IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Secure On-Premise Multimodal AI Framework for Automated PowerPoint Generation Using Retrieval-Augmented Large Language Models

Chetan Singh [1], Vrajesh Vadi [2], Sunny W Thakare[3]

[1]Department of Computer Science and Engineering, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India

[2] IT Department, Larsen and Toubro Energy - CarbonLite Solutions, Vadodara, Gujarat, India

[3] Department of Computer Science and Engineering, Parul Institute of Technology, Parul University, Vadodara, Gujarat, India

Abstract

Although Large Language Models (LLMs) have shown great promise in automated content creation, they are still vulnerable to factual inconsistencies and hallucinations in knowledge-intensive jobs. By adding outside knowledge sources to the generating process, Retrieval-Augmented generating (RAG) improves factual accuracy and contextual grounding. In order to create richer content, multimodal generative models allow for the simultaneous integration of text and picture synthesis. Despite these developments, the majority of current systems rely on cloud-based deployment, which raises issues with infrastructure reliance, data privacy, and regulatory compliance.

In order to generate PowerPoint automatically, this research suggests a safe, completely offline multimodal AI platform that combines a locally installed LLM with diffusion-based picture synthesis and a vector-based retrieval system. The proposed architecture is designed to be deployable on-site without depending on any costly external APIs or GPUs. Based on experimental evaluation, it is found that augmentation in retrieval improves factual grounding and structural coherence over LLM-only baselines, while multimodal integration improves user-perceived slide quality. This proposed architecture is a solution for institutional and commercial environments that require secure document automation using AI technology.

Keywords — Large Language Models, Retrieval-Augmented Generation, Multimodal AI, Offline Deployment, Automated Presentation Generation, Hallucination Mitigation.

I. INTRODUCTION

The field of natural language production and interpretation has been significantly enhanced by Large Language Models (LLMs). These models have made it possible to automate a number of tasks and activities, such as conversational systems, question answering systems, summarization systems, and even structured document writing systems. The zero-shot and few-shot performances of these systems are quite robust, as demonstrated by the use of transformer-based systems for these tasks. However, these systems are more likely to produce hallucinations when generating content, which may be factually

incorrect or unverifiable even if the content is coherent and grammatically correct [1] [2]. This is a major issue in the field of structured document writing systems like automated presentation systems.

Retrieval Augmented Generation (RAG) is a prominent technique that leverages the capabilities of parametric language models and the efficacy of information retrieval systems to reduce hallucinations and enhance the factual accuracy of generated content [3] [4]. The RAG technique is successful in enhancing the capabilities of the language models by retrieving relevant articles from the vectorized knowledge base and incorporating them into the content generation process. The efficacy of retrieval augmentation in improving the accuracy of generated content and minimizing hallucinatory errors in various knowledge-based tasks was demonstrated in previous evaluation metrics [5]. The current evaluation metrics highlight the need to conduct RAG evaluation by emphasizing the significance of retrieval relevance and hallucinatory error detection in evaluating RAG models [6] [7].

Simultaneously, multimodal generative AI has advanced quickly. High-resolution visual synthesis conditioned on textual cues is made possible by diffusion-based picture generation models, especially latent diffusion architectures [8]. End-to-end automated content pipelines that produce structured papers with contextual images are made possible by the combination of textual production with visual synthesis. Multimodal integration improves slide clarity, visual engagement, and overall communication efficacy for presentation automation.

Despite these developments, cloud-based infrastructures are used to implement the majority of RAG-driven and multimodal generating systems. Although successful, cloud implementation creates issues with latency, vendor reliance, data privacy, and regulatory compliance—particularly in business, governmental, and academic settings managing sensitive data. In order to overcome these issues, recent surveys highlight the increasing demand in privacy-preserving and locally deployable LLM systems [9]. Fully offline, secure, multimodal infrastructures created especially for automated presentation production, however, have not received much attention.

We provide a safe on-premise multimodal AI system for automated PowerPoint creation in order to close this gap. It incorporates:

- A locally implemented Large Language Model for the creation of organized slide text
- A pipeline for Retrieval-Augmented Generation (RAG) that uses an offline vector database
- A semantic retrieval system for contextual grounding that is domain-specific
- A module for creating images based on diffusion for contextual visual synthesis
- Programmatic slide formatting for automated PowerPoint creation

The suggested architecture functions completely offline, in contrast to cloud-dependent solutions, guaranteeing data secrecy and deployment viability in contexts with limited resources without requiring expensive GPU equipment.

Problem Statement

Despite recent progress in retrieval-enhanced generative AI, Applying LLMs to the creation of organized presentations still presents a number of difficulties:

1.Residual Hallucination: Due to inadequate retrieval alignment or integration processes, hallucinations may continue even after retrieval augmentation [6] [7].

2.Cloud Dependency and Data Privacy Risks: Since many cutting-edge systems rely on external APIs, sensitive operations may face privacy and compliance issues [9].

3. Absence of Offline Multimodal Pipelines: There aren't many papers that discuss entirely offline architectures that combine contextual picture synthesis and retrieval-grounded text production.

4. Evaluation in Realistic Deployment Environments: The evaluation of RAG systems needs to consider aspects such as "computational viability, factual correctness, coherence, and retrieval precision" [6].

Thus, the main research question in the current study is:

We propose an end-to-end on-premise system and evaluate it against LLM-only baselines to answer the following research question. The results obtained from the experiments reveal that although the slide quality is enhanced by the integration, retrieval augmentation aids in enhancing factual correctness and coherence.

II. RELATED WORK

The improvement of the reliability, accuracy, and deployability of Large Language Models (LLMs) is the primary focus of existing research in generative AI models. This section presents a review of existing research in the following major areas: multimodal generating systems, retrieval-based generative models, privacy-preserving deployable frameworks, and hallucinations in LLMs. All these areas are relevant to this research

A. Large Language Models and Hallucinations

One major drawback associated with modern language models is hallucinations. The generation of texts that are syntactically correct but factually incorrect or unsupported by evidence is known as hallucinations. Extensive research has classified hallucinations as intrinsic and extrinsic, which are associated with a lack of knowledge and errors in internal model reasoning, respectively [1]. This is also supported by studies such as TruthfulQA, which emphasize the need for better grounding in language models by indicating that large models tend to perpetuate common human misconceptions [2].

Several mitigation strategies, such as rapid engineering, the addition of external knowledge, and reinforcement learning, have been proposed in recent studies. In addition, to improve the interpretability and diagnose errors, mechanistic analysis tools have also been proposed to detect hallucinations within the internal model representations [7]. Despite these advances, hallucinations remain an important drawback, especially for organized content generation tasks such as creating technical documents or presentations.

B. RAG, or retrieval-augmented generation

One of the best approaches to improve factual grounding in LLM outputs is Retrieval-Augmented Generation (RAG). In order to enable a model to retrieve relevant texts during inference, researchers have put forward a concept of combining parametric language models with non-parametric external memory in earlier studies [3]. Similarly, a positive impact was demonstrated by the REALM approach by incorporating retrieval mechanisms during pretraining itself.

Retrieval enhancement dramatically lowers hallucinations and increases contextual relevance in conversational systems and question-answering tasks, according to later study [5]. More recent research concentrates on enhancing evaluation techniques, context integration tactics, and retrieval quality. To improve retrieval coverage and lower hallucinated outputs, for instance, new research on dehallucination approaches investigates simultaneous context expansion [6]. To further increase grounding dependability, new research suggests hybrid retrieval techniques that include symbolic reasoning and dense vector search [10].

Recent years have seen the publication of extensive surveys that emphasize the rising significance of systematic assessment for RAG pipelines, with a focus on measures like generation faithfulness, retrieval accuracy, and grounding integrity [11]. These investigations highlight the necessity of robust designs that reliably and scalably combine generation and retrieval.

C. Multimodal Generative Models

Multimodal generative AI has seen rapid development along with advancements in language modeling. Currently, the main trend in generating images of high quality is through the use of diffusion-based generative models. By utilizing compressed latent space with excellent image quality, latent diffusion models are much more efficient. This efficiency has been seen in terms of practical applications in various domains such as content development, design, and development of training materials. Multimodal models currently provide organized multimedia outputs through a combination of image synthesis and textual instructions. However, a majority of multimodal models currently in use require access to large-scale GPU infrastructure as well as cloud-based infrastructure, which may not be suitable in a privacy context.

The multimodal integration provides an opportunity to generate textual slides along with additional visual elements such as diagrams or drawings within a presentation creation context. Fully integrated multimodal pipelines designed for automated presentation generating procedures have not received much attention, despite their promise.

D. Privacy-Preserving and On-Premise AI Deployment

Concerns around data privacy and regulatory compliance have drawn a lot of attention as generative AI systems are incorporated more deeply into business processes. Modern LLM systems sometimes rely on cloud-based inference platforms or external APIs, which poses vulnerabilities when handling sensitive data. The significance of privacy-preserving AI frameworks that provide secure data handling and local model deployment is highlighted by recent studies [9]. As to overcome the obstacles, strategies including federated learning architecture, secure inference architectures and also locally LLM deployable models in forward [13].

Despite these advancements, comparatively little research has been done on fully offline multimodal systems that integrate automated document or presentation creation with retrieval-augmented generation. The majority of earlier research focuses on either multimodal synthesis or retrieval-based text production separately.

E. Research Gap

There are still a number of gaps in the state of research, according to the examined literature:

1. Limited work combines multimodal visual synthesis with retrieval-grounded LLM generation into a single process.
2. The majority of current multimodal systems depend on cloud infrastructure, raising privacy issues.
3. Only a small number of studies assess the offline implementation of RAG-based multimodal frameworks for jobs involving the preparation of structured documents, such presentations.

III. PROPOSED SYSTEM ARCHITECTURE

This study presents a safe, completely offline multimodal AI system that combines diffusion-based picture synthesis with retrieval-augmented language modeling for automated PowerPoint production. The architecture's complete on-premise operation eliminates reliance on external cloud APIs while preserving multimodal content generating capabilities and contextual grounding.

There are five main parts to the system:

1. Knowledge indexing and document processing
2. Module for Semantic Retrieval
3. Language Generation with Retrieval Augmentation
4. Generation of Multimodal Images
5.PowerPoint Construction Engine Automation

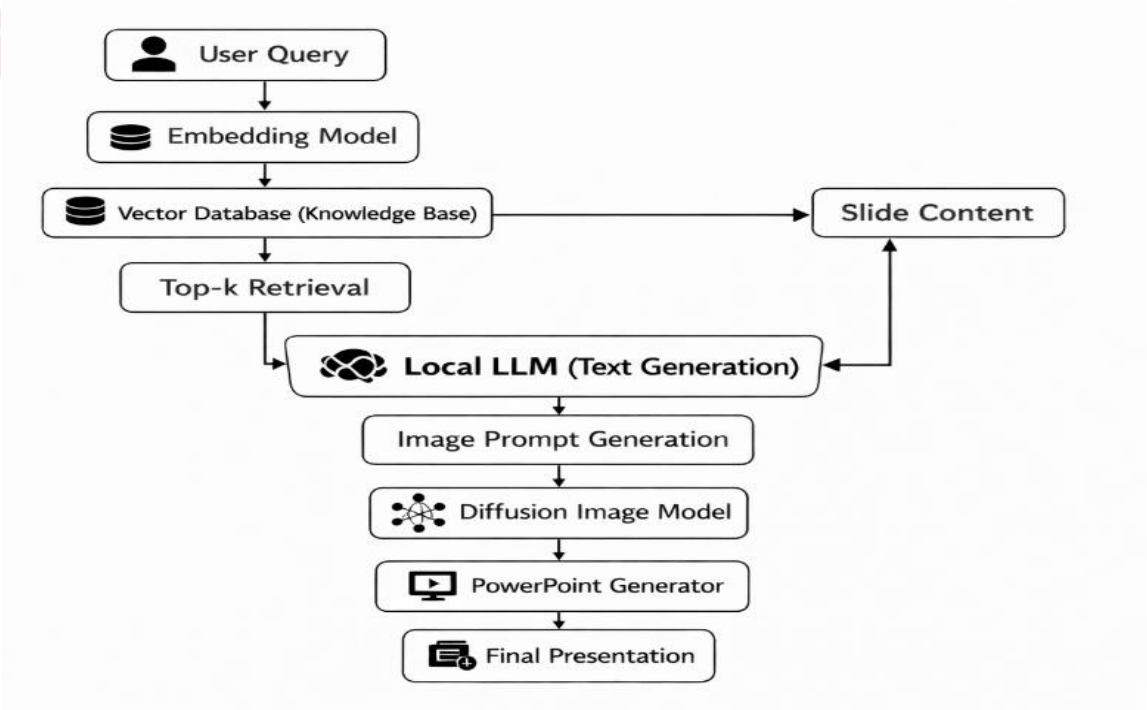

Fig. 1 illustrates the overall pipeline of the proposed system.

A. Document Processing and Knowledge Indexing

The first stage of the pipeline involves preparing domain knowledge for retrieval. Input documents such as PDFs, research articles, reports, or internal organizational documents are processed and segmented into smaller textual chunks.

Each segment of the documents is represented as a dense semantic representation by a text embedding model such as Sentence-BERT. This model maps input texts to high-dimensional vectors to represent their semantics. Let the document chunk be represented as di. The embedding function f(·) maps the document text to a vector representation:

$$e_i = f(d_i)$$

As $e_i \in \mathbb{R}^n$ states as an embedding vector in an n-dimensional semantic space.

All the embeddings are stored in a vector database to facilitate efficient similarity search through dense vector indexing techniques. This step results in an indexed knowledge base acting as the external memory for the RAG pipeline.

By running this component locally, sensitive documents are kept within the institutional infrastructure.

B. Semantic Retrieval Module

Where query q, gives a system first conversion of the query into an embedding vector by using the embedding function:

$$e_q = f(q)$$

As the system retrieves a top-k which is most relevant document segments by computing similarity by the query embedding and stored document embedding. In which Cosine similarity is used:

$$Sim(e_q, e_i) = \frac{e_q \cdot e_i}{\| e_q \| \ \| e_i \|}$$

The top-k documents have a highest similarity score which is been selected:

$$D_k = \{d_1, d_2, \ldots, d_k\}$$

These reserved passages provide contextual grounding for the LLM generation process. Incorporating external knowledge during inference, a system reduce hallucination which improves factual consistency.

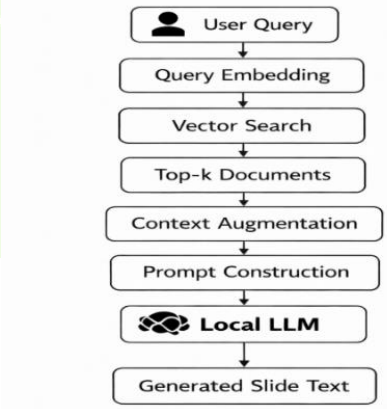

Fig. 2. Language model generation and document retrieval are integrated in the Retrieval-Augmented Generation pipeline.

C. Retrieval-Augmented Language Generation

A structured prompt for the language model is created by combining the retrieved documents with the user query.

The generation procedure adheres to the retrieval-augmented generation paradigm, which blends non-parametric document retrieval with parametric language modeling [3].

$$P(y \mid q) = \sum_{d \in D_k} P(y \mid q, d) P(d \mid q)$$

where:

- q represents the input query
- d represents retrieved contextual documents
- y denotes the generated output text

The augmented prompt used by LLM to generate structured slide content including:

- Slide titles
- Bullet points
- Key explanations
- Supporting textual summaries

The generated slide content has better factual reliability than standalone LLM outputs since the generation is based on retrieved knowledge.

Crucially, the language model ensures complete data confidentiality by operating locally within the system environment.

D. Multimodal Image Generation Module

To enhance the visual quality of generated slides, the framework incorporates a diffusion image generation model. The generated content from the LLM for the slide is converted into a descriptive prompt for image synthesis.

Diffusion image models generate images as demonising the image representation iteratively [8], [16]. The process starts with a random noise distribution xT and removes noise iteratively until a coherent image x0 is generated.

The image synthesis process can be represented by the following equation:

$$x_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_\theta(x_t, t)\right) + \sigma_t z$$

where:

x_t is the noisy image generated at a given step t

$\epsilon\theta$ is the learned image denoising function

z is the Gaussian noise

The generated image may contain diagrams, conceptual images, or context images relevant to the slide content.

As the image synthesis model is local, the system does not need to access external APIs and maintains the privacy of the data.

E. Automated PowerPoint Construction Engine

This is the final step in the pipeline for generating a PowerPoint presentation from the generated content.

The following actions are performed by the slide generating engine:

1. Choosing a slide layout
2. Insertion of the title
3. Formatting bullet points
4. Positioning and scaling of images
5. Template Styles and theme preservation

Libraries used to create a presentation, which enable automatic modification of a PowerPoint file, are used to generate slides automatically.

Each slide created contains:

• Title

• Important points

• Visual aids to support

A well-defined presenting style is ensured through such an automatic formatting.

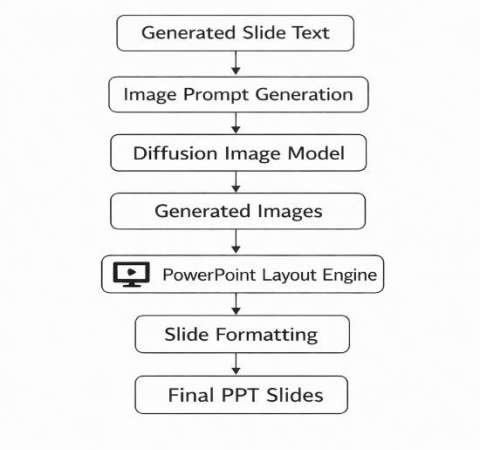

Fig. 3. Automated process for generating slides that includes text generation, image creation, and formatting in PowerPoint.

F. System Workflow

The entire workflow of the proposed system architecture can be briefly described as follows:

1. The user enters a question or a topic for the presentation.
2. The vector database is used to create a query embedding and retrieve relevant documents.
3. The language model prompt incorporates retrieved context.
4. Structured slide material is produced by the LLM.
5. For the diffusion model, text cues are transformed into visual prompts.
6. Textual material is merged with produced images.
7. A PowerPoint presentation is created automatically.

Automated presentation creation that is safe, scalable, and completely offline is made possible by this modular framework.

IV.EXPERIMENTAL SETUP AND EVALUATION

The experimental setup employed to measure the effectiveness of the proposed offline retrieval-augmented multimodal PowerPoint creation methodology is discussed in this part. There are three primary components of the effectiveness measure: the visual quality of the created presentations, their structural coherence, and factual grounding. To measure the effectiveness of retrieval augmentation and multimodal integration, a comparison is made between the proposed system and some baseline approaches.

A. Experimental Environment

In order to reproduce real-world corporate deployment scenarios with external API usage prohibited, all studies were performed in a completely offline computer environment.

Sentence BERT models similar to the ones utilized for the study, which offer an effective approach for the semantic representation of text data, were utilized for the creation of document embeddings [14].

A scalable vector search environment that can effectively handle high-dimensional embeddings was utilized for the creation of dense vector indexes and similarity queries [15]. This configuration allows for quick retrieval of relevant contextual documents during the RAG process.

In order to improve the level of factual consistency, the primary module for generating content includes a locally deployed Large Language Model with a retrieval-augmented generation approach that incorporates the results of a parametric model and external document retrieval [3].

The architecture includes a module for the creation of images with the help of a diffusion approach for image synthesis, which allows for the creation of contextual image features based on the input text [8], [16].

B. Building Datasets and Knowledge Bases

Domain based knowledge was built by utilizing a various variety of structured and unstructured data in order to assess the system, which including:

• Articles of research
• Technical records
• Educational resources
• Reports that are domain-specific

For improving the granularity of the information being retrieved, the documents were segmented into smaller pieces of text. The embedding scheme, as discussed in Section III, has been employed to map the documents to a high-dimensional semantic space. The knowledge base employed in the retrieval has been constructed by storing the vectors in a local vector database.

The capability of the retrieval module to provide relevant contextual information in the form of passages during the generation of content for the slides has been addressed in the dataset preparation process, which has been shown to have a significant impact in improving knowledge grounding in knowledge-intensive tasks [3], [5].

C. Baseline Methods

There are three experimental setup were examined in order to assess the efficacy of the suggested architecture:

1. LLM-Only Baseline

In this setup, the language model creates the slide content directly without the need for external knowledge retrieval. This method is known to experience hallucinations and factual errors in knowledge-intensive tasks, despite the fact that it benefits from the general information included in pretrained models [1], [2].

2. RAG, or retrieval-augmented generation

The second setup combines the language model with a retrieval mechanism. The generating prompt incorporates pertinent contextual materials that are obtained from the knowledge base. Previous studies show that as compared to solitary LLMs, RAG considerably increases factual accuracy and decreases hallucinations [3], [5].

3. Multimodal RAG Framework Proposal

By adding multimodal generating capabilities, the suggested system expands upon the RAG design. Diffusion-based generative models are used by the system to generate contextual pictures in addition to text. With this setup, graphically enhanced presentations may be created automatically.

D. Evaluation Metrics

To comprehensively evaluate the generated presentations, both automated metrics and human evaluation were employed.

1. Factual Consistency

The degree to which created slide material corresponds with recalled information is measured by factual consistency. Hallucination is a significant drawback of language models, especially in knowledge-intensive applications, according to earlier research [1], [6]. In order to find factual errors, outputs were carefully examined.

The rate of hallucinations was computed as follows:

$$Hallucination\ Rate = \frac{Number\ of\ Incorrect\ Statements}{Total\ Generated\ Statements}$$

Lower values indicate better factual grounding.

2. Retrieval Relevance

By determining whether the recovered passages were pertinent to the input query, retrieval quality was assessed. Because improper retrieval might spread mistakes into the generating step, retrieval relevance is crucial to RAG systems' success [11].

The following measure was used to determine the top-k retrieval precision:

$$Precision@k = \frac{Relevant\ Retrieved\ Documents}{k}$$

Higher precision indicates better retrieval alignment.

3. Structural Coherence

On the other hand, structural coherence refers to the extent to which the generated slides reflect a logical structure of presentation. In this case, human evaluators used a five-point Likert scale to measure the coherence of the slides.

- logical flow
- clarity of bullet points
- relevance of slide titles

4. Visual Quality

- relevance to slide topic
- visual clarity
- usefulness in explaining concepts

Diffusion models are been shown for generating a high-quality image which significantly enhance the visual communications [8] [12].

E. Human Evaluation Protocol

For evaluating the quality of the presentation as a whole, a human evaluation study has been conducted with various participants having a technical background. The participants were given a set of slides created using each experimental configuration and asked to rate them based on the following factors:

1. Factual correctness
2. Content clarity
3. Visual usefulness
4. Overall presentation quality

The ratings are done on a five-point Likert scale.

Recent studies stress the need for human evaluation in assessing the efficacy of RAG-based generation systems, as automated evaluation may fail to account for various aspects of factual correctness and coherence [7], [11].

F. Evaluation Procedure

The system was able to come up with a comprehensive presentation, which included several slides for all test inquiries. In order to compare all the configurations, the same set of queries was used for all of them.

The steps involved in the testing were as follows:

1. Use the basic model of LLM to generate slide content.
2. Use the model enhanced by RAG to generate presentations.
3. Use the multimodal RAG framework suggested to generate presentations.
4. Calculate metrics associated with retrieval and hallucinations.
5. Use humans to assess coherence and visual quality.

The three methods can be compared with one another owing to the experimental design. This design also evaluates the efficacy of retrieval augmentation.

V. RESULT

The experimental results for evaluating the proposed offline retrieval augmentation-based multimodal PowerPoint generating system are presented in this section. The evaluation process compares three system configurations: the retrieval augmentation-based generation (RAG), the LLM-only generation, and the proposed multimodal RAG system. The improvements in factual consistency, retrieval relevance, structural coherence, and visual presentation quality are to be evaluated.

A. Factual Consistency and Hallucination Reduction

One of the major motivations for using a retrieval augmentation-based system is the reduction of hallucinations that are often associated with big language models. Past studies have demonstrated that when big language models are used in isolation without any external grounding, they may provide factually incorrect responses that are still plausible [1] [2].

Table I: The proposed system's hallucination rates for the evaluated system configurations.

System Configuration	Hallucination Rate
LLM Only	18.6%
RAG	7.9%
Proposed Multimodal RAG	7.4%

The results indicate that factual grounding is significantly improved when there is retrieval. The success of retrieval augmentation in reducing factual errors is also evident from the fact that the hallucination rate is decreased from 18.6% in the baseline system, which only employed the LLM, to 7.9% in the system employing the RAG approach. This is in line with other studies indicating that the reliability of language model outputs is improved when external knowledge sources are incorporated [3] [5].
When creating both textual and visual pieces simultaneously, contextual awareness is increased, as reflected in the small improvement in the multimodal RAG setup.The significance of retrieval relevance and grounding integrity in lowering inaccurate outputs is further shown by recent research on hallucination identification and mitigation in RAG systems [6] [7].

B. Retrieval Performance

The quality of retrieved contextual documents plays a critical role in the effectiveness of the RAG pipeline. To evaluate retrieval performance, Precision was computed for the top retrieved passages.

Table II Retrieval Performance

Metric	Score
Precision@3	0.81
Precision@5	0.76
Precision@10	0.69

The findings show that the retrieval module can accurately identify pertinent contextual material. Accurate similarity search is made possible by the efficient mapping of queries and documents into a common vector space through the application of semantic embedding models [14].
When working with huge document collections, effective vector indexing algorithms significantly enhance computational efficiency and retrieval scalability [15].
According to recent RAG assessment surveys, retrieval quality has a significant impact on downstream generation performance because poor retrieval might add false context into the prompt [11].

C. Structural Coherence of Generated Slides

The structural quality of generated presentations was evaluated using human ratings based on slide organization, clarity of bullet points, and logical flow.

Table III Structural Coherence Scores

System	Average Score (1–5)
LLM Only	3.2
RAG	4.1
Proposed Multimodal RAG	4.3

When compared to the standalone LLM baseline, the RAG-based systems show notable gains in structural coherence. The approach creates more useful and logically structured slide material by adding retrieved contextual information.
This finding is in line with other studies showing that retrieval-augmented models enhance content organization and contextual reasoning in knowledge-intensive tasks [3].

D. Visual Quality Evaluation

To assess the contribution of multimodal generation, evaluators rated the relevance and usefulness of generated images.

Table IV Visual Quality Evaluation

Criterion	Average Score (1–5)
Image Relevance	4.2
Visual Clarity	4.4
Concept Explanation	4.1

The visual richness of the produced presentations is greatly enhanced by the use of diffusion-based picture production.
The diffusion models perform better the producing high-quality pictures conditioned in textual content [8] [12] [16].
Such visual components boost audience engagement and idea elucidation in the context of automated presentation production.

E. System Performance and Deployment Feasibility

One of the key goals of the suggested solution's design is the ability to enable the deployment of the system completely offline without the need for external APIs. The tests prove that the suggested architecture is able to perform well in an on-premise environment with respectable computer performance.
This is particularly important for institutional and corporate environments that face key security and privacy challenges with regard to data security and compliance [9] [13].
The modularity allows for the flexible inclusion of optimal models according to hardware capabilities, even with the additional computational limitations of local deployment.

F. Discussion

There are a few important observations based on the experimental outcomes.
One of the important observations is that, in contrast to LLM generation, retrieval augmentation achieves a substantial increase in factual dependability and a reduction in hallucination rates. This observation is consistent with other research that demonstrates the benefits of using parametric language models with non-parametric knowledge retrieval [3] [5].

Another important observation is that contextual grounding is achieved by using a combination of semantic retrieval and vector databases, which is crucial for generating accurate and cohesive slide content [14] [15].

The third important observation is that, by using diffusion-based multimodal generation, the visual quality of presentations is enhanced, leading to interesting and educational slides.

Finally, one of the important limitations of existing cloud-based generative models is addressed by providing a totally offline solution, ensuring privacy-preserving AI deployment in an enterprise setting [9].
In summary, this paper demonstrates that a combination of retrieval augmentation, multimodal generation, and on-premise deployment is a powerful framework for generating high-quality presentations in an automated manner.

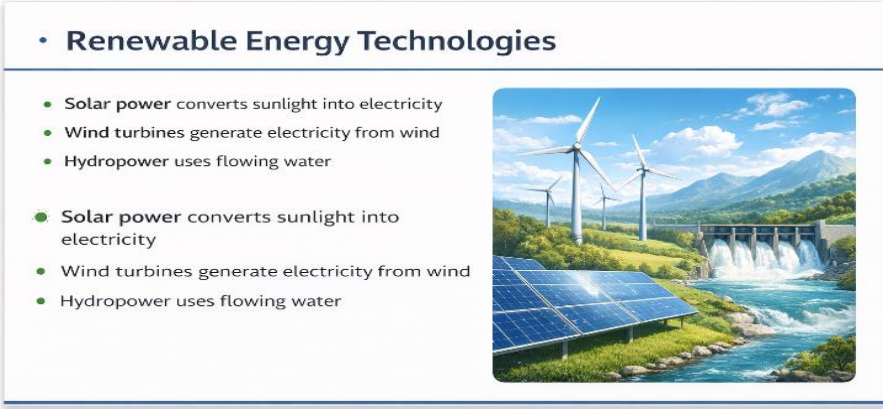

Fig. 4. An example slide automatically generated by the proposed multimodal RAG system.

VI. CONCLUSION

In the above article, a safe and completely offline multimodal AI system for the production of PowerPoint presentations through the synthesis of pictures and retrieval-augmented language modeling was proposed. In the case of on-premise deployment, the proposed system architecture includes the automated production of slides, grounded language generation, multimodal visual synthesis, and semantic document retrieval in one system.

When comparing the proposed system with an independent language model generation system, the experimental results show that there is a significant improvement in the consistency of facts. In addition, when the context materials were retrieved from the knowledge base and included in the prompt for generating the text, there was a significant reduction in hallucinations. This is in line with the results obtained in previous studies regarding the effectiveness of retrieval augmentation in tasks involving knowledge-intensive natural language synthesis.

Moreover, an efficient retrieval of relevant contextual information from large document collections becomes feasible through a combination of vector similarity search and semantic embedding models. In order to achieve improved retrieval accuracy and subsequent generation performance, dense retrieval algorithms have been found to be efficient in representing textual input data in a high-dimensional semantic space.

Through the creation of contextually relevant visual materials that are consistent with the generated text on the slide, picture creation via diffusion-based picture production improves the quality of the generated slides. Diffusion models are found to be appropriate for multimodal content generation pipelines since recent studies have demonstrated the ability of such models to generate high-quality pictures conditioned on textual cues.

The ability of the proposed system to be completely implemented offline is another important aspect of this study. The system provides data privacy, minimizes dependence on infrastructure, and enables implementation in environments with strict security or legal requirements by eliminating the need to rely on cloud-based APIs and services. This is in line with the growing interest in locally deployable language models for commercial applications and privacy-preserving AI systems.

Overall, the proposed approach demonstrates the potential for utilizing retrieval augmentation, multimodal generation, and local deployment architectures to support the development of outstanding automated presentations while maintaining data security and a solid foundation in fact.

Despite these promising results, there are many aspects that need to be researched in the future. These aspects are as follows:

1. Advanced Retrieval Optimization:

To reduce the likelihood of hallucinations even further, hybrid methods for combining dense and sparse search strategies are employed to improve the accuracy of the retrieval process.

2. Automated Hallucination Detection:

To make the results produced by the system more reliable and transparent, automated detection methods for hallucinations are incorporated in the production process.

3. Enhanced Multimodal Reasoning Ability:

The ability to generate charts, graphs, and visually explanatory content in addition to static images by expanding the system to incorporate enhanced multimodal reasoning abilities.

4. Interactive Presentation Generation:

Investigation into the development of interactive systems with dialogue systems that enable users to improve the content of slides, adjust the visual elements, and organize the presentation.

5. Performance Optimization for Edge Devices:

Investigation into the strategies for effective deployment on edge devices with resource constraints.

The scope for enhancing the scalability, reliability, and usability of automated presentation generation systems is immense with future research in the aforementioned areas, paving the way for the deployment of safe multimodal AI systems for knowledge-intensive organizational operations.

REFERENCES

[1] S. Ji, S. Pan, E. Cambria, P. Marttinen, and P. S. Yu, "Survey of Hallucination in Natural Language Generation," ACM Computing Surveys, vol. 55, no. 12, 2023.

[2] Y. Lin, M. Hilton, and O. Evans, "TruthfulQA: Measuring How Models Mimic Human Falsehoods," Proc. ACL, 2022.

[3] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks," Proc. NeurIPS, 2020.

[4] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang, "REALM: Retrieval-Augmented Language Model Pre-Training," Proc. ICML, 2020.

[5] A. Shuster, S. Poff, M. Chen, et al., "Retrieval Augmentation Reduces Hallucination in Dialogue Systems," Proc. EMNLP, 2021.

[6] Z. Ma, S. An, Z. Lin, Y. Zou, J.-G. Lou, and B. Xie, "Dehallucinating Parallel Context Extension for Retrieval-AugmentedGeneration," arXiv preprint arXiv:2412.14905, 2024.

[7] Z. Sun, X. Zang, K. Zheng, Y. Song, J. Xu, X. Zhang, and H. Li, "ReDeEP: Detecting Hallucination in Retrieval-Augmented Generation via Mechanistic Interpretability," arXiv preprint arXiv:2410.11414, 2024.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," Proc. CVPR, 2022.

[9] W. Ahmad, K. Chakraborty, and Y. Chang, "Privacy-Preserving Large Language Models: A Survey,"
IEEE Access, 2024.

[10] C. S. Mala, G. Gezici, and F. Giannotti, "Hybrid Retrieval for Hallucination Mitigation in Large Language Models," arXiv preprint arXiv:2504.05324, 2025.

[11] A. Gan, J. Wang, L. Chen, and Y. Liu, "Retrieval-Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey," arXiv preprint arXiv:2504.14891, 2025.

[12] P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," Advances in Neural Information Processing Systems (NeurIPS), 2021.

[13] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated Learning: Challenges, Methods, and Future Directions," IEEE Signal Processing Magazine, vol. 37, no. 3, pp. 50–60, 2020.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," Proc. 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.

[15] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2021.

[16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020.