MATIC: Multilingual Accurate Textual Image Customization via Joint Generative Artificial Intelligence

Chiao-Hsin Wu*
National Taiwan Normal University

I-Wei Lai†
National Taiwan Normal University

Abstract

The advancements in diffusion model enables the creation of highly detailed images. However, concurrently fusing texts and images poses significant challenges, often struggling with the maintenance of text accuracy across languages, optimal placement, and appropriate typography. To address these challenges, we introduce the Multilingual Accurate Textual Image Customization (MATIC) framework. MATIC employs the Chain-of-Thought (CoT) concept to decompose the textual image generation process into multiple steps, leveraging diverse generative artificial intelligence, including Multimodal Large Language Model (MMLLM) and diffusion model. The framework first generates the desired text and a corresponding prompt for the diffusion model based on user input. The diffusion-generated image is then examined to remove any undesired text. Meanwhile, the typographic elements are designed to align with the visual content. Finally, the textual image is fused with the aid of a grid coordinate system, evaluated by MMLLM, and further customized by the user through natural language. Experimental results demonstrate that MATIC can produce accurate, high-quality, multilingual textual images that meet user requirements across various domains, including digital marketing, graphic design, and educational content creation.

Index Terms: Computing methodologies—Artificial intelligence—Natural language processing; Computing methodologies—Artificial intelligence—Computer Vision

1 Introduction

Diffusion models have greatly improved image synthesis, especially in generating rich visual content [6, 7, 15, 16]. As technology advances, there is an increasing demand for more controllable models to precisely manipulate elements such as scene, layout, and content perspective [1, 4, 8]. However, generating *textual images*, i.e., image with textual information like movie poster, still presents challenges with multilingual support and the accurate rendering of small text elements or long text [2, 10, 14, 21].

To resolve the issues mentioned above, we propose a system named Multilingual Accurate Textual Image Customization (MATIC), a comprehensive framework combines MMLLMs and a diffusion model. We employ a CoT approach [20] to decompose an input-output task into manageable steps. Our method begins by analyzing user input to produce the desired text along with a diffusion model prompt. The diffusion model then generates an image, which is immediately processed to identify and remove any unintended text. Subsequently, the text-free image is fused with the textual element designed by MMLLM. Finally, the textual image is evaluated initially by the system and then facilitating user modifications through natural language requests.

It should be emphasized that, this framework serves as an example of the generalization of CoT. Specifically, by decomposing the process into distinct steps—such as element generation, styling, and allocation—each can be independently optimized. For example, while the current implementation utilizes a finite typographic database, our modular framework allows seamless integration of text-specific diffusion models like FontCLIP [18] or GenType [5]. By automatically assigning tasks to various generative AI, In summary, the proposed MATIC can not only provide superior textual images, but also offer potential for further advancement. The key benefits of this approach are outlined below:

- **Versatile Multilingual Text Synthesis**: Enables accurate text generation across various languages, accommodating diverse font styles, sizes, and text lengths.
- **Harmonized Hierarchical Text Placement**: Organizes textual content into clear hierarchies with precise positioning, ensuring optimal visual balance and coherence.
- **Image-Oriented Typographic Adaptation**: Dynamically adjusts typographic elements to integrate seamlessly with the image's visual characteristics for cohesive compositions.
- **Natural Language-Driven Refinement**: Enhances the generated text-image through natural language inputs, ensuring alignment with user requirements and facilitating intuitive personalization.

2 RELATED WORK

2.1 Text Rendering in Diffusion model

The evolution of diffusion models has significantly advanced text-to-image generation, yet challenges in precise text control remain critical for conveying accurate information. While notable progress has been made, limitations persist across various models. For instance, TextDiffuser2 [3] improves text accuracy with input layout planning and character-level guidance, but is constrained to English. Diff-Text [21] offers multi-language support by utilizing a contrastive image-level prompt as input, yet struggles with small and lengthy texts. These enhancements primarily focus on input modification or technological augmentation within the diffusion process, aiming to simultaneously generate text and images. Our objective of MATIC is to provide multilingual support and accommodate both lengthy text and fine glyph.

2.2 Multimodal Large Language Model

Advancements in Large Language Models (LLMs) have inaugurated a transformative era for Multimodal Language Models (MMLLMs). These models exhibit the capability to process multiple data modalities, thereby offering substantial improvements in the fields of computational linguistics and image processing. A prominent example is Lumos [17], which leverages multilingual functionalities to enhance the precision of text detection in images. Similarly, DOCLLM [19] marks a significant progression in automating text layout generation across diverse document formats. Within this context, MMLLMs are employed to address a range of challenges associated with text rendering, encompassing text detection, object recognition, and layout recommendation.

*e-mail: 61175021h@ntnu.edu.tw
†e-mail: iweilai@ntnu.edu.tw

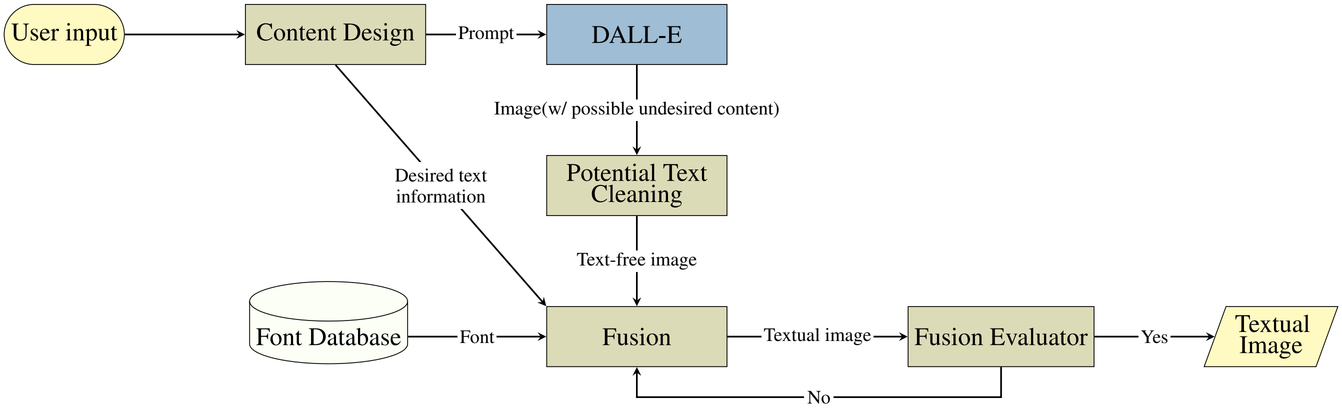

Figure 1: Architecture of MATIC illustrating the four core modules: Content Design, Potential Text Cleaning, Fusion, and Fusion Evaluator.

3 METHOD

Our research leverages the CoT concept to break down complex process into manageable stages. The MATIC system integrates a diffusion model, specifically DALL-E [12], along with four sequential modules: Content Design, Potential Text Cleaning, Fusion, and Fusion Evaluator (see Figure 1). Each module employs MMLLMs, particularly GPT-4o [13], to address distinct challenges within the text-image fusion workflow.

3.1 Content Design

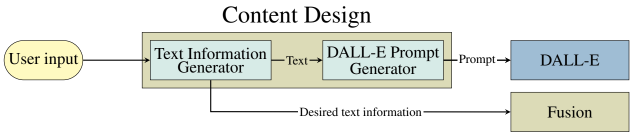

Figure 2: The Content Design stage analyzes user input to generate both the desired text and DALL-E prompt.

The Content Design module, depicted in Figure 2, comprises two core components: the Text Information Generator and the DALL-E Prompt Generator. Together, these components ensure the seamless integration of text and image elements while adhering to user specifications and linguistic requirements.

The Text Information Generator is designed to produce the required text while providing detailed information regarding its structure and the relationships within the text. This component handles line breaking and organizes text hierarchies to facilitate refined layout designs in subsequent stages. For multilingual adaptation, the generator follows established protocols: it outputs content in the specified language, infers linguistic context when explicit language instructions are not provided, and defaults to the language of the input in the absence of additional cues. By adhering to these processes, the system effectively supports the creation of multilingual textual images, addressing diverse linguistic and cultural requirements.

The DALL-E Prompt Generator is designed to produce detailed, text-free image generation instructions, addressing scenarios where implicit textual suggestions may arise, such as in holiday cards or signage. This component ensures that the generated visuals remain free of names, logos, brands, or other textual content, focusing instead on creating scenes with placeholder or symbolic thematic elements. This approach maintains unobstructed visual areas, ensuring clarity and readability for future text integration. Moreover, it prevents the unintended generation of embedded text, a common challenge in AI-driven image synthesis.

3.2 Potential Text Cleaning

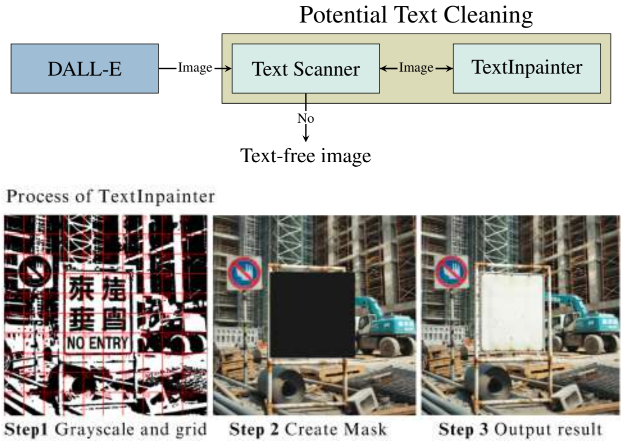

Figure 3: Potential Text Cleaning ensures images are devoid of any text.

In this module, the absence of text within the image is verified, as illustrated in Figure 3. The process begins with the Text Scanner, which determines whether text is present. If no text is detected, the workflow transitions to the subsequent module. However, if text is identified, the TextInpainter removes it. The TextInpainter utilizes inpainting techniques [11, 12], which typically depend on manually created masks to define areas for regeneration. Distinctively, this system automates the mask generation process.

To achieve this, a grid system is employed. The image is first converted to grayscale and overlaid with a grid consisting of 100 uniquely numbered squares. The system identifies squares containing text and directly uses their corresponding numbers to locate precise positions for mask creation. This division into smaller grid sections addresses a critical limitation: scanning the entire image simultaneously often results in errors for MMLLM. By processing the image square by square, the computational load is reduced, enhancing accuracy. Moreover, this grid-based approach surpasses Cartesian coordinate methods in efficiency, as it eliminates the complexities of axis-based positional alignment, which poses significant challenges for MMLLM systems.

After the TextInpainter completes text removal, the process returns to the TextScanner for iterative verification, ensuring that all text elements have been successfully removed. This cyclical verification guarantees a clean and text-free output.

3.3 Fusion

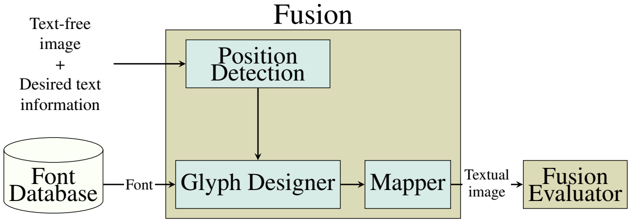

Figure 4: The Fuse module consists of three components collaboratively generate a cohesive visual and textual result.

The Fusion module consists of three principal components, as illustrated in Figure 4. The first component, the Position Detection module, processes the image by analyzing grid coordinates to identify low-detail areas and evaluating the relationship information derived from the Content Design output. This evaluation determines the most suitable placement for text. The module also accommodates language-specific layout arrangements, such as right-to-left formatting for Arabic, which differs from the left-to-right orientation of most other languages. This analysis ensures that the text layout adheres to principles of information hierarchy and visual balance while avoiding coverage of key visual content, ultimately determining the optimal text placement for enhanced clarity and aesthetic coherence.

Secondly, the Glyph Designer selects an appropriate font that aligns with the visual style of the image and the intended text application. The font dataset, developed as an open-source and commercially usable resource, supports multiple languages, including English, Arabic, Cyrillic, Japanese, Mandarin, and others. Fonts are cataloged based on language and font weight, organized into a font atlas for efficient selection. The Glyph Designer also determines the optimal text size, colors, and shadow effects to ensure visual harmony and readability. Finally, it generates the output parameters for all text attributes, enabling seamless integration into the final composition.

Finally, these parameters are send to the Mapper to accurately fuse text into the image within the designated boundaries. Upon completion, the textual image is forwarded to the fusion evaluator for assessment.

3.4 Fusion Evaluator

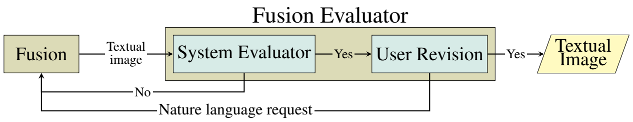

Figure 5: Fusion evaluator will go through system evaluation and User Revision before provide final images.

In the Fusion Evaluator, as illustrated in Figure 5, the System Evaluator first assesses whether the text obscures any critical visual content. It then evaluates the style, color, and consistency of the text. If inconsistencies are identified, the system reverts to the Fusion module for further adjustments. Beyond automated alterations, the system allows users to customize all aspects of the textual design embedded in the image. Customizable features include language, textual content, positioning, color, size, shadow effects, and transparency, providing users with the flexibility to refine the visual and stylistic presentation of the text to meet their specific preferences. However, if the user chooses to modify the image itself, the system automatically regenerates the image. This approach is necessary because changes to the image often require corresponding adjustments to the text to maintain overall stylistic cohesion and balance. Given the significant impact such alterations can have on the composition, automated regeneration ensures that the text and image remain visually integrated and harmonious.

4 RESULTS

This section provides a comprehensive analysis, encompassing both quantitative and qualitative aspects. The evaluation includes the accuracy of text generation across multiple languages and varying text lengths, a detailed assessment of stylistic attributes, an overview of results from user-driven revisions, and a comparison of visualization outcomes.

4.1 Text Accuracy Across Different Language

The EasyOCR system [9] was utilized to assess text recognition accuracy within images. Data for Diff-Text [21] were sourced directly from its original study, which excluded Japanese-language data.

As shown in Table 1, the proposed system, MATIC, outperforms other systems by achieving superior OCR accuracy across multiple languages. MATIC consistently maintains accuracy rates exceeding 95%, even for languages with sophisticated glyph structures such as Russian, Arabic, and Mandarin. A detailed manual review indicates that the remaining inaccuracies stem from inherent limitations of OCR technique. Although only five languages are showcased here, MATIC can reliably and accurately generate textual images for any language, provided the corresponding font file is available. These results highlight MATIC's robustness and suggest its significant potential for applications in education, information dissemination, and other communication media.

Table 1: *OCR Accuracy (%) Comparison Across Languages: Stable Diffusion (SD), DALL-E, TextDiffuser2 (TD2), Diff-Text, and MATIC*

Language	SD	DALL-E	TD2	Diff-Text	MATIC
English	9.5	33.21	52.11	61.03	96.83
Russian	0	0	1.38	39.29	95.45
Arabic	0	0	0	33.13	95.28
Mandarin	0	0	0	32.40	95.21
Japanese	0	0	0	N/A	95.89

4.2 Text Accuracy Across Different Text Length

Since textual image generation often involves varying text lengths, and image-generation models typically experience reduced text accuracy as text length increases, a comparison was conducted for text lengths ranging from 1 to 10 words. To minimize confounding factors unrelated to text length, GPT-generated prompts were restricted to simple terms to reduce the impact of text complexity on the results. Accuracy was measured by the number of images required to produce 10 correct outputs, with a maximum limit of 100 images per evaluation. Correctness was verified manually. The evaluated models included DALL-E, TextDiffuser2, and MATIC. Stable Diffusion was excluded due to its consistently low accuracy, and Diff-Text was omitted as it lacks custom text generation capabilities. Only English was analyzed, as other systems offer limited multilingual support. Results, depicted in Fig. 6, show that while DALL-E and TextDiffuser2 exhibit declining accuracy as text length increases, MATIC maintains stable performance even with longer text inputs. These findings underscore MATIC's versatility and reliability for diverse applications requiring precise and consistent text-to-image generation across varying text lengths and levels of complexity.

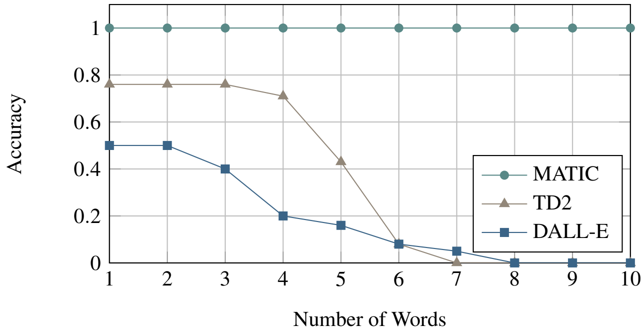

Figure 6: Text accuracy across different word lengths for DALL-E, TextDiffuser2(TD2), and MATIC

4.3 Style Analysis

The style evaluation compared the performance of DALL-E, TextDiffuser2, and MATIC. Standardized prompts was applied across all three systems to generated four images, which were compiled into a survey for evaluation. Due to the limitations of the other systems, which only support English and short text, the evaluation was restricted to English short-text images.

Participants rated each aspect on a scale from 1 to 5. Text style diversity measured the variety of fonts and formatting styles employed, while text quality assessed clarity, legibility, and accuracy. Layout aesthetics focused on the spatial arrangement of text and images. Ease of understanding evaluated the intuitive clarity of the generated content, and text-image consistency assessed the semantic alignment between textual and visual elements.

A total of 100 responses were collected, forming a robust dataset for comparative analysis. The results, depicted in Figure 7, indicate that while TextDiffuser2 performs relatively well in terms of text quality, it exhibits notable differences from the other two systems across the remaining evaluation criteria. The findings suggest that the incorporation of CoT techniques extends the strengths of large models such as DALL-E, enabling the generation of stylistically refined images. Furthermore, the system demonstrates slight advantages in information conveyance, text quality, and text-image coherence, reflecting its potential for delivering high-quality and contextually aligned outputs.

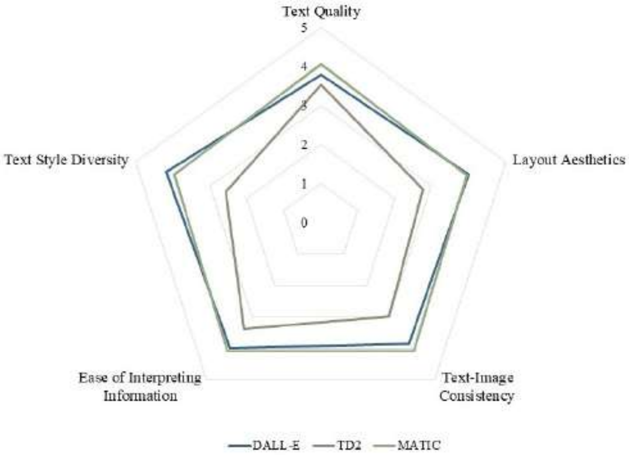

Figure 7: Score result across five criteria for DALL-E, TextDiffuser2(TD2), and MATIC-generated images.

4.4 User Revision Result

Following the System Evaluator process, users can refine generated images through natural language instructions, with the User Revision component playing a crucial role in interpreting these inputs and implementing the desired adjustments. Figure 8 illustrates this process, showcasing an originally generated movie poster used as a reference image alongside three variations created in response to distinct user-requested modifications.

The system's key advantage lies in its accessibility, enabling users to refine images with simple natural language inputs without requiring advanced technical or design skills. The MMLLM processes these instructions, translating them into actionable parameters to modify images intuitively and efficiently. This user-friendly approach simplifies image editing, offering high customization for diverse applications ranging from personal projects to professional design tasks, and highlights the transformative potential of multimodal language models in advancing intuitive and adaptable image editing technologies.

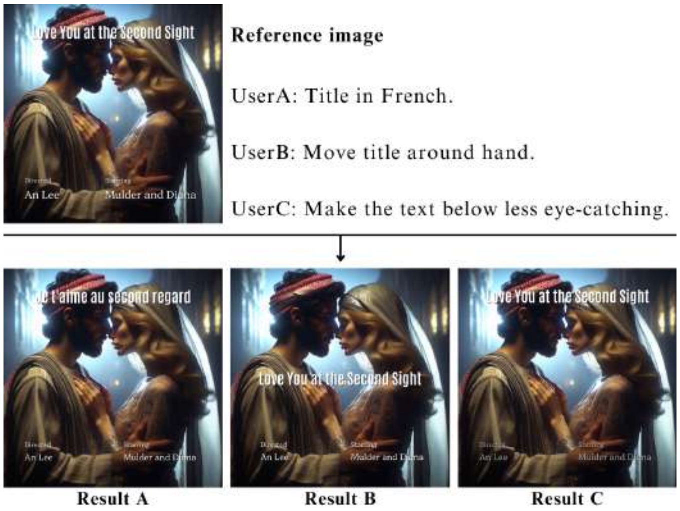
Love You at the Second Sight
An Lee
Mulder and Diana
Reference image
UserA: Title in French.
UserB: Move title around hand.
UserC: Make the text below less eye-catching.
Je t'aime au second regard
Love You at the Second Sight
Result A
Result B
Result C

Figure 8: Reference image and modified results base on various user request.

4.5 Visualization Outcomes

The visual results in Figure 9 illustrate outputs generated by three systems using identical prompts, highlighting their comparative performance. DALL-E excels in producing fine stylistic details, delivering visually appealing outputs. However, it struggles with text generation, often producing inaccurate or illegible text, which limits its utility for text-intensive tasks. Similarly, TextDiffuser2 demonstrates weaknesses in understanding prompts and frequently generates text with errors. Additionally, its image styles are limited, further restricting its versatility and effectiveness.

In contrast, MATIC vividly demonstrates its robust capabilities in handling multiple languages while effectively organizing text into clear hierarchical structures with aesthetically balanced visual layouts. This design enhances both the absorption and retention of information. Moreover, the innovative Content Design module enables the whole system to automatically generate contextually rich and relevant text from provided keywords, significantly simplifying the process of creating comprehensive materials. Notably, MATIC maintains exceptional text accuracy across varying lengths and font sizes, consistently delivering high-quality, readable content. This reliability is essential for addressing diverse user needs, making MATIC a powerful and versatile tool for generating visually and linguistically coherent outputs.

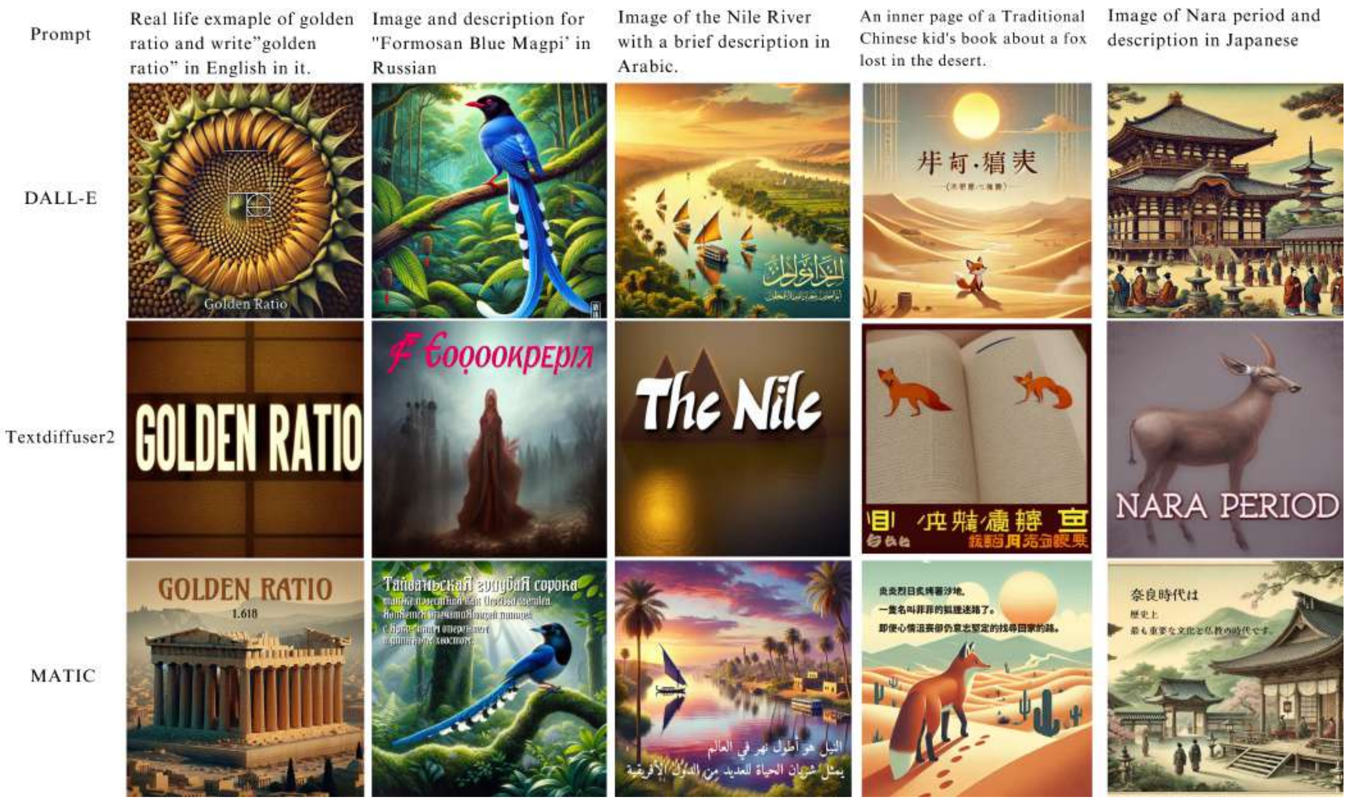
Prompt
Real life exmaple of golden ratio and write”golden ratio” in English in it.
Image and description for "Formosan Blue Magpi' in Russian
Image of the Nile River with a brief description in Arabic.
An inner page of a Traditional Chinese kid's book about a fox lost in the desert.
Image of Nara period and description in Japanese
DALL-E
Golden Ratio
Textdiffuser2
GOLDEN RATIO
The Nile
NARA PERIOD
MATIC
GOLDEN RATIO
1.618
Тайваньская голубая сорока
النيل هو أطول نهر في العالم
يمثل شريان الحياة للعديد من الدول الأفريقية
炎炎烈日炙烤著沙地。
一隻名叫菲菲的狐狸迷路了。
即使心情沮喪卻仍意志堅定的找尋回家的路。
奈良時代は
歴史上
最も重要な文化と仏教の時代です。

Figure 9: Comparison of Visualization Outcomes Generated by DALL-E, TextDiffuser2, and MATIC Across Subjects and Languages

5 CONCLUSION

This study demonstrates that the proposed MATIC system significantly enhances the quality of textual images by seamlessly integrating multilingual support and achieving coherent text-image synthesis. Employing a modular architecture that incorporates Chain-of-Thought reasoning within both the MMLLM and the diffusion model, MATIC enables targeted optimization of individual components, ensuring outstanding performance in generating precise and complex outputs. Furthermore, the use of a grid system to segment images allows the MMLLM to perform more precise analyses. This integration ensures that the generated content maintains both aesthetic appeal and linguistic accuracy, effectively addressing diverse application scenarios. Additionally, the incorporation of natural language-based customization offers unparalleled design flexibility, enabling seamless refinement and adaptation of content to meet specific requirements.

The system's distinctive capability to generate visually compelling and contextually relevant images not only addresses the requirements of contemporary professional and creative domains but also creates opportunities for innovative applications. These include the development of educational materials, interactive media, and multilingual digital storytelling, thereby broadening its utility across diverse contexts. Future advancements are anticipated to center on the integration of generative text functionalities, which will enhance outputs by providing deeper contextual richness and improved semantic coherence. Additionally, the introduction of a more intuitive, user-friendly interface will ensure greater accessibility for users with varying levels of technical proficiency. Such advancements are expected to further consolidate MATIC's position as a pioneering platform, amplifying its influence across a wide range of industries and use cases.

References

[1] G. Batzolis, J. Stanczuk, C.-B. Schönlieb, and C. Etmann. Conditional image generation with score-based diffusion models. *arXiv preprint arXiv:2111.13606*, 2021. 1

[2] J. Chen, Y. Huang, T. Lv, L. Cui, Q. Chen, and F. Wei. Textdiffuser: Diffusion models as text painters. In *Advances in Neural Information Processing Systems*, vol. 36, 2024. 1

[3] J. Chen, Y. Huang, T. Lv, L. Cui, Q. Chen, and F. Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. In *European Conference on Computer Vision*, pp. 386–402. Springer, 2025. 1

[4] O. Gafni, A. Polyak, O. Ashual, S. Sheynin, D. Parikh, and Y. Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pp. 89–106. Springer, 2022. 1

[5] Google Research. Gentype. `https://labs.google/gentype`, 2024. 1

[6] S. Gu, D. Chen, J. Bao, F. Wen, B. Zhang, D. Chen, L. Yuan, and B. Guo. Vector quantized diffusion model for text-to-image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR '22)*, pp. 10696–10706, 2022. 1

[7] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *Advances in neural information processing systems*, vol. 33, pp. 6840–6851, 2020. 1

[8] N. Inoue, K. Kikuchi, E. Simo-Serra, M. Otani, and K. Yamaguchi. Layoutdm: Discrete diffusion model for controllable layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10167–10176, 2023. 1

[9] JadedAI. Easyocr. `https://github.com/JaidedAI/EasyOCR`, 2020. 3

[10] J. Ma, M. Zhao, C. Chen, R. Wang, D. Niu, H. Lu, and X. Lin. Glyphdraw: Seamlessly rendering text with intricate spatial structures in text-to-image generation. *arXiv preprint arXiv:2303.17870*, 2023. 1

[11] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew,

I. Sutskever, and M. Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[12] OpenAI. Dalle-3. https://openai.com/index/dall-e-3/, 2024. 2

[13] OpenAI. Gpt-4o. https://openai.com/index/hello-gpt-4o/, 2024. 2

[14] S. Paliwal, A. Jain, M. Sharma, V. Jamwal, and L. Vig. Customtext: Customized textual image generation using diffusion models. *arXiv preprint arXiv:2405.12531*, 2024. 1

[15] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 1

[16] C. Saharia, W. Chan, H. Chang, C. Lee, J. Ho, T. Salimans, D. Fleet, and M. Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022. 1

[17] A. Shenoy, Y. Lu, S. Jayakumar, D. Chatterjee, M. Moslehpour, P. Chuang, A. Harpale, V. Bhardwaj, D. Xu, S. Zhao, et al. Lumos: Empowering multimodal llms with scene text recognition. *arXiv preprint arXiv:2402.08017*, 2024. 1

[18] Y. Tatsukawa, I.-C. Shen, A. Qi, Y. Koyama, T. Igarashi, and A. Shamir. Fontclip: A semantic typography visual-language model for multilingual font applications. In *Computer Graphics Forum*, p. e15043. Wiley Online Library, 2024. 1

[19] D. Wang, N. Raman, M. Sibue, Z. Ma, P. Babkin, S. Kaur, Y. Pei, A. Nourbakhsh, and X. Liu. Docllm: A layout-aware generative language model for multimodal document understanding. *arXiv preprint arXiv:2401.00908*, 2023. 1

[20] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, D. Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, vol. 35, pp. 24824–24837, 2022. 1

[21] L. Zhang, X. Chen, Y. Wang, Y. Lu, and Y. Qiao. Brush your text: Synthesize any scene text on images via diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, pp. 7215–7223, 2024. 1, 3