aoff123 at ROCLING-2026 Shared Task: A Validation-Guided Protocol for Staged Domain Adaptation in Chinese Dimensional Sentiment Analysis

Chih-Yuan Cheng **Yi-Hsuan Huang** **Bor-Jiunn Hwang**
Ming Chuan University, Taiwan
andy42790380@gmail.com, h20040323@gmail.com
bjhwang@mail.mcu.edu.tw

Abstract

Dimensional sentiment analysis assigns continuous Valence and Arousal scores to text. The DSA-NIFT shared task focuses on Chinese reflections written by new immigrants. We present a validation-guided protocol in which a bounded MacBERT Valence regressor is trained on the CVAT and CVAS portions of Chinese EmoBank and then adapted to educational and medical reflection data. Checkpoints and three fixed interpolation candidates are compared using Valence mean absolute error subject to a Pearson-correlation floor on held-out DSA-MST data, while a frozen reference supplies Arousal predictions. This dimension-specific design separates Valence adaptation from Arousal preservation and treats model selection as a constrained two-metric decision. Convex interpolation preserves the score range, admits an MAE upper bound, and leaves Arousal metrics unchanged across the candidate family. The selected equal-weight blend was used in the official submission, which placed `aoff123` third overall with a mean rank of 3.75. The metric profile motivates joint consideration of error and correlation when developing systems for Chinese dimensional sentiment analysis.

Keywords: dimensional sentiment analysis, domain adaptation, Valence, Arousal, Chinese NLP

1 Introduction

Dimensional sentiment analysis represents affective states as continuous numerical values on dimensions such as Valence and Arousal (Lin et al., 2026; ROCLING-SIGAI, 2026). Computational affect detection spans multiple models and methods across disciplines (Calvo and D'Mello, 2010), while the circumplex model organizes affective concepts in a spatial structure (Russell, 1980). DSA-NIFT extends Chinese dimensional sentiment analysis to reflections written by new immigrants (Lin et al., 2026; ROCLING-SIGAI, 2026), following resources and tasks centered on general affect, educational comments, and medical self-reflections (Lee et al., 2022; Yu et al., 2021; Lee et al., 2025).

The task asks systems to predict scores from 1 to 9 for both dimensions (Lin et al., 2026; ROCLING-SIGAI, 2026). Two design pressures follow. General-purpose annotations provide affective scale supervision, while the target discourse reflects a distinct population and communicative setting. In addition, a candidate that improves one dimension may change the other differently. Prior medical-reflection work used general and domain data for Valence while using domain supervision for Arousal (Huang and Shao, 2025). We therefore treat Valence as the trained and selected stream and Arousal as a frozen reference stream.

We study how a related-domain proxy can guide staged adaptation and dimension-specific composition for a two-output prediction protocol. The paper makes three contributions:

1. We specify bounded, validation-guided Valence fine-tuning from a general Chinese affect resource to related reflective domains, with a joint pointwise-and-correlation training objective.

2. We formalize dimension-specific composition as a constrained, lexicographic rule: Valence candidates are ordered by MAE subject to a PCC floor while the reference Arousal stream remains fixed.

3. We derive interpolation and gate-stability properties, report recovered stage and candidate comparisons, and analyze the official four-metric result profile.

Proceedings of the 38th Conference on Computational Linguistics and Speech Processing (ROCLING 2026), pages 466–476
November 6-7, 2026 ©2026 Association for Computational Linguistics

2 Related Work

2.1 Chinese dimensional sentiment resources

Earlier Chinese resources included a 1,653-word lexicon and a 2,009-sentence corpus annotated in the Valence–Arousal dimensions (Yu et al., 2016). Chinese EmoBank was later constructed to address the scarcity of Chinese resources with real-valued Valence and Arousal annotations. It includes 5,512 words, 2,998 phrases, 2,582 sentences, and 2,969 multi-sentence texts (Lee et al., 2022). It supplies broad affective granularity, with reflective-discourse specialization provided by the later adaptation stage.

Earlier estimators used distributional semantic similarity (Malandrakis et al., 2013), graph neighborhoods (Wang et al., 2016), and structured CNN–LSTM models (Wang et al., 2020). Chinese shared tasks later moved to educational self-evaluation comments (Yu et al., 2021) and medical self-reflections (Lee et al., 2025). Across these settings, pretrained encoders, external affect supervision, validation-set comparison, and prediction composition became directly relevant design choices; Table 1 gives the system-level details.

Recent DimABSA work extends continuous Valence–Arousal modeling to aspect-level annotations across six languages and four domains (Lee et al., 2026). DSA-NIFT poses a complementary text-level problem: each Chinese reflection receives one Valence–Arousal pair, and the educational, medical, and new-immigrant corpora differ in population and communicative purpose. This setting makes discourse-level domain transfer and dimension-specific composition central to the system design.

2.2 Domain and dimension separation

Transfer learning addresses settings in which training and target data differ in feature space or distribution (Pan and Yang, 2010). For pretrained language models, intermediate-task transfer trains a model on an intermediate task before target-task fine-tuning; its behavior is studied across intermediate–target task combinations (Pruksachatkun et al., 2020). Our ordering follows this pattern, using general affect supervision before reflective-domain Valence adaptation.

In this setting, the protocol links validation-selected checkpoints, a finite interpolation family, and an explicit two-metric gate. The resulting decision concerns only the Valence stream; Arousal is fixed from the reference prediction.

2.3 Positioning among shared-task systems

Table 1 organizes the closest Chinese DSA precedents by the design choice each contributes.

Three method families emerge. Encoder fine-tuning supplies a Chinese representation, verification-set comparison supports candidate selection, and dimension-dependent data mixtures permit separate treatment of Valence and Arousal. Our contribution integrates these strands: staged transfer defines a Valence candidate family, and the two-metric gate defines candidate eligibility atop an established Chinese encoder.

3 Task and Data

3.1 Task definition

For each text x_i, the system predicts

$$\hat{\mathbf{y}}_i = (\hat{v}_i, \hat{a}_i), \qquad \hat{v}_i, \hat{a}_i \in [1, 9], \tag{1}$$

where v denotes Valence and a denotes Arousal. The official task contains 200 development texts and 1,100 test texts. Mean absolute error (MAE) and Pearson correlation coefficient (PCC) are computed independently for the two dimensions (Lin et al., 2026; ROCLING-SIGAI, 2026).

3.2 Training resources

Stage 1 uses the CVAT and CVAS portions of Chinese EmoBank as general Valence supervision (Lee et al., 2022; Hung et al., 2021). Stage 2 uses labeled reflective text from the ROCLING 2025 DSA-MST medical domain (Lee et al., 2025) together with ROCLING 2021 educational comments (Yu et al., 2021). These corpora add reflection-oriented supervision while retaining distinct populations and communicative settings.

Table 2 assigns each data source one empirical role: general supervision, reflection-domain adaptation, local selection, or target evaluation.

4 Method

The protocol combines general-domain Valence training, reflective-domain adaptation, full-split checkpointing, fixed interpolation, and dimension-specific composition: the rule selects one Valence path while Arousal remains fixed.

Figure 1 summarizes model fitting, local selection, and composition within the proposed protocol.

System	Reflection domain	Key method	Connection to this workflow
NCU 2021 (Hung et al., 2021)	Educational	Fine-tunes MacBERT using CVAT and CVAS portions of Chinese EmoBank.	Supports MacBERT and external Chinese affect supervision as a starting point.
NTUST 2021 (Wang et al., 2021)	Educational	Compares BERT and MacBERT on a verification set and averages several model settings.	Motivates validation before prediction composition.
TCU 2025 (Li and Lin, 2025)	Medical	Feeds contextual embeddings from large language models to regressors including SVR.	Demonstrates a representation–regression alternative for reflection text.
NTULAW 2025 (Huang and Shao, 2025)	Medical	Uses Chinese EmoBank plus domain data for Valence and domain supervision for Arousal.	Motivates dimension-dependent supervision for reflection text.
This work	New-immigrant	Specifies general and reflection-domain Valence training, a fixed two-metric gate, and Arousal from a frozen multiseed reference.	Combines transfer and dimension-specific selection in one validation-guided protocol.

Table 1: Closest shared-task precedents and the design thread carried into DSA-NIFT.

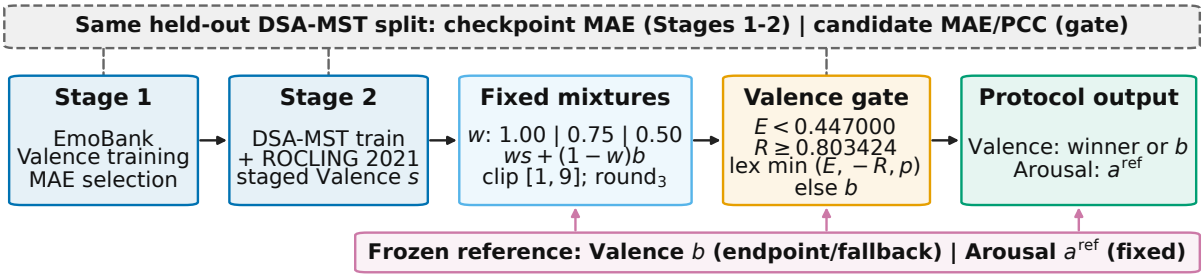

Figure 1: Staged-Valence training, interpolation, and selection workflow. The predefined weight order supplies the final tie-break.

Resource	Role in the study
EmoBank CVAT/CVAS	General Chinese Valence supervision in Stage 1
ROCLING 2021	Educational-reflection adaptation data in Stage 2
DSA-MST 2025	Medical-reflection adaptation data and a disjoint local selection split
DSA-NIFT 2026	Unlabeled target inputs for official evaluation; official metrics used only for reporting

Table 2: Separation of supervision, local selection, and target evaluation.

4.1 Frozen reference prediction

The reference averages two independently fine-tuned, two-output MacBERT regressors initialized from `Pectics/vad-macbert`, using seeds 42 and 13. Each head has the same dropout–linear–GELU–linear architecture as the staged regressor below, with two bounded outputs instead of one. The recorded configuration uses the 0.8 Smooth L1 plus 0.2 PCC objective, two epochs, learning rate 1.5×10^{-5}, maximum length 128, batch size 4, accumulation 2, and the final six encoder layers trainable. These settings come from retained training commands and model artifacts; complete per-run configuration logs were not saved.

For local selection, each reference component uses 5,553 processed EmoBank rows plus 2,028 DSA-MST training rows. The test-time counterparts use the same EmoBank input plus all 2,535 DSA-MST rows; they do not use DSA-NIFT test labels. Thus local and target reference vectors come from separate fits with the same model family. Equal-weight averaging, clipping, and three-decimal rounding reproduce both retained reference CSVs exactly from their two component prediction files. The ROCLING 2021 educational data enter the staged branch, not this reference.

The serialized local reference has Valence MAE 0.447008 and PCC 0.813425, and Arousal MAE 0.835033 and PCC 0.627744. Earlier coarse-report values were 0.447000, 0.813424, 0.835036, and 0.627739, respectively; the small differences arise from serialization/rounding. We retain the historically configured gate constants, $E < 0.447000$ and $R \geq 0.803424$, while reporting metrics recomputed from the actual CSVs. Reference Arousal is copied unchanged into every candidate.

Why preserve Arousal? Freezing was a conservative scope choice: compare Valence adaptation without changing Arousal in the submitted candidate family. It was not established as an optimal Arousal policy. A new, post-acceptance experiment applies the same staged training configuration to Arousal (Section 6.3). Its Stage-2 model improves both local Arousal metrics relative to the frozen reference. We therefore distinguish the historical submission choice from the stronger Arousal alternative found after acceptance; the official scores still belong to the frozen-Arousal submission.

4.2 Bounded Transformer regression

The protocol initializes the encoder from `Pectics/vad-macbert`, a MacBERT checkpoint pretrained for Chinese VAD regression (Pectics, 2026). This model follows the pretrained Transformer and downstream fine-tuning paradigm established by BERT (Devlin et al., 2019); MacBERT changes the masking strategy to use masked language modeling as correction (Cui et al., 2020). Earlier work also demonstrated its use for Chinese dimensional sentiment regression (Hung et al., 2021).

The regressor consumes the encoder's pooled output when available and otherwise uses an attention-mask-weighted mean of the final hidden states. Its head applies dropout 0.1, projects the encoder width to half that width, applies GELU and a second dropout 0.1, and produces one scalar. We bound the prediction by

$$\hat{v} = 1 + 8\,\sigma(z), \tag{2}$$

so the model output respects the task range. The pointwise term is Smooth L1; Huber's formulation provides its robust-estimation foundation (Huber, 1964).

The staged model optimizes a single Valence output with

$$\mathcal{L} = 0.8\,\mathcal{L}_{\mathrm{SL1}} + 0.2\,(1 - \rho_B), \tag{3}$$

where ρ_B is PCC computed across the current training batch, with its denominator lower-bounded by 10^{-8} for numerical stability. The first term controls pointwise error, while the second rewards within-batch linear association. The PCC term is applied only to batches containing at least three examples; the configuration uses four and drops an undersized final batch.

The two loss terms resolve different aspects of the regression problem. For a nondegenerate batch on which the numerical floor is inactive, ordinary correlation is unchanged by a positive affine transformation, so it cannot identify the score location or scale; Section 7 formalizes this property. Smooth L1 anchors predictions to the annotated units, while the sigmoid head enforces the task support. When a batch has nearly zero variance, the denominator floor bounds the correlation term and the pointwise term continues to provide a learning signal. Full-split MAE checkpointing then evaluates the quantity given primary priority by the gate.

Setting	Stage 1	Stage 2
Training data	EmoBank CVAT + CVAS	DSA-MST + ROCLING 2021
Initialization	`vad-macbert`	Stage-1 checkpoint
Learning rate	1.5×10^{-5}	8×10^{-6}
Epoch cap	4	4
Validation	DSA-MST selection split	Same split
Batch / accumulation	4 / 2	4 / 2
Length / seed	128 / 42	128 / 42
Trainable encoder	Final six layers	Final six layers

Table 3: Training configuration for the two Valence stages. Both use early stopping with patience 2 and minimum delta 0.002.

Both stages use AdamW with weight decay 0.01, 8% linear warmup, gradient-norm clipping at 1.0, and two four-example mini-batches per optimizer step; an odd final mini-batch forms its own step. The head and final six encoder layers are trainable, while lower layers remain frozen. Early stopping uses the minimum delta, whereas checkpoint saving tracks the exact lowest MAE and retains the earlier epoch on a tie.

4.3 Objective-to-selection alignment

The mini-batch objective guides parameter updates; with batch size four, its PCC component acts as a stochastic association signal. After each epoch, exact full-split MAE selects the checkpoint. The gate then evaluates each fixed interpolation candidate by full-split MAE and PCC, preserving the priority of absolute error at selection time.

The sigmoid head keeps learned Valence outputs in $[1, 9]$; clipping before three-decimal rounding ensures submitted scores remain in the valid range.

4.4 Staged Valence adaptation

The first stage learns a general Chinese text-to-Valence mapping from the CVAT and CVAS portions of Chinese EmoBank. The second initializes from the lowest-MAE Stage-1 checkpoint and adapts on combined DSA-MST and ROCLING 2021 training data at the lower learning rate in Table 3. Each epoch is evaluated on the held-out DSA-MST split, and training restores the state with the lowest Valence MAE, following the validation-based early-stopping principle (Prechelt, 1998).

This ordering encodes a concrete transfer hypothesis: broad affect supervision initializes the score scale, and the later stage specializes the representation for reflective discourse. Fixing the ordering also makes subsequent candidate comparisons share the same training path.

4.5 Fixed interpolation and dimension-specific composition

Let s_i be the staged Valence prediction and b_i the reference Valence prediction. The three candidate values are

$$\tilde{v}_i(w) = \mathrm{round}_3\Big(\mathrm{clip}_{[1,9]}(ws_i + (1-w)b_i)\Big), \tag{4}$$

$$w \in \{1.00, 0.75, 0.50\}. \tag{5}$$

The candidate policy fixes these weights before target evaluation. For every w, the Arousal output is $\tilde{a}_i = a_i^{\mathrm{ref}}$; only Valence changes. This construction makes the Valence comparison independent of the fixed reference Arousal stream.

4.6 Gated decision rule

For candidate set $\mathcal{C}$, define $E(c)$ as Valence MAE and $R(c)$ as Valence PCC on the DSA-MST selection split. Let $p(c) \in \{1,2,3\}$ be the position of c in the fixed weight order $(1.00, 0.75, 0.50)$. The eligible set and selected candidate are

$$\mathcal{G}_E = \{c \in \mathcal{C} : E(c) < 0.447000\}, \tag{6}$$

$$\mathcal{G}_R = \{c \in \mathcal{C} : R(c) \geq 0.803424\}, \tag{7}$$

$$\mathcal{G} = \mathcal{G}_E \cap \mathcal{G}_R, \tag{8}$$

$$c^\star = \begin{cases} \underset{c\in\mathcal{G}}{\overset{\mathrm{lex}}{\arg\min}} \big(E(c), -R(c), p(c)\big), & \mathcal{G} \neq \emptyset, \\ c_{\mathrm{ref}}, & \mathcal{G} = \emptyset. \end{cases} \tag{9}$$

where $c_{\mathrm{ref}} = (b, a^{\mathrm{ref}})$ denotes the unblended reference output. The tuple prioritizes lowest MAE, highest PCC, then the earlier predefined weight on an exact metric tie. The MAE boundary uses the historical 0.447000 proxy constant; the PCC floor permits 0.01 below its 0.813424 value. Equivalently, $m_E(c) = 0.447000 - E(c) > 0$ and $m_R(c) = R(c) - 0.803424 \geq 0$. If no candidate is eligible, the selector retains the unblended reference.

Geometrically, $\mathcal{G}$ is an axis-aligned feasible region in the MAE–PCC plane. When it is nonempty, the selected candidate is Pareto-undominated within it: a candidate with lower MAE, or equal MAE and higher PCC, would precede it. Full-precision metrics determine eligibility near either boundary because rounded display values can alter either threshold comparison.

4.7 Local guarantees of selection

Every nonfallback output satisfies the configured MAE and PCC thresholds by construction. Fallback returns the serialized reference, whose recomputed MAE is 0.447008 rather than exactly 0.447000. Thus the MAE envelope including fallback is $E(c^\star) \leq \max(0.447000, E_{\mathrm{ref}})$, with strict improvement over the reference for an eligible staged candidate. All candidates and fallback retain exactly the same serialized Arousal vector. These are properties on the selection split, not guarantees on target data.

4.8 Analytical properties of composition

Let $u_i(w) = ws_i + (1-w)b_i$ denote the unrounded Valence mixture. Because $[1,9]$ is convex, $s_i, b_i \in [1,9]$ implies $u_i(w) \in [1,9]$ for every $w \in [0,1]$. Moreover, for any gold score v_i,

$$|v_i - u_i(w)| \leq w|v_i - s_i| + (1-w)|v_i - b_i|. \tag{10}$$

Averaging this pointwise inequality and writing $M_x = \mathrm{MAE}(v, x)$ gives

$$M_{u(w)} \leq wM_s + (1-w)M_b. \tag{11}$$

Thus interpolation has a direct absolute-error bound; the ordering of an interior weight and the two endpoints remains an empirical question.

Projection onto $[1,9]$ cannot increase distance from a gold score in that interval. Three-decimal rounding introduces at most 5×10^{-4} per prediction, so the reverse triangle inequality yields

$$|\mathrm{MAE}(v, \mathrm{round}_3(u)) - \mathrm{MAE}(v, u)| \leq 5\times10^{-4}. \tag{12}$$

PCC has no corresponding convex ordering: both the covariance with v and the variance of $u(w)$ depend on the weight and on covariance between s and b. This difference motivates optimizing MAE while retaining a separate correlation floor.

The dimension-specific construction supplies a second exact property. Let the four-metric vector for a candidate weight be

$$\mathbf{F}(w) = \big(M_V(w), R_V(w), M_A(w), R_A(w)\big). \tag{13}$$

Since $\tilde{a}_i = a_i^{\mathrm{ref}}$ for all w,

$$\mathbf{F}(w) = \big(M_V(w), R_V(w), M_A^{\mathrm{ref}}, R_A^{\mathrm{ref}}\big). \tag{14}$$

Thus candidate comparison can change only the two Valence coordinates; Arousal MAE and PCC

are constant across the candidate family. The two-metric gate is therefore the potentially nonconstant part of the four-metric comparison, rather than a trade-off against Arousal within that family.

4.9 Finite-grid implications

The original grid samples the staged endpoint and two increasingly conservative moves toward the reference. The three weights and PCC tolerance were engineering choices, not empirically established optima. Before rounding, MAE is convex in the weight, but PCC need not be. We therefore supplement this grid with a post-acceptance sensitivity analysis over the recovered prediction vectors in Section 6.2; this analysis does not alter the submitted system.

5 Experimental Setup

5.1 Selection split and evaluation roles

The protocol joins the released DSA-MST validation and test text/answer tables one-to-one, verifies exact coverage and finite labels in $[1, 9]$, pools them, and constructs a deterministic 80/20 train–selection split with seed 42.

Cut points $[1, 3, 5, 7, 9]$ form a joint 4×4 Valence–Arousal grid; cells with fewer than two examples enter a rare-cell stratum. Stratification is used when every stratum can contribute to both partitions; otherwise seed 42 defines a shuffled unstratified split. The resulting split contains 2,028 training and 507 selection rows, with disjoint IDs. A post-acceptance exact-text audit found one selection text also present in the Stage-2 training input under a different ID; the historical split was not text-deduplicated. Section 6.4 reports a training-side cleaned rerun. Stage 2 adds 1,600 educational rows for 3,628 training rows. The processed EmoBank input contains 2,970 CVAT and 2,583 CVAS rows; these are the retained file counts, which differ slightly from the published resource counts in Section 2.

The DSA-MST split supports checkpoint and blend decisions; official target metrics support only reporting. The recovered configuration specifies the seed, binning rule, epoch cap, candidate weights, thresholds, and tie-break. Earlier candidate exploration also used this split, so these results are selection statistics rather than an independent estimate. Section 7.3 examines this selection depth and proxy-to-target interpretation.

5.2 Metrics

The primary metric is Valence MAE:

$$\mathrm{MAE}(v, \hat{v}) = \frac{1}{n} \sum_{i=1}^{n} |v_i - \hat{v}_i|. \tag{15}$$

PCC is computed as

$$\rho(v, \hat{v}) = \frac{\sum_i (v_i - \bar{v})(\hat{v}_i - \bar{\hat{v}})}{\sqrt{\sum_i (v_i - \bar{v})^2} \sqrt{\sum_i (\hat{v}_i - \bar{\hat{v}})^2}}. \tag{16}$$

MAE is sensitive to absolute calibration, whereas PCC measures centered linear association. The official evaluation computes both measures independently for Valence and Arousal, yielding four complementary views of each submission (Lin et al., 2026; ROCLING-SIGAI, 2026).

5.3 Gate stability and evaluation scope

The gate margins quantify decision stability. Suppose a recomputation changes every candidate metric by at most ϵ_E in MAE and ϵ_R in PCC:

$$|E'(c) - E(c)| \le \epsilon_E, \qquad |R'(c) - R(c)| \le \epsilon_R. \tag{17}$$

Candidate c keeps the same eligibility status whenever $|m_E(c)| > \epsilon_E$ and $|m_R(c)| > \epsilon_R$. If all memberships are stable and the selected candidate has an MAE lead greater than $2\epsilon_E$ over every other eligible candidate, its identity is also stable. The first condition protects threshold membership, while the second protects the ordering among eligible candidates. Together, they expose both threshold proximity and winner separation and provide a direct interpretation of numerical sensitivity around the gate.

Local decisions use DSA-MST, whereas official evaluation uses DSA-NIFT. The recovered decision record binds the local truth, reference, staged predictions, and selected Stage-2 checkpoint by SHA-256. Recomputing all three candidates reproduces its full-precision metrics and selects $w = 0.50$. On target inputs, blending the retained staged and reference Valence vectors at this weight and copying reference Arousal reproduces all 1,100 official submitted rows exactly. The official CSV hash also matches the organizer-downloaded artifact for submission #849556. This identifies the submitted construction without treating the aggregate ranking as a component ablation.

Valence configuration	MAE	PCC	Gate
Frozen reference	0.447008	0.813425	fallback
Stage 1	0.543568	0.764630	–
Stage 2 ($w = 1.00$)	0.449939	0.806845	fail
Blend ($w = 0.75$)	0.443213	0.812371	pass
Blend ($w = 0.50$)	**0.440430**	**0.815087**	selected

Table 4: Recovered local selection results. The gate tests only the three original Stage-2 candidates; Stage 1 is an intermediate diagnostic.

n	Arousal predictor	MAE	PCC
507	Frozen reference	0.835033	0.627744
507	Stage 1	0.930815	0.556935
507	Stage 2	**0.815242**	**0.653563**
506	Frozen reference	0.834420	0.626401
506	Stage 1	0.929913	0.555987
506	Stage 2	**0.814871**	**0.652147**

Table 5: Post-acceptance Arousal comparison. The 506-row subset excludes the one overlapping text; it is not a new untouched split. All predictions are exported to three decimals.

6 Results

6.1 Recovered stage and candidate comparisons

Table 4 reports metrics on the same 507 selection rows. Stage-1 inference was recomputed from the preserved checkpoint after acceptance; other rows were recomputed from retained prediction files. The Stage-1 training log records best epoch 3 with unrounded MAE 0.543559, and the MAE-selected Stage-2 log records epoch 2 with 0.449928. Three-decimal prediction export explains the small differences from the table.

Stage 2 reduces MAE relative to Stage 1 by 0.093629, but does not beat the reference MAE alone. Interpolation at $w = 0.50$ improves reference MAE by 0.006578 and PCC by 0.001662. The gate excludes the pure staged endpoint; among the original three weights, it selects the same candidate as MAE alone, so these results do not demonstrate an additional performance gain caused by the PCC constraint. All three candidates have Arousal MAE 0.835033 and PCC 0.627744 because their Arousal vectors are identical. These comparisons are conditional on one reused selection split, without a significance claim.

6.2 Post-acceptance sensitivity analysis

We evaluated 101 weights $w \in \{0, 0.01, \ldots, 1\}$ using the recovered local endpoints, with the same clipping and three-decimal rounding. The lowest-MAE weight is 0.52 (MAE 0.440410, PCC 0.814973), only 0.000020 lower in MAE than the submitted 0.50. This small difference does not establish a meaningful generalization advantage for the finer grid.

We also replaced the PCC floor with $0.813424 - \delta$ for $\delta \in \{0, 0.0025, 0.005, 0.01, 0.02, 0.05\}$, retaining the strict MAE threshold. All six settings select 0.50 on the original grid and 0.52 on the finer grid. At zero tolerance, the 0.75 candidate becomes ineligible, but the selected weight is unchanged. Thus the local selection is insensitive to these tested tolerances; this retrospective sweep is not evidence of stability across seeds, splits, or target domains. Excluding the one overlapping text leaves 506 rows: the 0.50 blend still has the lowest original-grid MAE (0.439073; PCC 0.815101), compared with reference MAE 0.445573. This post-hoc exclusion does not undo possible checkpoint-selection exposure.

6.3 Matched staged-Arousal ablation

After acceptance, we trained a single-output Arousal regressor using the same initialization, data files, seed 42, two learning rates, epoch caps, loss weights, batch/accumulation sizes, length limit, and trainable layers as the Valence protocol. Only the prediction target and target-specific MAE checkpoint criterion change. Stage 2 initializes from the best Stage-1 Arousal checkpoint. Both stages ran four epochs and selected epoch 2. This matches the historical hyperparameters, not necessarily the floating-point trajectory across software environments; the new run used PyTorch 2.5.1, Transformers 5.9.0, and a GTX 1650.

Table 5 shows that general-domain Stage 1 alone is worse than the reference, whereas Stage 2 lowers MAE by 0.019791 and raises PCC by 0.025819 on all 507 rows. Excluding the overlapping text retains this ordering. The result does not support superiority of freezing Arousal. This single-seed comparison reuses selection data; Section 6.4 addresses the training-side overlap by retraining. No target labels or new official evaluation were used, and these improvements are not attributed to the submitted system.

6.4 Training-side overlap removal

We next removed training texts matching any of the unchanged 507 selection texts after Unicode NFKC normalization and whitespace removal. One row was removed from each of the Stage-2 and local-

Target	Clean-training predictor	MAE	PCC
V	Reference	0.452260	0.811924
V	Stage 2 ($w = 1.00$)	0.446036	0.811219
V	Blend ($w = 0.75$)	0.442609	0.814783
V	Blend ($w = 0.50$)	**0.441742**	**0.815829**
A	Reference	0.841850	0.621976
A	Stage 2	**0.810347**	**0.648269**

Table 6: Post-acceptance retraining after training-side normalized-text overlap removal; all 507 selection rows remain. V and A denote Valence and Arousal.

reference inputs, leaving 3,627 and 7,580 rows. Stage-1 data had no such matches, so its target-specific checkpoints were retained. Both Stage-2 models were retrained with seed 42 and the same settings; each stopped after epoch 3 and selected epoch 1. Both joint-output reference components were retrained with seeds 42 and 13 for the documented two epochs, without validation checkpointing, then averaged.

Using the retrained reference MAE and PCC minus 0.01 as the gate thresholds, all three Valence candidates pass and $w = 0.50$ again wins (Table 6). The same candidates pass the historical thresholds. The blend improves reference MAE by 0.010519; staged Arousal improves MAE by 0.031503 and PCC by 0.026293. Input hashes, prediction alignment, metrics, and gate decisions were independently checked. This removes the identified training-side text overlap, but not selection reuse, possible near-duplicates, or pretraining overlap. Runtime differences also prevent attributing changes from historical scores solely to removing one row. No target labels or new official scores were used.

6.5 Official evaluation

The organizers' official ranking snapshot, distributed to participating teams on July 17, 2026, lists `aoff123` third overall among the seven displayed teams.[1] Each of the four metrics is ranked independently, and mean metric rank determines the overall ordering (Lin et al., 2026; ROCLING-SIGAI, 2026). Table 7 transcribes the snapshot; $(3 + 3 + 7 + 2)/4 = 3.75$.

The table reports the byte-verified $w = 0.50$ submission, not the causal contribution of each component or the new retraining runs. Valence ranks third under both metrics; Arousal combines rank 7 under MAE with rank 2 under PCC. This metric-specific profile motivates examining calibration separately from association.

[1] ROCLING-2026 DSA-NIFT Shared Task organizers, *Official Ranking*, distributed to participating teams on July 17, 2026.

7 Analysis

7.1 Metric geometry and calibration

PCC and MAE respond differently to calibration. For any $\alpha > 0$ and constant β,

$$\mathrm{PCC}(y, \alpha\hat{y} + \beta) = \mathrm{PCC}(y, \hat{y}), \tag{18}$$

whereas MAE generally changes under the same transformation. Preserved linear association can therefore coexist with a location or scale mismatch, a pattern compatible with the official Arousal ranks of second for PCC and seventh for MAE. The aggregate ranks identify this calibration question without establishing its cause.

Any future offset or scale calibration must be fitted on a separate calibration partition and evaluated after freezing. Clipping to $[1, 9]$ can alter correlation even when the unclipped transformation preserves it.

7.2 Residual complementarity

The source of a possible interpolation gain can be stated at the example level. Let $e_i^s = s_i - v_i$ and $e_i^b = b_i - v_i$ be signed residuals. Before rounding, the mixed residual is

$$e_i(w) = we_i^s + (1 - w)e_i^b. \tag{19}$$

When component residuals have the same sign, $|e_i(w)|$ lies between their absolute errors. Opposite signs permit partial cancellation, with zero-error weight

$$w_i^\circ = \frac{|e_i^b|}{|e_i^s| + |e_i^b|}. \tag{20}$$

At split level,

$$L(w) = \frac{1}{n} \sum_i \left| (1 - w)e_i^b + we_i^s \right| \tag{21}$$

is a convex piecewise-linear function of w before clipping and rounding. Its interior breakpoints are precisely the w_i° values for examples whose two endpoint residuals have opposite signs. No single corpus-level weight matches every breakpoint; the fixed grid instead evaluates three predefined points on this path. Gold residuals on the local selection split determine whether these cancellations reduce aggregate error.

Team	V MAE ↓	V PCC ↑	A MAE ↓	A PCC ↑	Mean rank	Overall
CYUT	0.5366 (1)	0.8971 (1)	0.8658 (1)	0.4376 (1)	1.00	1
SCUNLP	0.5793 (2)	0.8815 (2)	0.8951 (2)	0.4050 (3)	2.25	2
aoff123	**0.6018 (3)**	**0.8688 (3)**	**0.9403 (7)**	**0.4052 (2)**	**3.75**	**3**
DonkiDonki	0.6200 (4)	0.8663 (4)	0.9259 (4)	0.3566 (6)	4.50	4
BEDO	0.6418 (5)	0.8619 (5)	0.8992 (3)	0.3204 (7)	5.00	5
wuhongyi	0.6495 (6)	0.8586 (6)	0.9294 (5)	0.3366 (6)	5.75	6
845310	0.6684 (7)	0.8542 (7)	0.9341 (6)	0.3581 (5)	6.25	7

Table 7: Official DSA-NIFT ranking snapshot distributed on July 17, 2026. Parentheses denote the metric ranks reported by the organizers.

For correlation, writing C_{xy} for sample covariance gives

$$\begin{aligned}\sigma^2_{u(w)} &= w^2\sigma_s^2 + (1-w)^2\sigma_b^2 \\ &\quad + 2w(1-w)C_{sb}, \\ \rho(v, u(w)) &= \frac{wC_{vs} + (1-w)C_{vb}}{\sigma_v \sigma_{u(w)}}. \end{aligned} \tag{22}$$

The numerator combines alignment with the gold labels, while the denominator also depends on agreement between predictors. Equation 22 explains why correlation may peak at an interior weight, decline alongside an MAE improvement, or remain nearly unchanged. The PCC floor therefore adds a constraint not supplied by the MAE bound.

7.3 Selection depth and transfer scope

The staged curriculum moves from heterogeneous Chinese affect text to educational and medical reflections, then targets new-immigrant reflections. The same DSA-MST selection partition participates in three successive decisions: choosing the Stage-1 checkpoint, choosing the Stage-2 checkpoint, and ordering the fixed mixtures. Its values are consequently decision statistics for the specified procedure rather than an untouched post-selection estimate. A finite candidate set limits the final decision, while repeated use of evaluation feedback still calls for this explicit interpretation (Dwork et al., 2015).

Transfer-learning analysis distinguishes proxy and target distributions (Pan and Yang, 2010; Ben-David et al., 2010). Proxy ordering alone does not establish target ordering; the official ranking supplies a separate team-level target result.

Dimension-specific composition is consistent with prior Chinese medical-reflection work that used different training-data mixtures for Valence and Arousal (Huang and Shao, 2025). Within the candidate family, Equation 14 proves Arousal invariance and characterizes the candidate construction.

7.4 Implications for metric-aware development

The official ranks $(3, 3, 7, 2)$ average to 3.75, while the raw scores retain gaps omitted by ranks. Arousal's seventh-place MAE and second-place PCC motivate separating calibration from association. The post-acceptance Arousal comparison additionally shows that reflective-domain adaptation can improve both proxy metrics, rather than requiring a choice between them.

Checkpoint selection, blending, and calibration should use explicit data roles. Text-deduplicated splits, nested validation, or cross-fitting can reduce overlap and decision reuse. Independent seeds and folds are needed to assess whether the observed gains persist beyond this particular selection split. The official ranking remains a result of the historical submission, not evidence for the newly trained Arousal model.

8 Conclusion

We specify a dimension-specific DSA-NIFT protocol with staged Valence adaptation, fixed blends, a two-metric gate, and a fixed reference Arousal stream. Recovered checkpoint and candidate results support the selected equal-weight blend, while a retrospective sensitivity sweep finds little benefit from a finer grid. The official evaluation placed `aoff123` third overall with a mean rank of 3.75. Recovered prediction endpoints reproduce the official artifact exactly. A post-acceptance matched Arousal experiment improves both proxy metrics over freezing, so the historical choice is not claimed to be optimal. Retraining after training-side text-overlap removal retains the equal-weight Valence winner and the staged-Arousal advantage. Selection reuse and a single staged seed remain limitations; independent multi-seed evaluation remains future work.

References

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. 2010. A theory of learning from different domains. *Machine Learning*, 79(1–2):151–175.

Rafael A. Calvo and Sidney D'Mello. 2010. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37.

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, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toni Pitassi, Omer Reingold, and Aaron Roth. 2015. Generalization in adaptive data analysis and holdout reuse. In *Advances in Neural Information Processing Systems 28*, pages 2350–2358. Curran Associates, Inc.

Sieh-chuen Huang and Hsuan-Lei Shao. 2025. NTULAW 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. Association for Computational Linguistics.

Peter J. Huber. 1964. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101.

Man-Chen Hung, Chao-Yi Chen, Pin-Jung Chen, and Lung-Hao Lee. 2021. NCU-NLP at ROCLING-2021 shared task: Using MacBERT transformers for dimensional sentiment analysis. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 380–384, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

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.

Lung-Hao Lee, Liang-Chih Yu, Natalia V. Loukachevitch, Ilseyar Alimova, Alexander Panchenko, Tzu-Mi Lin, Zhe-Yu Xu, Jian-Yu Zhou, Guangmin Zheng, Jin Wang, Sharanya Awasthi, Jonas Becker, Jan Philip Wahle, Terry Ruas, Shamsuddeen Hassan Muhammad, and Saif M. Mohammad. 2026. DimABSA: Building multilingual and multidomain datasets for dimensional aspect-based sentiment analysis. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 40492–40518, San Diego, California, United States. 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.

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*. Forthcoming; citation metadata supplied by the shared-task organizers; URL points to the official task page pending a proceedings record.

Nikolaos Malandrakis, Alexandros Potamianos, Elias Iosif, and Shrikanth Narayanan. 2013. Distributional semantic models for affective text analysis. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(11):2379–2392.

Sinno Jialin Pan and Qiang Yang. 2010. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359.

Pectics. 2026. vad-macbert. Hugging Face model repository. Accessed 2026-08-07; MIT license.

Lutz Prechelt. 1998. Early stopping—but when? In Genevieve B. Orr and Klaus-Robert Müller, editors, *Neural Networks: Tricks of the Trade*, volume 1524 of *Lecture Notes in Computer Science*, pages 55–69. Springer, Berlin, Heidelberg.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

ROCLING-SIGAI. 2026. ROCLING 2026 Shared Task. Official shared-task website. Accessed 2026-08-08.

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

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2016. Community-based weighted graph model for valence-arousal prediction of affective words. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 24(11):1957–1968.

Jin Wang, Liang-Chih Yu, K. Robert Lai, and Xuejie Zhang. 2020. Tree-structured regional CNN–LSTM model for dimensional sentiment analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:581–591.

Yi-Wei Wang, Wei-Zhe Chang, Bo-Han Fang, Yi-Chia Chen, Wei-Kai Huang, and Kuan-Yu Chen. 2021. ntust-nlp-1 at ROCLING-2021 shared task: Educational texts dimensional sentiment analysis using pretrained language models. In *Proceedings of the 33rd Conference on Computational Linguistics and Speech Processing (ROCLING 2021)*, pages 354–359, Taoyuan, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

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).