Adaptive Momentum and EMA-weighted Modeling for Imbalanced Label Distribution Learning

Appendix

A.1 Algorithm Details

The pseudo code of our method is presented in **Algorithm 1**.

Algorithm 1 AMEMA: Adaptive Momentum Allocation and EMA-weighted Modeling for Imbalanced Label Distribution Learning

```
Input: Dataset $\mathcal{D}$; learning rate $\eta$; initial momenta $\mu_s, \mu_{ns}$; EMA factor $m$; adaptation rate $\gamma$; sigmoid factor $\alpha$; baseline $b$; total steps $T$
Output: Parameters $\theta_s$, $\theta_{ns}$
final weights $w_s, w_{ns}$
Initialize model parameters $\theta_s^{(0)}, \theta_{ns}^{(0)}$
Initialize momentum buffers: $v_s^{(0)} \leftarrow 0,\ v_{ns}^{(0)} \leftarrow 0$
Initialize EMA losses: $E_s^{(0)} \leftarrow 0,\ E_{ns}^{(0)} \leftarrow 0$
for each training step $t = 1$ to $T$ do
    Sample a mini-batch $(x, y)$ from $\mathcal{D}$
    Compute the KL loss for the dominant branch:
    $\mathcal{L}_s^{(t)} \leftarrow \mathrm{KL}\big(\bar{D}_{\boldsymbol{x}} \parallel \bar{f}_\theta(\boldsymbol{x})\big)$
    Compute the KL loss for the non-dominant branch:
    $\mathcal{L}_{ns}^{(t)} \leftarrow \mathrm{KL}\big(\hat{D}_{\boldsymbol{x}} \parallel \hat{f}_\theta(\boldsymbol{x})\big)$
    Update EMA of losses:
    $E_s^{(t)} \leftarrow m\, E_s^{(t-1)} + (1-m)\, \mathcal{L}_s^{(t)}$
    $E_{ns}^{(t)} \leftarrow m\, E_{ns}^{(t-1)} + (1-m)\, \mathcal{L}_{ns}^{(t)}$
    Compute the loss ratio:
    $r^{(t)} \leftarrow E_{ns}^{(t)} / (E_s^{(t)} + \epsilon)$
    Compute dynamic branch weights:
    $w_{ns}^{(t)} \leftarrow b + (1-b)\, \sigma(\alpha(r^{(t)} - 1))$
    $w_s^{(t)} \leftarrow 1 - w_{ns}^{(t)}$
    Compute the total weighted loss:
    $\mathcal{L}^{(t)} \leftarrow w_s^{(t)} \mathcal{L}_s^{(t)} + w_{ns}^{(t)} \mathcal{L}_{ns}^{(t)}$
    Compute gradients for both branches:
    $g_s^{(t)} \leftarrow \nabla_{\theta_s} \mathcal{L}_s^{(t)}$
    $g_{ns}^{(t)} \leftarrow \nabla_{\theta_{ns}} \mathcal{L}_{ns}^{(t)}$
    Update momentum buffers:
    $\boldsymbol{v}_s^{(t)} \leftarrow \mu_s^{(t-1)} \boldsymbol{v}_s^{(t-1)} + g_s^{(t)}$
    $\boldsymbol{v}_{ns}^{(t)} \leftarrow \mu_{ns}^{(t-1)} \boldsymbol{v}_{ns}^{(t-1)} + g_{ns}^{(t)}$
    Update model parameters:
    $\boldsymbol{\theta}_s^{(t)} \leftarrow \boldsymbol{\theta}_s^{(t-1)} - \eta \boldsymbol{v}_s^{(t)}$
    $\boldsymbol{\theta}_{ns}^{(t)} \leftarrow \boldsymbol{\theta}_{ns}^{(t-1)} - \eta \boldsymbol{v}_{ns}^{(t)}$
    Compute the modulation factor:
    $\delta^{(t)} \leftarrow \sigma(\alpha(r^{(t)} - 1))$
    Update momentum coefficients:
    $\mu_{ns}^{(t)} \leftarrow \mu_{ns}^{(t-1)} + \gamma \delta^{(t)}$
    $\mu_s^{(t)} \leftarrow \mu_s^{(t-1)} - \gamma \delta^{(t)}$
end for
return $\boldsymbol{\theta}_s^{(T)}$, $\boldsymbol{\theta}_{ns}^{(T)}$, $w_s^{(T)}$, $w_{ns}^{(T)}$
```

A.2 Proof of the Derivation of Momentum Variance Bound and Convergence Rate

In this section, we provide a rigorous derivation of the variance bound for the momentum buffer and the convergence rate of momentum-based stochastic gradient descent (SGD). These theoretical insights form the foundation for our proposed momentum allocation strategy.

A.2.1 Variance Bound of Momentum Buffer

We consider the momentum update at iteration t given by:

$$\boldsymbol{v}^{(t)} = \mu \boldsymbol{v}^{(t-1)} + \nabla \mathcal{L}^{(t)},$$

where $\mu \in [0, 1)$ is the momentum coefficient, and $\nabla \mathcal{L}^{(t)}$ denotes the stochastic gradient at iteration t.

Assume the stochastic gradients are unbiased and have bounded variance:

$$\mathbb{E}[\nabla \mathcal{L}^{(t)}] = 0, \quad \mathrm{Var}[\nabla \mathcal{L}^{(t)}] = \sigma^2,$$

and are independent across iterations.

To derive the variance of $\boldsymbol{v}^{(t)}$, we apply the law of total variance under independence:

$$\mathrm{Var}[\boldsymbol{v}^{(t)}] = \mathrm{Var}[\mu \boldsymbol{v}^{(t-1)} + \nabla \mathcal{L}^{(t)}] = \mu^2 \mathrm{Var}[\boldsymbol{v}^{(t-1)}] + \mathrm{Var}[\nabla \mathcal{L}^{(t)}].$$

Let $V_t = \mathrm{Var}[\boldsymbol{v}^{(t)}]$. We obtain a recurrence relation:

$$V_t = \mu^2 V_{t-1} + \sigma^2,$$

with an initial condition $V_0 = 0$ (assuming zero initialization for momentum buffer).

This is a linear non-homogeneous recurrence. Its closed-form solution is:

$$V_t = \sigma^2 \sum_{i=0}^{t-1} \mu^{2i} = \sigma^2 \cdot \frac{1 - \mu^{2t}}{1 - \mu^2}.$$

As $t \to \infty$, the geometric sum converges:

$$\lim_{t \to \infty} V_t = \frac{\sigma^2}{1 - \mu^2}.$$

Hence, the asymptotic variance of the momentum buffer is bounded by:

$$\mathrm{Var}[\boldsymbol{v}^{(t)}] \leq \frac{\sigma^2}{1 - \mu^2}.$$

In practice, since μ is often chosen close to 1, we use the loose upper bound:

$$\frac{1}{1 - \mu^2} \leq \frac{1}{1 - \mu},$$

Leading to the practical bound:

$$\mathrm{Var}[\boldsymbol{v}^{(t)}] \leq \frac{\sigma^2}{1 - \mu}.$$

This corresponds to the specific form used in our main text for the non-dominant branch:

$$\mathrm{Var}[v_{ns}^{(t)}] \leq \frac{1}{1 - \mu_{ns}} \mathrm{Var}[\nabla \mathcal{L}_{ns}].$$

A.2.2 Convergence Rate of Momentum SGD

Consider minimizing a convex function $f(\boldsymbol{\theta})$ using momentum SGD:

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \eta \boldsymbol{v}^{(t)},$$

where η is the learning rate.

Under standard assumptions (convexity of f, bounded variance σ^2 of stochastic gradients, and proper choice of η and μ), it can be shown that the expected suboptimality satisfies

$$\mathbb{E}[f(\boldsymbol{\theta}^{(t)})] - f^* \leq \mathcal{O}\left(\frac{1}{(1 - \mu)t}\right) + \mathcal{O}(\eta \sigma^2),$$

where f^* is the global minimum of f.

This result implies that a larger momentum coefficient μ improves the convergence rate by effectively increasing the step size and reducing the noise variance via smoothing.

Therefore, assigning a larger momentum coefficient helps to suppress gradient noise variance and accelerate convergence, which theoretically justifies the design of assigning higher momentum to branches with greater learning difficulty.

Method	**Movie**	**SCUT-FBP**	**Emotion6**	**Flickr_LDL**	**RAF-ML**	**Natural Scene**
SA-BFGS	0.546±0.010•	0.244±0.026•	0.154±0.015•	0.099±0.005•	0.228±0.016•	0.247±0.020•
EDL-LRL	0.529±0.016•	0.573±0.032•	0.417±0.012•	0.209±0.005•	0.395±0.013•	0.394±0.025•
LDLSF	0.514±0.012•	0.416±0.038•	0.397±0.017•	0.253±0.013•	0.476±0.026•	0.359±0.021•
LDL-LCLR	0.550±0.010•	0.593±0.031•	0.347±0.014•	0.186±0.012•	0.559±0.012•	0.301±0.017•
Adam-LDL-SCL	0.240±0.049•	0.395±0.031•	0.295±0.031•	0.141±0.015•	0.259±0.016•	0.345±0.034•
LDL-LDM	0.437±0.031•	0.521±0.050•	0.361±0.015•	0.217±0.012•	0.339±0.027•	0.343±0.022•
OFR-FL	0.534±0.018•	0.569±0.049•	0.430±0.022•	0.219±0.010•	0.368±0.027•	0.390±0.022•
OFR-CB	0.540±0.023•	0.559±0.043•	0.430±0.017•	0.218±0.006•	0.373±0.018•	0.389±0.021•
OFR-DB	0.631±0.008•	0.577±0.052•	0.469±0.010•	0.288±0.031•	0.395±0.018•	0.396±0.020•
RDA	0.709±0.009•	0.630±0.014•	0.512±0.008•	0.306±0.011•	0.509±0.008•	0.403±0.0591•
DILDL	0.729±0.031•	0.656±0.020•	0.528±0.026•	0.320±0.019•	0.537±0.039•	0.409±0.011•
AMEMA	**0.7481±0.0143**	**0.7403±0.0180**	**0.5389±0.0096**	**0.3562±0.0066**	**0.5488±0.0084**	**0.4240±0.0262**

Table 1: Experimental results on ILDL datasets measured by Intersection Similarity (↑).

Method	**Movie**	**SCUT-FBP**	**Emotion6**	**Flickr_LDL**	**RAF-ML**	**Natural Scene**
SA-BFGS	0.612±0.011•	0.337±0.035•	0.253±0.024•	0.149±0.006•	0.336±0.022•	0.334±0.334•
EDL-LRL	0.589±0.018•	0.679±0.041•	0.467±0.017•	0.219±0.008•	0.429±0.019•	0.505±0.034•
LDLSF	0.573±0.015•	0.471±0.050•	0.459±0.020•	0.290±0.017•	0.570±0.032•	0.457±0.029•
LDL-LCLR	0.613±0.012•	0.705±0.042•	0.408±0.018•	0.224±0.016•	0.678±0.016•	0.384±0.026•
Adam-LDL-SCL	0.398±0.092•	0.438±0.038•	0.331±0.031•	0.156±0.016•	0.302±0.016•	0.420±0.042•
LDL-LDM	0.521±0.040•	0.616±0.067•	0.399±0.026•	0.238±0.018•	0.373±0.033•	0.446±0.029•
OFR-FL	0.589±0.021•	0.674±0.072•	0.482±0.030•	0.229±0.012•	0.386±0.037•	0.506±0.028•
OFR-CB	0.598±0.027•	0.659±0.061•	0.480±0.024•	0.226±0.008•	0.390±0.024•	0.504±0.026•
OFR-DB	0.709±0.011•	0.715±0.082•	0.543±0.015•	0.350±0.060•	0.422±0.030•	0.575±0.026•
RDA	0.816±0.012•	0.795±0.022•	0.622±0.009•	0.388±0.024○	0.659±0.012•	0.597±0.017○
DILDL	0.834±0.014•	0.817±0.034•	0.640±0.016•	0.407±0.018•	0.791±0.032•	0.610±0.020•
AMEMA	**0.8549±0.0143**	**0.8472±0.0133**	**0.6793±0.0107**	**0.4508±0.0136**	**0.7966±0.0098**	**0.6207±0.0256**

Table 2: Experimental results on ILDL datasets measured by Cosine Coefficient (↑).

A.3 More Experimental Results and Visualization Analysis

To provide a comprehensive evaluation of the AMEMA method, this section presents extensive quantitative results and in-depth analysis. We systematically compare AMEMA with classical and state-of-the-art methods across six public ILDL datasets, using four widely recognized evaluation metrics: Intersection similarity (Table 1), Cosine similarity (Table 2), Canberra distance (Table 3), and Clark distance (Table 4). In addition, Table 5 and Table 6 provide detailed ablation results to analyze the contribution of each module. Figures 1–7 visualize the training dynamics and momentum evolution under imbalanced label distributions.

A.3.1 Proof of the Quantitative Comparison

As shown in Table 1, AMEMA consistently achieves the highest Intersection similarity across all datasets. On the *Movie* dataset, AMEMA achieves 0.7481, clearly outperforming DILDL (0.729), RDA (0.709), and OFR-DB (0.631). On the *SCUT-FBP*, AMEMA achieves 0.7403, higher than DILDL (0.656) and LDL-LCLR (0.593). For the *Flickr_LDL* dataset, AMEMA achieves 0.3562, while DILDL and RDA achieve 0.320 and 0.306, respectively. Similar improvements are observed on *RAF-ML* and *Natural Scene*.

As shown in Table 2, AMEMA also achieves the highest Cosine similarity across all datasets. For example, on the *Movie*, AMEMA achieves 0.8549, higher than DILDL (0.834) and OFR-DB (0.709). On the *SCUT-FBP*, AMEMA reaches 0.8472, outperforming DILDL (0.817) and RDA (0.795). Even on the *Flickr_LDL* dataset, AMEMA achieves 0.4508, higher than DILDL (0.407) and RDA (0.388).

Table 3 shows that AMEMA achieves the lowest (i.e., best) Canberra distance on nearly all datasets. On the *Movie*, AMEMA achieves 1.2893, while DILDL, RDA, and OFR-DB achieve 1.529, 1.549, and 1.837. On the *SCUT-FBP*, AMEMA achieves 2.3679, outperforming DILDL (2.447). On the *Emotion6*, AMEMA reaches 3.7940, better than DILDL (3.832). It should be noted that on the *Flickr_LDL*, although SA-BFGS achieves a lower Canberra distance (4.100), this comes at the cost of much lower Intersection and Cosine similarity, indicating a trade-off among metrics.

Method	Movie	SCUT-FBP	Emotion6	Flickr_LDL	RAF-ML	Natural Scene
SA-BFGS	2.357±0.052•	4.240±0.083•	5.618±0.178•	**4.100±0.613•**	4.733±0.038•	7.107±0.064•
EDL-LRL	2.412±0.074•	2.988±0.080•	4.526±0.072•	6.955±0.044○	3.872±0.090•	7.111±0.187•
LDLSF	2.572±0.062•	3.244±0.180•	4.422±0.120•	4.851±0.065○	3.555±0.100•	7.494±0.092•
LDL-LCLR	2.298±0.048•	2.645±0.070•	5.241±0.064•	7.388±0.026○	3.865±0.065•	6.882±0.055•
Adam-LDL-SCL	4.438±0.080•	3.433±0.242•	5.459±0.384•	6.293±1.178○	5.186±0.235•	7.545±0.418○
LDL-LDM	3.010±0.151•	3.019±0.139•	4.873±0.061•	6.892±0.045○	4.267±0.150•	7.273±0.071○
OFR-FL	2.296±0.099•	2.718±0.113•	4.321±0.111•	6.803±0.067○	3.905±0.098•	7.103±0.115○
OFR-CB	2.270±0.115•	2.732±0.096•	4.321±0.085•	6.804±0.062○	3.884±0.081•	7.097±0.106○
OFR-DB	1.837±0.034•	2.576±0.113•	4.056±0.051•	6.354±0.099○	3.760±0.066•	6.962±0.124○
RDA	1.549±0.036•	2.479±0.057•	3.893±0.063•	6.232±0.048○	3.496±0.041•	6.932±0.111○
DILDL	1.529±0.047•	2.447±0.076•	3.832±0.083•	6.146±0.0061•	3.387±0.053•	6.907±0.087•
AMEMA	**1.2893±0.0255**	**2.3679±0.1308**	**3.7940±0.0973**	5.8434±0.1102	**2.9876±0.0345**	**6.8014±0.1116**

Table 3: Experimental results on ILDL datasets measured by Canberra Distance (↓).

Method	Movie	SCUT-FBP	Emotion6	Flickr_LDL	RAF-ML	Natural Scene
SA-BFGS	1.164±0.024•	1.935±0.028•	2.227±0.027•	**1.914±0.064•**	2.030±0.014•	2.506±0.017•
EDL-LRL	1.186±0.030•	1.516±0.028•	1.900±0.024•	2.498±0.065○	1.746±0.034•	2.531±0.049•
LDLSF	1.271±0.027•	1.630±0.065•	1.935±0.034•	2.104±0.018○	1.707±0.035•	2.653±0.020•
LDL-LCLR	1.127±0.020•	1.373±0.024•	2.121±0.020•	2.658±0.005○	1.767±0.023•	2.449±0.015•
Adam-LDL-SCL	2.030±0.026•	1.688±0.104•	2.132±0.110•	2.353±0.258○	2.160±0.104•	2.667±0.111○
LDL-LDM	1.469±0.065•	1.518±0.056•	2.004±0.020•	2.537±0.013○	1.882±0.051•	2.557±0.017○
OFR-FL	1.113±0.044•	1.394±0.038•	1.826±0.033•	2.517±0.016○	1.758±0.036•	2.527±0.029○
OFR-CB	1.102±0.048•	1.398±0.032•	1.827±0.0028•	2.516±0.019○	1.749±0.030•	2.524±0.026○
OFR-DB	0.915±0.014•	1.350±0.039•	1.734±0.021•	2.409±0.028○	1.695±0.025•	2.493±0.030○
RDA	0.796±0.061•	1.320±0.024○	1.699±0.023•	2.384±0.013○	1.602±0.015•	2.486±0.026○
DILDL	0.768±0.031•	1.283±0.028•	1.676±0.042•	2.366±0.046•	1.571±0.024•	2.479±0.047•
AMEMA	**0.6892±0.0128**	**1.1150±0.0612**	**1.6683±0.0379**	2.3019±0.0119	**1.3502±0.0145**	**2.410±0.0264**

Table 4: Experimental results on ILDL datasets measured by Clark Distance (↓).

Table 4 further confirms AMEMA's superiority in terms of Clark distance. On the *Movie*, AMEMA achieves a Clark distance of 0.6892, lower than DILDL (0.768), RDA (0.796), and OFR-DB (0.915). On the *SCUT-FBP*, AMEMA achieves 1.1150, also lower than DILDL (1.283). On the *Emotion6*, AMEMA achieves 1.6683, compared to DILDL's 1.676. On the *Flickr_LDL*, AMEMA's value of 2.3019 remains highly competitive without sacrificing other metrics. On the *RAF-ML*, AMEMA achieves 1.3502, clearly better than DILDL (1.571).

The superior performance of AMEMA over previous methods is primarily due to its ability to dynamically adapt to label distribution imbalance during training. Unlike traditional approaches, which often rely on static optimization strategies or uniform update rules, AMEMA adjusts its optimization focus throughout the training process based on the learning dynamics of both dominant and non-dominant branches. This dynamic adaptation helps prevent overfitting to dominant branches while ensuring sufficient learning for non-dominant branches, which are often difficult to model in imbalanced label distributions. Additionally, AMEMA's unified optimization process enables the model to more effectively leverage the intrinsic structure of the label distribution, resulting in better convergence, improved robustness, and enhanced generalization compared to existing methods.

A.3.2 Supplementary Ablation Study

We further conducted ablation studies on *Flickr_LDL*, *Emotion6*, *SCUT-FBP*, and *RAF-ML* to evaluate the necessity of the WEIG and MOME modules (see Table 5 and Table 6). For example, on *Flickr_LDL*, removing both modules reduces Intersection similarity to 0.306, Cosine similarity to 0.388, and increases KL divergence to 1.607. Introducing either module alone brings some improvement, but only the combination of both modules achieves optimal performance, with Intersection similarity reaching 0.356, Cosine similarity reaching 0.451, and KL divergence dropping to 1.177. Similar trends are observed on *Emotion6* (Intersection 0.539, Cosine 0.679), *SCUT-FBP* (Intersection 0.740, Cosine 0.847), and *RAF-ML* (Intersection 0.549, Cosine 0.797).

The above results show that AMEMA consistently outperforms existing methods across all datasets and metrics, demonstrating its robustness in imbalanced label distribution learning. This superiority mainly stems from: (1) **EMA-based adaptive loss**

WEIG	MOME	Cheb ↓	Clark ↓	Can ↓	KL ↓	Cos ↑	Inter ↑
		0.521	2.384	6.232	1.607	0.388	0.306
✓		0.472	2.321	5.861	1.711	0.427	0.327
	✓	0.474	2.330	5.868	1.692	0.431	0.331
✓	✓	**0.453**	**2.302**	**5.843**	**1.177**	**0.451**	**0.356**

(a) *Flickr_LDL* dataset.

WEIG	MOME	Cheb ↓	Clark ↓	Can ↓	KL ↓	Cos ↑	Inter ↑
		0.360	1.699	3.893	0.768	0.622	0.512
✓		0.341	1.671	3.806	0.672	0.661	0.529
	✓	0.339	1.670	3.801	0.670	0.663	0.531
✓	✓	**0.336**	**1.668**	**3.794**	**0.661**	**0.679**	**0.539**

(b) *Emotion6* dataset.

Table 5: Ablation results of WEIG and MOME on the *Flickr_LDL* (a) and *Emotion6* (b) datasets.

WEIG	MOME	Cheb ↓	Clark ↓	Can ↓	KL ↓	Cos ↑	Inter ↑
		0.285	1.320	2.479	0.431	0.795	0.630
✓		0.226	1.225	2.381	0.337	0.812	0.721
	✓	0.214	1.212	2.380	0.351	0.831	0.728
✓	✓	**0.212**	**1.115**	**2.368**	**0.325**	**0.847**	**0.740**

(a) *SCUT-FBP* dataset.

WEIG	MOME	Cheb ↓	Clark ↓	Can ↓	KL ↓	Cos ↑	Inter ↑
		0.376	1.602	3.496	0.706	0.659	0.509
✓		0.351	1.392	3.018	0.602	0.762	0.531
	✓	0.355	1.401	3.087	0.593	0.771	0.537
✓	✓	**0.343**	**1.305**	**2.988**	**0.546**	**0.797**	**0.549**

(b) *RAF-ML* dataset.

Table 6: Ablation results of WEIG and MOME on the *SCUT-FBP* (a) and *RAF-ML* (b) datasets.

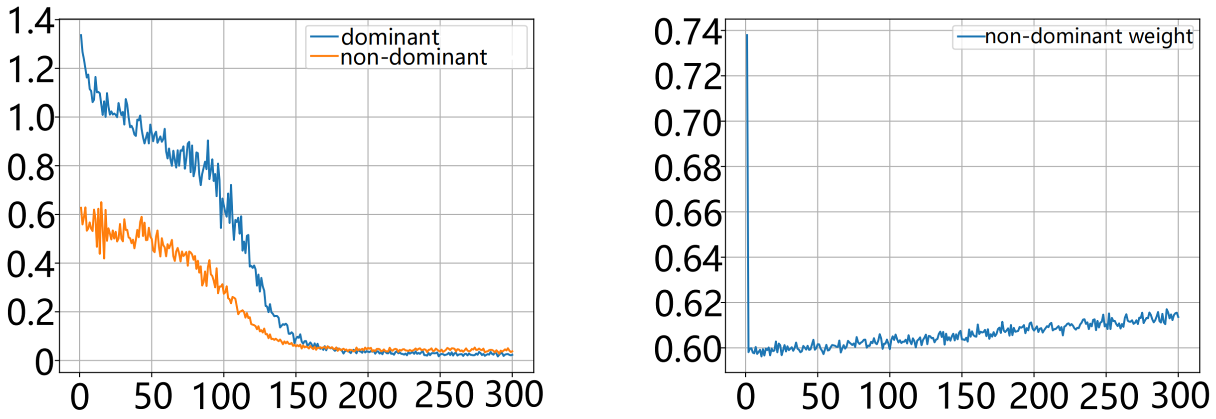
weighting, which dynamically balances the optimization between dominant and non-dominant branches and effectively alleviates the gradient vanishing problem in imbalanced label distribution learning; (2) **Branch-specific momentum allocation**, which improves the convergence stability of the non-dominant branch and reduces overfitting in the dominant branch. Ablation studies further demonstrate that both modules are indispensable for the overall performance improvements.

A.3.3 Visualization and Training Dynamics

Figures 1 to 5 visualize the training process. For instance, on the *Flickr_LDL* (Figure 1), the loss curves show that the dominant branch converges rapidly and stably, while the non-dominant branch converges more slowly, illustrating the advantage of decoupled optimization under imbalance. The weight variation plots show that the non-dominant branch weight increases rapidly in the early training phase and then stabilizes due to the EMA mechanism.

Momentum evolution, as shown in Figures 6 and 7, reveals that the non-dominant branch is assigned higher initial momentum, while the dominant branch's momentum decreases over time to prevent overfitting. This adaptive allocation ensures balanced optimization and robust learning from imbalanced data.

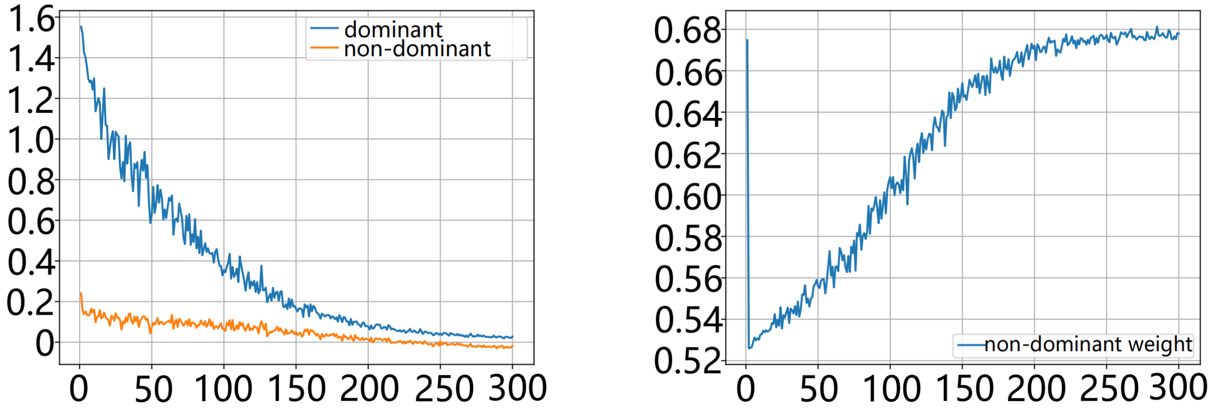
In summary, AMEMA consistently outperforms both traditional and state-of-the-art methods across all datasets and metrics, establishing a new benchmark for ILDL tasks. Ablation and visualization analyses further demonstrate the indispensable role of its innovative modules and adaptive mechanisms, showcasing the method's interpretability, robustness, and generalizability.



(a) Loss Evolution

(b) Weight Variation

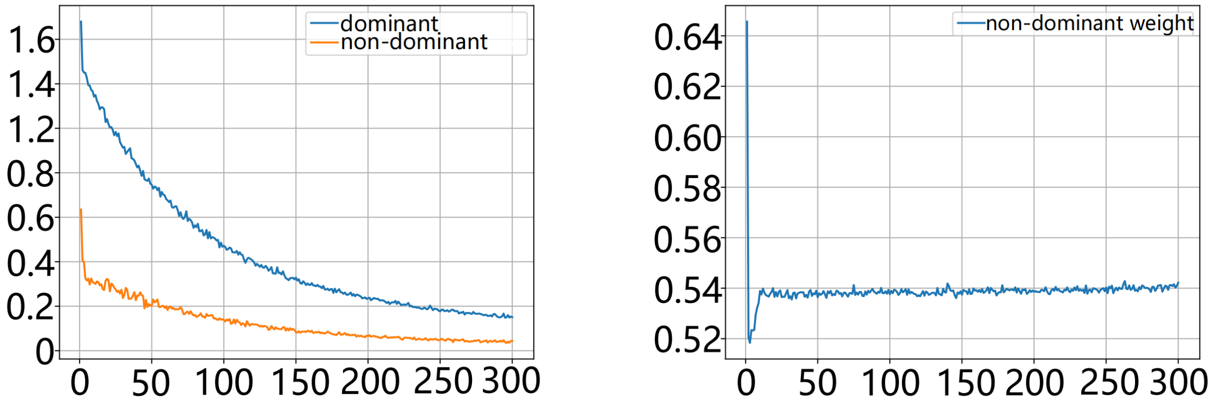
Figure 1: The curves of loss evolution for dominant and non-dominant branches during training (a), and weight variation for the non-dominant branch over time (b), on the imbalanced *Flickr_LDL* dataset.



(a) Loss Evolution

(b) Weight Variation

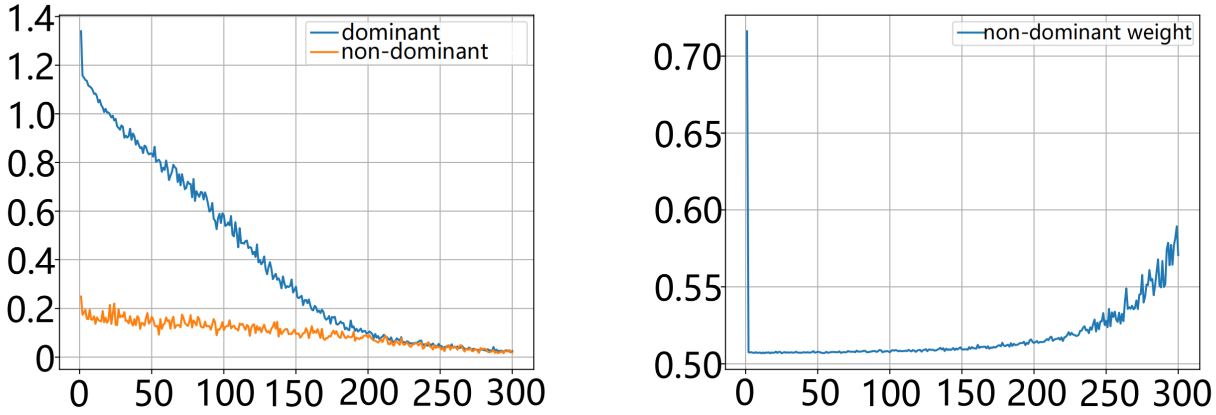
Figure 2: The curves of loss evolution for dominant and non-dominant branches during training (a), and weight variation for the non-dominant branch over time (b), on the imbalanced *Movie* dataset.



(a) Loss Evolution

(b) Weight Variation

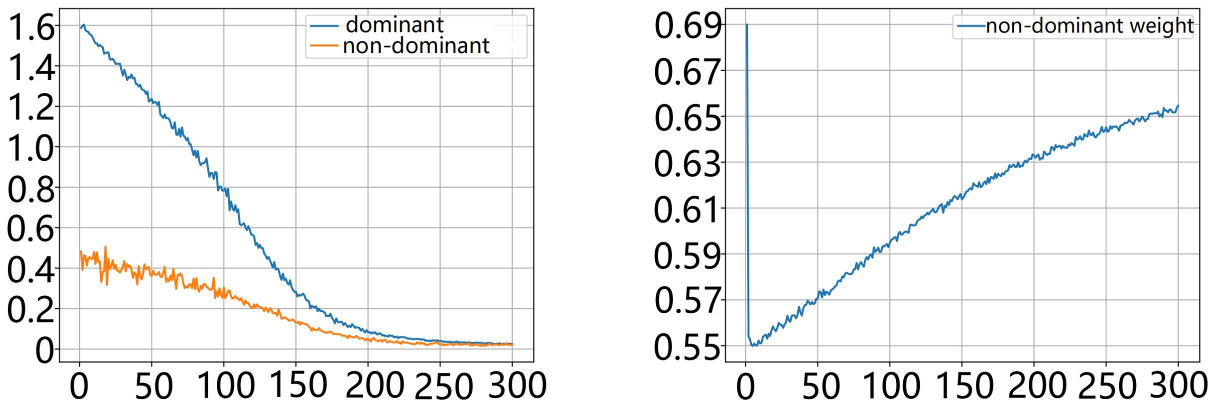
Figure 3: The curves of loss evolution for dominant and non-dominant branches during training (a), and weight variation for the non-dominant branch over time (b), on the imbalanced *SCUT-FBP* dataset.



(a) Loss Evolution

(b) Weight Variation

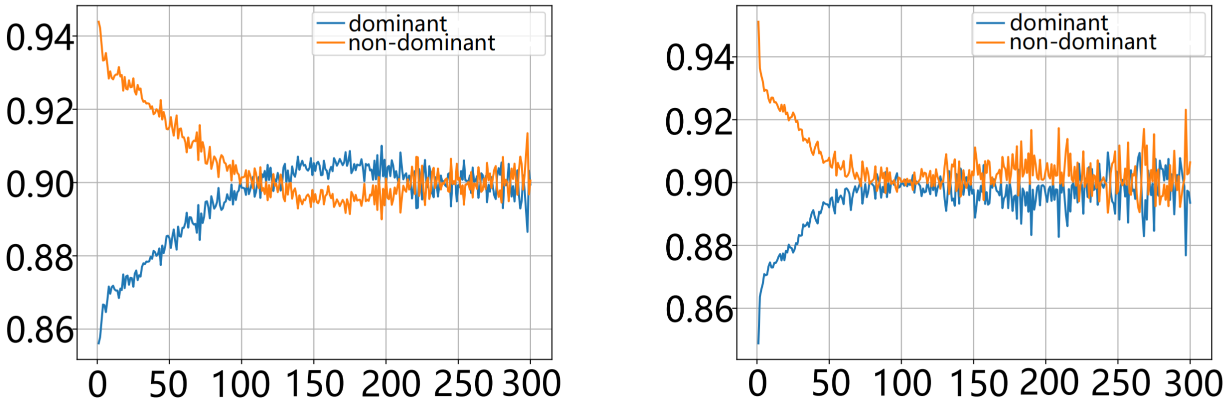
Figure 4: The curves of loss evolution for dominant and non-dominant branches during training (a), and weight variation for the non-dominant branch over time (b), on the imbalanced *Natural Scene* dataset.



(a) Loss Evolution

(b) Weight Variation

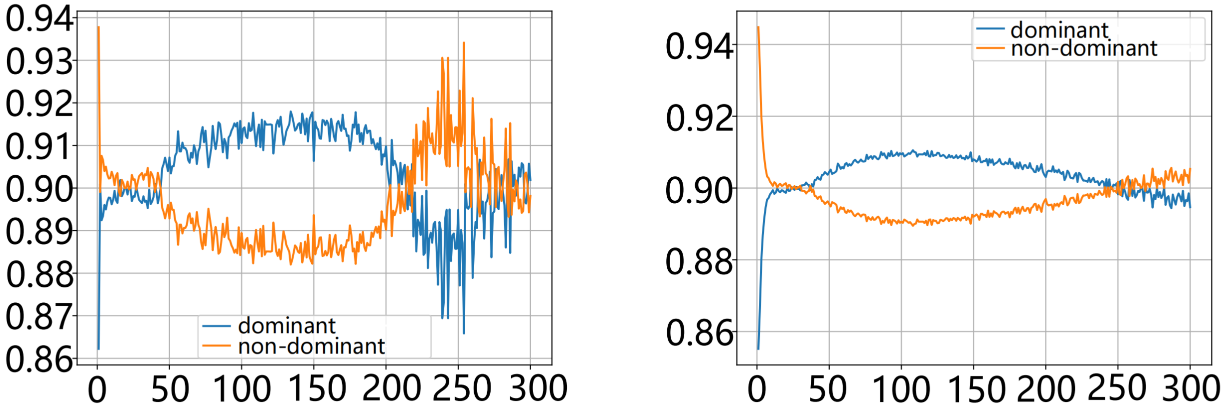
Figure 5: The curves of loss evolution for dominant and non-dominant branches during training (a), and weight variation for the non-dominant branch over time (b), on the imbalanced *RAF-ML* dataset.



(c) Momentum on *Emotion6*

(d) Momentum on *Flickr_LDL*

Figure 6: The curves of momentum evolution for the dominant and non-dominant branches during training on the *Emotion6* and *Flickr_LDL* datasets.



(c) Momentum on *RAF-ML*

(d) Momentum on *Nature Scene*

Figure 7: The curves of momentum evolution for the dominant and non-dominant branches during training on the *RAF-ML* and *Nature Scene* datasets.