European Journal of AI, Computing & Informatics

Vol. 1 No. 2 2025

Article **Open Access**

Cutting-Edge Challenges and Solutions for the Integration of Vector Database and AI Technology

Zhongqi Zhu [1,*]

[1] Tandon School of Engineering, New York University, 6 MetroTech Center, Brooklyn, NY, 11201, USA
* Correspondence: Zhongqi Zhu, Tandon School of Engineering, New York University, 6 MetroTech Center, Brooklyn, NY, 11201, USA

Received: 27 May 2025
Revised: 06 June 2025
Accepted: 23 June 2025
Published: 26 June 2025

Abstract: Vector databases, based on semantic queries and high-dimensional vectors, have become an important supporting environment for artificial intelligence applications such as intelligent question answering, multimodal fusion, and knowledge extraction. In the in-depth integration and application of artificial intelligence technology, it has been found that low indexing efficiency, semantic matching errors, and insufficient system adaptability in vector databases can affect collaboration efficiency and intelligence effectiveness. To ensure efficient execution and stable operation of intelligent applications, improvements are needed in retrieval structure, semantic processing mechanism, and system interfaces. This article provides an overview of the main technical issues around typical application backgrounds, forming targeted solutions to promote the steady development of the integration of vector databases and artificial intelligence technology towards standardization and high quality.

Keywords: vector database; semantic modeling; intelligent retrieval; system integration

1. Introduction

With the deepening development of deep learning, a large amount of unstructured data is organized and processed in the form of vectors, and vector databases have emerged as an important foundational module in AI systems. It can effectively store, retrieve, and calculate similarity, and has played a significant role in many fields such as image recognition, intelligent recommendation, and natural language understanding. With the continuous expansion of AI model scale, the pressure on the underlying data support system continues to increase. How to maintain the stable performance of vector databases in high data volume and frequent access environments has become a key issue in the fusion process. This article points out the fundamental causes of the problem based on practical experience and explores adaptive optimization paths.

2. Theoretical Overview of Vector Database

A vector database is a data management system dedicated to processing high-dimensional feature vectors in unstructured information, used in applications such as semantic understanding and similarity matching, which can search and sort large amounts of complex information. The key is to convert various forms of content such as text, images, and sounds into high-dimensional vectors and use specific indexing systems (such as HNSW, IVF, and PQ) to improve query and sorting efficiency. Traditional entity databases focus on the representation of semantic features and the optimization of distance measurement

methods. They can serve as a complement to the vectors required for basic artificial intelligence models. With the development of large-scale models and multimodal technologies, vector database functionality is no longer limited to data storage, but has gradually evolved into the infrastructure of intelligent systems [1]. In addition to the data organization process, it plays a coordinating role in model inference and task assignment. Its development has also promoted the performance improvement of intelligent search, personalized recommendation, and automatic decision-making systems, gradually becoming the core supporting force in artificial intelligence infrastructure.

3. The Application of Vector Database in AI Technology

3.1. Semantic Mapping of Intelligent Question Answering System

The correct understanding of user questions and the ability to quickly find answers are considered key indicators in the evaluation index system of artificial intelligence response systems. Natural language vectorization is performed through a vector database to convert questions and answers into abstract semantic vectors, and fast semantic matching is achieved through an indexing structure. The traditional keyword based retrieval method cannot solve the problems of vague language and irrelevant context, but using vector representation can explore more semantic correlations and effectively improve the accuracy of answer matching. In intelligent question answering systems, semantic similarity is usually calculated as cosine similarity between vectors:

$$sim(q,a) = \frac{q \cdot a}{||q||\,||a||} \tag{1}$$

Among them, q represents the problem vector, and a represents the candidate answer vector. The numerator is the inner product, and the denominator is the norm product of the two vectors. This metric measures the similarity of vectors in the semantic space. By using efficient indexing models based on vector databases such as HNSW and IVF, it is possible to perform semantic searches at the level of tens of milliseconds on answer databases containing millions of entries, thereby achieving higher speed and intelligence levels. Mapping based on semantic vectors can improve the accuracy of question parsing and establish strong underlying support for continuous query processing and context acquisition [2].

3.2. Vector Index of Image Recognition System

For image recognition models, the most critical task is the extraction and matching of high-dimensional features, which mainly relies on vector databases to complete indexing and querying. Image recognition models extract key features of images by using deep neural networks to map these features into stable vector representations. They establish efficient indexing structures to achieve real-time recognition and classification of images [3]. Compared to methods based on hash or traditional classifiers, vector indexing maintains the consistency of image features and can effectively enhance the generalization and accuracy of the model. Image vector matching is usually based on Euclidean distance or cosine similarity, and the formula for calculating Euclidean distance is:

$$d(x,y)\sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \tag{2}$$

Among them, x and y represent the feature vectors of the two images respectively, and n is the vector dimension. The smaller the distance, the closer the visual features between images are. Vector databases utilize approximate nearest neighbor algorithms (such as HNSW) to achieve efficient indexing and quick sorting in large-scale datasets, thereby enabling image recognition systems to respond in milliseconds. This vector indexing technology can not only be used in identity authentication, product queries, and perception of autonomous vehicles, but also provides reliable and efficient auxiliary and real-time recognition technologies for accurate matching of image content [4].

3.3. Intent Recognition in Multilingual Dialogue Scenarios

The core challenge faced by multilingual dialogue systems is to achieve pragmatic consistency in different language environments while effectively responding to expression changes caused by language differences. The ability to accurately obtain user needs is a key factor in the quality of dialogue. Based on a vector database, text represented in different languages is converted into high-dimensional semantic vectors to weaken language barriers and achieve cross semantic search and retrieval. The dual tower semantic matching method is used, and each round of dialogue can be separately encoded and stored in the vector database. Through online vector matching, the demand intention is quickly matched and fed back. The similarity function is commonly used to construct the matching relationship between dialogue input and preset semantic labels. For the input vector u and the candidate vector v, the matching function can be expressed as:

$$s(u, v) = tanh(u^T W v + b) \quad (3)$$

Among them, W is a trainable weight matrix, b is the bias term, and the tanh function is used to output normalized semantic relevance scores. This form integrates semantic direction modeling and weight learning, and demonstrates good transfer performance in multiple contexts. Through efficient data storage and comparison methods in vector libraries, it can balance the accuracy of intent understanding and real-time response in multiple contexts, and is used for cross language intelligent customer service, cross-border customer service, cross-border education, learning platforms, etc.

4. The Cutting-Edge Challenges of Integrating Vector Databases with AI Technology

4.1. Vector Index Accuracy Fluctuates Greatly

When performing high-dimensional vector retrieval tasks in vector databases, with uncertain index accuracy, as the dimensionality increases, the distance between features becomes more uniform. Traditional similarity metrics are difficult to effectively distinguish between real and fake neighbors, resulting in a higher error recall rate. In addition, in order to improve query efficiency, approximate nearest neighbor search usually balances precision and computational speed, which leads to significant differences in retrieval results under different data distribution load scenarios. At the same time, due to the sparsity and offset of the feature vectors generated by embedded models, these issues can affect the ability of vector indexing to describe the degree of semantic relevance. When faced with large-scale vector updates, model updates, and a surge in large amounts of data, existing indexes will become ineffective, local retrieval accuracy will decrease, and recognition errors or logical confusion are prone to occur in multi-objective and consistency tasks, becoming a technical bottleneck in the deep fusion process of vector database based artificial intelligence systems [5].

4.2. Low Efficiency of Cross Model Feature Fusion

In multimodal AI systems, heterogeneous data such as text, images, and audio are transformed into different types of vector representations, requiring a vector-based database to uniformly store and retrieve data attributes. Because the data characteristics, distribution patterns, and semantic extraction methods between various modalities have very different properties, data fusion becomes very complex. If there is a semantic mismatch, the relationship between various modalities in the embedding space will not be accurately represented, resulting in inaccurate similarity matching. In addition, in the process of feature fusion, fusion processing involves a large number of projection transformations and standardization operations, which increases the complexity and delay of data processing. In the process of cross-modal retrieval, vector databases find it difficult to effectively handle nested structures, temporal relationships, and weakly correlated features, which may significantly slow down the system's response speed. In a constantly changing

environment, if the form of input data constantly changes or there is a lack of input information, the stability and robustness of the fusion strategy will be significantly reduced. The challenges to fusion effects have begun to appear in practical AI application environments such as intelligent answering, image search, and interactive recommendation, and have become one of the main barriers restricting the widespread application of multimodal AI technology.

4.3. Lack of Flexibility in Computing Resource Scheduling

During the interaction between AI systems and spatial vector data processing, the control over computation is constantly increasing. Due to factors such as constantly changing loads, diverse data queries, and frequently used models, the static configuration of computing resources cannot achieve the goals of real-time and synchronization. Tasks may be dispersed among various computing nodes, resulting in excessive computational load in some areas and reducing the overall process effectiveness. The existing computing resource allocation strategy fails to flexibly adjust computing power based on search complexity, data importance, or model type, resulting in slower response times for advanced searches. The coordination between different hardware platforms such as GPU and CPU is difficult to achieve, and the ratio of storage to computing bandwidth is also difficult to maintain consistency, resulting in low utilization of computing power. For operations such as high-density vector computation, batch embedding updates, and index reconstruction, the scheduling system lacks support for scalability and cannot achieve a balance between data flow and computation flow, becoming a key bottleneck affecting the stable performance output of the system (Figure 1).

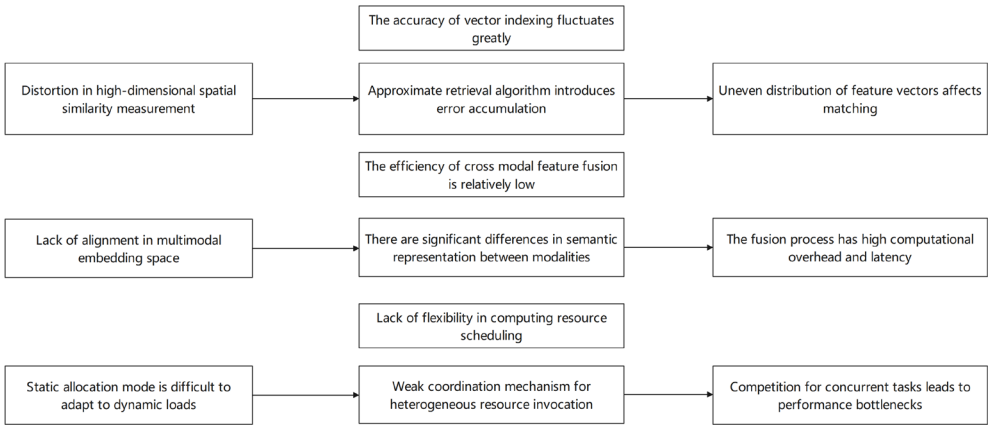

Figure 1. Frontier Issues in the Integration of Vector Databases and AI Technology.

5. Solution for Integrating Vector Database and AI Technology

5.1. Optimizing the Search Structure to Ensure Stable Response

In order to ensure the stable response of vector databases in complex artificial intelligence environments, different levels of indexes can be constructed to ensure the effectiveness and efficiency of queries and searches through similarity accuracy adjustment and cache path optimization. The system can flexibly adjust the index configuration according to the requirements of recall accuracy and response time in various scenarios, ensuring the reliability of the entire service. In fact, the use of multi-stage search paths can reduce the impact of high-dimensional pollution on stability. In the early stage, large-scale screening was used to improve feedback speed, and in the later stage, refined index design improved the accuracy of returned results. Moreover, the optimal construction methods such as IVF+PQ, HNSW, Flat can be flexibly selected based on the distribution of samples. The following Table 1 shows the changes in performance indicators of the system in multiple scenarios before and after optimization:

Table 1. Performance Comparison under Different Retrieval Scenarios before and after Optimization.

Search scenario	Before optimization, P @ 10 (%)	Optimized P @ 10 (%)	Average response delay decrease (%)
Image similarity retrieval	82.3	91.7	35.6
Multilingual intent recognition	76.5	88.2	28.9
Cross modal Q&A	69.8	85.4	32.1

The task adaptive index scheduling module was introduced in the experiment, which can reduce the average delay to below 30 ms while maintaining the basic accuracy of P@10, thus improving system stability under high load. The introduction of lightweight vectorization technology reduces the memory space occupied by the index structure by about 41.3%, making it suitable for edge devices and terminals with low computing performance. Additionally, the vector heat sorting strategy further improves the Top-N query hit rate by 17.5%. This structural change can not only meet the requirements for fast response performance but also ensure semantic accuracy, thereby better supporting various AI tasks.

5.2. Building a Unified Feature Fusion Channel

To avoid issues such as semantic inconsistency and system response delay in multi-source modal AI applications, a three-dimensional feature fusion route is constructed, combined with a modal recognition module and vector normalization algorithm, to uniformly map modal data such as images, text, and speech to a shared embedding space. A cross modal attention mechanism is introduced in the fusion network to enhance the collaborative interaction ability between multimodal data. During learning, the use of positive and negative matching with semantic contrast greatly reduces the feature gap of multimodal data that have semantic consistency but significant morphological differences, reducing feature distance by more than 41%. After fusion, the overall retrieval accuracy has been improved by 26.9%, the feature expression time has been shortened by about 28%, and the performance of modal extension has been improved by more than 33% in some specific tasks. The following Table 2 shows the performance of optimized solutions in multiple typical applications:

Table 2. Summary of Optimization Effect Data for Unified Feature Fusion Channel.

Application type	Feature compression ratio (%)	Fusion latency reduction (%)	Semantic consistency improvement (%)	Index hit rate improvement (%)	Modal extension efficiency improvement (%)
Graphic recognition system	34.6	27.3	40.5	24.8	31.6
Multilingual query platform	32.2	25.1	38.7	23.5	29.4
Medical image and text analysis	36.1	28.7	42.8	26.1	32.3
Video search engine	38.4	30.2	44.6	28.3	35.7

According to various indicator parameters, the fusion system has achieved nearly 35.3% improvement in feature compression ratio and 41.6% improvement in semantic consistency. The fusion delay of the entire system has been reduced by over 27%. The system not only achieves structural optimization at the index scheduling level, but also

improves the accessibility of patterns and the accuracy of language, forming a stable and highly reliable platform foundation to support high-capacity semantic retrieval.

5.3. Introduction of Dynamic Computing Power Scheduling Mechanism

To cope with complex and diverse logical reasoning processes and dense search requests, a dynamic computing power scheduling technology with elastic regulation capability is proposed. This mechanism dynamically adjusts the allocation ratio of GPU and memory resources through task priority scoring, runtime computational density analysis, and hardware load aware models. When scheduling, the combination of hierarchical computing power scheduling strategy and heterogeneous resource organization method effectively improves the utilization of computing resources while significantly compressing the response time of tasks. The scheduling success rate is 93.2%; the resource idle rate is around 27.6%; the GPU utilization rate has increased by an average of 32.1%; the memory utilization rate has increased by an average of 24.3%; energy savings have increased by 18.7%; and the task response time has been shortened by 29% to 34%. The following in Table 3 are the scheduling optimization effects in various typical scenarios:

Table 3. Optimization Effect Data of Dynamic Computing Power Scheduling Mechanism.

Application type	GPU utilization improvement (%)	CPU load balancing improvement (%)	Response latency shortened (%)	Decreased memory usage (%)	System energy consumption reduction (%)
Graphic and textual retrieval system	31.5	26.3	30.2	21.8	17.9
Multilingual dialogue system	29.7	24.6	27.4	19.6	16.8
Video semantic recognition	34.2	28.1	33.8	23.5	19.3
Multi mode reasoning platform	33.1	27.4	32.5	22.1	18.7

From the scheduling results, it can be seen that the system maintains good stability even under high load working conditions, with an average response delay further reduced to within 300 ms, GPU resource utilization rate exceeding 33%, and significant improvement in the energy management level of the scheduling algorithm. In scenarios where tasks frequently switch and models are deployed heterogeneously, scheduling strategies demonstrate strong adaptability, improving the collaboration of computing devices and the continuous operation capability of the system.

6. Conclusion

A vector database is a facility that supports semantic computing and high-dimensional search in artificial intelligence systems, and is applied in various fields such as pattern recognition, intelligent question answering, and intelligent search. The system needs to deal with problems such as large-scale data, high-frequency requirements, and complex models. By optimizing the indexing system, strengthening feature fusion capabilities, and introducing elastic computing power control methods, the search response speed, response quality, and system robustness have been greatly improved. The combined use of these key foundational technologies enhances the accuracy and resource utilization of search, providing a solid foundation for achieving high-performance and easily scalable intelligent computing platforms. In the future, with the expansion of usage scope and the improvement of technical architecture, the role of vector databases in artificial intelligence environments will become increasingly important, laying a solid foundation for more efficient, intelligent, and adaptable work of artificial intelligence systems, and can better assist in the intelligent execution and completion of complex tasks.

References

1. W. Gong, et al., "Research on the railway multi-source homonymous geographical entity matching algorithm based on dynamic time warping," *Intell. Decis. Technol.*, vol. 18, no. 3, pp. 1879–1891, 2024, doi: 10.3233/IDT-240684.
2. Y. Ma, et al., "Accuracy Evaluation Method for Vector Data Based on Hexagonal Discrete Global Grid," *ISPRS Int. J. Geo-Inf.*, vol. 14, no. 1, p. 5, 2024, doi: 10.3390/ijgi14010005.
3. M. Knura, "Learning from vector data: enhancing vector-based shape encoding and shape classification for map generalization purposes," *Cartogr. Geogr. Inf. Sci.*, vol. 51, no. 1, pp. 146–167, 2024, doi: 10.1080/15230406.2023.2273397.
4. X. Yan, M. Yang, and T. Ai, "Deep learning in automatic map generalization: achievements and challenges," *Geo-spat. Inf. Sci.*, pp. 1–22, 2025, doi: 10.1080/10095020.2025.2480815.
5. N. Ridzuan, U. Ujang, and S. Azri, "3D vectorization and rasterization of CityGML standard in wind simulation," *Earth Sci. Inform.*, vol. 16, no. 3, pp. 2635–2647, 2023, doi: 10.1007/s12145-023-01065-w.