NEURAL INFORMATION PROCESSING SYSTEMS

北京理工大學(珠海)
BEIJING INSTITUTE OF TECHNOLOGY , ZHUHAI

Language-Bias-Resilient Visual Question Answering via Adaptive Multi-Margin Collaborative Debiasing

Biomedical & AI Group，BAI

Huanjia Zhu
Beijing Institute of Technology, Zhuhai
bvyih3@gmail.com

Shuyuan Zheng
The University of Osaka
zheng@ist.osaka-u.ac.jp

Yishu Liu
Harbin Institute of Technology, Shenzhen
liuyishu@stu.hit.edu.cn

Sudong Cai*
Beijing Institute of Technology, Zhuhai
caisudong.ai@gmail.com

Bingzhi Chen*
Beijing Institute of Technology, Zhuhai
chenbingzhi@bit.edu.cn

目 录

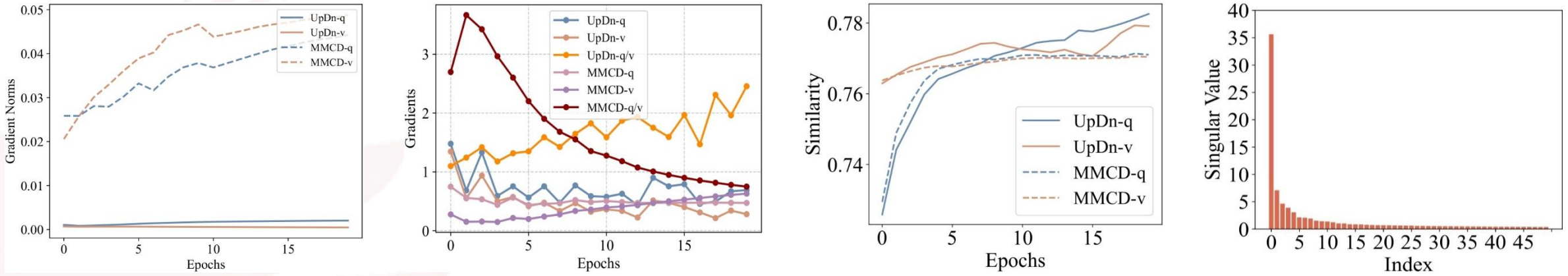
CONTENTS

Introduction-Background

- Visual Question Answering (VQA) has emerged as a challenging task that blends computer vision and natural language processing to provide answers to natural language questions about images.

- Networks still suffer from language bias, where the model learns spurious correlations between questions and answers. This bias occurs when models overly rely on common patterns in questions and answers, neglecting crucial visual information.

Introduction-Challenges

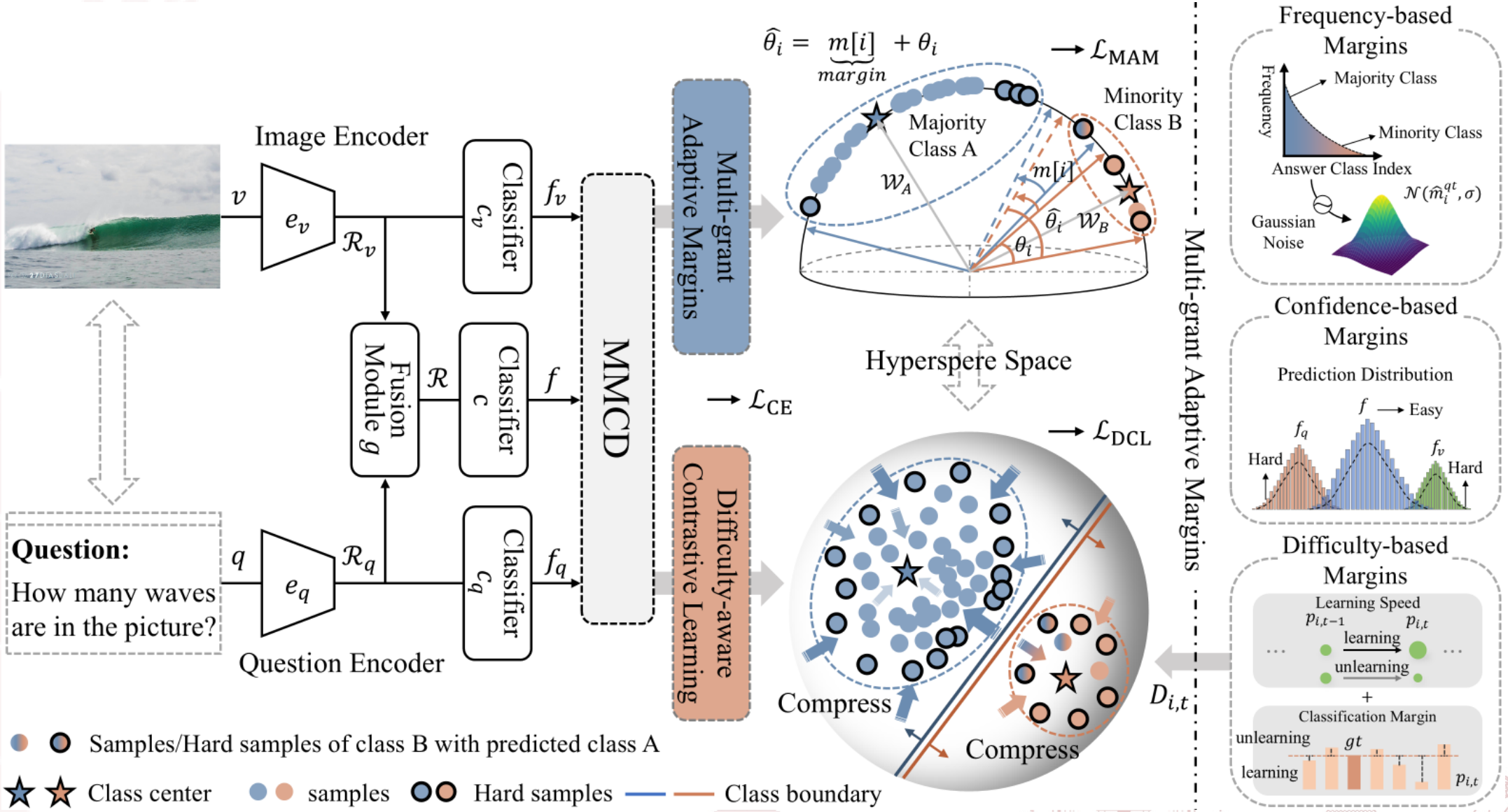


(a) (b) (c) (d)

(a) The gradient norms across modalities differ considerably, reflecting an imbalance in how learning signals are propagated. (b) This imbalance leads to modality-specific optimization deviations, where the question modality of baseline disproportionately accumulates gradient updates, thereby amplifying its influence. (c) As a result, the fused representation becomes skewed, with question features occupying a dominant share of the multimodal space and suppressing contributions from visual features. (d) Furthermore, the classifier weights reveal directional bias: the singular value spectrum is highly uneven, suggesting that the model primarily aligns with directions that capture biased cues while overlooking secondary directions that encode meaningful information.

In this work, we conduct a comprehensive investigation into bias formation. First, we identify a modality gradient optimization deviation (see Fig.1(b)), where the image modality is under-optimized and the question modality is over-optimized. Second, we observe a feature fusion component deviation (see Fig.1(c)), in which question features dominate the joint representation and image features are marginalized. These phenomena culminate in directional deviation of the classifier weights (see Fig.1(d)), amplifying primary (question-driven) directions and attenuating secondary (vision-driven) axes.

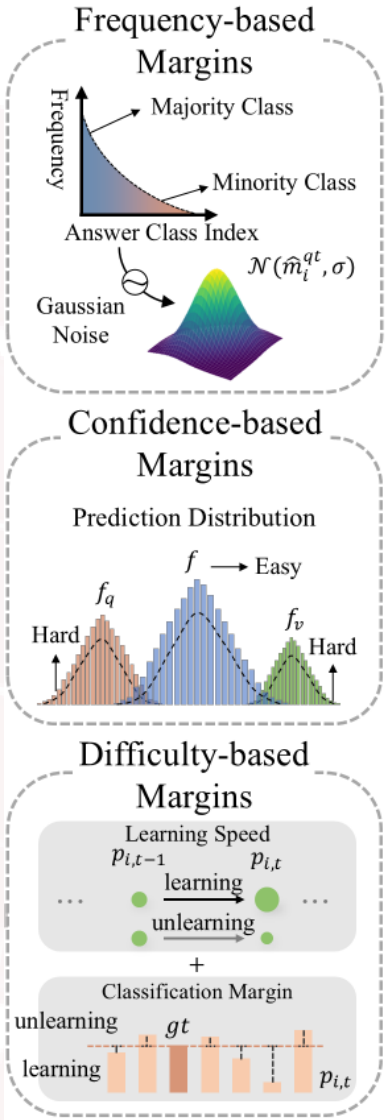
Methodology

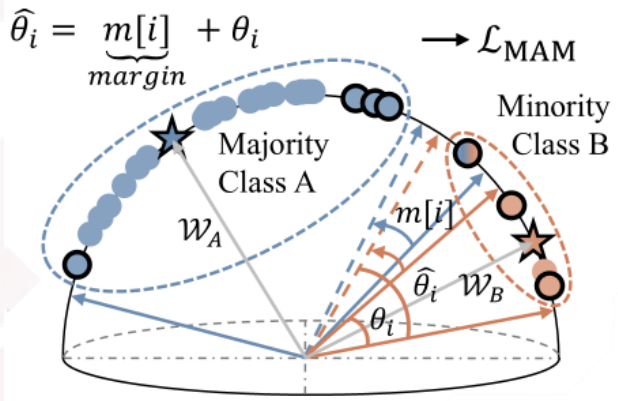


Methodology

NEURAL INFORMATION PROCESSING SYSTEMS Biomedical & AI Group

1.Multi-Grained Adaptive Margins

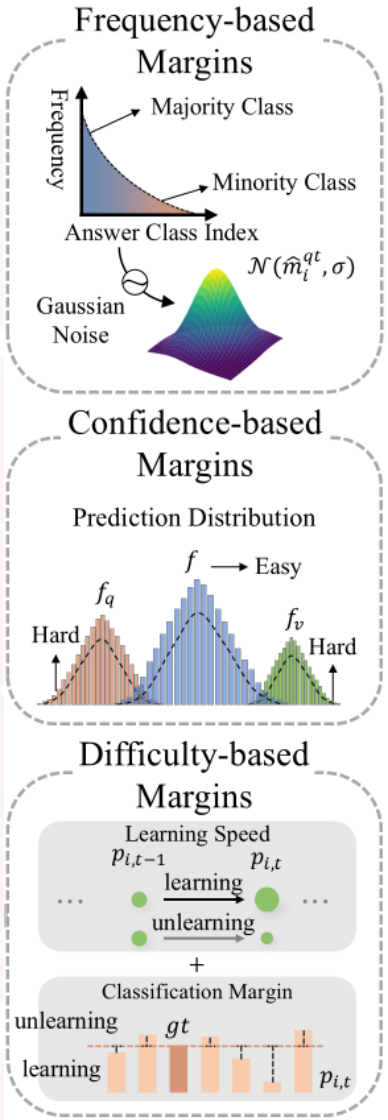


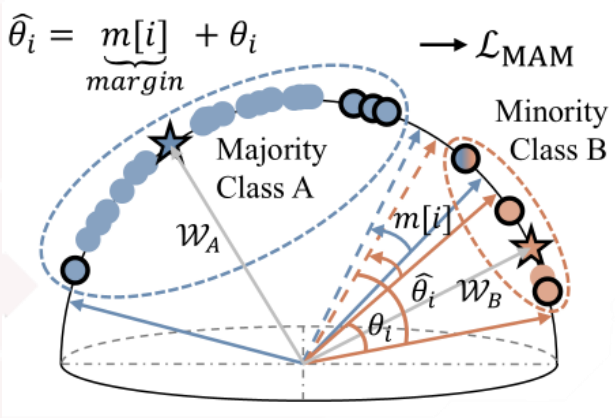


- Our MAM mechanism aims to address the challenge of chaotic class boundaries posed by imbalanced data.
- By considering answer frequency and evaluating instance difficulty from coarse-grained and fine-grained perspectives, MAM enhances intra-class compactness and inter-class separation, thus refining a discriminative and robust feature space.
- Specifically, MAM integrates three components: frequency-aware, confidence-aware, and difficulty-aware margins.

Methodology

1.Multi-Grained Adaptive Margins





Frequency-aware margins impose larger margin penalties on minority classes for driving their representations closer to the respective class centers and smaller margin penalties on majority classes.

$$\hat{m}_i^{qt} = \frac{n_i^{qt} + \epsilon}{\sum_{j=1}^{|\mathcal{A}|} n_j^{qt} + \epsilon},$$

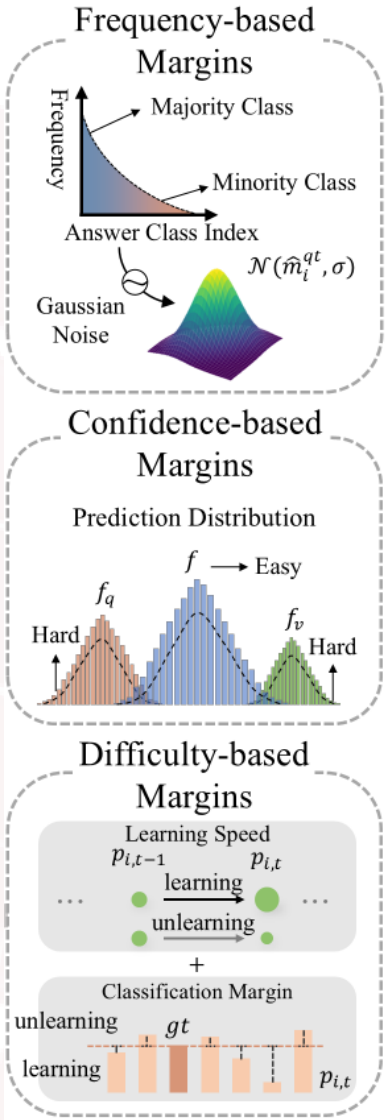
Fixed margins fail to adapt to the dynamic inter- and intra-class variances in real-world data, thereby impairing the model's discriminability and generalization. We incorporate random Gaussian noise into frequency-aware margins

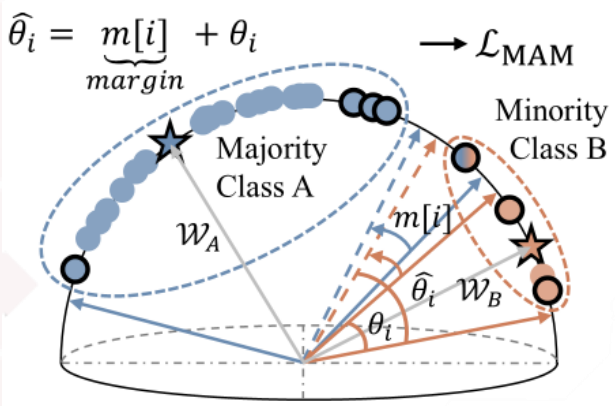
$$m_i^{qt} = \mathcal{N}(\hat{m}_i^{qt}, \sigma),$$

Methodology

NEURAL INFORMATION PROCESSING SYSTEMS Biomedical & AI Group

1.Multi-Grained Adaptive Margins





Sample difficulty significantly affects the discriminative decision margin and class separability. A simple yet effective measure of difficulty is the prediction logits. We incorporate auxiliary branches dedicated solely to the question and image modalities. These branches promote multimodal integration by deliberately introducing controlled modality bias. This strategy not only boosts ID performance but also prevents excessive bias correction.

We introduce a question-only branch and an image-only branch:

$$f_q(q) = c_q(e_q(q)), \quad f_v(v) = c_v(e_v(v)).$$

Recognizing the challenges of imbalanced multimodal learning, we leverage the posterior distributions as the weights for unimodal logits:

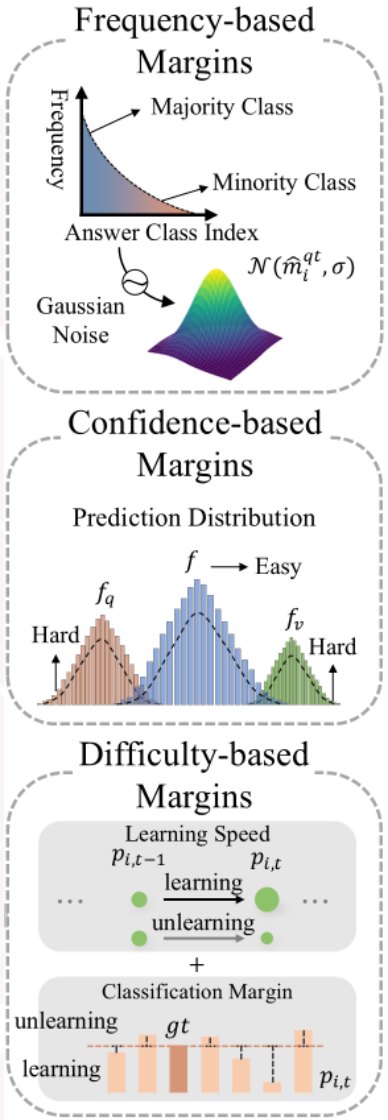
$$s_q = \text{softmax}(\mathcal{W}_q \cdot e_q(q) + \frac{b}{2})[gt], \quad s_v = \text{softmax}(\mathcal{W}_v \cdot e_v(v) + \frac{b}{2})[gt],$$

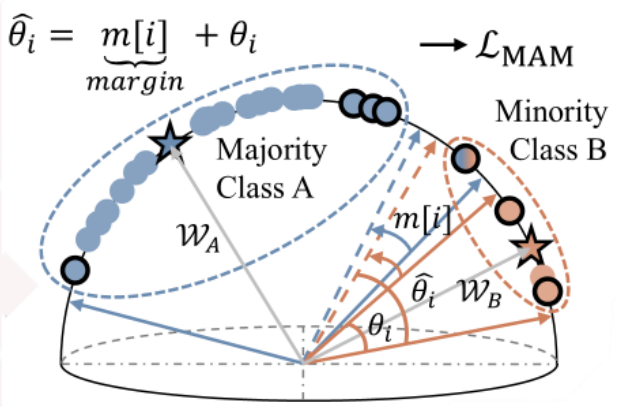
The weighted hybrid confidence and confidence-aware margins are formulated as follows:

$$f_m = \frac{s_f \cdot f + s_q \cdot f_q + s_v \cdot f_v}{s_f + s_q + s_v}, \quad m_{conf} = \text{softmax}(f_m/\tau_1),$$

Methodology

1.Multi-Grained Adaptive Margins



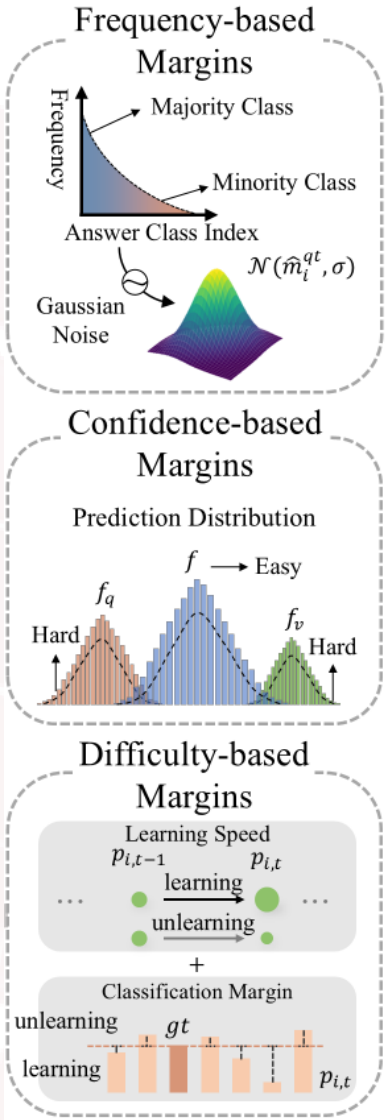


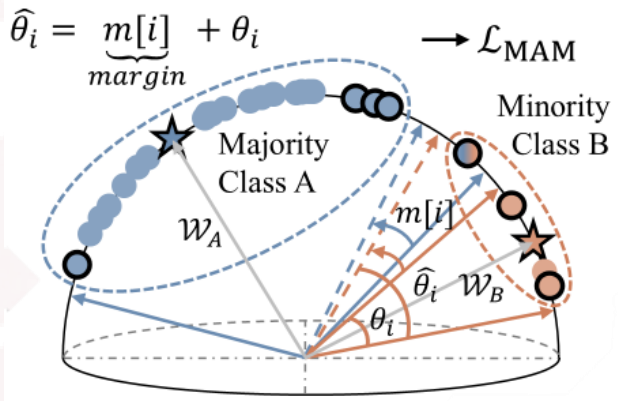
Although logits can simply and intuitively reflect sample difficulty, their static nature limits the fine-grained mining of intrinsic sample difficulty and makes it difficult to modulate stubborn decision boundaries. We develop a fine-grained difficulty model that evaluates instance difficulty from two perspectives: a) Learning Rate: akin to human learning, where easy samples are learned quickly, and b) Classification Margins: reflecting relative confidence, where smaller margins indicate closer proximity to the decision boundary. Specifically, given an instance representation, its difficulty after t iterations is estimated as:

$$D_{i,t} = \underbrace{\alpha \cdot \frac{vu_{i,t} + c}{vl_{i,t} + c}}_{\text{learning speed}} + \underbrace{(1-\alpha) \cdot \frac{mu_{i,t} + c}{ml_{i,t} + c}}_{\text{classification margins}},$$

Methodology

1.Multi-Grained Adaptive Margins





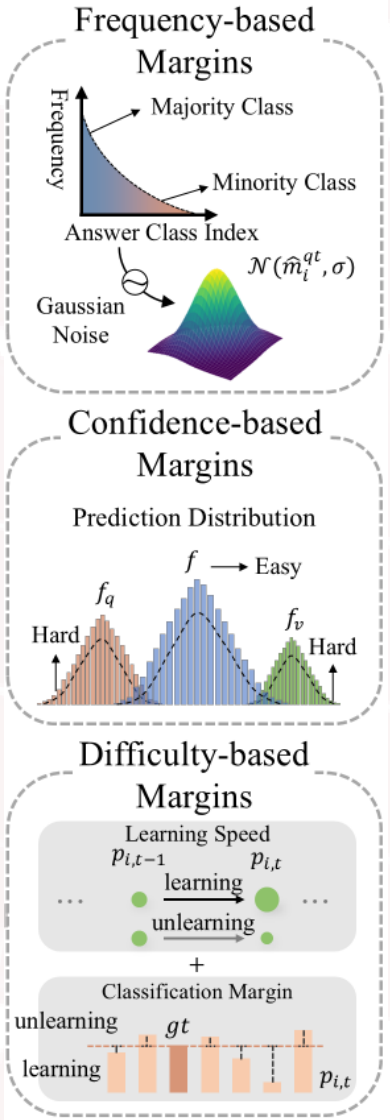
The quantification method for learning rate is as follows:

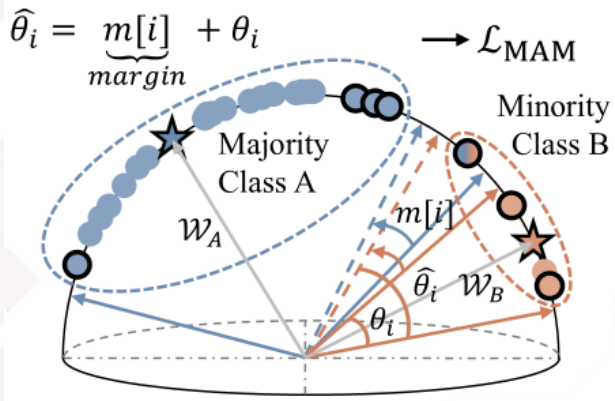
$$v_{i,t} = \frac{1}{2}\mathrm{KL}(p_{i,t-1} \| q_{i,t}) + \frac{1}{2}\mathrm{KL}(p_{i,t} \| q_{i,t}),$$

$$\begin{cases} vu_{i,t} = \beta \cdot vu_{i,t-1} + (1-\beta) \cdot vu'_{i,t}, \quad vl_{i,t} = \beta \cdot vl_{i,t-1} + (1-\beta) \cdot vl'_{i,t}, \\ vu'_{i,t} = \min(p^{gt}_{i,t} - p^{gt}_{i,t-1}, 0) v_{i,t}[gt] + \sum_{j=1, j\neq gt}^{C} \max(p^{j}_{i,t} - p^{j}_{i,t-1}, 0) v_{i,t}[j], \\ vl'_{i,t} = \max(p^{gt}_{i,t} - p^{gt}_{i,t-1}, 0) v_{i,t}[gt] + \sum_{j=1, j\neq gt}^{C} \min(p^{j}_{i,t} - p^{j}_{i,t-1}, 0) v_{i,t}[j], \end{cases}$$

Methodology

1.Multi-Grained Adaptive Margins





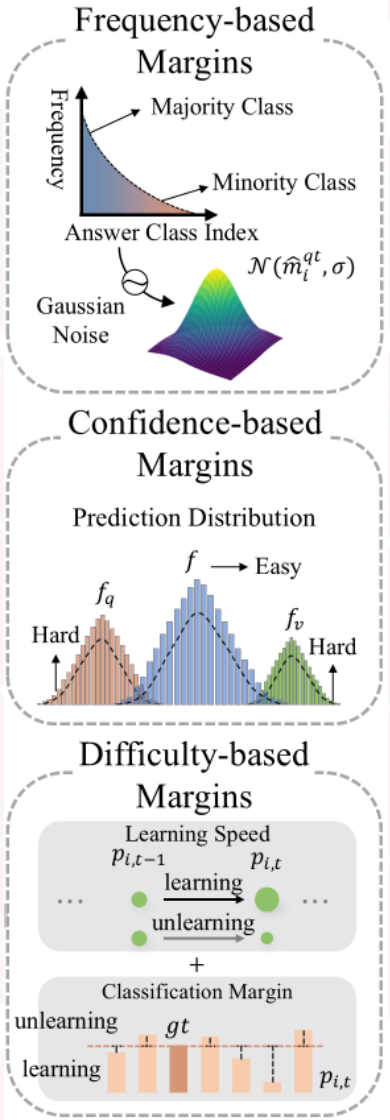
The quantification method for classification margins is as follows:

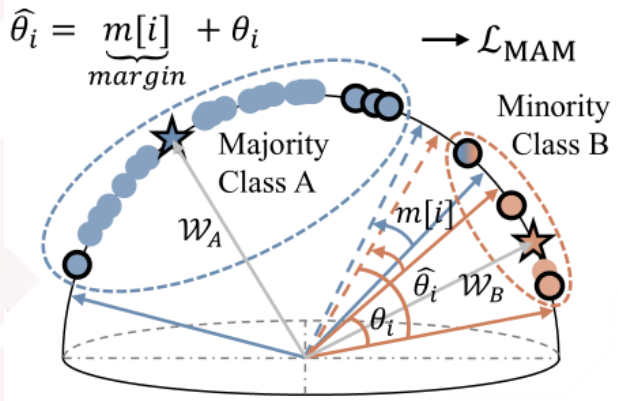
$$m_{i,t} = |p_{i,t}^{gt} - p_{i,t}^{j}|, \quad j = 1 \ldots C, j \neq gt,$$

$$\begin{cases} mu_{i,t} = \beta \cdot mu_{i,t-1} + (1-\beta) \cdot mu'_{i,t}, & ml_{i,t} = \beta \cdot ml_{i,t-1} + (1-\beta) \cdot ml'_{i,t}, \\ mu'_{i,t} = \log(\frac{1}{|\Psi|} \sum_{j \in \Psi} \exp(m_{i,t}^{j})), & ml'_{i,t} = \log(\frac{1}{|\Omega|} \sum_{j \in \Omega} \exp(m_{i,t}^{j})), \end{cases}$$

Methodology

1.Multi-Grained Adaptive Margins



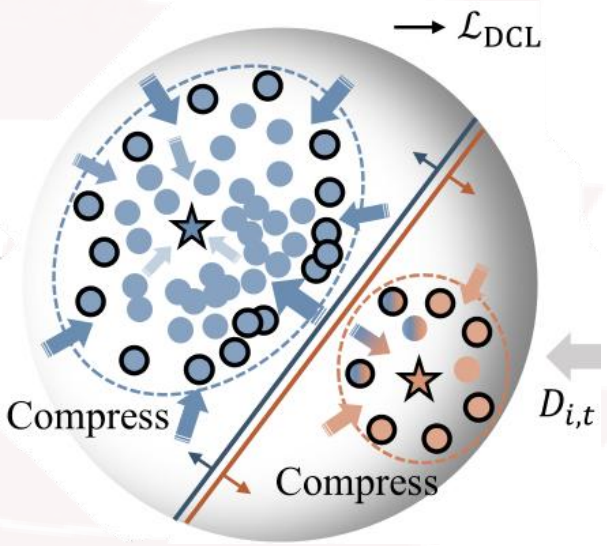


Ultimately, the various margin terms are aggregated in a cohesive manner to form the multi-grained adaptive margins:

$$
\begin{cases}
m_{\text{MAM}} = m_{freq}, \\
m_{\text{MAM}}[gt] = (1-\lambda_1)\cdot m_{\text{MAM}}[gt] + \lambda_1 \cdot m_{conf}[gt], \\
m_{\text{MAM}}[gt] = (1-\lambda_2)\cdot m_{\text{MAM}}[gt] + \lambda_2 \cdot m_{diff}, \quad \text{epoch} \geq w, \\
m_{\text{MAM}} = 1 - m_{\text{MAM}},
\end{cases}
$$

Methodology

2.Difficulty-aware Contrastive Learning



We further propose the DCL mechanism that integrates our instance difficulty model into a supervised contrastive paradigm, which dynamically emphasizes hard samples by difficulty-adaptive weighting, effectively enhancing intra-class compactness and inter-class separation to form a discriminative feature space.

$$\mathcal{L}_{\mathrm{DCL}} = \sum_{j \in \mathcal{B}} \frac{-1}{|P_j|} \sum_{p \in P_j} \log \frac{\exp(D_{p,t}) \exp(x_j^\top x_p / \tau_2)}{\sum_{n \in N_j} \exp(D_{n,t}) \exp(x_j^\top x_n / \tau_2)},$$

Experiments

Table 1: Accuracy comparisons with other methods on the VQA-CP v2 and VQA-CP v1 datasets.

Datasets		VQA-CP v2				VQA-CP v1			
Methods		All	Y/N	Num	Others	All	Y/N	Num	Others
UpDn [2]	CVPR'18	39.74	42.27	11.93	46.05	37.96	42.79	12.41	42.53
RUBi [7]	NeurIPS'19	47.11	68.65	20.28	43.18	-	-	-	-
LMH [13]	EMNLP'19	52.15	70.29	44.10	44.86	55.73	78.59	24.68	45.47
GGE-iter [18]	ICCV'21	57.12	87.35	26.16	49.77	59.82	85.52	28.93	46.67
AdaVQA [17]	IJCAI'21	54.02	70.83	49.00	46.29	61.20	**91.17**	41.34	39.38
COB [21]	WACV'23	57.53	88.36	28.81	49.27	60.98	87.41	32.02	46.34
GENB [11]	CVPR'23	59.15	88.03	40.05	49.25	62.74	86.18	43.85	**47.03**
GGD [19]	TPAMI'23	59.37	88.23	38.11	49.82	-	-	-	-
CVIV [33]	TMM'24	60.08	88.85	40.77	**50.30**	-	-	-	-
PWVQA [41]	TMM'24	59.06	88.26	52.89	45.45	-	-	-	-
MMCD	Ours	**61.34**	**88.93**	**55.68**	48.44	**63.62**	90.72	**52.67**	41.08

Experiments

Table 2: Performance of our approach with different network architectures

Methods	All	Y/N	Num	Other	Increased ↑
SAN	26.88	35.34	11.34	24.70	33.24
SAN+MMCD	**60.12**	**86.97**	**54.62**	**47.56**	
S-MRL	38.46	42.85	12.81	43.20	22.23
S-MRL+MMCD	**60.69**	**88.40**	**55.44**	**47.61**	
LXMERT	48.66	47.49	22.24	**56.52**	18.29
LXMERT+MMCD	**66.95**	**91.79**	**63.28**	54.39	

We further evaluate the generalizability and robustness of MMCD across additional architectures, including SAN, S-MRL, and LXMERT. As shown in Table 2, the MMCD approach consistently outperforms the corresponding baselines, demonstrating strong adaptability and model-agnostic performance across diverse network designs. In particular, applying MMCD to LXMERT, a widely adopted vision-language pre-trained model commonly used in various multimodal downstream tasks, yields a notable 18.29% performance improvement, further highlighting its effectiveness in enhancing a broad range of model families.

Conclusion

- In this paper, we investigated the origin of language bias in VQA and elucidated why margin-based mechanisms effectively mitigate it.
- Building on these insights, we propose MMCD, an adaptive multi-margin framework that incorporates sample frequency and difficulty to reshape decision boundaries and enhance feature discrimination via difficulty-aware contrastive learning.
- Extensive experiments confirm the superior robustness of MMCD, with potential implications for broader challenges such as shortcut learning, long-tail recognition, and class imbalance.

Thank you for your attention!