MISSRAG: Addressing the Missing Modality Challenge in Multimodal Large Language Models

Vittorio Pipoli[*1,2], Alessia Saporita[*1,3], Federico Bolelli[1], Marcella Cornia[1], Lorenzo Baraldi[1], Costantino Grana[1], Rita Cucchiara[1], Elisa Ficarra[1]

[1]University of Modena and Reggio Emilia, Italy

[2]University of Pisa, Italy [3]University of Bologna, Italy

{name.surname}@unimore.it

Abstract

*Recently, Multimodal Large Language Models (MLLMs) have emerged as a leading framework for enhancing the ability of Large Language Models (LLMs) to interpret non-linguistic modalities. Despite their impressive capabilities, the robustness of MLLMs under conditions where one or more modalities are missing remains largely unexplored. In this paper, we investigate the extent to which MLLMs can maintain performance when faced with missing modality inputs. Moreover, we propose a novel framework to mitigate the aforementioned issue called retrieval-augmented generation for missing modalities (*MISSRAG*). It consists of a novel multimodal RAG technique alongside a tailored prompt engineering strategy designed to enhance model robustness by mitigating the impact of absent modalities while preventing the burden of additional instruction tuning. To demonstrate the effectiveness of our techniques, we conduct comprehensive evaluations across five diverse datasets, covering tasks such as audio-visual question answering, audio-visual captioning, and multimodal sentiment analysis. Our source code is available at* https://github.com/aimagelab/MissRAG.

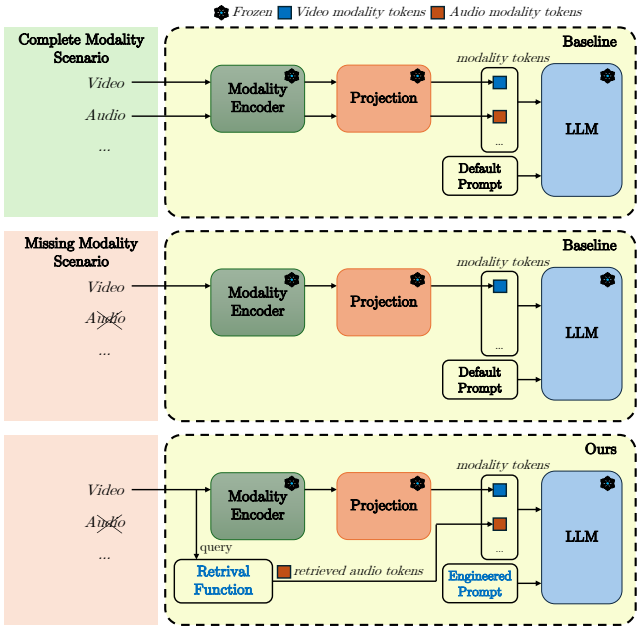


Figure 1. From top to bottom: (*i*) the complete modality scenario, where the MLLM receives all modality tokens, (*ii*) the missing modality scenario where one modality is absent, and (*iii*) our proposed multimodal RAG framework, enhanced with prompt engineering, to mitigate the missing modality problem.

1. Introduction

Multimodal learning, which integrates diverse data types such as text, images, audio, and video, is gaining prominence in Artificial Intelligence research. By leveraging these complementary modalities, multimodal models have achieved remarkable success in tasks like image captioning [2, 5, 16, 29], audio-video question answering [11, 17, 28], and cross-modal retrieval [1, 9, 35, 40].

Recently, with the advent of Large Language Models (LLMs), the field has witnessed a trend of aligning multimodal information with textual data to harness the strong language generation and zero-shot transfer abilities of LLMs [10, 21, 36, 53]. These models are designed to generate representations of non-textual modalities that are compatible with LLMs. This is achieved by creating token representations for non-textual modalities, which we refer to as *modality tokens* in this paper. These tokens maintain the same representational depth as textual tokens, allowing them to be fed directly into an LLM as input. Consequently, these modality tokens can be seamlessly incorporated into the LLM input through simple concatenation, enabling the LLM to interpret non-textual data and perform tasks such as

[*]Equal contribution. Authors are allowed to list their name first on their CVs.

multimodal captioning and multimodal question answering.

For example, X-LLM [10] and ChatBridge [53] integrate pre-trained modality-specific image, video, and audio encoders with LLMs by employing separate models such as Q-Former [30] or Perceiver [23]. In contrast, OneLLM [22] addresses the alignment challenge by employing modality-agnostic encoders, enabling direct mapping of inputs from eight modalities into the LLM embedding space. This design makes OneLLM one of the most versatile Multimodal Large Language Models (MLLMs) in terms of supported modalities. However, in real-world applications, multimodal systems often face the challenge of handling cases where certain data modalities are missing or incomplete [31]. This situation arises due to various factors such as sensor malfunctions, hardware limitations, privacy concerns, environmental interference, and data transmission issues. In the literature, these challenges are collectively referred to as the *missing modality problem*.

To address the missing modality problem, numerous studies have been conducted over the years [25, 33, 34, 37, 42, 52]. However, these works primarily focus on encoder-only architectures that learn to solve tasks in a closed-form manner by leveraging conventional training methods or fine-tuning existing backbones. In contrast, MLLMs contain billions of parameters, making fine-tuning computationally prohibitive. This has led to increased interest in parameter-efficient alternatives such as zero-shot learning, where models generalize without task-specific training. Recent efforts favor techniques that avoid updating LLM parameters, such as prompt engineering [8, 14, 39, 46], which involves manually modifying the textual prompt of an MLLM to induce a desired behavior, and Retrieval-Augmented Generation (RAG) [6, 26, 41, 49], which integrates a retrieval system to support the MLLM.

To the best of our knowledge, there are no studies that investigate the robustness of MLLMs under missing modality conditions. Hence, in this paper, we assess the robustness of MLLMs that accept at least two non-textual modalities as input, such as those addressing audio-video tasks [13, 22, 53]. We extend our evaluation by incorporating a textual modality, thereby conducting the first examination of such models on three-modality-input tasks like audio-video-text multimodal sentiment analysis.

Furthermore, we propose MISSRAG, a novel retrieval framework empowered with a tailored prompt engineering strategy that mitigates the missing modality problem in MLLMs without requiring additional fine-tuning. In particular, MISSRAG substitutes the missing modality with a candidate retrieved from a prototype pool by querying based on the available modalities. Specifically, our proposed multimodal RAG for addressing missing modalities is the first framework capable of concurrently managing three modalities and retrieving all possible combinations of single or multiple modality inputs. Also, the proposed framework is enhanced with tailored prompts to inform the model about the absent modality and condition the text generation process, as illustrated in Fig. 1. We validate the significance of our findings through extensive experiments on five multimodal datasets, covering a broad spectrum of tasks such as, audio-visual question answering (MUSIC-AVQA [27]), audio-visual captioning (VALOR-32K [11], CharadesEgo [43]), and audio-video-text sentiment analysis (MOSI [50], MOSEI [3]). In summary, our contributions are as follows:

- We are the first to assess the robustness of MLLMs under missing modality conditions across a wide range of tasks involving audio-video, and audio-video-text data;
- This is the first work in which RAG is employed to address the missing modality problem;
- We introduce MISSRAG, the first multimodal RAG framework that concurrently operates across three distinct modalities. This framework is capable of retrieving complementary audio-visual-text data for any given combination of audio, visual, and textual inputs;
- Through extensive experiments and ablation studies, we show that our proposed MISSRAG enhanced with the proposed prompt engineering strategy effectively mitigates the missing modality problem for MLLMs.

2. Related Work

Multimodal Learning and Missing Modalities. The integration of heterogeneous data streams in Deep Learning models is a demanding problem in multimodal learning [7, 19, 24, 28]. Among them, the *missing modality problem*, wherein one or more modalities may be absent during inference or even training, is one of the most challenging.

Contemporary studies [33, 34, 38, 51, 52] have been directed towards the development of multimodal frameworks capable of handling absent modalities. The SMIL approach [33] has been introduced to infer the latent features of data with incomplete modalities using Bayesian meta-learning. Zeng *et al.* [51] designed a tag-encoding mechanism that aids the training of Transformer encoders to cope with absent modalities. Ma *et al.* [34] explored multimodal Transformers under missing modality conditions, improving robustness by automatically learning optimal fusion strategies. However, these studies mainly focus on encoder-only models that are designed to solve tasks in a closed-form manner using standard training techniques or by fine-tuning existing backbones, an experimental setup that differs substantially from that addressed in this paper.

Multimodal Large Language Models. Early efforts to inject visual cues into LLMs [4, 30, 32] have recently been extended to additional modalities such as audio, video, and point clouds, with the goal of unifying multiple modalities within a single LLM framework. In this context,

X-LLM [10] pioneered the use of modality-specific Q-Formers and adapters to bridge pre-trained image, audio, and video encoders with frozen LLMs. Follow-up works such as ChatBridge [53] and AnyMAL [36] adopt similar designs but with varying connector modules, including Perceiver and linear layers. VideoLLaMA 2 [13] exploits a pretrained visual encoder in combination with a learnable Spatial-Temporal Convolution Connector to process visual information, while the audio modality is encoded via the BEATs model [12] and aligned to text with an MLP. Qwen2.5-Omni [47] jointly models perception and generation of text and speech across interleaved video and audio inputs, via a thinker-talker architecture and novel positional encoding. VITA [18] enables speech-to-speech dialogue while maintaining remarkable visual-language understanding capabilities. Also, OneLLM [22] removes the need for modality-specific encoders by introducing a universal encoder and projection module, supporting eight modalities.

In this work, we focus our experiments on three representative MLLMs (*i.e.*, ChatBridge, OneLLM, and VideoLLaMA 2) selected for their public availability and demonstrated robustness in audio-visual-text scenarios.

RAG and Prompt Engineering. Retrieval-Augmented Generation (RAG) [6, 26, 41, 49] has emerged as a powerful paradigm for enriching LLMs and MLLMs with external knowledge. By retrieving relevant snippets from a pre-built knowledge base and incorporating them into the prompt, RAG complements the intrinsic capabilities of generative models with dynamic, context-aware information from external information repositories. In this context, Lewis *et al.* [26] demonstrated the effectiveness of this approach by conditioning the generation on retrieved textual documents. More recently, Xu *et al.* [6] introduced VisualRAG, combining image retrieval with text generation for multimodal tasks. Despite these recent advancements, our proposed multimodal RAG-based framework is the first to operate concurrently across three distinct modalities.

Prompt engineering [8, 14, 39, 46] utilizes task-specific instructions to elicit desired behaviors without modifying model weights. Radford *et al.* [39] demonstrated that carefully designed prompts enable zero-shot learning, where models perform unseen tasks without labeled examples. Wei *et al.* [46] further introduced chain-of-thought prompting to encourage step-by-step reasoning. Building on its success, in this work we develop a tailored prompt engineering strategy to handle missing modalities within our multimodal RAG framework.

3. Proposed Method

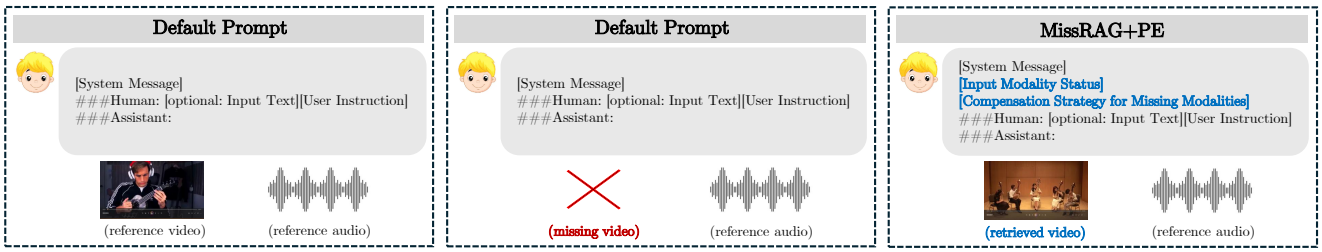
This work addresses the challenge of missing modalities in the field of multimodal learning, specifically for MLLMs. Since no previous literature has tackled the missing modality problem for MLLMs, we establish a benchmark to evaluate the robustness of these models across a wide range of tasks, including audio-visual question answering, audio-visual captioning, and audio-video-text sentiment analysis. In this setting, we evaluate the robustness of MLLMs under six missing modality scenarios: *Missing Audio* (MA), *Missing Video* (MV), *Missing Text* (MT), *Missing Audio-Video* (MAV), *Missing Audio-Text* (MAT), and *Missing Video-Text* (MVT). Additionally, we consider a *Complete* (C) scenario in which all modalities are available. Accordingly, in our experimental design, once a specific missing modality scenario is established, we remove the corresponding modalities from all samples in the analyzed dataset to simulate conditions of missing modalities with 100% of missing rate. Thus, we work in the worst-case scenario where all samples strictly adhere to the specified missing modality condition. Consequently, in a hypothetical scenario of lower missing rates, the overall performance would be higher.

Within this scope, our main contribution is a novel multimodal RAG framework for missing modalities called MISSRAG, the first RAG-based solution that simultaneously manages three different modalities, thereby mitigating the missing modality problem. MISSRAG exploits pre-trained contrastive backbones to query a pool of prototypes of the same modality as the missing ones, retrieving the corresponding modality tokens to ensure that the model consistently receives complete input pairs or triplets. Moreover, we enhance MISSRAG with a specific prompt engineering strategy that informs the model about which modalities are present and which are absent. It also forces the model to infer the likely context of the missing modality through a customized prompt. By integrating both techniques, we reconstruct the input data and induce the model to adapt its behavior in scenarios with missing modalities.

3.1. Preliminaries

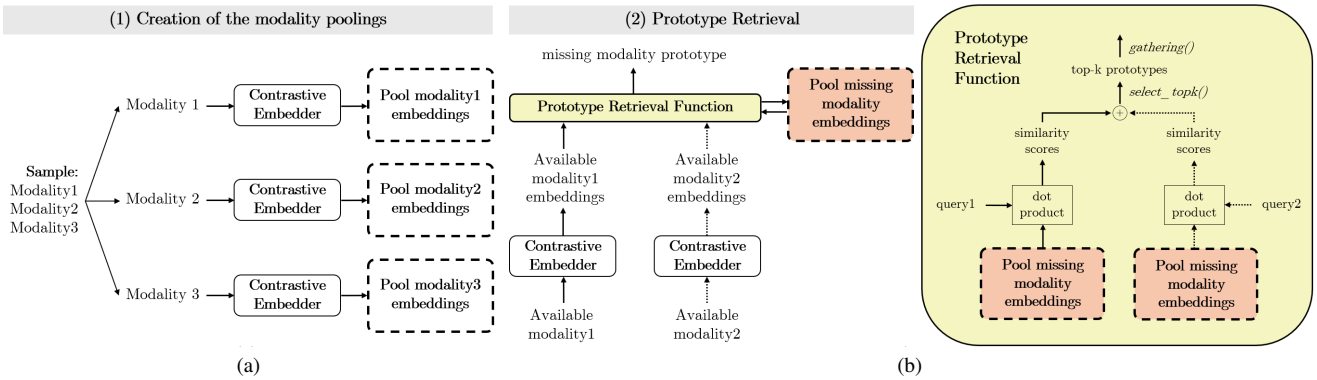
Problem Statement. In our setting, we employ multimodal datasets comprising M modalities, where $M \in \{2, 3\}$. To assess the robustness of our models under missing modality conditions, we omit one or two modalities at a time across all samples within each dataset. Consequently, for a dataset comprising M modalities, there exist $2^M - 1$ potential missing modality scenarios, each corresponding to a non-empty subset of the M modalities excluding the case where all modalities are absent. Specifically, for audio-visual question answering and audio-visual captioning datasets, the missing modality scenarios include *Missing Audio*, *Missing Video*, and *Complete*. In the case of audio-video-text sentiment analysis, we consider seven scenarios: *Missing Audio*, *Missing Video*, *Missing Text*, *Missing Audio-Video*, *Missing Video-Text*, *Missing Audio-Text*, and *Complete*.

Contrastive Backbone. A contrastive backbone is designed to map data from different modalities into a unified embedding space by pulling together the representa-



(a) (b) (c)

Figure 2. Overview of three different scenarios: (a) complete modality scenario where both reference video and reference audio are available; (b) missing video scenario without compensation for the missing modality; (c) missing video scenario where our proposed MISSRAG+PE approach retrieves a prototype video while employing a designed prompt to mitigate the impact of the missing modality.



(a) (b)

Figure 3. Overview of our MISSRAG framework with three modalities. (a) Creation of modality embeddings through a contrastive embedder, pooled for retrieval. (b) Retrieval of the top-k most similar prototypes by computing similarity scores between the embeddings of available modalities (*i.e.*, query) and the stored embeddings of the missing modality via dot product, then aggregated to obtain the missing modality representation. Dashed arrows indicate that the second modality may be unavailable.

tions of similar data points while pushing apart those of dissimilar ones. This process encourages the formation of distinct clusters in the shared embedding space, where semantically similar instances are grouped closely together. This approach facilitates effective cross-modal retrieval, as it allows for direct similarity computations between data points from disparate modalities. For example, elements with a high dot product are likely to represent similar concepts.

In our work, we use ImageBind [20] as contrastive backbone. It aligns seven modalities (*i.e.*, audio, image, video, text, IMUs, thermal, and depth) into a shared embedding space, making it useful for retrieval tasks where establishing semantic similarity between modalities is essential.

MLLM Backbone. For what concerns the MLLM, we select three publicly available models, namely OneLLM [22], VideoLLaMA 2 [13], and ChatBridge [53], due to their broad support to different input modalities and state-of-the-art zero-shot performance on audio-visual question-answering and audio-visual captioning tasks.

In particular, OneLLM, VideoLLaMA 2, and ChatBridge employ LLaMA-7B [44], Qwen2-7B [48], and Vicuna-13B [15], respectively, as their underlying LLM backbones, alongside a framework that maps non-textual modalities into the LLM embedding space. MLLMs are indeed engineered to translate each non-textual modality into a format interpretable by the LLM. Specifically, the LLM processes text input as token embeddings, where each token represents a subcomponent of a sentence and is embedded using an embedding layer that maps the token to a learnable representation of fixed dimension d_{model}. As a result, the input to the LLM has a dimensionality of (n_{tokens}, d_{model}). Moreover, to map each non-textual modality into the input embedding space of the LLM, a token representation must be generated, which we refer to as *modality tokens*. These modality tokens are created by employing a modality encoder and a projection module, which process the non-textual modality to produce a token representation that can be concatenated with the textual tokens in the LLM input.

MLLM Prompt Structure. The input prompts for MLLMs adhere to a specific structure designed to enhance model performance. Specifically, each input prompt consists of two primary components: the system message and the user instruction. An example of an input prompt is provided in Fig. 2a.[1] The system message provides contextual information that guides the behavior of the LLM, while the user instruction simulates a human request to the machine, which

[1] Additional details on the employed input prompts are reported in the supplementary material.

Task	Retr.	Miss. Scenario	Prompt
Audio-Video	✗	MA	The audio is missing. The assistant must use visual data to infer a probable audio context.
		MV	The video is missing. The assistant must use audio data to infer a probable visual context.
		C	Both video and audio are present.
	✓	MA	The assistant receives an approximate audio data and uses it to generate a response as accurately as possible.
		MV	The assistant receives an approximate visual data and uses it to generate a response as accurately as possible.
Audio-Video-Text	✗	MA	The audio is missing. The assistant must use visual and textual data to infer a probable audio context.
		MV	The video is missing. The assistant must use audio and textual data to infer a probable video context.
		MT	The input text is missing. The assistant must use audio and visual data to infer a probable textual context.
		MAT	The audio and input text are missing. The assistant must use visual data to infer a probable audio and textual context.
		MVT	The video and input text are missing. The assistant must use audio to infer a probable video and textual context.
		MAV	The video and audio are missing. The assistant must use textual data to infer a probable visual and audio context.
		C	Audio, visual and textual data are all present.
	✓	MA	The assistant receives an approximate audio data and uses it to generate a response as accurately as possible.
		MV	The assistant receives an approximate visual data and uses it to generate a response as accurately as possible.
		MT	The assistant receives an approximate input text data and uses it to generate a response as accurately as possible.
		MAT	The assistant receives an approximate audio and input text data and uses them to generate a response as accurately as possible.
		MVT	The assistant receives an approximate visual and input text data and uses them to generate a response as accurately as possible.
		MAV	The assistant receives an approximate visual and audio data and uses them to generate a response as accurately as possible.

Table 1. Prompt Engineering (PE) details for all missing modality scenarios and retrieval modes.

the LLM must respond to while also considering the information contained in the system message.

3.2. Multimodal RAG for Missing Modalities

In this study, we propose MISSRAG, the first multimodal RAG-based framework that simultaneously processes three distinct modalities. This framework is capable of retrieving complementary audio, visual, and textual data for any given combination of these input modalities. Specifically, when a modality is absent, we utilize the information from the available modalities to retrieve optimal candidates for substitution. This approach aims to achieve performance comparable to scenarios where all modalities are present, thereby mitigating the missing modality problem.

To implement this, we require a repository of prototypes for each modality that needs to be retrieved during the testing phase. In practice, we construct this prototype pool using the training set of each respective dataset independently. This approach does not make the proposed prototype retrieval technique restricted or sensitive to the tasks.

Subsequently, to establish an effective retrieval system, we employ a function that maps the samples of each modality into a shared embedding space where representations of similar concepts are proximally located. Consequently, we expect that within this embedding space, the modality representations of a given sample will exhibit greater similarity to one another than to those of different samples.

A pre-trained multimodal backbone facilitates this mapping by ensuring that semantically similar samples are closely aligned in the embedding space. Specifically, we use a contrastive backbone to generate key embeddings for each modality in the training set, as shown in Fig. 3, which can then be queried using corresponding embeddings derived from the available modalities in the test set.

We then calculate the dot product between these queries and the embeddings of the missing modality. The resulting similarity scores are concatenated, and the top-k highest scores are selected to retrieve the most relevant samples. If only one modality is available, there is no concatenation of similarity scores, and the process is executed two times to obtain candidates for each of the two missing modalities. Once retrieved, the modality tokens that correspond to the retrieved samples are computed and averaged. It is important to note that there are no modality tokens for the text modality, as it is the native modality of the LLM. Therefore, we concatenate multiple retrieved texts if k is greater than 1. The retrieved information is subsequently used to complete the missing sections of the LLM input prompt, ensuring that the input becomes fully comprehensive.

3.3. Prompt Engineering

In this work, we leverage Prompt Engineering (PE) techniques to improve the comprehension of our MLLM with respect to its input data, inducing it to cope with cases of missing modalities. In particular, our prompt engineering strategy has two objectives: informing the model about the input modality status and inducing it to mitigate the missing modality problem with a *Compensation Strategy for Missing Modalities* (CSMM) prompt.

Specifically, we append to the system message a string that describes the input modality status, composed of a concatenation of input descriptors for each of the possible input modalities of the task defined as:

`{Modality}: {Present/Prototype/Missing}.`

Taking an audio-video task as an example, the string used to indicate that the audio is original and the video is retrieved is "`Audio: Present; Video: Prototype`".

Subsequently, we append the compensation strategy, *i.e.*, a handcrafted string designed to induce behavior in the

LLM that helps the mitigation of the missing modality problem. These prompts are detailed in Table 1 for each missing modality scenario and retrieval mode. Specifically, the handcrafted prompt for the *Complete* (C) scenario is utilized in both retrieval and non-retrieval modes, as the latter scenario does not require any data retrieval.

3.4. Multimodal RAG and Prompt Engineering

Our proposal to mitigate the missing modality problem consists in our proposed multimodal RAG for missing modalities, called MISSRAG, enhanced with our proposed prompt engineering strategy. Such techniques aim to mitigate the missing modality problem by working on two different and independent levels. The former improves the input quality by substituting missing input with optimal candidates, while the latter is designed to induce the MLLM to change its behavior in front of missing or retrieved modalities. An enhanced visualization of the application of our method is depicted in Fig. 2c.

4. Experimental Results

We evaluate our proposed methodology in a zero-shot setting by applying it to three baseline models, ChatBridge, OneLLM, and VideoLLaMA 2. The only exception is VideoLLaMA 2 on the MUSIC-AVQA dataset, as it has been exposed to data from this benchmark during training. These baselines use their default system messages and user instructions and, by augmenting them with our proposed MISSRAG framework enhanced with the design prompt engineering strategy, we achieve remarkable performance gains. We assess the effectiveness of our methodology across three multimodal downstream tasks: question-answering, captioning, and sentiment analysis.

Audio-Visual Question Answering. We evaluate our methods on audio-visual question answering on MUSIC-AVQA [27], a dataset which contains 45K question-answer pairs. The problem requires a comprehensive multimodal understanding of both audio and visual cues.

Audio-Visual Captioning. We evaluate our methods on audio-visual captioning tasks on VALOR-32K [11] and CharadesEgo [43]. VALOR-32K is a large-scale dataset that contains 32K videos with rich audio-visual captions. CharadesEgo contains 7,860 first and third-person videos annotated with textual descriptions, making it suitable for audio-visual captioning.

Audio-Video-Text Sentiment Analysis. We evaluate our methods on audio-video-text sentiment analysis, testing them on CMU-MOSI [50] and CMU-MOSEI [3], where the goal is to predict the sentiment of the video. CMU-MOSI is a collection of 2,199 opinion video clips, and CMU-MOSEI is an extension of the CMU-MOSI dataset that contains 23,453 annotated video segments. For both datasets, each video segment is paired with a text describing spoken words and annotated with a negative, neutral, or positive label.

Metrics. To assess the performance on these diverse tasks, we employ accuracy for audio-video-text sentiment analysis and audio-visual question answering and the CIDEr metric [45] for the captioning task.

Evaluation Protocol. For captioning, we compute CIDEr using the COCO caption evaluation toolkit,[2] a publicly accessible tool that supports multiple captioning evaluation metrics. To assess the accuracy on the MOSI and MOSEI datasets, we implement a customized evaluation function designed to maximize conservativeness. Specifically, we categorize as missed any prediction that includes multiple labels, and we invert a label polarity if it is preceded by the negation term "not" within the prediction. For further details, please refer to the supplementary material.

4.1. Main Results

As shown in Table 2, our method, MISSRAG+PE, consistently improves ChatBridge in 22 out of 23 experiments, OneLLM in 21 out of 23 experiments, and VideoLLaMA 2 in 17 out of 23 experiments, demonstrating the effectiveness of our approach in missing modality scenarios, especially when the absent modality is the dominant one for the task. In fact, in audio-video tasks such as MUSIC-AVQA, VALOR-32K, and CharadesEgo, our method achieves significant gains over ChatBridge (+15.93, +5.90, +2.86), OneLLM (+7.18, +13.23, +1.30), and VideoLLaMA 2 (+0.54, +3.28, +2.60), in the most challenging scenarios where visual information is absent.

Regarding MOSI and MOSEI, our methods generally lead to improved performance across all evaluated MLLMs. However, to contextualize these results, it is important to emphasize that the models are assessed in a zero-shot setting and have not been explicitly trained on tasks analogous to sentiment analysis—tasks which demand complex reasoning across textual, visual, and auditory inputs. This leads to two potential issues: (*i*) a model may lack understanding of the task and thus perform poorly overall (*e.g.*, ChatBridge); or (*ii*) a model may struggle to reason across multiple modalities, and perform best when prompted exclusively with text—the most natural input for an MLLM (*e.g.*, VideoLLaMA 2 and OneLLM). Furthermore, our approach tries to align the performances of the missing modality scenarios with that of the *Complete* one, regardless of whether the latter leads to the optimal performance for the MLLM. Consequently, if our method does not yield improvements in every missing modality scenario, we do not attribute this to a shortcoming of the proposed methodology, but rather to inherent limitations of current MLLMs in addressing complex multimodal tasks such as sentiment analysis across au-

[2] https://github.com/tylin/coco-caption

Task	Audio-Video Question Answering			Audio-Video Captioning						Audio-Video-Text Sentiment Analysis													
Dataset	MUSIC-AVQA			VALOR32K			CharadesEgo			MOSI							MOSEI						
Metric	Accuracy			CIDEr			CIDEr			Accuracy							Accuracy						
Method	MA	MV	C	MA	MV	C	MA	MV	C	MA	MV	MT	MAV	MAT	MVT	C	MA	MV	MT	MAV	MAT	MVT	C
ChatBridge [53]	41.80	27.25	43.87	18.36	4.30	20.22	9.47	1.51	9.96	18.51	29.01	17.49	23.18	15.16	11.37	26.38	40.20	21.76	35.20	28.38	**39.02**	21.76	36.81
MISSRAG+PE	**45.14**	**43.18**	**46.10**	**19.09**	**10.20**	**20.66**	**10.85**	**4.37**	**10.25**	**24.78**	**30.09**	**18.22**	**23.76**	**16.18**	**17.93**	**28.12**	**41.21**	**40.95**	**37.05**	**40.46**	36.19	**37.33**	**38.94**
Improvement	+3.34	+15.93	+2.23	+0.73	+5.90	+0.44	+1.38	+2.86	+0.29	+6.27	+1.08	+0.73	+0.58	+1.02	+6.56	+1.74	+1.01	+19.19	+1.85	+12.08	-2.83	+15.57	+2.13
OneLLM [22]	48.89	39.48	49.86	20.41	3.34	26.40	2.98	1.20	3.76	48.54	73.91	40.23	75.88	38.34	41.69	56.41	30.82	**57.70**	40.22	**58.75**	38.14	46.10	40.09
MISSRAG+PE	**50.85**	**46.66**	**51.32**	**27.19**	**16.57**	**26.42**	**4.03**	**2.50**	**4.17**	**78.57**	**74.20**	**46.65**	**76.68**	**54.08**	**45.48**	**75.51**	**59.18**	51.66	**45.16**	52.35	**45.89**	**47.66**	**61.56**
Improvement	+1.96	+7.18	+1.46	+6.78	+13.23	+0.02	+1.05	+1.30	+0.41	+30.03	+0.29	+6.42	+0.80	+15.74	+3.79	+19.10	+28.36	-6.04	+4.94	-6.40	+7.75	+1.56	+21.47
VideoLLaMA 2 [13]	77.20	59.57	79.94	19.36	7.02	22.40	10.17	0.69	12.20	**46.06**	**43.59**	10.50	**74.34**	10.20	4.37	25.80	**41.36**	**44.24**	28.55	**53.17**	28.07	22.00	36.62
MISSRAG+PE	**77.53**	**60.11**	**79.97**	**21.42**	**10.30**	**22.64**	**11.96**	**3.29**	**12.62**	33.38	37.76	**23.76**	40.52	**11.02**	**23.34**	**34.69**	38.63	39.47	**29.20**	40.14	**30.31**	**27.15**	**39.17**
Improvement	+0.33	+0.54	+0.03	+2.06	+3.28	+0.24	+1.79	+2.60	+0.42	-12.68	-6.23	+13.26	-33.82	+0.82	+18.97	+8.89	-2.73	-4.77	+0.65	-13.03	+2.24	+5.15	+2.55

Table 2. Main results comparing the baseline versions of ChatBridge, OneLLM, and VideoLLaMA 2 with their enhanced counterparts incorporating our MISSRAG+PE method. In this experiment, we assess the robustness of the MLLMs under seven missing modality scenarios: MA (*Missing Audio*), MV (*Missing Video*), MT (*Missing Text*), MAV (*Missing Audio-Video*), MAT (*Missing Audio-Text*), MVT (*Missing Video-Text*), and C (*Complete*). All experiments are conducted using $k = 1$ retrieved elements. Best results in **bold**. Gray color indicates that the model has seen the dataset during training.

dio, video, and text. Notably, the *Complete* scenario never yields optimal performance for all MLLMs in this task.

In particular, ChatBridge consistently demonstrates relatively poor performance in this task. Indeed, qualitative analysis shows that its responses are often nonsensical compared to those generated by OneLLM and VideoLLaMA 2, highlighting its limitations. Nevertheless, our method generally contributes to performance improvements.

Conversely, VideoLLaMA 2 shows discrete performance when operating with text-only inputs, but underperforms in the multimodal setting. This suggests that while its underlying language model possesses the necessary capabilities, the overall multimodal framework struggles to effectively integrate and reason over multimodal inputs. As a result, our methodology encourages the model to behave similarly to the *Complete* scenario, which, in the case of VideoLLaMA 2, is suboptimal and may lead to a reduction in the final results.

On the other hand, OneLLM is the only model capable of achieving satisfactory performance in the *Complete* scenario when prompted and does not exhibit performance degradation in the presence of the audio modality. Consequently, when integrated with our proposed method, it exhibits substantial performance gains in both the *Missing Audio* and *Complete* scenarios—achieving improvements of +30.03 and +19.10 on the MOSI dataset, and +28.36 and +21.47 on the MOSEI dataset, respectively. However, similar to the other models, OneLLM experiences performance degradation both when the video modality is present and when the text is absent. Accordingly, as shown in Table 3, when our PE is used alone, yields the highest accuracy under the *Missing Video* scenario. Hence, by explicitly informing the MLLM of the absence of the video modality, the PE improves its reasoning ability, resulting in a 6.88-point increase in accuracy. Furthemore, a plausible explanation for the performance degradation observed when incorporating the video modality lies in the MLLMs visual encoder. Hence, video encoders utilized for sentiment analysis tasks extract highly specific visual features based on facial landmarks, as opposed to general-purpose visual encoders employed by our MLLM backbone. For what concerns the absence of text—the typically worst-case scenario—our method consistently improves OneLLM.

Finally, although the evaluated models natively struggle with audio-video-text sentiment analysis, these experiments provide valuable insights. They serve to validate existing literature and offer a robust benchmark for assessing the multimodal reasoning capabilities of current MLLMs across three heterogeneous input modalities.

5. Ablation Studies

Our ablation studies offer a thorough evaluation of the proposed MISSRAG framework and prompt engineering strategy. For consistency, experiments are conducted using OneLLM [22], which in general demonstrates greater robustness across various missing modality scenarios.

MISSRAG: on the Effect of the k Parameter. The objective of this ablation study is to investigate the impact of the hyperparameter k in MISSRAG, which determines the number of prototypes retrieved by our RAG-based framework. We conduct experiments both with and without isolating the MISSRAG from the PE component across all tasks and datasets discussed in this paper. The results, presented in Table 3, indicate that our solution consistently outperforms the baseline in 22 out of 23 cases. Moreover, enhancing our MISSRAG with our proposed PE yields performance improvements compared to utilizing the MISSRAG alone. These results suggest that the two techniques are complementary, thereby justifying their combined application. However, it is also important to note that if incorporating all modalities does not result in optimal performance for the MLLM, the most effective approach may be to employ only PE. This is because PE is designed to enhance the model behavior regardless of the available input

Task	Audio-Video Question Answering			Audio-Video Captioning						Audio-Video-Text Sentiment Analysis													
Dataset	MUSIC-AVQA			VALOR32K			ChardesEgo			MOSI							MOSEI						
Metric	Accuracy			CIDEr			CIDEr			Accuracy							Accuracy						
Method	MA	MV	C	MA	MV	C	MA	MV	C	MA	MV	MT	MAV	MAT	MVT	C	MA	MV	MT	MAV	MAT	MVT	C
OneLLM [22]	48.89	39.48	49.86	20.41	3.34	26.40	2.98	1.20	3.76	48.54	73.91	40.23	75.88	38.34	41.69	56.41	30.82	57.70	40.22	**58.75**	38.14	46.10	40.09
PE	**51.72**	41.81	**51.32**	23.79	3.77	**26.42**	3.25	1.22	**4.17**	68.66	**79.74**	44.90	76.11	39.48	42.72	**75.51**	48.12	**64.58**	**49.43**	47.69	31.79	41.73	**61.56**
MISSRAG ($k = 1$)	49.61	45.55	-	25.00	15.83	-	3.93	2.2	-	65.60	63.41	39.50	56.71	49.97	38.78	-	42.20	36.34	28.89	31.53	26.23	22.04	-
MISSRAG ($k = 3$)	49.67	45.69	-	22.87	17.13	-	3.99	2.17	-	63.70	63.85	46.79	53.35	40.96	44.46	-	41.43	31.70	26.89	26.74	31.36	22.84	-
MISSRAG ($k = 5$)	49.50	46.27	-	21.01	17.16	-	4.22	2.14	-	62.68	60.50	47.52	51.60	49.56	48.98	-	41.34	30.78	24.55	25.80	31.96	21.59	-
MISSRAG+PE ($k = 1$)	50.85	46.66	**51.32**	**27.19**	16.57	**26.42**	4.03	2.5	**4.17**	76.09	74.49	47.81	**76.68**	**54.08**	45.48	**75.51**	**59.18**	51.66	46.16	52.35	**45.89**	**47.66**	**61.56**
MISSRAG+PE ($k = 3$)	50.86	46.80	**51.32**	24.86	17.34	**26.42**	4.00	2.31	**4.17**	76.38	74.05	48.69	72.01	49.71	**51.02**	**75.51**	57.67	49.04	41.21	43.79	44.47	37.48	**61.56**
MISSRAG+PE ($k = 5$)	50.83	**47.29**	**51.32**	23.50	**17.49**	**26.42**	**4.25**	**2.60**	**4.17**	**76.53**	72.74	**49.98**	70.85	49.42	50.73	**75.51**	57.07	47.91	37.33	42.91	41.83	31.90	**61.56**
MISSRAG+PE ($k = 1$)†	50.71	46.57	51.32	26.73	16.37	26.42	3.97	2.46	4.17	78.43	73.85	48.25	72.68	54.96	45.34	75.51	55.42	49.33	48.89	52.16	45.78	46.51	61.56

Table 3. Ablation study on the impact of the number k of retrieved prototypes in MISSRAG conducted on OneLLM. The symbol † denotes that the prototype pool is constructed by merging the prototype pools from all datasets.

Task	Question Answering			Captioning		
Dataset	MUSIC-AVQA			VALOR32K		
Metric	Accuracy			CIDEr		
Method	MA	MV	C	MA	MV	C
OneLLM [22]	48.89	39.48	49.86	20.41	3.34	26.40
System Message	**51.72**	**41.81**	**51.32**	**23.79**	**3.77**	26.42
User Instruction	48.68	41.02	49.27	23.00	3.58	**27.30**

Table 4. Ablation study on system message *vs.* user instruction, performed on OneLLM. Best results in **bold**.

Task	Question Answering			Captioning		
Dataset	MUSIC-AVQA			VALOR32K		
Metric	Accuracy			CIDEr		
Method	MA	MV	C	MA	MV	C
OneLLM [22]	48.89	39.48	49.86	20.41	3.34	26.4
PE	**51.72**	**41.81**	**51.32**	**23.79**	**3.77**	**26.42**
PE w/o CSMM	49.68	40.79	51.01	23.06	3.69	26.27

Table 5. Ablation study on the prompt engineering compensation strategy, performed on OneLLM. Best results in **bold**.

modalities, whereas MISSRAG aims to improve the model performance to closely approximate that of the *Complete* modality scenario. Consequently, when the *Complete* scenario represents the optimal condition, performance gains are maximized. Otherwise, a performance decline may be observed. Nevertheless, this decline is not attributable to our methodology but rather stems from inherent limitations within the MLLM backbone (*i.e.*, OneLLM).

Robustness of MISSRAG. We design an experiment to evaluate the robustness of our MISSRAG in a scenario where the prototype pool is generic and not customized for each dataset. To this end, we create a unified pool by merging the prototype pools from all datasets and repeat the experiment with $k = 1$. The results, presented in the last row of Table 3, indicate that there is no significant deviation compared to the MISSRAG+PE ($k = 1$) case.

Prompt Engineering: in the System Message *vs.* in the User Instruction. A key research question is whether our PE technique is more effective when applied to the system message or the user instruction. To explore this, we evaluate both variants on the MUSIC-AVQA and VALOR-32K datasets, as presented in Table 4. The results demonstrate that applying our PE technique to the system message yields superior results compared to applying it to the user instruction across both tasks. This indicates that enhancing the initial system message with the PE method contributes to increased robustness against missing modalities.

Prompt Engineering: Compensation Strategy. The objective of this ablation study is to demonstrate that incorporating the *Compensation Strategy for Missing Modalities* (CSMM) immediately following the string indicating the input modality status within the system message leads to enhanced performance. To isolate the impact of CSMM from that of prototypes, we employ only PE in this setting. Specifically, within the input modality string, the possible values for each modality are `Present/Missing`, whereas the CSMM prompts, as detailed in Table 1, correspond to prompts without retrieval and are designed to induce the model to infer a plausible context for the missing modality using the available ones. Our results, presented in Table 5, indicate that, across both datasets, adding the instructions to infer missing modalities after the input modality status yields superior performance.

6. Conclusion

In this work, we have presented the first comprehensive study on the robustness of MLLMs under missing modality conditions. To address this challenge, we have introduced MISSRAG, a novel multimodal RAG-based framework, enhanced with a tailored Prompt Engineering (PE) strategy, designed to effectively handle missing modalities. Notably, our approach is the first to jointly manage three distinct modalities within a RAG-based paradigm. Extensive experiments across five diverse datasets spanning audio-visual question answering, audio-visual captioning, and audio-video-text sentiment analysis have demonstrated that the proposed solution significantly outperforms baseline methods. Specifically, MISSRAG enhances input quality by substituting missing information with contextually appropriate candidates, while PE improves model robustness to incomplete inputs. Combined, MISSRAG+PE consistently yields superior performance in complex multimodal scenarios.

Acknowledgments

This work has been supported by the EU Horizon project "ELLIOT - European Large Open Multi-Modal Foundation Models For Robust Generalization On Arbitrary Data Streams" (GA No. 101214398), by PNRR project "Fit4MedRob - Fit for Medical Robotics" funded by the Italian Ministry of University and Research, and by the University of Modena and Reggio Emilia and Fondazione di Modena through the "Fondo di Ateneo per la Ricerca - FAR 2024" (CUP E93C24002080007) and FARD-2024. The work also received funding from DECIDER, the European Union's Horizon 2020 research and innovation programme under GA No. 965193 and "AIDA: explAinable multImodal Deep learning for personAlized oncology" (Project Code 20228MZFAA).

References

[1] Hassan Akbari, Yin Yuan, Wei Feng, Wei Hu, Yanjun Wang, M Javad Roshtkhari, and Greg Mori. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. In *NeurIPS*, 2021. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, et al. Flamingo: a Visual Language Model for Few-Shot Learning. In *NeurIPS*, 2022. 1

[3] 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 *ACL*, 2018. 2, 6

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[5] Federico Bolelli, Lorenzo Baraldi, Federico Pollastri, and Costantino Grana. A Hierarchical Quasi-Recurrent approach to Video Captioning. In *IPAS*, 2018. 1

[6] Mirco Bonomo and Simone Bianco. Visual RAG: Expanding MLLM visual knowledge without fine-tuning. *arXiv preprint arXiv:2501.10834*, 2025. 2, 3

[7] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-End Referring Video Object Segmentation with Multimodal Transformers. In *CVPR*, 2022. 2

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, et al. Language Models are Few-Shot Learners. In *NeurIPS*, 2020. 2, 3

[9] Davide Caffagni, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Recurrence-Enhanced Vision-and-Language Transformers for Robust Multimodal Document Retrieval. In *CVPR*, 2025. 1

[10] Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, , and Bo Xu. X-LLM: Bootstrapping Advanced Large Language Models by Treating Multi-Modalities as Foreign Languages. *arXiv preprint arXiv:2305.04160*, 2023. 1, 2, 3

[11] Sihan Chen, Xingjian He, Longteng Guo, Xinxin Zhu, Weining Wang, Jinhui Tang, and Jing Liu. VALOR: Vision-Audio-Language Omni-Perception Pretraining Model and Dataset. *arXiv preprint arXiv:2304.08345*, 2023. 1, 2, 6

[12] Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. BEATs: Audio Pre-Training with Acoustic Tokenizers. In *ICML*, 2023. 3

[13] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*, 2024. 2, 3, 4, 7, 11

[14] Yew Ken Chia, Guizhen Chen, Luu Anh Tuan, Soujanya Poria, and Lidong Bing. Contrastive Chain-of-Thought Prompting. *arXiv preprint arXiv:2311.09277*, 2023. 2, 3

[15] Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality, 2023. 4, 11

[16] Jaemin Cho, Jiasen Lu, Dustin Schwenk, Hannaneh Hajishirzi, and Aniruddha Kembhavi. X-LXMERT: Paint, Caption and Answer Questions with Multi-Modal Transformers. In *EMNLP*, 2020. 1

[17] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. In *CVPR*, 2019. 1

[18] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Yuhang Dai, Meng Zhao, Yi-Fan Zhang, Shaoqi Dong, Yangze Li, Xiong Wang, et al. VITA: Towards Open-Source Interactive Omni Multimodal LLM. *arXiv preprint arXiv:2408.05211*, 2024. 3, 11, 12

[19] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-Modal Fusion Transformer for Video Retrieval. In *ECCV*, 2020. 2

[20] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. ImageBind: One Embedding Space To Bind Them All. In *CVPR*, 2023. 4

[21] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. ImageBind-LLM: Multi-modality Instruction Tuning. *arXiv preprint arXiv:2309.03905*, 2023. 1

[22] Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. OneLLM: One Framework to Align All Modalities with Language. In *CVPR*, 2024. 2, 3, 4, 7, 8, 11

[23] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General Perception with Iterative Attention. In *ICML*, 2021. 2

[24] Wonjae Kim, Bokyung Son, and Ildoo Kim. VILT: Vision-and-Language Transformer Without Convolution or Region Supervision. In *ICML*, 2021. 2

[25] Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. Multimodal Prompting with Missing Modalities for Visual Recognition. In *CVPR*, 2023. 2

[26] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*, 2020. 2, 3

[27] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to Answer Questions in Dynamic Audio-Visual Scenarios. In *CVPR*, 2022. 2, 6

[28] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. In *NeurIPS*, 2021. 1, 2

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 2022. 1

[30] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models, 2023. 2

[31] Paul Pu Liang and Louis-Philippe Morency. Tutorial on Multimodal Machine Learning: Principles, Challenges, and Open Questions. In *ICMI*, 2023. 2

[32] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved Baselines with Visual Instruction Tuning. In *CVPR*, 2024. 2

[33] Mengmeng Ma, Jian Ren, Long Zhao, Sergey Tulyakov, Cathy Wu, and Xi Peng. SMIL: Multimodal Learning with Severely Missing Modality. In *AAAI*, 2021. 2

[34] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are Multimodal Transformers Robust to Missing Modality? In *CVPR*, 2022. 2

[35] Tanvir Mahmud, Shentong Mo, Yapeng Tian, and Diana Marculescu. MA-AVT: Modality Alignment for Parameter-Efficient Audio-Visual Transformers. In *CVPR Workshops*, 2024. 1

[36] Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Tushar Nagarajan, Matt Smith, Shashank Jain, Chun-Fu Yeh, Prakash Murugesan, Peyman Heidari, Yue Liu, and et al. AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model. In *EMNLP*, 2024. 1, 3

[37] Vittorio Pipoli, Federico Bolelli, Sara Sarto, Marcella Cornia, Lorenzo Baraldi, Costantino Grana, Rita Cucchiara, and Elisa Ficarra. Semantically Conditioned Prompts for Visual Recognition under Missing Modality Scenarios. In *WACV*, 2025. 2

[38] Vittorio Pipoli, Alessia Saporita, Kevin Marchesini, Costantino Grana, Elisa Ficarra, and Federico Bolelli. IM-Fuse: A Mamba-based Fusion Block for Brain Tumor Segmentation with Incomplete Modalities. In *MICCAI*, 2025. 2

[39] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. In *Open-AI blog*, 2019. 2, 3

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 1

[41] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-Context Retrieval-Augmented Language Models. *TACL*, 11: 1316–1331, 2023. 2, 3

[42] Alessia Saporita, Vittorio Pipoli, Federico Bolelli, Lorenzo Baraldi, Andrea Acquaviva, and Elisa Ficarra. Tracing Information Flow in LLaMA Vision: A Step Toward Multimodal Understanding. In *CAIP*, 2025. 2

[43] Gunnar A Sigurdsson, Abhinav Gupta, Cordelia Schmid, Ali Farhadi, and Karteek Alahari. Charades-Ego: A Large-Scale Dataset of Paired Third and First Person Videos. *arXiv preprint arXiv:1804.09626*, 2018. 2, 6

[44] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023. 4, 11

[45] Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based Image Description Evaluation. In *CVPR*, 2015. 6

[46] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *NeurIPS*, 2022. 2, 3

[47] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2.5-Omni Technical Report. *arXiv preprint arXiv:2503.20215*, 2025. 3, 11, 12

[48] An Yang et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2024. 4, 11

[49] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing Reasoning and Acting in Language Models. In *ICLR*, 2023. 2, 3

[50] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal Sentiment Intensity Analysis in Videos: Facial Gestures and Verbal Messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 2, 6, 12, 13

[51] Jiandian Zeng, Tianyi Liu, and Jiantao Zhou. Tag-assisted Multimodal Sentiment Analysis under Uncertain Missing Modalities. In *SIGIR*, 2022. 2

[52] Jinming Zhao, Ruichen Li, and Qin Jin. Missing Modality Imagination Network for Emotion Recognition with Uncertain Missing Modalities. In *ACL*, 2021. 2

[53] Zijia Zhao, Longteng Guo, Tongtian Yue, Sihan Chen, Shuai Shao, Xinxin Zhu, Zehuan Yuan, and Jing Liu. ChatBridge: Bridging Modalities with Large Language Model as a Language Catalyst. *arXiv preprint arXiv:2305.16103*, 2023. 1, 2, 3, 4, 7, 11

MISSRAG: Addressing the Missing Modality Challenge in Multimodal Large Language Models

Supplementary Material

A. Additional Implementation Details

MLLM Backbones. In our experiments, we employ OneLLM [22], ChatBridge [53], and VideoLLaMA 2 [13] as our MLLMs. OneLLM maps non-textual modalities to token representations of shape $(n_{\text{tokens}}, d_{\text{model}})$ with $n_{\text{tokens}} = 30$ and $d_{\text{model}} = 4{,}096$, reflecting the representation depth of LLaMA-7B [44] and supporting a maximum input size of 2,048 tokens. In contrast, ChatBridge maps non-textual modalities to token representations of shape $(n_{\text{tokens}}, d_{\text{model}})$ with $n_{\text{tokens}} = 32$ and $d_{\text{model}} = 5{,}120$, corresponding to the representation depth of Vicuna-13B [15] and also supporting a maximum input size of 2,048 tokens. Finally, VideoLLaMA 2 is based on Qwen2 [48], operating with $d_{\text{model}} = 4{,}096$, and maps the video modality to $n_{\text{tokens}} = 676$ and the audio modality to $n_{\text{tokens}} = 1{,}496$.

Default System Message and User Instructions. To disclose our prompts, Table 6 provides the default system messages for each input modality combination and Table 7 provides the default user instructions for each task.

Enhanced Visualization of the Prompt Engineering Details. In Table 9, we provide an enhanced visualization of the PE details, also reported the main paper.

Custom Evaluation Metric. In our experiments, we employ a custom evaluation metric to assess the performance in audio-video-text sentiment analysis on the MOSI and MOSEI datasets. A pseudo-code version of the aforementioned metric is exemplified in Algorithm 1.

Algorithm 1 Custom evaluation function for classification.

```
TASK_CLASSES ← ["class_1", . . ., "class_n"]
correct ← 0
total ← 0
for each sample in samples do
    total ← total + 1
    label ← sample["label"]
    negative_label ← concat("not", label)
    prediction ← sample["prediction"]
    count_match ← 0
    for each ground_truth in TASK_CLASSES do
        if ground_truth is in prediction then
            count_match ← count_match + 1
        end if
    end for
    if count_match > 1 then
        continue
    else if (label is in prediction) and not(negative_label is in prediction)
    then
        correct ← correct + 1
    end if
end for
accuracy ← correct / total
```

B. Additional Quantitative Results

Beyond validating our approach on ChatBridge, OneLLM, and VideoLLaMA 2 as done in the main paper, we here provide additional experiments with other MLLMs, namely VITA [18], and Qwen2.5-Omni [47]. Specifically, VITA is based on the Mixtral 8×7B LLM and is trained using a bilingual instruction tuning strategy. Instead, Qwen2.5-Omni, based on the Qwen2.5-7B LLM, handles interleaved video and audio inputs using a thinker-talker architecture and a time-aligned positional encoding scheme.

Results are reported in Table 8 for the audio-visual question answering (*i.e.*, MUSIC-AVQA) and audio-visual captioning (*i.e.*, VALOR32K and CharadesEgo) tasks.[3] As shown, MISSRAG+PE consistently improves both VITA and Qwen2.5-Omni across all tasks and metrics. When applied to VITA, our method yields gains of up to +6.92, +3.70, and +0.43 on MUSIC-AVQA, VALOR32K, and CharadesEgo, respectively, when the visual information is missing (*i.e.*, MV). Similarly, for Qwen2.5-Omni, we observe consistent improvements of +4.92, +1.53, and +3.15 under the same MV scenario. These results further validate the generality and effectiveness of our approach across diverse MLLM architectures.

[3] For these MLLMs, we exclude MOSI and MOSEI due to the poor performance of the models on the audio-video-text sentiment analysis task.

Input Modalities	Prompt
	A chat between a curious human and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the human's questions,
Audio-Video	**combining visual and audio data.**
Audio-Video-Text	**combining visual, audio and textual data.**

Table 6. Default system messages for all modality combinations.

Task	Prompt
Audio-Video Question Answering	{Question} Answer the question using a single word or phrase.
Audio-Video Captioning	Provide a detailed description for the given video in one sentence.
Audio-Video-Text Sentiment Analysis	Input text: Text. Given the class set [ClassList] What is the sentiment of this video?

Table 7. Default user instruction prompt table for all tasks.

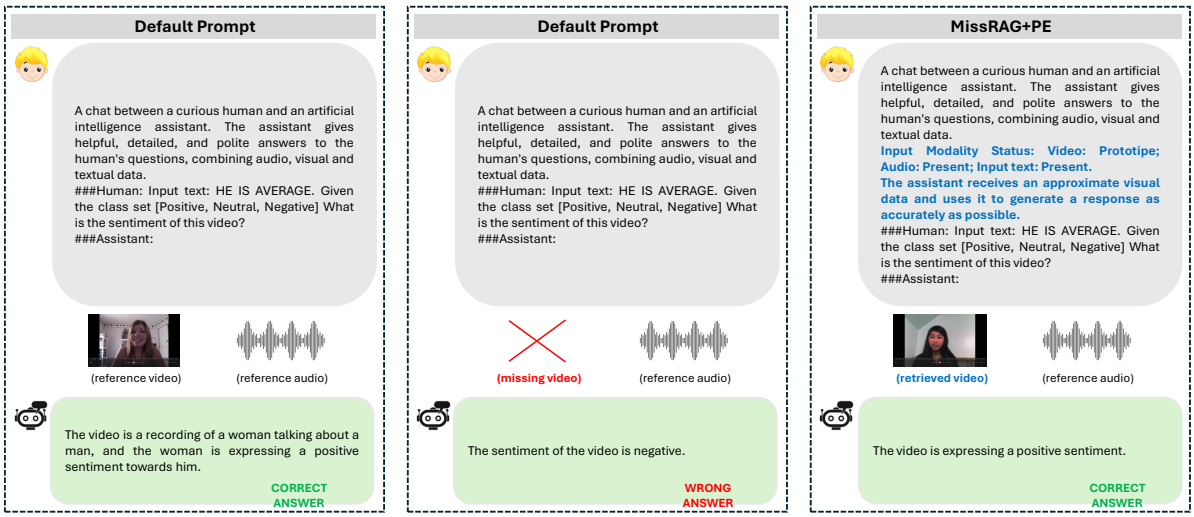
Dataset Metric	MUSIC-AVQA Accuracy			VALOR32K CIDEr			CharadesEgo CIDEr		
Method	MA	MV	C	MA	MV	C	MA	MV	C
VITA [18]	33.24	28.12	42.70	22.30	6.90	25.60	10.62	3.21	11.31
MISSRAG+PE	**39.38**	**35.04**	**42.80**	**24.30**	**10.60**	**25.70**	**11.64**	**3.64**	**11.33**
Improvement	+6.14	+6.92	+0.10	+2.00	+3.70	+0.10	+1.02	+0.43	+0.02
Qwen2.5-Omni [47]	62.77	51.78	62.42	8.75	7.38	10.10	9.30	2.24	14.15
MISSRAG+PE	**64.21**	**56.70**	**64.42**	**9.45**	**8.91**	**13.64**	**12.33**	**5.39**	**16.29**
Improvement	+1.44	+4.92	+ 2.00	+0.70	+1.53	+3.54	+3.03	+3.15	+2.14

Table 8. Results with additional state-of-the-art MLLMs. Best results in **bold**.

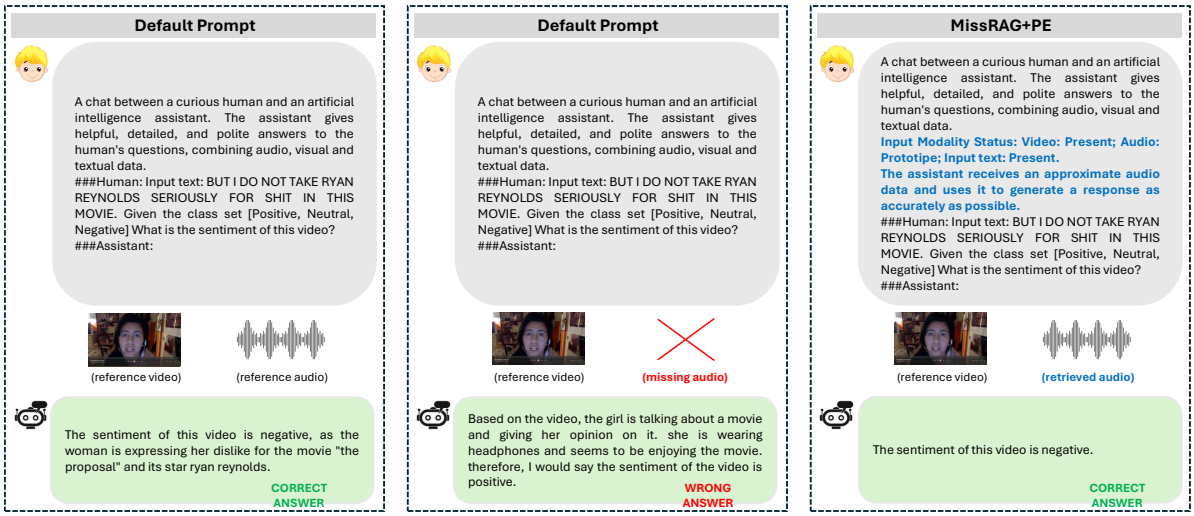
C. Qualitative Results

Fig. 4 illustrates sample qualitative results that demonstrate the effectiveness of our proposed framework in the context of audio-visual-text sentiment analysis, specifically evaluated on the MOSI dataset [50].

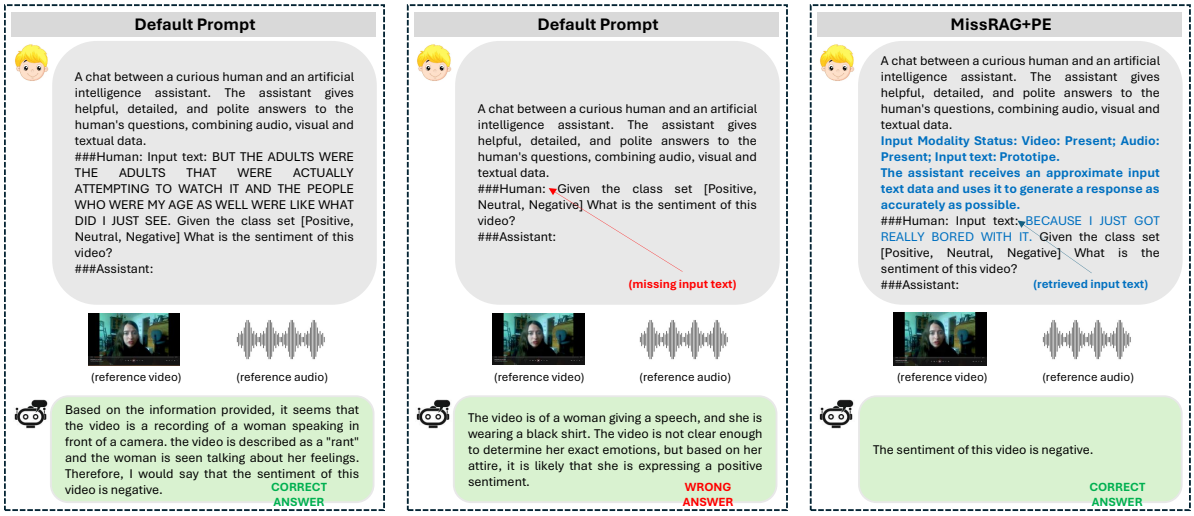
The subfigures in the leftmost column depict three examples illustrating scenarios where all modalities (*i.e.*, video, audio, and text) are present. Under these conditions, the MLLM model successfully interprets the sentiment of the reference video and provides the correct output. The subfigures in the middle column provide examples where one modality is absent. Specifically, the top example lacks video input, the middle example lacks audio, and the bottom example lacks the input text. In these cases, the model struggles to accurately determine sentiment, resulting in wrong answers. Finally, the subfigures in the rightmost column demonstrate how our proposed approach (*i.e.*, MISSRAG+PE) effectively addresses the challenge of missing modalities. By retrieving an appropriate prototype and employing a carefully designed prompt, the MLLM is better equipped to interpret the available inputs, thus mitigating the effects of the missing modality input.



(a) Missing Video Scenario



(b) Missing Audio Scenario



(c) Missing Text Scenario

Figure 4. Qualitative results on the MOSI dataset [50] evaluated under three missing modality scenarios using OneLLM as underlying model. From top to bottom, the scenarios are: missing video, missing audio, and missing text. In each scenario, the subfigures are organized into three columns. The leftmost column depicts the baseline method under the complete scenario; the middle column shows the baseline method under the missing modality scenario; and the rightmost column illustrates our proposed technique, namely MISSRAG+PE.

Table 9. Prompt Engineering (PE) details for all missing modality scenarios and retrieval modes.

Task	Retr.	Missing Scenario	Prompt
Audio-Video	✗	MA	The audio is missing. The assistant must use visual data to infer a probable audio context.
		MV	The video is missing. The assistant must use audio data to infer a probable visual context.
		C	Both video and audio are present.
	✓	MA	The assistant receives an approximate audio data and uses it to generate a response as accurately as possible.
		MV	The assistant receives an approximate visual data and uses it to generate a response as accurately as possible.
Audio-Video-Text	✗	MA	The audio is missing. The assistant must use visual and textual data to infer a probable audio context.
		MV	The video is missing. The assistant must use audio and textual data to infer a probable video context.
		MT	The input text is missing. The assistant must use audio and visual data to infer a probable textual context.
		MAT	The audio and input text are missing. The assistant must use visual data to infer a probable audio and textual context.
		MVT	The video and input text are missing. The assistant must use audio to infer a probable video and textual context.
		MAV	The video and audio are missing. The assistant must use textual data to infer a probable visual and audio context.
		C	Audio, visual and textual data are all present.
	✓	MA	The assistant receives an approximate audio data and uses it to generate a response as accurately as possible.
		MV	The assistant receives an approximate visual data and uses it to generate a response as accurately as possible.
		MT	The assistant receives an approximate input text data and uses it to generate a response as accurately as possible.
		MAT	The assistant receives an approximate audio and input text data and uses them to generate a response as accurately as possible.
		MVT	The assistant receives an approximate visual and input text data and uses them to generate a response as accurately as possible.
		MAV	The assistant receives an approximate visual and audio data and uses them to generate a response as accurately as possible.