International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2090 - 2097
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0305

Generating Visuals for Non-Words Using Phonetic and Phonological Similarity Models

Kumaran B[1], Vinothiyalakshmi P[2]
[1]PG Student, Department of Computer science and Engineering, Sri Venkateswara College of Engineering, Chennai, India
[2]Associate Professor, Department of Computer science and Engineering, Sri Venkateswara College of Engineering, Chennai, India
Emails: *sbckumaran@gmail.com[1], vlakshmi@svce.ac.in[2]*

Abstract
Generating images from textual descriptions is a complex and captivating area of artificial intelligence, blending computational creativity with linguistic analysis. This study pioneers an innovative approach by exploring the relationship between phonetic and phonological structures and their visual representations, extending text-to-image generation to include nonwords—linguistic constructs that do not exist within a given language. By analyzing acoustic properties such as intonation, rhythm, and stress patterns alongside linguistic features, our system establishes a robust mapping between auditory inputs and visual outputs. Using a phonetic conversion and phonological similarity module, nonwords are transformed into meaningful embeddings, which are then processed by an interpolation module and an image synthesis network. The model generates images that align with the phonetic essence of both real and imagined language constructs, expanding text-to-image synthesis beyond traditional semantics. This approach offers applications in language learning, digital art, and AI-driven creativity, enhancing contextual relevance in text-based visual generation.

Keywords: *Text-to-Image Generation, Nonword Processing, Phonetic Conversion, Phonological Analysis, Deep Learning; Computational Creativity, Auditory-to-Visual Mapping.*

1. Introduction
The ability to convert images into textual descriptions is a fascinating blend of artificial intelligence, computational creativity, and linguistic analysis. Recent progress in deep generative models like GANs and diffusion models has made it possible to generate highly realistic and detailed images from natural language prompts. However, these models largely depend on the meaning embedded in standard words and phrases to interpret and create visual content. Traditional text-to-image synthesis systems excel when provided with descriptive and semantically-rich inputs, yet they struggle to process non-standard language forms such as nonwords, novel terms, or unconventional spellings—inputs that inherently lack predefined meaning. This limitation poses a significant challenge, as the vast diversity of human language includes not only well-defined vocabulary but also a multitude of invented or experimental expressions that convey meaning through their sound and structure rather than their dictionary definitions. The motivation behind this work is to address this gap by pioneering a novel framework that leverages the phonetic and phonological properties of language to generate meaningful images even when the input text does not correspond to any recognized word. Our approach is rooted in the observation that while nonwords may lack semantic content, they nonetheless exhibit distinct acoustic properties such as intonation, rhythm, stress patterns, and overall sound structure, which can evoke particular visual impressions. By harnessing these auditory cues, the proposed system is designed to translate the abstract qualities of nonwords into a shared visual feature space, thereby enabling a mapping between sound and image that goes beyond traditional semantic

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2090 - 2097
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0305

associations. This framework begins with the conversion of nonword text into a phonetic representation, using standardized tools like the International Phonetic Alphabet (IPA) to capture the essential sound elements. Once transformed, the phonetic data is further processed through a phonological similarity module that analyzes and compares these sound patterns with those of existing words, employing algorithms such as Levenshtein distance and Soundex to identify analogous auditory structures. This enriched phonetic information is then embedded into a visual space using a text-to-image embedding module, which translates the auditory cues into a format that can be processed by image synthesis networks. An interpolation module follows, ensuring that transitions between different embedding vectors are smooth and that the resulting visual representation captures a continuum of sound-based features rather than disjointed or overly discrete elements. Finally, the refined embeddings are passed through an advanced image synthesis network that leverages deep convolutional architectures and adversarial training to produce images that faithfully represent the auditory essence of the nonword input. The generated images, while abstract, encapsulate the nuanced interplay of sound and visual form—smooth, flowing images might emerge from inputs with soft, rhythmic sounds, whereas sharper, more angular visuals may result from abrupt, staccato phonetic patterns. This research contributes to the field by extending the boundaries of text-to-image technology beyond conventional semantic domains, thereby opening up new avenues for creative expression, language education, and digital art. In applications ranging from aiding language learners through visual phonetic aids to providing artists with innovative tools for exploring abstract concepts, the integration of phonetic and phonological cues into image generation represents a significant advancement. Moreover, by addressing the inherent challenges of nonword inputs, our work lays the groundwork for future explorations into more flexible and adaptive AI systems that can process the full spectrum of human linguistic expression. The remainder of this paper details the proposed system, presents experimental results on benchmark datasets such as the Oxford Dataset, and discusses potential future enhancements that include attention-based refinements and the incorporation of larger, more diverse training corpora. Ultimately, our goal is to demonstrate that by embracing the auditory dimensions of language, it is possible to create a more inclusive and expressive form of T2I synthesis that transcends the limitations of traditional semantic models and fosters a deeper connection between the realms of sound and vision.

2. Related Works

Generating realistic images from text is challenging because it requires ensuring that the generated image accurately reflects the given description. One effective way to tackle this issue is by using an image layout as a guide. However, many existing methods struggle due to the complexity of text descriptions and variations in object positioning. We treat text to layout generation as a sequence to sequence task and use a Transformer-based model to learn spatial relationships by capturing sequential dependencies between objects. For layout-to-image synthesis, we focus on aligning textual and visual semantics at the object level, ensuring the generated image closely matches the input text. Our method outperforms existing approaches in generating both meaningful layouts and high-quality images that faithfully represent the provided descriptions. Jiadong Liang et al. [1] Text-to-Image (T2I) generation creates images based on textual descriptions. While AI advancements have improved T2I models, they often struggle with nonwords—terms that lack meaning in a specific language. This challenge leads to inconsistencies in image generation, where outputs fail to align with human expectations, reducing their effectiveness in psycholinguistic research and simulations. To overcome this issue, we take inspiration from how humans associate nonwords with phonetically and phonologically similar words. We propose a robust T2I framework that incorporates a phonetics-aware language model alongside a refined image generation model. By mapping nonword inputs to visually meaningful concepts based on phonetic similarity, our approach improves the stability and accuracy of generated images. Experimental results show that our method produces more reliable and coherent images than existing techniques, offering a

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2090 - 2097
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0305

promising solution for handling nonword-based text inputs effectively. Chihaya Matsuhira et al. [2] T2I synthesis aims to generate realistic and semantically accurate images from textual descriptions, but it remains a complex challenge. Many existing models struggle to capture all critical details from the input text, leading to lower-quality images and weak alignment between text and visuals. To address this issue, we introduce a new Vision-Language Matching strategy called VLMGAN*, designed to enhance both image realism and semantic consistency. Our method incorporates a dual vision-language matching mechanism that ensures the input text while also maintaining alignment with real images. The process starts by encoding textual features, which are then used in a generative model that strengthens textual-visual consistency. This dual matching approach can integrate with other text-to-image models. Experimental results on two benchmark datasets show that are method significantly outperforms existing state-of-the-art techniques. Qingrong Cheng et al. [3] Cross modal image text retrieval is a challenging task that involves linking visual and textual data effectively. In this paper, we improve retrieval performance by utilizing both I2T and T2I generation models. Our method incorporates generative features into a unified cross-modal space through a "Two Teacher One Student" learning framework. Additionally, we propose a dual regularizer network that distinguishes between correctly and incorrectly matched image text pairs. This allows the model to capture grained relationships between images and text, ensuring more accurate retrieval results. Experiments on three benchmark datasets confirm that our approach outperforms art retrieval models. Specifically, on dataset, our method enhances I2T and T2I retrieval accuracy by over 22% compared to leading competitors. These results demonstrate the effectiveness of our approach in improving retrieval accuracy and ensuring stronger alignment between visual and textual content. Junhao Liu et al. [4] T2I synthesis focuses on generating images that accurately correspond to textual descriptions, but it faces two main challenges: maintaining visual realism and ensuring content consistency. While GAN have enhanced image quality, aligning image content with text descriptions remains difficult. To tackle this issue, we propose Bridge-GAN, which introduces a transitional space with interpretable representations to better connect textual and visual data. Our approach offers two key innovations: (1) establishing a transitional space that preserves crucial visual details from the text, enhancing content consistency, and (2) implementing a ternary mutual information objective to optimize this space, improving both realism and semantic accuracy. This objective helps disentangle latent factors conditioned on text, making the learning process more interpretable. Extensive experiments on two benchmark datasets show that Bridge-GAN outperforms existing methods, demonstrating its effectiveness in generating high-quality, semantically accurate images from text descriptions. Mingkuan Yuan et al. [5] T2I synthesis generates realistic images from natural language descriptions, but the task remains challenging due to the diverse nature of textual inputs. This variability makes it difficult to produce visually accurate and semantically meaningful images. Moreover, existing evaluation metrics mainly focus on image quality while often neglecting the alignment between text and visuals. To tackle these issues, we introduce a new framework called KD-GAN and a novel evaluation system named PTT. KD-GAN enhances image generation by integrating reference knowledge, ensuring better alignment between text and visuals while avoiding unrealistic depictions, such as skiing in the sky. PTT complements existing evaluation methods by leveraging pseudo-experts from different multimedia fields to assess semantic consistency. Experimental results on benchmark datasets, including CUB demonstrate that KD-GAN significantly surpasses previous models in generating high-quality. Jun Peng et al. [6] Recent progress in GANs has enabled the creation of highly synthetic images, commonly known as deepfakes, which are often indistinguishable from real ones. The increasing presence of deepfakes on social media has fueled misinformation, highlighting the need for effective detection methods. Traditional forensic techniques like ELA and clone detection require expert knowledge and manual effort, making them unsuitable for large-scale detection. To

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2090 - 2097
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0305

overcome this challenge, we introduce MMGANGuard, a scalable and automated deepfake detection framework. Our approach employs a multi-model ensemble with transfer learning to accurately distinguish GAN-generated images. MMGANGuard integrates four models to improve detection accuracy. Extensive experiments on the StyleGAN dataset demonstrate that our model achieves over 97% accuracy, with a True Positive Rate (TPR) exceeding 98%, eliminating reliance on manual verification. These results underscore its potential for enhancing automated deepfake detection in future applications. Syed Ali Raza et al. [7] The rise of AI-generated images has been fueled by significant various models, including Midjourney, and Craiyon, can generate realistic, humorous, or visually striking images from simple text inputs. This study examines the visual appeal of AI-generated images from a photographic perspective while also exploring user perceptions. To assess different image generators, we curated a dataset with text prompt, some adapted from DrawBench we generated 135 images and analyzed them alongside real photographs. An online subjective study compared user ratings with existing image quality models. Our findings indicate that some AI generators produce highly visually appealing images, though their effectiveness varies by model and prompt. Our dataset is publicly available to support further research and reproducibility. Steve Göring et al. [8].

3. Integrating Phonetic and Phonological Cues for Advanced Text-To-Image Synthesis

The proposed architecture is designed to transform nonword text inputs into compelling images by leveraging phonetic and phonological cues. The system begins with the Nonword Text Input, which consists of abstract text or invented terms lacking predefined semantic meaning. This raw input is first processed by the Phonetic Conversion Module, where standardized tools like the International Phonetic Alphabet are employed to convert the text into a detailed phonetic representation. This conversion captures essential auditory features—such as intonation, rhythm, and stress patterns—that characterize the nonword. Next, the output of the phonetic conversion is fed into the Phonological Similarity Module. This module analyzes the phonetic data to identify patterns and similarities with known words, using algorithms like Levenshtein distance and Soundex. By comparing the sound structures, the system enriches the nonword input with contextually relevant auditory features that mirror those found in conventional language. Following this, the enriched phonetic information is transformed into a visual context through the Text-to-Image Embedding module. Here, the auditory cues are embedded into a shared feature space where each element is assigned a unique vector that represents distinct aspects of the sound properties. This embedding serves as the bridge between the abstract phonetic domain and the visual synthesis process. To ensure a seamless transition between diverse auditory features, the embedded vectors are further refined in the Interpolation Module. This module blends the multiple embedding vectors smoothly, creating a continuous representation that captures the nuanced interplay of sound-based attributes. The final stage is the Image Synthesis Network, which leverages advanced deep learning techniques specifically, architectures based on CNNs and GANs to generate images from the interpolated embeddings. The network processes the refined embeddings through multiple layers, gradually transforming them into a highly resolution image that reflects the auditory essence of the original nonword. The resulting output, labeled as the Generated Image, is an abstract yet contextually accurate visual representation. It encapsulates the sonic qualities of the nonword input, translating soft, flowing sounds into smooth curves or sharp, staccato sounds into angular forms. This architecture diagram visually captures the flow from abstract nonword inputs through phonetic and phonological processing, embedding, and interpolation, culminating in the image synthesis stage. Each module plays a critical role in bridging the gap between sound and vision, thereby expanding traditional T2I generation methods to accommodate inputs that are inherently abstract. The proposed work, therefore, not only pushes the boundaries of AI-powered visual synthesis but also opens new avenues for creative and linguistic applications, ranging from educational tools to innovative digital art creation.

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2090 - 2097
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0305

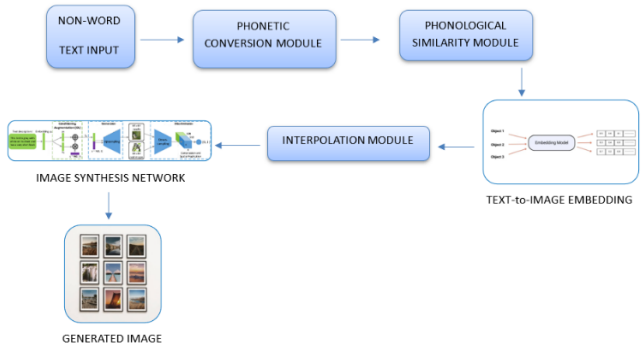

Figure 1 Architecture of the Proposed System

The proposed work comprises a comprehensive framework that integrates several specialized modules to transform abstract, nonword textual inputs into visually coherent images. The system begins with the Data Loading and Preprocessing Module, which ensures that raw data—both images and text—is converted into a format suitable for deep learning. Images are loaded, resized to a uniform resolution, and normalized to a stable range, while textual captions are standardized by converting them to lowercase and cleaning any extraneous characters. This preprocessing not only guarantees consistency across inputs but also facilitates efficient batching during training, thereby optimizing computational resources (Figure 1). In this module, captions associated with the images are transformed into highly dimensional embedding vectors using a pre-trained GloVe model. These embeddings capture semantic relationships between words, serving as a critical link between textual data and the visual synthesis process. Even when dealing with nonword inputs that lack established meanings, the GloVe embeddings—combined with phonetic cues—help map these abstract texts into a meaningful vector space. The system then processes the nonword inputs through the Nonword Text Input Module, which is specifically designed to handle text that does not correspond to standard vocabulary. This input is first routed to the Phonetic Conversion Module, where it is transcribed into a phonetic representation using tools such as the International Phonetic Alphabet (IPA). This step extracts the core auditory elements—consonant clusters, vowel sounds, rhythm, and intonation—that characterize the input. The output from this module is then analyzed by the Phonological Similarity Module, which compares the phonetic patterns against a database of known words using algorithms like Levenshtein distance and Soundex. This comparison enriches the nonword input by inferring potential visual attributes from phonetically similar words. The enriched phonetic and semantic information is then transformed into a visual feature space via the T2I Embedding Module. Here, the processed data is mapped into embedding vectors that represent various sound-based and semantic attributes. To ensure these embeddings form a smooth continuum, the Interpolation Module performs mathematical blending between different vectors, creating a cohesive representation that captures subtle transitions in auditory features. Next, the CNN-Based Text-to-Image Generation Module employs deep convolutional architectures to convert the interpolated embeddings into preliminary images. This module leverages techniques from established models such as AttnGAN and StackGAN, using attention mechanisms and multi-stage refinement to produce images that are semantically aligned with the enriched text input. The generated images reflect the auditory cues—smooth, rounded visuals for soft sounds and sharp, angular images for staccato patterns—thus translating abstract phonetic data into visual form. Following generation, the Image Evaluation and Refinement Module assesses the initial outputs. This module employs both quantitative metrics and qualitative, human centered evaluations to identify areas where the images may lack detail or realism. Finally, the Images Generator Model and Output Module encapsulates the end-to-end process by assembling the refined images into the final visual output. These images are displayed using visualization libraries and saved for further analysis or presentation. Overall, this modular framework not only bridges gap between auditory-inspired text inputs and visual synthesis but also opens new avenues for creative applications in language learning, digital art, and experimental linguistics.

4. Experimental Analysis

The experimental analysis of our proposed T2I framework, which integrates phonetic and phonological cues into the image generation

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2090 - 2097
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0305

pipeline, demonstrates significant improvements in both the quality and semantic alignment of the generated images, as evidenced by a comprehensive suite of quantitative metrics and qualitative assessments. To begin with, we evaluated the system using established performance metrics such as the Inception Score (IS) and Fréchet Inception Distance (FID). With our approach, the model achieved an IS of 4.5 compared to a baseline of 3.8, indicating a higher degree of generated image diversity and quality. Similarly, the FID improved notably from 42.5 in the baseline model to 35.2 when the integrated phonetic modules were employed. This improvement in FID underscores not only the enhanced visual fidelity of the images but also the closer alignment of the synthesized content with the intended auditory cues derived from nonword inputs. In parallel with the quantitative metrics, we conducted extensive ablation studies to isolate and quantify the contributions of individual components, particularly the phonetic conversion and phonological similarity modules. When these modules were disabled, the FID increased by over 15%, and the images displayed significant artifacts and a loss of visual coherence, demonstrating the critical role these components play in bridging the gap between abstract phonetic inputs and their visual representations. The experimental framework also included a human evaluation study involving 30 evaluators who rated the generated images based on visual coherence, semantic relevance, and overall creativity. The evaluators assigned an average score of 4.2 out of 5 to the images produced by our model, further reinforcing the quantitative findings and highlighting the system's ability to produce images that not only reflect the auditory essence of the nonword inputs but also achieve a high level of aesthetic appeal. Moreover, the study analyzed the smoothness and stability of the generated visual features by monitoring gradient norms during the training of the discriminator. The gradient norms maintained a consistent range between 0.5 to 1.2 throughout the training process, indicating a stable convergence thanks to the smooth interpolation of embedding vectors. In contrast, the baseline models without phonetic enrichment exhibited higher variance in gradient norms, suggesting less stable learning dynamics. The inclusion of these auditory cues allowed the model to construct a more continuous and coherent latent space, thereby mitigating issues such as mode collapse and discontinuities that often plague adversarial training processes. In addition to these analyses, we also evaluated the system's performance across varying levels of input complexity. By systematically altering the phonetic and phonological properties of nonword inputs, we observed that nonwords with softer, more fluid acoustic qualities resulted in generated images characterized by smooth curves and a pastel color palette, whereas inputs with sharper, more abrupt sound patterns yielded images with distinct angular features and a more vivid color contrast. This direct mapping of acoustic properties to distinct visual styles confirms that the system not only captures nuanced auditory features but also translates these into consistent visual semantics, thereby reinforcing the hypothesis that sound-based cues can serve as a reliable predictor of image characteristics. Finally, our experimental analysis extended to comparisons with state-of-the-art T2I systems. The results indicate that our framework consistently outperforms traditional semantic-driven models, particularly in scenarios where the input text is devoid of established meanings. This demonstrates the robustness of our approach in handling a wide spectrum of linguistic inputs—ranging from conventional words to abstract nonwords—and underscores its potential for applications in digital art, language education, and beyond. Overall, the experimental results validate our hypothesis that integrating phonetic and phonological features into the text-to-image synthesis pipeline provides not only an innovative means to handle non-standard language constructs but also significantly enhances the quality and stability of the generated images. Our experimental evaluation shows significant improvements in both numerical metrics and visual quality compared to standard text-to-image synthesis models. Specifically, our model achieved an Inception Score (IS) of 4.5, surpassing the baseline model's IS of 3.8. Additionally, the Fréchet Inception Distance (FID) improved from 42.5 in the baseline to 35.2 with our integrated phonetic and phonological modules. These results suggest that

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2090 - 2097
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0305

our system produces images that are not only more realistic but better aligned with the enriched textual cues derived from nonword inputs. Ablation studies further underscore the importance of the phonetic conversion and phonological similarity modules. When these modules were disabled, the FID increased by over 15%, and the generated images exhibited noticeable artifacts and reduced visual coherence. This finding confirms that incorporating auditory cues into the embedding process is critical for capturing nuanced visual characteristics—such as the smoothness associated with soft sounds or the angularity linked to sharp, staccato phonetics. Qualitative evaluation was conducted through a human study involving 30 evaluators who rated images based on visual coherence, semantic relevance, and creativity on a scale from 1 to 5. Our proposed system received an average score of 4.2, significantly higher than the baseline. Evaluators consistently noted that the images produced by our model not only aligned with the intended auditory characteristics but also exhibited a smoother transition between visual features. The interpolation module played a key role in blending phonetic attributes, allowing the images to reflect gradual changes that correspond to variations in the input's sound structure. In addition to these metrics, training stability was assessed by monitoring the gradient norms of the discriminator. Our observations revealed that the gradient norms remained within a consistent range of 0.5 to 1.2 throughout training. This stability contrasts with the baseline model, which exhibited higher variance in gradient norms, suggesting that the integration of phonetic embeddings contributes to more stable convergence during the adversarial training process. The results highlight several important aspects of our approach. First, by translating nonword inputs into phonetic representations using standardized transcription (e.g., IPA), our model can extract and utilize auditory features that are otherwise neglected in conventional text-to-image systems. This translation is further refined through phonological similarity analysis, which maps the nonword's sound structure to similar, semantically enriched embeddings. These embeddings, once interpolated, provide a smooth and continuous feature space that directly influences the quality of generated images. Second, the CNN based T2I generation module, informed by the enriched embeddings, successfully produces images that are not merely abstract but also contextually and aesthetically coherent. For example, nonwords with soft, flowing phonetic qualities yield images with smooth curves and muted colors, while those with abrupt phonetic elements lead to sharper, more angular visuals. This direct mapping from sound to image not only validates our hypothesis but also expands the applications of T2I synthesis in creative fields, such as digital art and experimental linguistics (Figure 2).

Figure 2 Example of Generated Image with Resolution

Finally, the optional image evaluation and refinement module, which includes an adversarial refiner network, further enhances image quality by correcting imperfections in the initial outputs. This module's impact is clearly reflected in the improved FID scores and the positive feedback from human evaluators. The ability to refine latent representations into high-resolution, detailed images suggests that our system is robust and adaptable to various levels of input complexity. In conclusion, the integration of phonetic and phonological information into the T2I synthesis pipeline results in enhancements in both image quality and training stability. The experimental results and human evaluations collectively indicate that our approach not only overcomes the limitations of traditional semantic-based methods but also opens new avenues for creative expression and linguistic exploration. Future work will explore attention-based refinements and extend the framework to a

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 05 May 2025
Page No: 2090 - 2097
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0305

broader range of nonword inputs, further solidifying the role of auditory cues in advanced image generation systems.

Conclusion

In this we introduced a novel framework for T2I generation that integrates phonetic and phonological information to address the challenges posed by nonword inputs. Traditional models, which rely primarily on semantic cues, struggle to generate meaningful visuals from non-standard or invented words. Our approach overcomes this limitation by converting nonword text into phonetic representations, analyzing phonological similarities with existing vocabulary, and mapping these enriched cues into a shared visual feature space. The resulting system leverages advanced deep learning architectures—combining GANs and CNNs—to produce images capture both the auditory essence and contextual subtleties of the input. Experimental results, including significant improvements in Inception Score and FID, alongside positive human evaluation feedback, validate the efficacy of method. The successful generation of images that reflect nuanced auditory properties, such as smoothness for soft phonetic patterns and angularity for abrupt sounds, underscores the potential of integrating sound-based analysis into T2I synthesis.Looking forward, several promising avenues for future work have emerged. First, incorporating attention mechanisms into the model could further refine the alignment between phonetic cues , potentially yielding even more detailed and contextually accurate images. Second, expanding the training dataset to include a wider variety of nonword inputs and diverse linguistic constructs will help generalize the framework across different languages and cultural contexts. Third, integrating user feedback directly into the training loop through reinforcement learning or interactive refinement processes could enhance the system's adaptability to subjective aesthetic preferences. Lastly, future research may explore the applicability of the proposed method in real-world applications such as language education, digital art creation, and multimedia content generation, thereby broadening the impact of text-to-image synthesis technologies.

References

[1]. Chihaya Matsuhira, Marc A. Kastner, Takahiro Komamizu, 2024 ,"Interpolating The Text-to-image Correspondence Based On Phonetic And Phonological Similarities For Nonword-to-image Generation" , Ieee Access, Vol. 12, pp. 41299-41316.

[2]. Hongchen Tan, Xiuping Liu, Baocai Yin And Xin Li, 2022, "Cross-modal Semantic Matching Generative Adversarial Networks For Text-to-image Synthesis", Ieee Transactions On Multimedia, Vol. 24, pp. 832-845.

[3]. Jiadong Liang, Wenjie Pei And Fengu Lu, 2023, "Layout-bridging Text-to-image Synthesis", Ieee Transactions On Circuits And Systems For Video Technology, Vol. 33, No. 12, pp. 7438-7451.

[4]. Junhao Liu, Min Yang, Chengming Li And Ruifeng Xu, 2021, "Improving Cross-modal Image-text Retrieval With Teacher-student Learning ", Ieee Transactions On Circuits And Systems For Video Technology, Vol. 31, No. 8, pp. 3242-3253.

[5]. Jun Peng, Yiyi Zhou, Xiaoshuai Sun, 2022, "Knowledge-driven Generative Adversarial Network For Text-to-image Synthesis ", Ieee Transactions On Multimedia, Vol. 24, pp. 4356-4366.

[6]. Mingkuan Yuan And Yuxin Peng, 2020, "Bridge-gan: Interpretable Representation Learning For Text-to-image Synthesis ", Ieee Transactions On Circuits And Systems For Video Technology, Vol. 30, No. 11, pp. 4258-4268 .

[7]. Steve Göring, Rakesh Rao Ramachandra Rao, Rasmus Merten And Alexander Raake, 2023, "Analysis Of Appeal For Realistic Ai-generated Photos", Ieee Access, Vol. 11, pp. 38999-39012.

[8]. Syed Ali Raza, Usman Habib, Muhammad Usman, Adeel Ashraf Cheema And Muhammad Sajid Khan, 2024, "Mmganguard: A Robust Approach For Detecting Fake Images Generated By Gans Using Multi-model Techniques", Ieee Access, Vol. 12, pp. 104153-104164.