Multi-modal

LlamaIndex offers capabilities to not only build language-based applications but also **multi-modal** applications - combining language and images.

Types of Multi-modal Use Cases

This space is actively being explored right now, but some fascinating use cases are popping up.

RAG (Retrieval Augmented Generation)

All the core RAG concepts: indexing, retrieval, and synthesis, can be extended into the image setting.

- The input could be text or image.
- The stored knowledge base can consist of text or images.
- The inputs to response generation can be text or image.
- The final response can be text or image.

Check out our guides below:

- GPT-4V Multi Modal
- Multi-modal retrieval with CLIP
- Image to Image Retrieval
- Structured Image Retrieval
- Chroma Multi-Modal
- Gemini Multi-Modal
- Ollama Multi-Modal

Structured Outputs

You can generate a `structured` output with the new OpenAI GPT4V via LlamaIndex. The user just needs to specify a Pydantic object to define the structure of the output.

Check out the guide below:

- Multi-Modal Pydantic Program

Retrieval-Augmented Image Captioning

Oftentimes understanding an image requires looking up information from a knowledge base. A flow here is retrieval-augmented image captioning - first caption the image with a multi-modal model, then refine the caption by retrieving it from a text corpus.

Check out our guides below:

- Llava + Testla 10Q

Agents

Here are some initial works demonstrating agentic capabilities with GPT-4V.

- Multi-Modal Agents
- GPT-4V Experiments

Evaluations and Comparisons

These sections show comparisons between different multi-modal models for different use cases.

LLaVa-13, Fuyu-8B, and MiniGPT-4 Multi-Modal LLM Models Comparison for Image Reasoning

These notebooks show how to use different Multi-Modal LLM models for image understanding/reasoning. The various model inferences are supported by Replicate or OpenAI GPT4-V API. We compared several popular Multi-Modal LLMs:

- GPT4-V (OpenAI API)
- LLava-13B (Replicate)
- Fuyu-8B (Replicate)
- MiniGPT-4 (Replicate)
- CogVLM (Replicate)

Check out our guides below:

- Replicate Multi-Modal
- GPT4-V

Simple Evaluation of Multi-Modal RAG

In this notebook guide, we'll demonstrate how to evaluate a Multi-Modal RAG system. As in the text-only case, we will consider the evaluation of Retrievers and Generators separately. As we alluded to in our blog on the topic of Evaluating Multi-Modal RAGs, our approach here involves the application of adapted versions of the usual techniques for evaluating both Retriever and Generator (used for the text-only case). These adapted versions are part of the llama-index library (i.e., evaluation module), and this notebook will walk you through how you can apply them to your evaluation use cases.

- Multi-Modal RAG Evaluation

Model Guides

Here are notebook guides showing you how to interact with different multimodal model providers.

- OpenAI Multi-Modal
- Replicate Multi-Modal
- Ollama Multi-Modal