Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Generation of 3D Image by Editing the Background Using Only Prompt

Indresh Upadhyay[1], Prof. Prakash Devale[2], Prof. Sumedh Vithalrao Dhole[3], Prof.P. A Dixit[4], Prof. Chetan more[5], Prof. Milind Gayakwad[6*]

[1]*Bharati Vidyapeeth Deemed to Be University, College of Engineering, Pune 411043, India*
[2]*Bharati Vidyapeeth Deemed to Be University, College of Engineering, Pune 411043, India*
[3]*Bharati Vidyapeeth Deemed to Be University, College of Engineering, Pune 411043, India*
[4]*Bharati Vidyapeeth Deemed to Be University, College of Engineering, Pune 411043, India*
[5]*Bharati Vidyapeeth Deemed to Be University, College of Engineering, Pune 411043, India*
[6*]*Bharati Vidyapeeth Deemed to Be University, College of Engineering, Pune 411043, India*

ARTICLE INFO

Received: 31 Dec 2024

Revised: 20 Feb 2025

Accepted: 28 Feb 2025

ABSTRACT

This study explores an innovative approach to creating 3D visuals by altering backdrops with textual cues, leveraging recent advancements in AI, machine learning, and generative models. Traditionally, 3D modeling required expert knowledge and sophisticated tools, but AI-powered generative models now allow users to create realistic 3D scenes using simple text descriptions. The study aims to develop a flexible model capable of multimodal shape denoising, conditional synthesis, and shape interpolation, catering to applications in virtual reality (VR), gaming, design, and entertainment. A key challenge in 3D shape development is integrating global and local information, with hierarchical latent spaces offering a solution by capturing both large-scale structures and fine-grained details. Hierarchical DE denoising diffusion models (DDMs) are employed to produce high-quality 3D shapes through iterative refinement. Surface reconstruction ensures smooth, realistic models for practical applications, such as digital art and VR. Additionally, prompt-based 3D creation enables quick prototyping, enhancing creative workflows in industries like filmmaking and gaming. Ethical considerations, including copyright and ownership of AI-generated content, are also discussed.

Keywords: 3D Image Generation, Background Editing, Textual Prompts, Generative Models, Multimodal Shape DE noising, Conditional Synthesis, Shape Interpolation

I. INTRODUCTION

Image-based modelling is becoming a more comprehensive, cost-effective, portable, and adaptable method for many applications as digital cameras and scanners are becoming commonplace sources of input data in many application areas. Reconstructing 3D pictures from 2D photographs has long been a difficult task. Computer vision, computer graphics, and image processing are all involved [1]. Enhancing the input photos requires image processing, restoring and interpolating the scene requires computer vision, and synthesizing the scene requires computer graphics. The visualization research field offers a wide range of methods, from common sceneries to medical imaging. We have attempted to apply domain knowledge for modelling in the suggested study [1]. We postulate that it is feasible to see the whole scene if there are many photographs of a certain domain accessible. The key to proving this theory is image-based rendering that makes use of domain expertise. In essence, the concept involves observing, contemplating, and visualizing the scene supported by scene information. Multiple photos of an environment are combined to create a view using three-dimensional reconstructions [2]. To create a synthetic impression of panoramic nature, we first patched together many photographs. One of the most important issues is determining the right stitch locations in the photographs and the image sequence needed to create composite scenes. For this, we have put forth plans [3]. Missing scene interpretation to make the created scene comprehensive is another issue. According to the current methods drawing a panoramic view based on the user's preference requires at least two stages. In the first step, we must sequentially stitch many photographs together based on certain places in the images [4]. In the second step, we must interpolate the missing scene by making an educated guess as to where the missing portions of the images are. The

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

issue with difficult to rebuild a 3D vision from 2D photographs without understanding the fundamentals of the surroundings and the characteristics of camera. In light of this, the rehabilitation project must address three main issues. The first step is to create a panoramic view from many photos, followed by the interpolation of missing details from specific scenes and, lastly, the use of volume rendering to create a 3D picture of the scene [5]. The panorama may be effectively saved and displayed on contemporary graphics technology. The Point Grey Ladybird 5.0 spherical camera, which is shown in Fig 1.1, makes it simple to overcome issues with picture capturing and creating 360-degree, non-graphic real-world photos [6]. The quick convergence of artificial intelligence (AI), natural language processing (NLP), and sophisticated computer vision methods is best shown by the creation of 3D graphics via background editing with textual cues. AI has evolved dramatically in recent years, allowing machines to interpret and synthesize information across modalities with ease [4] [5]. New techniques for creating digital material have emerged as a result of this interdisciplinary convergence, such as the capacity to produce three-dimensional (3D) visuals from just textual inputs [7]. Once limited to highly qualified specialists with strong modelling software and technological know-how, consumers may now transform 3D sceneries, change surroundings, and create realistic spatial pictures by just inputting a descriptive query. The increasing need for user-friendly and accessible tools for creating 3D material is what led to this breakthrough. In addition to a thorough grasp of perspective, lighting, and spatial design, traditional 3D modelling techniques needed extensive familiarity with sophisticated software like Autodesk Maya, Blender, or CAD tools. Teams or people without technical skills faced substantial obstacles due to the time, effort, and expertise required[8].

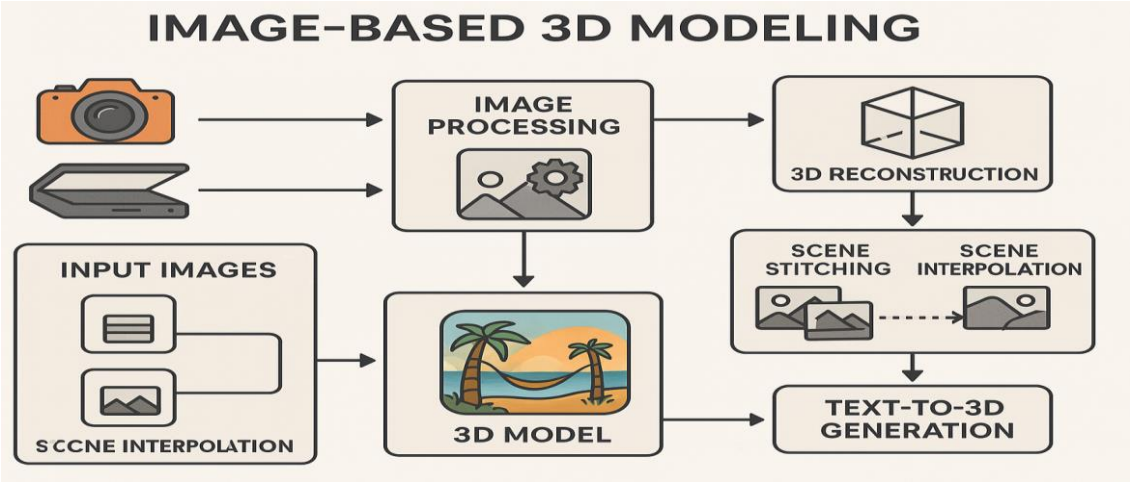


Fig.1. Image Based 3D Modeling

Computer Vision

The concept of machine vision, or computer vision, involves enabling computers to interpret and understand images, similar to human vision [9][10]. This technology uses image sensors, such as cameras, to capture electromagnetic radiation and relies on algorithms for tasks like object recognition, video summarization, and industrial quality control. Computer vision integrates several fields, including artificial intelligence, robotics, signal processing, physics, neurobiology, mathematics, and geometry. It involves automating processes like image feature matching, segmentation, and statistical analysis. Machine learning plays a key role in teaching computers to recognize patterns and mimic cognitive behavior [11].

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

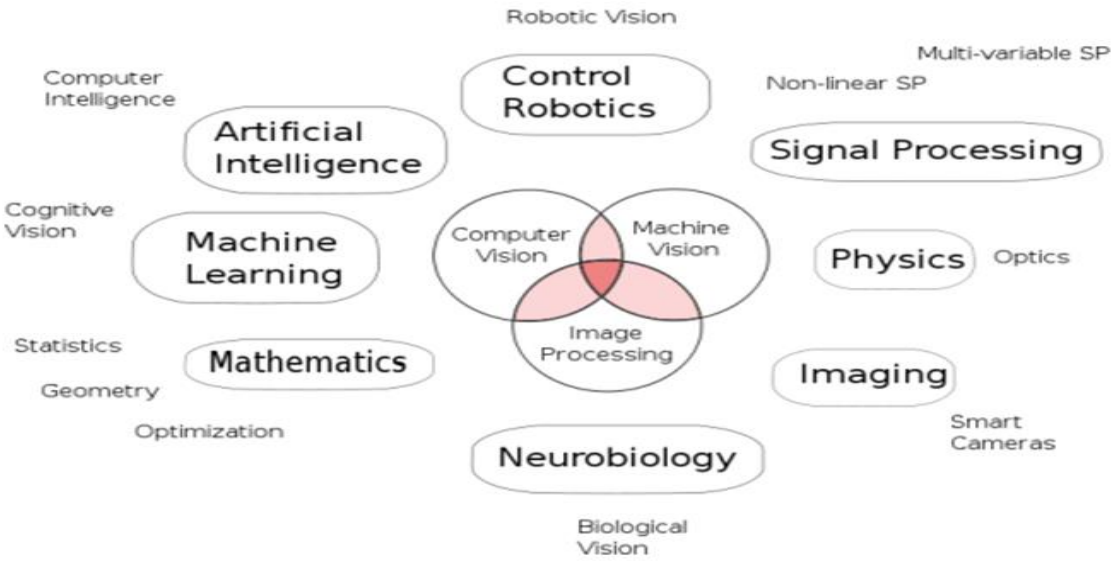


Fig 2. Fields associated with Computer Vision

Depth Perception

Vision is crucial for perceiving and understanding the environment, with depth perception being key to determining an object's distance in 3D space [12]. Typically, 3D space is represented along the x, y, and z axes, but when real-world scenes are captured by cameras, the depth (z-axis) is lost due to perspective projection. The projection process involves the object's visual rays intersecting a view plane, with terms like "horizon line" and "ground line" describing the relationship between the camera, ground, and image plane. While cameras and human eyes follow the same principle of perspective transformation, restoring the lost depth dimension from a single image is challenging[14]. Techniques like motion cues and stereo vision, requiring multiple images from different perspectives, are used to address this [13] [14]. This ability to estimate depth from a single image is gaining importance in computer vision applications, such as for autonomous driving [15] [16].

II. RELATED WORK

Text-to-Image Generation Models

Reference	Technique Used	Dataset (Method)	Research Gaps
Andrew Melnik et al. (2024)	Deep learning techniques for StyleGAN face creation and modification.	StyleGAN, GAN inversion, face restoration, deep fake applications.	Lack of multi-modal control and robust training metrics for face editing.
Bo Li et al. (2024)	End-to-end 3D-aware picture generating and editing model with multi-modal conditional inputs (text, reference images, noise).	3D GANs, style transfer, attribute updates using text descriptions.	Poor disentanglement of shape and appearance; multi-modal control still weak.
Zhengzhe Liu et al. (2023)	DreamStone: text-guided 3D shape creation using images, CLIP feature mapping, and SVR model.	CLIP features, pre-trained single-view reconstruction (SVR) model.	Lack of generalization in multi-domain applications; limited stylization beyond SVR models.
Zhengzhe Liu et al. (2023)	DreamStone: A text-guided 3D shape creation method with enhanced texture and structure	CLIP features, SVR model, pre-trained text-to-image diffusion models.	Limited control over appearance and shapes beyond the pre-trained models; scalability issues.

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	mapping using pre-trained text-to-image diffusion models.		
Zhijie Wang et al. (2024)	Prompt Charm: A mixed-initiative solution for easier text-to-image prompting using multi-modal prompts.	Stable Diffusion model, image generation via prompt engineering.	Challenges with improving user interaction and prompt optimization.
Alex Nichol et al. (2022)	Text-to-image diffusion models for quick 3D object creation.	Single-view diffusion models, 3D point cloud generation.	Lack of state-of-the-art sample quality; slower processing time compared to modern generative models.
Feng-Lin Liu et al. (2024)	Sketch Dream: Text-driven 3D creation and editing with sketch-based input and NeRF (Neural Radiance Fields) creation.	3D ControlNet, sketch-based multi-view picture creation, depth guidance.	Ambiguity in 2D-to-3D translation, limited free-view editing; requires better multi-modal integration and more control.
Yiying Yang et al. (2024)	Scene123: 3D scene generation model combining video generation models with implicit neural representations.	Video generation models, Masked Autoencoders (MAE), GAN-based loss.	Difficulties in large-scale scene generation, ensuring consistency across multiple views.
Wa James Tam et al. (2011)	Study of visual comfort in stereoscopic 3D TV, examining factors affecting viewer comfort.	No dataset specified, theoretical research on 3D-TV comfort.	Limited real-world testing for varied 3D-TV setups and viewer experiences.
Xiaozhi Chen et al. (2017)	3D object identification for autonomous driving using stereo images, CNNs, and LIDAR data.	KITTI dataset, stereo images, LIDAR data.	Limitations in handling different environmental conditions and non-ideal sensor data.

2.2 3D Image Generation

Reference	Technique Used	Dataset (Method)	Research Gaps
Yang Chen et al. (2023)	Control3D: Text-to-3D generation with additional hand-drawn sketches for enhanced control.	NeRF, ControlNet, pre-trained differentiable photo-to-sketch model.	Lack of interactivity in text-to-3D systems for creative manipulation; limited user control in existing methods.
Vivian Liu et al. (2022)	Study of hyperparameters and prompt phrases for improving text-to-image generation results.	Evaluation of 5493 generated images based on 51 issues and 51 styles.	Inexperienced users struggle with generating optimal results; unclear reasoning for success/failure of prompts.
Jingbo Zhang et al. (2024)	Text2NeRF: Text-driven 3D scene generation with complex geometry and high-fidelity textures using NeRF.	NeRF, text-to-image diffusion model, monocular depth estimation.	Lack of high-quality, multi-view consistency; improvement needed in generating diverse, realistic 3D environments.
Yuhan Guo et al. (2024)	Image Variant Graph: A visual aid for analyzing prompt-image	Prompt history, Image Variant Graph for	Limited visual insight for users to improve their prompts; need

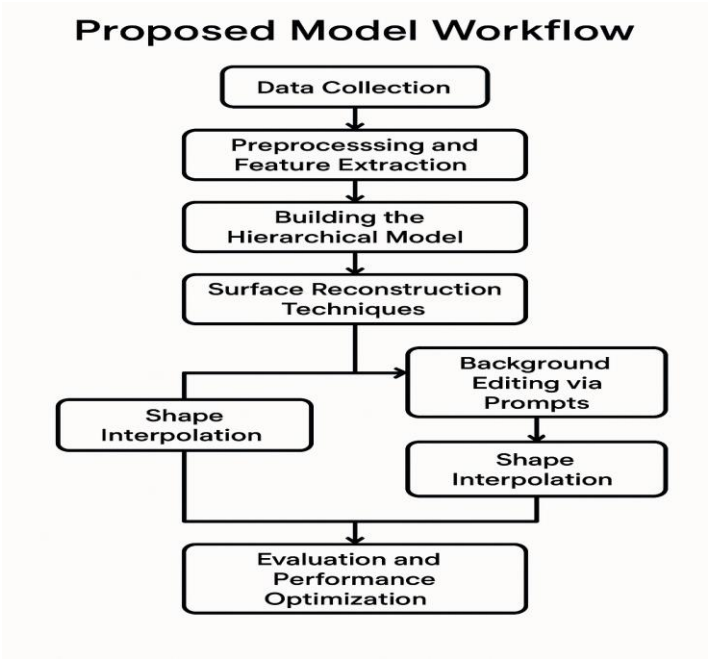
Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

	history and enhancing control over image generation.	analyzing text-to-image results.	for more intuitive and adaptive prompting systems.
Xiaoqin Peng et al. (2024)	FocusSculp: Text and image prompt-driven 3D object editing with precise adjustments and NeRF creation.	3D ControlNet, point clouds, text and image prompts.	Struggles with 2D-to-3D translation ambiguity; better control over shape and geometry required for user-driven editing.
Quan Zhou et al. (2024)	Semi-supervised learning for 3D medical image segmentation with instance-specific adaptation.	Public datasets, internal dataset, semi-supervised learning (SSL) models.	Model performance deteriorates due to the lack of accurate segmentation guidance; limited availability of 3D medical data.
Tiankai Hang et al. (2024)	Language-guided face animation: Using motion data to animate a still face picture.	StyleGAN, recurrent motion generator, multiple domain testing (faces, anime, dogs).	Limited research on using motion semantics in language; challenges with high-quality video synthesis from still images.
Jingyu Zhuang et al. (2023)	Dream Editor: Text-driven neural field editing for 3D scene modification.	Mesh-based neural fields, score distillation sampling, text-to-image diffusion models.	Difficulty in editing neural fields without sacrificing realism; challenges with local area editing in complex 3D environments.

III. METHODOLOGY

The methodology focuses on generating high-quality visual outputs from text prompts, ensuring accurate data handling, tracking, and analysis. Key steps include text-to-image/3D generation, data preprocessing, experiment tracking, and optional background editing for enhanced flexibility.



The proposed methodology for 3D object generation from text prompts involves several key steps. First, diverse 3D datasets are collected, paired with textual descriptions. The data is then preprocessed, with 3D features extracted into formats like point clouds or meshes, and text processed using NLP techniques. A hierarchical model is built to capture both global and detailed features, trained with a denoising diffusion model to align shapes with text prompts. Surface reconstruction techniques ensure smooth, realistic meshes, and background editing allows seamless scene

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

integration. Shape interpolation enables dynamic object creation, and evaluation optimizes performance. The model is fine-tuned to balance visual quality, efficiency, and consistency for real-time applications.

System Architecture

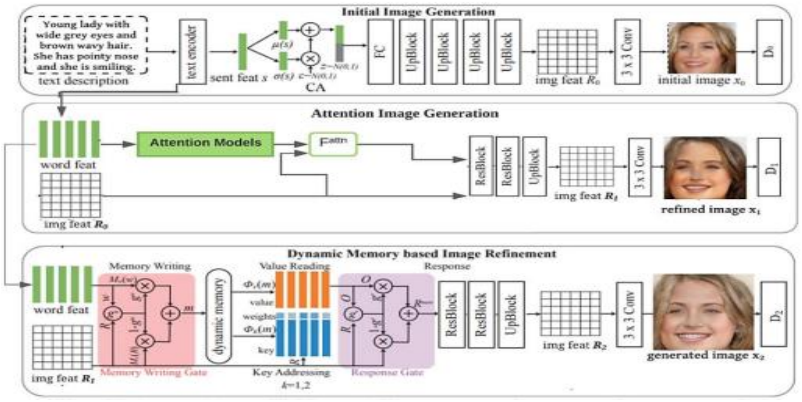


Fig 3. System Architecture

Text-to-Image/3D Model Generation:

- **Text-to-Image/3D Model Generation**: This module converts user text prompts into images or 3D models using AI tools like Stable Diffusion or Shap-E, relying on pre-trained neural networks for accurate and creative outputs.
- **Data Transformation and Processing**: Text and image data are preprocessed using functions like get_title_prompts, ensuring clarity and accessibility of generated outputs, along with dynamically creating paths for easy visualization and access.
- **Experiment Tracking and Logging**: Integrated with Weights & Biases (W&B), this module tracks experiments, logs metadata, and manages generated artifacts, ensuring reproducibility, collaboration, and real-time logging for transparency.
- **Visualization and Analysis**: This module visualizes generated outputs with tools like Matplotlib and integrates W&B for logging visualizations, performance metrics, and user insights, aiding in detailed analysis and feedback.
- **Optional Background Editing**: This feature allows background modification of generated images based on updated prompts, enabling users to change scenes while maintaining object integrity, ideal for creative applications.

Algorithms Used

- **Generative AI:** Stable Diffusion uses a diffusion probabilistic model, combining Transformer-based text encoders and latent space optimization to generate high-quality images or 3D models from text prompts.
- **Logging and Sharing:** Generated images and associated metadata are logged to Weights & Biases (W&B) using their Image API, ensuring efficient storage, organization, and easy sharing for collaborative analysis.
- **Model Fine-Tuning:** Generative models like Stable Diffusion are fine-tuned using domain-specific datasets, leveraging transfer learning and regularization to enhance model performance while retaining pre-trained knowledge for accurate results.

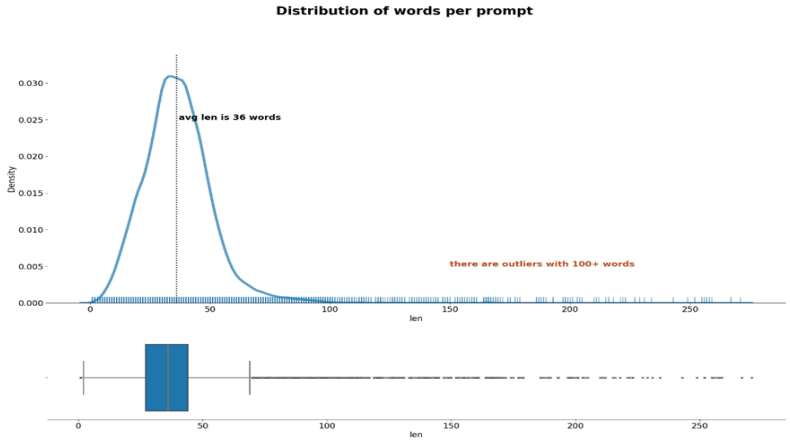
IV. RESULT AND DISCUSSION

The study's results provide valuable insights, highlighting trends, patterns, and relationships. Statistical significance, performance metrics, and comparative analysis validate findings, while anomalies suggest areas for further exploration. The implications support practical applications, guiding future research and improvements in methodologies.

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

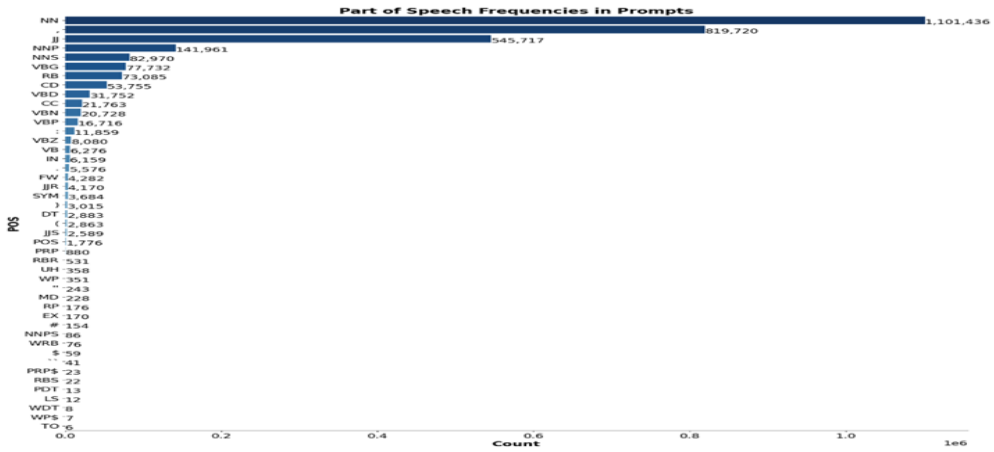
Research Article

Distribution of Words Per Prompt



The analysis of prompt lengths shows that most prompts are concise, averaging 36 words, with a right-skewed distribution. While most prompts cluster around the average, some outliers exceed 100 words, suggesting a need for handling longer prompts efficiently. Strategies like truncation or summarization could optimize processing, as the model performs best with shorter prompts. Future research could categorize prompts by length to improve system responses.

Part of Speech (POS) frequencies

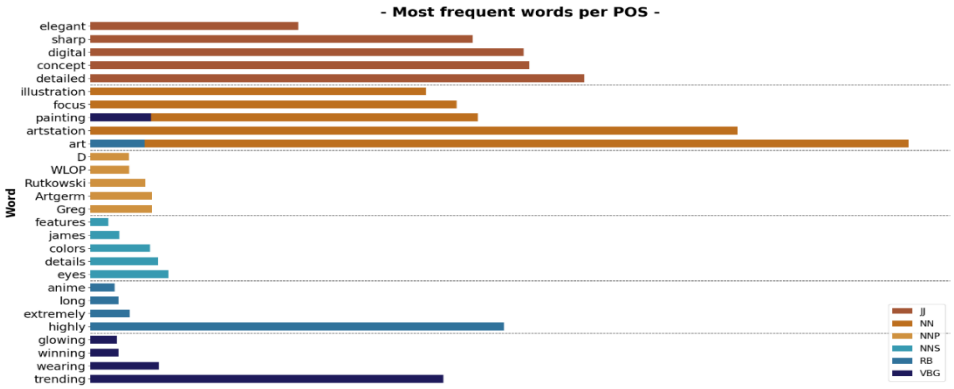


The analysis of Part of Speech (POS) frequencies reveals that nouns, particularly singular ones, dominate the dataset, indicating that prompts focus more on objects and entities than actions. Verbs, pronouns, and determiners are less frequent, suggesting descriptive rather than conversational prompts. Adjectives and adverbs appear moderately, while conjunctions and prepositions are rare, indicating short, direct prompts. This suggests that prompts are primarily noun-centric, which can guide NLP models to efficiently handle entity-heavy inputs with minimal complexity.

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Frequent Words Categorized by Part of Speech



The visualization of frequent words categorized by Part of Speech (POS) reveals a strong focus on artistic and conceptual themes. Nouns dominate, with terms like "concept," "illustration," and "painting," indicating an emphasis on visual and creative subjects. References to renowned artists (e.g., WLOP, Rutkowski) reinforce the art-centric nature of the prompts. Plural nouns such as "features," "colors," and "details" highlight the focus on multiple elements in digital or visual aesthetics. Adjectives like "elegant," "sharp," and "glowing" suggest users often specify artistic qualities, while adverbs such as "extremely" emphasize intensity. Verbs are less frequent, indicating that the prompts are primarily descriptive. Overall, the dataset leans toward art and design, focusing on styles and well-known artists, rather than action-oriented content, suggesting its application in AI-generated art and visual design.

Mask

This alteration isolates the glider, allowing the viewer to pay close attention to its movement and design without any distractions from the environment. The absence of the natural backdrop changes the perception of the video, making it feel more abstract. The focus shifts entirely to the object—highlighting the simplicity and fluidity of the glider's flight. This implementation could evoke a sense of pure freedom or solitude, as it removes the context of the surrounding world, emphasizing the experience of flight itself.

Original

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

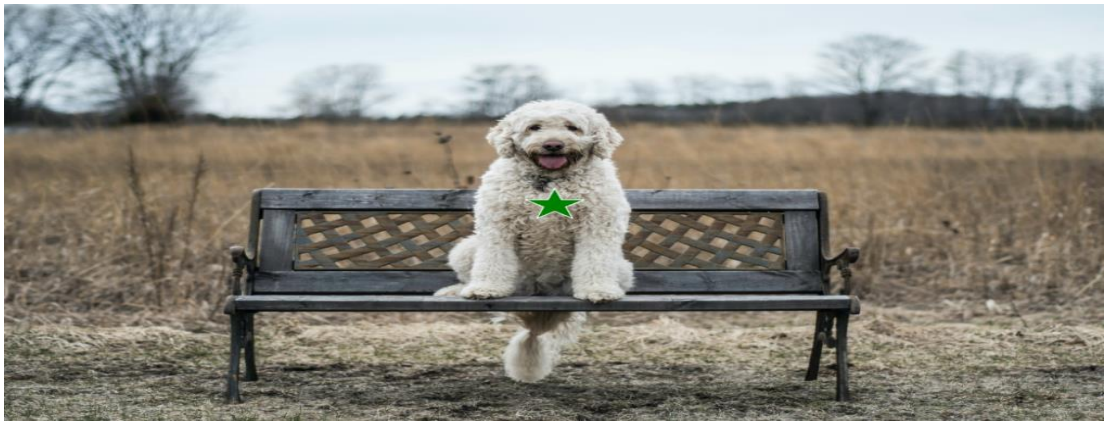
where the paraglider is shown with its original colors and set against a scenic landscape. The combination of the bright colors of the paraglider and the picturesque natural environment creates a dynamic and immersive experience. This video captures the thrill of paragliding in a real-world context, showing the interaction between the human experience of flight and the beauty of nature. The presence of the surrounding hills and the village below gives the viewer a sense of scale and perspective, emphasizing the vastness of the environment and the exhilaration of flying through it.

Removed

the parachute has been removed, leaving only the glider itself. This implementation gives a surreal and minimalist feel to the scene, as the glider seems to be suspended or floating without the usual element of the parachute that defines paragliding. The removal of the parachute shifts the focus purely to the object and its movement through space. This alteration can symbolize a sense of freedom, detachment, or even weightlessness, as it evokes a more abstract representation of flight. The absence of the parachute invites the viewer to focus on the motion and form of the glider, removing the usual context of paragliding and offering a more conceptual, artistic view of the experience.

The three videos depict paragliding with varying visual elements. The original video shows a balanced scene with both the paraglider and landscape. The mask video isolates the paraglider, focusing on the object. The third video removes the parachute, creating a surreal effect and emphasizing the glider's form. These changes shift the viewer's interpretation of the gliding experience.

Results: dog with points



In this image, the dog is sitting on a bench in a natural outdoor setting. The dog’s chest is marked with a green star. The star seems to add an extra layer of emphasis to the dog, potentially highlighting it as the focal point of the image. This could symbolize recognition or importance, suggesting that the dog is the "star" of the scene. The background remains natural and untouched, which draws attention to the dog's playful and calm nature in the environment.

Results: dog replaced with mask

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

In this image, the background has been replaced with an artificial scene, and the dog is placed on a swing. The replacement of the background removes the connection to the natural environment, creating an altered, surreal context. The swing seems to be placed in a controlled or stylized setting, perhaps evoking a more whimsical or abstract interpretation of the scene. The dog, now in a less natural context, becomes a subject of focus in an environment that is not as grounded in reality.

Results: dog replaced with mask 2

Similar to the second image, the dog is now placed on a swing, but this time, the swing itself seems to be slightly altered, potentially making it appear as if the dog is floating or suspended in an unusual way. The altered background removes it further from reality, and the dog now appears in an environment that feels even more conceptual. This change emphasizes the dog's playful nature but, in an abstract, imaginative setting.

Results: dog replaced with mask 3

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

This image features the dog sitting on a swing with a similarly altered, surreal background. The swing's altered position or the lack of clear attachment further amplifies the sense of abstraction, allowing the viewer to focus on the idea of the dog's behavior in an environment detached from its natural context. The presence of the swing but the absence of realistic grounding for the scene adds a surreal, playful, and thought-provoking element to the image.

In summary, the progression from the first image (realistic with emphasis on the dog) to the last (abstract with altered context) showcases how changes in background and object manipulation can shift the viewer's perception. Each variation takes the original concept of a dog sitting in an outdoor space and gradually removes it from reality, ultimately creating a playful, conceptual space that invites the viewer to explore the scene through a different lens. The idea of "masking" the background and changing the context challenges the viewer's understanding of the dog's natural setting, emphasizing the imaginative aspect of the visual composition.

CONCLUSION

The experiment on generating 3D models from text prompts using Shap-E has demonstrated significant potential in bridging natural language and 3D object creation, but also highlighted areas where optimization is necessary for enhancing model quality and accuracy. Strategies such as fine-tuning the model's response to different prompt structures, enhancing training datasets, and employing advanced AI techniques like multi-view imagery and depth maps have shown promise in improving the overall output. The integration of multi-view images for constructing Neural Radiance Fields (NeRF) models provides an efficient method for 3D visualization, though the process requires careful alignment of these images to ensure consistency and optimize for computational limitations, particularly when constrained by GPU capabilities. Balancing the speed and quality of this process remains a challenge, but it's clear that further advancements in hardware optimization and algorithm refinement will improve real-time performance. Additionally, the combination of RGB images with depth maps is essential for creating accurate and textured 3D models increasing patient outcomes.

Statements and Declarations

Ethical Approval

"The submitted work is original and not have been published elsewhere in any form or language (partially or in full), unless the new work concerns an expansion of previous work."

Consent to Participate

"Informed consent was obtained from all individual participants included in the study."

Consent to Publish

"The authors affirm that human research participants provided informed consent for publication of the research study to the journal."

Funding

"The authors declare that no funds, grants, or other support were received during the preparation of this manuscript."

Competing Interests

"The authors have no relevant financial or non-financial interests to disclose."

Availability of data and materials

"The authors confirm that the data supporting the findings of this study are available within the article."

Acknowledgements

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of competing interest

Journal of Information Systems Engineering and Management
2025, 10(51s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

REFERENCES

[1] Melnik et al., "Deep learning techniques for StyleGAN face creation and modification," StyleGAN, GAN inversion, face restoration, deep fake applications, 2024.
[2] Li et al., "End-to-end 3D-aware picture generating and editing model with multi-modal conditional inputs (text, reference images, noise)," 3D GANs, style transfer, attribute updates using text descriptions, 2024.
[3] Z. Liu et al., "DreamStone: text-guided 3D shape creation using images, CLIP feature mapping, and SVR model," CLIP features, pre-trained single-view reconstruction (SVR) model, 2023.
[4] Z. Liu et al., "DreamStone: A text-guided 3D shape creation method with enhanced texture and structure mapping using pre-trained text-to-image diffusion models," CLIP features, SVR model, pre-trained text-to-image diffusion models, 2023.
[5] Z. Wang et al., "Prompt Charm: A mixed-initiative solution for easier text-to-image prompting using multi-modal prompts," Stable Diffusion model, image generation via prompt engineering, 2024.
[6] Nichol et al., "Text-to-image diffusion models for quick 3D object creation," Single-view diffusion models, 3D points cloud generation, 2022.
[7] F.-L. Liu et al., "Sketch Dream: Text-driven 3D creation and editing with sketch-based input and NeRF (Neural Radiance Fields) creation," 3D ControlNet, sketch-based multi-view picture creation, depth guidance, 2024.
[8] Y. Yang et al., "Scene123: 3D scene generation model combining video generation models with implicit neural representations," Video generation models, Masked Autoencoders (MAE), GAN-based loss, 2024.
[9] W. J. Tam et al., "Study of visual comfort in stereoscopic 3D TV, examining factors affecting viewer comfort," No dataset specified, theoretical research on 3D-TV comfort, 2011.
[10] X. Chen et al., "3D object identification for autonomous driving using stereo images, CNNs, and LIDAR data," KITTI dataset, stereo images, LIDAR data, 2017.
[11] Y. Chen et al., "Control3D: Text-to-3D generation with additional hand-drawn sketches for enhanced control," NeRF, ControlNet, pre-trained differentiable photo-to-sketch model, 2023.
[12] V. Liu et al., "Study of hyperparameters and prompt phrases for improving text-to-image generation results," Evaluation of 5493 generated images based on 51 issues and 51 styles, 2022.
[13] J. Zhang et al., "Text2NeRF: Text-driven 3D scene generation with complex geometry and high-fidelity textures using NeRF," NeRF, text-to-image diffusion model, monocular depth estimation, 2024.
[14] Y. Guo et al., "Image Variant Graph: A visual aid for analyzing prompt-image history and enhancing control over image generation," Prompt history, Image Variant Graph for analyzing text-to-image results, 2024.
[15] X. Peng et al., "FocusSculp: Text and image prompt-driven 3D object editing with precise adjustments and NeRF creation," 3D ControlNet, point clouds, text and image prompts, 2024.
[16] Q. Zhou et al., "Semi-supervised learning for 3D medical image segmentation with instance-specific adaptation," Public datasets, internal dataset, semi-supervised learning (SSL) models, 2024.
[17] T. Hang et al., "Language-guided face animation: Using motion data to animate a still face picture," StyleGAN, recurrent motion generator, multiple domain testing (faces, anime, dogs), 2024.