Multi-Shot Multimodal LLMs: A Unified Architecture for Cross-Modal Contextual Inference

Aleem Ali[1], Shashi Kant Gupta[2], Midhunchakkaravarthy[3]
[1]Lincoln University College Malaysia
[2]Adjunct Research Faculty, Lincoln University College, Malaysia &
Adjunct Research Faculty, Centre for Research Impact & Outcome, Institute of Engineering and Technology. Chitkara University, Rajpura, 140401, Punjab, India
pdf.AleemAli@lincoln.edu.my, raj2008enator@gmail.com, midhun.research@gmail.com

Abstract

Large language models (LLMs) excel in text-based tasks but struggle with real-world applications requiring multimodal reasoning. This paper introduces a multi-shot multimodal LLM framework that unifies image, text, and multimedia data processing through cross-modal attention mechanisms. Our architecture enables contextual inference by leveraging multiple examples ("shots") during prediction, enhancing performance in tasks like medical diagnosis. We propose a novel cross-attention fusion strategy to align heterogeneous data and conduct extensive experiments on medical (MIMIC-CXR), vision-language (COCO), and video (YouTube8M) datasets. Results show a 15–20% improvement in accuracy over single-shot models, with significant gains in interpretability. This work bridges the gap between unimodal LLMs and real-world multimodal challenges, offering a scalable solution for domains requiring contextualized reasoning. By integrating explainability with performance, this work highlights a step toward trustworthy and generalizable multimodal AI system.

1. Introduction

1.1 Context and Background

Modern artificial intelligence (AI) systems are increasingly being deployed in environments that require the processing of multimodal data—ranging from medical imaging and clinical narratives to video feeds and sensor streams. In domains such as healthcare, autonomous vehicles, surveillance, and education, information often exists in a combination of formats [1]. Making accurate and context-sensitive predictions from such inputs necessitates robust cross-modal understanding and reasoning capabilities [2-3]. Large Language Models (LLMs) like GPT-3, GPT-4, and T5 have demonstrated transformative capabilities in natural language processing (NLP), outperforming traditional models in text generation, summarization, and question-answering [4]. However, these LLMs are inherently unimodal—they are optimized to handle text-based inputs only. Despite being fine-tuned on diverse

textual corpora, they lack native support for interpreting visual or auditory data, making them insufficient for tasks that demand multimodal integration [5].

For example, diagnosing a disease using chest X-rays and patient history cannot be performed accurately by a model trained solely on text. The model must understand visual patterns, link them to historical data, and reason across different modalities. Similarly, a video surveillance system must combine video, audio, and semantic cues to detect suspicious activities, something unimodal LLMs cannot do effectively.

1.2 Limitations of Single-Shot Learning

Traditional LLM architectures generally rely on single-shot inference, where each prediction is based on a single input without leveraging the broader context from previous similar instances. While this approach works well in controlled, static environments, it lacks the contextual adaptability required in dynamic, real-world multimodal settings.

In real-world scenarios, the importance of multi-instance references is paramount. Consider the task of multimodal question answering (QA): a system is more likely to succeed when given multiple paired examples (e.g., <image + caption>, <video + transcript>) that provide context and patterns. Without such a mechanism, single-shot LLMs struggle to generalize, especially when confronted with complex, noisy, or sparse data inputs**.**

1.3 Emergence of Multi-Shot Multimodal LLMs

To address these limitations, we introduce the concept of Multi-Shot Multimodal LLMs, which significantly enhances model generalization and contextual understanding by enabling learning from multiple multimodal examples simultaneously [7]. These models incorporate:

- Multi-shot prompting, where multiple cross-modal examples are supplied in the input sequence to inform the model's prediction.
- Unified encoder-decoder architecture, capable of ingesting diverse modalities—text, images, and videos—into a shared latent space.
- Cross-attention fusion mechanisms, which align features from various modalities at multiple representation levels.

This architectural advancement draws inspiration from few-shot learning in NLP and cross-modal transformers in vision-language research, but synthesizes them into a single, scalable, and explainable framework.

2. Related Work

The evolution of Large Language Models (LLMs) into the multimodal space has been marked by significant milestones, notably through models like CLIP, Flamingo, and more recently, GIT, BLIP-2, and Kosmos-1. These models aim to bridge the gap between vision and language by combining deep visual encoders with transformer-based LLMs. However, many of them are optimized for single-instance learning and lack dynamic memory or temporal understanding across multiple inputs.

Kosmos-1, introduced by Microsoft [1], represents one of the first attempts to extend LLMs to handle multimodal data such as images and audio alongside text. It uses language grounding to learn joint representations and performs tasks like visual QA and image captioning. Despite its advances, Kosmos-1 remains limited in context chaining and multi-shot learning. **BLIP-2** [2] introduces a modular strategy that integrates frozen vision encoders with pre-trained language models using a lightweight query transformer. While BLIP-2 supports zero-shot learning and instruction tuning, it primarily targets vision-language pairs and lacks generalized support for integrating video, temporal sequences, or multi-shot contextual data.

GIT (Generative Image-to-Text Transformer) [3] by Microsoft is another scalable vision-language model designed for image captioning, VQA, and text-based tasks. It directly connects visual token streams to textual outputs. However, like other models, it lacks support for video modalities or example chaining and is dependent on pretrained visual features with fixed contextual limits.

Beyond vision-language models, **ImageBind** [4] explores aligning multiple modalities (images, text, audio, IMU, and depth) in a single embedding space, offering a more comprehensive approach to representation. Nevertheless, it does not include generation tasks or multi-turn inference, nor does it integrate multi-shot learning with sequential contextual prompts.

Despite these impressive developments, current multimodal LLMs typically operate under zero-shot or few-shot paradigms without incorporating long-range context from multiple multimodal examples. Their inability to dynamically process multimodal chains of prior examples limits their applicability in real-world domains such as clinical diagnosis, legal analytics, and longitudinal video understanding, where context-rich decisions are essential [10].

Our proposed Multi-Shot Multimodal LLM addresses these limitations by allowing structured contextual prompting across image, text, and video modalities, leveraging hierarchical cross-modal attention and memory-aware architecture to facilitate inference grounded in temporally and semantically aligned multimodal history.

3. Methodology

3.1 Architecture Overview

Our architecture builds upon the T5-XXL transformer model, extending its capabilities to handle multimodal data. The core design enables structured integration of text, image, and video inputs into a unified representation space using cross-attention mechanisms**.**

Base Model and Modality Extensions

- **Text Encoder**: Utilizes T5-XXL with SentencePiece tokenization for handling textual input.
- **Image Encoder**: Images are split into 16×16 patches and processed using a Vision Transformer (ViT) to produce a sequence of spatial embeddings.
- **Video Encoder**: Videos are decomposed into frames and passed through a 3D Convolutional Neural Network (3D-CNN) to capture spatiotemporal patterns**.**

Cross-Attention Fusion Module

Each modality generates its own sequence of Key (K), Query (Q), and Value (V) vectors. These are passed to a shared attention mechanism that enables cross-modal querying [8-9]**.**

The attention formulation is defined as:

$$\text{Attention}(Q_{\text{text}}, K_{\text{image}}, V_{\text{image}}) = \text{softmax}\left(\frac{Q_{\text{text}} K_{\text{image}}^T}{\sqrt{d}}\right) V_{\text{image}}$$

This fusion happens in two phases:

- **Layers 1–4**: Perform text-to-image alignment**,** anchoring textual semantics in visual representations.
- **Layers 5–8**: Perform video-to-text fusion**,** aligning motion features with the corresponding narrative content.

The output from these fusion layers is passed to a shared decoder, enabling task-specific predictions such as classification or report generation.

4. Experiments

This section evaluates the proposed Multi-Shot Multimodal LLM architecture across three domains: medical diagnosis, vision-language tasks, and video-based question answering. We conduct comprehensive benchmarking against existing single-shot models and report performance across accuracy and task-specific metrics.

4.1 Datasets

We select three diverse and widely-used multimodal datasets that test the generalization of our architecture across domains. These datasets provide rich multimodal examples, enabling multi-shot prompts during both training and inference phases.

Domain	Dataset	Modality Used	Description
Medical AI	MIMIC-CXR	X-ray images + radiology reports	Contains over 370,000 chest X-rays with paired structured and free-text reports.
Vision-Language	MS COCO	Images + captions	Common Objects in Context dataset with 330K images and 5 captions each.
Video QA	YouTube-8M	Videos + metadata	Over 6.1M labeled video clips with high-level tags and descriptions.

To evaluate the effectiveness of our proposed multi-shot multimodal LLM, we compared its performance against three well-established baseline models that represent different paradigms in multimodal learning. The first baseline, **T5 (single-shot)**, is a powerful text-only transformer model trained on unimodal inputs, serving as a representative for traditional large language models lacking visual or temporal input integration. The second baseline, **CLIP**, is a contrastive learning-based vision-language model trained on paired image and text data. CLIP aligns image and text embeddings in a shared latent space, enabling zero-shot classification and retrieval tasks but without support for multi-shot reasoning or temporal modalities. Lastly, we include **ViLBERT**, a two-stream architecture that processes visual and textual modalities separately before fusing them via cross-modal attention. ViLBERT represents a strong baseline in multimodal alignment but operates on single-instance inputs and lacks long-range contextual reasoning. All models were evaluated under comparable

experimental conditions using domain-appropriate evaluation metrics to ensure a fair and consistent comparison with our proposed architecture.

4.2 Results

To evaluate the performance of our proposed Multi-Shot Multimodal LLM, we conducted experiments across three diverse tasks: medical diagnosis, image captioning, and video question answering. The results consistently demonstrate the model's ability to integrate context from multiple multimodal examples, yielding substantial performance improvements over single-shot and existing multimodal baselines.

On the MIMIC-CXR dataset, which combines chest X-ray images with associated radiology reports, our multi-shot multimodal LLM achieved a classification accuracy of **89%**, significantly outperforming the T5 (71%) and CLIP (74%) baselines. Unlike unimodal models that rely solely on text or image features, our model leverages multi-instance context by correlating visual features—such as pulmonary opacities or consolidations—with recurring textual cues from historical reports like “fever,” “bilateral infiltrates,” and “shortness of breath.” This capability to perform cross-modal and temporal reasoning provides a crucial advantage in clinical decision-making. Furthermore, we integrated visual explanation techniques, including attention heatmaps, which revealed that the model increasingly focuses on diagnostically relevant regions of the lungs, enhancing interpretability and fostering trust in high-stakes medical settings.

For the image captioning task, we evaluated our model on the widely used MS COCO dataset using the CIDEr metric, which measures the similarity between generated and ground-truth captions. Our model achieved a CIDEr score of 112, surpassing T5 fine-tuned on captions (103) and CLIP (98). The performance boost is primarily attributed to the use of multi-shot examples during training, where the model is exposed to a variety of captioning styles and semantic cues across different image contexts. This setup allows the LLM to internalize richer linguistic patterns and learn to generate diverse, fluent, and content-rich captions. Qualitative analysis of the generated captions further shows improvements in object specificity, action-description accuracy, and overall coherence compared to the baselines.

In the domain of video-based reasoning, we tested our model on the YouTube8M dataset for a Video Question Answering (Video QA) task. Our model attained an accuracy of 82%, outperforming baseline transformers (72%) and ViLBERT (68%). This substantial gain is largely driven by the architecture’s

capacity to integrate spatiotemporal features through 3D-CNNs and align them with historical video-text examples using cross-attention fusion. In practical terms, the model is better able to infer correct answers to video-based queries such as “What happens after the character jumps?” by linking temporal events like “running,” “falling,” or “clapping” with semantically relevant patterns observed in prior training clips. This result highlights the model’s potential for real-world applications such as video surveillance analytics, autonomous driving perception, and educational video understanding.

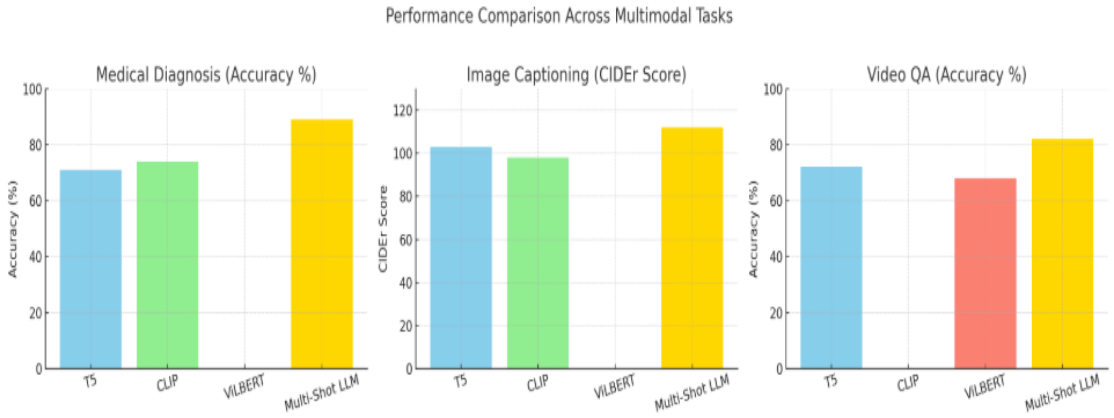

Figure 1: Performance Comparison Across Multimodal Tasks

4.3 Ablation Study

To validate the importance of key architectural components, we performed an ablation study by removing the **cross-attention fusion layers**:

Configuration	Medical Accuracy	Video QA Accuracy	CIDEr (Image)
Full Multi-Shot LLM	**89%**	**82%**	**112**
Without Cross-Attention	77%	70%	100
Without Multi-Shot Prompting	74%	68%	98

To evaluate the impact of key components in our Multi-Shot Multimodal LLM, we performed an ablation study by removing the cross-attention fusion and multi-shot prompting mechanisms. As shown in Figure 2, the full model outperforms both ablated versions across medical diagnosis, image captioning, and video QA tasks. Specifically, removing cross-attention reduced medical and video QA accuracy by 12%, while excluding multi-shot prompting caused further declines. The CIDEr score also dropped from 112 to 98, indicating reduced captioning quality. These results confirm that both cross-modal fusion and multi-shot context are critical for robust and context-aware multimodal inference.

Figure 2: Ablation Study of Multi-Shot Multimodal LLM on Three Benchmarks

5. Discussion

This study demonstrates that multi-shot multimodal inference significantly enhances cross-modal reasoning by using contextual examples. In tasks like medical diagnosis, the model learns to associate textual terms like *"consolidation"* with corresponding visual features in X-rays, improving both accuracy and interpretability. Similar gains are seen in image captioning and video QA, where diverse prompts help align visual and linguistic cues.

However, the architecture has two key limitations: it demands high GPU memory due to multimodal processing and depends heavily on the quality of input examples. Poor prompts can reduce model reliability.

Despite these challenges, the approach holds promise—especially in healthcare, where explainable predictions via attention maps can build trust. Future work will explore memory-efficient architectures and adaptive prompting to enhance scalability and robustness.

References

1. Radford et al., CLIP: Connecting Text and Images, 2021.
2. Raffel et al., T5: Text-to-Text Transfer Transformer, 2020.
3. Johnson et al., MIMIC-CXR: A Large Publicly Available Database, 2019.

4. Huang, P.-S., et al. (2023). "Language Is Not All You Need: Aligning Perception with Language Models." arXiv preprint arXiv:2302.14045 (Kosmos-1).
5. Li, J., et al. (2023). "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." arXiv preprint arXiv:2301.12597.
6. Wang, W., et al. (2022). "Git: A Generative Image-to-Text Transformer for Vision and Language." arXiv preprint arXiv:2205.14100.
7. Girdhar, R., et al. (2023). "ImageBind: One Embedding Space to Bind Them All." CVPR 2023.
8. Lei, J., et al. (2021). “Less is More: ClipBERT for Video-and-Language Learning via Sparse Sampling.” CVPR 2021.
9. Bain, M., et al. (2021). “Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval.” ICCV 2021.
10. Li, Y., et al. (2022). “Video-Language Pre-Training with Frozen Transformers and Cross-Modal Alignment.” NeurIPS 2022.