SimMLM: A Simple Framework for Multi-modal Learning with Missing Modality

Supplementary Material

Implementation of the gating network The gating network G in our DMoME framework is designed with a generic and efficient architecture to ensure adaptability across diverse tasks and modalities. At a high level, it consists of a feature extraction block (e.g., 2D or 3D convolutional neural networks (CNNs), transformers, etc.), followed by a linear layer that generates low-dimensional weight vectors. This architecture strikes a balance between simplicity and flexibility, enabling it to adjust modality contributions effectively across tasks. Furthermore, its modular design allows seamless accommodation of varying task requirements and diverse input modalities, ensuring broad applicability and adaptability.

As shown in Table A1, for the BraTS 2018 segmentation task, the input modalities are four input imaging modalities ($4 \times 128 \times 128 \times 128$ with zeros if certain modalities are missing). Here, since they are all images sharing the same dimensionality, we simply extract $256 - dim$ features using a 5-layer 3D CNN to directly fuse visual information and send the features to an linear layer to get a 4×3 gating vector (# modalities $\times$ #tasks). The vector will be transformed using the softmax function to generate gating weights, which adjust the contribution of each modality per task. In contrast, the avMNIST classification task involves heterogeneous modalities: the image modality ($1 \times 28 \times 28$) and the audio modality ($1 \times 20 \times 20$). To account for this heterogeneity, two separate 2-layer CNNs are used to extract 128-dimensional feature vectors for each modality. These features are concatenated to form the same size of $256 - dim$ representation, which is then processed by a similar linear layer to generate a 2×1 gating vector. This modular approach ensures flexibility and robust feature integration across diverse data types.

In this work, we simply designed these two lightweight gating networks empirically, which provides substantial improvements. Thanks to the flexibility of the framework, readers are encouraged to explore more advanced architectures for their customized tasks.

General design	BraTS 2018	avMNIST	
Feature extractor	4-modality MRIs 4×128^3	image 1×28^2	audio 1×20^2
	5-layer CNNs (3D)	2-layer CNNs (2D)	2-layer CNNs (2D)
	Flatten	Flatten & concat	
	256-dim feature	256-dim feature	
Linear layer	Linear layer Output size: 4×3 (4 modalities $\times$ 3 tasks)	Linear layer Output size: 2×1 (2 modalities $\times$ 1 task)	

Table A1. Designs of the gating network G for the BraTS 2018 segmentation task and the avMNIST classification task.

	Dice score ↑		
Method	ET	TC	WT
DMoME w/o expert pretraining	62.24	77.42	86.63
DMoME w/ static averaging	61.40	77.49	86.70
DMoME w/ dynamic weighting (SimMLM)	**63.22**	**78.54**	**87.21**

Table A2. Importance of the design choice and the training recipe in SimMLM. Reported values are the average segmentation performance on the BraTS 2018 validation set, across all missing & full modality settings.

Importance of the design choice and the training recipe in SimMLM In this section, we would like to highlight the contribution of pretraining and dynamic weighting by comparing our proposed DMoME with its two variants:

- **DMoME w/o expert pretraining**: Here, all modality experts and the gating network G were directly co-trained from random initialization, without pretraining. The total number of training epochs is the same as the proposed two-stage one for fair comparison.
- **DMoME w/ static averaging**: In this setting, the gating network G is replaced with a simple static averaging operation. The final output can be simply viewed as the mean of the available modality experts' outputs.

Results in Table A2 demonstrate that both expert pretraining and dynamic reweighting contribute to the success of SimMLM. Skipping expert pretraining (**DMoME w/o expert pretraining**) results in a noticeable performance drop, highlighting the critical role of independent pretraining. This step enables modality experts to focus on task-relevant knowledge without interference from cross-expert interactions, which could be noisy especially at the beginning of training. Replacing the dynamic gating mechanism with static averaging (**DMoME w/ static averaging**) also causes a significant performance drop across all subregions (ET, TC, WT). This underscores the effectiveness of dynamically assigning weights to modality experts, which is not just a simple model ensembling strategy.

Impact of performing dynamic weighting at different levels in DMoME As shown in Figure A1, we propose a simple yet effective mixture of modality experts strategy, which directly applies expert weighting at the logit level (before softmax) instead of probabilities.

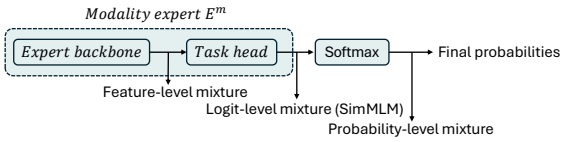

Figure A1. Three types of weighted sum micro design: feature-level, logit-level (SimMLM), and probability-level.

Here, we tested the importance of this design choice, by comparing it to the other two alternatives with the same gating network design:

- **Feature-level mixture**: Here, the dynamic weighting is applied at hidden feature maps (before passing through the segmentation/classification head) rather than the output space, which is adopted in MoMKE [3]. The feature maps from modality experts are later mixed with the weights produced by the gating network. A segmentation head is followed to generate the final results given the mixed feature map.
- **Probability-level mixture**: Here the dynamic weighting is applied to the probabilities from each expert (after softmax).

Results in Table A3 show that performing logit-level mixture generally achieves better results, especially on the more challenging enhancing tumor (ET) and tumor core (TC) segmentation tasks. We believe this improvement stems from SimMLM's logit-level weighting, where the weighting parameter functions as a temperature to rescale the confidence of each modality's prediction in the final output. This approach aligns with the principles of temperature rescaling [1] for model calibration, which can effectively adapt model predictions to better reflect the likelihood of ground truth correctness, leading to more reliable multi-modal inference especially in challenging scenarios. In our main paper, we have already demonstrated that our DMoME framework, when used standalone, can achieve much lower calibration errors compared to MoMKE in Table 5, thanks to the logit-level dynamic weighting.

	Dice score ↑		
Method	ET	TC	WT
Feature-level mixture	62.76	78.19	87.27
Probability-level mixture	62.72	78.31	**87.33**
Logit-level mixture (SimMLM)	**63.22**	**78.54**	87.21

Table A3. Importance of performing dynamic weighting at logit level. Reported values are the average segmentation performance on the BraTS 2018 validation set, across all missing & full modality settings.

SimMLM's robustness against varied levels of missing modalities at training time To further investigate

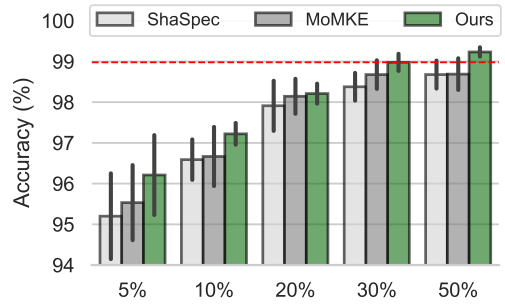

Figure A2. Model performance comparison with varying rates of missing audio data during training on av-MNIST.

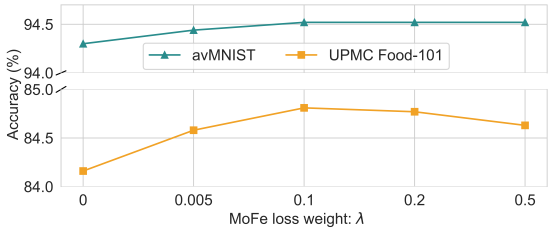

Figure A3. MoFe weight λ tuning on avMNIST, UPMC Food-101.

SimMLM's resilience to training-time missingness, we conducted experiments where the audio coverage rate was reduced to 5%, 10%, 20%, 30%, 50% during the training. We then evaluated the model on the full modality evaluation set and compared its performance to ShaSpec [2] and MoMKE [3]. As shown in Figure A2, our model consistently achieves the highest accuracy in all settings, demonstrating substantial improvement in handling training-time modality absence. Remarkably, our model achieves even better accuracy comparable to competing methods that use 50% of the audio data, with only 30% of it during training (see red dashed line). These results demonstrate our method's potential to handle training-time modality missingness as well.

MoFe coefficient tuning on avMNIST and UPMC Food-101 Setting the MoFe loss coefficient to 0.1 generalizes well across tasks. As shown in Fig. A3, where MoFe consistently improves performance across a range of values, with 0.1 yielding stable results.

Discussion: How well does the unimodal performance align with the importance derived from the learned gating weights? Interestingly, the gating weights show weak to moderate alignment with unimodal performance *at the sample level*. On UPMC Food-101, the Pearson correlation between each modality's gating weight and its unimodal confidence on the correct class is 0.29 for image and 0.41

for text. This partial alignment suggests that while unimodal confidence reflects individual modality strength, it may not fully capture cross-modal complementarity or sample difficulty—both of which influence the model's reliance during inference. We consider this behavior desirable, as it indicates the gating mechanism is context-aware rather than simply mirroring unimodal confidence.

Clinical value of SimMLM In real-world clinical scenarios, especially in low-resource settings, certain imaging modalities may be unavailable due to technical issues, equipment limitations, or patient factors. SimMLM's ability to generate more reliable and accurate segmentation under these conditions is essential. The model not only provides direct importance values for each input modality but also generates more informative confidence maps, including entropy maps, for voxel-wise predictions in the output space. This capability is particularly valuable in critical applications like tumor detection and diagnosis, where accurate and detailed segmentation is crucial for treatment planning and patient care. By enabling clinicians to make confident decisions—whether by providing insights or suggesting further investigation or special care for uncertain regions—even in the presence of incomplete data, SimMLM greatly enhances the robustness and reliability of clinical workflows. In the future, we will explore SimMLM's application in broader, high-stakes scenarios, such as surgical scene reconstruction, action segmentation, and surgical planning.

References

[1] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 2

[2] Hu Wang, Yuanhong Chen, Congbo Ma, Jodie Avery, Louise Hull, and Gustavo Carneiro. Multi-modal learning with missing modality via shared-specific feature modelling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15878–15887, 2023. 2

[3] Wenxin Xu, Hexin Jiang, et al. Leveraging knowledge of modality experts for incomplete multimodal learning. In *ACM Multimedia 2024*, 2024. 2