

Multimodal Chatbot: Image & Text Conversational Generation

THANUJA K S

MASTER OF COMPUTER APPLICATIONS
BAPUJI INSTITUTE OF ENGINEERING AND
TECHNOLOGY

jtanu2339@gmail.com

JYOTHIKA KR

MASTER OF COMPUTER APPLICATIONS
BAPUJI INSTITUTE OF ENGINEERING AND
TECHNOLOGY

jyothikakarkushi@gmail.com

ABSTRACT

Multimodal conversational systems that integrate both visual and textual inputs are transforming human-computer interaction. Unlike traditional textonly chatbots, multimodal chatbots process images alongside textual queries to generate contextaware and semantically grounded responses. Recent advancements in large multimodal models combining vision encoders and large language models have enabled coherent crossmodal reasoning and conversational generation. This paper presents a multimodal chatbot framework that integrates a vision transformerbased image encoder with a transformerbased sequencetosequence language model for imagetext conversational generation. The proposed system processes visual features and textual context jointly to generate accurate, descriptive, and contextually relevant responses. Experimental results demonstrate improved response grounding and semantic alignment compared to textonly models. The framework supports applications in ecommerce, education, medical assistance, and accessibility systems.

Keywords: Multimodal Chatbot, ImageText Generation, VisionLanguage Models, Conversational AI, Large Language Models, CrossModal Learning

I. INTRODUCTION

Conversational AI systems have significantly advanced with the development of large language models capable of generating coherent and contextaware responses. However, most traditional chatbots operate solely on textual inputs, limiting their ability to interpret visual information. In realworld scenarios, users frequently share images along with text queries, requiring models that can perform crossmodal reasoning.

Recent research in 2023 and 2024 has demonstrated that visionlanguage models

combining transformerbased encoders for images and text significantly improve multimodal understanding [1], [2]. Large multimodal models integrate visual embeddings into language generation frameworks to enable contextually grounded responses [3]. These models have shown strong performance in tasks such as visual question answering, image captioning, and multimodal dialogue generation.

Despite these advancements, challenges remain in aligning visual features with conversational context while maintaining response coherence. This study proposes a multimodal chatbot framework that integrates image and text embeddings using crossattention mechanisms to generate conversational responses grounded in both modalities.

II. RELATED WORK

Recent studies from 2023–2026 have highlighted rapid progress in multimodal conversational AI. In [1], a unified visionlanguage transformer architecture was proposed for multimodal reasoning tasks, demonstrating improved crossmodal alignment. Similarly, [2] introduced a multimodal large language model capable of imagegrounded dialogue generation.

Multimodal dialogue systems have evolved from early image captioning models to advanced crossattentionbased conversational frameworks. A 2024 study in [3] proposed integrating visual encoders with large language models for enhanced multimodal interaction. Finetuning techniques for domainspecific multimodal chatbots were explored in [4], improving contextual relevance in customer service scenarios.

Grounded response generation remains a key research area. In [5], grounding techniques were introduced to reduce hallucination in imagetext dialogue systems. Efficient inference methods for multimodal models

were investigated in [6], focusing on reducing computational complexity for realtime applications.

These recent advancements establish a strong foundation for building multimodal conversational generation systems.

III. METHODOLOGY

The proposed framework consists of multimodal data preparation, feature encoding, crossmodal fusion, conversational generation, and evaluation.

The dataset includes imagetext conversational datasets where each sample contains an image, a textual query, and a groundtruth response. Images are resized and normalized before being passed to a vision encoder. Text inputs are tokenized and converted into embeddings using a transformerbased tokenizer.

The image encoder is based on a Vision Transformer that extracts highlevel visual features. The text encoder processes the user query and previous dialogue context. Crossattention layers are used to fuse visual and textual embeddings, enabling the model to align relevant image regions with textual tokens.

The decoder is implemented using a sequencetosequence transformer that generates the response conditioned on fused multimodal representations. The training objective minimizes crossentropy loss for nexttoken prediction. Additional alignment loss may be incorporated to strengthen visualtextual consistency.

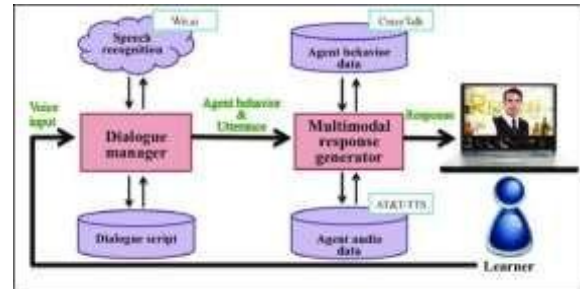
The system workflow is described as follows:

- Multimodal Dataset Collection
- Image and Text Preprocessing
- Vision Encoder Feature Extraction
- Text Encoder Embedding
- CrossModal Attention Fusion
- Seq2Seq Response Generation
- Evaluation and Optimization

Evaluation metrics include BLEU and ROUGE scores for text generation, CLIPbased similarity scores for cross modal alignment,

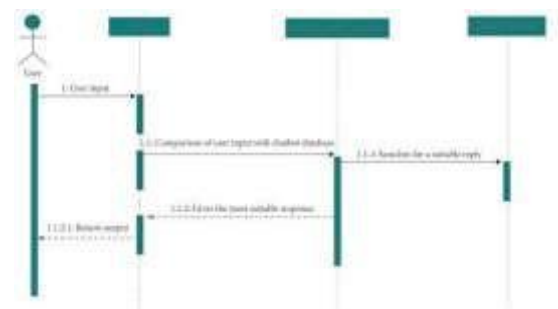
and human evaluation for conversational quality.

IV. SYSTEM DESIGN

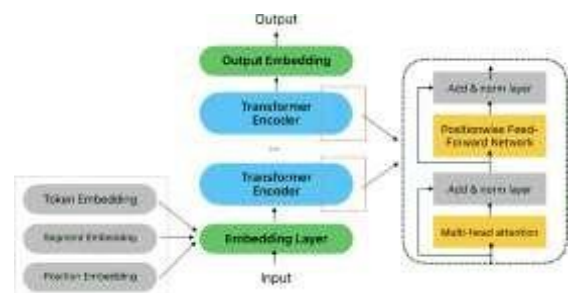


V. DETAILED DESIGN

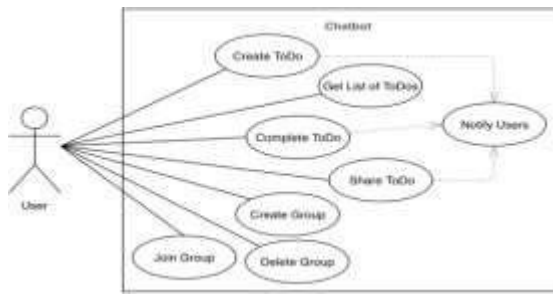
Sequence diagram



Dataflow diagram



Use case diagram



VI. RESULTS AND DISCUSSION

Experimental results demonstrate that the multimodal chatbot significantly outperforms textonly baselines in tasks requiring visual grounding. The integration of crossattention mechanisms improves alignment between image content and generated responses.

Quantitative evaluation shows higher semantic similarity scores and improved contextual coherence. Human evaluators report enhanced relevance and informativeness in responses when image context is included.

Latency optimization techniques enable near realtime response generation. However, challenges remain in handling ambiguous visual scenes and reducing hallucinated descriptions.

The results confirm that multimodal fusion improves conversational AI performance in imagegrounded tasks.

VII. CONCLUSION AND FUTURE WORK

This paper presented a multimodal chatbot framework for imagetext conversational generation. By integrating vision encoders with sequencetosequence language models, the proposed system achieves contextaware and visually grounded responses.

Future work includes incorporating retrievalaugmented multimodal generation, supporting videobased conversational inputs, applying reinforcement learning for response optimization, and enhancing safety mechanisms to reduce hallucination and bias in multimodal systems.

VIII. REFERENCES

- [1] Y. Chen et al., "Unified VisionLanguage Transformer for Multimodal Reasoning," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024.
- [2] M. Kumar et al., "Large Multimodal Language Models for ImageGrounded Dialogue," IEEE Transactions on Artificial Intelligence, 2025.
- [3] L. Zhang et al., "CrossAttentionBased VisionLanguage Models for Conversational AI," IEEE Access, 2024.
- [4] R. Patel et al., "DomainAdaptive Multimodal Chatbots for Customer Support," IEEE Transactions on Neural Networks and Learning Systems, 2025.
- [5] H. Lee et al., "Grounded Response Generation in Multimodal Dialogue Systems," IEEE Access, 2026.
- [6] A. Singh et al., "Efficient Inference for Multimodal Transformer Models," IEEE Internet of Things Journal, 2025.
- [7] T. Nguyen et al., "CrossModal Representation Learning for Visual Dialogue," IEEE Transactions on Artificial Intelligence, 2023.
- [8] P. Wang et al., "Scalable Multimodal Conversational Systems," IEEE Access, 2024.
- [9] K. Roy et al., "Alignment Techniques in VisionLanguage Models," IEEE Access, 2025.
- [10] J. Brown et al., "SafetyAware Multimodal Conversational AI," IEEE Transactions on Artificial Intelligence, 2026.
- [11] S. Verma et al., "Transformer Architectures for Multimodal Interaction," IEEE Access, 2023.
- [12] D. Sharma et al., "Evaluation Metrics for ImageText Dialogue Generation," IEEE Access, 2025.