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.

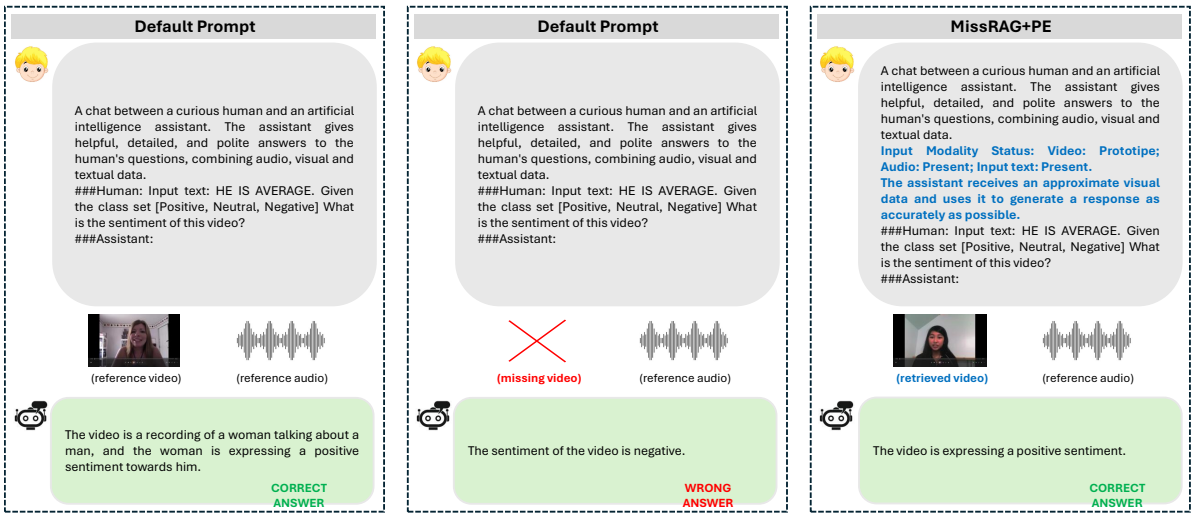
Dataset Metric	MUSIC-AVQA Accuracy			VALOR32K CIDEr			CharadesEgo CIDEr		
Method	MA	MV	C	MA	MV	C	MA	MV	C
VITA [18]	33.24	28.12	42.70	22.30	6.90	25.60	10.62	3.21	11.31
MISSRAG+PE	**39.38**	**35.04**	**42.80**	**24.30**	**10.60**	**25.70**	**11.64**	**3.64**	**11.33**
Improvement	+6.14	+6.92	+0.10	+2.00	+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**.

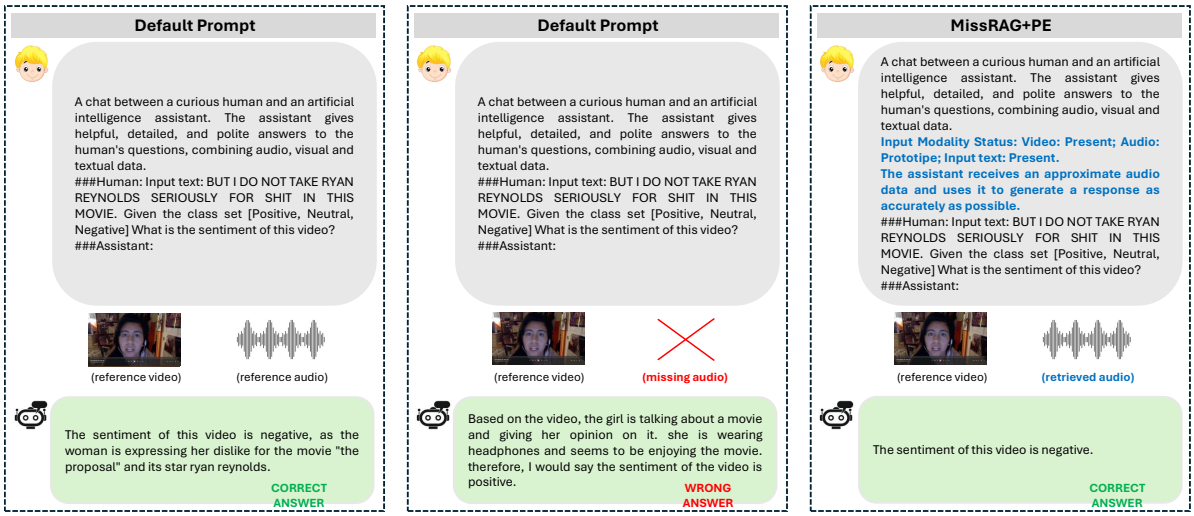
C. Qualitative Results

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].

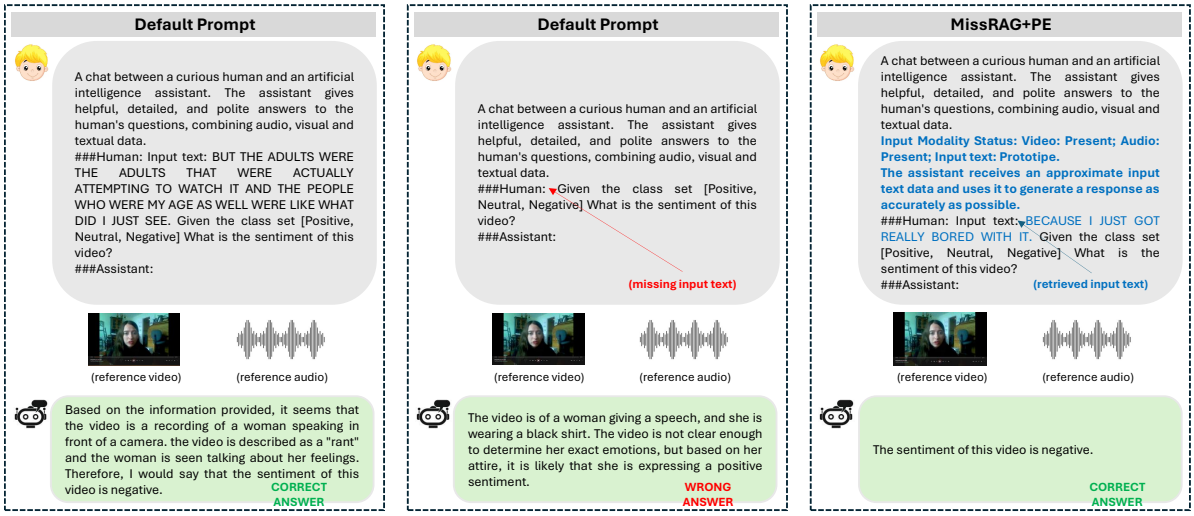
The subfigures in the leftmost column depict three examples illustrating scenarios where all modalities (*i.e.*, video, audio, and text) are present. Under these conditions, the MLLM model successfully interprets the sentiment of the reference video and provides the correct output. The subfigures in 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.