Towards a Study on Exploring Missing Modalities through the Lens of Token Embeddings

R E Zera Marveen Lyngkhoi
(2311AI06)

under the guidance of

Dr. Sriparna Saha

Department of Computer Science and Engineering
Indian Institute of Technology Patna

May 20, 2025

Overview

Introduction

Problem Statement: Multimodal models often suffer when one or more input modalities, especially visual information are missing at inference time. Existing solutions like GANs, diffusion models, or architecture modifications incur high computational costs and require pre-training.

Our Solution: We propose a token-level modality imputation method that uses precomputed vision token embeddings from an offline database to substitute missing visual inputs during inference eliminating the need for retraining or generative pipelines.

Why It Matters: Our approach improves **efficiency** and safeguards **privacy**, particularly in sensitive domains like **finance** and **healthcare**, where raw visual data may expose confidential or personally identifiable information. By relying on offline embeddings, no sensitive images need to be stored or processed during inference.

Validated On:

- **MM-Advice** — a new multilingual, code-mixed financial dialogue advisory dataset in financiald domain.

We also evaluated **MM-IMDB**, a benchmark for multimodal classification to test for robustness and generalizability

Methodology

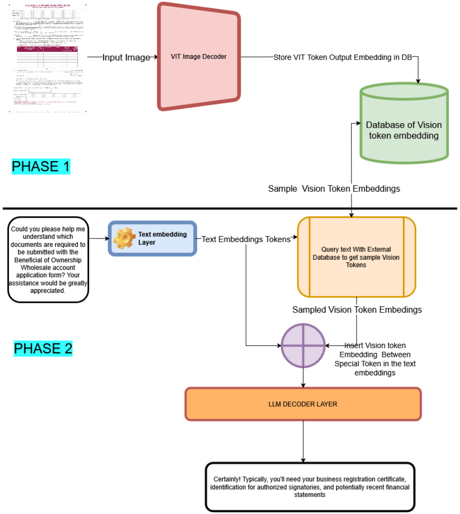

Figure 1: Workflow of the Posposed Method

Results: Performance on the MM-Advice Dataset

Table 1: Performance of Different Techniques on the MM-Advice Dataset For generation evalauted using Rouge(R1,R2,RL),Bleu(B1,B2,B3,B4) and Bert Score(Bert)

Method	Model	B1	B2	B3	B4	R1	R2	RL	Bert
Baseline	Gemma3-4B-it+SFT	0.343	0.217	0.149	0.107	0.308	0.117	0.240	0.851
	Qwen2.5-VL-3B-it+SFT	0.240	0.121	0.073	0.047	0.222	0.061	0.160	0.825
SFT + Image Retrieval+100% Missing Images	Gemma3-4B-it+Top_1 Image	0.338	0.212	0.145	0.105	0.309	0.116	0.238	0.850
	Gemma3-4B-it+Top_2 Image	0.341	0.212	0.143	0.102	0.306	0.112	0.234	0.850
	Qwen2.5-VL-3B-it+Top_1 Image	0.236	0.119	0.071	0.045	0.218	0.060	0.158	0.825
	Qwen2.5-VL-3B-it+Top_2 Image	0.237	0.119	0.070	0.046	0.218	0.059	0.157	0.824
Zero Shot+100% Missing Images	Qwen2.5-VL-3B-it	0.137	0.057	0.031	0.020	0.159	0.042	0.109	0.797
	Gemma3-4B-it	0.124	0.043	0.021	0.012	0.154	0.038	0.107	0.792
	Gemma3-4B-it+Top_1 Image	0.124	0.044	0.022	0.013	0.153	0.039	0.107	0.792
	Gemma3-4B-it+Top_2 Image	0.123	0.041	0.020	0.011	0.151	0.037	0.104	0.792
	Qwen2.5-VL-3B-it+Top_2 Image	0.139	0.058	0.032	0.020	0.159	0.042	0.110	0.797
	Qwen2.5-VL-3B-it+Top_1 Image	0.138	0.057	0.031	0.019	0.161	0.042	0.110	0.797
50% Missing Images Train + SFT+100% Missing Images Testing	Gemma3-4B-it+SFT	0.339	0.216	0.150	0.108	0.324	0.123	0.253	0.855
	Qwen2.5-VL-3B-it+SFT	0.281	0.169	0.114	0.082	0.272	0.095	0.210	0.840
SFT + Token Retrieval+100% Missing Images	Gemma3-4B-it+Top_150 Tokens	**0.358**	**0.232**	**0.161**	**0.116**	**0.334**	**0.129**	**0.259**	**0.860**

Results: Ablation Study on Token Sampling

Table 2: Abalation Study for different Token Sampling Techniques on the MM-Advice Dataset evaluated on Rouge, Bleu and Bert score. The best score for each Method is highloghted in bold

Method	Model	B1	B2	B3	B4	R1	R2	RL	Bert
Baseline	Gemma3-4B-it+SFT	0.343	0.217	0.149	0.107	0.308	0.117	0.240	0.851
	Qwen2.5-VL-3B-it+SFT	0.240	0.121	0.073	0.047	0.222	0.061	0.160	0.825
Img_Retrieval+Token Sampling+100% Missing Images	Qwen2.5-3B-it+SFT+top_1	0.247	0.129	0.079	0.051	0.234	0.068	0.169	0.829
	Qwen2.5-3B-it+SFT+top_2	0.248	0.131	0.081	0.053	0.235	0.070	0.169	0.829
	Gemma3-4B-it+top_2	0.341	0.217	0.149	0.106	0.311	0.118	0.241	0.851
	Gemma3-4B-it+top_1	0.347	0.221	0.152	0.109	0.313	0.119	0.242	0.852
Token Sampling+100% Missing Images	Qwen2.5-3B-it+SFT+30	0.247	0.131	0.081	0.054	0.235	0.070	0.170	0.829
	Qwen2.5-3B-it+SFT+60	0.243	0.127	0.078	0.052	0.234	0.069	0.168	0.829
	Qwen2.5-3B-it+SFT+100	0.245	0.128	0.079	0.052	0.237	0.069	0.169	0.829
	Qwen2.5-3B-it+SFT+120	0.243	0.127	0.077	0.051	0.235	0.068	0.169	0.829
	Qwen2.5-3B-it+SFT+150	0.244	0.126	0.078	0.051	0.233	0.067	0.168	0.829
	Gemma3-4B-it+30	0.353	0.228	0.158	0.114	0.329	0.127	0.257	0.856
	Gemma3-4B-it+60	**0.358**	0.231	**0.161**	**0.117**	**0.334**	**0.129**	**0.261**	0.859
	Gemma3-4B-it+120	0.356	0.230	0.159	0.114	0.331	0.126	0.257	**0.860**
	Gemma3-4B-it+150	**0.358**	**0.232**	**0.161**	0.116	**0.334**	**0.129**	0.259	**0.860**

Results: MM-IMDB Genre Classification Results

Table 3: Testing For Generalization For MM-IMDB dataset

Methods	File	Micro F1	Macro F1	Weighted F1
Img_Retrieval+Token Sampling+100% Missing Images	Gemma3-4B-it_+Top_1 Similar Image + Token Sampling	0.6121	0.1720	0.6092
	Gemma3-4B-it_+Top_2 Similar Image + Token Sampling	0.5920	0.0588	0.5839
Full Modality	Gemma3-4B-it	0.6842	0.3371	**0.6867**
	MoE MaxoutMLP	0.601	0.516	0.592
	GMU	0.630	0.541	0.617
	MM-GATBT	**0.685**	**0.645**	0.683
Token Sampling+100% Missing Images	Gemma3-4B-it + Top_100 Tokens	0.6465	0.2902	0.6443
	Gemma3-4B-it + Top_120 Tokens	0.6509	0.3755	0.6470
	Gemma3-4B-it + Top_150 Tokens	0.6430	0.1373	0.6400
	Gemma3-4B-it + Top_30 Tokens	<u>0.6585</u>	<u>0.4062</u>	<u>0.6559</u>
	Gemma3-4B-it + Top_60 Tokens	0.6563	0.1855	0.6546
Other Baseline+70% Missing Images	VILT	0.647	0.553	0.644

Publications

- **R. E. Zera Marveen Lyngkhoi**, Sriparna Saha. *Towards a Study on Exploring Missing Modalities through the Lens of Token Embeddings.***CIKM 2025: The 33rd ACM International Conference on Information and Knowledge Management**. **(**Communicated)
- **R. E. Zera Marveen Lyngkhoi**, Sarmistha Das, and Sriparna Saha. *Transforming Hours into Insights: A Next-gen Multimodal Summarization with Multimodal Output Framework for Financial Advisory Videos.***ACM MM 2025: The ACM International Conference on Multimedia**. **(**Communicated)

References I

- L. Tran, X. Liu, J. Zhou, and R. Jin, "Missing Modalities Imputation via Cascaded Residual Autoencoder," in *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 11405–11413.
- A. Kebaili, J. Lapuyade-Lahorgue, P. Vera, and S. Ruan, "AMM-Diff: Adaptive Multi-Modality Diffusion Network for Missing Modality Imputation," *arXiv preprint arXiv:2501.12840*, Jan. 2025.
- X. Hao, Y. Wang, and Z. Li, "Semi-Supervised Multimodal Image Translation for Missing Modality Imputation," in *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.
- T. Zhou, P. Vera, S. Canu, and S. Ruan, "Missing Data Imputation via Conditional Generator and Correlation Learning for Multimodal Brain Tumor Segmentation," *Pattern Recognition Letters*, vol. 158, pp. 12–21, Apr. 2022.

References II

- R. Wu, H. Wang, H.-T. Chen, and G. Carneiro, "Deep Multimodal Learning with Missing Modality: A Survey," *arXiv preprint arXiv:2409.07825*, Sep. 2024.
- J. E. Arevalo Ovalle, T. Solorio, M. MontesyGómez, and F. A. González, "Gated Multimodal Units for Information Fusion," *CoRR*, vol. abs/1702.01992, 2017.
- K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "BLEU: a Method for Automatic Evaluation of Machine Translation," in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2002, pp. 311–318.
- C.-Y. Lin, "ROUGE: A Package for Automatic Evaluation of Summaries," in *Text Summarization Branches Out: Proc. of the ACL Workshop*, 2004, pp. 74–81.
- T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating Text Generation with BERT," in *Proc. of ICLR*, 2020.

Thank You