TD-AFGW: TONE-AWARE MULTIMODAL EMOTION TRANSFER WITHOUT TARGET EMOTION LABELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Transferring emotion recognition to tonal languages with scarce annotations requires addressing cross-lingual representation shift and the overlap between lexical tone and emotional prosody. We introduce TD-AFGW, a multimodal framework trained with labeled source conversations and target conversations without emotion labels. A text-conditioned pitch predictor, ToneNet, estimates a lexical-tone baseline and supplies residual pitch features and feature-wise acoustic conditioning. Anchored fused Gromov–Wasserstein alignment combines within-domain relations with fixed facial and paralinguistic cues. The transport plan propagates source labels, while a tone-swap consistency gate selects target soft labels for distillation. A visual-to-audio-to-text curriculum progressively enables adaptation. We study English-to-Mandarin transfer from MELD to EmotionTalk, using Mandarin as an established tonal-language benchmark for task-specific annotation scarcity. Under shared source-prior logit calibration, the full model achieves 31.42 ± 0.38 weighted F1 and 18.01 ± 0.35 macro F1 across three seeds, versus 27.59 ± 0.32 and 15.29 ± 0.28 for source-only transfer. Ablations support the contributions of anchored geometry, ToneNet, and consistency filtering. Per-class results show particularly strong gains for anger and surprise, alongside persistent minority-class weaknesses. The findings support tone-aware transfer on this benchmark; broader low-resource generalization and fully target-label-blind model selection remain to be established.

1 INTRODUCTION

Emotion annotations are costly to obtain, particularly when the same utterance must be interpreted across language, speech, and facial behavior. Transferring a model from a labeled source language is therefore attractive for multimodal emotion recognition (MER). For tonal target languages, however, transfer must contend with more than a change in vocabulary: lexical tone and affective prosody share the fundamental-frequency (F_0) channel. A model can consequently exploit a pitch pattern that is informative about emotion in one corpus but primarily linguistic in another.

Our motivating setting is a tonal language with limited emotion annotation. We investigate its zero-target-emotion-label endpoint using English MELD (Poria et al., 2019) as the source and Mandarin EmotionTalk (Sun et al., 2026) as the target. Mandarin is selected because it provides a tonal speech system and an established annotated multimodal benchmark. Hiding its target emotion labels during adaptation allows quantitative evaluation without collecting a new corpus. This simulates *task-specific annotation scarcity*; it does not imply that Mandarin is generally a low-resource language or reproduce the absence of transcripts and pretrained models in many low-resource settings.

Two challenges shape the method. First, cross-domain distribution matching need not preserve emotion semantics. Adversarial alignment and kernel discrepancy minimization (Ganin et al., 2016; Gretton et al., 2012) can reduce a marginal mismatch while leaving incorrect class correspondences. Relational optimal transport offers a complementary view: Gromov–Wasserstein (GW) compares within-domain pairwise structure (Peyre et al., 2016). Yet structure alone may admit multiple couplings under a symmetry. Fixed affective cues can help distinguish such alternatives. Second, pitch combines lexical, pragmatic, speaker, and affective variation. Mandarin tones (Chao, 1968) and emotional prosody (Scherer, 2003) share pitch cues. Acoustic representations should account for predictable tone variation while retaining other prosody.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

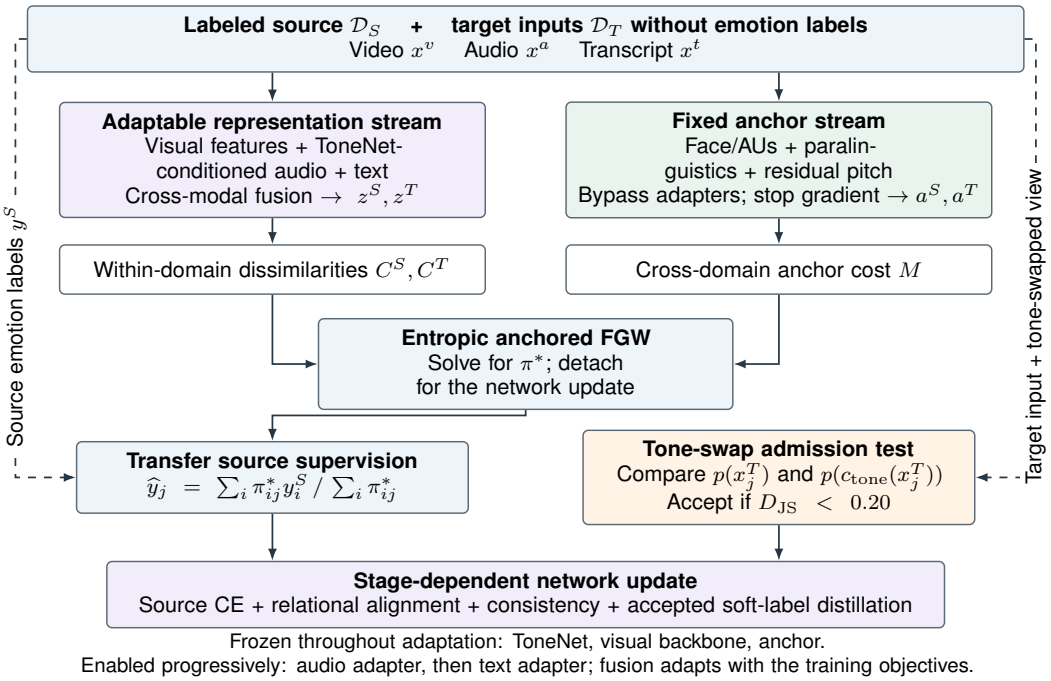

Figure 1: **TD-AFGW during adaptation.** Both domains supply video, audio, and transcripts; only the source supplies emotion labels. Target text additionally drives ToneNet (Figure 2). The two representation streams provide within-domain relations and cross-domain anchor costs. Couplings transfer source labels, while the tone-swap gate determines which target soft labels enter distillation. Dashed connectors denote conditioning or training control, not extra emotion supervision.

We propose TD-AFGW, short for Tone-Disentangled Anchored Fused Gromov–Wasserstein. Here, "disentangled" describes the intended modeling separation, not a guarantee of identifiable latent factors. ToneNet predicts a baseline contour from linguistic features alone. Its hidden states condition the audio branch, and residual pitch contributes to a fixed anchor. Anchored FGW combines this anchor with relational alignment. Source labels are then propagated through the coupling and admitted to target distillation only when predictions are stable under a lexical-pitch intervention. Figure 1 summarizes the framework.

Our contributions are threefold: (i) a tone-aware acoustic pathway combining text-conditioned pitch prediction, residual features, and FiLM; (ii) an anchored relational transfer objective with an explicit characterization of when anchors distinguish symmetry-equivalent couplings; and (iii) coupling-based soft-label distillation with tone-swap filtering and progressive modality adaptation. We evaluate the combined framework and its components on English-to-Mandarin MER under withheld target emotion supervision.

2 Related Work

Multimodal and cross-lingual emotion recognition. Cross-modal attention is a standard means of integrating heterogeneous streams (Tsai et al., 2019). MELD and EmotionTalk support conversational emotion recognition in English and Mandarin, respectively. Cross-lingual speech studies also examine the transferability of self-supervised encoders and the role of linguistic background (Han et al., 2024). Our contribution concerns adaptation of multimodal emotion representations under unavailable target emotion supervision, rather than a new generic fusion architecture.

Relational and paralinguistic adaptation. Optimal transport has been used for domain adaptation (Courty et al., 2017); FGW combines feature matching and relational costs (Vayer et al., 2019). Recent

work on low-resource multilingual speech, NOVA-ARC, uses non-verbal supervision, geometry-aware representations, and transport-based soft supervision (Girish et al., 2026). Thus neither transport-based pseudo-labeling nor paralinguistic transfer alone is new. Our formulation combines a text-conditioned lexical-pitch baseline, a fixed audiovisual/paralinguistic anchor, a GW relation term, and a tone-swap admission test. NOVA-ARC concerns non-verbal-to-verbal speech transfer, whereas our benchmark transfers labeled English multimodal conversations to Mandarin conversations. We do not claim an empirical advantage over methods absent from our comparison.

Tone-aware conditioning. FiLM modulates intermediate features using a conditioning vector (Perez et al., 2018). We derive that vector from text-based tone features, avoiding direct acoustic input to the baseline predictor. This prevents that predictor from reading the observed emotional contour, but does not establish independence from emotion: lexical content, speaking style, and tone realization can remain correlated. Residualization is consequently a modeling bias to be tested, rather than a causal identification result.

3 Problem Setup

Let $\mathcal{D}_S = \{(x_i^S, y_i^S)\}_{i=1}^{N_S}$ denote labeled source utterances and $\mathcal{D}_T = \{x_j^T\}_{j=1}^{N_T}$ target adaptation utterances. An input $x = (x^v, x^a, x^t)$ contains video, audio, and a transcript; $y_i^S \in \{0,1\}^K$ is a one-hot emotion label. The shared label vocabulary has $K = 7$: neutral, surprise, fear, sadness, joy, disgust, and anger. The task is *unsupervised domain adaptation* (UDA), because target inputs are available during adaptation. Source-only transfer is a separate baseline.

The low-resource assumption concerns *target emotion labels*. The method still uses transcripts, linguistic tone processing, speech alignment, pretrained encoders, and neutral Mandarin speech for ToneNet. Neither target emotion labels nor emotion-bearing captions are inputs to the adaptation objectives. A deployment without these auxiliary resources is outside the evaluated setting.

The backbone extracts visual, acoustic, and textual sequences, combines them with cross-modal fusion, and returns a fused vector $z = f_\theta(x)$ and classifier distribution $p_\theta(x)$. The visual pathway uses facial-video features and action units (Cai et al., 2023; Baltrusaitis et al., 2018); the acoustic and text pathways use pretrained speech and multilingual text representations (Chen et al., 2022; Babu et al., 2022; Conneau et al., 2020). Lightweight target adapters are progressively enabled. We distinguish these adaptable representations from the fixed anchor below.

4 Method

4.1 A text-conditioned lexical-pitch baseline

For syllable r, ToneNet receives

$$q_r = (\tau_r, \mathrm{ini}_r, \mathrm{fin}_r, u_r, b_r), \tag{1}$$

where τ_r is surface tone after sandhi processing, ini_r and fin_r are its initial and final, $u_r \in [0,1]$ is relative sentence position, and b_r is a boundary indicator. All features are derived from text. ToneNet receives neither measured pitch nor energy, absolute syllable duration, or speaker embeddings.

A four-layer pre-norm Transformer uses model width 256, four attention heads, feed-forward width 1024, and dropout 0.1. A linear head produces ten semitone values per syllable at relative locations $0, 1/9, \ldots, 1$, giving $\widehat{F} \in \mathbb{R}^{N\times 10}$ for N syllables. This fixed relative-time grid lets the network predict shape without taking observed duration as an input.

Neutral speech from AISHELL-3 (Shi et al., 2021) supplies training targets. Harvest (Morise, 2017) extracts pitch, and Montreal Forced Aligner (McAuliffe et al., 2017) supplies syllable boundaries. With speaker reference pitch b_s estimated from voiced frames, normalized pitch is $F(t) = 12\log_2(F_0(t)/b_s)$. Resampling aligned syllables gives targets on the same ten-point grid. With Δ denoting first differences and $\lambda_\Delta = 0.3$, ToneNet uses the following masked contour and slope objective. The masks exclude unreliable pitch samples. Appendix A gives further preprocessing details.

$$\mathcal{L}_{\mathrm{tone}} = \|F - \widehat{F}\|_{1,\Omega} + \lambda_\Delta \|\Delta F - \Delta\widehat{F}\|_{1,\Omega_\Delta}, \tag{2}$$

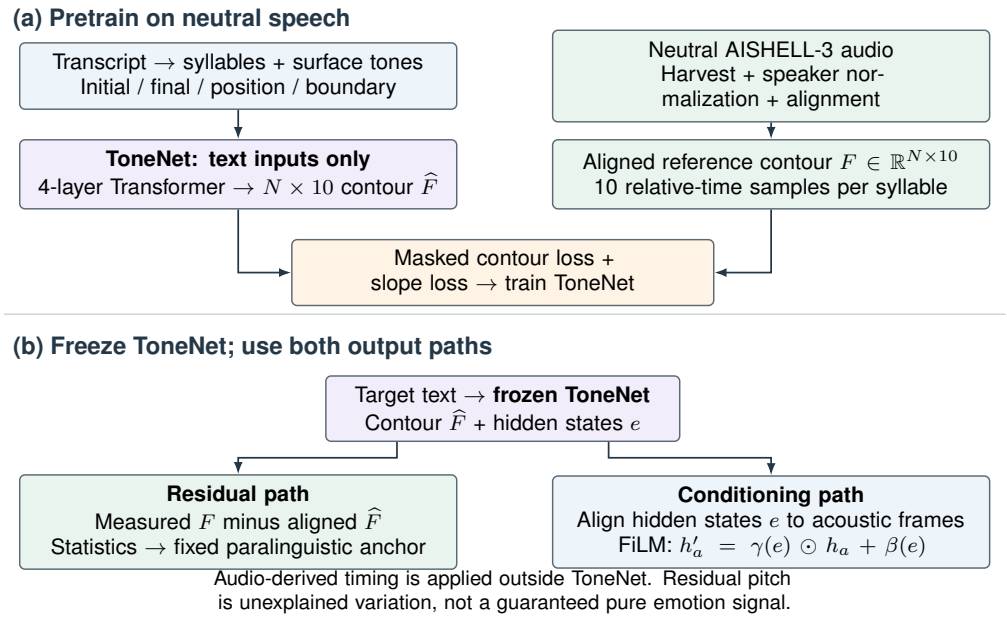

Figure 2: **ToneNet and its two uses.** The predictor takes linguistic features only. Neutral audio supplies regression targets during pretraining, then ToneNet is frozen. At adaptation and inference, external alignment maps its predictions and hidden states to audio time. Residual statistics enter the fixed anchor; hidden states drive FiLM. Measured pitch is never fed into ToneNet.

After pretraining, ToneNet is frozen and serves two paths (Figure 2). First, the aligned residual $R(t) = F(t) - \widehat{F}(t)$ supplies paralinguistic statistics. It contains variation not explained by the predictor, including affect, other linguistic prosody, and prediction error. Second, its hidden states e_r are aligned to acoustic frames and mapped to FiLM parameters:

$$h'_{a,t} = \gamma(e_t) \odot h_{a,t} + \beta(e_t). \tag{3}$$

The conditioning functions let the acoustic pathway account for expected lexical-pitch patterns. Alignment uses audio-derived timing outside ToneNet; "text-conditioned" refers to the predictor's inputs, not to an audio-free end-to-end system. For non-tonal English, the lexical-tone baseline is zero in normalized pitch coordinates. This convention does not assert that English has no linguistic intonation.

4.2 A FIXED AFFECTIVE ANCHOR

The anchor concatenates frozen facial-video and action-unit features with paralinguistic descriptors:

$$a = g(x) = \big[h_{\text{face}}(x^v);\ h_{\text{para}}(x^a, R)\big]. \tag{4}$$

The latter include energy, speaking rate, pauses, voice quality, and residual-pitch summaries. The anchor bypasses language adapters and is detached from the alignment gradient. Its role is to provide comparable cues across domains while the fused representation changes. Fixed weights alone do not make these cues culturally invariant; their cross-domain reliability remains an assumption.

4.3 ANCHORED FUSED GROMOV–WASSERSTEIN ALIGNMENT

For a batch with n source and m target utterances, let $p_i = 1/n$, $q_j = 1/m$, and define the feasible couplings

$$\Pi(p, q) = \{\pi \geq 0 : \pi \mathbf{1}_m = p,\ \pi^\top \mathbf{1}_n = q\}. \tag{5}$$

Using cosine dissimilarity $d(u, v) = 1 - \cos(u, v)$, construct $M_{ij} = d(a_i^S, a_j^T)$, $C_{ik}^S = d(z_i^S, z_k^S)$, and $C_{j\ell}^T = d(z_j^T, z_\ell^T)$. These are dissimilarity matrices; no triangle-inequality claim is needed.

The coupling minimizes

$$\mathcal{J}(\theta,\pi) = (1-\alpha)\langle \pi, M\rangle + \alpha \sum_{ijk\ell} \pi_{ij}\pi_{k\ell}(C^S_{ik} - C^T_{j\ell})^2 - \varepsilon H(\pi), \tag{6}$$

$$H(\pi) = -\sum_{ij} \pi_{ij} \log \pi_{ij}, \qquad \pi^* \in \arg\min_{\pi \in \Pi(p,q)} \mathcal{J}(\theta,\pi).$$

The anchor term provides cross-domain correspondence cues, and the relational term penalizes mismatched pairwise structure. Entropic FGW can be solved using iterative transport routines in POT (Flamary et al., 2021), with Sinkhorn scaling for entropic OT subproblems (Cuturi, 2013). Entropy promotes diffuse couplings; it does not generally make the quadratic FGW problem globally convex.

For the network update, we use $\mathcal{L}_{\mathrm{AFGW}} = \mathcal{J}(\theta, \mathrm{sg}(\pi^*))$, where sg stops the gradient. For an exact optimizer, an envelope argument can justify this partial derivative under suitable regularity conditions. A finite, locally converged FGW solve instead yields an approximate alternating update, not a guarantee of the derivative of a global optimum. Alignment gradients update enabled target adapters and fusion; the visual backbone and anchor stay fixed.

Scope of symmetry breaking. If a target permutation P preserves masses and relations, $Pq = q$ and $PC^T P^\top = C^T$, then π and $\pi P^\top$ have identical relational and entropy costs. For $\alpha < 1$, the anchor distinguishes this pair precisely when

$$\langle \pi P^\top, M\rangle \neq \langle \pi, M\rangle. \tag{7}$$

Appendix D proves this statement. Non-invariance of the anchor matrix alone does not guarantee a nonzero difference for every coupling, nor does distinguishing a pair guarantee the semantically correct match or uniqueness.

4.4 Soft-label propagation and tone-swap admission

The detached coupling transfers source labels to a target utterance:

$$\widehat{y}_j = \frac{\sum_i \pi^*_{ij} y^S_i}{\sum_i \pi^*_{ij}}. \tag{8}$$

Thus supervision comes from source labels and batch-level correspondences rather than directly from a target classifier's argmax. It can still inherit errors from the learned geometry and anchor; propagation does not eliminate confirmation bias.

To test sensitivity to lexical-pitch variation, a tone-swap intervention replaces the predicted baseline contour while retaining residual pitch. In normalized coordinates, the intended edit is $F^{\mathrm{cf}}(t) = R(t) + \widehat{F}_{\mathrm{swap}}(t)$. WORLD resynthesis (Morise et al., 2016) changes the pitch channel while keeping the other synthesis streams fixed. The transcript and video are unchanged. This intervention tests model sensitivity; it is not asserted to be a natural, semantically equivalent spoken sentence. In particular, lexical changes and synthesis artifacts can affect its validity.

Let $p_j = p_\theta(x^T_j)$ and $p^{\mathrm{cf}}_j = p_\theta(c_{\mathrm{tone}}(x^T_j))$. An example is admitted if

$$w_j = \mathbf{1}\{D_{\mathrm{JS}}(p_j, p^{\mathrm{cf}}_j) < \tau\}, \qquad \tau = 0.20. \tag{9}$$

Jensen–Shannon divergence is the mean KL divergence to the mixture distribution (Lin, 1991). The gate requires stability, not correctness; an invariant but incorrect prediction may pass. Soft-label distillation uses

$$\mathcal{L}_{\mathrm{distill}} = \frac{\sum_j w_j \,\mathrm{KL}(\mathrm{sg}(\widehat{y}_j) \,\|\, p_\theta(x^T_j))}{\max(1, \sum_j w_j)}. \tag{10}$$

A rejected example contributes no distillation loss in that round. A complementary representation consistency term is

$$\mathcal{L}_{\mathrm{inv}} = \mathbb{E}_{x,c}\big[1 - \cos(f_\theta(x), f_\theta(c(x)))\big]. \tag{11}$$

The consistency term uses frame dropout and Gaussian noise; tone swapping supplies the admission test.

4.5 TRAINING STAGES AND INFERENCE

Stage 1 learns the source emotion task. Stage 2 adds target alignment and progressively enables the visual-to-audio-to-text ($V \to A \to T$) adaptation schedule. The initial V phase establishes the coupling with language adapters frozen; A enables the tone-conditioned acoustic adapter; T enables the text adapter. V denotes visual-led initialization, not unfreezing the visual backbone. Stage 3 adds coupling-label distillation, recomputing pseudo-labels as representations change. These three training *stages* are distinct from the three modality *phases* inside Stage 2.

The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{CE}}^{S} + \beta_1 \mathcal{L}_{\mathrm{AFGW}} + \beta_2 \mathcal{L}_{\mathrm{inv}} + \beta_3 \mathcal{L}_{\mathrm{distill}}. \tag{12}$$

Terms are enabled according to their stage, and source supervision remains available to limit task drift. At inference, a target utterance passes through preprocessing, ToneNet, the adapted encoders, fusion, and the classifier, followed by the calibration below. It requires neither source samples, a transport solve, nor counterfactual resynthesis. Appendix B gives the training and inference flow.

Source-prior calibration at inference. All learned target predictors are evaluated with source-prior logit correction. Let $\ell_c(x)$ be the logit for class c, and ρ_c^S its empirical frequency in the source training subset. The decision rule is

$$\widehat{c}(x) = \arg\max_c \left[\ell_c(x) - \lambda_{\mathrm{cal}} \log \rho_c^S + \delta_c \right]. \tag{13}$$

Here λ_{cal} controls prior correction and δ_c is a class offset; neither is the tone-swap threshold τ. The balanced-margin variant changes the inference margins and is reported separately. Calibration is a decision-stage component, not evidence that the encoders alone eliminate class imbalance. Shared calibration across learned baselines separates the comparison of adaptation methods from a raw-argmax versus calibrated comparison.

5 EXPERIMENTS

5.1 DATA, PROTOCOL, AND IMPLEMENTATION

The source run uses the aligned multimodal MELD subset: 4,175 training, 621 development, and 1,459 test utterances. Filtering requires usable facial, audio, and text inputs. EmotionTalk provides 15,413 unlabeled adaptation utterances, a separate 1,908-utterance validation split, and 1,929 test utterances. The reported adaptation/test partition is speaker-disjoint, and test inputs are excluded from transport and gradient updates. Target class supports appear in Table 2. Happy and sad map to joy and sadness. Recording conditions and interaction styles differ alongside language, so this experiment measures cross-corpus as well as cross-lingual transfer.

Target emotion labels do not enter training losses. Source development loss guides Stage 1 selection; the stated adaptation criteria are FGW cost/stability in Stage 2 and counterfactual stability/prediction entropy in Stage 3. Target test trajectories were also monitored for analysis. This distinction matters: label-free objectives alone do not establish a target-label-blind selection protocol (Section 6).

Training uses 10, 12, and 8 epochs in Stages 1–3, with batch sizes 32, 64, and 64. Stage 1 uses learning rate 2×10^{-4} and $\beta_2 = 0.1$. Stage 2 uses $\alpha = 0.6$ and $\beta_1 = 0.3$, with V/A/T phases at epochs 1–4/5–8/9–12. Stage 3 uses $\beta_3 = 0.5$ and $\tau = 0.20$. The reported hardware is one NVIDIA GeForce RTX 5060 Ti with 16 GB memory. Appendix B summarizes the schedule and inference path.

We compare source-only transfer, DANN, MMD, anchor-only OT, confidence self-training ($p > 0.8$), Stage 2 AFGW, and the full pipeline. Anchor-only OT learns from coupling-derived soft labels: with a fixed detached anchor and $\alpha = 0$, the direct relational gradient vanishes, but distillation still trains the predictor. The no-relation ablation instead retains the other full-pipeline components. A constant neutral predictor supplies a majority reference.

Accuracy, weighted F1 (W-F1), and macro F1 (M-F1) use a 0–100 scale; changes are percentage points. Learned-model aggregate scores report mean $\pm$ standard deviation over seeds $\{42, 100, 2024\}$. For class supports n_c, W-F1 is $\sum_c (n_c/N) F_{1,c}$ and M-F1 is $K^{-1} \sum_c F_{1,c}$. No statistical significance claim is inferred from standard deviations alone.

Table 1: **Target EmotionTalk results.** Mean ± standard deviation over three seeds on 1,929 utterances. All learned methods use source-prior calibration. Bold marks the highest mean in each metric; the majority predictor is deterministic.

Method	Accuracy	W-F1	M-F1
Always neutral	$\mathbf{41.52}$	24.37	8.38
Source-only (Stage 1)	34.21 ± 0.35	27.59 ± 0.32	15.29 ± 0.28
DANN	34.90 ± 0.44	28.15 ± 0.40	15.80 ± 0.36
MMD	34.60 ± 0.40	27.95 ± 0.37	15.65 ± 0.33
Anchor-only OT ($\alpha = 0$)	35.10 ± 0.38	28.50 ± 0.35	16.10 ± 0.31
Confidence self-training	34.80 ± 0.41	28.25 ± 0.39	15.90 ± 0.34
Anchored FGW (Stage 2)	28.72 ± 0.48	28.09 ± 0.39	16.76 ± 0.35
TD-AFGW (full)	33.90 ± 0.45	$\mathbf{31.42 \pm 0.38}$	18.01 ± 0.35
TD-AFGW (balanced margin)	28.10 ± 0.42	29.38 ± 0.36	$\mathbf{18.98 \pm 0.32}$

Table 2: **Classwise target F1.** Source-only, full TD-AFGW, and its balanced-margin variant use the same test supports. Aggregate means are in Table 1; class scores are shown to two decimals.

Emotion	Support	Source-only	Full	Balanced margin
Neutral	801	52.85	55.10	48.49
Surprise	161	7.32	13.57	15.16
Fear	125	9.01	8.97	13.48
Sadness	123	4.03	4.26	8.54
Joy	246	9.37	11.55	12.42
Disgust	134	12.21	6.14	10.69
Anger	339	12.24	26.51	24.11
Macro F1	1,929	15.29	18.01	18.98
Weighted F1	1,929	27.59	31.42	29.38

5.2 Target performance and class-level behavior

Table 1 reports the calibrated target comparison. Full TD-AFGW reaches 31.42 W-F1 and 18.01 M-F1, exceeding source-only transfer by 3.83 and 2.72 points. Against DANN, the gains are 3.27 and 2.21 points; against anchor-only OT, they are 2.92 and 1.91 points. Stage 2 improves macro F1 to 16.76, but its W-F1 remains below DANN and anchor-only OT. Distillation is therefore important to the full system's advantage.

Full-model accuracy is 0.31 points below source-only and 7.62 below always-neutral prediction. Balanced margins increase macro F1 by 0.97 but reduce W-F1 by 2.04 and accuracy by 5.80. We report this variant as a separate operating point, retaining the full configuration as the primary result. The majority reference illustrates why both class-averaged and frequency-weighted F1 are needed alongside accuracy.

Table 2 localizes the gains. Anger improves by 14.27 points and surprise by 6.25. All seven classes have positive F1 under calibrated inference, including the source-only baseline. This is not evidence that adaptation alone prevents class collapse. Fear changes little (9.01 to 8.97), and disgust decreases from 12.21 to 6.14. The aggregate improvement therefore coexists with important class-specific weaknesses.

5.3 Ablations, diagnostics, and source retention

Every component ablation reduces W-F1 (Table 3). Removing the anchor gives the largest decline, 4.00 points, followed by removing relations, 3.10. Removing ToneNet costs 2.30 points; replacing tone-swap admission with confidence filtering costs 1.70; simultaneous modality adaptation costs 1.30. These conditional effects support the combined design without proving which geometric or acoustic failure caused each error.

Table 3: **Component ablations.** Mean $\pm$ standard deviation over three seeds; ΔW-F1 is relative to the full model. The no-relation ablation retains the rest of the full pipeline.

Configuration	Accuracy	W-F1	ΔW-F1
Full TD-AFGW	33.90 ± 0.45	31.42 ± 0.38	–
No anchor ($\alpha = 1$)	29.80 ± 0.48	27.42 ± 0.42	-4.00
No relation ($\alpha = 0$)	30.70 ± 0.44	28.32 ± 0.39	-3.10
No ToneNet	31.80 ± 0.39	29.12 ± 0.35	-2.30
Confidence gate	32.50 ± 0.41	29.72 ± 0.37	-1.70
No modality curriculum	32.90 ± 0.43	30.12 ± 0.36	-1.30

The threshold comparison reports 68.40% admission at $\tau = 0.20$ and 31.42 W-F1 after distillation, versus 28.78 without filtering. Appendix C lists the sweep and a vector visualization of the ablations. This sensitivity analysis describes observed performance; it is not an independent validation of threshold selection without target labels.

On 6,148 held-out AISHELL-3 syllables, ToneNet's contour prediction has RMSE 4.199 semitones, MAE 2.685 semitones, and Pearson $r = 0.357$ (Appendix C). These results assess baseline-pitch prediction, not complete removal of lexical information from residuals.

On the 1,459-utterance MELD test subset, source accuracy changes from 53.80 to 53.25 after Stage 3, while W-F1 falls from 49.80 to 46.26 and M-F1 from 30.27 to 28.30. The small accuracy change thus accompanies a 3.54-point W-F1 cost. Table 5 reports all stages; source retention should not be reduced to accuracy alone.

6 Discussion and Limitations

Scope of low-resource transfer. The experiment isolates absent target emotion supervision on a resource-supported Mandarin benchmark. It does not demonstrate transfer to an unseen tonal language, performance as a function of labeled-data budget, or robustness without transcripts and neutral speech. Extending the approach requires language-appropriate tone and pronunciation processing, aligned neutral speech, and checks that the residualization preserves useful affective cues. Merely replacing the tone inventory is insufficient when tone interacts with phonation or when alignment quality differs.

Mechanism and robustness. The ToneNet ablation supports a system-level gain, not complete removal of lexical tone. Tone and emotion need not combine additively, and residuals retain model error and other prosody. Likewise, facial and paralinguistic anchors may shift with culture, speaker, and recording conditions. Tone-swapped audio can create transcript mismatch and vocoder artifacts; stability under this perturbation is an imperfect proxy for useful pseudo-labels. The held-out pitch diagnostic measures prediction quality; stronger disentanglement claims require controlled tone-predictability and affect-preservation evaluations.

Selection and reproducibility. Unlabeled adaptation inputs are separated from test inputs, but target test trajectories and threshold comparisons were inspected. The available evidence therefore does not establish a fully target-label-blind selection protocol. Coincidence between a high test score and a stated unsupervised selection criterion is insufficient to prove independence. A frozen checkpoint/threshold/calibration rule evaluated on a fresh holdout is needed to test that stronger claim. Exact calibration coefficients, class offsets, divergence convention, and solver settings are also necessary to reproduce the operating points.

Statistical and computational limits. Evidence is limited to one language pair and three seeds. Macro F1 remains low in absolute terms, and the classwise results expose uneven transfer. The calibration and balanced-margin settings also influence minority-class performance. Balanced transport also transfers the source batch's class mass before filtering, which can be restrictive under label-prior shift (Appendix D). Dense FGW stores quadratic-sized relation matrices; square-loss matrix contractions cost $O(n^2m + nm^2)$ per outer iteration, in addition to transport scaling.

Preprocessing and counterfactual resynthesis add training cost, although transport and filtering are absent at inference.

7 Conclusion

TD-AFGW combines text-conditioned lexical-pitch modeling, anchored relational transport, and consistency-filtered soft-label distillation for multimodal emotion transfer without target emotion labels. Under shared calibrated inference on MELD-to-EmotionTalk transfer, it improves weighted and macro F1 by 3.83 and 2.72 points over source-only transfer. Ablations support the combined design, while classwise results and source retention reveal remaining trade-offs. Mandarin provides a controlled tonal benchmark for studying annotation scarcity; broader low-resource and cross-language generalization remains an open empirical question.

Ethics statement

The study uses existing conversational datasets and introduces no new human-subject data collection. Emotion categories reflect annotation conventions and should not be treated as objective measurements of a person's internal state. Cultural and demographic bias, consent, and privacy remain relevant when adapting across populations. The results do not justify consequential decisions about individuals. Use and redistribution of underlying data remain subject to their licenses and access conditions.

AI use statement

Generative AI assisted with manuscript drafting and revision, title and abstract wording, methodological critique, mathematical exposition, literature and reference checks, and figure code and layout. Quantitative results reported here originate from the authors' experimental records. The authors are responsible for the final text, methods, claims, and artifacts.

Reproducibility statement

Section 4 defines the model, transport objective, and pseudo-label admission rule. Section 5 describes the benchmark and aggregate comparisons. Appendix A records ToneNet architecture and preprocessing, Appendix B separates adaptation from inference, Appendix C provides diagnostic and source-retention results, and Appendix D gives the symmetry argument. Section 6 states the remaining selection and reproducibility limits.

Acknowledgments

We would like to express our sincere gratitude to the Posts and Telecommunications Institute of Technology (PTIT) for its academic support throughout this research. We also extend our appreciation to the management and colleagues at PThub Company for providing a supportive professional environment, valuable resources, and technical guidance that contributed to the completion of this study.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, et al. XLS-R: Self-supervised Cross-lingual Speech Representation Learning at Scale. In *Interspeech*, pp. 2278–2282, 2022.

Tadas Baltrusaitis, Amir Zadeh, Yao Chong Lim, and Louis-Philippe Morency. OpenFace 2.0: Facial Behavior Analysis Toolkit. In *13th IEEE International Conference on Automatic Face & Gesture Recognition*, pp. 59–66, 2018.

Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. MARLIN: Masked Autoencoder for Facial Video Representation Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1493–1504, 2023.

Yuen Ren Chao. *A Grammar of Spoken Chinese*. University of California Press, 1968.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-Scale Self-Supervised Pre-Training for Full Stack Speech Processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6): 1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzman, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 8440–8451, 2020. doi: 10.18653/v1/2020.acl-main.747.

Nicolas Courty, Remi Flamary, Amaury Habrard, and Alain Rakotomamonjy. Optimal Transport for Domain Adaptation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(9): 1853–1865, 2017. doi: 10.1109/TPAMI.2016.2615921.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In *Advances in Neural Information Processing Systems*, volume 26, 2013.

Remi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurelie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, et al. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, Francois Laviolette, Mario Marchand, and Victor Lempitsky. Domain-Adversarial Training of Neural Networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Girish, Mohd Mujtaba Akhtar, and Muskaan Singh. Prosody as Supervision: Bridging the Non-Verbal–Verbal for Multilingual Speech Emotion Recognition. *arXiv preprint arXiv:2604.17647*, 2026. URL https://arxiv.org/abs/2604.17647.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schoelkopf, and Alexander Smola. A Kernel Two-Sample Test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Zhichen Han, Tianqi Geng, Hui Feng, Jiahong Yuan, Korin Richmond, and Yuanchao Li. Cross-lingual Speech Emotion Recognition: Humans vs. Self-Supervised Models. *arXiv preprint arXiv:2409.16920*, 2024. URL https://arxiv.org/abs/2409.16920.

Jianhua Lin. Divergence Measures Based on the Shannon Entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Michael McAuliffe, Michaela Socolof, Sarah Mihuc, Michael Wagner, and Morgan Sonderegger. Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. In *Interspeech*, pp. 498–502, 2017. doi: 10.21437/Interspeech.2017-1386.

Masanori Morise. Harvest: A High-Performance Fundamental Frequency Estimator from Speech Signals. In *Interspeech*, pp. 2321–2325, 2017.

Masanori Morise, Fumiya Yokomori, and Kenji Ozawa. WORLD: A Vocoder-Based High-Quality Speech Synthesis System for Real-Time Applications. *IEICE Transactions on Information and Systems*, E99-D(7):1877–1884, 2016. doi: 10.1587/transinf.2015EDP7457.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual Reasoning with a General Conditioning Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, pp. 3942–3951, 2018. doi: 10.1609/aaai.v32i1.11671.

Gabriel Peyre, Marco Cuturi, and Justin Solomon. Gromov-Wasserstein Averaging of Kernel and Distance Matrices. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2664–2672, 2016.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019. doi: 10.18653/v1/P19-1050.

Klaus R. Scherer. Vocal Communication of Emotion: A Review of Research Paradigms. *Speech Communication*, 40(1–2):227–256, 2003.

Yao Shi, Hui Bu, Xin Xu, Shaoji Zhang, and Ming Li. AISHELL-3: A Multi-Speaker Mandarin TTS Corpus. In *Interspeech*, pp. 2756–2760, 2021. doi: 10.21437/Interspeech.2021-755.

Haoqin Sun, Jinghua Zhao, Xuechen Wang, Shiwan Zhao, Jiaming Zhou, Hui Wang, Xi Yang, Yequan Wang, and Yonghua Lin. EmotionTalk: An Interactive Chinese Multimodal Emotion Dataset With Rich Annotations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9054–9071, 2026. doi: 10.18653/v1/2026.findings-acl.440.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal Transformer for Unaligned Multimodal Language Sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558–6569, 2019. doi: 10.18653/v1/P19-1656.

Titouan Vayer, Laetitia Chapel, Remi Flamary, Romain Tavenard, and Nicolas Courty. Optimal Transport for Structured Data with Application on Graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6275–6284, 2019.

Kun Zhou, Berrak Sisman, Rui Liu, and Haizhou Li. Emotional Voice Conversion: Theory, Databases and ESD. *Speech Communication*, 137:1–18, 2022. doi: 10.1016/j.specom.2021.11.006.

A TONENET ARCHITECTURE AND PREPROCESSING

Table 4: ToneNet architectural specification. Pitch targets and outputs use the same speaker-relative semitone scale.

Component	Specification
Linguistic inputs	Surface tone, initial, final, relative sentence position, boundary indicator
Input representation	Feature embeddings projected to width 256, with positional encoding
Transformer	4 pre-norm encoder layers; 4 heads; feed-forward width 1024; dropout 0.1
Regression head	Linear $256 \rightarrow 10$ per syllable; no output activation
Output	$N \times 10$ real-valued contour for N syllables
Supervision	Neutral AISHELL-3 speech; masked contour and first-difference losses
Downstream state	Frozen predictor; hidden-state FiLM and explicit residual statistics

Tone labels are surface forms derived from the transcript after pronunciation and sandhi processing. Initials and finals are encoded separately rather than as a whole-syllable identifier. Mandarin processing includes third-tone sequences and context-dependent pronunciations of *yi* and *bu*. Relative position and boundaries allow the predictor to model contextual contour changes; they do not encode measured duration.

Harvest extracts pitch at 5-ms spacing. The specified preprocessing addresses octave jumps, median-filters isolated spikes, and interpolates short unvoiced gaps below 50 ms; longer gaps remain masked. Pitch is represented relative to a speaker median, rather than normalized to zero mean and unit variance within each utterance. This retains utterance-level pitch offsets. Syllables with fewer than 50% voiced frames are excluded from the training loss. Each aligned syllable is resampled to ten relative-time points; slope differences use adjacent valid pairs. At downstream use, both measured and predicted pitch must be expressed on the same scale before subtraction.

The reference speaker median and alignment are external preprocessing quantities, not inputs to ToneNet. For a held-out evaluation protocol, preprocessing statistics fitted on adaptation data must be separated from evaluation-derived statistics. A new speaker without a usable reference requires an explicit normalization policy; the current model does not by itself resolve that deployment case.

The model produces a baseline prediction and hidden states in a single text-conditioned forward pass. Alignment maps both to acoustic time. The explicit residual is summarized by statistics such as mean, standard deviation, slope, and range. The conditioning path uses hidden states to generate channelwise FiLM scales and shifts. Disabling FiLM corresponds to $\gamma = 1$, $\beta = 0$; removing ToneNet additionally removes its contribution to residual-based anchor features.

ESD (Zhou et al., 2022), with parallel emotional utterances, offers an appropriate setting for emotion-preservation probes. A useful evaluation would jointly test pitch prediction, tone predictability from residuals, and affect predictability before and after residualization. Appendix C reports held-out pitch prediction. Predictive accuracy, nuisance removal, and affect preservation are separate properties.

B TRAINING, GRADIENT PATHS, AND INFERENCE

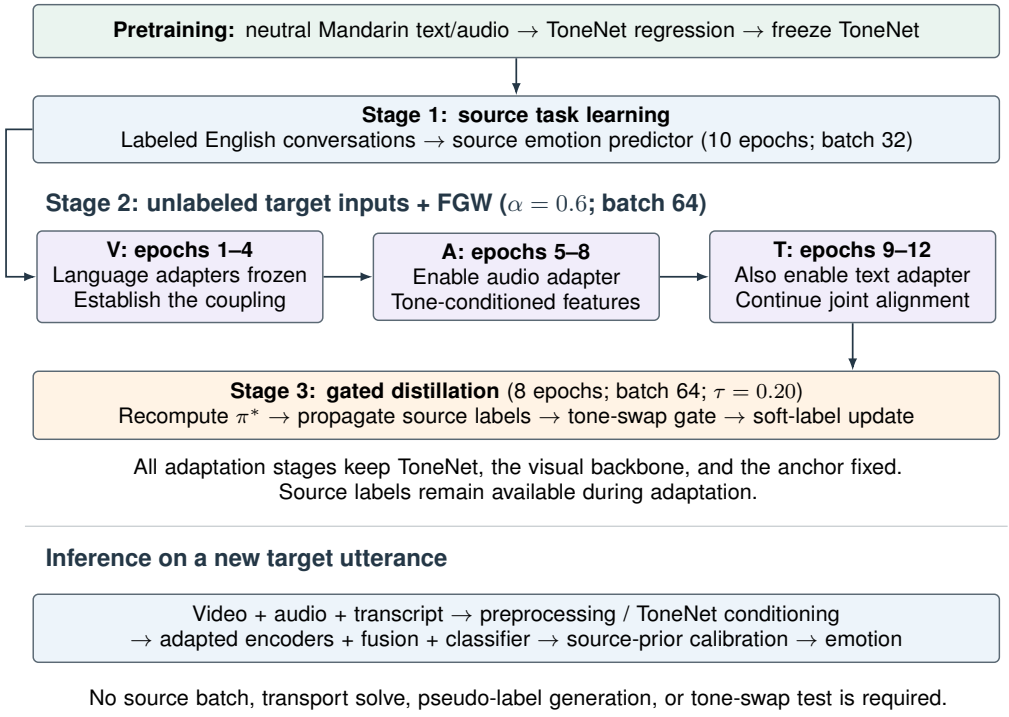

Figure 3: **Training stages and target inference.** ToneNet pretraining precedes the MER stages. The V/A/T phases occur within Stage 2. The final predictor has no dependency on source examples or a transport plan at inference.

One adaptation update. Sample source and target batches, compute representations and detached anchors, construct M, C^S, C^T, and solve FGW. In Stage 3, propagate source labels and apply the tone-swap gate. Evaluate Eq. equation 12, treating the coupling, soft labels, and admission decision as fixed for this update. Update enabled parameter groups, then recompute transport and admission after representation changes.

Gradient paths. Source cross-entropy trains the emotion predictor. The fixed anchor influences coupling selection but has no direct encoder gradient after the coupling is detached. Relational alignment acts through fused representations; distillation acts through the classifier and enabled target pathways. ToneNet stays frozen. At $\alpha = 0$, the direct relational gradient disappears.

Coupling indices. The coupling is $n \times m$ over examples, not intrinsically a 7×7 class matrix. Class aggregation requires a separate evaluation protocol. Coupling averages or drift estimates require aligned row/column example identities; elementwise averaging across unrelated minibatches is not meaningful.

Divergence convention. For $r = (u + v)/2$, $D_{\mathrm{JS}}(u, v) = \frac{1}{2}\mathrm{KL}(u\|r) + \frac{1}{2}\mathrm{KL}(v\|r)$. A fixed log convention is necessary because it changes the scale of the threshold 0.20. Admission certifies low prediction divergence for the intervention, not pseudo-label correctness or preservation of affect.

C Additional Experimental Results

C.1 Source retention

Table 5: **MELD source test results.** Evaluation uses the aligned 1,459-utterance subset; values are mean ± standard deviation over three seeds.

Checkpoint	Accuracy	W-F1	M-F1
Source supervised (Stage 1)	53.80 ± 0.32	49.80 ± 0.35	30.27 ± 0.28
After Stage 2	51.20 ± 0.41	48.88 ± 0.38	29.15 ± 0.33
After Stage 3	53.25 ± 0.30	46.26 ± 0.32	28.30 ± 0.29

Stage 3 recovers 2.05 accuracy points relative to Stage 2, while source W-F1 decreases by a further 2.62 points. Relative to Stage 1, final accuracy, W-F1, and M-F1 change by -0.55, -3.54, and -1.97 points. These different trends preclude a blanket claim that adaptation eliminates forgetting.

C.2 Tone-swap threshold sensitivity

Table 6: **Gate sensitivity.** Reported admission rates and downstream calibrated target W-F1 for the threshold comparison. The unfiltered run is distinct from the confidence-gate ablation.

Admission rule	Accepted (%)	Target W-F1
Unfiltered	100.00	28.78
$\tau = 0.10$	42.10	28.92
$\tau = 0.15$	55.70	29.64
$\tau = 0.20$ (main)	68.40	31.42
$\tau = 0.25$	79.20	29.85
$\tau = 0.30$	88.60	29.12

Tighter admission discards more target examples; looser admission admits more examples without ensuring label correctness. The peak in this comparison is at 0.20. Because the comparison exposes target test performance, it cannot itself justify a claim of target-label-independent threshold tuning. No statistical significance or seed variability is asserted for these sweep values.

C.3 Held-out lexical-pitch prediction

Table 7: **ToneNet contour prediction on held-out AISHELL-3.** Errors are in speaker-relative semitones (st). Counts refer to syllables, not EmotionTalk utterances. Overall Pearson correlation is $r = 0.357$.

Surface tone	Syllables	RMSE (st)	MAE (st)
Tone 1	1,280	3.488	2.196
Tone 2	1,440	4.509	2.871
Tone 3	1,068	5.101	3.324
Tone 4	1,920	3.840	2.502
Tone 5 (neutral tone)	440	4.417	2.766
Overall	6,148	4.199	2.685

The reported overall errors are given separately from tone-specific errors; syllable counts alone do not specify the valid-frame weighting. Tone 3 has the highest reported prediction error. Residual features therefore retain baseline-model error, and correlation with measured pitch is not a measure of emotion preservation.

C.4 VISUAL SUMMARY OF COMPONENT ABLATIONS

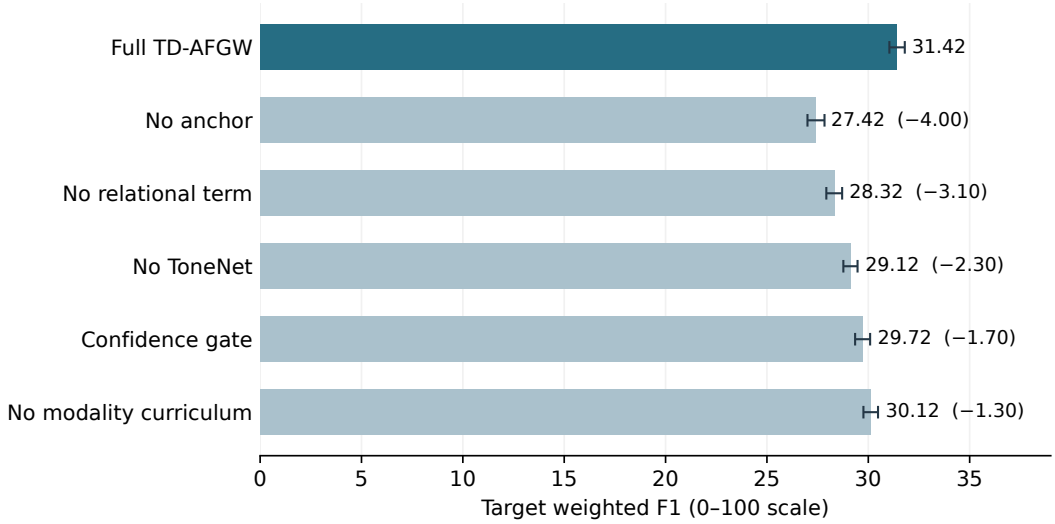

Figure 4: **Ablation results under calibrated inference.** Bars are mean W-F1 and error bars are standard deviations across three seeds, matching Table 3. Labels report each reduction relative to the full model. The axis begins at zero; the figure does not imply statistical significance.

The largest observed ablation cost is associated with removing the anchor. Removing the relational term, ToneNet, the counterfactual gate, or the curriculum also lowers the mean. Ablation differences are conditional on the remaining system and should not be summed to estimate a total causal contribution.

D SYMMETRY AND LABEL-MASS PROPERTIES

Proposition 1 (pairwise symmetry distinction). Let P be a permutation over target indices satisfying $Pq = q$ and $PC^T P^\top = C^T$. Define

$$G(\pi) = \sum_{ijk\ell} \pi_{ij}\pi_{k\ell}(C^S_{ik} - C^T_{j\ell})^2. \tag{14}$$

For every $\pi \in \Pi(p, q)$, $\pi P^\top$ is feasible, $G(\pi P^\top) = G(\pi)$, and $H(\pi P^\top) = H(\pi)$. Consequently,

$$\mathcal{J}(\theta, \pi P^\top) - \mathcal{J}(\theta, \pi) = (1 - \alpha)\langle \pi P^\top - \pi, M\rangle. \tag{15}$$

For $\alpha < 1$, the two objective values differ if and only if the inner product on the right is nonzero.

Proof. The row sums satisfy $(\pi P^\top)\mathbf{1} = \pi\mathbf{1} = p$. Column sums satisfy $(\pi P^\top)^\top\mathbf{1} = P\pi^\top\mathbf{1} = Pq = q$. Nonnegativity is preserved. Reindexing the two target coordinates of the quadratic form leaves it unchanged by $PC^T P^\top = C^T$. Entropy is invariant to permutations of entries. Only the anchor term can therefore change, giving the stated difference. □

Limits of the proposition. The condition depends on both M and π. Even if the anchor costs change under P, a uniform or otherwise symmetric coupling can satisfy $\pi P^\top = \pi$. Moreover, a distinguishing anchor may favor the wrong correspondence. The proposition establishes neither a unique optimum nor semantic correctness, statistical generalization, or a target-risk bound. Such claims would need additional assumptions and evidence.

Proposition 2 (label mass before filtering). For an exactly feasible coupling and the propagated labels in Eq. equation 8,

$$\sum_j q_j \widehat{y}_j = \sum_i p_i y^S_i. \tag{16}$$

Proof. Since $\sum_i \pi_{ij} = q_j$, multiplying Eq. equation 8 by q_j, summing over j, and using $\sum_j \pi_{ij} = p_i$ yields Eq. equation 16. □

Thus balanced coupling preserves the source batch's emotion mass in the target soft-label average before gating. It does not discover the target class prior. Admission filtering can change this average; batch sampling also affects the prior being transferred. This is a limitation when source and target emotion frequencies differ, even if the matching geometry is informative.