MAE, 0.866 V-PCC, 0.926 A-MAE, and 0.357 A-PCC, ranking fourth of seven teams by mean rank across the four metrics. E19 was selected using public validation and evaluated once on the test set; this is the formal submission result, whereas the later controlled test experiments are post-hoc diagnostics.

6.3 Sampling and Seed Diagnostics

On the labeled medical-reflection development proxy, 5,000 bootstrap resamples of size 200 give an average 95% PCC interval width of approximately 0.083 for valence and 0.193 for arousal, a $2.31\times$ difference; arousal PCC is therefore more sample-sensitive on the proxy, although this is neither an official validation interval nor a paired comparison between systems. Training-seed variation is smaller at the leaderboard's displayed precision: five batch-64 E4-family checkpoints obtain test V-PCC 0.870 ± 0.000 and A-PCC 0.372 ± 0.004 (sample standard deviation). Because the leaderboard rounds most post-hoc scores to two decimals,

Table 4: Official public-validation results ($n = 200$). Lower MAE and higher PCC are better; bold marks the best value in each column. These results reconstruct the development decision and do not establish test efficacy. E22 was scored after the evaluation period and was not available at submission time.

System	V-MAE	V-PCC	A-MAE	A-PCC
EmoBank baseline	0.654	0.867	0.985	0.412
+ DAPT	0.649	0.866	1.009	0.395
+ reflective sources	0.627	0.870	0.944	0.388
E4: + 10-d lexicon	**0.600**	0.880	0.882	0.426
Raw synthetic labels	0.611	0.870	1.100	**0.461**
Synthetic label shrinkage	0.632	0.874	1.058	0.460
Five-seed ensemble	0.614	0.881	0.907	0.415
Multi-encoder ensemble	0.613	**0.882**	0.906	0.418
E13: teacher pseudo-labels	0.617	0.878	0.898	0.423
E18: 31-d intensity	0.641	0.865	0.909	0.408
E19: + source-aware (submitted)	0.666	0.869	**0.870**	0.452
E22: 10-d lexicon + source-aware	0.608	0.881	0.904	0.431
RoBERTa-large	0.691	0.869	0.914	0.407
E20: pairwise synthetic loss	0.631	0.877	0.923	0.407
E21b: E20 + dimension attention	0.638	0.865	0.923	0.394

Table 5: Source-weight sensitivity on the DSA-MST internal development proxy (mean ± sample SD over three seeds). Weights are arousal weights listed as CVAS/CVAT/DSA-MST/EDU2021; valence weights remain 1.0. Lower A-MAE and higher A-PCC are better.

Setting	Weights	A-MAE↓	A-PCC↑
Uniform	1.00/1.00/1.00/1.00	.838±.014	.606±.013
Mild	.50/.75/1.00/1.00	.833±.015	.609±.014
Current	.25/.50/1.00/.75	**.830±.017**	**.612±.016**
Refl. swap	.25/.50/.75/1.00	**.830±.014**	.611±.015

these values are resolution-limited and show only that visible seed variation does not by itself explain the full E19 validation–test change.

6.4 Source-Weight Sensitivity

We performed a small post-hoc sensitivity analysis to test whether E19's behavior depends on its exact heuristic weights. Table 5 compares four predeclared settings on the fixed 253-document DSA-MST internal development split. All settings use MacBERT, L1++, batch size 32, learning rate 2×10^{-5}, four epochs, maximum length 256, the same checkpoint criterion, and seeds 42, 1, and 2. This comparison was not used to choose the formal submission and is not a full hyperparameter search.

Every non-uniform setting improves both mean arousal metrics over uniform weighting (A-MAE by 0.005–0.008, A-PCC by 0.003–0.007), and the current weights give the best unrounded means. The differences are nevertheless smaller than the between-seed sample standard deviations, and current and reflection-swap differ by only 0.0006 A-MAE and 0.0015 A-PCC. The analysis therefore supports down-weighting general-domain arousal supervision without identifying the exact weight values or a strict ordering between the two reflective sources.

6.5 Lexicon Comparison

The L1 lexicon matches at least one entry in 99.82% of official test documents, with mean token coverage 15.59%. Under the matched protocol in Table 6, L1 changes mean A-PCC from 0.363 to 0.370 while mean A-MAE remains 0.923. Mean V-PCC is unchanged at 0.870, and V-MAE changes from 0.617 to 0.613. High coverage therefore corresponds to only a small observed change in target-test performance.

6.6 Synthetic Augmentation Components

The lower block of Table 6 reports the controlled batch-32 augmentation ladder. Within that ladder, adding arbitrary seed words changes mean A-PCC from 0.353 to 0.367. VA filtering (A→C) and graph topology (C→E) each show an approximately 0.003 increase in mean A-PCC based on rounded leaderboard scores. This small downstream difference contrasts with a strong intrinsic effect: graph seed coherence—the proportion of seed word pairs connected by an edge—is 0.179, versus 0.0126 for lookup, or $14.2\times$ larger. The graph changes the control signal, but that change does not clearly transfer to PCC.

We also test the intermediate text-level signal with the same frozen three-teacher ensemble used by E13. C has the smallest mean Euclidean dis-

Table 6: Matched official-test arousal ablation (mean ± sample SD). Unless noted, all rows use MacBERT, three seeds (42/1/2), batch size 32, learning rate 2×10^{-5}, four epochs, maximum length 256, and the same checkpoint-selection criterion. The lower block adds 400 synthetic texts to the L1 reference and varies one prompt component at a time. Valence metrics vary little across all rows (V-MAE .613–.630, V-PCC .867–.870) and are reported in the text. Differences below 0.01 are descriptive because the leaderboard scores are resolution-limited.

Condition	Seeds	A-MAE↓	A-PCC↑
No L1	3	.923±.023	.363±.021
L1 (no aug.)	3	**.923±.040**	.370±.000
N: no seeds	3	1.067±.047	.353±.012
A: random seeds	3	1.090±.070	.367±.015
C: VA lookup	3	1.080±.056	.370±.017
E: graph	3	1.057±.025	.373±.012
F: + style	1	1.080	.380
F2: + length	3	1.073±.115	**.387±.012**

tance between teacher predictions and intended centers (1.631) and the largest teacher-predicted high–low arousal separation (2.173). For E, these values are 1.701 and 2.020, respectively, despite its much higher seed coherence. Thus, under this scorer, graph expansion makes prompt seeds more coherent without making the generated texts more tightly aligned to their VA targets. The estimates are descriptive rather than gold annotations and may inherit the source-domain biases of the teachers.

The matched comparison confirms a PCC–MAE trade-off. E changes mean A-PCC from 0.370 to 0.373 and mean A-MAE from 0.923 to 1.057; F2 changes them to 0.387 and 1.073, respectively (Table 6). No synthetic condition gives a clearly better joint outcome than the no-augmentation reference: F2 provides the largest arousal association, but still incurs a substantial calibration cost relative to that reference. Moreover, F2 only partially matches target length: 31.5% of F2 texts are shorter than 50 characters, compared with 6.0% of the official public-validation texts.

7 Discussion and Limitations

Feature availability versus supervision quality. The validation comparison suggests that adding explicit intensity and migration cues is insufficient by itself. E18 performs worse than E4 on all four official public-validation metrics, indicating that feature availability does not guarantee useful supervision. In contrast, E19 improves both arousal metrics while holding the L1++ feature vector fixed. This pattern suggests that the reliability assigned to source-domain arousal labels may matter more than adding further lexical indicators, although the comparison does not establish a causal mechanism.

Prompt controllability versus label quality. The graph results distinguish prompt-level control from label quality. Graph expansion substantially increases seed coherence, but the teacher-based diagnostics do not show tighter alignment between generated texts and their intended VA centers. One possible explanation is that semantically coherent affective words still permit diverse narrative realizations, while the assigned bin-center labels remain too coarse for regression. Thus, improving the seed selection mechanism alone may not resolve synthetic-label noise.

8 Conclusion

We studied cross-domain valence–arousal prediction without labeled target-domain training data. Our formal system combines affective-lexicon fusion with source-aware arousal supervision and ranked fourth of seven teams in the ROCLING-2026 DSA-NIFT shared task. Across validation and post-hoc diagnostics, arousal models exhibited a recurring PCC–MAE trade-off, suggesting that correlation and calibration should be considered jointly during model selection.

Graph-guided augmentation substantially improved seed-level coherence but produced only small downstream gains, while stronger synthetic conditions improved A-PCC at the cost of MAE. These findings suggest that controllable generation does not necessarily provide reliable regression labels. Future work should explore stronger target-domain validation, comparative arousal supervision, and trusted target annotations for calibrating synthetic data.

References

Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. "you are an expert annotator": Automatic best–worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7924–7936. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and

Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Gavin C. Cawley and Nicola L. C. Talbot. 2010. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11(70):2079–2107.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for Chinese natural language processing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 657–668. Association for Computational Linguistics.

Yu-Chih Deng, Yih-Ru Wang, Sin-Horng Chen, and Lung-Hao Lee. 2023. Toward transformer fusions for Chinese sentiment intensity prediction in valence-arousal dimensions. *IEEE Access*, 11:109974–109982.

Shaz Furniturewala and Kokil Jaidka. 2024. Empaths at WASSA 2024 empathy and personality shared task: Turn-level empathy prediction using psychological indicators. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 404–411. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Sieh-Chuen Huang and Hsuan-Lei Shao. 2025. NTU-LAW at ROCLING-2025 shared task: Domain-adaptive modeling of implicit emotions in medical reflections. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 390–398, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Yi-Min Jian, An Yu Hsiao, and Shih-Hung Wu. 2025. CYUT-NLP at ROCLING-2025 shared task: Valence–arousal prediction in physicians' texts using BERT, RAG, and multi-teacher pseudo-labeling. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 381–389, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Maximilian Köper, Evgeny Kim, and Roman Klinger. 2017. IMS at EmoInt-2017: Emotion intensity prediction with affective norms, automatically extended resources and deep learning. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 50–57. Association for Computational Linguistics.

Lung-Hao Lee, Jian-Hong Li, and Liang-Chih Yu. 2022. Chinese EmoBank: Building valence-arousal resources for dimensional sentiment analysis. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 21(4):65:1–65:18.

Lung-Hao Lee, Tzu-Mi Lin, Hsiu-Min Shih, Kuo-Kai Shyu, Anna S. Hsu, and Peih-Ying Lu. 2025. ROCLING-2025 shared task: Chinese dimensional sentiment analysis for medical self-reflection texts. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 375–380, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Hsin-Chieh Li and Wen-Cheng Lin. 2025. TCU at ROCLING-2025 shared task: Leveraging LLM embeddings and ensemble regression for Chinese dimensional sentiment analysis. In *Proceedings of the 37th Conference on Computational Linguistics and Speech Processing (ROCLING 2025)*, pages 399–406, National Taiwan University, Taipei City, Taiwan. Association for Computational Linguistics.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

Tzu-Mi Lin, Wen-Ni Liu, Yen-Chen Tang, and Lung-Hao Lee. 2026. ROCLING-2026 shared task: Chinese dimensional sentiment analysis for new immigrants' feeling texts. In *Proceedings of the 38th Conference on Computational Linguistics and Speech Processing*.

Michail Mitsios, Georgios Vamvoukakis, Georgia Maniati, Nikolaos Ellinas, Georgios Dimitriou, Konstantinos Markopoulos, Panos Kakoulidis, Alexandra Vioni, Myrsini Christidou, Junkwang Oh, Gunu Jho, Inchul Hwang, Georgios Vardaxoglou, Aimilios Chalamandaris, Pirros Tsiakoulis, and Spyros Raptis. 2024. Improved text emotion prediction using combined valence and arousal ordinal classification. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 808–813. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—a survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855. International Committee on Computational Linguistics.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Liang-Chih Yu, Lung-Hao Lee, Shuai Hao, Jin Wang, Yunchao He, Jun Hu, K. Robert Lai, and Xuejie

Zhang. 2016. Building Chinese affective resources in valence-arousal dimensions. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 540–545, San Diego, California. Association for Computational Linguistics.

Liang-Chih Yu, Jin Wang, Bo Peng, and Chu-Ren Huang. 2021. ROCLING-2021 shared task: Dimensional sentiment analysis for educational texts. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 385–388, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Suyang Zhu, Shoushan Li, and Guodong Zhou. 2019. Adversarial attention modeling for multi-dimensional emotion regression. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 471–480. Association for Computational Linguistics.