Supplementary Material for "Boosting Multi-modal Model Performance with Adaptive Gradient Modulation"

A. Experiment Details

A.1. Datasets

AV-MNIST [9]. The dataset is collected for multi-media classification tasks by assembling visual and audio features. The first modality, disturbed image, is made of the 28×28 PCA-projected MNIST images. The second modality, audio, is made of audio samples on 112×122 spectrograms. The whole dataset includes $70,000$ samples, and the division of the training set and validation set is $6/1$. We randomly selected 10% samples from the training set and validation set to create a development set.

UR-Funny [3]. The dataset is created for affective computing tasks that detect humor by the usage of words (text), gestures(vision) and prosodic cues (acoustic). This dataset is collected from the TED talks and uses an equal number of binary labels for each sample. In the experiments, the split of the dataset follows[5].

CREMA-D [1]. The dataset is devised for speech emotion recognition with facial and vocal emotional expressions. This dataset contains 6 most usual emotions: angry, happy, sad, neutral, discarding, disgust, and fear. The whole dataset is randomly divided into $6,027$-sample training set and 669-sample validation set, as well as 745-sample testing set.

AVE [8]. The dataset is an $Audio$-$Visual\ Event$ (AVE) dataset for audio-visual event localization. This dataset consists of $4,143$ ten-second video clips and has 28 event classes for each clip together with frame-level annotations. All videos are collected from YouTube. In the experiments, we follow [8] in splitting and pre-processing the dataset.

CMU-MOSEI [10]. This dataset is collected for sentence-level sentiment analysis and emotion recognition, containing $23,454$ movie review clips with more than 65.9 hours of YouTube video by $1,000$ speakers. In our experiments, we only use text and audio modalities, and the train/valid/test set is split into $16,327/1,871/4,662$ samples, respectively.

Kinetics-Sound [4]. The dataset is a multi-modal dataset for human action recognition in videos. The original dataset contains 400 human action classes with at least 400 video clips for each class. In our experiments, we randomly select 30 classes, of which the number of classes is close to OGM-GE [6]. This dataset contains 25956 video clips (21545 training, 1494 validation, 2917 test).

A.2. Implementation details

For the AV-MNIST dataset, we use ResNet18-based networks as the audio and visual encoders. Following [2], we reduce the number of input channels from 3 to 1. For the UR-Funny dataset, we use a 4-layer Transformer as the encoder for each modality. The number of attention heads is 8 and the hidden dimension is 768. In the experiments on the above two datasets, models are trained using the the SGD optimizer with a 0.9 momentum and a 1e-4 weight decay. The initial learning rate is 1e-4, and it decays with a rate of 0.9 every 70 epochs. The batch size is set to 64.

For the CREMA-D and Kinetics-Sound dataset, we follow the experimental settings used in OGM-GE [6], except for the CREMA-D decay rate in the learning rate scheduler. This decay rate is now set to 0.9 to make our training more stable.

For the AVE and CMU-MOSEI datasets, we adopt the same experimental settings in [11] and [10], respectively.

The linear predictor in Section 3.2.2 is implemented with the `sklearn` package. Specifically, we use ridge regression with the regularization strength $\lambda = 120$ for all the situations. The value of λ is chosen so that the competition strength converges on the validation sets across all the datasets.

In all the experiments in the main text, the random seed is set to 999 for reproducibility.

B. Sanity Check

In this section, we justify the definition of the proposed competition strength metric. As linear probing is a standard

AV-MNIST			Acc	Acc_a	Acc_v	d^a	d^v
Late fusion	seed=99	$\mathcal{C}^a$	-	39.73	-	-	-
		$\mathcal{C}^v$	-	-	65.30	-	-
		Joint-Train	69.77	16.05	55.83	0.7903	0.1341
	seed=999	$\mathcal{C}^a$	-	39.61	-	-	-
		$\mathcal{C}^v$	-	-	65.14	-	-
		Joint-Train	69.77	16.05	55.83	0.7838	0.1408
Early fusion	seed=99	$\mathcal{C}^a$	-	41.21	-	-	-
		$\mathcal{C}^v$	-	-	65.27	-	-
		Joint-Train	71.15	24.28	60.14	0.7219	0.1999
	seed=999	$\mathcal{C}^a$	-	41.60	-	-	-
		$\mathcal{C}^v$	-	-	65.46	-	-
		Joint-Train	71.15	24.28	60.14	0.7668	0.1825

Table 5. Comparing the effect of differently randomly initialized mono-modal concepts on competition strength in the AV-MNIST dataset joint-training. $seed$ is the random seed we set in our experiments. $\mathcal{C}^a$ and $\mathcal{C}^v$ indicate the performance of audio and visual modality concepts, respectively.

AV-MNIST		Acc	Acc_a	Acc_v	d^a	d^v
zero-pad	$\mathcal{C}^a$	-	41.60	-	-	-
	$\mathcal{C}^v$	-	-	65.46	-	-
	Joint-Train	71.15	24.28	60.14	0.7668	0.1825
rand-pad	$\mathcal{C}^a$	-	40.63	-	-	-
	$\mathcal{C}^v$	-	-	65.26	-	-
	Joint-Train	71.15	24.28	60.14	0.7147	0.2324

Table 6. Comparing the impact of mono-modal concept with different padding methods on competition strength in the AV-MNIST dataset early fusion joint-training. **zero-pad** indicates padding the input modality with zero vector and **rand-pad** pad input modality with normal distribution.

technique, we are mostly concerned about the robustness of the mono-modal concept.

To this end, We first train the mono-modal concept with different random seeds in initialization on the AV-MNIST dataset. The result is shown in Table 5. As expected, corresponding competition strengths are of similar magnitudes.

We then compare the cases where the mono-modal concepts are computed using different padding methods. Recall that we have adopted zero-padding for $\mathbf{0}^m$ to represent the absence of the modality m. In this control experiment, we use the random-padding instead. In other words, all the elements in $\mathbf{0}^m$ are drawn independently from the normal distribution $N(0, 1)$. It is arguable that both the zero-padding and random-padding stand for the competition-less state as they carry no task-relevant information. Note that the padding method only matters in the early and hybrid fusion cases. Table 6 summarises the results on the AV-MNIST dataset with early fusion models. Clearly, the values of competition strength in the zero-padding case are close to the corresponding ones in the random-padding case.

At last, we compare the performance of the mono-modal concept in different fusion strategies. Recall that the mono-modal concept is a function that maps the mono-modal input to a vector in $\mathbb{R}^K$, which can be used for prediction. The performance of the mono-modal concept refers to its prediction accuracy and, hence, represents the amount of task-relevant information in the corresponding modality. From the results in Table 1 to 3, we find that the performance of the mono-modal concept is very similar in the late and early fusion cases on each dataset. It is noteworthy that the performance of mono-model concepts in Tables 5 and 6 are all close to each other as well. This is desirable since the amount of task-relevant information should be independent of specific models.

In summary, the results verify the robustness of the mono-modal concept under different situations and indicate that the competition strength is a well-defined metric.

C. Additional Results

Kinetics-Sound		Acc	Acc_a	Acc_v	d^a	d^v
Late Fusion	$\mathcal{C}^a$	-	42.06	-	-	-
	$\mathcal{C}^v$	-	-	49.23	-	-
	Joint-Train	52.78	39.92	23.84	0.6392	0.7064
	AGM	56.93	31.01	37.04	0.7726	0.5916
FiLM	$\mathcal{C}^a$	-	41.86	-	-	-
	$\mathcal{C}^v$	-	-	48.76	-	-
	Joint-Train	51.17	34.76	25.32	0.6416	0.6691
	AGM	55.73	48.56	51.57	0.6861	0.5045

Table 7. Experiments on the Kinetics-Sound dataset with late fusion and FiLM [7] strategies.

In this section, we present additional experiment results on the Kinetics-Sound dataset with both the later fusion and the FiLM fusion [7] strategies. Apart from the implementation of the fusion module for the FiLM case, the encoder network and training parameters are the same as those in the AVMNIST late fusion setting.

Table 7 shows the result on the Kinetics-Sound dataset with late fusion and FiLM, the improvement on which are 3.15% and 3.56%, respectively. Comparing joint-train and AGM, the competition strengths of the visual modality decrease for both fusion strategies, which demonstrates that AGM pushes the model to rely on the more informative modality. These additional results further demonstrate the universal effectiveness of AGM.

			Acc	Acc_a	Acc_v	d^a	d^v
AV-MNIST	L-f	OGM-GE(RA)	70.43	18.81	55.87	0.7329	0.1362
		AGM(**1**)	71.63	38.35	63.50	0.6849	0.1313
		AGM-GE	72.03	40.24	64.52	0.7006	0.1215
	E-f	AGM(**1**)	71.72	67.89	66.53	0.7640	0.1813
		AGM-GE	71.88	35.88	67.89	0.7368	0.1798
CREMA-D	L-f	OGM-GE(RA)	64.28	60.69	25.41	0.4436	0.7423
		AGM(**1**)	72.05	39.46	44.39	0.6370	0.6103
		AGM-GE	78.03	45.44	50.22	0.6254	0.5152
	E-f	AGM(**1**)	71.15	69.66	73.24	0.6507	0.6726
		AGM-GE	81.02	75.49	77.73	0.8421	0.7583

Table 8. Experiments on AV-MNIST and CREMA-D with different ablation experiments. OGM-GE(RA) indicates the OGM-GE method discrepancy ratio toward the running average. AGM(1) is our AGM method tunning toward 1. AGM-GE is our AGM with Generalization Enhancement(GE).

D. Ablation Study

In this section, we provide an in-depth comparison between AGM and OGM-GE as their performance outstands in our experiments. Specifically, we tune the AGM discrepancy ratio towards 1 instead of the running average to justify the usefulness of the running average as the reference. On the other hand, we try to tune the discrepancy ratio in OGM-GE toward the running average instead of simply 1 to see whether it could improve the performance. We also integrate our AGM with the generalization enhancement (GE) technique in OGM-GE and run additional experiments to test its comparability with our modulation method.

Table 8 shows the result of the above-mentioned experiments on the AV-MNIST and CREMA-D datasets. The running average of AGM tuning toward 1 improves the performance compared to the joint-training case while being worse than the one using the running average. It reflects that the running average push model uses the modality with more information. We find that the running average does not improve the OGM-GE method, which attributes to that AGM and OGM-GE adopt different ways to compute the discrepancy ratio, the latter may not be compatible with the running average. Unlike OGM-GE, GE does not improve our AGM. One possible reason is that the running average introduces additional fluctuations in the gradient which is similar to the effect of the noise term in GE.

References

[1] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[2] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[3] Md Kamrul Hasan, Wasifur Rahman, Amir Zadeh, Jianyuan Zhong, Md Iftekhar Tanveer, Louis-Philippe Morency, et al. Ur-funny: A multimodal language dataset for understanding humor. *arXiv preprint arXiv:1904.06618*, 2019.

[4] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

[5] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A Lee, Yuke Zhu, et al. Multibench: Multiscale benchmarks for multimodal representation learning. *arXiv preprint arXiv:2107.07502*, 2021.

[6] Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced Multimodal Learning via On-the-fly Gradient Modulation. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8228–8237, New Orleans, LA, USA, June 2022. IEEE.

[7] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[8] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 247–263, 2018.

[9] Valentin Vielzeuf, Alexis Lechervy, Stéphane Pateux, and Frédéric Jurie. Centralnet: a multilayer approach for multimodal fusion. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

[10] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

[11] Jinxing Zhou, Liang Zheng, Yiran Zhong, Shijie Hao, and Meng Wang. Positive sample propagation along the audio-visual event line. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8436–8444, 2021.