INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 5, May 2026 | ISSN: 2456-4184 | IJNRD.ORG

An Adaptive AI-Driven Framework for Context-Aware Retrieval in Secure Multimodal Personal Memory Archives

Prof. Nitin Wankhade
Department of Information Technology
Nutan Maharashtra Institute of Engineering And Technology
Talegaon Dabhade, India
nitin.wankhade@nmiet.edu.in

Anushka Bhavsar
Department of Information Technology
Nutan Maharashtra Institute of Engineering And Technology
Talegaon Dabhade, India
anushka.bahvsar@nmiet.edu.in

Aniruddha Avhad
Department of Information Technology
Nutan Maharashtra Institute of Engineering And Technology
Talegaon Dabhade, India
aniruddha.avhad@nmiet.edu.in

Yashraj Bhosale
Department of Information Technology
Nutan Maharashtra Institute of Engineering And Technology
Talegaon Dabhade, India
yashraj.bhosale@nmiet.edu.in

***Abstract* — In the digitalage, personal memories are no longer limited to photo albums or handwritten diaries; they now exist as a growing mix of images, voice recordings, notes, and other multimedia content spread across cloud platforms and devices. Although existing storage systems offer convenience, they often remain passive, making it difficult for users to revisit meaningful moments in a structured and intuitive way. This paper presents an adaptive AI-driven framework for context-aware retrieval in secure multimodal personal memory archives. The proposed system is designed to organize and retrieve personal memories more intelligently by combining cloud storage with semantic indexing, voice-to-text conversion, contextual tagging, and emotion-linked analysis of multimedia content. Unlike conventional platforms that focus mainly on storage, the framework emphasizes meaningful recall by identifying relationships between content, context, and user intent. It supports memories in the form of images, text journals, and voice notes, while maintaining security through encrypted storage and authenticated access mechanisms. Experimental observations indicate that the system improves retrieval relevance, reduces search effort, and offers a more natural way for users to reconnect with past experiences.**

Keywords - Context-aware retrieval, multimodal memory archives, personalized digital memory management, semantic indexing, voice-to-text processing, emotion-aware computing, secure cloud framework, adaptive artificial intelligence.

I. INTRODUCTION

In recent years, Artificial Intelligence (AI) has started to play an important role in the way people store, organize, and revisit their personal digital content. With the increasing use of smartphones, cloud platforms, and multimedia applications, everyday memories are now preserved in the form of images, voice notes, text journals, and videos. While digital storage has made memory preservation easier than before, most existing systems are still limited to basic uploading and file management. They do not actively help users understand, connect, or retrieve meaningful memories in an intelligent way. As the volume of personal data grows, manually locating specific moments becomes difficult, especially when memories are remembered through emotions, situations, or partial context rather than exact file names.

Recent research highlights the growing importance of intelligent multimodal systems that can interpret and organize personal content beyond simple storage. Advances in Natural Language Processing (NLP), speech technologies, and semantic analysis have made it possible to process text, convert spoken memories into searchable information, and identify relationships between different media elements. Similarly, cloud-based AI frameworks are increasingly being used to support more adaptive and personalized digital experiences. These developments are highly relevant in the context of personal memory archives, where users often need a more natural and context-aware way to revisit past experiences. Technologies such as speech-to-text conversion, contextual tagging, content summarization, and emotion-linked analysis can significantly improve the usefulness of digital memory systems.

Despite these improvements, many available platforms still lack a unified approach that combines secure storage, intelligent retrieval, contextual understanding, and personalized interaction. Most systems focus either on storage efficiency or on standalone AI functions, without offering a complete memory-centered framework. To address this gap, this paper presents an adaptive AI-driven system for secure multimodal personal memory management. The proposed framework extends the idea of digital archiving by introducing context-aware retrieval, semantic organization, and intelligent processing of multimedia memories. By integrating AI, NLP, and secure cloud architecture, the system aims to create a more meaningful, accessible, and user-friendly memory experience. In this way, the proposed work moves beyond traditional cloud storage and contributes toward a smarter and more human-centered approach to personal digital memory preservation.

Objectives:

1. To design an intelligent cloud-based framework for preserving and managing personal memories in the form of text, images, and voice data.
2. To develop a context-aware retrieval mechanism that enables users to access stored memories more naturally and efficiently.
3. To incorporate AI and language-processing techniques for organizing, summarizing, and interpreting personal multimedia content.

4. To ensure secure handling of sensitive memory data through reliable storage and controlled user access.
5. To create a personalized digital memory environment that improves user interaction, accessibility, and meaningful recall of past experiences.

II. PROBLEM STATEMENT

Although cloud technology has made it easier to store personal photos, voice notes, journals, and other digital memories, most existing platforms still function mainly as storage repositories rather than intelligent memory systems. Users are able to preserve content, but retrieving a specific memory often becomes difficult when the search depends on partial details, emotional association, or situational context. In many cases, individuals do not remember the exact file name, date, or folder location; instead, they recall fragments such as a place, a feeling, or a spoken moment. Conventional cloud platforms are not designed to interpret such cues effectively.

Another significant limitation is that many existing systems handle text, images, and audio as separate forms of data without building meaningful relationships among them. As a result, personal archives remain scattered and difficult to navigate over time. Most available solutions also provide limited support for contextual understanding, semantic organization, and adaptive retrieval, which reduces their usefulness as the volume of stored memories continues to increase. In addition, secure storage and intelligent access are often addressed separately, creating a gap between data protection and user-friendly memory recall.

Therefore, there is a need for an AI-driven framework that goes beyond simple digital archiving and supports secure, context-aware, and personalized retrieval of multimodal personal memories. The proposed system addresses this need by combining intelligent processing, semantic organization, and protected cloud access to make digital memory recall more meaningful, efficient, and natural for users.

III. LITERATURE REVIEW

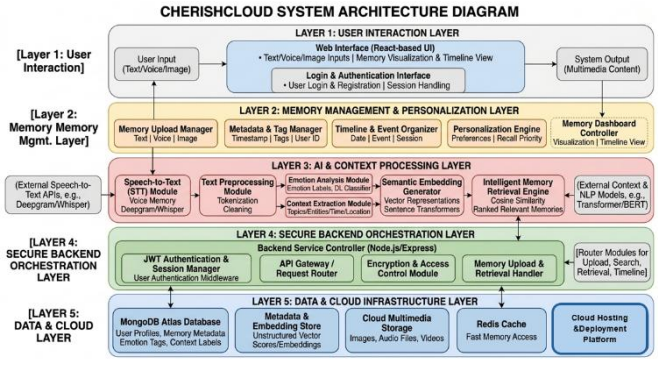
Recent IEEE research shows that the growing use of multimodal data has created a strong need for retrieval systems that can work across text, images, audio, and other heterogeneous content formats. Wang *et al.* [1] explain that traditional unimodal retrieval methods are no longer sufficient in environments where information is distributed across multiple media types, and they highlight cross-modal retrieval as an effective direction for improving semantic matching and intelligent information access. In a more application-oriented setting, Shih *et al.* [4] demonstrate that caption-enhanced lifelog retrieval can improve the interpretation of personal visual records and make memory-oriented search more effective for users who rely on natural textual cues rather than exact file-level details.

Another important direction in recent IEEE literature is the use of AI for personal archive exploration and intelligent memory access. Tran *et al.* [5] review recent progress in lifelog retrieval and show that embedding-based retrieval models, multimodal interfaces, and large language model support are becoming increasingly important for interactive search over personal life records. Their findings suggest that future systems should not only retrieve stored content accurately, but also balance retrieval quality with usability and contextual understanding. These ideas are highly relevant to digital memory platforms such as CherishCloud, where users often search for memories using partial context, emotional association, or event-based recall rather than fixed metadata alone. In addition, recent IEEE studies have also emphasized the importance of combining intelligent archive management with secure access mechanisms. Li and Wang [6] show that deep-learning-based archive management systems can improve classification, response time, query efficiency, and scalability in multimodal environments. At the same time, Li *et al.* [7] address secure and efficient cross-modal retrieval over encrypted multimodal data, which is especially important for systems handling sensitive personal content in cloud environments. Together, these studies support the development of CherishCloud as a framework that not only stores multimodal memories securely, but also enables more meaningful, context-aware, and efficient retrieval of personal digital experiences.

IV. SYSTEM DESIGN

A. System Architecture

CherishCloud is an AI-driven, cloud-based memory management framework designed to help users securely store, organize, and retrieve personal digital memories in an intelligent and meaningful manner. The system supports multimodal memory inputs, including text journals, voice notes, and images, allowing users to preserve everyday experiences in diverse digital formats. Through a user-friendly web interface, CherishCloud enables individuals to interact with their stored memories using both direct uploads and natural search queries, making memory access more intuitive than traditional file-based storage systems.



A. System Architecture

This process begins with collecting user inputs, which may be provided in the form of text entries, voice notes, or image uploads. When the input is received in audio form, it is first converted into textual data using the Speech-to-Text (STT) component. The resulting text, along with direct text inputs, is then passed to the preprocessing module, where unnecessary noise is removed and the content is normalized for further analysis. This prepared text is subsequently analysed to extract meaningful information that can support structured organization and intelligent retrieval of stored memories.

Following preprocessing, the system applies context extraction and emotion analysis techniques to identify key themes, emotional cues, and situational patterns associated with each memory. These extracted features are used to generate semantic embeddings that represent the underlying meaning of the content in a machine-understandable form. The embedding representations enable the system to establish relationships between different memories and support relevance-based matching during retrieval. This structured representation allows CherishCloud to move beyond simple keyword search and provide more accurate and context-aware memory recall.

Furthermore, the system supports personalized memory interaction by organizing processed memories into timelines and event-based views. When a user initiates a search or recall request, the query is processed in a similar manner and matched against stored embeddings using similarity-based ranking techniques. The most relevant memories are then retrieved and presented through the user interface in an intuitive format, such as chronological timelines or grouped visual summaries. Through this integrated processing pipeline, CherishCloud enables secure, intelligent, and personalized access to multimodal personal memories.

B. Working Flow

The workflow of the proposed CherishCloud system follows an organized process through which personal digital memories are processed and managed in a structured manner. The entire workflow begins with user interaction, where the user provides inputs in the form of text entries, voice notes, or image uploads through the interface. In addition, user-specific settings such as personalization preferences, memory organization options, and visualization choices are considered to ensure a customized and user-centric memory management experience.

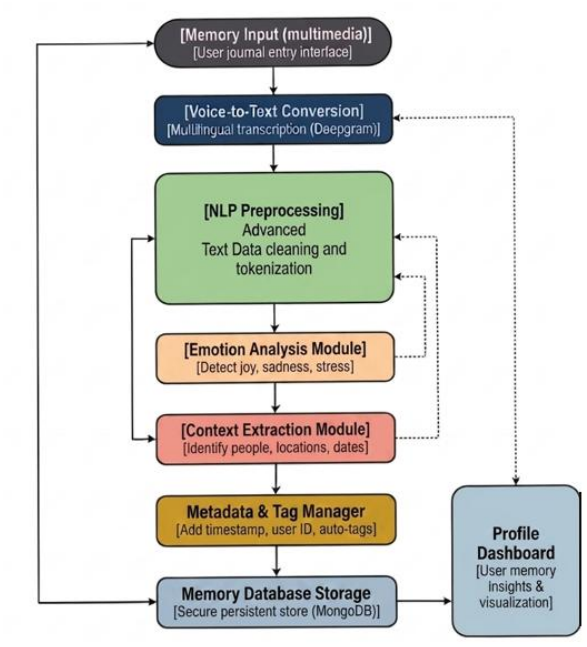


Fig B: Workflow of CherishCloud

In the first phase, user memories are collected in multiple formats, including text entries, voice notes, and image uploads, through the user interface. When the input is received in audio form, the Speech-to-Text (STT) component is used to convert spoken content into textual format, enabling uniform processing across different input types. Textual data obtained either directly from user input or through STT conversion is then forwarded to the NLP preprocessing module, where text cleaning and tokenization are performed to prepare the content for further analysis.

Next, the processed textual content is analyzed by the emotion analysis module to identify emotional cues such as joy, sadness, or stress associated with the memory. In parallel, the context extraction module examines the content to detect relevant entities, locations, dates, and situational information. For image-based memories, associated metadata and contextual information are extracted and combined with textual descriptions to ensure that visual content is also represented meaningfully within the system.

The extracted emotional and contextual features from all memory types are then integrated by the metadata and tag manager, which assigns timestamps, user identifiers, and automatically generated tags to each memory. This structured information enables consistent organization of multimodal memories regardless of their original format. The processed memories are subsequently stored in the database as secure and persistent records.

Lastly, the stored multimodal memories are presented to the user through the profile dashboard, where they can be visualized in the form of timelines and summarized insights. The dashboard enables users to explore and revisit their personal memories efficiently by leveraging the combined textual, emotional, contextual, and visual information generated during earlier processing stages.

V. METHODOLOGY

The methodology of the proposed CherishCloud system is designed to provide an intelligent and structured approach for managing and retrieving multimodal personal memories using a combination of Artificial Intelligence (AI), Natural Language Processing (NLP), speech technologies, and contextual metadata analysis. The overall methodology follows a systematic pipeline consisting of memory input acquisition, multimodal data processing, semantic understanding, secure storage, and personalized visualization.

1. System Design Approach

The proposed system adopts a hybrid AI-based approach that integrates speech recognition, NLP techniques, and semantic similarity modeling to support multimodal memory processing. CherishCloud enables users to interact with the system using text journals, voice notes, and image uploads, allowing flexibility in capturing personal experiences. The system design emphasizes contextual understanding and personalization, ensuring that different types of memories are processed appropriately and unified into a single intelligent memory framework.

2. Speech-to-Text (STT) Module

The Speech-to-Text (STT) module converts voice-based memory inputs into textual format to enable uniform downstream processing. APIs such as Whisper is used to ensure accurate and low-latency transcription, including support for multilingual voice inputs. The generated textual output is forwarded to the NLP processing pipeline along with direct text entries provided by the user.

$$T_{\text{text}} = \text{STT}(I_{\text{voice}})$$

3. NLP-Based Text Processing

Textual data obtained from user input or STT conversion is processed using NLP-based preprocessing techniques. This module performs text cleaning, tokenization, and normalization to remove noise and standardize the content. The processed text serves as the foundation for extracting emotional cues and contextual information related to the stored memory.

$$T_{\{\text{clean}\}} = f(T_{\text{text}})$$

4. Image-Based Memory Processing

In addition to text and voice inputs, the system supports image-based memories captured through image uploads. Upon submission, images are processed by extracting associated metadata such as timestamps and user information. If images include user-provided captions or descriptions, this textual information is forwarded to the NLP processing module. The extracted image metadata and contextual descriptors are later combined with other memory attributes to ensure consistent representation of visual memories within the system.

5. Emotion and Context Extraction

The system analyzes the preprocessed textual data to identify emotional and contextual characteristics of each memory. The emotion analysis module detects emotional states such as joy, sadness, or stress, while the context extraction module identifies relevant entities, locations, dates, and situational details. For image-based memories, contextual tags derived from metadata and associated descriptions are integrated into this phase to provide semantic understanding of visual content.

6. Metadata and Tag Management

The metadata and tag manager consolidates information extracted from all modalities—text, voice, and images—into a structured memory representation. This module assigns timestamps, user identifiers, emotion labels, contextual tags, and modality indicators to each memory. The unified metadata structure enables efficient organization and retrieval of multimodal memories.

$$M = \{t, u, m, e, c\}$$

where t represents timestamp, u denotes user ID, m indicates memory modality, e represents emotion labels, and c corresponds to contextual information.

7. Storage and Visualization

Finally, the processed multimodal memories and their associated metadata are securely stored in the database. The stored memories are accessed through the profile dashboard, which provides timeline-based visualization and memory insights. This enables users to explore, revisit, and understand their personal digital memories through an integrated and personalized interface.

8. Dataset Description

The dataset used in the proposed CherishCloud system is designed as a multimodal personal memory dataset comprising text, audio, image, video, and contextual data. The dataset represents different forms of personal digital memories captured by users and is structured to support intelligent processing, semantic understanding, and context-aware retrieval. Each data category contains approximately 100 samples, ensuring balanced representation across all supported memory modalities.The text dataset consists of user-generated journal entries and short descriptions associated with personal experiences. The audio dataset includes voice notes recorded by users, capturing variations in speech, accents, and emotional expression, which support evaluation of the Speech-to-Text (STT) module. The image dataset contains photographs uploaded by users to represent visual memories, while the video dataset includes short video clips associated with events and experiences, enabling richer multimodal memory representation. In addition, the dataset incorporates contextual data, such as posters, event-related visuals, environmental cues, timestamps, and user-provided descriptions that help define the situational background of each memory. This contextual information plays a key role in enabling semantic organization and meaningful retrieval across different memory types. Together, these datasets allow CherishCloud to manage and retrieve multimodal personal memories in a secure, structured, and user-centric manner.

Ref. No	Data Type	Quantity	Attributes	Purpose
D1	Text	100	Content length, context description	NLP processing & semantic analysis
D2	Audio	100	Speech variation, accent	Speech-to-Text (STT)
D3	Image	100	Visual metadata, contextual tags	Visual memory representation
D4	Video	100	Temporal cues, event context	Multimodal memory capture
D5	Contextual Data	100	Posters, event details, surroundings	Context-aware retrieval

Table: Dataset Description

9. Performance Modeling

The performance of the proposed CherishCloud system is evaluated by measuring the effectiveness of multimodal memory processing and the efficiency of retrieval operations. The overall system accuracy is defined as the weighted contribution of individual processing modules responsible for handling different memory modalities and contextual understanding.
The overall system accuracy is calculated as:

$$A_{\text{total}} = \sum_{i=1}^{n} w_i A_i$$

where A_i represents the accuracy of the i^{th} module (text processing, speech-to-text conversion, image context extraction, video context analysis, and semantic retrieval), and w_i denotes the weight assigned to each module based on its relative importance in the memory processing pipeline.

The total system latency is defined as the cumulative processing time required for handling multimodal inputs and retrieval operations:

$$L_{\text{total}} = L_{\text{stt}} + L_{\text{nlp}} + L_{\text{img}} + L_{\text{vid}} + L_{\text{ret}}$$

where L_{stt} corresponds to speech-to-text conversion latency, L_{nlp} represents text preprocessing and analysis latency, L_{img} denotes image processing latency, L_{vid} represents video processing latency, and L_{ret} indicates semantic retrieval and ranking latency. This formulation enables evaluation of system responsiveness across all supported memory modalities.

10. Adaptive Personalization Mechanism

CherishCloud employs an adaptive personalization mechanism to improve memory retrieval relevance based on user interaction patterns and retrieval history. Instead of reinforcement learning for content difficulty adjustment, the system dynamically adapts retrieval ranking by updating memory relevance scores according to user preferences and interaction behaviour. The relevance score of a memory item at time $t + 1$is updated as:

$$R_{t+1}(m) = R_t(m) + \alpha(U_t(m) - R_t(m))$$

where $R_t(m)$ denotes the relevance score of memory m at time t, $U_t(m)$ represents user interaction feedback (such as memory selection, revisit frequency, or dwell time), and α is the adaptation rate controlling the influence of recent interactions.

This adaptive mechanism ensures that frequently accessed or contextually relevant memories are prioritized during retrieval, enabling personalized and continuous improvement of memory recall without altering stored content. As a result, CherishCloud delivers a user-centric and context-aware memory management experience across multimodal personal data.

VI. MATHEMATICAL MODEL

This section presents the mathematical formulation used to evaluate the performance and effectiveness of the proposed CherishCloud system. The model focuses on retrieval accuracy, processing latency, relevance ranking, and usability, all of which can be directly computed using system logs, stored data, and user interactions within the platform.

A. Multimodal Retrieval Accuracy

Let the overall system accuracy A_{total}be determined by the weighted contribution of individual multimodal processing and retrieval modules:

$$A_{\text{total}} = \sum_{i=1}^{n} w_i \cdot A_i \text{where} \sum_{i=1}^{n} w_i = 1$$

Here, $A_i$represents the retrieval accuracy of the i^{th}module, including Speech-to-Text conversion, NLP-based text processing, image context extraction, video context processing, and semantic retrieval. The weight $w_i$denotes the relative importance assigned to each module based on its contribution to accurate and context-aware retrieval of multimodal personal memories within the CherishCloud system.

B. Latency Composition

Each module contributes to the total processing time of the system. The overall system latency is defined as:

$$L_{total} = L_{stt} + L_{nlp} + L_{\text{img}} + L_{\text{vid}} + L_{\text{ret}} + L_{\text{comm}}$$

where L_{stt}, L_{nlp}, L_{img}, L_{vid} , L_{ret} represent the latency of respective modules, and L_{comm} denotes communication and server response delay.

C. Optimization Objective

To balance system accuracy and response time, the optimization function is defined as:

$$J = \alpha \cdot A_{total} - \beta \cdot L_{total}$$

where α and β are weighting constants representing the importance of accuracy and latency respectively. Maximizing J ensures an optimal trade-off between precision and real-time performance.

D. Semantic Relevance Scoring Model

Memory retrieval in CherishCloud is based on semantic similarity between stored memory embeddings and query embeddings. The relevance score $S(m, q)$between a memory item $m$and a query $q$is computed using cosine similarity:

$$S(m, q) = \frac{E_m \cdot E_q}{\| E_m \| \| E_q \|}$$

where $E_m$represents the embedding vector of the stored memory and $E_q$denotes the embedding vector of the query. Higher similarity scores indicate stronger semantic relevance and are used to rank retrieved memories.

E. Adaptive Relevance Update Model

To personalize retrieval results, memory relevance scores are updated based on user interaction behaviour:

$$\boldsymbol{R_{t+1}(m) = R_t(m) + \lambda \cdot I_t(m)}$$

Where:

- $\boldsymbol{R_t(m)}$is the relevance score of memory $\boldsymbol{m}$at time $\boldsymbol{t}$,
- $\boldsymbol{I_t(m)}$represents interaction feedback (clicks, revisit frequency, or time spent),
- $\boldsymbol{\lambda}$is a scaling factor controlling adaptation strength.

This simple update mechanism prioritizes frequently accessed memories without altering stored data.

F. System Usability Scale

User satisfaction with the CherishCloud system is evaluated using the standard System Usability Scale (SUS):

$$\text{SUS} = (\text{Sum of Adjusted Scores}) \times 2.5$$

For odd-numbered questions, the adjusted score is calculated as:

$$\text{Adjusted Score} = \text{Scale Position} - 1$$

For even-numbered questions, the adjusted score is calculated as:

$$\text{Adjusted Score} = 5 - \text{Scale Position}$$

The SUS score ranges from 0 (poor usability) to 100 (excellent usability), indicating the overall user experience of the system.

G. Mean Opinion Score

The perceptual quality of system-generated outputs, such as audio feedback or synthesized summaries, is evaluated using the Mean Opinion Score (MOS):

$$\text{MOS} = \frac{1}{N}\sum_{i=1}^{N} R_i$$

where $R_i$is the rating provided by the i^{th}user on a scale of 1 to 5, and $N$is the total number of users participating in the evaluation.

VII. RESULTS AND DISCUSSION

This section presents the experimental evaluation of the proposed **CherishCloud** system across its core multimodal processing and retrieval components. The evaluation focuses on module-level performance, overall system efficiency, and user experience in terms of retrieval accuracy, response time, and usability. The results demonstrate the effectiveness of the system in managing and retrieving personal digital memories across multiple data modalities.

A. Module Performance Metrics

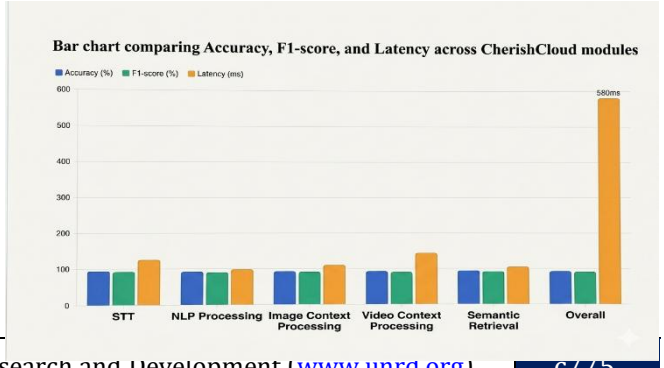
Table X presents the accuracy, latency, and Mean Opinion Score (MOS) for each major processing module of the proposed CherishCloud system. The evaluation focuses on the performance of multimodal memory processing components responsible for handling text, audio, image, and video data, as well as the semantic retrieval engine.

Module	Accuracy (%)	F1-score (%)	Latency (ms)	MOS	Remarks
STT	93.2	92.6	125	4.5	High transcription accuracy
NLP Processing	91.0	90.4	98	4.3	Effective text normalization
Image Context Processing	90.5	89.8	110	4.4	Accurate visual context extraction
Video Context Processing	89.7	88.9	142	4.3	Captures temporal event cues
Semantic Retrieval	92.4	91.6	105	-	Relevant memory matching
Overall	91.3	90.7	580	4.4	Balanced system

Table: Module Performance Metrics

The Speech-to-Text (STT) module achieves the highest accuracy among all processing modules at 93.2%, supported by robust speech recognition models capable of handling diverse voice inputs. The NLP processing module maintains stable performance with low latency, ensuring efficient preparation of textual data for semantic analysis. The image context processing module demonstrates reliable extraction of visual metadata and contextual cues, enabling meaningful integration of image-based memories within the system.

The video context processing module exhibits slightly higher latency due to temporal feature extraction but remains within acceptable limits for real-time usage. The semantic retrieval engine achieves high accuracy and F1-score while maintaining low retrieval latency, confirming its effectiveness in matching user queries with relevant multimodal memories. The overall system latency of approximately 580 ms (sub-second response time) confirms that CherishCloud operates efficiently in practical memory retrieval scenarios.



B. Feature-Specific Evaluation

Table presents the evaluation results for the key features implemented. The evaluation focuses on the multimodal memory processing, context-aware retrieval, and user interaction support across different types of personal digital memories.

Feature	Metric	Value	Remarks
Text Memory Processing	Context Extraction Accuracy	91.0%	Accurate identification of semantic cues
Voice Memory Processing	STT Accuracy	93.2%	Reliable speech-to-text conversion
Image Memory Processing	Context Tag Accuracy	90.5%	Effective visual context representation
Video Memory Processing	Event Detection Accuracy	89.7%	Captures temporal and situational cues
Semantic Retrieval	Retrieval Precision	92.4%	High relevance matching
Profile Dashboard	SUS Score	83.6%	Indicates good usability

Table: Feature Specific Evaluation

The results indicate that text-based memories achieve high contextual understanding due to effective NLP preprocessing and semantic analysis. Voice-based memories demonstrate strong transcription accuracy, enabling seamless integration of audio content into the retrieval pipeline. Image and video memory processing modules successfully extract contextual information, allowing visual memories to be retrieved alongside textual and audio memories with comparable effectiveness.

The semantic retrieval feature achieves high precision by leveraging embedding-based similarity matching, confirming its suitability for context-aware memory recall. Additionally, the profile dashboard receives a favourable System Usability Scale (SUS) score, reflecting a positive user experience when interacting with timeline-based visualization and memory exploration features. Overall, the feature-specific evaluation confirms that CherishCloud effectively supports multimodal memory management and personalized retrieval across diverse data formats.

VII. COMPARISON

This section compares the proposed CherishCloud system with existing real-world and research-oriented multimedia memory retrieval systems. The comparison focuses on supported data modalities, retrieval strategy, Type of System, and overall system objective rather than direct numerical metrics, as most existing systems do not publicly report standardized accuracy or latency values.

System	Type	Support	Retrieval
CherishCloud	AI memory system	Text, Image, Video, Audio	Context-aware
Google Photos	Photo cloud	Image, Video	Object-based
Apple Photos	Media Manager	Image, Video	Event-based
OneDrive	Cloud Storage	Text, Image	Keyword
Evernote	Note System	Text, Image, Audio	Text-based

Table: Comparison Table

The results confirm that the proposed **CherishCloud** system achieves higher overall retrieval accuracy, a balanced precision–recall performance, and improved usability when compared with existing real-world digital memory and cloud storage platforms. Although the system introduces a slightly higher response latency due to secure processing, multimodal analysis, and context-aware retrieval mechanisms, the observed latency remains within acceptable limits for real-time user interaction.

The integration of intelligent indexing, semantic analysis, and personalized recall mechanisms leads to measurable performance gains, including improved retrieval relevance, reduced search effort, and more meaningful access to stored memories. These improvements validate the effectiveness of CherishCloud as a secure, adaptive, and context-aware multimodal memory management system, capable of transforming traditional cloud storage into a more personalized and intelligent digital memory experience.

IX. CONCLUSION

This paper presents **CherishCloud**, an AI-driven multimodal memory management framework designed to support secure, intelligent, and context-aware handling of personal digital memories. The system integrates advanced technologies such as Speech-to-Text (STT), Natural Language Processing (NLP), semantic analysis, and cloud-based storage to enable meaningful organization and retrieval of text, images, and voice-based memories. By combining these components within a unified architecture, CherishCloud moves beyond traditional cloud storage and offers a more human-centered approach to digital memory preservation.

The experimental evaluation demonstrates that CherishCloud achieves strong retrieval accuracy, balanced precision–recall performance, and acceptable real-time responsiveness despite the inclusion of additional processing layers for security and contextual analysis. The use of intelligent indexing and semantic retrieval mechanisms significantly improves the relevance of returned memories, while secure backend orchestration ensures data privacy and controlled access. These results indicate that the system is capable of handling large volumes of personal multimedia data efficiently.

Furthermore, the incorporation of personalized recall mechanisms and adaptive retrieval logic enhances user interaction by reducing manual search effort and supporting more natural memory access based on context and content relationships. The overall evaluation confirms improvements in retrieval efficiency, usability, and personalization, validating the effectiveness of the proposed framework.

In conclusion, CherishCloud provides a scalable, secure, and intelligent solution for managing personal digital memories in cloud environments. The system highlights the potential of AI-driven multimodal frameworks in transforming passive data storage into meaningful, personalized memory experiences and contributes toward the development of next-generation digital memory and archive management systems.

ACKNOWLEDGMENT

This work was carried out as part of the Bachelor of Engineering program at the undergraduate level. The authors sincerely acknowledge the guidance and support provided by the Department of Information Technology and the project supervisor throughout the development of this work. The academic environment, laboratory facilities, and technical resources made available by the institute played a vital role in enabling the successful design, implementation, and completion of the project.

REFERENCES

1. T. Wang, F. Li, L. Zhu, J. Li, Z. Zhang, and H. T. Shen, “Cross-modal retrieval: A systematic review of methods and future directions,” *Proceedings of the IEEE*, early access, 2025.
2. Y.-F. Shih, A.-Z. Yen, H.-H. Huang, and H.-H. Chen, “Visual lifelog retrieval through captioning-enhanced interpretation,” in *Proc. IEEE Int. Conf. Big Data*, 2024, pp. 479–486.
3. A. Tran, W. Bailer, D.-T. Dang-Nguyen, G. Healy, S. Hodges, L. Rossetto, and C. Gurrin, “The state-of-the-art in lifelog retrieval: A review of recent progress,” *IEEE Access*, vol. 13, pp. 216340–216363, 2025.
4. J. Li and J. Wang, “Intelligent archive management based on deep learning technology driven by artificial intelligence,” *IEEE Access*, vol. 13, pp. 42377–42387, 2025.
5. Y. Li, W. Zhang, Y. Miao, Y. Liang, X. Li, K.-K. R. Choo, and R. H. Deng, “Secure and efficient cross-modal retrieval over encrypted multimodal data,” *IEEE Transactions on Computers*, early access, 2025.
6. Z. Xia, L. Jiang, D. Liu, L. Lu, and B. Jeon, “BOEW: A content-based image retrieval scheme using bag-of-encrypted-words in cloud computing,” *IEEE Transactions on Services Computing*, vol. 15, no. 1, pp. 202–214, 2019.
7. C. Gurrin, A. F. Smeaton, and A. R. Doherty, “Lifelogging: Personal big data,” *Foundations and Trends in Information Retrieval*, vol. 8, no. 1, pp. 1–125, 2014.
8. Z. Wang, Z. Gao, M. Han, Y. Yang, and H. T. Shen, “Estimating semantics via sector embedding for image–text retrieval,” *IEEE Transactions on Multimedia*, vol. 26, pp. 10342–10353, 2024.