ProRAG: Towards Reliable and Proficient AIGC-Based Digital Avatar

Yongkang Zhou[1], Muyang Yan[1], Junjie Yao[1,2✉], and Gang Xu[1,3]

[1] East China Normal University, Shanghai.
[2] Shanghai Key Laboratory of Trustworthy Computing.
[3] National Trusted Embedded Software Engineering Technology Research Center.
junjie.yao@sei.ecnu.edu.cn

Abstract. The concept of a Virtual Human represents an advanced interactive interface that bridges users with digital information, offering an increasingly realistic experience. Recent breakthroughs in Large Language Models (LLMs) and AI-Generated Content (AIGC) have significantly improved the lifelike nature of virtual humans, making them increasingly indistinguishable from real humans. However, this rapid progress raises significant concerns regarding the ethical implications and the reliability of virtual human interactions, particularly in high-stakes, domain-specific scenarios where factual accuracy and trustworthiness are paramount.

In response to these challenges, we introduce ProRAG, a novel framework designed to enhance the trustworthiness and reliability of digital avatars. ProRAG combines domain-specific LLMs with innovative strategies to address key challenges such as hallucinations, computational inefficiency, and context stability. Our approach integrates a multimodal knowledge base, consisting of textual, visual, and auditory data, to improve retrieval accuracy and content consistency. Furthermore, ProRAG supports multimodal digital human interactions, facilitating voice, visual, and text communication, which ensures high trust for critical applications. By leveraging adaptive data representation techniques, ProRAG resolves the "Lost in the Middle" challenge, enhancing hallucination suppression and promoting structured knowledge integration. This framework is designed to be scalable and versatile, demonstrating its potential across diverse domains such as education, cultural preservation, and legal consultation, while ensuring the generation of reliable, context-aware content in mission-critical decision-making environments.

Keywords: Large Language Models · Retrieval Augmented Generation · Digital Avatar · Knowledge Integration · Multi-modal Interaction

1 Introduction

Recent advancements in AI-driven content generation, particularly through Artificial Intelligence-Generated Content (AIGC), have significantly impacted the

development of digital avatars. These avatars, which integrate multimodal capabilities—such as speech, text, and visuals—are increasingly being deployed in interactive applications across diverse domains. However, ensuring their reliability and effectiveness in specialized tasks presents a range of challenges, particularly when it comes to handling domain-specific knowledge and reasoning.

Digital avatars, especially those powered by Large Language Models (LLMs), often struggle with the representation and application of specialized information. LLMs, while effective in many general tasks, are constrained by their parameterized knowledge representations, leading to limited semantic understanding and reduced logical reasoning capacity in specialized contexts. These limitations undermine their factual accuracy and reliability, particularly in knowledge-intensive domains.

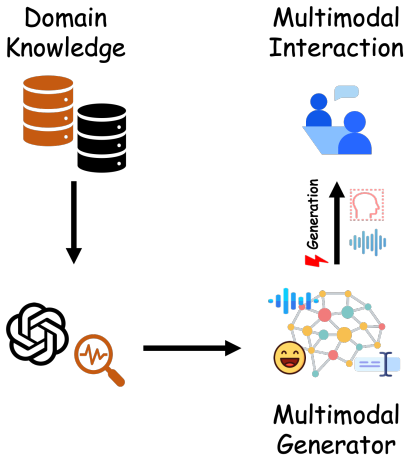


Fig. 1: Architecture of the ProRAG Framework.

When deployed in specialized domains, such as healthcare or legal consultation, LLM-based digital avatars face additional risks, including hallucinations during domain-specific queries that can result in inaccurate or unreliable outputs. Furthermore, achieving seamless alignment between modalities—speech, text, and visuals—requires not only extensive and diverse training data but also a robust system design, which can affect the system's adaptability and interaction reliability.

Retrieval-Augmented Generation (RAG) frameworks have been proposed as a promising solution by integrating external document retrieval into the content generation process [3] [12]. While RAG frameworks have shown potential in improving content generation, they face challenges such as the "Lost in the Middle" phenomenon [7], where lengthy retrieved documents compromise the coherence and relevance of the generated content [11]. Additionally, issues related

to the reliability of retrieved documents and the limited capacity to process structured data further constrain the effectiveness of RAG in domains requiring precise, rule-based reasoning.

Addressing these challenges requires balancing efficiency and trustworthiness. Systems must deliver low-latency, real-time responses while ensuring credibility through high-fidelity multimodal outputs. Features like natural speech synthesis, lip synchronization, and expressive facial gestures are crucial for user trust and engagement in interactive applications. Despite progress, current technologies face significant barriers in knowledge-intensive, high-reliability domains. Overcoming these limitations will require both technical innovation and strategic advancements in system design to realize the full potential of digital avatars in specialized applications.

To address these challenges, we propose a comprehensive solution: the Proficient and Reliable Framework for Multimodal Retrieval-Augmented Generation (ProRAG), as illustrated in Fig. 1. The framework is centered on three key enhancements:

(i) Comprehensive Data Representation: We propose an offline multimodal knowledge base integrating textual, visual, and auditory data for efficient document retrieval and knowledge verification, ensuring factual accuracy and contextual relevance.
(ii) Reliable LLM Answer Generation: To mitigate the "Lost in the Middle" phenomenon [7], we implement adaptive data governance and indexing to reduce hallucinations and improve knowledge integration, particularly in domain-specific tasks.
(iii) Efficient Avatar Response: Our system integrates voice, visuals, and text to optimize user engagement, ensuring seamless interactions with low latency, suitable for high-trust, multimodal applications.

ProRAG provides a framework to address challenges in AIGC-driven digital avatars, ensuring accuracy, relevance, and trustworthiness in content generation and multimodal interaction. It integrates a multimodal knowledge base, adaptive governance, and efficient response mechanisms to mitigate issues like the "Lost in the Middle" phenomenon [7] and optimize domain-specific reasoning. Its low-latency, high-efficiency design enhances user engagement, making it suitable for real-time, knowledge-intensive applications. ProRAG overcomes traditional LLM limitations and offers a scalable solution for complex, cross-domain scenarios.

2 System Architecture and Implementation

2.1 System Overview

This paper introduces ProRAG, a systematic framework for developing reliable, domain-adaptive, and efficient digital avatars, addressing critical challenges in large language models (LLMs), such as hallucinations and inaccuracies arising from parameterized memory. The ProRAG framework consists of two primary

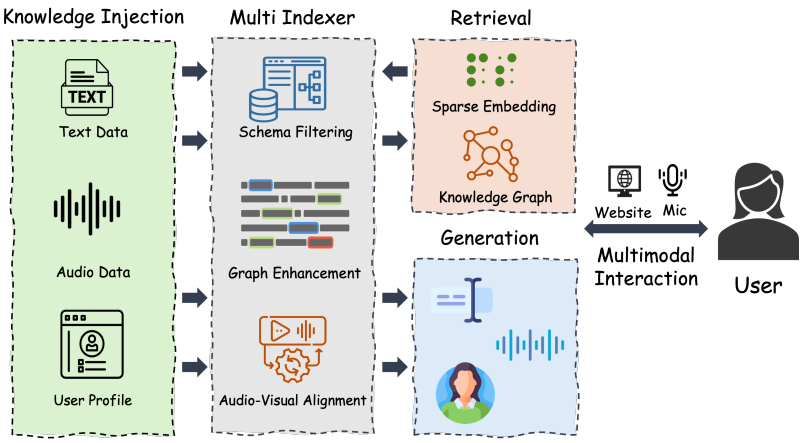


Fig. 2: Pipeline of the ProRAG for Knowledge Retrieval,Generation and Multimodal Interaction.

components: a system model for constructing domain-specific knowledge representations and an online response mechanism based on retrieval-augmented generation, enabling real-time avatar generation and interaction. This integration improves the reliability and contextual consistency of generated content, fostering trust in multimodal interactions.

The ProRAG system model extracts and structures domain-specific knowledge, as shown in Fig. 2. It employs techniques such as semantic segmentation and entity extraction to convert unstructured textual and multimodal data into a structured knowledge base for retrieval and generation tasks. Knowledge graph techniques capture entities and their relationships, while graph embedding and community detection refine the topological structure. This bottom-up approach mitigates the "Lost in the Middle" phenomenon [7], preserving critical contextual information. Sparse embeddings [8] emphasize semantically important features, improving retrieval accuracy and supporting complex knowledge extraction. In constructing digital avatars, ProRAG incorporates high-precision speech synthesis and lip synchronization, enhancing user immersion and trust.

The ProRAG online response mechanism optimizes retrieval and real-time avatar generation. By utilizing sparse embeddings, it enhances the retrieval process, retaining only semantically relevant information for generation. The system integrates structured knowledge from the knowledge graph during retrieval, capturing complex semantic relationships between textual elements. The generation process combines entity recognition with summary generation, ensuring accurate mapping from local text to global knowledge. The system's audio and visual models enable efficient, real-time audiovisual generation, facilitating reliable multimodal interactions.

ProRAG provides a robust RAG framework that addresses hallucinations and inaccuracies in LLMs. By combining the system model with the online response

mechanism, it improves content reliability, contextual consistency, and domain-specific knowledge management, enhancing the performance and credibility of digital avatar interactions in complex, domain-specific scenarios.

2.2 System Modules

The ProRAG framework integrates two core model categories: an efficient and reliable Retrieval-Augmented Generation (RAG) model for knowledge management and a high-fidelity avatar generation model for real-time multimodal synthesis.

The RAG model transforms unstructured multimodal data into structured knowledge through semantic segmentation, entity extraction, and knowledge graphs, ensuring robust representation of entities and their semantic relationships. Sparse embeddings and hierarchical graph embeddings improve retrieval efficiency and accuracy, while dynamic re-ranking algorithms maintain contextual relevance and semantic consistency of retrieved documents. This effectively addresses challenges such as long-tail knowledge and the "Lost in the Middle" phenomenon [7]. The avatar generation model produces high-fidelity multimodal outputs by integrating advanced audio-visual synthesis techniques. The BERT-VITS2 model generates natural, emotionally expressive speech adapted to specific roles via voice cloning. The Wav2Lip model ensures precise lip synchronization, while diffusion-based portrait generation and 3D facial construction enable lifelike expressions and movements. These modules work in synergy to deliver synchronized, credible avatar interactions in real-time.

Retrieval Enhancement Design: As detailed in the right panel of Fig. 2, the first step in ProRAG involves a segmentation method guided by semantic consistency, transforming unstructured text into structured units suitable for knowledge extraction and retrieval. This method preserves contextual semantics [2] [5] [9], capturing key entities and their relationships, thereby establishing a foundation for domain-specific knowledge management.

In role-adaptation scenarios, high textual similarity and complex semantic relationships within documents pose challenges for segmentation tasks. For instance, character biographies or career trajectories often feature highly similar semantics but include irrelevant information. Traditional dense embeddings struggle with such cases, leading to semantic noise and inaccuracies in retrieval and generation. To address this, we propose a semantic segmentation method optimized for role adaptation, which integrates context filtering and pronoun replacement to enhance text relevance and clarity. This approach ensures the accurate extraction of semantically relevant units from similar documents, improving the reliability of knowledge representation and generation while enhancing the system's adaptability to role-specific tasks.

To optimize knowledge base management and retrieval, ProRAG integrates sparse embeddings and hierarchical knowledge graphs as complementary technologies. Sparse embeddings highlight key semantic features while suppressing irrelevant information, mitigating issues of high textual similarity and reducing the "middle-loss" phenomenon [6] [1]. This enhances retrieval accuracy, in-

terpretability, and reliability in domain-specific tasks. Hierarchical knowledge graphs complement sparse embeddings by organizing knowledge across multiple granularities, from local entities to global semantics. Documents are segmented into semantic units, with entities, relationships, and assertions standardized, summarized, and disambiguated for precision. These entities are further structured into hierarchical communities using algorithms such as the Leiden algorithm, with Node2Vec embeddings enhancing graph structure and query performance [4]. This framework captures implicit cross-document relationships, enabling multi-level knowledge organization and dynamic updates.

By integrating semantic segmentation, sparse embeddings, and hierarchical knowledge graphs, ProRAG establishes an efficient knowledge governance methodology. Semantic segmentation ensures contextual relevance, sparse embeddings provide precise and interpretable representations, and knowledge graphs enable structured, multi-granularity management. This integrated approach systematically enhances semantic management and dynamic content generation, addressing complex domain-specific challenges in avatar systems.

Audio Model Training: The system employs an advanced generative model for speech synthesis training tailored to digital avatars. The training dataset consists of a manifest file (e.g., "vo_2.wav|XG|EN|Christmas is coming soon. Wishing everyone a Merry Christmas!") paired with corresponding raw audio files (e.g., v_n.wav). The raw audio data undergoes a meticulous preprocessing pipeline, which includes noise reduction, speech quality assessment, and comprehensive data cleaning. Audio samples exhibiting excessive background noise, unclear articulation, significant dialectal influence, or low word recognition accuracy are excluded, ensuring dataset integrity and reliability.

Subsequent to filtering, the audio files are segmented into individual sentences, with each segment aligned to its corresponding text, forming a standardized audio-text training set. This preprocessed dataset is then input into a generative speech synthesis model that utilizes a combination of deep neural networks and latent variable models. The text input is processed to extract semantic features, while the speech synthesis component generates high-quality speech waveforms through joint training. By leveraging shared latent semantic representations, the model integrates both speech and text during training, generating synthesized speech that is both natural and semantically accurate.

Video Model Training: Building upon the speech audio generated by the previous component, the system integrates a video synthesis model to synchronize the audio with video datasets. The mouth movements are precisely aligned with the audio, while facial expressions are dynamically generated from the video dataset. A facial animation model based on CNN and GAN ensures accurate lip synchronization, achieving high alignment between speech articulation and mouth movements.

To optimize system efficiency, the model preloads common speech and motion features, reducing processing time for subsequent tasks. During the integration process, image blending techniques, such as Poisson blending, are employed to refine boundary gradients, facilitating the seamless fusion of mouth movements

with the facial area in the video. This technique eliminates visible artifacts, ensuring a natural and coherent visual appearance. By combining these approaches, the system generates digital human facial expressions and articulation that are perfectly synchronized with speech content, significantly enhancing the realism and expressiveness of virtual avatars.

2.3 Online Response

Building on the overall models of the ProRAG system, the online response module is designed to enhance retrieval-augmented generation (RAG) for supporting real-time digital avatar interactions. By employing sparse embedding techniques and multi-layered retrieval strategies, the system efficiently processes domain-specific tasks, enabling the optimized integration of multimodal and structured data. These methodologies ensure that only semantically relevant information is retained for generation, addressing dynamic semantic requirements in complex domains. This capability allows the audio-visual model to perform high-precision, real-time avatar role reasoning and knowledge generation, facilitating seamless multimodal interactions with users.

Retrieval-Augmented Generation: ProRAG's online response benefits from its multi-layered RAG framework, as shown in the middle section of Fig. 2. This framewok allows for granular control over information extraction and processing. It is particularly effective for supporting multi-turn, multimodal dialogues, where the system must handle a variety of data types, from structured, pattern-based information to unstructured textual data. Sparse embedding technology ensures that the system prioritizes data with high semantic consistency, filtering out irrelevant or redundant content. This improves the interpretability of retrieved information and enhances retrieval efficiency, especially in high-precision and real-time scenarios.

ProRAG employs a dual-layer retrieval strategy, combining global and local retrieval. The global retrieval phase performs broad reasoning across the entire dataset, optimizing retrieval accuracy through community-level summarization and semantic relationship enhancement. This strategy is particularly useful for cross-domain queries involving multiple entities, where a comprehensive understanding of the broader context is necessary. In contrast, local retrieval focuses on extracting highly relevant information from specific communities, ensuring deep, contextually aligned queries. This multi-layered framework provides the system with flexibility, enabling it to address queries of varying complexity and granularity effectively. Furthermore, an advanced query generation mechanism integrates historical dialogue context, synthesizing multimodal content, including prior conversations and domain-specific pattern knowledge. This integration ensures that relevant information, retrieved from both general knowledge bases and domain-specific repositories, supports dynamic dialogues between the digital avatar and the user, regardless of the task's complexity or scope.

Avatar Generation:Building upon the insights regarding "hallucinations" in large language models (LLMs), the avatar generation process emphasizes the

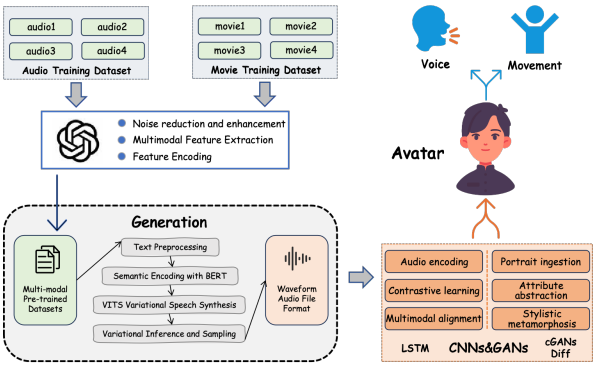


Fig. 3: Illustration of the Avatar Construction Process.

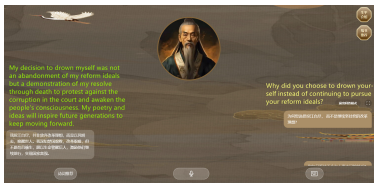
authenticity and consistency of both auditory and visual outputs, enabling seamless user interactions. As illustrated in the upper section of Fig. 3, the system leverages advanced voice cloning techniques to achieve context-adaptive voice generation, ensuring accurate emotional conveyance and vocal consistency. By utilizing the BERT-VITS2 model, the system excels in producing natural intonations, enhancing both the credibility of the avatar and the fluidity of user interactions. This approach mitigates inaccuracies in voice generation typically observed in LLMs, delivering richer, more emotionally expressive speech, thereby fostering user trust and engagement.

For visual presentation, as shown in the lower section of Fig. 3, the system integrates the Wav2Lip [10] model for precise lip synchronization, ensuring facial animations align seamlessly with audio output to enhance realism and immersion. Additionally, the portrait generation framework, based on a diffusion model, translates audio input into natural facial expressions. Combined with 3D facial construction, this facilitates lifelike facial movements and emotionally nuanced expressions during interactions. By harmoniously integrating these audio and visual components, the system effectively addresses challenges such as audio-visual desynchronization and emotional incongruence, providing a coherent and trustworthy interaction experience. This approach significantly improves the system's usability and reliability in dynamic user scenarios.

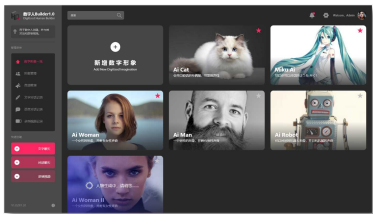
3 System Demonstrations

3.1 Applied Scenarios

Building on ProRAG, we developed a digital avatar demonstration system for the book domain, showcasing its capabilities in large-scale knowledge base construction, multimodal interaction, and personalized services. By processing extensive book data and applying ProRAG's knowledge extraction and filtering technologies, the system constructs a context-aware private knowledge base, enabling precise retrieval and personalized book recommendations. Through interaction



(a) Book Domain Digital Human



(b) Digital Human Builder

(c) Micro-Lesson Workshop

(d) Micro-Lesson Workshop

Fig. 4: Illustration Of Various Scenarios Showcasing The Application Of Our Digital Avatar.

history and preference analysis, the system tailors recommendations to user interests, demonstrating ProRAG's adaptability and scalability. Additionally, the system recreates historical figures as digital avatars, presenting their life stories and works for high-fidelity, real-time interactions in audio, video, or text formats, ensuring seamless and immersive engagement.

Designed for scalability, the system is adaptable to other knowledge-intensive domains, offering ultra-low latency in content generation and multimodal output. Thanks to ProRAG's efficient multimodal generation, the audio and video models occupy less than 1GB of GPU memory, enabling a 13B-parameter LLM to run on a single NVIDIA RTX 4090 GPU (24GB). This reduces reliance on cloud infrastructure, minimizes operational costs, and ensures robustness in knowledge management, user interaction, and personalization.

In the education domain, we developed the Micro-class Workshop, a system that generates digital teacher avatars and voice explanations from PowerPoint (PPT) content. Teachers can customize their digital avatars' voice, expressions, and gestures to closely match their real-life personas. Additionally, the Digital Person Builder supports the flexible creation and management of digital avatars by uploading multimodal data, enabling customization of both visual and auditory features for a wide range of applications.

3.2 Generation Quality

We compare ProRAG's retrieval mechanism with other methods, focusing on performance across the Alpaca and MSMARCO datasets. As shown in Table 1, ProRAG outperforms other indices in Recall@1 and Recall@5 on the

Alpaca dataset, with HNSW achieving a marginally higher score for Recall@10. ProRAG's consistent performance across all recall metrics can be attributed to its third-layer search mechanism, which re-ranks results using a top-k algorithm to prioritize the most relevant outputs.

Table 1: Default Evaluation of Retrieval Mechanisms Across Two Datasets

Datasets	**Retrieve method**	**Recall@1**	**Recall@5**	**Recall@10**
Alpaca	pqindex (milvus)	75.75%	79.94%	80.25%
	Chroma	94.23%	94.30%	94.40%
	ProRAG	95.65%	95.65%	95.65%
MSMARCO	pqindex (milvus)	67.19%	70.13%	71.86%
	Chroma	90.11%	90.39%	90.56%
	ProRAG	90.97%	90.97%	90.97%

For the larger MSMARCO dataset, retrieval accuracy declines across all methods due to its increased complexity. While the IVFPQ algorithm could not be tested due to memory constraints, ProRAG, leveraging a quantization-based representation, successfully handled the dataset and maintained robust performance, particularly excelling in Recall@5.

3.3 System Efficiency

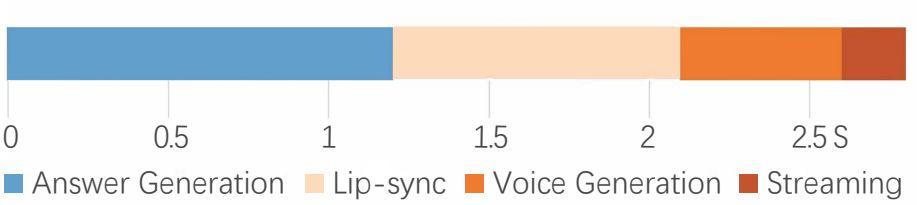


Fig. 5: Time Latency of ProRAG Components.

Additionally, we evaluate the retrieval quality of the methods, as detailed in Fig 6.In terms of retrieval speed, ProRAG demonstrates clear advantages. On the Alpaca dataset, it achieves 0.29 ms/query, significantly outperforming pqindex

(5.31 ms/query) and Chroma (0.84 ms/query), while remaining competitive with the faster HNSW index (0.003 ms/query). On the MSMARCO dataset, ProRAG maintains strong performance at 5.16 ms/query, outperforming Chroma (20.6 ms/query) and pqindex (35.82 ms/query). These results highlight ProRAG's ability to balance high-speed retrieval with adaptability across diverse datasets, underscoring its scalability, robustness, and suitability for reliable and efficient retrieval tasks. Fig. 5 illustrates the distribution of time latency across the key

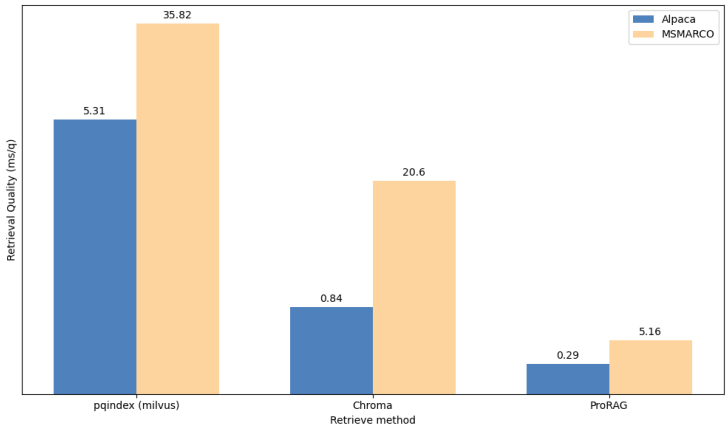


Fig. 6: Efficiency Evaluation Across Two Datasets

components of the ProRAG framework, highlighting the temporal efficiency of its multimodal processing pipeline. The chart reveals that the majority of the latency is attributed to Answer Generation, which takes approximately 1 second. Following this, the combined processes of Lip-sync and Voice Generation account for an additional 1 second, with Lip-sync being slightly faster. The final step, Streaming, exhibits the lowest latency, taking less than 0.5 seconds. This breakdown demonstrates the framework's ability to efficiently manage multimodal tasks while maintaining low latency, making it well-suited for real-time applications that require fast and seamless user interactions.

.

4 Conclusion

ProRAG presents a robust framework for Artificial Intelligence Generated Content (AIGC)-based digital avatars, addressing key challenges in large language models and multimodal generation, including hallucinations, inaccuracies, and vividness. By improving content credibility and consistency, the framework is designed to meet the requirements of high-stakes application scenarios. Its innovative retrieval strategy, which combines sparse embeddings with semantic

re-ranking, ensures the accurate retrieval of contextually relevant documents. Furthermore, by integrating multimodal generation, ProRAG enables real-time interactions through precise speech synthesis and synchronized facial expressions. This work contributes significantly to the advancement of reliable digital avatar systems in the context of Large Language Models and AIGC, laying the foundation for their future development.

Acknowledgments

This work is sponsored by National Key Research and Development Program of China (2022ZD0120302), and National Natural Science Foundation of China (61972151).

References

1. Doshi, M., Kumar, V., Murthy, R., Sen, J., et al.: Mistral-splade: Llms for for better learned sparse retrieval. arXiv preprint arXiv:2408.11119 (2024)
2. Edge, D., Trinh, H., Cheng, N., Bradley, J., Chao, A., Mody, A., Truitt, S., Larson, J.: From local to global: A graph rag approach to query-focused summarization. arXiv preprint arXiv:2404.16130 (2024)
3. Fan, W., Ding, Y., Ning, L., Wang, S., Li, H., Yin, D., Chua, T.S., Li, Q.: A survey on rag meeting llms: Towards retrieval-augmented large language models. In: Proc. of SIGKDD. pp. 6491–6501 (2024)
4. Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proc. of SIGKDD. pp. 855–864 (2016)
5. Hu, Y., Lei, Z., Zhang, Z., Pan, B., Ling, C., Zhao, L.: Grag: Graph retrieval-augmented generation. arXiv preprint arXiv:2405.16506 (2024)
6. Kong, W., Dudek, J.M., Li, C., Zhang, M., Bendersky, M.: Sparseembed: Learning sparse lexical representations with contextual embeddings for retrieval. In: Proc. of SIGIR. pp. 2399–2403 (2023)
7. Liu, N.F., Lin, K., Hewitt, J., Paranjape, A., Bevilacqua, M., Petroni, F., Liang, P.: Lost in the middle: How language models use long contexts. TACL **12**, 157–173 (2024)
8. Nguyen, H.V., Patel, V.M., Nasrabadi, N.M., Chellappa, R.: Sparse embedding: A framework for sparsity promoting dimensionality reduction. In: ECCV. pp. 414–427. Springer (2012)
9. Peng, B., Zhu, Y., Liu, Y., Bo, X., Shi, H., Hong, C., Zhang, Y., Tang, S.: Graph retrieval-augmented generation: A survey. arXiv preprint arXiv:2408.08921 (2024)
10. Prajwal, K., Mukhopadhyay, R., Namboodiri, V.P., Jawahar, C.: A lip sync expert is all you need for speech to lip generation in the wild. In: Proc. of ACM Multimedia. pp. 484–492 (2020)
11. Xu, Z., Jain, S., Kankanhalli, M.: Hallucination is inevitable: An innate limitation of large language models. arXiv preprint arXiv:2401.11817 (2024)
12. Zhao, P., Zhang, H., Yu, Q., Wang, Z., Geng, Y., Fu, F., Yang, L., Zhang, W., Jiang, J., Cui, B.: Retrieval-augmented generation for ai-generated content: A survey. arXiv preprint arXiv:2402.19473 (2024)