ijamrsd.com
ISSN: 2583-6404
Journal of Advanced Multidisciplinary Research Studies and Development
Peer Reviewed Referred Journal | Bi-monthly | Open Access

DIMENSION AI - 3D: PROMPT-BASED MULTIMODAL AI FOR TEXT AND VISUAL CONTENT GENERATION

Durvesh Raneja[1], Gaurav Shinde[2], Abdullah Shaikh[3], Sujal Patil[4]
[1,2,3,4] *Computer Engineering, Trinity College of Engineering and Research, Pune, Maharashtra.*
Email: durveshraneja13@gmail.com[1], gauravgms8588@gmail.com[2], abdshk007@gmail.com[3], sujalpatil1313@gmail.com[4]

Abstract
The advancement of multimodal artificial intelligence has enabled systems capable of processing and generating diverse forms of digital content from unified inputs. This paper presents Dimension AI – 3D, a prompt-based multimodal framework designed to generate textual responses, 2D images, and structured 3D models from a single user query. The system integrates transformer-based Question Answer Generation (QAG) models for contextual reasoning, Conditional Generative Adversarial Networks (cGAN) and diffusion-based techniques for image synthesis, and U-Net++ combined with Open3D for 3D reconstruction. A modular client–server architecture ensures scalable deployment, efficient AI inference, and secure data handling. The proposed pipeline enables seamless transformation from natural language prompts to coherent multimodal outputs. Experimental validation demonstrates stable system performance under concurrent usage, acceptable inference latency, and structured geometric reconstruction quality. The system establishes a scalable foundation for next-generation prompt-driven multimodal AI platforms integrating textual, visual, and spatial intelligence.
Keywords: Dimension AI, QAG, cGAN, Multimodal AI, 2D Model, 3D Model.

► *Corresponding Author: Durvesh Raneja*

I. Introduction

Artificial intelligence technologies have rapidly evolved across domains such as natural language processing, computer vision, and 3D modelling. Transformer-based architectures have improved contextual reasoning in text generation tasks, while generative models such as GANs and diffusion networks have enhanced visual synthesis capabilities. Similarly, segmentation networks and geometric processing frameworks have enabled structured 3D reconstruction from 2D representations.

Despite these advancements, most AI systems operate within isolated modalities. Text-based models generate responses without visual context, image generators lack structured spatial reasoning, and 3D reconstruction systems function independently of semantic input. This separation restricts the development of unified intelligent systems capable of seamless multimodal transformation.

Dimension AI – 3D proposes an integrated prompt- driven architecture that generates three coordinated outputs from a single user input: a contextual textual explanation, a structured 2D image, and a reconstructed 3D model. The framework combines transformer-based reasoning, conditional image synthesis, and geometric reconstruction into a unified scalable system.

II. Literature Review

Transformer architectures introduced self-attention mechanisms that significantly improved contextual text understanding and generation. These models form the foundation of Question Answer Generation systems capable of semantic reasoning.

Generative Adversarial Networks, particularly Conditional GANs, allow controlled image synthesis by conditioning outputs on textual or structured prompts. Diffusion-based generative models further enhance image diversity and realism through iterative denoising processes.

For 3D reconstruction, encoder–decoder segmentation networks such as U-Net and U-Net++ improve boundary detection and feature extraction accuracy. Open3D provides efficient tools for point cloud processing, mesh reconstruction, and geometry refinement.

However, existing research primarily focuses on single- modality systems. Integrated pipelines combining transformer reasoning, generative image synthesis, and 3D reconstruction remain limited. Dimension AI addresses this gap by designing a unified multimodal architecture.

III. System Architecture

The architecture follows a layered client–server framework consisting of:

A. User Interaction Layer

Accepts user prompts and displays multimodal outputs including text, image, and 3D model.

B. Application Layer

The application layer processes user requests and ensures proper formatting and validation before forwarding them to the backend server. It manages secure communication between the frontend interface and AI services.

C. Server Layer

The server handles incoming requests and directs them to the appropriate AI components. It manages coordination between the QAG model, image generation module, 3D reconstruction module, and storage systems to maintain smooth data flow.

D. AI & Image Generation Module

This module performs transformer-based Question Answer Generation (QAG) for contextual reasoning. The generated response is then used as a conditioned prompt for image synthesis using Conditional GAN or diffusion-based models, producing structured 2D outputs.

E. 3D Model Generation Module

This component converts generated 2D representations into structured 3D models. U-Net++ is used for feature extraction and segmentation, while Open3D techniques enable point cloud processing and mesh reconstruction to produce accurate 3D outputs.

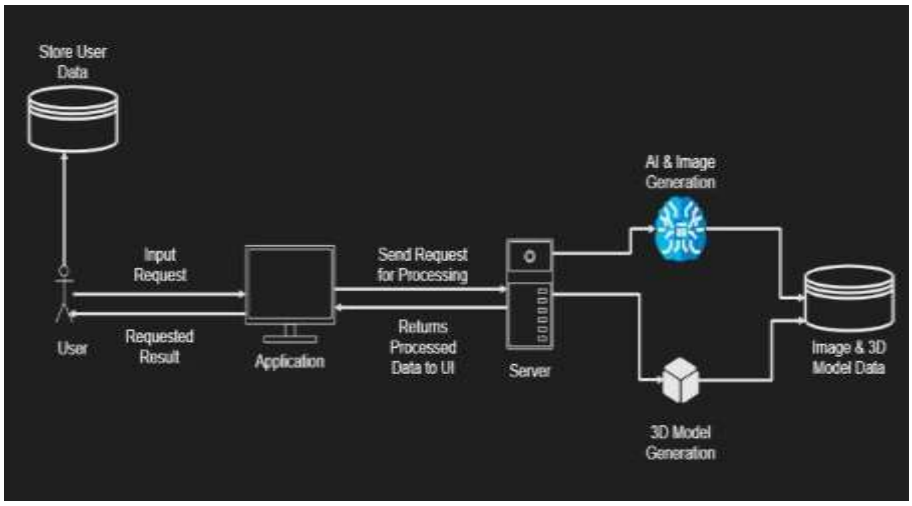

Figure 1: System Architecture

IV. Methodology

The system follows a sequential multimodal pipeline:

A. User submits a natural language prompt.

B. Transformer-based QAG generates contextual response.

C. Generated text serves as conditioned input for image synthesis.

D. Synthesized image undergoes segmentation using U- Net++.

E. Open3D performs point cloud processing and mesh reconstruction.

F. Final outputs are stored and displayed.

G. 2D images, and interactive 3D models. Basic controls allow users to view, rotate, and interact with generated content.

V. Mathematical Model

To Evaluate Multimodal Output Quality, A Composite Scoring Function is Introduced:

$$Q = wTT + wII + wFF + wAA$$

Where:

T = Text Generation Accuracy

I = Image Coherence Score

F = Feature Extraction Reliability

A = 3d Reconstruction Accuracy

Weights Satisfy:

$wT + wI + wF + wA = 1$

This Weighted Evaluation Ensures Balanced Assessment Across Modalities.

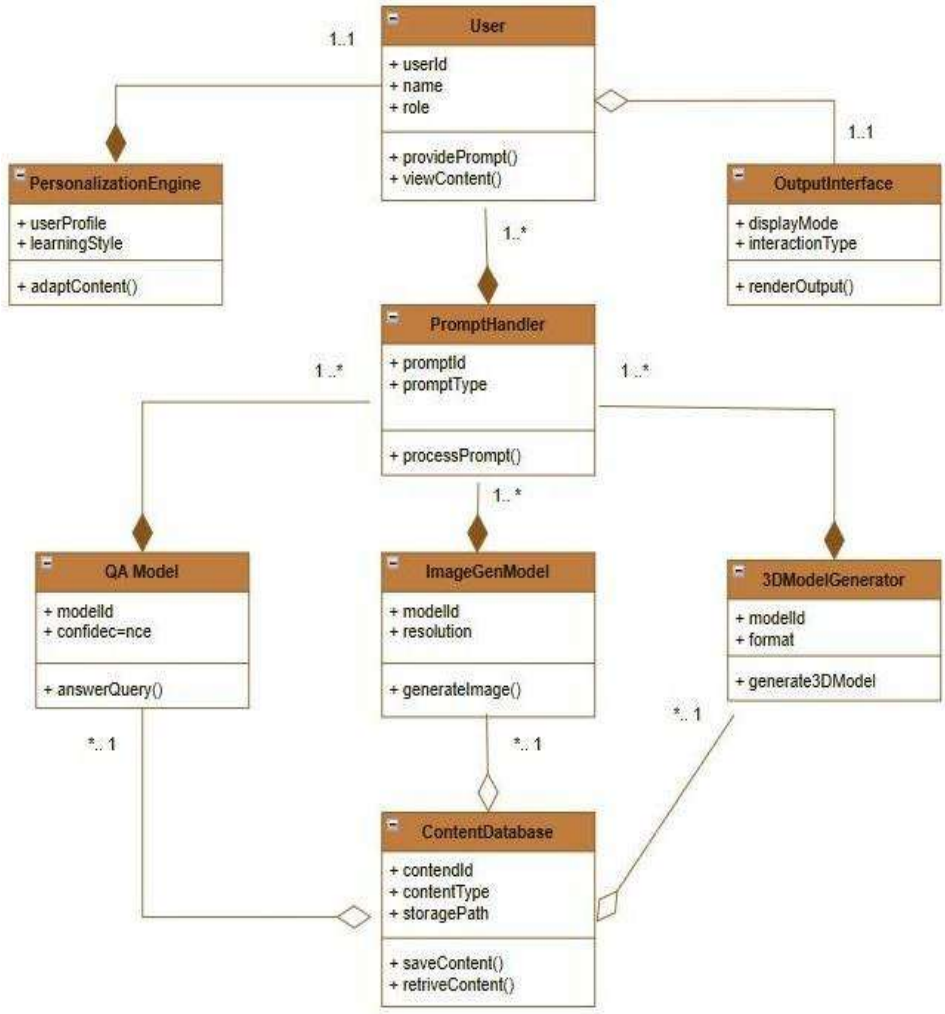

Figure 2: Class Diagram

VI. Experimental Results and Discussion

Prototype testing demonstrated successful generation of contextual textual responses, semantically aligned 2D images, and structured 3D models. The system maintained stable API communication and coherent cross-modal outputs across multiple prompts.

Performance evaluation indicated acceptable inference latency for text reasoning and image synthesis, while 3D reconstruction required moderate computational resources. Load testing confirmed stable system behaviour under concurrent requests.

Challenges included optimizing inference speed and improving mesh precision. Continuous model tuning enhanced structural consistency and output reliability.

Figure 3: Home Page

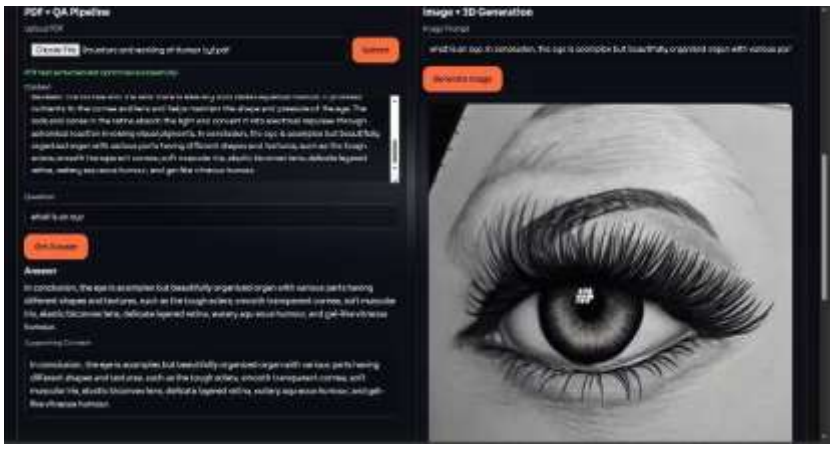
Figure 4: QA and 2D Image Generation

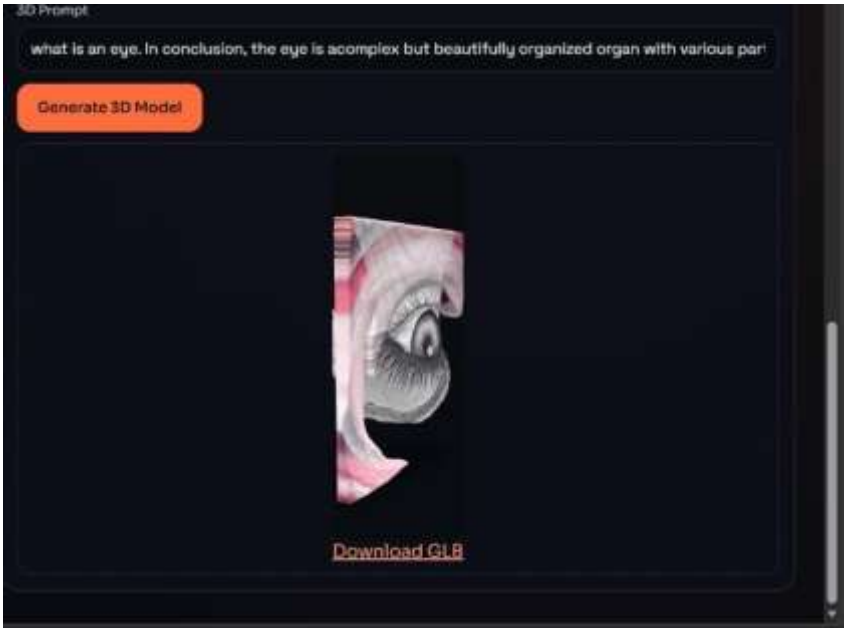

Figure 5: 3D Model Generation

VII. Comparative Analysis

The Comparison of Multimodal Systems

Table 1: Comparison Table

Feature	Text Systems	Image GAN	3D Systems	Dimension AI
Text Generation	Yes	No	No	Yes
Image Generation	No	Yes	No	Yes
3D Reconstruction	No	No	Yes	Yes
Unified Pipeline	No	No	No	Yes
Prompt-Based	Partial	Partial	Limited	Full

The comparison highlights the integrated nature of Dimension AI.

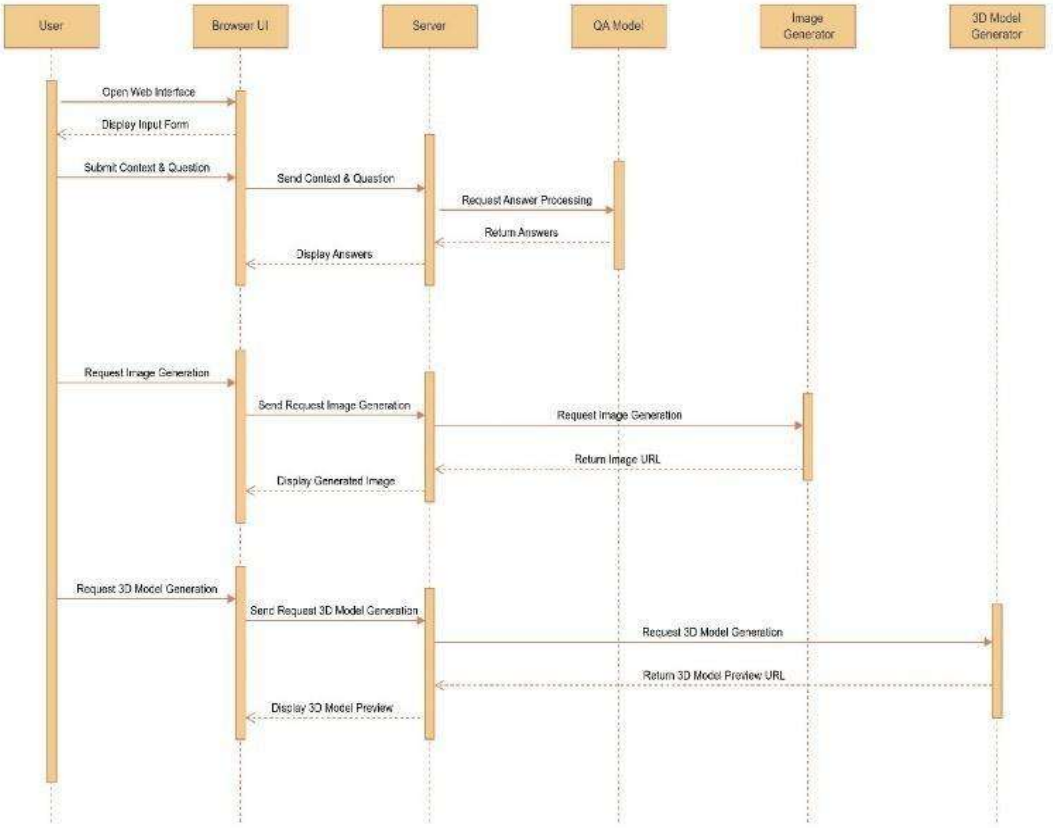

Figure 6: Sequence Diagram

VIII. Conclusion

Dimension AI – 3D demonstrates the feasibility of integrating transformer-based reasoning, conditional image synthesis, and structured 3D reconstruction within a unified prompt-driven architecture. The modular design ensures scalability, stable performance, and secure data handling. The system bridges isolated AI modalities and establishes a foundation for future multimodal intelligent systems.

Acknowledgment

The authors sincerely thank the project supervisor and the Department of Computer Engineering for their guidance and support in completing this research work. The institutional infrastructure and technical resources helped us to implement the proposed system.

References

1. Ushio, F. Alva-Manchego, And J. Camacho- Collados, “A Practical Toolkit For Multilingual Question And Answer Generation,” In Proc. 61st Annual Meeting Of The Association For Computational Linguistics (Volume 3: System Demonstrations), Toronto, Canada, Jul. 2023, Pp. 86-94, Doi: 10.18653/V1/2023.Acl-Demo.8.
2. M. Momen-Tayefeh, “Text-To-Image With Generative Adversarial Networks,” Arxiv Preprint Arxiv:2410.08608, Oct. 2024.
3. H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, And D. Metaxas, “Stackgan: Text To Photo-Realistic Image Synthesis With Stacked Generative Adversarial Networks,” Arxiv Preprint Arxiv:1612.03242, Dec. 2016.
4. Q. Xu, W. Wang, D. Ceylan, R. Mech, And U. Neumann, “Disn: Deep Implicit Surface Network For High-Quality Single-View 3d Reconstruction,” Arxiv Preprint Arxiv:1905.10711, May 2019.
5. A. Singh Et Al., “Pixel2point: 3d Object Reconstruction From A Single Image Using Cnn And Initial Sphere,” Ieee Access, Vol. 9, Pp. 110- 121, 2020.
6. S. Ramzan Et Al., “Text-To-Image Generation With Enhanced Gans: Bridging Semantic Gaps Using Rnn And Cnn,” Pmc, 2026.
7. M. Fathallah, S. Eletriby, M. Alsabaan, M. I. Ibrahem, And G. Farok, “Advanced 3d Face Reconstruction From Single 2d Images Using Enhanced Adversarial Neural Networks And Graph Neural Networks,” Sensors, Vol. 24, No. 19, Art. No. 6280, 2024.