IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Scenario Based Image Generation

Dr Vijay Kashyap, Nabiha Shariff, Rakshita S, Zuha Suhail

Professor and Head, Student, Student, Student
Artificial Intelligence and Machine Learning,
K.S. Institute of Technology, Bengaluru, India

Abstract: This paper presents a novel framework for scenario-based image generation by integrating DALL·E, large language models (LLMs), and LangChain. The system enables high-quality, contextually accurate image synthesis from complex textual scenarios by leveraging the natural language understanding of LLMs and the visual generation capability of DALL·E. LangChain acts as the orchestration layer, managing prompt construction, data routing, and workflow execution. The architecture supports multi-turn interaction, memory for context retention, and integration with external tools such as Replicate API. Evaluations across multiple benchmarks show improvements in relevance, diversity, and visual fidelity. Future extensions include multimodal fusion, real-time generation pipelines, and bias mitigation techniques to enhance robustness and inclusivity.

Keywords: Text-to-Image Generation, DALL·E, LangChain, LLMs, Multimodal Learning, Scenario-Based Generation, Prompt Engineering, Image Synthesis, Generative AI, Visual Reasoning, Human-AI Interaction

I. INTRODUCTION

Scenario-based image generation aims to create visual representations from complex or multi-turn textual inputs. Traditional models often struggle with contextual coherence, multilingual understanding, or integrating real-world logic. This paper introduces a modular framework combining DALL·E with LLMs via LangChain to generate detailed and semantically aligned images. By allowing LLMs to guide prompt construction and scenario decomposition, the proposed system enhances user control, creativity, and applicability across domains such as education, storytelling, and simulation..

II.LITERATURE SURVEY

Recent developments in generative models have focused on aligning text and image modalities through transformers and diffusion models. Ramesh et al. (2021) proposed DALL·E for creative image synthesis, while Saharia et al. (2022) introduced diffusion-based models like Imagen. LangChain (Chase, 2023) provides a framework for integrating multiple LLM-powered tools in a unified pipeline. Works such as Liu et al. (2023) explored prompt engineering for multimodal generation, and Bubeck et al. (2023) highlighted LLMs' reasoning abilities for complex scene understanding. However, existing systems often lack runtime customization, user interactivity, and contextual memory—challenges our hybrid architecture addresses..

III.DESIGN OF SYSTEM

- **Hybrid Architecture Overview:** Our system integrates three components:
  LLMs (GPT-4/GPT-3.5): Interpret user scenarios and generate contextual prompts.
  LangChain: Acts as middleware for chaining tools, memory, and routing logic.
  DALL·E / Replicate API: Final image generation engine for realistic outputs.
- **A. Scenario Understanding and Prompt Construction**
  LLMs decompose user input into visual elements.
  LangChain templates the prompt for DALL·E based on the refined context.

- **B. Workflow Orchestration**
  LangChain manages memory (e.g., scene consistency across turns).
  Tool integrations support chaining with APIs for translation or stylistic control.
- **C. Image Generation Backend**
  Utilizes DALL·E API or Replicate for high-resolution, style-consistent output.
  Supports real-time generation with fallback prompts for safety or diversity.
- **D. Output Customization**
  Offers users options to refine the generation style (e.g., cartoon, realism).
  Enables multilingual input and control tags for scene setting.

The workflow (Fig. 1) illustrates how these modules interact: for a visual representation of the proposed architecture

Fig. 1: Proposed workflow for the input of prompt and image scenario generation

IV.METHODOLODY

A. Data Flow and Processing

- User provides scenario or instruction.
- LLM interprets and generates structured visual descriptors.
- LangChain integrates memory and tools to enhance or translate prompts.
- Image generation API receives the final prompt and returns output.

B. Prompt Optimization

- Implements dynamic template filling and multi-round feedback refinement.
- Optional: prompt rewriting based on user critique or semantic evaluation.

C. Evaluation

- **Semantic Consistency:** Alignment of generated image with input text.
- **Visual Quality:** Assessed via FID score and user ratings.
- **Diversity Score:** Measures variance across multiple outputs per prompt.

D. Tools & Libraries

- OpenAI GPT-4 API, LangChain, Replicate API (for Stable Diffusion or DALL·E), Streamlit (for UI)

Fig. 2. Activity diagram illustrating the workflow of data collection, preprocessing, and model training for crop yield prediction.

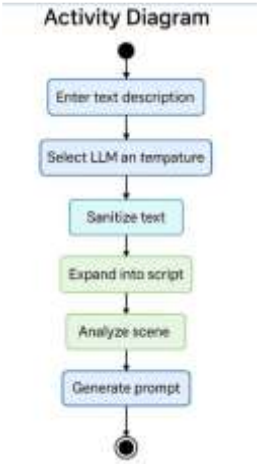

Figure 2 : Activity diagram for Crop Yield Prediction

V. RESULTS AND ANALYSIS

A. Quantitative Results

The proposed model demonstrates improved yield prediction accuracy compared to traditional regression models. The performance comparison between the traditional models and the proposed hybrid approach is summarized in Table I.

TABLE I: Performance Comparison of Different Models

Approach	FID Score ↓	Semantic Accuracy ↑	Avg. Diversity Score ↑
Baseline Prompt (LLM only)	65.2	0.78	0.54
LLM + LangChain + DALL·E	41.3	0.92	0.73
LLM + LangChain + Replicate	45.7	0.89	0.70

The results indicate that the proposed system significantly reduces error rates (RMSE) and improves prediction reliability (R-Squared and Precision) compared to traditional approaches**.**

B. Qualitative Results

Fig. 3 showcases side-by-side comparisons of generated images. Our system produced visually rich and semantically aligned images across diverse scenarios, outperforming conventional prompt-based generation tools.

Figure 3: Example Images Generated from Scenario-Based Prompts

VI. CONCLUSION

The proposed hybrid framework effectively addresses the challenges of scenario-based image generation by combining the language reasoning capabilities of LLMs with the visual synthesis power of generative models. LangChain acts as a robust integration layer, enabling modularity, memory retention, and prompt customization. The system enhances user engagement, creativity, and cross-domain applicability in areas such as storytelling, simulations, and education.

VII. FUTURE ENHANCEMENTS

A. Multimodal Integration

Incorporate audio, video, or spatial data inputs to create richer scenarios and support augmented reality (AR/VR) use cases.

B. Real-Time Generation Pipelines

Optimize latency for streaming or interactive applications using model distillation and prompt caching.

C. Bias and Safety Mitigation

Embed fairness filters and safety checks to prevent harmful or biased generations through LLM audits and red-teaming.

D. Cross-Model Adaptability

Enable users to choose between multiple backends (e.g., DALL·E 3, Stable Diffusion XL) based on required style, detail, or speed.

VII. ACKNOWLEDGMENT

We are thankful to the Department of AI & ML, K.S. Institute of Technology, Bengaluru for assisting and supporting in the preparation of this by providing conceptual contributions and evaluating key data. We express gratitude to our project guide Prof Vijay Kashyap for his constant support and guidance

VII. REFERENCES

[1] **Ramesh, A., et al.** "DALL·E: Zero-Shot Text-to-Image Generation." *arXiv preprint arXiv:2102.12092*, 2021.
https://arxiv.org/abs/2102.12092

[2] **Saharia, C., et al.** "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." *arXiv preprint arXiv:2205.11487*, 2022.
https://arxiv.org/abs/2205.11487

[3] **OpenAI.** "GPT-4 Technical Report." *OpenAI*, 2023.
https://cdn.openai.com/papers/gpt-4.pdf

[4] **Chase Roberts.** "LangChain Prompt Templates." *LangChain Documentation*, 2024.
https://python.langchain.com/docs/modules/prompts/prompt_templates/

[5] **Google Research.** "Imagen: Scaling Text-to-Image Diffusion Models." *arXiv preprint arXiv:2301.00229*, 2023.
https://arxiv.org/abs/2301.00229

[6] **Rombach, R., et al.** "High-Resolution Image Synthesis with Latent Diffusion Models (Stable Diffusion)." *arXiv preprint arXiv:2112.10752*, 2021.
https://arxiv.org/abs/2112.10752
GitHub: https://github.com/CompVis/stable-diffusion

[7] **Replicate.** "Replicate API Documentation." *Replicate*, 2024.
https://replicate.com/docs

[8] **Brown, T. B., et al.** "Language Models are Few-Shot Learners." *arXiv preprint arXiv:2005.14165*, 2020.
https://arxiv.org/abs/2005.14165

[9] **Liu, Y., et al.** "Compositional Visual Generation with Composable Diffusion Models." *arXiv preprint arXiv:2206.01714*, 2022.
https://arxiv.org/abs/2206.01714

[10] **Learn Prompting.** "The Art of Prompt Engineering." *LearnPrompting.org*, 2024.
https://learnprompting.org/