MMP: Towards Robust Multi-Modal Learning with Masked Modality Projection

Niki Nezakati[1], Md Kaykobad Reza[1], Ameya Patil[2], Mashhour Solh[2], M. Salman Asif[1]

[1]University of California, Riverside [2]Amazon

***Abstract*—Multimodal learning seeks to combine data from multiple input sources to enhance the performance of different downstream tasks. In real-world scenarios, performance can degrade substantially if some input modalities are missing. Existing methods that can handle missing modalities involve custom training or adaptation steps for each input modality combination. These approaches are either tied to specific modalities or become computationally expensive as the number of input modalities increases. In this paper, we propose Masked Modality Projection (MMP), a method designed to train a single model that is robust to any missing modality scenario. We achieve this by randomly masking a subset of modalities during training and learning to project available input modalities to estimate the tokens for the masked modalities. This approach enables the model to effectively learn to leverage the information from the available modalities to compensate for the missing ones, enhancing missing modality robustness. We conduct a series of experiments with various baseline models and datasets to assess the effectiveness of this strategy. Experiments demonstrate that our approach improves robustness to different missing modality scenarios, outperforming existing methods designed for missing modalities or specific modality combinations.**

***Index Terms*—Multimodal learning, missing modality robustness**

I. Introduction

Multimodal learning (MML) [1], [2] leverages information from multiple input sources to enhance task performance [3], [4]. However, these models are typically trained assuming all modalities are present at inference. In real-world settings, modalities may be missing due to sensor malfunction, privacy constraints, or limited acquisition, causing significant performance degradation [5], [6]. This paper investigates this missing modality problem and shows that a single robustly trained model can outperform existing methods across different missing-modality scenarios.

Existing approaches include robust training [7], [8], modality masking [9], [10], and knowledge distillation [11]–[13]. Prompt-based methods [6], [14] require separate prompts per scenario and do not scale, while imputation via GANs/VAEs [15]–[17] adds overhead. In this paper, we propose **M**asked **M**odality **P**rojection (MMP), a method for training a single model robust to any missing-modality case. As illustrated in Figure 1, MMP applies modality masking during training and uses modality projection to predict tokens for masked modalities from available ones. An alignment loss encourages projected tokens to match actual tokens, while projected tokens replace missing inputs at inference. MMP is architecture-agnostic and requires no per-scenario retraining. Experiments across three models, five datasets, and three tasks (Section IV) show that MMP significantly improves performance under missing modalities while maintaining strong performance when all modalities are present. Our main contributions are:

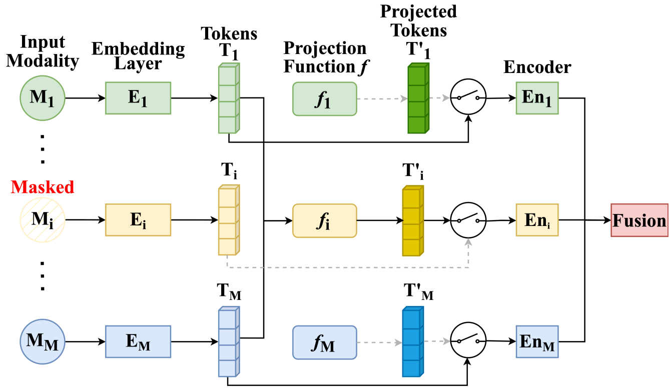


Fig. 1. Architecture of the proposed MMP approach for training a single multimodal model that is robust to missing modalities. Input modalities are passed through embedding layers, generating tokens. For a masked modality $\mathbf{M}_i$, a projection function utilizes the tokens from the available modalities to generate projected tokens. These projected tokens are then passed to the masked modality branch.

- We introduce MMP, a novel approach to predict missing modality tokens from available modalities, improving robustness to missing data.
- MMP achieves strong performance under missing modalities, often matching or exceeding networks trained for specific modality subsets.
- MMP requires minimal architectural changes, making it broadly applicable across multimodal tasks and models.
- We validate MMP on image segmentation, image-text classification, and text-visual-audio sentiment analysis across five diverse datasets.

II. Related Work

Robust model design approaches include learning modality-specific and shared features [18], modality masking with knowledge distillation [10], dynamic token replacement [19], and robust fusion strategies [20]–[22]. [23] proposed a Transformer-based approach for MRI missing modality tasks. However, these models are generally task-specific and difficult to generalize.

Robust training approaches apply modality dropout [7], [8], complementary random masking [10], [21], masked autoencoders [9], masked cross attention [5], and modality perturbation [24]. While these methods improve robustness, they cannot fully compensate for performance drops with missing modalities.

Model adaptation methods learn prompts for each modality combination [6] or per modality [14]. Recent work in [25] utilized parameter-efficient adaptation for generic robustness, while [26] utilizes cross-modal proxy tokens along with the parameter-efficient adaptation of the encoders. The main limitation is the requirement for separate learnable parameters for each modality combination.

Generation and knowledge distillation approaches use GANs [15], [16], [27], VAEs [17], diffusion models [28], or feature generation [29] to handle missing modalities. Knowledge distillation methods [10], [12], [13], [30] transfer comprehensive multimodal information. These approaches require training or utilizing additional models.

In this paper, we train a single model robust to any missing modality scenario by generating tokens for missing modalities from available inputs, without tuning or adapting for specific modality combinations.

III. Method

In this section, we introduce Masked Modality Projection (MMP), a novel approach for training a single multimodal model that is robust to missing modalities. During training, a subset of modalities is randomly masked out, and we introduce projection functions that learn to map tokens from available modalities to the missing modality tokens. These projected tokens are aligned with actual tokens using an alignment loss objective.

A. Modality Masking

In the MMP framework, each modality is randomly available or masked at each iteration with probability 0.5, providing balanced exposure to different modality combinations. Masked modalities are replaced with placeholders (zeros for visual/audio data or empty strings for text), following standard practices from prior works [6], [25].

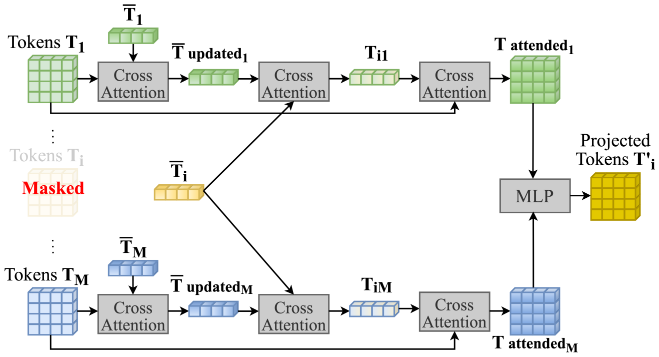
B. Modality Projection

We propose a modality projection approach for masked modality i, illustrated in Figure 2. The embedding layers generate tokens from each input modality as

$$\mathbf{T}_i = \text{EmbeddingLayer}(\mathbf{I}_i), \tag{1}$$

where $\mathbf{I}_i$ represents input for modality $i \in \{1, 2, \ldots, M\}$, $\mathbf{T}_i \in \mathbb{R}^{N\times d}$ represents tokens for the corresponding modality, N is the number of tokens, and d is the embedding dimension. For simplicity, we assume N and d are consistent across modalities (see Section III-C for handling variability).

Following [31], we use eight learnable aggregated tokens $\overline{\mathbf{T}}_i$ per modality to summarize information compactly. When modality j is available, its aggregated tokens are updated via cross-attention:



$\overline{\mathbf{T}}_1$
Tokens $\mathbf{T}_1$
$\overline{\mathbf{T}}$ updated$_1$
Cross Attention
$\mathbf{T}_{i1}$
Cross Attention
Cross Attention
$\mathbf{T}$ attended$_1$
Tokens $\mathbf{T}_i$
Masked
$\overline{\mathbf{T}}_i$
Projected Tokens $\mathbf{T'}_i$
MLP
$\overline{\mathbf{T}}_M$
Tokens $\mathbf{T}_M$
$\overline{\mathbf{T}}$ updated$_M$
Cross Attention
$\mathbf{T}_{iM}$
Cross Attention
Cross Attention
$\mathbf{T}$ attended$_M$

Fig. 2. Visualization of the modality projection approach. Available modality tokens are processed through cross-attention to update their aggregated tokens. These aggregated tokens are combined with those of the masked modality through another cross-attention step. The resulting cross-modal relationships are used to attend to the actual tokens of the available modalities. The final output tokens are passed through an MLP to generate the projected tokens of the masked modality.

$$\begin{aligned}\overline{\mathbf{T}}_{\text{updated}_j} &= \text{CrossAttention}\left(\overline{\mathbf{T}}_j, \mathbf{T}_j \mid j \in \mathcal{A}\right), \\ &= \text{softmax}\left(\frac{\overline{\mathbf{T}}_j \mathbf{W}_\text{q} \mathbf{W}_\text{k}^\top \mathbf{T}_j^\top}{\sqrt{d}}\right) \mathbf{T}_j \mathbf{W}_\text{v},\end{aligned} \tag{2}$$

where $\mathcal{A}$ is the set of available modalities and $\mathbf{W}_\text{q}$, $\mathbf{W}_\text{k}$, and $\mathbf{W}_\text{v}$ are learnable weight matrices. For missing modality i, we perform cross-attention between its aggregated tokens and those of each available modality:

$$\mathbf{T}_{ij} = \text{CrossAttention}(\overline{\mathbf{T}}_i, \overline{\mathbf{T}}_{\text{updated}_j}), \tag{3}$$

where $\mathbf{T}_{ij}$ represents the attended tokens for available modality j in relation to missing modality i. These attended tokens are refined with the original tokens of available modalities:

$$\mathbf{T}_{\text{attended}_j} = \text{CrossAttention}(\mathbf{T}_j, \mathbf{T}_{ij}). \tag{4}$$

The refined tokens are concatenated and fed through an MLP to produce projected tokens:

$$\mathbf{T}'_i = \text{MLP}\left(\text{Concat}\left(\left\{\mathbf{T}_{\text{attended}_j} \mid j \in \mathcal{A}\right\}\right)\right). \tag{5}$$

These projected tokens $\mathbf{T}'_i$ replace the missing modality tokens and are passed to their respective branch.

C. Token and Dimension Variability

When embedding dimensions differ across modalities, we apply a linear layer to map tokens to a common dimension. For varying token counts, we incorporate a linear layer in the MLP to align the projected token count with the missing modality.

D. Alignment Loss Objective

To minimize discrepancy between projected and real tokens, the alignment loss is computed as

$$\mathcal{L}_{\text{alignment}} = \frac{1}{N_{\text{masked}}} \sum_{i\in\text{masked}} \mathcal{L}_{\text{alignment}_i}(\mathbf{T}'_i, \mathbf{T}_i), \tag{6}$$

where $\mathcal{L}_{\text{alignment}_i}$ represents Smooth L1 loss. The total loss combines task-specific and alignment losses:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \mathcal{L}_{\text{alignment}}. \tag{7}$$

IV. Experiments and Results

We evaluate our proposed method on multimodal segmentation and classification tasks across five datasets, comparing with established baseline methods for missing modalities.

A. Datasets

We evaluate on five datasets: MCubeS [32], NYUDv2 [33], and FMB [34] for multimodal segmentation; UPMC Food-101 [35] for image-text classification; and CMU-MOSI [36] for text-visual-audio sentiment analysis.

B. Implementation Details

We use CMNeXt [37] for segmentation, ViLT [38] for classification, and multimodal transformer [39] for sentiment analysis with standard training configurations and AdamW optimizer [40].

C. Results on Multimodal Segmentation

Table I compares performance across MCubeS, NYUDv2, and FMB datasets using CMNeXt [37]. **Pretrained** indicates training without dropout augmentation, **Modality Dropout** uses dropout during training, and **MMP** uses our modality projection approach.

Effects of missing modalities. Pretrained models show significant performance drop with missing modalities. While modality dropout improves robustness, MMP outperforms both approaches in every missing modality scenario. The slightly lower MMP performance with all modalities available is due to the base model being pretrained with modality dropout.

For MCubeS, when RGB is missing and A-D-N are available, MMP achieves 38.57 mIoU versus 33.88 for modality dropout and 1.45 for pretrained. With only A available, MMP achieves 31.31 versus 26.3 and 1.13. For NYUDv2 with only Depth available, MMP achieves 41.08 versus 29.79 and 6.01. For FMB with only Thermal, MMP achieves 51.73 versus 39.66 and 23.35, demonstrating consistent superiority in handling missing modalities.

Comparison with other methods. Table II shows MMP achieves superior average performance on NYUDv2. When Depth is missing, MMP ranks second to Reza et al. [25] with minimal difference (-0.78%), while outperforming in average (+1.79%) and RGB-missing (+4.36%) scenarios without requiring separate adaptation for each modality combination. When RGB is missing, MMP is second to MMANet [30], which relies on a teacher model, yet MMP achieves better performance in other cases without this complexity.

D. Visualization of Predictions

Figure 3 visualizes predictions from pretrained CMNeXt and MMP. The pretrained model struggles when modalities are missing: failing to detect bikes and cars on MCubeS with missing RGB (Figure 3a), showing reduced accuracy for kitchen objects on NYUDv2 (Figure 3b), and failing to detect cars, bicyclists, and humans on FMB (Figure 3c). MMP successfully detects these objects even with missing modalities, with predictions either better or comparable to the pretrained model with all modalities available.

TABLE I

Performance comparison (mIoU) on multimodal segmentation. A, D and N denote angle of linear polarization, degree of linear polarization, and near-infrared. Best and second-best results are shown as **bold** and underlined, respectively.

Dataset	Input	Pretrained	Dropout	**MMP**
MCubeS	All	**51.54**	48.56	48.95
	RGB	42.32	47.64	**48.65**
	A-D-N	1.45	33.88	**38.57**
	A-D	0.93	33.15	**37.74**
	A	1.13	26.30	**31.31**
NYUDv2	All	**56.30**	51.12	53.81
	RGB	51.05	48.80	**52.04**
	Depth	6.01	29.79	**41.08**
FMB	All	**62.68**	54.11	60.03
	RGB	22.20	48.32	**55.83**
	Thermal	23.35	39.66	**51.73**

TABLE II

Performance (mIoU) comparison on NYUDv2 dataset. RGB, Depth and Avg. columns show RGB only, Depth only and average performance. Best and second-best results are shown as **bold** and underlined, respectively.

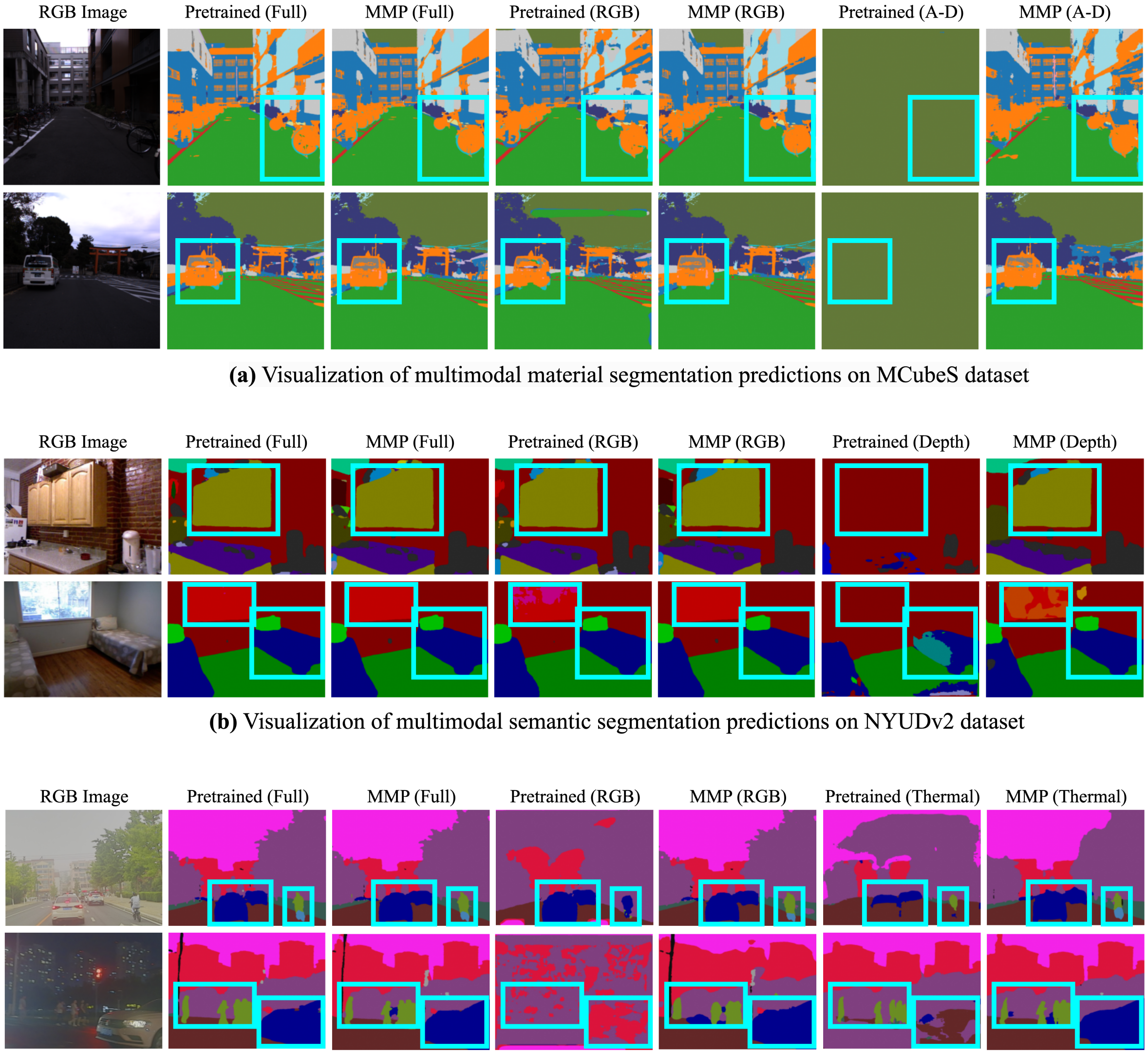
Methods	Backbone	RGB	Depth	Avg.
AsymFusion [41]	ResNet-101	46.50	34.30	40.40
CEN [42]	ResNet-101	39.59	19.32	29.46
TokenFusion [19]	MiT-B3	49.32	36.84	43.08
Reza et al. [25]	MiT-B4	**52.82**	36.72	44.77
MMANet [30]	ResNet-50	44.93	**42.75**	43.84
HeMIS [43]	ResNet-50	33.23	31.23	32.23
CMNeXt [37]	MiT-B4	51.19	5.26	28.23
RFNet [44]	ResNet-50	42.89	40.76	41.82
MMP (Ours)	MiT-B4	52.04	41.08	**46.56**

E. Results on Multimodal Classification

Table III compares MMP against missing-aware prompts [6] on UPMC Food-101 [35] using ViLT [38]. MMP outperforms prompting-based methods in most scenarios, particularly when all modalities are available, when 35% of both modalities are missing, when 70% of images are missing, and when no images are available. MMP shows slight decreases in two cases: 0.21% lower when 70% of text is missing, and 2.04% lower when no text is available, as prompting methods learn dedicated prompts for each scenario. Notably, MMP maintains strong performance without requiring training for every modality combination.

F. Results on Multimodal Sentiment Analysis

Table IV presents results on CMU-MOSI [36] using multimodal transformer (MulT) [39]. When text is present, missing audio or video has minimal impact. However, performance



(a) Visualization of multimodal material segmentation predictions on MCubeS dataset

(b) Visualization of multimodal semantic segmentation predictions on NYUDv2 dataset

(c) Visualization of multimodal semantic segmentation predictions on FMB dataset

Fig. 3. Visualization of predicted segmentation maps for the Pretrained (CMNeXt) model and our MMP approach. Title above each image indicates the method name (available modalities). Blue boxes mark the areas where the differences are more prominent. A and D denote angle and degree of linear polarization, respectively.

TABLE III
PERFORMANCE COMPARISON ON UPMC FOOD-101 DATASET. AVAILABLE MODALITY SHOWS PERCENTAGE AVAILABLE DURING INFERENCE. † INDICATES RESULTS GENERATED FROM AUTHORS' CODE. BEST AND SECOND-BEST RESULTS ARE SHOWN AS **BOLD** AND UNDERLINED, RESPECTIVELY.

Available Modality		ViLT	Missing Prompts [6]		MMP
Image	Text	[38]	Attention	Input	(Ours)
100%	100%	92.71†	92.71	92.71	**92.87**
100%	30%	66.29	72.57	**74.53**	74.32
100%	0%	23.70†	67.70	**68.10**	66.06
65%	65%	69.25	78.09	79.08	**80.28**
30%	100%	76.66	86.05	86.18	**87.71**
0%	100%	82.65†	85.30	84.80	**85.37**

drops significantly without text, where MMP provides substantial improvements: 5.8% in accuracy and 10.29% in F1 when text is missing, and 6.72% in accuracy and 13% in F1 when only audio is available. MMP outperforms existing methods in both metrics across all scenarios except when only text is missing, where it ranks second with minimal difference from [25], while surpassing this method by large margins in other scenarios.

G. Ablation Studies

Table V presents ablation results on NYUDv2. Adding linear projection (LP) to modality dropout improves performance across all scenarios. Adding alignment loss (Align) provides further improvements, particularly when RGB is missing. Finally, replacing linear projection with cross-attention based (CA) projection achieves the best performance, demonstrating

TABLE IV

PERFORMANCE (BINARY ACCURACY AND F1 SCORE) COMPARISON WITH EXISTING METHODS FOR MULTIMODAL SENTIMENT ANALYSIS ON CMU-MOSI DATASET. COLUMN NAMES INDICATE AVAILABLE MODALITIES. BEST AND SECOND-BEST RESULTS ARE SHOWN AS **BOLD** AND UNDERLINED, RESPECTIVELY.

Method	Backbone	Text-Visual-Audio		Visual-Audio		Audio		Average	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
MulT [39]	Transformer	79.57	79.67	48.93	41.95	48.31	40.98	58.93	54.20
TFN [45]	LSTM	73.90	73.40	42.23	25.07	42.23	25.07	52.78	41.18
LMF [46]	LSTM	76.40	75.70	43.29	27.61	42.23	25.07	53.97	42.79
Reza et al. [25]	Transformer	79.57	79.67	**55.49**	**53.96**	50.00	46.71	61.68	60.11
MMP (Ours)	Transformer	**80.03**	**80.04**	54.73	52.24	**55.03**	**53.98**	**63.26**	**62.08**

TABLE V

ABLATION STUDIES ON NYUDV2 DATASET. PERFORMANCE INCREASES WITH LINEAR PROJECTION (LP), ALIGNMENT LOSS (ALIGN), AND CROSS ATTENTION (CA).

Methods	RGB-Depth	RGB	Depth	Average
Dropout	51.12	48.80	29.79	43.23
Dropout + LP	51.31	51.08	35.48	45.95
Dropout + LP + Align	52.84	50.73	40.60	48.05
Dropout + CA + Align	**53.81**	**52.04**	**41.08**	**48.97**

TABLE VI

COMPARISON OF TRAINING TIME IN HOURS (H) AND PARAMETER COUNT IN MILLION (M) FOR BASELINE MODELS AND WITH MMP. MMP PARAMETERS SHOW ADDITIONAL PARAMETERS; MMP TRAINING TIME SHOWS TOTAL TRAINING TIME.

Model (Dataset)	Parameters (M)		Training Time (H)	
	Baseline	MMP	Baseline	MMP
CMNeXt [37] (MCubeS)	58.73	0.63	7.1	7.2
CMNeXt [37] (NYUDv2)	117.00	0.20	25.9	27.6
CMNeXt [37] (FMB)	90.01	0.20	11.0	13.5
ViLT [38] (UPMC Food-101)	112.26	11.05	15.1	15.4
MulT [39] (CMU-MOSI)	2.57	14.61	0.9	2.5

that each component in MMP contributes to enhanced robustness.

H. Computational Cost Analysis

MMP's computational cost arises from cross-attention operations and the MLP layer. For M modalities with N tokens and embedding dimension d, the total parameter count is $O(Md^2)$ and computation complexity is $O(NM^2d^2 + 3MNd)$. The number of cross-attention blocks is $2m + nm$, where m is available modalities and n is missing modalities.

Table VI shows MMP adds minimal parameters except for ViLT [38] and MulT [39] due to large text embedding dimensions, though training time overhead remains low. MMP introduces minimal computational cost while providing strong robustness with a single training procedure.

V. CONCLUSION

In this paper, we introduced Masked Modality Projection (MMP), a novel approach designed to enhance missing modality robustness of multimodal models. Our approach eliminates the need for training or adapting models for specific missing modality scenarios. We demonstrate that a single model can effectively handle any missing modality scenario and outperform current baselines. Thus, it reduces both time and computational overhead. Experimental results across several baseline models and datasets validate that MMP significantly improves performance and robustness compared to existing baseline methods. Future work will focus on further refining MMP and exploring its applicability to other multimodal tasks and datasets. Additionally, while our method is agnostic to modality quality, the quality of available modalities can influence MMP's performance. Investigating these effects represents an interesting direction for future research. We believe that MMP offers an efficient and effective solution to the challenge of missing modalities.

Acknowledgment: This work is supported in part by NSF awards 2046293 and 2406199, AFOSR award FA9550-21-1-0330, and an Amazon Gift award. This work used Indiana Jetstream2 through allocation CIS220128 from the ACCESS program supported by NSF grants 2138259, 2138286, 2138307, 2137603, and 2138296.

REFERENCES

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[2] P. Xu, X. Zhu, and D. A. Clifton, "Multimodal learning with transformers: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.

[3] Y. Huang, C. Du, Z. Xue, X. Chen, H. Zhao, and L. Huang, "What makes multi-modal learning better than single (provably)," *Advances in Neural Information Processing Systems*, vol. 34, pp. 10 944–10 956, 2021.

[4] Z. Lu, "A theory of multimodal learning," *Advances in Neural Information Processing Systems*, vol. 36, 2024.

[5] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.

[6] Y.-L. Lee, Y.-H. Tsai, W.-C. Chiu, and C.-Y. Lee, "Multimodal prompting with missing modalities for visual recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 14 943–14 952.

[7] N. Neverova, C. Wolf, G. Taylor, and F. Nebout, "ModDrop: Adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1692–1706, 2015.

[8] A. Hussen Abdelaziz, B.-J. Theobald, P. Dixon, R. Knothe, N. Apostoloff, and S. Kajareker, "Modality dropout for improved performance-driven talking faces," in *Proceedings of the 2020 International Conference on Multimodal Interaction*, 2020, pp. 378–386.

[9] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "MultiMAE: Multimodal multi-task masked autoencoders," in *European Conference on Computer Vision*, 2022.

[10] U. Shin, K. Lee, I. S. Kweon, and J. Oh, "Complementary random masking for rgb-thermal semantic segmentation," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 11 110–11 117.

[11] A. Tarvainen and H. Valpola, "Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results," *Advances in neural information processing systems*, vol. 30, 2017.

[12] H. Maheshwari, Y.-C. Liu, and Z. Kira, "Missing modality robustness in semi-supervised multi-modal semantic segmentation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE Computer Society, 2024, pp. 1009–1019.

[13] R. Wu, H. Wang, F. Dayoub, and H.-T. Chen, "Segment beyond view: Handling partially missing modality for audio-visual semantic segmentation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 6100–6108.

[14] J. Jang, Y. Wang, and C. Kim, "Towards robust multimodal prompting with missing modalities," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 8070–8074.

[15] B. Yu, L. Zhou, L. Wang, J. Fripp, and P. Bourgeat, "3D cGAN based cross-modality MR image synthesis for brain tumor segmentation," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*. IEEE, 2018, pp. 626–630.

[16] A. Sharma and G. Hamarneh, "Missing MRI pulse sequence synthesis using multi-modal generative adversarial network," *IEEE Transactions on Medical Imaging*, vol. 39, no. 4, pp. 1170–1183, 2019.

[17] R. Dorent, S. Joutard, M. Modat, S. Ourselin, and T. Vercauteren, "Hetero-modal variational encoder-decoder for joint modality completion and segmentation," in *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part II 22*. Springer, 2019, pp. 74–82.

[18] H. Wang, Y. Chen, C. Ma, J. Avery, L. Hull, and G. 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*, 2023, pp. 15 878–15 887.

[19] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 12 186–12 195.

[20] J.-H. Choi and J.-S. Lee, "EmbraceNet: A robust deep learning architecture for multimodal classification," *Information Fusion*, vol. 51, pp. 259–270, 2019.

[21] S. Fan, Z. Wang, Y. Wang, and J. Liu, "SpiderMesh: Spatial-aware demand-guided recursive meshing for RGB-T semantic segmentation," *arXiv:2303.08692*, 2023.

[22] B. Lin, Z. Lin, Y. Guo, Y. Zhang, J. Zou, and S. Fan, "Variational probabilistic fusion network for RGB-T semantic segmentation," *arXiv preprint arXiv:2307.08536*, 2023.

[23] S. Karimijafarbigloo, R. Azad, A. Kazerouni, S. Ebadollahi, and D. Merhof, "Mmcformer: Missing modality compensation transformer for brain tumor segmentation," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1144–1162.

[24] D. Hazarika, Y. Li, B. Cheng, S. Zhao, R. Zimmermann, and S. Poria, "Analyzing modality robustness in multimodal sentiment analysis," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Jul. 2022, pp. 685–696.

[25] M. K. Reza, A. Prater-Bennette, and M. S. Asif, "Robust multimodal learning with missing modalities via parameter-efficient adaptation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–13, 2024.

[26] M. K. Reza, A. Patil, M. Solh, and S. Asif, "Robust multimodal learning via cross-modal proxy tokens," *Transactions on Machine Learning Research*, 2025. [Online]. Available: https://openreview.net/forum?id=Wtc6wvcYJ0

[27] Y. Zhang, C. Peng, Q. Wang, D. Song, K. Li, and S. K. Zhou, "Unified multi-modal image synthesis for missing modality imputation," *IEEE Transactions on Medical Imaging*, 2024.

[28] J. Qu, Y. Yang, W. Dong, and Y. Yang, "Lds2ae: Local diffusion shared-specific autoencoder for multimodal remote sensing image classification with arbitrary missing modalities," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, 2024, pp. 14 731–14 739.

[29] S. Woo, S. Lee, Y. Park, M. A. Nugroho, and C. Kim, "Towards good practices for missing modality robust action recognition," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 2776–2784.

[30] S. Wei, C. Luo, and Y. Luo, "Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 20 039–20 049.

[31] S. Mo and P. Morgado, "Unveiling the power of audio-visual early fusion transformers with dense interactions through masked modeling," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 186–27 196.

[32] Y. Liang, R. Wakaki, S. Nobuhara, and K. Nishino, "Multimodal material segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022, pp. 19 800–19 808.

[33] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *European Conference on Computer Vision*, 2012.

[34] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2023, pp. 8115–8124.

[35] X. Wang, D. Kumar, N. Thome, M. Cord, and F. Precioso, "Recipe recognition with large multimodal food dataset," in *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2015, pp. 1–6.

[36] A. Zadeh, R. Zellers, E. Pincus, and L.-P. Morency, "Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages," *IEEE Intelligent Systems*, vol. 31, no. 6, pp. 82–88, 2016.

[37] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, "Delivering arbitrary-modal semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1136–1147.

[38] W. Kim, B. Son, and I. Kim, "Vilt: Vision-and-language transformer without convolution or region supervision," in *International conference on machine learning*. PMLR, 2021, pp. 5583–5594.

[39] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proceedings of the conference. Association for computational linguistics. Meeting*, vol. 2019. NIH Public Access, 2019, p. 6558.

[40] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[41] Y. Wang, F. Sun, M. Lu, and A. Yao, "Learning deep multimodal feature representation with asymmetric multi-layer fusion," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 3902–3910.

[42] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Advances in neural information processing systems*, vol. 33, pp. 4835–4845, 2020.

[43] M. Havaei, N. Guizard, N. Chapados, and Y. Bengio, "Hemis: Hetero-modal image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II 19*. Springer, 2016, pp. 469–477.

[44] Y. Ding, X. Yu, and Y. Yang, "Rfnet: Region-aware fusion network for incomplete multi-modal brain tumor segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 3975–3984.

[45] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2017, pp. 1103–1114.

[46] Z. Liu, Y. Shen, V. B. Lakshminarasimhan, P. P. Liang, A. Bagher Zadeh, and L.-P. Morency, "Efficient low-rank multimodal fusion with modality-specific factors," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2018, pp. 2247–2256.