Multimodal Generative AI: The Future of Human-AI Creativity

Name: ABHINAND.S

Roll No: 02

Reg No: MES22AD002

Guide: MR. SHERIKH KK

Department: ADS, MESCE

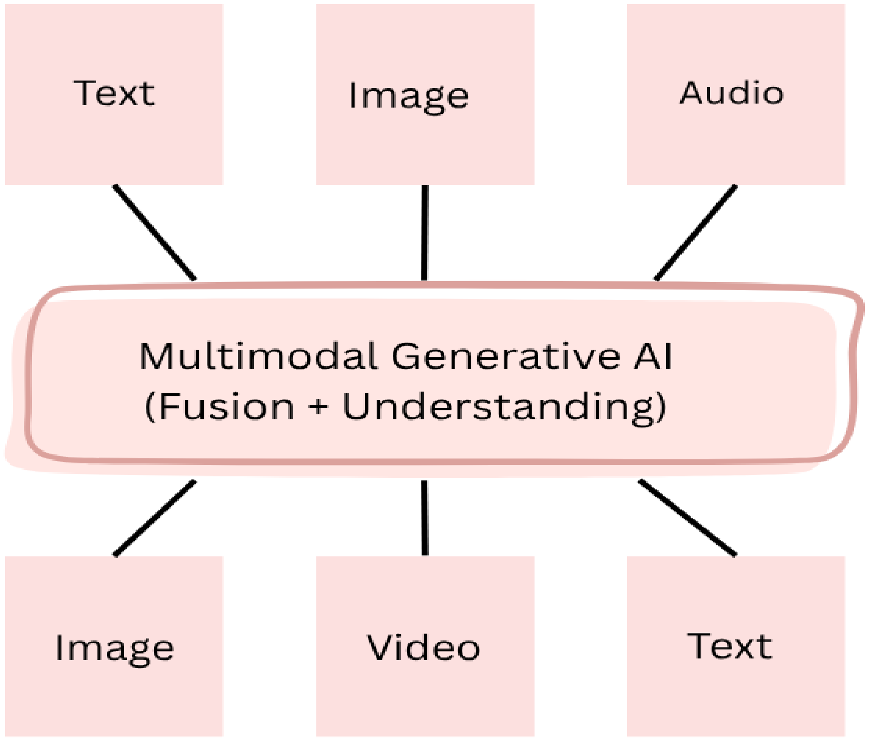
Date:07-08-2025

Agenda

1. Introduction
2. Problem Statement
3. Objectives
4. Literature Review
5. Methodology
6. Tools & Technologies
7. Applications
8. Pros & Cons
9. Future Scope
10. Conclusion
11. References

Introduction

- Combines data from text, image, audio, and video
- Enables cross-modal understanding
- Use cases: storytelling, sketch-to-image
- Facilitates creative human-AI collaboration
- Enhances multimodal content creation
- Examples: GPT-4o, Sora
- Industry-driven adoption



Img 1 – Multimodal AI: Combines text, image, and audio to generate diverse outputs.

Problem Statement

- Existing AI systems are unimodal
- Lack of unified multimodal architecture
- Limited creative interaction
- High resource demand
- Issues with real-time processing
- Poor interpretability in generation
- Constrains innovation

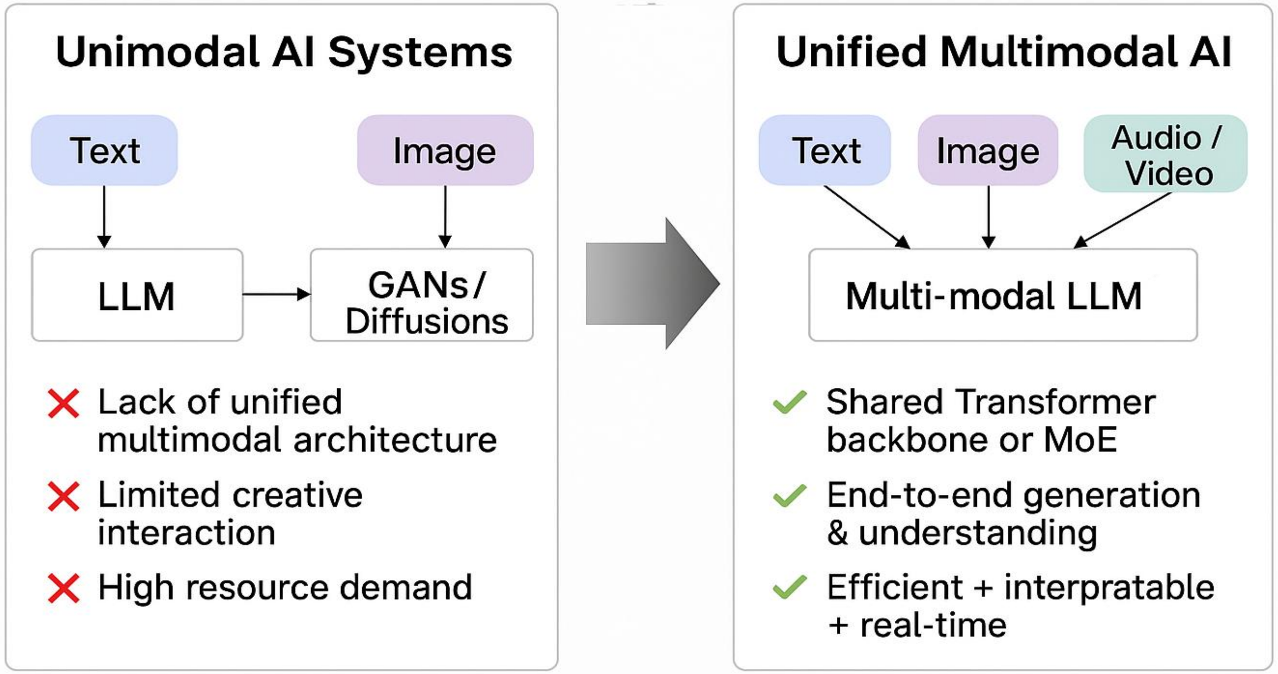


Fig 2 – Unified Models: Transitions from separate to unified multimodal AI systems.

Objectives

- Explore working of multimodal AI models
- Analyze creative and design applications
- Identify gaps in current approaches
- Evaluate methods like diffusion and transformers
- Study tools enabling multimodal generation

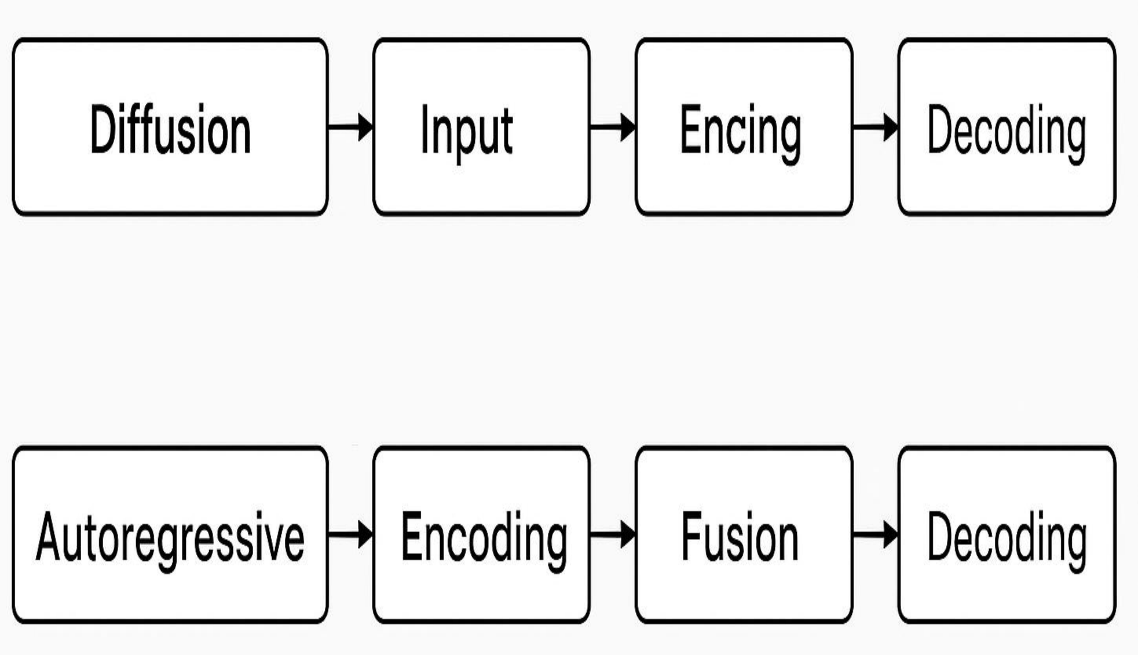


Fig 3 – Pipelines: Compares diffusion and autoregressive generation workflows.

Literature Review

Title	Authors, Year	Method	Result
Multi-modal Generative AI: Multi-modal LLMs, Diffusions and the Unification[1]	Xin Wang et al., 2025 (IEEE)	Unified transformer + diffusion	Framework for multimodal prompt-based generation
Multimodal Image Synthesis and Editing: The Generative AI Era[2]	Zhan et al., 2024 (IEEE TPAMI)	Diffusion & GANs	High-quality image synthesis; fast GAN inference
Sketch-to-Image via Diffusion Model for Superior Visual Synthesis[3]	Roy et al., 2023 (UTS, IIT KGP)	Sketch-guided diffusion	Photorealistic images from sketches
DiffSketching: Sketch-Controlled Diffusion Models[4]	Wang et al., 2023	Sketch-controlled diffusion	Precise sketch-to-image synthesis
Multimodal Explainable Artificial Intelligence: A Comprehensive Review[5]	Nikolaos Rodis et al., 2024 (IEEE Access)	Multimodal XAI (Grad-CAM, SHAP, DME)	Interpretable explanations for VQA, captioning

Summary of recent research on multimodal generative AI methods and results

Working Principle

- Based on Xin Wang et al. (2025) – unified multimodal LLM + diffusion framework
- Handles diverse inputs: text, sketch, and audio
- Uses modality-specific encoders for flexible input processing
- Aligns modalities using attention-based fusion
- Supports high-quality, prompt-based generation
- Ideal for creative and stylized multimodal outputs

Working Principle (Input & Encoding)

- Input: Text, sketch, or audio prompts
- Modality-specific encoders: transformer, CNN, spectrogram
- Text: Tokenized and encoded by LLM
- Image: Processed via convolutional encoder
- Audio: Converted to spectrograms
- Ensures semantic alignment
- Prepares for fusion

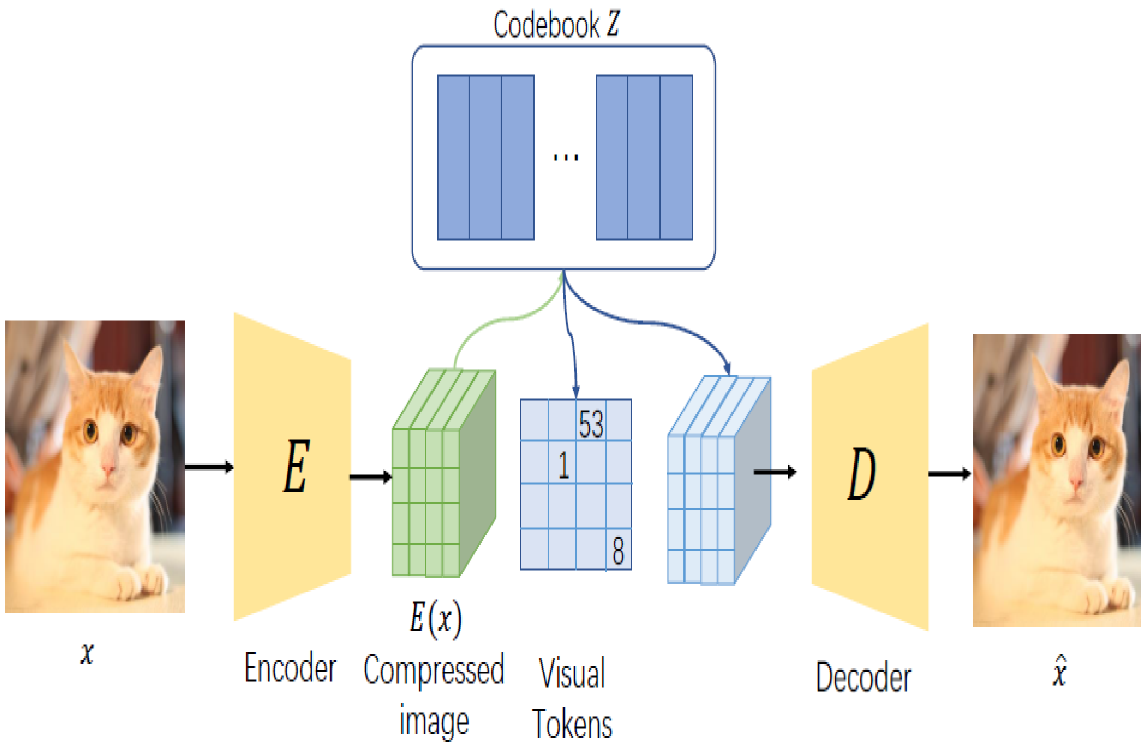


Fig 4 – Visual Tokenization: Encodes and decodes images through discrete visual tokens.

Working Principle (Fusion)

- Fusion via cross-attention/token fusion
- Aligns latent representations
- Combines embeddings across modalities
- Transformer layers capture dependencies
- Enhances contextual understanding
- Flexible fusion techniques
- Core step in multimodal generation

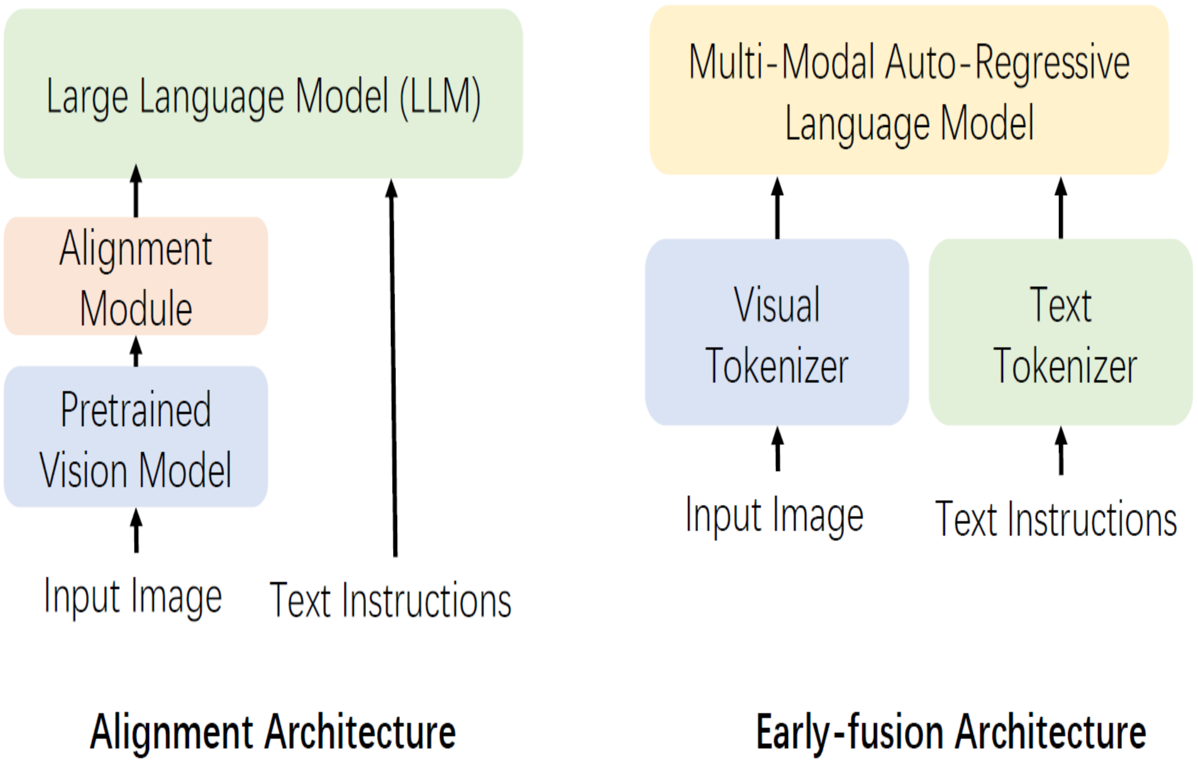


Fig 5 – Model Architectures: Contrasts alignment vs. early-fusion multimodal models.

Working Principle (Decoding)

- Decoding via diffusion or transformer decoder
- LoRA, ControlNet for fine-tuning
- Generates image/text/video output
- Supports real-time generation
- Ensures coherence across modalities
- Domain adaptation for specific tasks
- Achieves high fidelity outputs

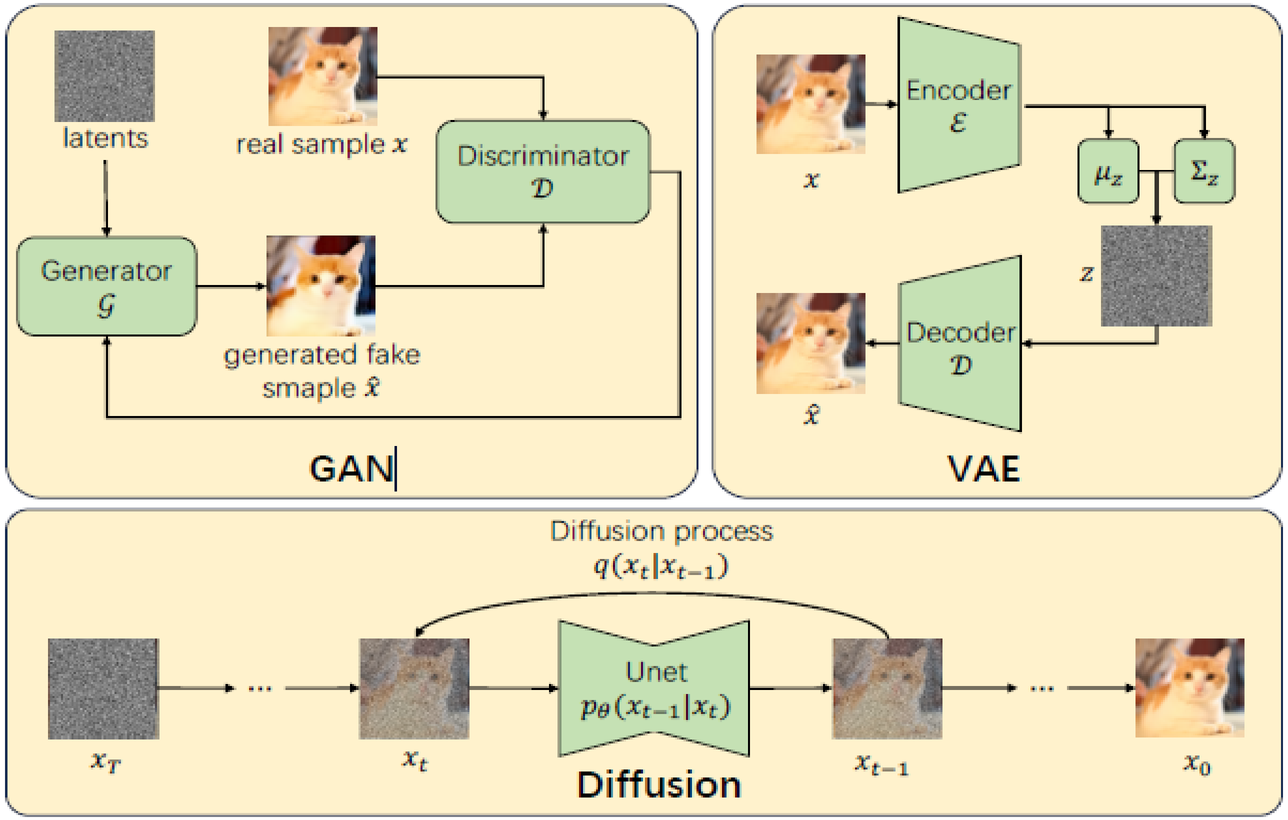


Fig 6 – Generative Models: Shows how GANs, VAEs, and diffusion models work.

Tools & Technologies

- **GPT-4o**: Unified vision + audio + text LLM
- **Gemini**: Google's multimodal model
- **CLIP**: Vision-language alignment
- **LLaVA**: Language-Vision Assistant
- Combines perception and reasoning
- Transformer-based alignment
- Real-world deployment ready

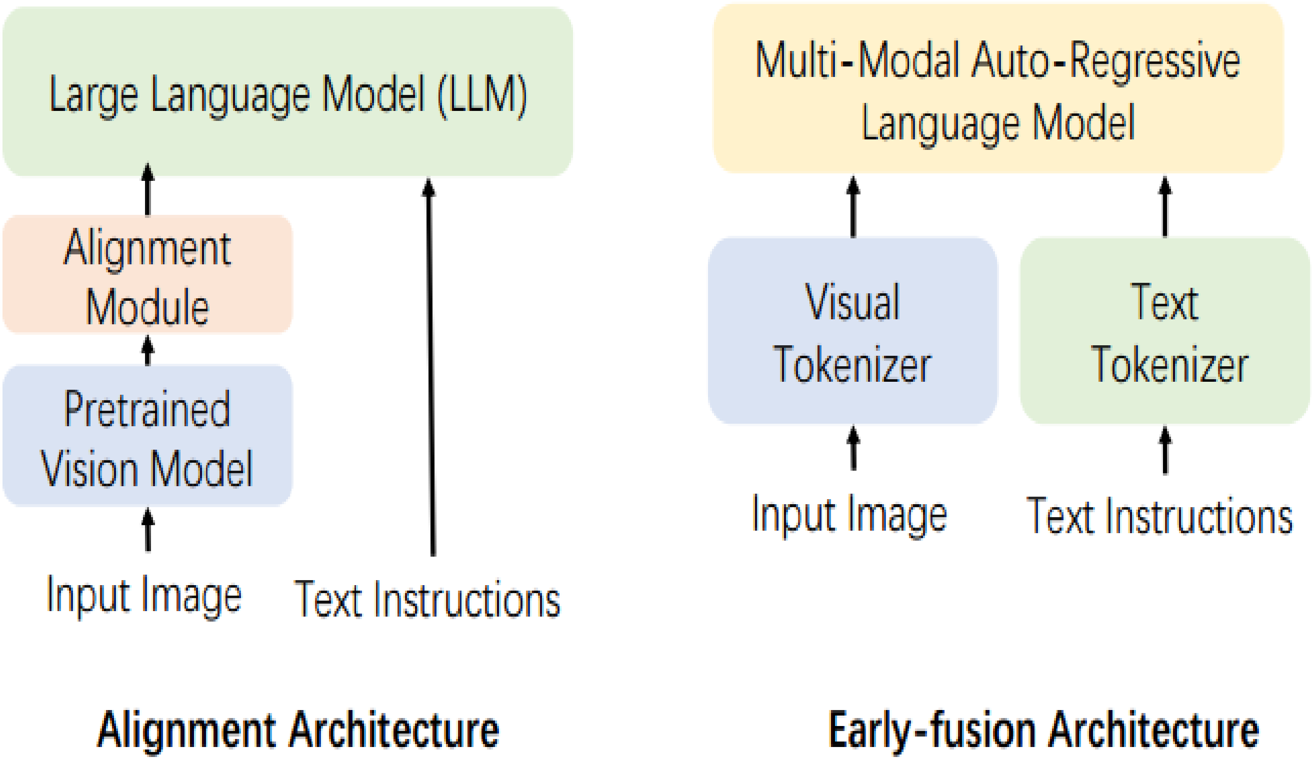


Fig 7 – Comparison of Alignment and Early-fusion multimodal AI architectures for processing images and text.

Tools & Technologies

- **Stable Diffusion**: Text-to-image generation
- **ControlNet**: Conditioning on inputs
- **Sora**: Text-to-video by OpenAI
- **LoRA**: Lightweight fine-tuning
- **LCMs**: Low-latency diffusion models
- Enables fast generation
- Supports real-time applications

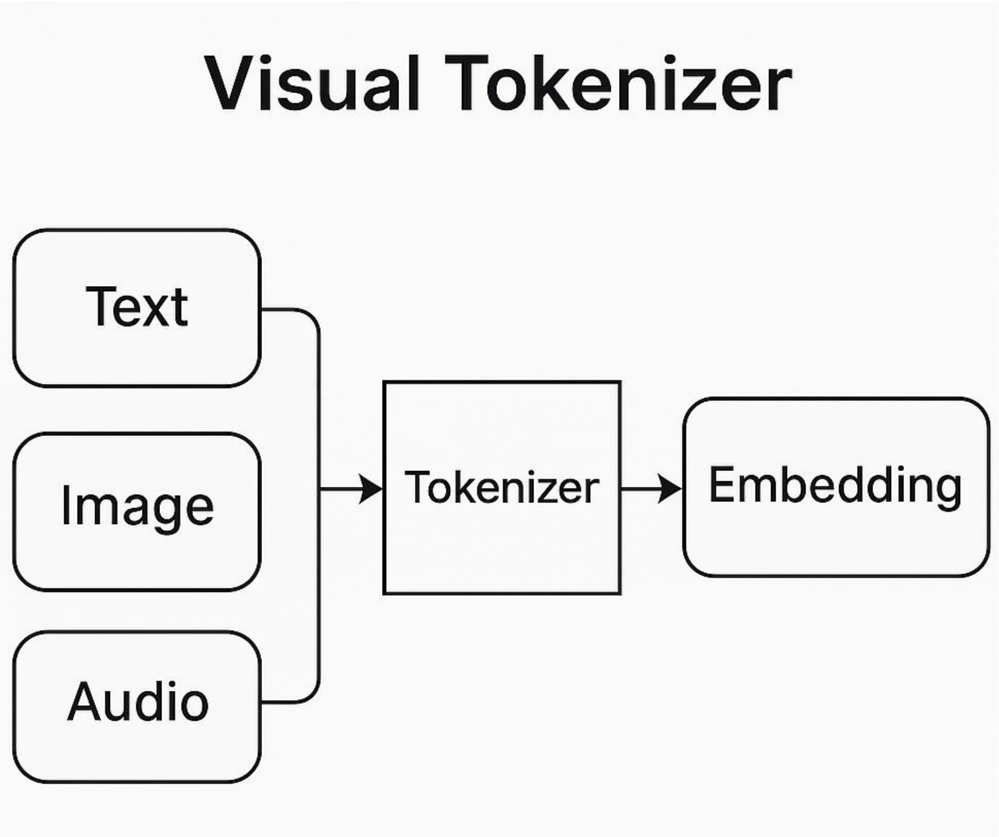


Fig 8 – Visual Tokenizer: Converts text, image, and audio into tokens for multimodal transformers.

Tools & Technologies

Datasets Used in Multimodal AI

Dataset Type	Modalities	Examples
Captions	Text-Image / Video	MSCOCO, CC-3M, LAION, WebVid
Conversation	Text-Image / Video	VQAv2, TextVQA, WebVidQA, EgoQA
Reasoning	Text-Image / Video	CLEVR, NExT-QA, CLEVRER
Integration	Multimodal	LLaVA-Instruct, Video-LLaVA, VideoChat2

Common multimodal datasets used for training and evaluating generative and understanding models

Applications

- AI storytelling with visuals + narration
- Educational tools with interactive visuals
- Sketch-to-image design assistants
- Generative art and concept visuals
- Media content automation
- Marketing and branding material
- Virtual assistants with multimodal input

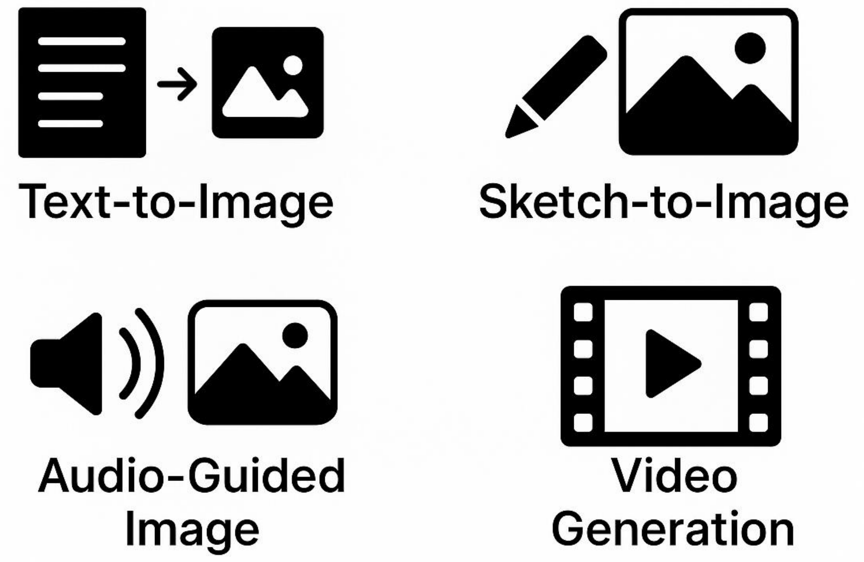


Fig 7 – Generation Modes: Examples: text-to-image, sketch-to-image, and more.

Applications

- Gaming:character generation, environments
- Smart assistants: voice and visual inputs
- Prototyping in fashion and interior design
- Creative coding environments
- Virtual worlds: Metaverse integration
- Accessibility: converting modalities
- Custom avatars and 3D objects

Advantages

- Cross-modal creativity
- Real-time generation support
- Productivity boost
- User-guided content generation
- Aligns with creative goals
- High flexibility in output
- Emerging industry standard

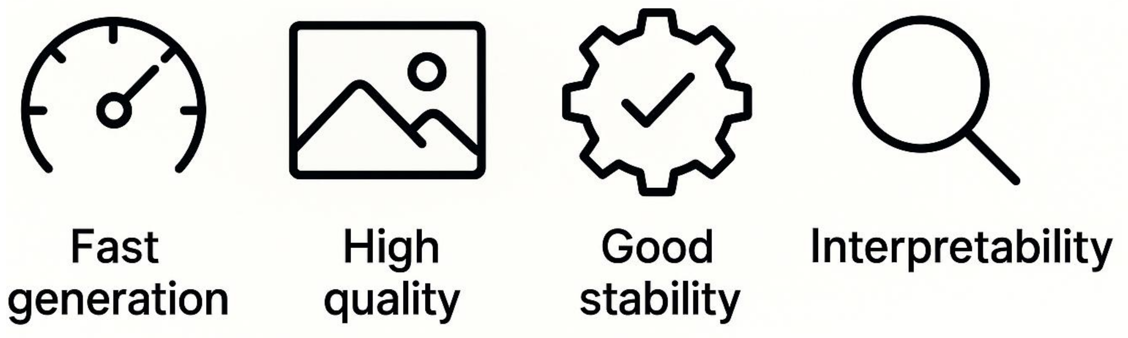


Fig 8 – Model Traits: Lists key qualities: speed, quality, stability, interpretability.

Disadvantages

- High computational cost
- Bias and hallucination risks
- Lack of transparency
- Difficult to interpret outputs
- Requires large training datasets
- Data privacy concerns
- Complexity in architecture

	GANs	VAEs	Diffusion
Speed	Medium	High	Low
Quality	High	Low	High
Stability	Low	Low	High
Interpretability	Low	Low	Low

Fig 9 – Model Comparison: Table comparing GANs, VAEs, and diffusion models.

Future Scope

- Real-time multimodal AI agents
- On-device deployment with LCMs
- Edge computing support
- Explainable Multimodal AI (XMAI)
- Deeper AR/VR integration
- Voice-image-sensor fusion
- Human-AI creativity interfaces

Conclusion

- Multimodal AI drives future creativity
- Tools like GPT-4o, Sora show unified AI
- Enables seamless prompt-to-output flows
- Potential in all creative domains
- Research remains fast evolving
- Calls for responsible innovation

References

1. X. Wang, Y. Zhou, B. Huang, H. Chen, and W. Zhu, “Multi-modal generative AI: Multi-modal LLMs, diffusions and the unification,” *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2025.
2. F. Zhan, Y. Yu, R. Wu, J. Zhang, S. Lu, L. Liu, A. Kortylewski, C. Theobalt, and E. Xing, “Multimodal image synthesis and editing: The generative AI era,” *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, 2024.
3. S. Roy, A. Das, R. Ahuja, and A. Banerjee, “Sketch-to-image synthesis via diffusion models,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Paris, France, Oct. 2023, pp.
4. L. Wang, Y. Liu, Z. Zhao, and X. Wu, “DiffSketching: Sketch-controlled image generation,” *arXiv preprint*, arXiv:2304.00877, Apr. 2023.
5. N. Rodis *et al*., “Multimodal explainable artificial intelligence: A comprehensive review of methodological advances and future research directions,” *IEEE Access*, vol. 12, pp. 830792–830845, Nov. 2024. doi: 10.1109/ACCESS.2024.3467062

References

6. X. Zhang *et al*., “Unified multimodal understanding and generation models: Advances, challenges, and opportunities,” *Journal of LaTeX Class Files*, vol. 14, no. 8, Oct. 2024.

7. Y. Dang *et al*., “Towards explainable and interpretable multimodal large language models: A comprehensive survey,” *arXiv preprint*, 2024.

8. T. Baltrušaitis, C. Ahuja, and L.-P. Morency, “Multimodal machine learning: A survey and taxonomy,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019. doi: 10.1109/TPAMI.2018.2798607

Thank You

Questions & Discussion Welcome

Feel free to ask anything related to the topic.

Presented by:

- ❑ *ABHINAND S*
- ❑ *MES22AD002*
- ❑ **Dept. of Artificial Intelligence & Data-Science**
- ❑ **MES College of Engineering,Kuttippuram**