Label What Matters: Modality-Balanced and Difficulty-Aware Multimodal Active Learning (Supplementary Material)

Yuqiao Zeng[1], Xu Wang[1], Tengfei Liang[1], Yiqing Hao[1], Yi Jin[1]*, Hui Yu[2]
[1]Key Laboratory of Big Data and Artificial Intelligence in Transportation, Ministry of Education;
State Key Laboratory of Advanced Rail Autonomous Operation;
School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China
[2]School of Psychology & Neuroscience, University of Glasgow, Glasgow, UK
{yuqiaozeng, xu.wang, tengfei.liang, yiqinghao, yjin}@bjtu.edu.cn
Hui.Yu@glasgow.ac.uk

A. Theoretical Analysis

Problem as an MDP. Let $\mathcal{L}_t$ and $\mathcal{U}_t$ be the labeled and unlabeled sets at round t. Given a fixed backbone architecture and a fixed training/evaluation protocol with a fixed validation split, define the operators

$$\textsc{Train}_E(\mathcal{L}) \quad \text{and} \quad \textsc{Eval}(\cdot) \mapsto \big(g, \phi, \bar{u}, \bar{d}, \rho\big),$$

where E is the number of epochs, g collects validation metrics (Top-1, NLL, ECE), $\phi \in \mathbb{R}^M$ are modality contributions (e.g., Top-1 gaps), $(\bar{u}, \bar{d})$ summarize uncertainty/diversity, and ρ are training diagnostics. We instantiate the MDP by

$$s_t = \big[g_t \,\|\, \phi_t \,\|\, \bar{u}_t \,\|\, \bar{d}_t \,\|\, \rho_t\big], \;\; a_t \subset \mathcal{U}_t, \;\; |a_t| = b,$$

and the transition

$$s_{t+1} = T(s_t, a_t) := \textsc{Eval}(\textsc{Train}_E(\mathcal{L}_t \cup a_t)), \quad (1)$$

with $\mathcal{L}_{t+1} = \mathcal{L}_t \cup a_t$ and $\mathcal{U}_{t+1} = \mathcal{U}_t \setminus a_t$. Because s_{t+1} is computed only from (s_t, a_t) via the fixed protocol (deterministic up to optimisation randomness), the process is Markov and the MDP is well-defined with finite horizon T.

Reward and shaping. Our per-round reward is

$$r_t = \text{Top-1}_t^{\text{RL–MBA}} - \frac{1}{|\mathcal{E}|} \sum_{h \in \mathcal{E}} \text{Top-1}_t^{(h)}, \quad (2)$$

where $\mathcal{E}$ denotes a fixed set of strong baselines. The baseline scores are *precomputed offline* under the same protocol and treated as constants during RL–MBA training; thus the shaping term is action-independent at each round. Therefore maximizing r_t is equivalent to maximizing Top-$1_t^{\text{RL–MBA}}$ up to a constant offset:

$$r_t = \text{Top-1}_t^{\text{RL–MBA}} - c_t.$$

Let the discounted return be $G_t = \sum_{k=0}^{T-t} \gamma^k r_{t+k}$. The policy gradient estimator with an action-independent baseline b^{bl} is

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \mathbb{E}\left[\sum_{t=1}^{T} \big(G_t - b^{\text{bl}}\big) \nabla_{\boldsymbol{\theta}} \log \pi_{\boldsymbol{\theta}}(a_t \mid s_t)\right]. \quad (3)$$

Lemma 1 (Unbiasedness and policy invariance). (i) Subtracting any action-independent baseline b^{bl} leaves $\nabla_{\boldsymbol{\theta}} J$ unbiased. (ii) Replacing the per-round reward by $r'_t = r_t - c_t$ with c_t independent of a_t leaves the optimal policy unchanged. *Sketch.* (i) $\mathbb{E}[b^{\text{bl}} \nabla \log \pi_{\boldsymbol{\theta}}(a_t \mid s_t)] = b^{\text{bl}} \cdot 0$. (ii) Adding constants per round only shifts G_t by an action-independent amount; the induced preference over actions is unchanged.

Dynamic feedback emerges from the transition. By (1), the next state packs *feedback* from the labels acquired at t: modality contributions ϕ_{t+1}, calibration/uncertainty summaries $\bar{u}_{t+1}$, and training diagnostics ρ_{t+1} are all functions of a_t through retraining and reevaluation. Hence the policy optimizes *long-term* returns by trading off immediate gains and how the choice reshapes future states.

A. Validity: Markov property and closed-loop coupling

Proposition 1 (Markovity). Under a fixed training/evaluation protocol (optimizer, epochs E, and a fixed validation split), the process $(s_t, a_t) \mapsto s_{t+1}$ defined in (1)

*Corresponding author.

is Markov. *Proof sketch.* Given $(\mathcal{L}_t, \mathcal{U}_t)$ encoded in s_t and a_t, the pair uniquely determines the training set $\mathcal{L}_t \cup a_t$ and thus the trained weights and validation statistics (in expectation over optimisation randomness).

Proposition 2 (Closed-loop AMCB/EFDA). AMCB produces weights $\boldsymbol{w}_t = \mathrm{softmax}(\boldsymbol{\Delta}_t/\tau)$ from ϕ_t (e.g., Top-1 gaps), and EFDA computes fused evidential uncertainty $\mathsf{U}_t(x)$ from $\boldsymbol{\alpha}_f$ parameterized by $\boldsymbol{w}_t$. Both $\boldsymbol{w}_t$ and summary statistics $(\bar{u}_t, \bar{d}_t)$ appear in s_t and affect scoring/selection; after retraining, they update to $(\boldsymbol{w}_{t+1}, \bar{u}_{t+1}, \bar{d}_{t+1})$ through T. Thus AMCB/EFDA are endogenously coupled to the policy via s_t, forming a closed loop.

B. Safety: reward shaping and variance reduction

Proposition 3 (Safe shaping). Using (2) is equivalent to using absolute accuracy reward up to a per-round constant; hence the optimal policy is preserved. Together with Lemma 1(i), employing a moving-average baseline b^{bl} reduces estimator variance without bias.

C. Adaptivity: policy-gradient alignment reduces imbalance

We quantify imbalance by a dispersion functional $D(\phi)$. For $M{=}2$, $D(\phi) = |\phi_1 - \phi_2|$; for $M{>}2$,

$$D(\phi) = \tfrac{1}{M} \sum_{m=1}^{M} \big(\phi_m - \bar{\phi}\big)^2, \qquad \bar{\phi} = \tfrac{1}{M} \sum_{m} \phi_m. \tag{4}$$

We consider the standard smoothness setting for policy optimization.

Assumption 1 (Smooth landscape). $J(\boldsymbol{\theta})$ is β-smooth and $\|\nabla J(\boldsymbol{\theta})\| \le G$ within the region visited by training.

Assumption 2 (Alignment). There exists $\varepsilon > 0$ such that

$$\big\langle \nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}} \big(- D(\phi(\boldsymbol{\theta})) \big) \big\rangle \ge \varepsilon \tag{5}$$

along the training trajectory.

Lemma 2 (Monotone decrease under small steps). With step size $\eta > 0$,

$$\begin{aligned} D(\phi(\boldsymbol{\theta}{+}\eta\nabla J)) - D(\phi(\boldsymbol{\theta})) &\approx \eta \,\langle \nabla D, \nabla J \rangle + O(\eta^2) \\ &\le -\eta\,\varepsilon + O(\eta^2). \end{aligned} \tag{6}$$

Theorem 1 (Dynamic feedback adaptation). Under Assumptions 1–2 and sufficiently small η, the policy-gradient update increases the expected return and reduces $D(\phi)$ up to $O(\eta^2)$, thereby adapting modality emphasis across rounds through the closed-loop transition T.

D. Why the reward supports long-horizon adaptivity

Let $r_t = \Delta\text{Top-1}_t - c_t$ with c_t independent of a_t. Then

$$G_t = \sum_{k=0}^{T-t} \gamma^k \, \Delta\text{Top-1}_{t+k} - \sum_{k=0}^{T-t} \gamma^k \, c_{t+k}. \tag{7}$$

The second summation is action-independent, while the first accumulates future accuracy gains induced by current choices through T.

E. Auxiliary regularization and alignment

To stabilise the alignment in (5), we use two auxiliary losses:

$$\mathcal{L}_{\text{modality}} = \tfrac{1}{M} \sum_{m=1}^{M} \big(\phi_{t,m} - \bar{\phi}_t\big)^2, \tag{8}$$

$$\mathcal{L}_{\text{difficulty}} = \tfrac{1}{|a_t|} \sum_{x \in a_t} \big[\tau_u^t - \mathsf{U}(x)\big]_+, \tag{9}$$

where τ_u^t is the q-quantile of fused evidential uncertainty within round t and $[z]_+ = \max(0, z)$.

F. Modality contributions in practice

In the main paper we use two types of modality contribution signals: (i) *Top-1 gaps* Δ_m between modality-m heads and the multimodal head, which directly drive AMCB; and (ii) *Shapley-style* contributions ϕ_m used only for analysis/visualisation. Shapley-style scores are computed by Monte Carlo sampling over modality coalitions on the validation split and are never used during training.

B. Additional Experimental Details

We follow the same datasets, backbones, optimisers, and active-learning protocol as in the main paper (Sec. 4). This supplementary provides implementation details omitted due to space.

State construction. The policy state s_t concatenates: (i) validation statistics g_t (Top-1, NLL, ECE), each normalised to $[0, 1]$; (ii) modality contributions ϕ_t (Top-1 gaps or their normalised variants); (iii) round-wise averages of fused uncertainty and diversity $(\bar{u}_t, \bar{d}_t)$; and (iv) training diagnostics ρ_t such as moving averages of training loss and gradient norm. All features are standardised across rounds using running means and variances.

Evidential heads and EFDA in practice. Each modality-specific head outputs non-negative evidence $e_m(x) \in \mathbb{R}^C_{\ge 0}$ via a `softplus` transformation. Dirichlet parameters are constructed as $\boldsymbol{\alpha}_m(x) = e_m(x) + \mathbf{1}$. Before fusion, we apply temperature scaling on a held-out validation split (one scalar per modality) to improve calibration, and use the calibrated evidences in EFDA to compute fused uncertainty scores $\mathsf{U}(x)$.

Policy network and optimisation. The policy is a lightweight two-layer MLP with ReLU activations. We train it with REINFORCE using discount $\gamma{=}0.9$ and a moving-average baseline (smoothing coefficient 0.1). The

policy optimiser is AdamW with weight decay 10^{-4} and global gradient clipping. Unless otherwise stated, we use α=0.1, λ_1=0.5, and λ_2=0.5 in the total loss.

References

[1] Jordan T Ash, Chicheng Zhang, Akshay Krishnamurthy, John Langford, and Alekh Agarwal. Deep batch active learning by diverse, uncertain gradient lower bounds. *arXiv preprint arXiv:1906.03671*, 2019.

[2] Razvan Caramalau, Binod Bhattarai, and Tae-Kyun Kim. Sequential graph convolutional network for active learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9583–9592, 2021.

[3] Melanie Ducoffe and Frederic Precioso. Adversarial active learning for deep networks: a margin based approach. *arXiv preprint arXiv:1802.09841*, 2018.

[4] Yarin Gal, Riashat Islam, and Zoubin Ghahramani. Deep bayesian active learning with image data. In *International conference on machine learning*, pages 1183–1192. PMLR, 2017.

[5] Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach. *arXiv preprint arXiv:1708.00489*, 2017.

[6] Burr Settles. From theories to queries: Active learning in practice. In *Active learning and experimental design workshop in conjunction with AISTATS 2010*, pages 1–18. JMLR Workshop and Conference Proceedings, 2011.

[7] Meng Shen, Yizheng Huang, Jianxiong Yin, Heqing Zou, Deepu Rajan, and Simon See. Towards balanced active learning for multimodal classification. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3434–3445, 2023.