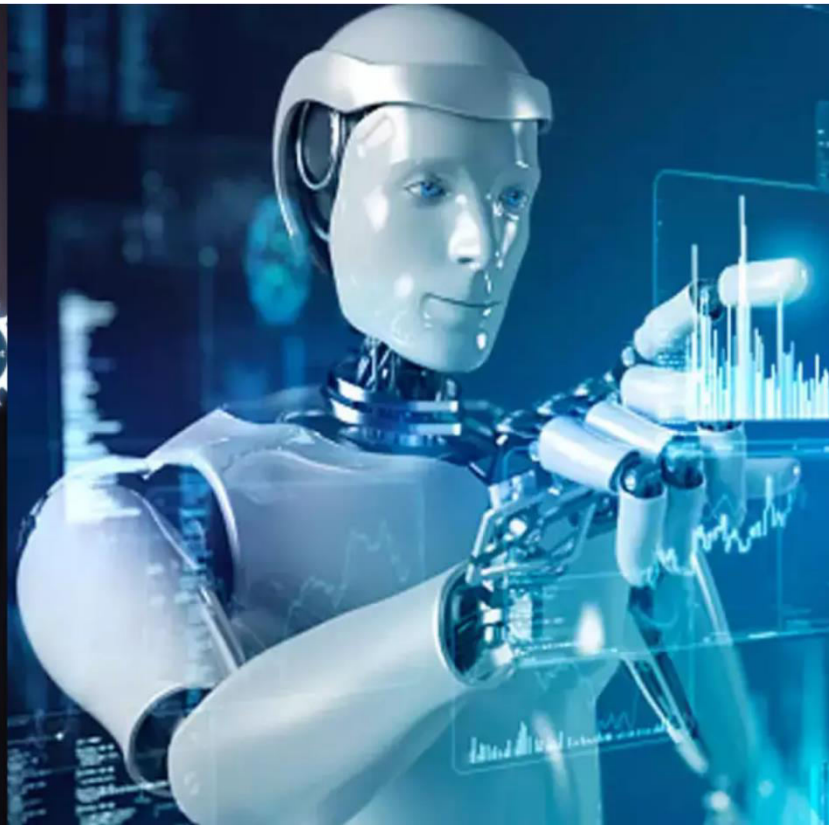
IJIRSET
International Journal of Innovative Research in Science | Engineering | Technology

ISSN(O): 2319-8753
ISSN(P): 2347-6710

International Journal of Innovative Research in Science Engineering and Technology (IJIRSET)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)



Impact Factor: 8.699

Volume 14, Issue 11, November 2025

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 11, November 2025

|DOI: 10.15680/IJIRSET.2025.1411029|

Prompt-to-3D Model Generation

Mr. B N Babar[1], Ms. Nikita Nevase[2], Atharva Hajare[3], Jignesh Randhe[4], Atharava Pawar[5], Shyam Lade[6]

Department of Artificial Intelligence and Data Science, Suman Ramesh Tulsiani Technical Campus- Faculty of Engineering.Khamshet, Pune, Maharashtra, India[1]

Department of Artificial Intelligence and Data Science, Suman Ramesh Tulsiani Technical Campus- Faculty of Engineering, Khamshet, Pune, Maharashtra, India [2-6]

ABSTRACT: This paper presents a structured pipeline for prompt-to-multi-view image generation, transforming a natural language description into a set of high-quality 2D images from different angles. The process begins by interpreting a user-provided textual prompt as a design specification, followed by the generation of a 2D image using the Stable Diffusion XL (SDXL) text-to-image model, which captures the visual semantics of the input. To prepare this image for multi-view generation, the background is removed to isolate the primary object, ensuring a clean and distraction-free input. Using the MV Adapter, a set of consistent multi-angle 2D views of the object is synthesized, enhancing geometric accuracy and enabling richer 3D understanding in later stages.

This end-to-end pipeline addresses key challenges in text-to-3D generation as multi-view consistency, object isolation, and semantic alignment with user intent without the need for paired text-3D datasets. The proposed approach is well-suited for real-world applications such as game development, virtual and augmented reality (XR), digital twins, product visualization, and educational content creation. By simplifying the visual asset creation process, this work contributes toward making 3D design more accessible, efficient, and user-driven.

I. INTRODUCTION

The creation of 3D models traditionally requires complex tools and significant manual effort. With the rise of generative AI, it is now possible to automate the early stages of 3D content creation. The proposed pipeline simplifies this process by converting text prompts into multiple 2D views of an object, which can later assist in building a 3D model.

Starting from a user's textual description, the system uses **Stable Diffusion XL (SDXL)** to generate a high-quality 2D image that captures the concept visually. The image is then processed to remove its background, isolating the main subject. The **MV Adapter** tool is then employed to create multiple images of the same object from different angles. These multi-angle images enable a more complete understanding of object geometry and can serve as a foundation for future 3D reconstruction.

This work aims to make 3D content creation more accessible, faster, and less dependent on manual modelling skills.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

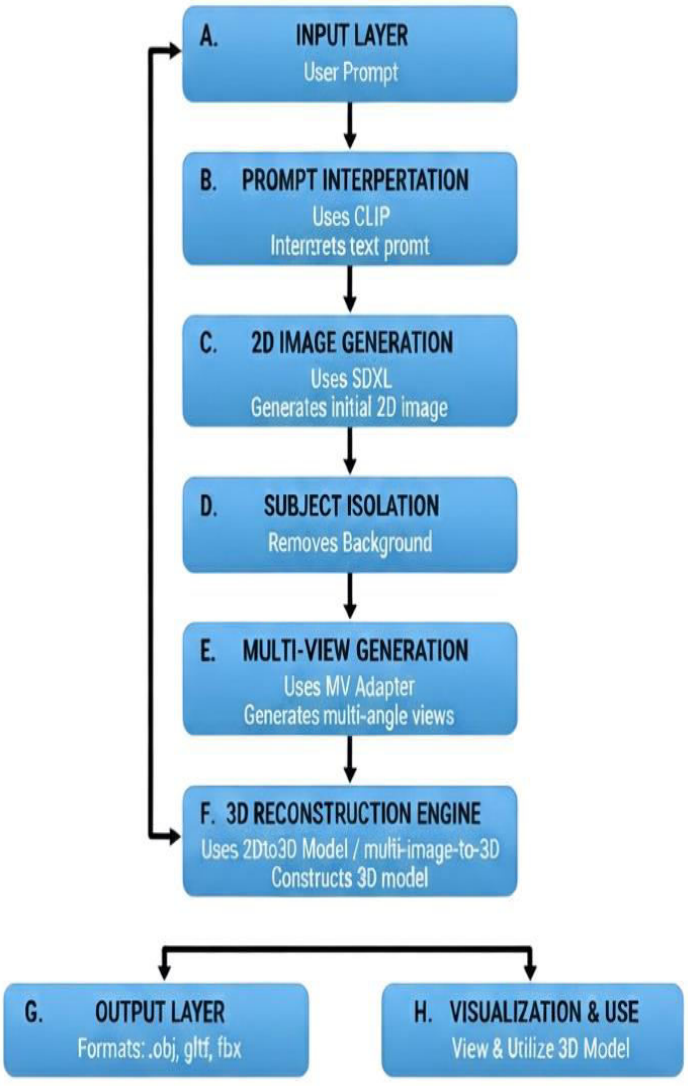
|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 11, November 2025

|DOI: 10.15680/IJIRSET.2025.1411029|

II. PROPOSED METHODOLOGY

PROMPT-TO-3D MODEL GENERATION SYSTEM ARCHITECTURE



The proposed pipeline follows a structured, modular approach to transform a text prompt into multiple clean, high-quality 2D images of an object viewed from different angles. Each step in the pipeline plays a distinct role in ensuring visual consistency, geometric clarity, and readiness for future 3D reconstruction.

1. Text-to-Image Generation (SDXL)

The process begins with a **natural language prompt** provided by the user. This input is interpreted as a design specification and processed using the **Stable Diffusion XL (SDXL)** model—a state-of-the-art, open-source text-to-image diffusion model known for producing **high-resolution (1024×1024px)** and **semantically accurate** outputs.

SDXL generates a **visually rich 2D representation** of the described object, capturing essential visual attributes such as shape, proportion, colour, lighting, and texture. This image serves as the **foundation for all subsequent stages** of the pipeline.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 11, November 2025

|DOI: 10.15680/IJIRSET.2025.1411029|

Why SDXL was chosen:

- **Open-source** and deployable locally, providing full control and no dependency on paid APIs.
- **High-resolution outputs** (1024×1024px) with detailed texture and structure.
- **Superior prompt understanding** compared to previous versions and competitors.
- Best balance between **quality, flexibility, and pipeline integration**.

Alternative Options Considered:

- **DALL·E 3**: Closed API with limited customization and commercial restrictions.
- **Midjourney**: Not open-source, lacks API support, limiting automation.
- **Stable Diffusion 1.5**: Lower quality, especially in resolution and semantic accuracy.
- **Imagen (Google)**: Not publicly available for implementation.
- **Kandinsky 2.2**: Weaker performance in prompt adherence and visual output.

2. Background Removal (rembg)

The image generated by SDXL is then passed through a **background removal stage** to isolate the object from any surrounding noise or scenery. For this step, the **rembg Python library** is employed. This tool removes the background, ensuring only the **primary subject** remains, enhancing **clarity and focus**.

This step is critical not only for **visual enhancement** but also for **downstream tasks**, particularly **multi-view synthesis**, where a clean foreground improves boundary detection and consistency across generated views.

Why rembg was chosen:

- **Local, fast, and automated**—ideal for scalable, automated pipelines.
- **Free and open-source**, based on the **U²-Net** architecture.
- **Seamless Python integration**, making it easy to plug into custom workflows.
- **Zero-cost**, avoiding reliance on external services or paid APIs.

Alternative Options Considered:

- **Remove.bg**: High-quality results but relies on a paid API and internet access.
- **Adobe Photoshop (Select Subject)**: Manual and unsuitable for automation or batch processing.
- **PhotoRoom API**: Requires payment and lacks flexibility for large-scale pipelines.
- **MODNet / FBA Matting**: Good accuracy, but more complex to integrate and overkill for this use case.

3. Multi-View Image Generation (MV Adapter)

The isolated object image is then passed into the **MV Adapter (Multi-View Adapter)**, a specialized model designed to generate **angle-consistent 2D views** of the object. It typically synthesizes **front, back, side, and top** perspectives, while maintaining consistency in **colour, lighting, and geometry**.

This multi-view dataset provides a **richer and more complete visual representation** of the object, addressing the limitations of single-view images and laying the groundwork for **accurate 3D reconstruction**.

Why MV Adapter was chosen:

- Designed specifically to generate **geometrically consistent** views.
- **Lightweight and efficient** compared to full 3D diffusion models.
- Enables **better 3D model fidelity** by preserving structural coherence.
- **Easy integration** into automated image pipelines.

Alternative Options Considered:

- **Using SDXL to generate multiple angles**: Results in **inconsistent object geometry** across views.
- **Zero-1-to-3 / DreamFusion**: Full 3D diffusion models, computationally heavy and slower.
- **3D-aware GANs (e.g., EG3D, StyleNeRF)**: Require training, complex setup, and high compute.
- **LGM / 3DiM**: Still experimental, lacking maturity and reliability for production use.

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 11, November 2025

|DOI: 10.15680/IJIRSET.2025.1411029|

4. Output

The final output of the proposed pipeline is a set of six or more consistent, high-quality 2D images, each representing the same object from a unique angle.

These images are background-free, geometrically aligned, and semantically consistent, making them ideal for:

- Input into 3D reconstruction tools (e.g., multi-view-to-3D systems)
- Visualization and rendering in XR, gaming, or design applications
- Dataset creation for training future text-to-3D models

III. DISCUSSION

The proposed pipeline leverages strengths of modern 2D generative models while mitigating their weaknesses for 3D reconstruction. Key advantages include:

- **Practicality without paired data:** By using 2D generative priors and multi-view synthesis, the method avoids needing large paired text–3D datasets.
- **Modularity and extensibility:** Each stage can be upgraded independently—swap SDXL for a stronger text-to-image model, or replace Lemonator with a different multi-view reconstruction method.
- **Improved multi-view coherence:** MV Adapter reduces view inconsistencies that typically derail multi-image reconstruction.

However, there are limitations and open challenges:

- **Fine-grained geometry:** For objects with thin structures (e.g., chain links, translucent materials), 2D-based multi-view reconstruction may fail to recover accurate geometry without real multi-view photo capture or stronger priors.
- **Failure modes from hallucination:** SDXL may hallucinate details inconsistent with physical plausibility; these can propagate through multi-view synthesis. Careful prompt engineering and validation are required.
- **Evaluation scale:** Lack of large-scale, diverse benchmarks for text→2D→multi-view→3D pipelines limits comprehensive evaluation
  across varied object classes and styles.

IV. FUTURE WORK

Future work will focus on extending the system to generate complete 3D models from the produced multi-view images. Efforts will be made to improve geometric precision and texture accuracy for complex objects. Additionally, integration and ensuring ethical, unbiased content generation will be key priorities.

REFERENCES

1. Kopper, J. (2024). *SDXL debiasing mechanism: a fairness improvement to image generation through diffusion kernels* (Master's thesis, Pontifícia Universidade Católica do Rio Grande do Sul).
2. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., ... & Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
3. Huang, Zehuan, et al. "Mv-adapter: Multi-view consistent image generation made easy." *arXiv preprint arXiv:2412.03632* (2024).
4. Shi, R., Chen, H., Zhang, Z., Liu, M., Xu, C., Wei, X., ... & Su, H. (2023). Zero123++: a single image to consistent multi-view diffusion base model. *arXiv preprint arXiv:2310.15110*.
5. Baldridge, J., Bauer, J., Bhutani, M., Brichtova, N., Bunner, A., Castrejon, L., ... & Malik, C. (2024). Imagen 3. *arXiv preprint arXiv:2408.07009*.
6. Kuoppala, P. (2025). Real-time image generation for interactive art: developing an artwork platform with ComfyUI and TouchDesigner.
7. Bianchi, F. Kalluri, P. Durmus, E. Ladhak, F. Cheng, M. Nozza, D. Hashimoto, T. Jurafsky, D. Zou, J. and Caliskan, A. (2023). Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency.
8. Friedrich, F. Brack, M. Struppek, L. Hintersdorf, D. Schramowski, P. Luccioni, S. and Kersting, K. (2023). Fair

International Journal of Innovative Research of Science, Engineering and Technology (IJIRSET)

Impact Factor 8.699

|www.ijirset.com |A Monthly, Peer Reviewed & Refereed Journal| e-ISSN: 2319-8753| p-ISSN: 2347-6710|

Volume 14, Issue 11, November 2025

|DOI: 10.15680/IJIRSET.2025.1411029|

diffusion: Instructing text-to-image generation models on fairness. arXiv preprint arXiv:2302.10893.

9. Podell, D. English, Z. Lacey, K. Blattmann, A. Dockhorn, T. Müller, J. Penna, J. and Rombach, R. (2023). Sdxl: Improving latent diffusion models for high-resolution image synthesis. arXiv preprint arXiv:2307.01952.
10. Rombach, R. Blattmann, A. Lorenz, D. Esser, P. and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. arXiv preprint arXiv:2112.10752
11. Schramowski, P. Brack, M. Deiseroth, B. and Kersting, K. (2023). Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.
12. Shen, X. Du, C. Pang, T. Lin, M. Wong, Y. and Kankanhalli, M. (2023). Finetuning text-to image diffusion models for fairness. arXiv preprint arXiv:2311.07604

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

INNO SPACE SJIF Scientific Journal Impact Factor

doi cross ref

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details