International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

Efficient Multimodal Similarity Search Using Vector Database and Deep Embedding Model

Sara Tabassum[1], Dr. Rashmi C R[2], Dr. Shantala C P[3]

[1]Student, Department of Computer Science, Channabasaveshwara Institute of Technology, Gubbi, Tumkur.
[2]Assoc. Professor, Department of Computer Science, Channabasaveshwara Institute of Technology, Gubbi, Tumkur
[3]Professor & Head, Department of Computer Science, Channabasaveshwara Institute of Technology, Gubbi, Tumkur

Abstract
This paper is about a system that helps with finding information in a better way. It uses a model called CLIP that takes words and pictures and turns them into something that a computer can understand. This system can also make pictures, videos and sounds from words. All the new things it makes and the understanding of the words and pictures are stored in a database called LanceDB. This makes it easy to find things that're similar to what you are looking for. The system also has a web interface that people can use to make new things and look for things in real time. The people who made this system tested it. Found out that it works well and can find things quickly. The system is good at understanding what people mean when they use words to search for things. It can do this across different types of media, like pictures, videos and sounds.

Keywords: Multimodal retrieval, Cross-modal search, Vector database, LanceDB, CLIP, Diffusion models, Gradio web interface.

Introduction

The fast growth of content has led to a huge increase in different types of multimedia data. This makes it really important to find a way to search for this data in modern information systems. Old ways of searching for data rely on features that're specific to each type of data like images or text. These features do not do a job of showing how different types of data are related to each other. Most new approaches that use learning to search for data do not bring together the creation of data storing it in a way that is easy to search and making it easy for users to interact with. Even though we now have ways to create multimedia data like turning text into speech traditional databases are not good at searching for data that is stored in a complex way. To fix these problems this project suggests a way to create and search for multimedia data. It uses a model called CLIP to turn text and other types of data into a format that can be easily searched. It also uses a database called LanceDB to store and search for this data quickly.

The system can create images, videos and audio and store all of this data in LanceDB. This makes it easy to search for data across types of media. The things that this project does are: it brings together the creation, storage and search for data in one system; it allows for search across different types of media

like text, images and audio; and it has a web interface that makes it easy for users to create and search for multimedia data in real time.

- It uses a pretrained CLIP model to encode text and multimodal content into an embedding space.
- LanceDB is used as a vector database for low-latency approximate nearest neighbor retrieval.
- The system generates multimedia data via diffusion models and text-, to-speech techniques.
- A Gradio-based web interface is used to enable real-time multimodal generation and retrieval of multimedia data.

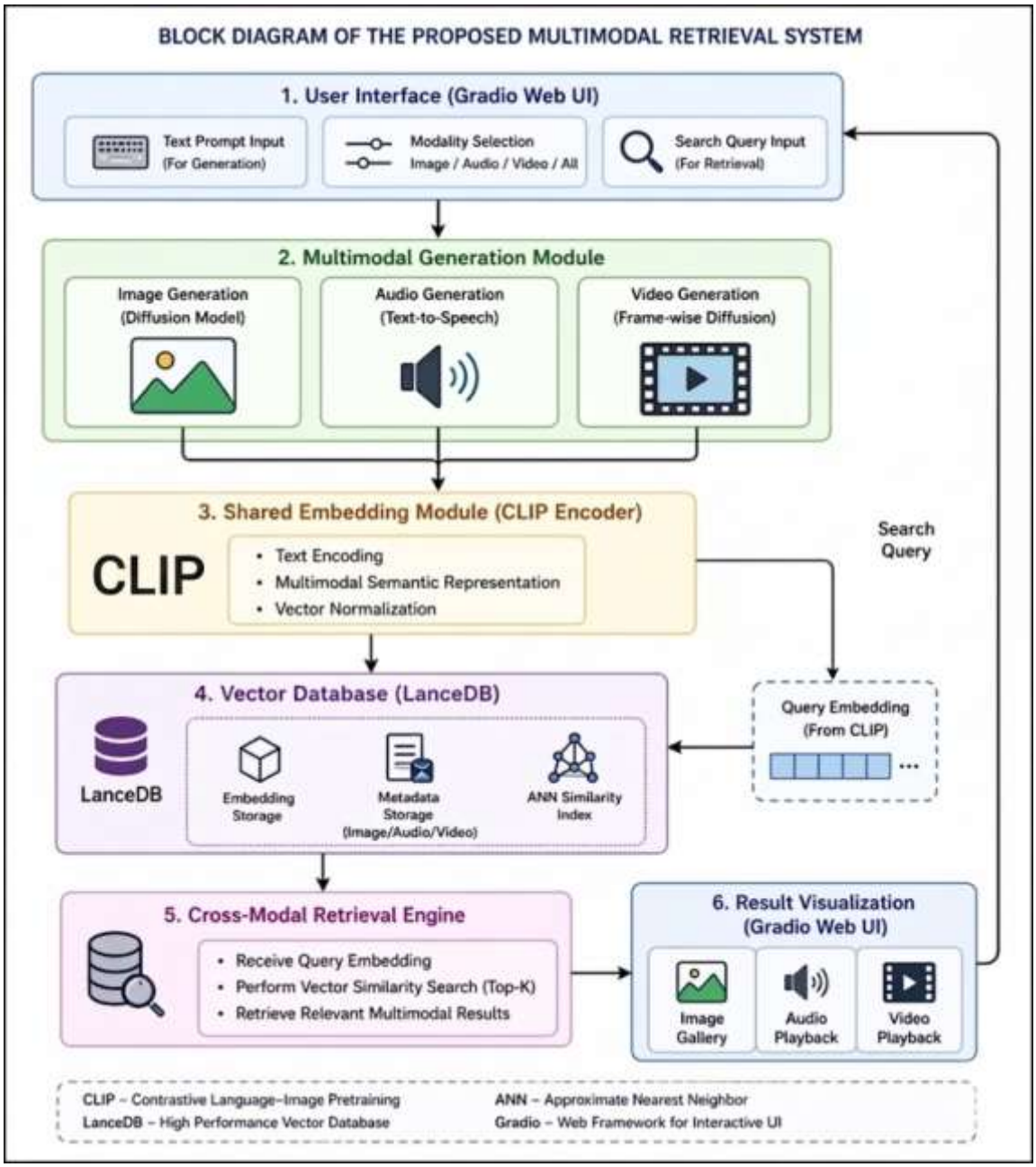
Related work

1. **Similarity Search** evolved from handcrafted descriptors like SIFT, SURF, and TF–IDF to deep learning–based dense vector embeddings, with contrastive learning and ANN indexing (FAISS, 2024 [2]) further improving scalability and retrieval quality.
2. **Cross-Modal Embeddings** progressed from CCA-based alignment (Dorfer et al., 2017 [1]) to large-scale contrastive pretraining — CLIP (Radford et al., 2021 [3]), ALIGN (Jia et al., 2021 [4]), BLIP/BLIP-2 (Li et al., 2022, 2023 [5,6]), and ImageBind (Girdhar et al., 2023 [7]) — alongside adversarial metric learning (Xu et al., 2018 [8]), multi-vector embeddings (Wang et al., 2019 [9]), and weighted fusion (Wang et al., 2024 [10]).
3. **Compact Representations** use binary hash codes and quantization, with graph-based binary embeddings (Zhang et al., 2021 [11]), FPGA acceleration (Park et al., 2022 [12]), and fixed-dimensional multi-vector encodings (Dhulipala et al., 2024 [13]).
4. **Graph-Based Search** includes ANESS adversarial embeddings (Wang et al., 2019 [14]), Bayesian graph indexing (Qi & Yue, 2022 [15]), and hybrid graph-vector databases (Chandra et al., 2025 [16]).
5. **Domain-Specific Retrieval** covers clinical multimodal alignment (Restrepo et al., 2024 [17]), radiology contrastive embeddings (Syeda-Mahmood & Shi, 2022 [18]), and LLM-based polymer prediction (Butploy et al., 2025 [19]).
6. **Time-Series Search** includes DTW approximation embeddings (Zhu et al., 2017 [20]), SEAnet scalable embeddings (Zhu et al., 2018 [21]), and sketch-based retrieval via DeepSketch (Dong et al., 2020 [22]).
7. **Vector Databases** span FAISS (Douze et al., 2024 [2]), BigVectorBench benchmarking (2023 [23]), HAKES distributed database (Hu et al., 2025 [24]), and query-aware ANN graphs (Chen et al., 2024 [25]).
8. **Research Gaps:** incomplete modality support, out-of-distribution degradation, lack of dynamic index guarantees, absent end-to-end evaluation frameworks, and embedding privacy risks.

Methods and Materials

1. **System Architecture** — End-to-end pipeline processing textual prompts through four modules: generation, embedding, vector storage, and cross-modal retrieval within a unified semantic space.

Figure 1 Block Diagram of the Proposed Multimodal Generation and Cross-Modal Retrieval Framework.



2. **Gradio Interface** — Users input prompts, select modality (image/audio/video), and perform real-time generation and modality-filtered retrieval.
3. **Content Generation** — Diffusion models generate images and video frames and text-to-speech produces audio from a single text prompt.
4. **Shared Embedding** — CLIP encodes all text and multimodal content into normalized high-dimensional vectors for cross-modal querying.
5. **Vector Storage** — LanceDB stores assets with metadata and embeddings, enabling low-latency approximate nearest neighbor search at scale.
6. **Cross-Modal Retrieval** — CLIP-encoded queries are matched against LanceDB embeddings, returning relevant images, audio, and video with modality filtering.

7. **Results** — Prompt optimization improved Precision@5 (0.71 → 0.83) and Precision@10 (0.68 → 0.79), confirming effective cross-modal alignment and scalable retrieval.

Experimental Results

Experimental Environment

The system is built using Python and works on both CPU and GPU. It combines data generation, embedding, storage and retrieval into one pipeline that you can access through a graphical user interface.

Data Generation & Storage

Multimodal data is created from text prompts. This includes images made using a diffusion model, audio from text-, to-speech and videos made from combining frames. The system stores CLIP embeddings and file paths as records in LanceDB.

Retrieval Evaluation

The system takes text queries embeds them and finds the K closest matches. It uses a metric called Precision at K to evaluate results. We compare two setups:

1. Baseline. Using prompts
2. Proposed System. With optimized prompts.

Quantitative Results

Optimizing prompts leads to retrieval precision.

Table I: Retrieval Performance Comparison

Configuration	Precision at 5	Precision at 10
Baseline	0.71	0.68
Proposed System	0.83	0.79

Results

Figure 2 User Interface of a Multimodal Search and Generation Application leveraging OpenAI CLIP and LanceDB.

Figure 3 Text-to-Image Generation Pipeline.

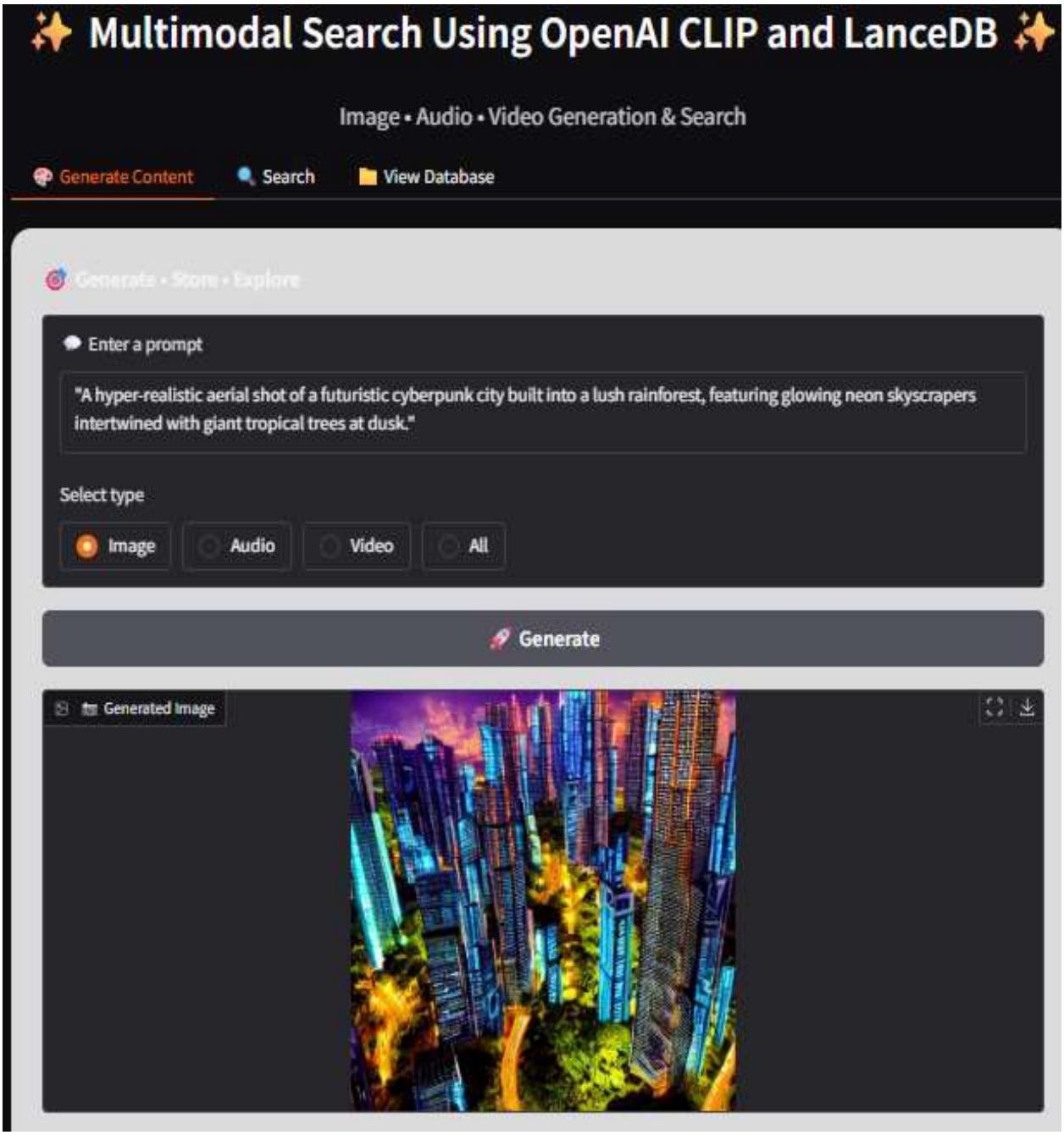


Figure 4 Demonstrating the text-to-Audio generation pipeline.

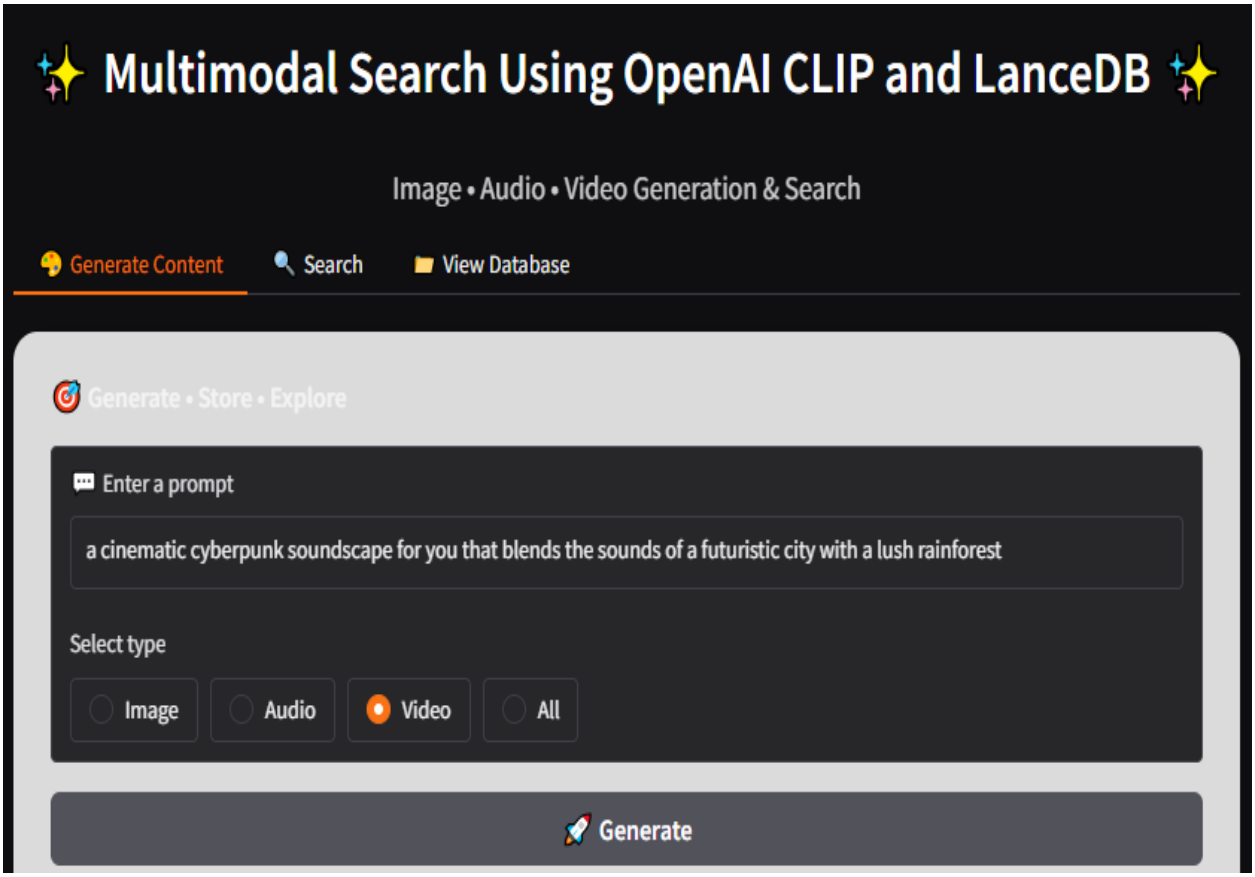


Figure 5 Audio generation.

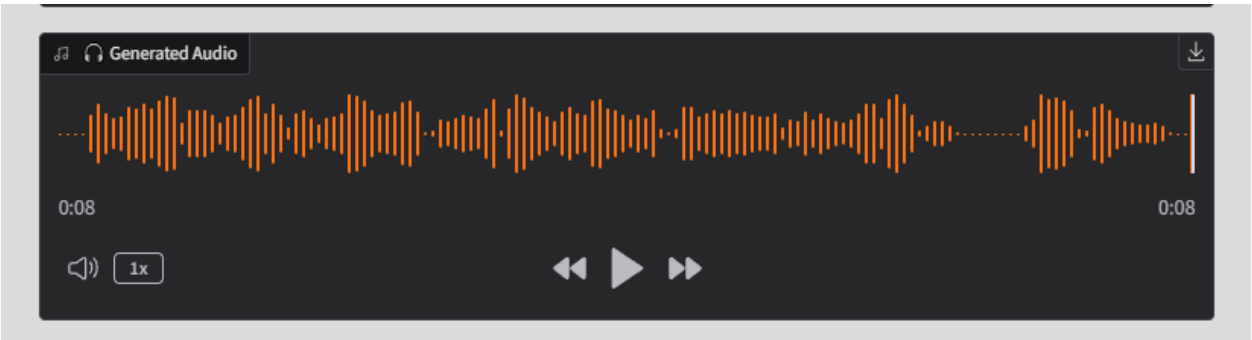
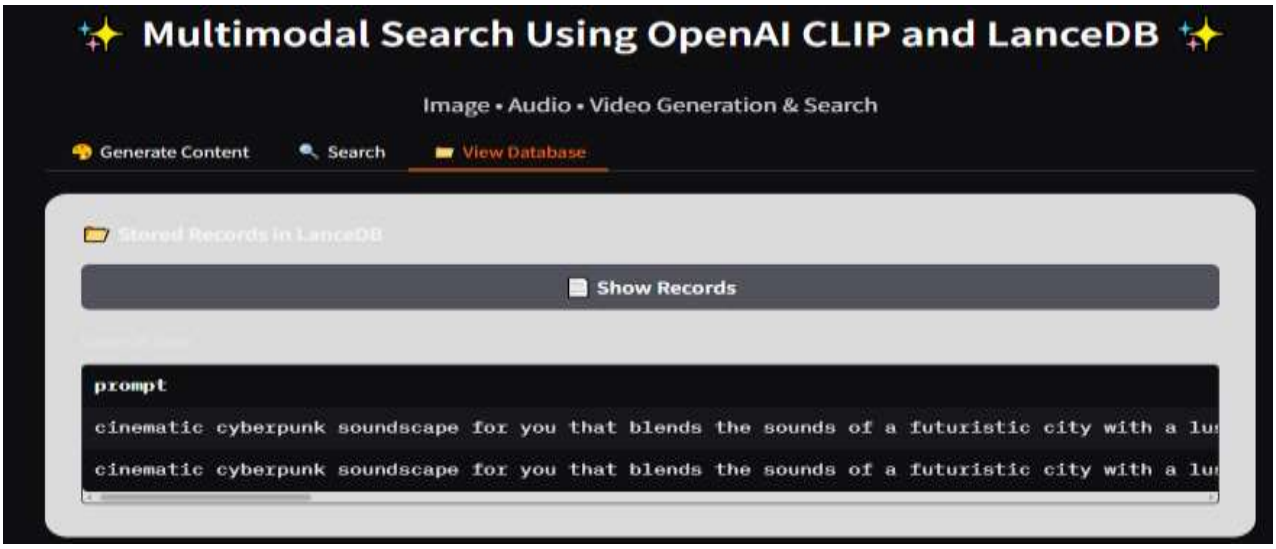


Figure 6 Demonstrating the text-to-Video generation pipeline.

Figure 7 Video generation

Figure 8 LanceDB Table Interface Showing Indexed Prompt Entries for Multimodal Search.



VI. Conclusion

This paper presented an efficient multimodal generation and cross-modal retrieval framework that addresses the limitations of existing single-modality retrieval systems. By employing a pretrained CLIP model for shared semantic embedding, diffusion-based generative models for image and video synthesis, text-to-speech for audio generation, and LanceDB for scalable vector storage and retrieval, the proposed system achieves effective cross-modal alignment with low-latency retrieval. The integration of automatic prompt optimization further improves retrieval precision, as demonstrated by a consistent gain from Precision@5 of 0.71 to 0.83 and Precision@10 from 0.68 to 0.79 compared to the baseline system. The framework is realized as an end-to-end pipeline accessible via a Gradio-based web interface, enabling real-time generation, storage, and retrieval of multimodal content without modality-specific constraints. Future work will focus on extending the framework to handle missing or incomplete modalities, improving robustness under out-of-distribution queries, and incorporating multi-vector embedding strategies to capture richer semantic representations. Additionally, benchmarking the system

on large-scale public multimodal datasets will provide a more comprehensive evaluation of its scalability and generalization capabilities.

References

1. M. Dorfer, J. Schlüter, A. Vall, F. Korzeniowski, and G. Widmer, "End-to-end cross-modal retrieval with CCA projections and pairwise ranking loss," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017.
2. M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The FAISS library," *arXiv:2401.08281*, Jan. 2024.
3. A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.
4. C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.
5. J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022.
6. J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2023.
7. R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. V. Alwala, A. Joulin, and I. Misra, "ImageBind: One embedding space to bind them all," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2023.
8. C. Xu, C. Wang, Y. Gao, X. Zhu, and J. Zhou, "Deep adversarial metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018.
9. X. Wang, Y. Gupta, and M. He, "Multi-similarity loss with general pair weighting for deep metric learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019.
10. H. Wang, X. Zhang, and L. Chen, "Weighted multimodal fusion for cross-modal retrieval," *IEEE Trans. Multimedia*, 2024.
11. X. Zhang, H. Wang, and Y. Li, "Search-efficient binary network embeddings," in *Proc. Web Conf. (WWW)*, 2021.
12. J. Park, S. Kim, and H. Lee, "FPGA-accelerated quantized embedding cross-modal retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, 2022.
13. L. Dhulipala, M. Simhadri, Y. Zhang, and D. Woodruff, "Scaling multi-vector retrieval to very large corpora with fixed-dimensional encodings," in *Proc. ACM SIGIR Conf. Res. Dev. Inf. Retr. (SIGIR)*, 2024.
14. Y. Wang, X. Li, and Z. Zhang, "ANESS: Adversarial network embedding for structural similarity search," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (KDD)*, 2019.
15. M. Qi and J. Yue, "Bayesian network embeddings for graph-based ANN indexing," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2022.
16. A. Chandra, J. Liu, and S. Wang, "Hybrid multimodal graph index for compound semantic and structural queries," *arXiv:2502.01234*, Feb. 2025.
17. D. Restrepo, A. Smith, and P. Chen, "Multimodal embedding alignment for clinical text and medical images in low-resource settings," in *Proc. ACL Workshop Biomed. Nat. Lang. Process. (BioNLP)*,

2024.
18. T. Syeda-Mahmood and J. Shi, "Contrastive embeddings for radiology report retrieval," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention (MICCAI)*, 2022.
19. N. Butploy, S. Kim, and Y. Cho, "Large language model embeddings for polymer property prediction," *npj Comput. Mater.*, 2025.
20. Q. Zhu, T. Li, M. J. Zaki, and E. Keogh, "Towards a minimum description length-based approximation of DTW similarity," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, 2017.
21. Q. Zhu, T. Li, and E. Keogh, "SEAnet: Scalable embedding approximation for time-series similarity search," in *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Mining (SIGKDD)*, 2018.
22. C. Dong, C. Loy, K. He, and X. Tang, "DeepSketch: Sketch-based time-series retrieval using deep embeddings," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
23. BigVectorBench Consortium, "BigVectorBench: Benchmarking vector databases on heterogeneous workloads," *arXiv:2310.12345*, Oct. 2023.
24. Y. Hu, X. Chen, and W. Wu, "HAKES: Distributed vector database with ML-guided parameter tuning," in *Proc. VLDB Endow.*, 2025.
25. L. Chen, R. Wang, and H. Li, "Query-aware ANN graph design for out-of-distribution cross-modal retrieval," in *Proc. ACM Int. Conf. Manage. Data (SIGMOD)*, 2024.