BEYOND FUSION CONFIDENCE: ADAPTIVE THREE-WAY CALIBRATION FOR SEMI-SUPERVISED MULTIMODAL LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Semi-supervised multimodal learning commonly uses fusion confidence to select or weight pseudo-labels. However, predictions with similar confidence can have different correctness depending on whether modality-specific heads support the predicted class. In this paper, we propose Adaptive Three-way Calibration (ATC), a teacher-student framework that incorporates modality-head support into training-time reliability assessment. ATC partitions unlabeled examples into Accept, Boundary, and Reject regions according to full, partial, and absent support for the teacher's fusion prediction. These regions integrate pseudo-label admission, weighting, and target construction to preserve partial agreement as a distinct supervision condition. By replaying a few labeled training examples, ATC further estimates regional reliability without additional annotations and stabilizes sparse estimates through shrinkage and guarded monotonic projection. The resulting reliability targets guide pseudo-label target concentration and confidence alignment while preserving the teacher's prediction. Experimental results on five benchmark datasets show that ATC remarkably outperforms both SSL baselines and multimodal or multiview baselines. Among the evaluated methods, ATC ranks first or second in 19 of 20 primary dataset-metric comparisons. Confidence-bin evaluation reveals both regional correctness stratification and distinct sample-count trends. Ablation studies show that regional supervision accounts for most of the classification gain, while retaining boundaries yield lower mean ECE than either binary partition. The results support using modality-head evidence to condition pseudo-label reliability rather than relying on fusion confidence alone.

1 INTRODUCTION

Multimodal and multi-view learning combine complementary information from images, text, video, audio, and metadata to improve prediction (Ngiam et al., 2011; Baltrusaitis et al., 2019; Liang et al., 2024). Obtaining labels aligned across views is costly and application-dependent, making semi-supervised learning (SSL) with unlabeled multimodal data attractive. Modern SSL builds on pseudo-labeling and consistency regularization (Laine & Aila, 2017; Tarvainen & Valpola, 2017; Sohn et al., 2020; Xie et al., 2020), commonly selecting or weighting unlabeled examples by confidence. Pseudo-label errors can alter optimization and reinforce confirmation bias (Arazo et al., 2020), making reliable confidence estimation central to multimodal SSL.

Model calibration asks whether predictive confidence reflects the empirical probability of being correct (Naeini et al., 2015; Guo et al., 2017; Kull et al., 2019; Minderer et al., 2021; Bae et al., 2025). A well-calibrated classifier is correct approximately c of the time for predictions assigned confidence c. In applications, calibration supports uncertainty-aware decisions. It can also affect pseudo-label admission, weighting, and target construction for SSL. Miscalibration can therefore alter the training trajectory, not merely reported uncertainty. In multimodal SSL, one fusion confidence can additionally conceal disagreement among modality heads.

Adaptive SSL adapts thresholds and weights (Zhang et al., 2021; Xu et al., 2021; Wang et al., 2023; Chen et al., 2023), but still operates mainly on scalar confidence; multimodal SSL uses agreement primarily for consistency or fusion (Du et al., 2024; 2025); and conventional calibration largely

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.

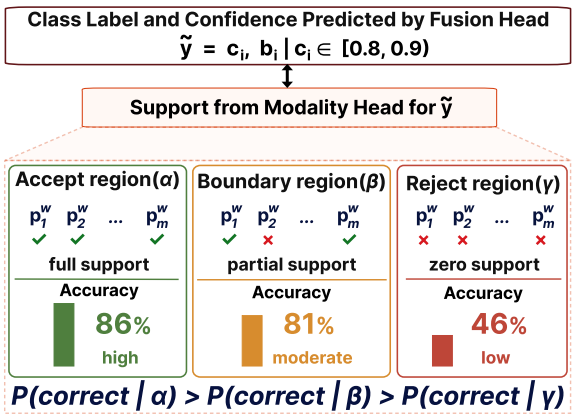

Figure 1: Fusion confidence alone does not identify pseudo-label reliability. Within the same confidence bin, modality heads provide full, partial, or no support for the fusion-proposed class. A case example on dataset CUB shows rounded correctness rates for these three regions in $[0.8, 0.9)$.

targets final predictive confidence (Guo et al., 2017; Kull et al., 2019; Minderer et al., 2021; Ma et al., 2023). Multimodal SSL additionally requires distinguishing evidence patterns hidden by the same confidence, preserving partial agreement, and estimating reliability under sparse support patterns. These observations motivate modality-head support as an additional reliability axis.

Figure 1 illustrates this ambiguity by holding the fusion prediction fixed while varying modality-head support. On DVM (Huang et al., 2022), within $\bar{c}_i \in [0.9, 1.0)$, full support yields 99.49% correctness versus 67.74% without modality-head support, a gap of 31.75 percentage points. On CUB (Wah et al., 2011), within $\bar{c}_i \in [0.8, 0.9)$, fully supported, partially supported, and unsupported samples achieve 86.15%, 81.23%, and 46.43% correctness, respectively. The Accept–Reject gap is 39.72 percentage points. These gaps show that support remains informative within fusion-confidence bins, motivating $P(z_i = 1 \mid \bar{c}_i, g_i)$ rather than $P(z_i = 1 \mid \bar{c}_i)$, where z_i indicates pseudo-label correctness and g_i denotes the support region.

The resulting design question is how to use support evidence without treating partial agreement as either full agreement or complete disagreement. We propose *Adaptive Three-Way Calibration (ATC)* in a teacher–student framework. The teacher fusion head proposes a class, while modality heads identify full, partial, or absent support for that class. These patterns define Accept, Boundary, and Reject as three levels of class-support evidence. Regional controls determine pseudo-label admission, weighting, and target construction. Replay on labeled training examples supplies an empirical regional reliability target for classification target shaping, confidence alignment, or both. ATC thus separates the source of the class proposal from the evidence used to regulate its supervision. Inference retains only the student fusion head.

We evaluate ATC on five heterogeneous benchmarks spanning image–tabular, image–attribute, image–text, image–metadata, and trimodal sequence classification. Under the shared protocol, ATC ranks first or second among the evaluated methods in 19 of 20 primary dataset–metric comparisons. Our diagnostics examine two complementary properties: correctness differences within confidence bins and regional sample-count trends across bins. Controlled CUB ablations distinguish regional supervision from partition granularity and replay refinement. The three-way partition achieves lower mean ECE than either binary alternative while preserving their classification gains.

Contributions. In summary, we make the following contributions:

- We examine a confidence–support mismatch: pseudo-labels in the same fusion-confidence bin can differ markedly in correctness across support regions.
- We propose an adaptive three-way calibration that retains partial support and uses regional reliability to regulate pseudo-label supervision and confidence.
- Experimental results on five benchmarks together with confidence-bin evaluation and ablation studies show the proposed method outperforms baselines markedly.

2 Related Work

2.1 Semi-supervised Multimodal Learning

Modern SSL combines pseudo-labeling and consistency regularization (Laine & Aila, 2017; Tarvainen & Valpola, 2017; Sohn et al., 2020; Xie et al., 2020) with distribution alignment, semantic matching, and adaptive selection or weighting (Berthelot et al., 2019; 2020; 2022; Zheng et al., 2022; Li et al., 2021; Zhang et al., 2021; Xu et al., 2021; Wang et al., 2023; Chen et al., 2023). Although these methods improve how confidence is used, confidently incorrect pseudo-labels remain problematic (Arazo et al., 2020), while learned estimators such as SemiReward introduce additional reliability signals (Li et al., 2024).

Multimodal learning combines modality-specific encoders and fusion mechanisms (Ngiam et al., 2011; Baltrusaitis et al., 2019; Liang et al., 2024). For example, image–tabular methods such as MMCL, TIP, and STiL demonstrate uneven but useful complementarity (Hager et al., 2023; Du et al., 2024; 2025); and recent multimodal SSL studies cross-view consistency and masked reconstruction (Chen et al., 2025; Cheng et al., 2023). Many existing methods use agreement for representation learning, consistency, or fusion. ATC uses modality-head support for the fusion-proposed class to condition pseudo-label training.

2.2 Model Calibration

Model calibration asks whether confidence matches empirical correctness when probabilities guide decisions or risk assessment. Classical methods include parametric and nonparametric probability mappings (Zadrozny & Elkan, 2001; Niculescu-Mizil & Caruana, 2005). Bayesian binning provides a probabilistic treatment (Naeini et al., 2015), while temperature scaling became standard for modern neural networks (Guo et al., 2017). Multiclass calibration and large-scale evaluations further extend this line (Kull et al., 2019; Minderer et al., 2021).

Calibration recently considers distribution shift, selective prediction, and failure detection (Ovadia et al., 2019; Chow, 1970; Geifman & El-Yaniv, 2017; 2019; Hendrycks & Gimpel, 2017; Corbiere et al., 2019; Jiang et al., 2018). Reliability-aware segmentation similarly separates reliable and unreliable pseudo-labels during training (Wang et al., 2022). These directions show that a single global confidence mapping can be insufficient when reliability varies across samples or prediction conditions, yet most methods still operate on a single output or final fused prediction.

Recent SSL methods such as CaliMatch explicitly calibrate confidence to improve pseudo-label selection (Bae et al., 2025), while ATC additionally conditions training-time reliability on modality-head support. Energy-based modeling also targets this interaction (Toba et al., 2024), whereas prior multimodal calibration has addressed the fused output (Ma et al., 2023).

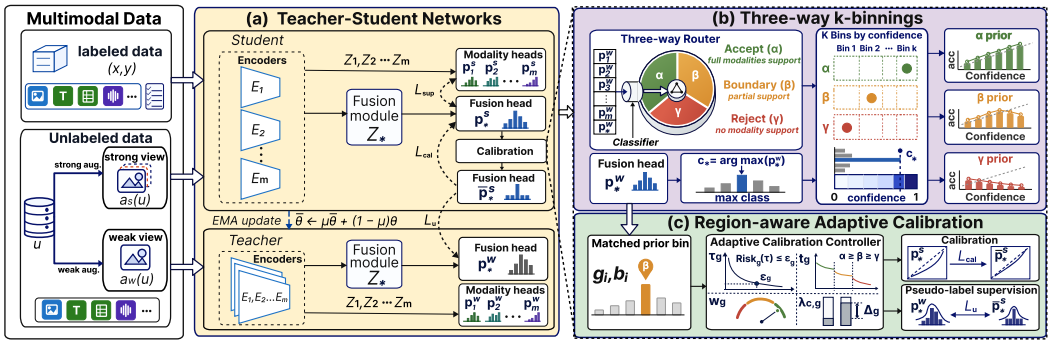

Figure 2: Schematic overview of ATC. (a) The EMA teacher produces weak-view fusion and modality predictions while the student learns from the strong view. (b) Modality-head support for the fusion pseudo-label routes samples into Accept, Boundary, or Reject, while region-conditioned confidence bins summarize replay reliability. (c) Regional controls regulate pseudo-label supervision and confidence alignment. In the evaluated configuration, replay estimates pooled regional targets after warm-up; thresholds, weights, and blending coefficients are fixed. Bins summarize local reliability.

3 METHODOLOGY

Overview. ATC separates three decisions: which class the teacher proposes, which modality heads support that class, and how strongly the student should learn from it. The EMA teacher supplies fusion and modality predictions on a weak unlabeled view; their support pattern determines the region. The student learns from the strong view using regional controls and replay-estimated reliability. Figure 2 summarizes this prediction, routing, and supervision path.

With $z_i = \mathbb{I}[\tilde{y}_i = y_i]$, the motivating reliability question is $P(z_i = 1 \mid \bar{c}_i, g_i)$ rather than $P(z_i = 1 \mid \bar{c}_i)$ alone. This conditional perspective motivates combining fusion-confidence admission with region-dependent supervision. Confidence-bin summaries characterize local variation, while regional pooling supplies a data-efficient reliability target for training. Inference uses only the fusion head.

3.1 MULTIMODAL SEMI-SUPERVISED LEARNING

We consider labeled $\mathcal{D}_l = \{(x_i, y_i)\}$ and unlabeled $\mathcal{D}_u = \{u_i\}$ sets with m modalities and C classes. Projection encoders produce $\mathbf{z}_i^j = E_j(x_i^j)$, and the fusion module forms $\mathbf{z}_i^* = F([\mathbf{z}_i^1; \ldots; \mathbf{z}_i^m])$. Modality heads $h_1, \ldots, h_m$ and fusion head h_* output class distributions. We use superscript s for student strong-view outputs and a bar for EMA-teacher weak-view outputs. On labeled data,

$$\mathcal{L}_{\text{sup}} = \text{CE}(\mathbf{p}_i^{*,s}, y_i) + \lambda_{\text{aux}} \sum_{j=1}^{m} \text{CE}(\mathbf{p}_i^{j,s}, y_i). \tag{1}$$

The auxiliary losses keep modality heads discriminative so that their predictions provide meaningful support evidence during unlabeled routing.

3.2 FUSION PSEUDO-LABEL AND MODALITY-HEAD SUPPORT

For an unlabeled sample, the EMA teacher outputs fusion distribution $\bar{\mathbf{p}}_i^{*,w}$ and modality distributions $\{\bar{\mathbf{p}}_i^{j,w}\}_{j=1}^m$. The fusion output defines

$$\tilde{y}_i = \arg\max_c \bar{p}_{i,c}^{*,w}, \qquad \bar{c}_i = \max_c \bar{p}_{i,c}^{*,w}. \tag{2}$$

The support indicator and set are

$$s_i^j = \mathbb{I}\Big[\arg\max_c \bar{p}_{i,c}^{j,w} = \tilde{y}_i\Big], \qquad S_i = \{j : \arg\max_c \bar{p}_{i,c}^{j,w} = \tilde{y}_i\}. \tag{3}$$

Consequently, $|S_i| = \sum_{j=1}^m s_i^j$ and $|S_i|/m$ is the support ratio. Support is defined for the fusion class, and high confidence may still co-occur with $|S_i| = 0$.

3.3 THREE-WAY REGION ASSIGNMENT

The number of supporting modality heads assigns the sample to one of three regions:

$$g_i = \begin{cases} \alpha, & |S_i| = m \quad \text{(Accept)}, \\ \beta, & 0 < |S_i| < m \quad \text{(Boundary)}, \\ \gamma, & |S_i| = 0 \quad \text{(Reject)}. \end{cases} \tag{4}$$

The regions describe support for the teacher's class proposal rather than replace it. Accept supplies fully supported anchors; Boundary retains useful agreement without treating conflicting evidence as either full support or no support. Reject is handled conservatively, but is not automatically discarded: sufficiently confident samples can still contribute under the regional controls. The following objectives translate these distinctions into admission, supervision strength, and target construction.

3.4 REGION-AWARE PSEUDO-LABEL TRAINING

Each regional control has a distinct role. The threshold τ_g determines pseudo-label admission, w_g weights its loss contribution, and b_g controls support-aware target blending. The coefficient λ_{cal} later scales scalar confidence alignment, while t_g provides its regional reliability target. Admission is

$$\xi_i = \mathbb{I}[\bar{c}_i \geq \tau_{g_i}]. \tag{5}$$

ATC averages the teacher fusion distribution and supporting modality heads:

$$\mathbf{r}_i = \frac{\bar{\mathbf{p}}_i^{*,w} + \sum_{j \in S_i} \bar{\mathbf{p}}_i^{j,w}}{1 + |S_i|}, \qquad \mathbf{q}_i = b_{g_i}\mathbf{r}_i + (1 - b_{g_i})\bar{\mathbf{p}}_i^{*,w}. \tag{6}$$

Accept averages all heads, Boundary uses only the supporting subset, and Reject reduces to the fusion distribution. Supporting heads reshape the target but never replace the fusion pseudo-label, while disagreeing heads are excluded. For Accept and Boundary, b_{g_i} controls this reshaping.

The unlabeled classification objective is

$$\mathcal{L}_u = \frac{1}{Z_u} \sum_i \xi_i w_{g_i} \, \mathrm{CE}(\mathbf{p}_i^{*,s}, \mathbf{q}_i), \qquad Z_u = \max\left(1, \sum_i \xi_i w_{g_i}\right). \tag{7}$$

Z_u normalizes the varying number and weight of admitted samples in each batch.

3.5 Regional Reliability Target

Replay uses the same labeled training pool: the teacher reproduces the pseudo-label and support region, while the label only evaluates pseudo-label correctness. Let $z_i = \mathbb{I}[\tilde{y}_i = y_i]$; replay computes

$$\begin{aligned} n_g &= \sum_i \mathbb{I}[g_i = g], \qquad a_g = \frac{\sum_i \mathbb{I}[g_i = g] z_i}{\max(1, n_g)}, \\ \bar{c}_g &= \frac{\sum_i \mathbb{I}[g_i = g] \bar{c}_i}{\max(1, n_g)}, \quad \Delta_g = |\bar{c}_g - a_g|. \end{aligned} \tag{8}$$

Here n_g, a_g, $\bar{c}_g$, and Δ_g measure regional support, correctness, confidence, and mismatch. Replay estimates t_g without using test labels or adding labeled examples.

ATC shrinks empirical accuracy toward a regional prior t_g^0 using shrinkage strength $\kappa > 0$:

$$\rho_g = \frac{n_g}{n_g + \kappa}, \qquad \hat{t}_g = \rho_g a_g + (1 - \rho_g) t_g^0. \tag{9}$$

Data-rich regions rely more on empirical accuracy, whereas sparse regions remain closer to t_g^0.

The estimates are projected as

$$(t_\alpha, t_\beta, t_\gamma) = \Pi_{\mathcal{C}}(\hat{t}_\alpha, \hat{t}_\beta, \hat{t}_\gamma), \quad \mathcal{C} = \{t_\alpha \geq t_\beta \geq t_\gamma\}, \tag{10}$$

with region-specific bounds. If $n_g < n_{\min,g}$ or the projected estimate falls outside its bound, ATC uses t_g^0. The projection encodes the expected Accept $\geq$ Boundary $\geq$ Reject ordering as a regional modeling constraint. Shrinkage and fallback limit the influence of sparse empirical estimates. For the reported runs, replay estimates targets once after warm-up and holds them fixed thereafter. Adaptation is therefore data-dependent and region-specific, while the EMA teacher continues to evolve.

For finer-grained diagnostics, replay also groups samples into K equal-width confidence bins within each region (Appendix B.2). These summaries characterize local reliability; the training target t_g pools regional evidence to stabilize supervision under sparse support.

Regional target placement. ATC exposes two complementary placements for the same replay-estimated regional reliability target. The classification slot uses t_{g_i} to control the concentration of the pseudo-label target, whereas the calibration slot aligns the student confidence with regional empirical correctness. This design separates reliability estimation from the target's role during optimization.

The classification slot uses

$$q_{i,c}^{(t)} = \begin{cases} t_{g_i}, & c = \tilde{y}_i, \\ \dfrac{1 - t_{g_i}}{C - 1}, & c \neq \tilde{y}_i, \end{cases} \tag{11}$$

where $\tilde{y}_i$ remains the teacher-proposed reference class and t_{g_i} sets its target mass. Retaining it as the unique maximum-probability class of $\mathbf{q}_i^{(t)}$ requires $t_{g_i} > 1/C$; at $t_{g_i} = 1/C$ the target is uniform. The reference class and the target's argmax are thus distinct concepts.

In the joint-target configuration, $\mathbf{q}_i^{(t)}$ replaces $\mathbf{q}_i$ for classification and t_{g_i} also supervises $\mathcal{L}_{\text{cal}}$. In the calibration-only configuration, $\mathbf{q}_i$ is retained and t_{g_i} acts only through confidence alignment. Both placements share the class proposal, architecture, regions, controls, and replay statistics.

The student's strong-view fusion confidence is $\hat{c}_i^s = \max_c p_{i,c}^{*,s}$. We fix the calibration scope as either $\xi_i^{\text{cal}} = 1$ for all unlabeled samples or $\xi_i^{\text{cal}} = \xi_i$ for admitted samples only. The calibration loss is

$$\mathcal{L}_{\text{cal}} = \frac{1}{Z_c} \sum_i \xi_i^{\text{cal}} w_{g_i} \left(\hat{c}_i^s - t_{g_i}\right)^2, \qquad Z_c = \max\left(1, \sum_i \xi_i^{\text{cal}} w_{g_i}\right). \tag{12}$$

The loss aligns maximum fusion confidence with regional replay reliability, allowing samples with similar confidence to receive different supervision according to their support evidence. ECE and Brier evaluate the resulting calibration and full-distribution quality on test data.

3.6 Objective Function and Optimization

The complete objective is

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \lambda_u \mathcal{L}_u + \lambda_{\text{cal}} \mathcal{L}_{\text{cal}}. \tag{13}$$

The three terms learn labeled class semantics, transfer region-conditioned pseudo-label supervision, and align confidence with replay reliability. Teacher predictions, support assignments, and controller targets are stop-gradient quantities. After each student update, the teacher follows $\bar{\theta} \leftarrow \mu\bar{\theta} + (1-\mu)\theta$, where μ is the EMA decay.

4 Experiments

4.1 Experimental Setup

Datasets and features. We evaluate DVM (Huang et al., 2022), CUB (Wah et al., 2011), Food101-MM derived from Food-101 (Bossard et al., 2014), PAD-UFES20 (Pacheco & Krohling, 2020), and MOSI (Zadeh et al., 2016) at labeled ratios of 10%, 20%, 10%, 20%, and 10%. Frozen inputs pair DVM image features with tabular attributes; CUB EfficientNet-B3 bounding-box features with semantic attributes; Food101-MM and PAD-UFES20 ConvNeXt-Tiny features with TF–IDF/SVD text or clinical metadata; and MOSI aligned text–audio–vision sequences.

Baselines. We adapt seven general SSL methods—FixMatch, FlexMatch, SoftMatch, FreeMatch, MixMatch, CoMatch, and SimMatch—to the shared multimodal architecture and include CoTrain and STiL as multi-view/multimodal baselines. The adapted general SSL methods, CoTrain, and ATC derive their unlabeled prediction and confidence information from the same EMA-teacher fusion head and then apply their respective unlabeled-learning rules under the shared protocol. STiL follows its multimodal implementation adapted to this protocol. ATC uses modality-head agreement to assign support regions and, in the calibration-only configuration, blends supporting-head distributions into the classification target.

Implementation Details. All input features remain frozen, and only projection, fusion, and prediction heads are optimized. Weak views use Gaussian feature perturbation, while strong views additionally apply feature dropout. All methods use the same feature protocol, augmentations, training epochs, and random seeds. Appendix A gives supplementary optimization settings; Appendix B describes the reported training schedule.

Label budget and model selection. The reported labeled ratio denotes the training-label budget available to every method. ATC estimates t_g by replaying samples from this same pool, introducing no additional labels; baselines use the same labels for supervised learning.

All methods share train/validation/test splits, labeled ratios, feature and augmentation protocols, training epochs, seeds, validation access, and checkpoint criteria. The same validation split selects checkpoints and, for ATC, one of two predefined target placements under a fixed rule shared across datasets. All configuration choices precede test evaluation; test labels are used only for reporting.

Evaluation metrics. We report ACC, macro-F1, 15-bin ECE, and the standard multiclass Brier score. ECE/Brier are scaled by 100; results are mean±std over three seeds.

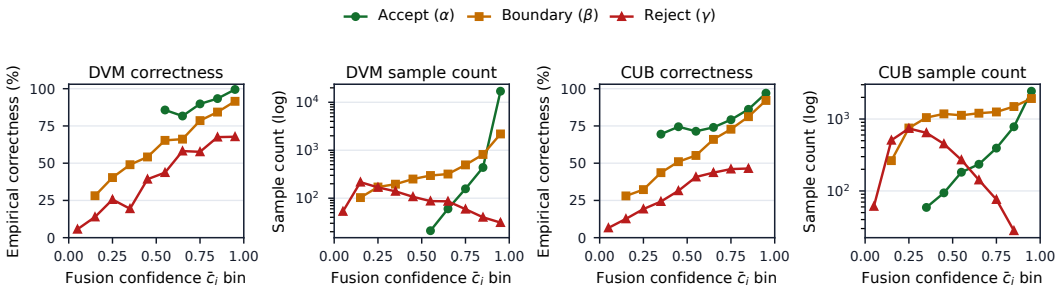

Figure 3: Two views of support-region differences on DVM (left pair) and CUB (right pair): correctness within fusion-confidence bins and sample counts across bins (log scale). The panels show regional reliability stratification and distinct confidence-distribution trends. Correctness is omitted for region–bin cells with fewer than 20 samples.

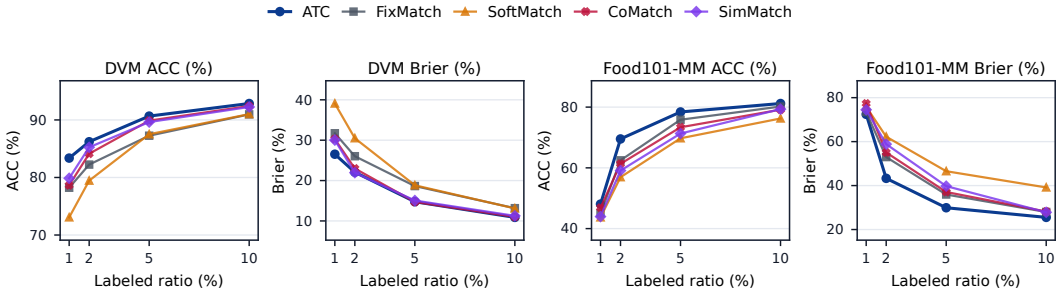

Figure 4: ACC (higher) and Brier (lower; ×100) on DVM and Food101-MM with 1/2/5/10% labels.

4.2 Main Results

Table 1 summarizes the main classification results. Among evaluated methods under the shared protocol, ATC has the highest reported mean ACC on all five datasets and macro-F1 on four, ranking top-two in every classification cell. On CUB, ATC outperforms the strong multimodal STiL baseline by 2.15 ACC and 3.05 macro-F1 points, consistent with the pronounced confidence–support separation in the correctness panels of Figure 3. The five benchmarks span image–tabular, image–attribute, image–text, image–metadata, and trimodal sequence settings.

Method	DVM		CUB		Food101-MM		PAD-UFES20		MOSI	
	ACC	F1	ACC	F1	ACC	F1	ACC	F1	ACC	F1
FixMatch	91.94$_{\pm 0.35}$	72.70$_{\pm 0.73}$	58.09$_{\pm 0.09}$	57.33$_{\pm 0.13}$	80.18$_{\pm 0.68}$	80.29$_{\pm 0.77}$	73.80$_{\pm 0.88}$	63.34$_{\pm 2.00}$	65.16$_{\pm 1.77}$	64.99$_{\pm 1.83}$
FlexMatch	92.38$_{\pm 0.45}$	73.28$_{\pm 1.20}$	58.96$_{\pm 1.02}$	57.95$_{\pm 0.54}$	80.59$_{\pm 0.20}$	80.92$_{\pm 0.17}$	70.64$_{\pm 4.50}$	61.68$_{\pm 5.31}$	66.38$_{\pm 1.24}$	**65.89**$_{\pm 1.37}$
SoftMatch	91.49$_{\pm 0.53}$	71.32$_{\pm 1.07}$	56.66$_{\pm 0.37}$	55.79$_{\pm 0.47}$	76.26$_{\pm 1.03}$	76.33$_{\pm 1.13}$	71.76$_{\pm 1.41}$	62.78$_{\pm 1.22}$	65.60$_{\pm 1.39}$	64.97$_{\pm 1.09}$
FreeMatch	91.86$_{\pm 0.37}$	71.41$_{\pm 1.22}$	60.26$_{\pm 0.44}$	58.82$_{\pm 0.64}$	79.84$_{\pm 0.55}$	80.16$_{\pm 0.36}$	74.41$_{\pm 1.27}$	65.15$_{\pm 6.07}$	60.45$_{\pm 0.84}$	60.37$_{\pm 0.84}$
MixMatch	92.56$_{\pm 0.57}$	72.79$_{\pm 2.24}$	61.28$_{\pm 0.79}$	59.92$_{\pm 0.77}$	79.94$_{\pm 0.47}$	79.95$_{\pm 0.41}$	73.39$_{\pm 2.21}$	64.48$_{\pm 2.66}$	65.06$_{\pm 3.23}$	64.46$_{\pm 2.81}$
CoMatch	92.54$_{\pm 0.42}$	73.05$_{\pm 1.57}$	58.61$_{\pm 0.82}$	57.69$_{\pm 0.63}$	79.16$_{\pm 0.58}$	79.15$_{\pm 0.56}$	73.90$_{\pm 0.98}$	64.07$_{\pm 3.25}$	64.14$_{\pm 1.51}$	63.15$_{\pm 1.95}$
SimMatch	92.47$_{\pm 0.51}$	73.67$_{\pm 1.08}$	58.34$_{\pm 0.27}$	57.60$_{\pm 0.03}$	79.33$_{\pm 0.18}$	79.39$_{\pm 0.04}$	73.39$_{\pm 1.33}$	60.42$_{\pm 4.82}$	63.95$_{\pm 2.48}$	63.66$_{\pm 2.31}$
CoTrain	88.88$_{\pm 1.21}$	67.67$_{\pm 1.89}$	58.95$_{\pm 1.04}$	57.12$_{\pm 0.61}$	82.19$_{\pm 0.25}$	82.09$_{\pm 0.30}$	74.72$_{\pm 1.79}$	64.42$_{\pm 4.83}$	64.92$_{\pm 2.19}$	64.79$_{\pm 2.16}$
STiL	92.31$_{\pm 0.84}$	73.53$_{\pm 1.56}$	62.80$_{\pm 0.85}$	61.06$_{\pm 0.86}$	80.83$_{\pm 0.35}$	80.88$_{\pm 0.32}$	73.50$_{\pm 0.47}$	63.04$_{\pm 2.58}$	66.42$_{\pm 1.13}$	65.44$_{\pm 0.45}$
ATC (ours)	**92.92**$_{\pm 0.38}$	**74.76**$_{\pm 1.18}$	**64.95**$_{\pm 0.82}$	**64.11**$_{\pm 0.88}$	**82.20**$_{\pm 0.36}$	**82.39**$_{\pm 0.39}$	**75.13**$_{\pm 2.38}$	**65.67**$_{\pm 8.39}$	**66.81**$_{\pm 1.10}$	65.70$_{\pm 0.46}$

Table 1: Main classification results (%; higher is better). Values are mean±std over three seeds; bold/underline mark the best and second-best means.

Raw Calibration Results. Table 2 reports raw ECE and Brier at the same checkpoints as Table 1. ECE measures calibration of the maximum confidence used for prediction, whereas Brier evaluates the full probability vector for each sample.

Method	DVM		CUB		Food101-MM		PAD-UFES20		MOSI	
	ECE	Brier	ECE	Brier	ECE	Brier	ECE	Brier	ECE	Brier
FixMatch	$1.41_{\pm 0.23}$	$11.52_{\pm 0.25}$	$4.62_{\pm 1.77}$	$55.42_{\pm 0.56}$	$6.18_{\pm 3.08}$	$27.99_{\pm 0.32}$	$9.37_{\pm 0.70}$	$38.27_{\pm 1.93}$	$27.48_{\pm 0.80}$	$59.82_{\pm 2.65}$
FlexMatch	$1.26_{\pm 0.49}$	$11.00_{\pm 0.63}$	$10.76_{\pm 3.66}$	$56.25_{\pm 1.18}$	$9.02_{\pm 1.22}$	$28.63_{\pm 0.66}$	$10.78_{\pm 2.77}$	$43.11_{\pm 6.96}$	$25.82_{\pm 1.03}$	$56.90_{\pm 1.92}$
SoftMatch	$6.47_{\pm 0.33}$	$12.96_{\pm 0.64}$	$9.35_{\pm 0.91}$	$58.58_{\pm 0.22}$	$21.34_{\pm 5.82}$	$39.15_{\pm 1.71}$	$9.11_{\pm 1.90}$	$41.95_{\pm 2.37}$	$11.14_{\pm 2.92}$	$45.63_{\pm 1.93}$
FreeMatch	$1.70_{\pm 0.35}$	$11.50_{\pm 0.36}$	$12.31_{\pm 0.46}$	$55.63_{\pm 0.57}$	$5.53_{\pm 2.03}$	$27.47_{\pm 0.48}$	$10.33_{\pm 1.39}$	$39.70_{\pm 2.72}$	$27.86_{\pm 4.24}$	$63.87_{\pm 4.83}$
MixMatch	$0.86_{\pm 0.11}$	$10.76_{\pm 0.65}$	$2.17_{\pm 0.12}$	$51.37_{\pm 0.99}$	$2.33_{\pm 1.53}$	$27.36_{\pm 0.81}$	$5.91_{\pm 1.81}$	$38.41_{\pm 4.20}$	$13.69_{\pm 1.44}$	$47.10_{\pm 2.14}$
CoMatch	$\mathbf{0.76}_{\pm 0.17}$	$10.74_{\pm 0.46}$	$4.02_{\pm 0.78}$	$54.87_{\pm 0.38}$	$3.03_{\pm 0.42}$	$28.18_{\pm 1.33}$	$8.90_{\pm 1.31}$	$38.78_{\pm 1.31}$	$14.09_{\pm 2.14}$	$49.31_{\pm 2.57}$
SimMatch	$0.91_{\pm 0.24}$	$11.00_{\pm 0.92}$	$5.62_{\pm 1.84}$	$55.49_{\pm 0.58}$	$3.13_{\pm 0.70}$	$27.85_{\pm 0.08}$	$8.36_{\pm 3.45}$	$39.35_{\pm 3.06}$	$22.94_{\pm 1.63}$	$55.56_{\pm 3.31}$
CoTrain	$1.73_{\pm 0.63}$	$15.92_{\pm 1.61}$	$5.98_{\pm 2.30}$	$54.81_{\pm 1.06}$	$3.03_{\pm 0.21}$	$\mathbf{23.81}_{\pm 0.76}$	$9.50_{\pm 2.32}$	$39.54_{\pm 2.93}$	$28.12_{\pm 1.86}$	$60.32_{\pm 3.47}$
STiL	$2.12_{\pm 0.45}$	$11.15_{\pm 0.98}$	$4.91_{\pm 0.54}$	$51.26_{\pm 0.67}$	$1.88_{\pm 0.69}$	$25.71_{\pm 0.75}$	$9.68_{\pm 2.08}$	$40.22_{\pm 2.22}$	$14.88_{\pm 0.78}$	$46.58_{\pm 0.99}$
ATC (ours)	$1.03_{\pm 0.16}$	$\mathbf{10.67}_{\pm 0.39}$	$\mathbf{1.96}_{\pm 0.22}$	$\mathbf{48.05}_{\pm 0.66}$	$\mathbf{1.55}_{\pm 0.30}$	$24.11_{\pm 0.62}$	$\mathbf{5.62}_{\pm 2.31}$	$\mathbf{38.06}_{\pm 2.40}$	$\mathbf{9.69}_{\pm 0.97}$	$\mathbf{44.31}_{\pm 0.93}$

Table 2: Raw ECE and Brier. Values are mean±std over three seeds; bold/underline mark the best and second-best means.

By reported mean, ATC is best among the evaluated methods in eight of ten calibration cells and top-two in nine. On CUB, ATC improves the best competing ECE and Brier by 0.21 and 3.21 points, respectively. STiL is second-best on Food101-MM ECE and CUB Brier, while ATC is best on both. On DVM, ATC obtains the best Brier while CoMatch gives the lowest ECE. Together with the classification results, these comparisons show competitive predictive and calibration performance under the shared protocol; the ablation below examines the contribution of confidence alignment.

Low-label Stress Test. Across 1/2/5/10% labels, Figure 4 shows that ATC maintains the highest or near-highest ACC and lowest or near-lowest Brier on DVM and Food101-MM, with the clearest advantage on Food101-MM.

Confidence Reliability and Sample Support. Figure 3 examines two complementary aspects of support regions on DVM and CUB. First, the correctness panels compare regions within the same fusion-confidence bins. Accept is generally most accurate, Boundary intermediate, and Reject least accurate, including at medium and high confidence. Thus similar fusion confidence can accompany different correctness depending on modality-head support. These within-bin gaps motivate conditioning reliability on both $\bar{c}_i$ and g_i; confidence is controlled at bin resolution.

Second, the count panels show how the prevalence of each region changes across confidence bins. Accept counts increase toward high confidence, whereas Reject counts generally decline after a low-confidence peak; Boundary follows a different, dataset-dependent profile. Their overlap shows that support regions span multiple confidence intervals. Together, the two views describe both how reliable predictions are within a bin and how samples are distributed across bins. Sparse high-confidence Reject cells additionally motivate shrinkage and fallback. Correctness is displayed only for cells with at least 20 samples.

Figure 5 complements the matched-bin analysis with aggregate regional summaries on DVM, CUB, and Food101-MM. Accept remains closest to the ideal confidence–correctness relation, Boundary occupies an intermediate position with non-negligible sample mass, and Reject shows the largest mismatch. This consistent geometry links modality-head support to both pseudo-label correctness and overconfidence, motivating separate regional targets rather than a single global correction.

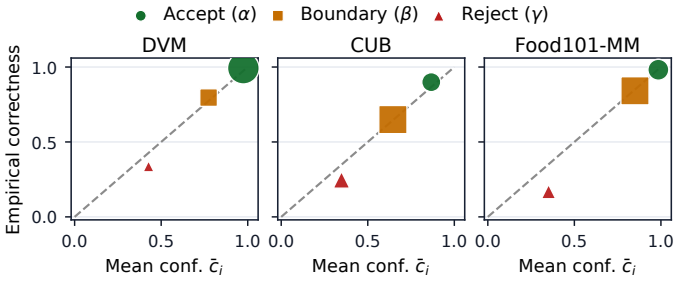

Figure 5: Regional confidence versus pseudo-label correctness on DVM, CUB, and Food101-MM. Bubble area denotes regional sample ratio; the diagonal is ideal calibration.

5 Ablation Studies

Representative Controller Ablation. CUB is used as a representative diagnostic case because its matched-confidence analysis clearly separates full, partial, and absent support. All variants use the same EMA fusion prediction source, so the comparison isolates how support structure is represented and used. The binary-partition, Three-way w/o L_{cal}, Full ATC entries are averaged over three seeds.

Variant	ACC↑	F1↑	ECE↓	Brier↓	PL Acc.↑	Eff. Cov.↑
Fusion-head confidence only	57.21	55.42	11.68	58.26	–	–
Binary-A: α vs. $\beta+\gamma$	64.72	63.78	2.79	48.23	–	–
Binary-B: $\alpha+\beta$ vs. γ	<u>64.86</u>	<u>63.86</u>	2.61	<u>48.09</u>	–	–
Three-way w/o $\mathcal{L}_{\text{cal}}$	64.18	63.39	2.49	49.18	70.24	44.46
Three-way fixed target	64.10	62.90	<u>2.00</u>	48.92	–	–
Full ATC (three-way)	**64.95**$(+0.09)$	**64.11**$(+0.25)$	**1.96**(-0.04)	**48.05**(-0.04)	**71.13**$(+0.89)$	**44.87**$(+0.41)$

Table 3: CUB controller ablation under the shared EMA fusion source and protocol. Binary-A/B denote α vs. $\beta+\gamma$ / $\alpha+\beta$ vs. γ. PL Acc. is the accuracy among admitted pseudo-labels; Eff. Cov. is the fraction of the unlabeled pool that is both admitted and correctly pseudo-labeled. Parenthesized Δ values after Full ATC are relative to the second-best result for ACC/F1/ECE/Brier and to Three-way w/o $\mathcal{L}_{\text{cal}}$ for PL Acc./Eff. Cov. Dashes mark unavailable diagnostics; ECE/Brier are $\times 100$.

Regional supervision. Table 3 distinguishes the effects of regional supervision, partition granularity, and replay refinement. Full ATC improves ACC/F1 over fusion confidence alone by 7.74/8.69 points. Both binary variants recover most of this classification gain, identifying support-conditioned supervision as the principal source of improvement in this comparison.

Retaining partial support. The three-way partition achieves the lowest mean ECE. Relative to Binary-A (Accept vs. Boundary+Reject) and Binary-B (Accept+Boundary vs. Reject), Full ATC reduces ECE from 2.79 and 2.61 to 1.96: improvements of 0.83 and 0.65 points on the $\times 100$ scale. At the reported means, Boundary yields lower ECE with similar classification performance.

Confidence alignment and replay. Adding $\mathcal{L}_{\text{cal}}$ reduces ECE/Brier from 2.49/49.18 to 1.96/48.05. It also raises pseudo-label accuracy from 70.24% to 71.13% and effective coverage from 44.46% to 44.87%, gains of 0.89 and 0.41 percentage points, respectively. This joint improvement shows that confidence alignment selects more accurate pseudo-labels without sacrificing usable unlabeled supervision. Relative to fixed targets, replay improves ACC/F1 by 0.85/1.21 points and Brier by 0.87 points, while ECE changes from 2.00 to 1.96. Together, regional supervision drives classification, confidence alignment improves pseudo-label accuracy and effective coverage, and replay improves all four reported performance/calibration metrics. These gains show that ATC jointly strengthens sample selection, supervision, and calibration.

Discussion and Limitations. Regional pooling stabilizes reliability estimation with limited labeled data, while Boundary retains partially supported samples. We use frozen features to focus the comparison on the learning strategy; end-to-end training remains to be studied. Correlated modality errors may still produce incorrect Accept samples, and replay on the labeled training set may yield optimistic reliability estimates. Shrinkage helps in sparse regions but does not remove this bias. Replay and routing are used only during training, so inference relies on the fusion head alone.

6 Conclusion

We presented an Adaptive Three-Way Calibration (ATC) method for semi-supervised multimodal learning. By modeling full, partial, and absent support for fusion predictions, ATC conditions pseudo-label supervision on Accept, Boundary, and Reject regions. Reliability estimated from replay data further enhances target shaping and confidence alignment without requiring additional annotations. Experimental results show that the proposed ATC ranks the top two in 19 of 20 primary comparisons, and matched-confidence evaluation and ablation studies show that incorporating modality-head evidence improves pseudo-label reliability over relying only on fusion confidence.

AI USE STATEMENT

A generative AI tool was used solely for language polishing, grammar correction, and formatting assistance. It was not used to generate scientific ideas, methods, experimental results, analyses, or conclusions. The authors manually reviewed all AI-assisted edits for technical accuracy and consistency with the original manuscript and take full responsibility for the final content of this work.

REFERENCES

Eric Arazo, Diego Ortego, Paul Albert, Noel E. O'Connor, and Kevin McGuinness. Pseudo-labeling and confirmation bias in deep semi-supervised learning. In *International Joint Conference on Neural Networks*, pp. 1–8, 2020.

Jinsoo Bae, Seoung Bum Kim, and Hyungrok Do. Calimatch: Adaptive calibration for improving safe semi-supervised learning. In *IEEE/CVF International Conference on Computer Vision*, pp. 2867–2876. IEEE, 2025.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019.

David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*, 2020.

David Berthelot, Rebecca Roelofs, Kihyuk Sohn, Nicholas Carlini, and Alex Kurakin. Adamatch: A unified approach to semi-supervised learning and domain adaptation. In *International Conference on Learning Representations*, 2022.

L. Bossard, M. Guillaumin, and L. Van Gool. Food-101: Mining discriminative components with random forests. In *European Conference on Computer Vision*, 2014.

Hao Chen, Ran Tao, Yue Fan, Yidong Wang, Jindong Wang, Bernt Schiele, Xing Xie, Bhiksha Raj, and Marios Savvides. Softmatch: Addressing the quantity-quality trade-off in semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Junchi Chen, Richong Zhang, and Junfan Chen. Semi-supervised multimodal classification through learning from modal and strategic complementarities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 15812–15820, 2025. doi: 10.1609/aaai.v39i15.33736.

Zebang Cheng, Yuxiang Lin, Zhaoru Chen, Xiang Li, Shuyi Mao, Fan Zhang, Daijun Ding, Bowen Zhang, and Xiaojiang Peng. Semi-supervised multimodal emotion recognition with expression mae. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 9436–9440, 2023. doi: 10.1145/3581783.3612840.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Charles Corbiere, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Perez. Addressing failure prediction by learning model confidence. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Siyi Du, Shaoming Zheng, Yinsong Wang, Wenjia Bai, Declan P. O'Regan, and Chen Qin. Tip: Tabular-image pre-training for multimodal classification with incomplete data. In *European Conference on Computer Vision*, pp. 478–496, 2024.

Siyi Du, Xinzhe Luo, Declan P. O'Regan, and Chen Qin. Stil: Semi-supervised tabular-image learning for comprehensive task-relevant information exploration in multimodal classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15549–15559, 2025.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pp. 4878–4887, 2017.

Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *International Conference on Machine Learning*, pp. 2151–2159, 2019.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 1321–1330, 2017.

Paul Hager, Martin J. Menten, and Daniel Rueckert. Best of both worlds: Multimodal contrastive learning with tabular and imaging data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23924–23935, 2023.

Zongbo Han, Changqing Zhang, Huazhu Fu, and Joey Tianyi Zhou. Trusted multi-view classification. In *International Conference on Learning Representations*, 2021.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*, 2017.

Jingmin Huang, Bowei Chen, Lan Luo, Shigang Yue, and Iadh Ounis. Dvm-car: A large-scale automotive dataset for visual marketing research and applications. In *IEEE International Conference on Big Data*, pp. 4140–4147, 2022.

Heinrich Jiang, Been Kim, Melody Guan, and Maya Gupta. To trust or not to trust a classifier. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Meelis Kull, Miquel Perello-Nieto, Markus Kängsepp, Telmo Silva Filho, Hao Song, and Peter Flach. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Samuli Laine and Timo Aila. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*, 2017.

Junnan Li, Caiming Xiong, and Steven C. H. Hoi. Comatch: Semi-supervised learning with contrastive graph regularization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9475–9484, 2021.

Siyuan Li, Weiyang Jin, Zedong Wang, Fang Wu, Zicheng Liu, Cheng Tan, and Stan Z. Li. Semireward: A general reward model for semi-supervised learning. In *International Conference on Learning Representations*, 2024.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Computing Surveys*, 56(10): 1–42, 2024.

Huan Ma, Qingyang Zhang, Changqing Zhang, Bingzhe Wu, Huazhu Fu, Joey Tianyi Zhou, and Qinghua Hu. Calibrating multimodal learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 23429–23450. PMLR, 2023.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems*, volume 34, pp. 15682–15694, 2021.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. Obtaining well calibrated probabilities using bayesian binning. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pp. 2901–2907, 2015.

Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning*, pp. 689–696, 2011.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pp. 625–632, 2005.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

A. G. C. Pacheco and R. A. Krohling. The impact of patient clinical information on automated skin cancer detection. *Computers in Biology and Medicine*, 116:103545, 2020.

Kihyuk Sohn, David Berthelot, Nicholas Carlini, Zizhao Zhang, Han Zhang, Colin Raffel, Ekin Dogus Cubuk, Alexey Kurakin, and Chun-Liang Li. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*, volume 33, pp. 596–608, 2020.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Masahito Toba, Seiichi Uchida, and Hideaki Hayashi. Pseudo-label learning with calibrated confidence using an energy-based model. In *International Joint Conference on Neural Networks*, pp. 1–8, 2024.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The caltech-ucsd birds-200-2011 dataset. Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*, 2023.

Yuchao Wang, Haochen Wang, Yujun Shen, Jingjing Fei, Wei Li, Guoqiang Jin, Liwei Wu, Rui Zhao, and Xinyi Le. Semi-supervised semantic segmentation using unreliable pseudo-labels. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4248–4257, 2022.

Mengyao Xie, Zongbo Han, Changqing Zhang, Yichen Bai, and Qinghua Hu. Exploring and exploiting uncertainty for incomplete multi-view classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19873–19882, 2023.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6256–6268, 2020.

Yi Xu, Lei Shang, Jinxing Ye, Qi Qian, Yu-Feng Li, Baigui Sun, Hao Li, and Rong Jin. Dash: Semi-supervised learning with dynamic thresholding. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 11525–11536, 2021.

A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016.

Bianca Zadrozny and Charles Elkan. Obtaining calibrated probability estimates from decision trees and naive bayesian classifiers. In *Proceedings of the 18th International Conference on Machine Learning*, pp. 609–616, 2001.

Bowen Zhang, Yidong Wang, Wenxin Hou, Hao Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*, volume 34, pp. 18408–18419, 2021.

Mingkai Zheng, Shan You, Lang Huang, Fei Wang, Chen Qian, and Chang Xu. Simmatch: Semi-supervised learning with similarity matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14471–14481, 2022.

A Implementation and Optimization Details

A.1 Aligned Protocol

The compared methods use the frozen features and label budgets described in Section 4. Replay draws only from labeled training examples and introduces no additional labels. Validation is reserved for checkpoint and predefined operating-point selection; test labels are used only for final reporting. The labeled/unlabeled split is fixed across the reported seeds 2023, 2024, and 2025. Randomness affects initialization, minibatch order, and stochastic feature augmentation, rather than changing the label budget. The implementation uses PyTorch with CUDA-enabled GPU execution.

Baseline adaptations share the EMA fusion source, features, splits, and training budget while retaining their method-specific selection, weighting, and consistency rules. Unlike evidential multi-view methods that model view reliability or missing-view uncertainty (Han et al., 2021; Xie et al., 2023), ATC uses modality-head agreement only to control pseudo-label training and retains the student fusion head at inference.

A.2 Optimization Settings

Table 4 records the reported optimization settings. LR denotes learning rate and WD denotes weight decay; batch is minibatch size. The loss coefficients correspond to the objective in the main text.

Dataset	LR	WD	Batch	Epochs	λ_u	λ_{cal}
CUB	3.8e-4	1.0e-3	64	46	0.30	0.060
DVM	4.5e-4	1.0e-3	64	46	0.18	0.140
Food101-MM	3.8e-4	1.0e-3	64	50	0.50	0.035
MOSI	3.0e-4	1.0e-4	64	44	0.26	0.045
PAD-UFES20	4.0e-4	1.5e-3	64	64	0.26	0.010

Table 4: Optimization settings for the reported operating points. All rows use seeds 2023, 2024, and 2025 and EMA teacher updates.

A.3 Fixed Regional Controls

Table 5 separates the fixed admission thresholds τ_g, classification weights w_g, and support-target blending coefficients b_g from the replay-estimated target t_g. Each triplet follows the order Accept/Boundary/Reject. The controller uses $K = 10$ confidence bins; κ is the regional shrinkage strength, not the bin-specific coefficient κ_b.

Dataset	$\tau_\alpha/\tau_\beta/\tau_\gamma$	$w_\alpha/w_\beta/w_\gamma$	$b_\alpha/b_\beta/b_\gamma$	κ
CUB	.903/.903/.903	1.00/.82/.82	.95/.86/.86	50
DVM	.540/.702/.771	1.00/.82/.82	.95/.86/.86	50
Food101-MM	.500/.547/.700	1.00/.82/.55	.00/.00/.00	50
MOSI	.640/.700/.700	1.00/.82/.25	.95/.86/.60	50
PAD-UFES20	.640/.700/.700	1.00/.95/.70	.95/.86/.86	50

Table 5: Fixed regional controls and regional shrinkage strengths. Triplets are ordered as $\alpha/\beta/\gamma$. The coefficient b_g blends supporting heads with the fusion distribution; it differs from the bin index.

For the reported operating points, replay estimates the regional target once after warm-up and holds it fixed for the remaining training epochs. This data-dependent estimate supplies stable regional supervision while the teacher continues to receive EMA updates. The fixed thresholds, weights, and blending coefficients are shared across the three seeds of each dataset. Seed variation therefore reflects initialization, minibatch order, and stochastic feature augmentation rather than seed-specific controller selection or retuning. No regional statistic is re-estimated from validation or test examples, preserving the common label budget.

B Controller Interpretation and Training Flow

B.1 Support-Conditioned Reliability

Let $Z = 1$ indicate a correct teacher pseudo-label. A confidence-only reliability estimate averages over support patterns:

$$P(Z = 1 \mid \bar{c}) = \sum_{g \in \{\alpha,\beta,\gamma\}} P(Z = 1 \mid \bar{c}, g) P(g \mid \bar{c}). \tag{14}$$

The same marginal confidence can therefore conceal different regional correctness probabilities. This identity motivates using support as an additional condition. The controller's ordering constraint expresses a regional inductive bias rather than an assumption needed for this identity.

Region–confidence bins summarize local replay evidence, whereas the target t_g in the loss pools evidence across the region. This separation combines fine-grained diagnostics with data-efficient training targets. Both target placements use the regional shrinkage estimator defined in the main text.

B.2 Region–Confidence Summaries

Replay statistics within each region are organized into K equal-width confidence bins $\{I_b\}_{b=1}^{K}$, with the final bin including confidence one:

$$I_b = \begin{cases} [(b-1)/K, b/K), & b = 1, \ldots, K-1, \\ [(K-1)/K, 1], & b = K. \end{cases} \tag{15}$$

For replay index set $\mathcal{R}_{g,b} = \{i : g_i = g, \bar{c}_i \in I_b\}$, the sufficient statistics are

$$\begin{aligned} n_{g,b} &= |\mathcal{R}_{g,b}|, \\ a_{g,b} &= \frac{\sum_{i \in \mathcal{R}_{g,b}} z_i}{\max(1, n_{g,b})}, \quad \bar{c}_{g,b} = \frac{\sum_{i \in \mathcal{R}_{g,b}} \bar{c}_i}{\max(1, n_{g,b})}. \end{aligned} \tag{16}$$

The shrunk region–bin reliability summary is

$$\hat{r}_{g,b} = \frac{n_{g,b} a_{g,b} + \kappa_b t_g^0}{n_{g,b} + \kappa_b}, \tag{17}$$

where $\kappa_b > 0$ controls shrinkage toward the regional prior t_g^0. Empty cells therefore equal the prior, while data-rich cells approach their empirical correctness. The pooled statistics used for training satisfy

$$n_g = \sum_{b=1}^{K} n_{g,b}, \qquad a_g = \frac{\sum_{b=1}^{K} n_{g,b} a_{g,b}}{\max(1, n_g)}, \qquad \bar{c}_g = \frac{\sum_{b=1}^{K} n_{g,b} \bar{c}_{g,b}}{\max(1, n_g)}. \tag{18}$$

Thus the bin summaries diagnose local variation, whereas the regional target uses all replay evidence assigned to g.

B.3 Shrinkage, Ordering, and Fallback

The unprojected regional target can be written directly as

$$\hat{t}_g = \frac{n_g a_g + \kappa t_g^0}{n_g + \kappa}, \tag{19}$$

which is equivalent to the convex-combination form in the main text. Let $[\ell_g, h_g]$ denote the allowed interval for region g. The joint feasible set and Euclidean projection are

$$\begin{aligned} \mathcal{F} &= \{\mathbf{u} : \ell_g \le u_g \le h_g \; \forall g, \; u_\alpha \ge u_\beta \ge u_\gamma\}, \\ \tilde{\mathbf{t}} &= \arg\min_{\mathbf{u} \in \mathcal{F}} \sum_{g \in \{\alpha,\beta,\gamma\}} (u_g - \hat{t}_g)^2, \\ t_g &= \begin{cases} t_g^0, & n_g < n_{\min,g}, \\ \tilde{t}_g, & n_g \ge n_{\min,g}. \end{cases} \end{aligned} \tag{20}$$

This form makes explicit that ordering is enforced jointly, whereas fallback is region-specific. The reported implementation estimates these targets once after warm-up and keeps them fixed during subsequent student updates.

B.4 Squared-Error Alignment

For an unconstrained scalar s and regional correctness probability $p_g = P(Z = 1 \mid g)$,

$$\mathbb{E}[(s - Z)^2 \mid g] = (s - p_g)^2 + p_g(1 - p_g). \tag{21}$$

The population minimizer is p_g, motivating squared-error reliability alignment. The implemented loss uses the replay estimate t_g as a surrogate for regional correctness. Since maximum class probability lies in $[1/C, 1]$, the attainable population and surrogate optima are

$$s_g^* = \mathrm{clip}\left(p_g, \frac{1}{C}, 1\right), \qquad \tilde{s}_g^* = \mathrm{clip}\left(t_g, \frac{1}{C}, 1\right). \tag{22}$$

The resulting test-set calibration is evaluated empirically with ECE and Brier rather than inferred from the population identity alone.

B.5 Training Schedule

Algorithm 1 separates warm-up, one-time replay estimation, and region-aware student–teacher updates. Teacher outputs, region assignments, and replay targets are stop-gradient quantities.

Algorithm 1 ATC training with once-after-warm-up replay

Require: Labeled set $\mathcal{D}_l$, unlabeled set $\mathcal{D}_u$; fixed controls $\{\tau_g, w_g, b_g\}$; priors and safeguards; EMA decay μ
Ensure: Trained student fusion predictor
Initialize student parameters θ and teacher $\bar{\theta} \leftarrow \theta$; complete the configured warm-up
Replay estimation
for all $(x_i, y_i) \in \mathcal{D}_l$ **do**
  Evaluate the teacher weak view; compute $(\tilde{y}_i, \bar{c}_i, g_i, b_i, z_i)$
  Accumulate regional and region–bin statistics
end for
Compute $\hat{t}_g$ and safeguarded t_g using Eqs. (19)–(20)
Region-aware updates
for each subsequent labeled–unlabeled minibatch **do**
  Evaluate the teacher weak view and student strong view
  Route each unlabeled sample, apply ξ_i, and construct $\mathbf{q}_i$ or $\mathbf{q}_i^{(t)}$
  Update θ using $\mathcal{L}_{\mathrm{sup}} + \lambda_u \mathcal{L}_u + \lambda_{\mathrm{cal}} \mathcal{L}_{\mathrm{cal}}$
  Update $\bar{\theta} \leftarrow \mu\bar{\theta} + (1 - \mu)\theta$
end for
return the student fusion predictor; discard training-only controls

C Additional Reliability Diagnostics

C.1 Aggregate Region Audit

Table 6 summarizes Full ATC across datasets and seeds, with region ratios normalized before averaging. Accept has the highest pseudo-label accuracy and admission rate, followed by Boundary and Reject. These descriptive aggregates complement the matched-confidence analysis in Figure 3.

Region	Ratio	PL Acc.	Confidence	Gap	Admitted
α/Accept	46.7	86.6	91.7	5.2	96.1
β/Boundary	47.4	67.3	81.9	14.6	78.8
γ/Reject	5.9	36.4	50.2	13.8	3.5

Table 6: Region audit averaged across datasets and seeds. PL Acc. denotes pseudo-label accuracy; Gap is the reported absolute confidence–accuracy gap, and Admitted is the admission percentage for each region. All values are reported on a percentage scale.