COSMO LENS

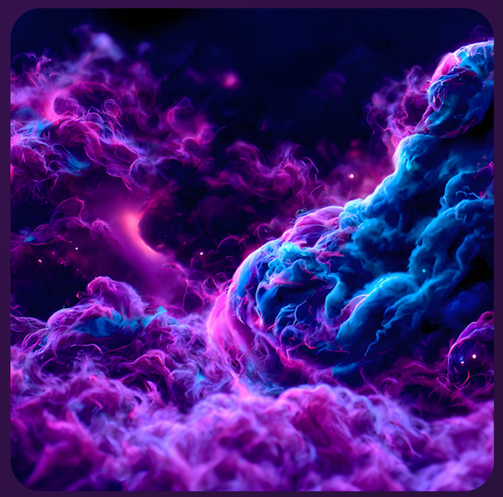
A MULTIMODAL VLM BASED RAG PIPELINE FOR ASTRONOMICAL ANALYSIS

ESA REVIEW

Team 10 (Sem 6 H Section):

Samarth Prakash	PES2UG22CS495
Sadhana Shashidhar	PES2UG22CS478
Sakshi Jineshbhai Rajanai	PES2UG22CS491

ABSRACT



Astronomical data proliferation (images, text) challenges traditional information retrieval. We introduce CosmoLens, a multimodal Retrieval-Augmented Generation (RAG) system for exploring public archives like ESA/Hubble and APOD. CosmoLens accepts natural language queries, optionally grounded by an image, using distinct CLIP (image) and Jina v3 (text) embeddings indexed in Pinecone (~30,000 pairs). It retrieves relevant multimodal context, which Google's Gemini Flash LLM synthesizes with the query to generate comprehensive, grounded answers. Our Python/FastAPI prototype demonstrates potential to accelerate discovery and democratize access to astronomical data.

LITERATURE SURVEY

Multimodal Universe dataset

E. Angeloudi et al., "The Multimodal Universe: Enabling Large-Scale Machine Learning with 100TB of Astronomical Scientific Data," in The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2024.

AstroCLIP

F. Lanusse et al., "AstroCLIP: Cross-Modal Pre-Training for Astronomical Foundation Models," in NeurIPS 2023 AI for Science Workshop, 2023.

Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks

P. Lewis et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," in The 34th Conference on Neural Information Processing Systems (NeurIPS 2020), 2020.

LITERATURE SURVEY

Qwen2-VL

P. Wang et al., "Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution," arXiv:2409.12191 [cs.CV], 2024.

LoRA Fine Tuning

E. J. Hu et al., "LoRA: Low-Rank Adaptation of Large Language Models," arXiv:2106.09685 [cs.CL], 2021.

Optimizing Retrieval-Augmented Generation for Space Mission Design via Multi-Task Learning

A. Balossino et al., "Optimizing Retrieval-Augmented Generation for Space Mission Design via Multi-Task Learning", Politecnico di Torino, 2024.

LITERATURE SURVEY

CosmoCLIP

R. Imam, M. T. Alam, U. Rahman, M. Guizani, and F. Karray, "CosmoCLIP: Generalizing Large Vision-Language Models for Astronomical Imaging," arXiv:2407.07315 [cs.CV], 2024.

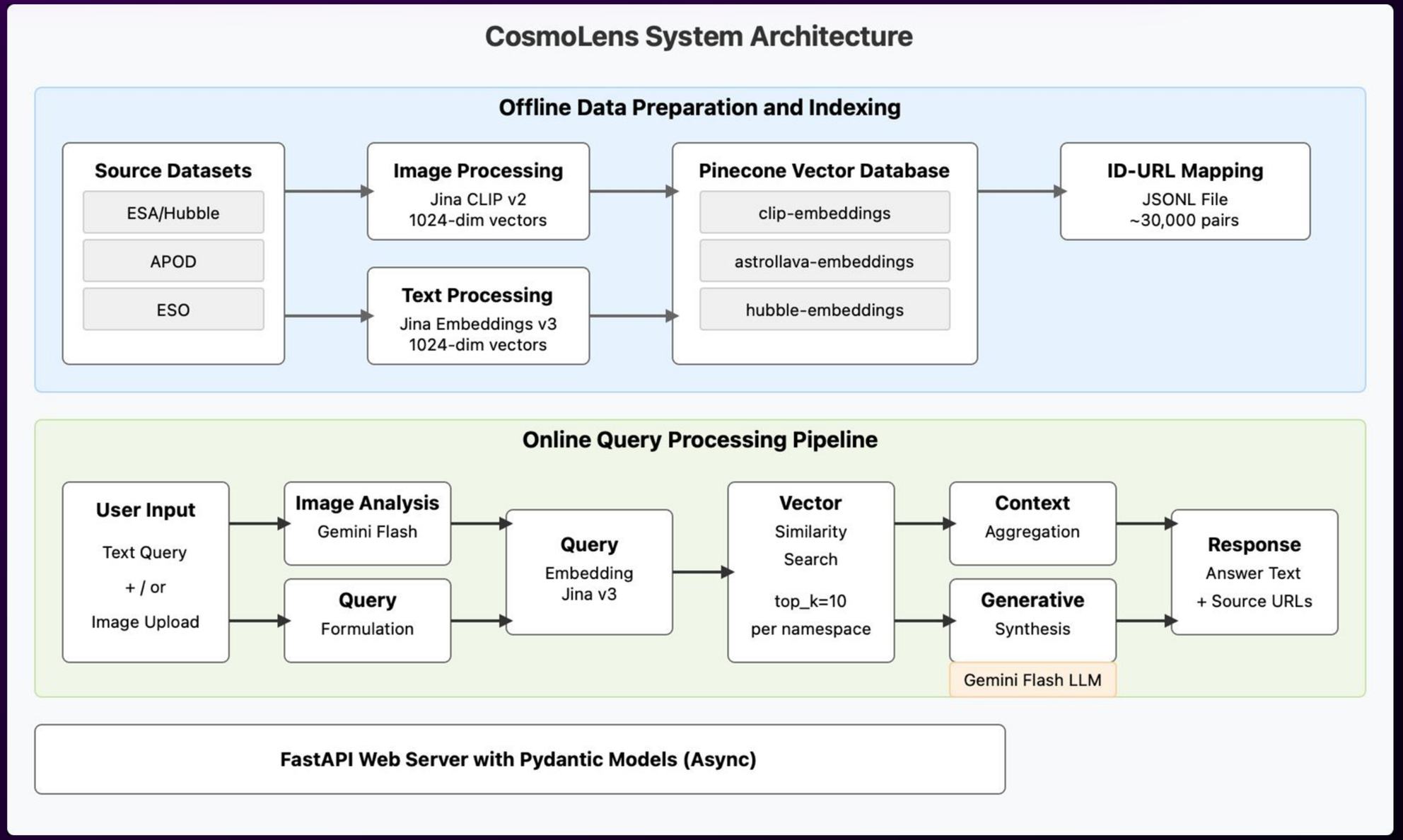
ColPali

M. Faysse, H. Sibille, T. Wu, B. Omrani, G. Viaud, C. Hudelot, and P. Colombo, "ColPali: Efficient Document Retrieval with Vision Language Models," in Proc. 13th Int. Conf. Learn. Represent. (ICLR), 2025.

Wiki-LLaVa: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs

D. Caffagni et al., "Wiki-LLaVa: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs," in IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW), 2024.

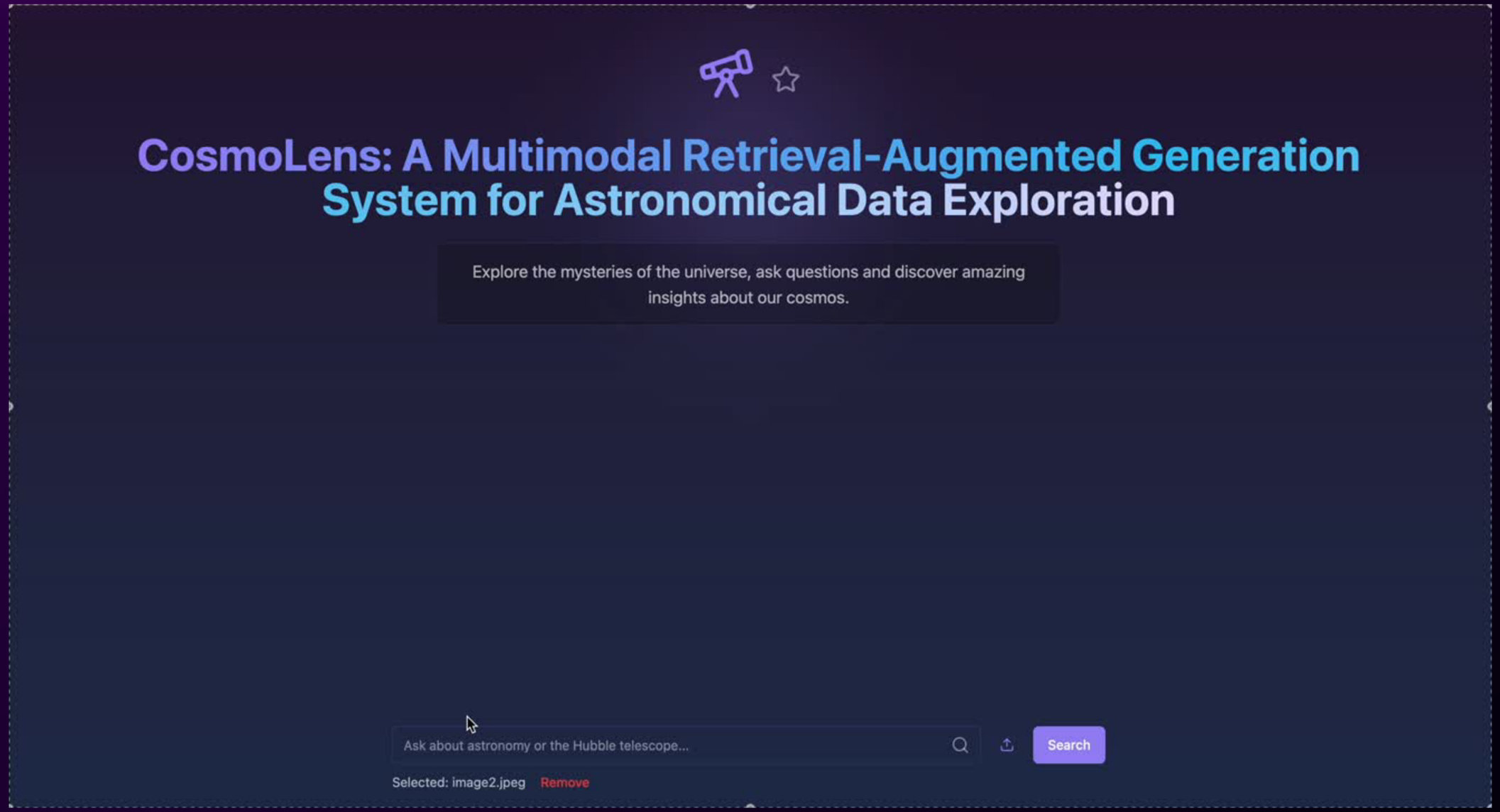
ARCHITECTURE DIAGRAM AND METHODOLOGY



NOVELTY OF APPROACH

- Domain-Specific Multimodal RAG Integration:
  - Applies end-to-end multimodal RAG specifically to astronomy archives (Hubble, APOD, ESO), tackling unique domain challenges (modality gap, accessibility).
- Hybrid Embedding Strategy:
  - Uses distinct, best-in-class embeddings (CLIP for vision, Jina v3 for text) instead of a single joint space, maximizing retrieval quality per modality.
- Structured Multimodal Knowledge Base:
  - Demonstrates practical Pinecone setup using multiple namespaces to organize different embedding types (image/text), enabling flexible retrieval.
- Shift in Interaction Paradigm for Astronomy:
  - Moves beyond expert interfaces (SQL/metadata search) to intuitive natural language interaction for complex, cross-modal astronomical queries.

RESULTS


CosmoLens: A Multimodal Retrieval-Augmented Generation System for Astronomical Data Exploration
Explore the mysteries of the universe, ask questions and discover amazing insights about our cosmos.
Ask about astronomy or the Hubble telescope...
Search
Selected: image2.jpeg
Remove

RESULTS

- RAG Outperforms Zero-Shot: CosmoLens, both text-only and multimodal, significantly improved Faithfulness (factual grounding) and Relevance for astronomy queries compared to directly prompting a large language model (LLM Zero-Shot).
- Multimodal Advantage: For queries involving both text and an input image, CosmoLens showed markedly superior Relevance compared to text-only RAG (which ignores the image) and zero-shot LLM (which lacks retrieved context).
- Effective Synthesis: CosmoLens successfully integrated visual information from user images with retrieved textual context, generating answers that maintained high Faithfulness.
- Validation: Preliminary findings strongly support the effectiveness of the implemented multimodal RAG architecture for answering complex, context-dependent astronomical questions involving both visual and textual data.
- Finetuning accuracy for multimodal universe gz-10 right now is 57% compared to base model of 24%, is a siginificant improvement

INDIVIDUAL CONTRIBUTION

Team Member	Primary Focus	Key Tasks	Outcome/Deliverable
Sakshi Rajani	Data & Vector Foundation	• Process/Clean Datasets (Hubble, APOD, Morphology) • Generate & Save Text Embeddings (Jina v3) • Generate & Save Image Embeddings (CLIP) • Setup Pinecone & Upsert Embeddings (Namespaces) • Manage ID-URL Mapping	Populated Pinecone KB & Prepared Data
Sadhana Shashidhar	RAG Pipeline & API Backend	• Setup FastAPI Backend & API Models • Implement Query Embedding (Jina) • Implement Pinecone Text Search • Integrate LLM (Gemini) for Text Generation • Build Text Query API Endpoint (/query/)	Functional Text-Based RAG API
Samarth P	Multimodal/VL Fine-tuning & Eval	• Fine-tune Qwen-VL (Galaxy Morphology) • Enhance API for Image Uploads • Integrate Image Analysis (Qwen/Gemini) • Implement Multimodal LLM Synthesis • Develop Evaluation Framework & Test • Lead Documentation & Presentation	Multimodal RAG, Fine-tuned VL Model, Eval, Docs
Shared Responsibilities		Initial Design, Data Strategy, Integration Testing, Debugging, Version Control (Git).	

REFERENCES

1] A. Accomazzi and G. Eichhorn, ”Astronomical data systems in the big data era,” in Handbook of Big Data, P. Buhlmann, P. Drineas, M. Kane, and M. van der Laan, Eds. Boca Raton, FL, USA: CRC Press, 2018, pp. 519–540.

[2] ESA/Hubble, ”Hubble images.” [Online]. Available: https://esahubble.org/images/ (accessed Aug. 2024).

[3] NASA, ”Astronomy picture of the day (APOD).” [Online]. Available: https://apod.nasa.gov/apod/ (accessed Aug. 2024).

[4] European Southern Observatory (ESO), ”ESO images.” [Online]. Available: https://www.eso.org/public/images/ (accessed Aug. 2024).

REFERENCES

[5] A. Radford et al., ”Learning transferable visual models from natural language supervision,” in Proc. 38th Int. Conf. Machine Learning (ICML), Jul. 2021, pp. 8748–8763.
[6] T. B. Brown et al., ”Language models are few-shot learners,” in Advances Neural Inf. Process. Syst., Dec. 2020, vol. 33, pp. 1877–1901.
[7] P. Lewis et al., ”Retrieval-augmented generation for knowledge-intensive NLP tasks,” in Advances Neural Inf. Process. Syst., Dec. 2020, vol. 33, pp. 9459–9474.
[8] J. Li et al., ”MultiModal-RAG: A framework for integrating multimodal information retrieval into large language models,” arXiv preprint arXiv:2311.16070, Nov. 2023.

REFERENCES

[9] M. Yasunaga et al., ”Retrieval-augmented multi-modal language modeling,” arXiv preprint arXiv:2305.03753, May 2023.
[10] Jina AI, ”Jina embeddings 2: World’s best open source text embedding model supporting 8192 sequence length,” Jina AI Blog, Oct. 2023. [Online]. Available: https://jina.ai/news/jina-embeddings-2-worlds-best-open-source-text-embedding-model-supporting-8192-sequence-length/
[11] Google DeepMind, “Gemini: A family of highly capable multi-modal models,” Google DeepMind, Tech. Rep., Dec. 2023. [Online]. Available: https://storage.googleapis.com/deepmind-media/gemini/gemini 1 report.pdf
[12] A. Ginsburg et al., “Astroquery: An astronomical web-querying package in Python,” Astron. J., vol. 157, no. 3, p. 98, Feb. 2019.


THANK YOU