OmnixR: Evaluating Omni-Modality Language Models on Reasoning across Modalities

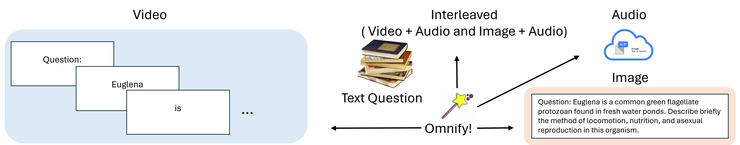
Lichang Chen[1,3], Hexiang Hu[1], Mingda Zhang[1], Yiwen Chen[2], Zifeng Wang[2], Yandong Li[1], Pranav Shyam[2], Tianyi Zhou[3], Heng Huang[3], Ming-Hsuan Yang[1], Boqing Gong[1]

[1]Google Deepmind [2]Google [3] University of Maryland, College Park

Motivations

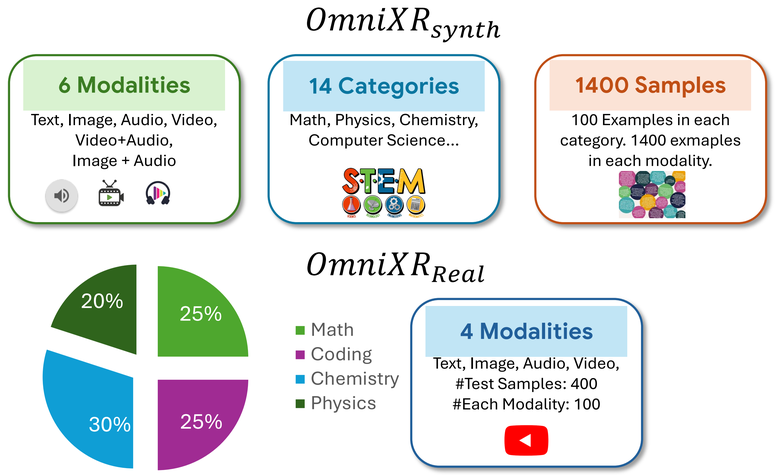
- **Evaluation Gaps.** Current benchmarks for Omni-modality Language Models (OLMs) like GPT-4o and Gemini focus on single- or dual-modality tasks (e.g., image+text), failing to evaluate the comprehensive cross-modal reasoning needed for real-world scenarios where inputs span multiple modalities (video, image, audio, text, etc.).
- **Performance Inconsistencies.** OLMs exhibit inconsistent reasoning and answers when processing the same question across different modalities (e.g., text vs. image vs. audio), exposing weaknesses in their ability to integrate and reason holistically across diverse inputs, which undermines their reliability.

Omnify! Transfer text data into other Modalities

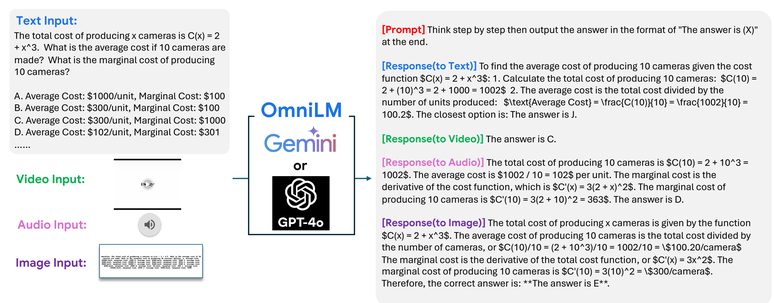


- **Purpose of Omnify!:** Omnify! is a scalable, low-cost method developed to transform text-based inputs into various modalities, such as audio, images, videos, and their combinations, enabling the creation of diverse, multi-modal datasets.
- **Practical Implementation:** Omnify! employs straightforward techniques like programmatic text rendering for images, text-to-speech for audio, and sequential frame assembly for videos, prioritizing simplicity and fidelity over complex generative models to maintain quality and focus on reasoning evaluation.
- **Enhancing Real-World Relevance:** Beyond synthetic data, Omnify! supports the generation of mixed-modality scenarios (e.g., video + audio), mimicking complex real-world inputs, which tests the ability to integrate and reason across interleaved modalities effectively.

OmnixR dataset



Behavior Inconsistency



1. **Variable Responses Across Modalities:** OLMs, such as Gemini-1.5-Flash, exhibit inconsistent reasoning and answers when processing the same question presented in different modalities (e.g., text, image, audio, or video), highlighting a lack of uniform understanding and integration across input types.

Main Results

Table 1: **Results on OmnixR-Synth** show different mixed modalities evaluations, including text, image, audio, video. Each modality (Image/Audio/Video) combines two input sources: the 'Question' provided by the respective image, audio, or video modality, and the 'CoT instruction' provided by the text. The numbers in green font, following the downward arrows, show the drops compared to the pure text input.

	Modality						
	Text	Image		Audio		Video	
Model	Perf.	Acc.	ΔAcc.	Acc.	ΔAcc.	Acc.	ΔAcc.
Gemini 1.5							
Pro	77.5	57.3	20.2↓	56.6	20.9↓	36.3	41.2↓
Flash	69.9	36.3	33.6↓	53.9	16.0↓	15.1	54.8↓
Claude							
Opus	77.7	26.9	50.8↓	-	-	-	-
Sonnet	77.4	18.8	58.6↓	-	-	-	-
Haiku	72.5	9.9	62.6↓	-	-	-	-
GPT							
4o	71.5	60.1	11.4↓	-	-	53.1	18.4↓
4o-mini	72.6	48.5	24.1↓	-	-	18.6	54.0↓

ETA Prompting

1. **Purpose:** Extract-Then-Answer(ETA) prompting is designed to enhance model performance on synthetic multi-modal tasks by decomposing them into a two-step process: first extracting the text from non-text modalities (image, audio, video), then reasoning over it.
2. **Method:** The prompt "Please extract the text from image/audio/video" is given first. The extracted text is then passed into the same OLM with a Chain-of-Thought (CoT) instruction to elicit the models' pure text reasoning.
3. **Effectiveness:** ETA prompting significantly boosts model accuracy and consistency on Omni×Rsynth, especially for modalities like images and audio where direct CoT reasoning often fails.