MetaDesigner: Advancing Artistic Typography through AI-Driven, User-Centric, and Multilingual WordArt Synthesis

Kaibo Wang

Peddie School, Hightstown NJ 08520, USA

Abstract. MetaDesigner introduces a groundbreaking approach to the synthesis of artistic typography by utilizing Large Language Models (LLMs). This system, designed to enhance user interaction, is built upon a multi-agent framework comprising the Pipeline, Glyph, and Texture agents. These agents work in tandem to offer personalized WordArt creations, from semantic enrichment to the application of intricate textures. The system integrates a detailed feedback loop that leverages insights from multimodal models and user feedback to continually refine the design process. By adjusting hyperparameters through this feedback loop, MetaDesigner precisely aligns with users' stylistic and thematic preferences, producing WordArt that surpasses expectations in both visual appeal and relevance to context. Empirical studies of MetaDesigner affirm its effectiveness in catering to a wide range of WordArt needs, delivering aesthetically pleasing and contextually appropriate outcomes. MetaDesigner thus marks a significant advancement in digital art creation, merging generative AI technologies with the nuanced requirements of artistic typography to open new avenues in WordArt synthesis.

1 Introduction.

Typography merges the expressiveness of language with the visual allure of design, highlighting its importance in fields like advertising [10-12,39], early childhood education [43], and historical tourism [1]. More than just displaying text, typography is a crucial channel for communication, artistic expression, and innovation. Yet, mastering typography is challenging, especially for amateurs, due to its dual demand for aesthetic intuition and understanding of complex design principles.

The emergence of generative models marks a transformative phase in typography design, promising to make this art form more accessible by accommodating varied aesthetic tastes. However, applying these models to meet the nuanced demands of typography design presents unique challenges:

Subjective Aesthetic Judgment. The appreciation of artistic typography is subjective, influenced by individual and cultural preferences, complicating the task of creating generative models that universally resonate with all user preferences.

1. *Dataset Shortcomings.* Artistic typography lacks extensive, well-annotated datasets, limiting generative models' ability to mimic and generate a broad spectrum of typographic styles.

Fig. 1. Overview of MetaDesigner, showcasing the interplay between the Pipeline, Glyph, and Texture agents

Current WordArt solutions, despite using advanced Latent Diffusion Models (LDMs), often fall short in versatility and performance due to their model limitations. These systems offer limited WordArt styles, failing to meet the diverse and changing needs of users.

MetaDesigner distinguishes itself by tackling the intricacies of typography design through a multi-agent system, aiming to produce artistic typography that resonates with a broad user base and enriches digital art. Employing LoRA techniques [20] and models from the civitai.com community, we strive to diversify WordArt synthesis, making AI-driven design accessible to non-professionals. The existence of Large Language Models (LLMs) introduces intelligent agents capable of advanced reasoning and adaptability [44, 46, 49], opening new paths for creative inspiration and design logic. This technological advancement raises the question: How can we leverage AI innovations to make WordArt creation more democratic, enabling easy creation of complex and varied designs? MetaDesigner addresses this by integrating LLMs with multi-agent systems, facilitating WordArt driven by user input. This blend of AI's technical capabilities with user creativity aims to meet the complex demands of a global audience. MetaDesigner focuses on four main aspects:

1. *User-Centric Design and Versatility*. Featuring a sophisticated multi-agent system with evaluation and optimization modules, MetaDesigner supports discovering and creating personalized and varied WordArt styles. It fine-tunes hyperparameters to ensure outputs reflect individual preferences.
2. *Expanding Artistic Expression*. By deploying glyph design agents and leveraging advanced font libraries and semantic transformation techniques, the system allows for a wide range of glyph transformations, enriching the artistic value of the generated WordArt.
3. *Feedback-Driven Refinement.* MetaDesigner employs a collaborative strategy, merging multi-modal model prompts with direct user feedback to meticulously enhance the quality of WordArt. This detailed process breaks down evaluation metrics into actionable insights, enabling targeted, iterative improvements to the design [35].
4. *Community Engagement and Resources*. Developing a LoRA model library and a comprehensive WordArt dataset demonstrates our commitment to supporting the artistic typography research community. These essential resources act as both a source of inspiration and a practical guide for inspiring WordArt creations.

In summary, MetaDesigner stands as a powerful automated platform for WordArt creation, deliberately tailored to meet the detailed preferences of its users. This research sets a foundation for future text synthesis advancements and heralds practical applications across various sectors. The upcoming release of the WordArt Dataset is expected to greatly improve the sophistication and accessibility of design methodologies for both general and professional use.

2 Related Work

Text-to-Image Synthesis. Recent advancements in text-to-image synthesis have been primarily driven by the development of denoising diffusion probabilistic models [5, 14, 19, 30, 32, 33, 35, 36, 38]. These models have evolved from generating basic images to facilitating interactive image editing [4, 16, 28] and adopting conditions like masks and depth maps for more detailed creations [35]. Additionally, the field is exploring multi-condition controllable synthesis [21, 29, 51] to enhance versatility. Techniques for integrating subjects into scenes have seen innovations through ELITE [45], UMM-Diffusion [27], and InstantBooth [37], which leverage CLIP image encoder capabilities for converting visual concepts into textual word embeddings. DreamIdentity [9] further refines this approach with a custom image encoder designed to improve word embedding enhancement schemes.

Visual Text Generation. Despite significant progress in image synthesis, integrating legible text within images remains a challenge [35, 36]. Current research on visual text generation focuses on:

Control Condition. Modern strategies often incorporate glyph conditions in latent space for text control. GlyphDraw [26] uses an explicit glyph image condition to render characters only in the center. GlyphControl [47] advances text alignment with attributes related to location, implicitly adjusting for font size and text box position. TextDiffuser [8] introduces character-level segmentation masks as a control condition, aiding text generation and in-painting by using a masked image.

Text Encoder. The accuracy of visual text generation greatly depends on the text encoder. While models like Imagen [36], eDiff-I [2], and Deepfloyd IF [23] show promising results with large-scale language models (e.g., T5-XXL), the challenge of accurately encoding non-Latin scripts (Chinese, Japanese, Korean, etc.) persists [25]. GlyphDraw specifically addresses Chinese character rendering by optimizing the text encoder with Chinese images and utilizing the CLIP image encoder for glyph embeddings [26]. DiffUTE innovates by replacing the text encoder with a pre-trained image encoder for glyph extraction in editing scenarios [6]. OCR-VQGAN employs a pre-trained OCR model to extract features for text generation, applying constraints across layers to guide the process [34]. TextDiffuser uses a character-level segmentation model to maintain character integrity in latent space, although it requires an additional model and restricts character class diversity [34].

WordArt Synthesis. Synthesizing WordArt [3, 22, 40, 41, 50] merges semantics with the need for artistic and legible text representation, posing a unique challenge. Initial approaches, such as the work by Tendulkar et al. [41], innovated by mapping letters to semantically similar icons in a joint embedding space. Typeface by Zhang et al. [50] took a semi-manual approach, segmenting letters to fit semantically relevant shapes, then refining the output by removing extra visual disturbances. The introduction of large generative models has significantly propelled the field forward. Word-As-Image by Iluz et al. [22] experimented with artistic typography for the Latin language, while DS-Fusion by Tanveer et al. [40] explored more complex text forms, including hieroglyphics, by applying additional constraints. Despite these advances, fully catering to the diverse preferences of users and the creation of innovative WordArt styles is an ongoing journey of exploration and development.

3 Methods

Overview. The MetaDesigner framework introduces an innovative, interactive multi-agent system aimed at creating WordArt that reflects user preferences. This system incorporates three distinct intelligent agents: the pipeline, glyph, and texture agents. Each agent contributes uniquely to the crafting of personalized WordArt. Specifically, the WordArt images generated by MetaDesigner, symbolized by Ψ, are mathematically formulated as follows:

$$\hat{I} = \Psi(s^{user}, \phi, \mathcal{P}, \mathcal{M})$$

Where s^{user} signifies a user's prompt, encapsulating their preferences and input. The term ϕ represents the collective functionality of the involved agents within MetaDesigner, specifically $\phi = \{\phi^{pip}, \phi^{gly}, \phi^{tex}\}$, corresponding to the pipeline, glyph, and texture agents, respectively. $\mathcal{M}$ is the library of models utilized by the teture agent, while $\mathcal{P}$ denotes a set of learnable hyperparameters designed to finely tune the system's outputs to closely match user preferences through interactive and context-aware learning. $P = \{P^{pip}, P^{gly}, P^{tex}\}$, with $P^{pip}, P^{gly}, and\ P^{tex}$ specifically allcated to the pipeline, glyph, and texture agents.

3.1 Agents of MetaDesigner

Pipeline Designer. The Pipeline Designer is a cornerstone of the MetaDesigner framework, skillfully converting visual tasks into a structured coding approach. This transformation is key to orchestrating the synthesis agents to work in tandem seamlessly. By generating the WordArt synthesis workflow, the Pipeline Designer plays a pivotal role in extending user prompts using the capabilities of the Large Language Model (LLM) and incorporating feedback to refine the synthesis process.

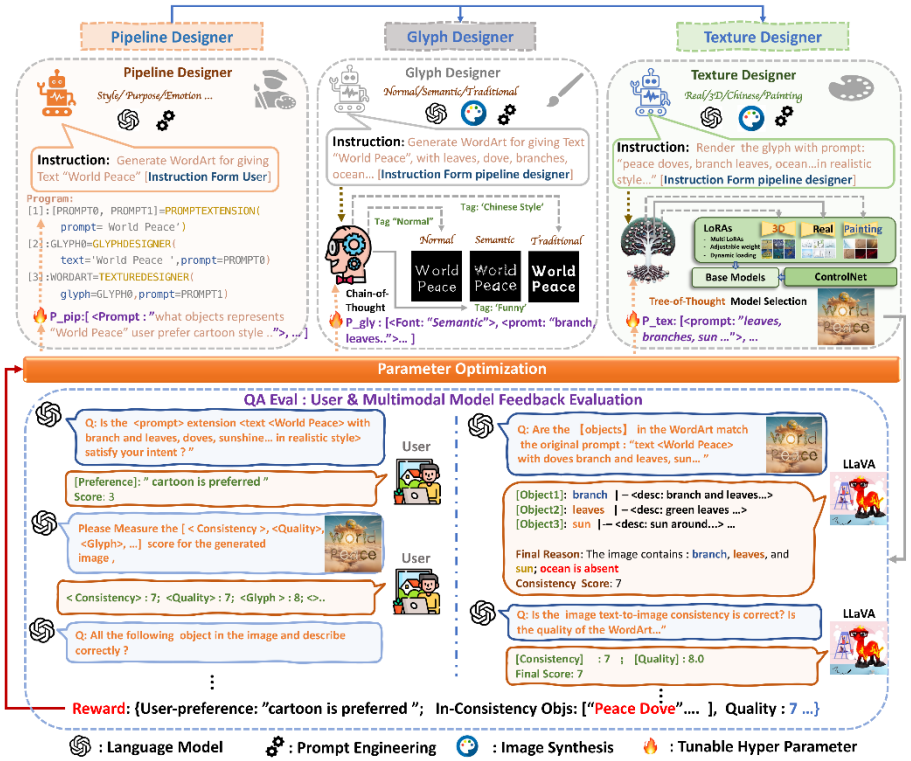

Fig. 2. Architectural Overview of MetaDesigner: This schematic illustrates the MetaDesigner system's structured integration of three core intelligent agents—pipeline, glyph, and texture—to facilitate the creation of personalized WordArt. Additionally, it features optimization and Q&A evaluation modules that synergistically enhance the quality of the generated WordArt.

- Visual Programming: Inspired by the principles of visual programming [17] and as demonstrated in Fig. 2, the Pipeline Designer leverages GPT-3.5-Instruct for generating visual programs based on pairs of instructions and the desired high-level outcome. Remarkably, this process does not necessitate finetuning GPT-3 for visual programming tasks. Utilizing the in-context learning prowess of GPT-3.5-Instruct, the designer can interpret natural language instructions to produce visual programs. These programs, akin to the in-context examples provided, are crafted manually and can typically be developed without needing a corresponding image. Each program step articulates a module's name, its input arguments and their values, and an output variable name, creating a clear and executable visual code sequence.
- Prompt Extension: To ensure the MetaDesigner system caters to a wide range of user preferences—spanning image style, application domain, and user backgrounds—the Pipeline Designer employs an innovative approach to prompt extension. By formulating a series of probing questions, the system engages the GPT-4 model in a Chain-of-Thought (CoT) reasoning process. This methodical questioning is designed to enhance the system's understanding and responsiveness to user prompts by guiding the

model through a step-by-step analytical process. This approach enables the decomposition of complex tasks into smaller, more manageable units, enriching the prompt details and improving the system's accessibility and utility for end-users. Through this refined interaction, the Pipeline Designer ensures that the synthesized WordArt not only aligns with user preferences but also benefits from a depth of understanding and specificity previously unattainable, thereby advancing the capabilities of the MetaDesigner framework in generating highly personalized and contextually relevant WordArt.

Glyph Designer. The Glyph Designer is a critical component of the MetaDesigner framework, dedicated to the specialized task of glyph rendering, which forms the foundation for WordArt generation. This module is adept at rendering three distinct types of glyphs—Normal, Traditional, and Semantic—each serving different purposes and catering to various application domains and aesthetic requirements.

- **Normal & Traditional Glyph.** For environments where formality and clarity are paramount, Normal and Traditional glyphs are the go-to choices. These glyph types find their applications in settings that demand a touch of elegance and precision, such as weddings, galas, and award ceremonies. Here, WordArt employing Traditional glyphs plays a significant role in enhancing invitations and announcements, imbuing them with a sense of sophistication and grace. The initial selection of the glyph shape is a crucial step that directly influences the texture rendering phase that follows. In scenarios where creativity is not the primary focus, opting for Normal and Traditional glyph styles ensures the synthesis of WordArt that resonates with the context's formality. To facilitate the rendering of these glyphs, the MetaDesigner leverages the FreeType font library [13], utilizing its comprehensive collection of text fonts to produce a variety of glyph styles. This versatility allows for a tailored approach to WordArt creation, ensuring the final product aligns with the intended aesthetic and thematic direction.
- **Semantic Glyph Transformation.** Venturing into more creative territories, the Semantic glyph transformation unlocks new dimensions of artistic expression within WordArt typography. This transformation is designed for contexts filled with levity and humor, such as comic books and animated content, where the visual representation plays a crucial role in conveying the narrative's tone. Drawing inspiration from pioneering work [18], the Glyph Designer employs a differential rasterization scheme to finely tune a vector graph (glyph) to closely resemble target objects. This process is complemented by the utilization of a depth-to-image SD model, which stylizes the semantic glyph in accordance with the provided prompt.

Texture Designer. The Texture Designer module represents an evolution in word art generation, moving away from reliance on a single model towards a multifaceted approach that combines controlled image synthesis with the innovative Tree-of-Thought (ToT) framework. This integration not only meets a wide range of user preferences but also enhances the creative and customization possibilities within the texture design process. The operation of this module is encapsulated in the following equation:

$$I_{tex} = \psi^{tex}\left(I_{gly}, s^{tex}, \mathcal{F}, \mathcal{M}\right)$$
$$= TexR\left(I_{gly}, s^{tex}, C, ToTSel(s^{tex}, \mathcal{F}, \mathcal{M})\right)$$

Here, $TexR$ symbolizes the controllable synthesis mechanism, significantly improved by $ToTSel$ for strategic model selection. This method with F, pretrained language model, alongside I_{glyph}, s^{tex}, and C, orchestrates the transformation of glyph images into textural artworks. This process is carefully guided by user input and predefined control conditions to achieve a tailored and high-quality result.

To better understand the Texture Designer's operation, we examine its three primary components that demonstrates the functionality:

- **Controllable Synthesis:** Incorporating ControlNet [51] within the texture design framework enhances the versatility and variety of texture styles at the designer's disposal. This component permits the adjustment of different parameters, such as Canny Edge, Depth, Scribble, and original font images, enabling the creation of unique and visually appealing textured fonts. The synthesis equation,

$$I_{tex} = TexR\left(I_{gly}, s^{tex}, C, \mathcal{W}\right),$$

  interacts dynamically with the texture s^{tex} and control parameters C, utilizing selected model weights W to realize the envisioned aesthetic effects.

- **Model Selection and Tree-of-Thought Framework:** The core of the Texture Designer is the Tree-of-Thought (ToT) approach, which methodically navigates through various reasoning paths to uncover solutions. This strategy is crucial for maintaining the originality and artistic quality of the textures produced, involving steps such as:
  - *Thought Decomposition and Generation*, these processes aid in deconstructing the initial prompt and generating a subsequent range of conceptual directions.
  - *State Evaluation and Model Search*, this phase assesses the viability of each conceptual direction, selecting the most suitable model using techniques like Depth-first search (DFS) [48].

  These phases guarantee a thorough evaluation and selection process, closely aligning the texture designs with user preferences and project goals.
- **Model Library Integration:** Incorporating a comprehensive library of 68 LoRA models introduces a systematic approach to model selection. This library, spanning categories from "General" to "Cartoon," enables precise model choices, ensuring the Texture Designer can faithfully translate the user’s creative intent into the corresponding textural output.

Fig. 3. Illustration of the feedback loop for hyperparameter tuning, showcasing the integration of user preference, glyph quality, text-to-image consistency, and image quality evaluations.

Algorithm 1 Hyper-Parameter Tuning

Input: prompt s^{user}, initial hyperparameter $\mathcal{P}$, max iteration threshold τ, score threshold θ model lib $\mathcal{M}$, and MetaDesigner Ψ;
Output: WordArt image $\hat{I}$;
while $i < \tau$ and $\mathcal{L} < \theta$ **do**
  $\hat{I} = \Psi(s^{user}, \phi, \mathcal{P}, \mathcal{M})$; ▷ Eq. (??)
  $G_m = \mathcal{H}(S^{eval}, \hat{I})$
  $G_u = \{g_u^{cos}, g_u^{qua}, g_u^{tex}, g_u^{pref}, \mathcal{L}_u\}$ ▷ User Feedback
  $G = Merge(G_m, G_u)$
  $\mathcal{P} = \mathcal{F}(G|s^{update})$
  $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_u$; $i = i + 1$
end while;
return $\hat{I}, \mathcal{P} = \{\mathcal{P}^{pip}, \mathcal{P}^{gly}, \mathcal{P}^{tex}\}$;

3.2 Evaluation & Optimization

Despite applying Chain-of-Thought and Tree-of-Thought methodologies to select the optimal parameters or models for WordArt synthesis, challenges such as user aversion and text-to-image inconsistency persist, due to biases inherent in the Language Model (LM) and the stable diffusion model. To address this, we engage a multi-modal model to provide feedback to the LM, enhancing image synthesis accuracy through in-context learning.

Q and A Evaluation Establishing a robust and effective feedback loop for hyperparameter tuning is vital for the MetaDesigner system. Our feedback collection and integration strategy encompass four principal aspects: text-to-image consistency, image quality, glyph feedback, and user preference. This comprehensive evaluation method ensures a holistic review, combining objective assessments from multi-modal models with subjective insights from user studies. The procedure for updating the algorithm based on this evaluation is detailed as Algorithm 1.

– **Multi-Modal Model Evaluation.** Our system utilizes multi-modal models, specifically LLaVA [24], for objective evaluations of text-to-image consistency and image quality. This evaluation process involves:

1. Crafting evaluation prompts for each synthesized image that reflect specific features or concepts depicted.

2. Analyzing these prompts using the LLaVA model to obtain feedback metrics: $G_m = \{g_m^{cos}, g_m^{qua}, L_m\}$, where g_m^{cos} and g_m^{qua} denote the scores for consistency and quality, respectively.
3. Compiling the feedback into a coherent score and rationale with an LLM, which aids in making informed decisions for hyperparameter adjustments.

— **User Evaluation.** In parallel to objective evaluations, MetaDesigner solicits direct user feedback. This subjective analysis revolves around questions posed by an LLM, designed to gauge user preferences g_{pref}^u and perceptions of glyph style g_{gly}^u:

$$G_u = \{g_{cos}^u, g_{qua}^u, g_{gly}^u, g_{pref}^u, L_u\}$$

Though participation is optional, user insights are crucial for ensuring the system's outputs match user expectations.

— **Feedback Fusion.** Feedback integration leverages a strategic voting method, with a preference for user insights to guarantee that the generated WordArt closely resonates with human aesthetics and preferences. The combined feedback, $G = \text{Merge}(G_m, G_u)$, merges evaluations from multi-modal models and user feedback, providing a solid basis for hyperparameter optimization.

Hyperparameter Optimization. The process of optimizing hyperparameters within the glyph and texture design components is critical for enhancing the quality and ensuring the generated WordArt accurately reflects user preferences. This meticulous procedure leverages feedback from both qualitative assessments and model performance evaluations, employing a strategic approach to refine WordArt synthesis.

— Objective Function Formulation: At the heart of our optimization strategy lies the objective function, represented as:

$$\mathcal{L} = \mathcal{L}_m + \mathcal{L}_u$$

This function amalgamates feedback from both the LLaVA model $\mathcal{L}_m$ and direct user input $\mathcal{L}_u$, aiming to adjust the hyperparameters $\mathcal{P}_{gly}$ and $\mathcal{P}_{tex}$ of the glyph and texture designers, respectively. The role of the LLaVA model's feedback is to guide the optimization process towards hyperparameters that elevate the WordArt's quality and relevance:

$$\mathcal{L}_m = argmax\ \{\mathcal{P}^{gly}, \mathcal{P}^{tex}\} \sim \mathcal{H}\left(s^{eval}, \hat{I}\right)$$

Here, $\mathcal{H}$ signifies the heuristic analysis from the LLaVA model, applied to the evaluative prompt s_{eval}, steering the optimization towards enhancing the WordArt's accuracy and aesthetic appeal.

– **Adaptive Feedback Integration and Tuning.** The set of hyperparameters $\mathcal{P}$, encompassing those for pipeline, glyph, and texture design, is dynamically refined based on a synthesis of feedback.
  - Preferences and textual feedback from users primarily influence the glyph design $\mathcal{P}^{glyph}$ and overall pipeline strategy.
  - Objective assessments of text-image coherence and visual quality shape the texture design parameters $\mathcal{P}_{tex}$.

  Through an iterative feedback loop, indicated as $\hat{\mathcal{P}} = \mathcal{F}(G|s_{update})$, this tuning method ensures the WordArt synthesis system adeptly aligns with both quantitative metrics and qualitative user expectations, promoting perpetual enhancement and personalization.

4 Experiment

4.1 Comparison with State-of-the-Art Methods

To assess the performance of our proposed MetaDesigner framework, we conducted comparative experiments against the latest state-of-the-art (SOTA) techniques. We benchmarked MetaDesigner alongside five methods and one online API: Stable Diffusion XL (SD-XL) [31], TextDiffuser [15], TextDiffuser-2 [7], Anytext [42], and DALL-E3. These selections represent a cross-section of current capabilities in the field. The findings are illustrated in Figure 4, focusing on three key aspects:

WordArt Synthesis Success. Figure 4 reveals varied capabilities among the tested methods. SD-XL exhibits limited capacity for accurate text rendering, particularly in English, lacking comprehensive support. The TextDiffuser series, while competent in English, struggles with languages like Chinese, Korean, and Japanese. Anytext outperforms in diversity but faces challenges with Korean and Japanese. DALL-E3's capabilities are confined to English. MetaDesigner demonstrates superior adaptability and accuracy across languages.

Quality and Diversity. The TextDiffuser series, constrained by the SD 1.5 model and limited data, shows poor diversity and quality. Anytext offers more variety but still falls short in quality. DALL-E3, with its cinematic and 3D style renderings, presents higher quality but limited diversity. MetaDesigner stands out, showcasing a broad spectrum of styles including realistic, cartoon, and 3D, evidencing its unparalleled versatility.

Creativity and Relevance. The benchmark for creativity was set with the prompt: "Create a stylish word ‘World Peace’ representing its meaning." SD-XL yielded stylistic but contextually irrelevant outputs. The TextDiffuser series added thematic elements like leaves and peace doves, showing improvement. Anytext, though innovative, failed to fully capture the essence of "World Peace." DALL-E3, benefitting from ChatGPT-4's

Fig. 4. WordArt Synthesis Comparison: The first two columns display results for "World Peace" in English. Columns three and four showcase the Chinese rendition of "World Peace". The final two columns present "World Peace" in Korean and Japanese, respectively. The first column displays WordArt generated using the basic prompt "Create a stylish word 'World Peace' representing its meaning." The rest of the comparison methods are generated with more keywords: "Sun, Peace Dove, leaves, cloud".

integration, produced a 3D cinematic portrayal that somewhat aligned with the thematic intent but lacked clarity. MetaDesigner, in contrast, excelled by capturing the essence of "World Peace" with cohesive and thematic elements, demonstrating its nuanced understanding and creativity.

Quantitative Analysis. A user study evaluated "Text Accuracy" and "Aesthetics & Creativity" across 20 diverse words, involving 11 participants. Our MetaDesigner outperformed SOTA techniques in both dimensions, confirming its excellence in text readability and aesthetic appeal, as summarized in Table 1.

Letter-Level Comparison. To rigorously evaluate MetaDesigner's efficacy, we conducted experiments to compare its performance against current state-of-the-art (SOTA) methods. For this comparative analysis, we selected three notable methods alongside

Table 1: User study. We present the results from user study conducted in two dimensions, “Text Accuracy” and “Aesthetics & Creativity”. All values are in percent and higher is better.

Evaluation Dimension	SDXL	TextDiffuser	TextDiffuser-2	Anytext	DALLE 3	Ours
Text Accuracy	7.1	45.0	37.5	76.7	37.9	**93.8**
Aesthetics & Creativity	2.3	1.4	0.9	2.3	19.5	**73.6**

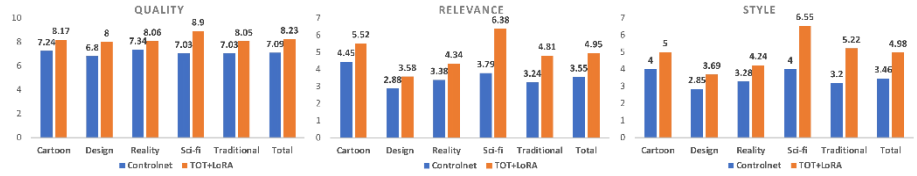

Fig. 5. The evaluation scores of the synthesis WordArt. From left to right are the “Quality", “Relevance", and “Style" scores generated by the ChatGPT-4 in the subcategories.

Fig. 6. The Comparison of WordArt texture rendering on the glyph “World Peace" (ControlNet vs. ToT-LoRA+ControlNet).

one widely used online API, specifically Google search, Stable Diffusion [35], DALL-E3, and DS-Fusion [40]. The outcomes of this comparison are illustrated in Table 1. Stable Diffusion [35] encounters difficulties in fully realizing all potential WordArt creations, indicating a challenge in covering the extensive variety of WordArt styles comprehensively. DS-Fusion [40] excels among the evaluated methods, particularly in terms of shape deformation and the preservation of letter legibility. However, despite its strengths, DS-Fusion's repertoire of WordArt styles is somewhat limited, suggesting a need for a broader stylistic range. DALL-E3 is notable for its high-quality renderings, with a strong emphasis on textural representations that resonate with the underlying semantic meanings. While impressive, its focus remains somewhat narrow, primarily concentrating on the textural aspects tied to semantics. In stark contrast, MetaDesigner distinguishes itself by producing WordArt that not only showcases a broader spectrum of stylistic creativity but also upholds an exceptional level of detail quality. This approach allows MetaDesigner to transcend the limitations observed in other methods, offering both versatility and precision in its WordArt creations.

Fig. 7. The examples are the optimization of the text-to-image consistency.

Fig. 8. Examples from the WordArt dataset.

4.2 Effect of Tree-of-thought

Quantitative Analysis. Evaluating the ToT-LoRA scheme against ControlNet revealed a 39% improvement in "Relevance," showcasing ToT-LoRA's effective domain adaptation and overall enhancement in "Quality," "relevance, " and "Style," as seen in Figure 5. The application of the ToT scheme has shown to yield significant improvements across various performance metrics, such as relevance, quality, and style. By leveraging this approach, MetaDesigner has demonstrated a marked advantage over both traditional methods and other state-of-the-art techniques. Specifically, the ToT scheme enhances the system's ability to generate WordArt that is more aligned with user expectations and preferences, as evidenced by improved scores in user studies and objective evaluations.

4.3 Effect of Optimization

The optimization techniques implemented within the MetaDesigner framework refines the generation of WordArt, ensuring that the final outputs adhere closely to user prompts and incorporate a deeper level of creativity and accuracy. Through a comprehensive case study, whose outcomes are visualized in Figure 7, we delve into the nuanced improvements brought about by these optimization strategies.

Identification and Inclusion of Omitted Elements. Utilizing the LLaVA (Language, Vision, and Action) system, the framework dynamically identifies elements that were mentioned in the user's prompt but inadvertently omitted in the initial WordArt generation. This detection prompts immediate updates to the generation process, ensuring that all essential components are included, thereby aligning the WordArt more closely with the user's expectations and the prompt's context.

Symbol Enhancement. This technique focuses on improving the visual clarity and appeal of symbols within WordArt. By enhancing symbol representation, the

MetaDesigner ensures that each symbol not only contributes to the overall aesthetic of the piece but also supports the intended message or theme.

Word Sequencing Adjustments Adjusting the sequence in which words and phrases appear within WordArt can dramatically affect its readability and thematic delivery. This optimization process reorders elements to maximize coherence and narrative flow, ensuring that viewers can easily understand and appreciate the artwork's message.

Keyword Repetition. Strategically repeating keywords within WordArt serves to emphasize central themes and concepts. This technique not only reinforces the message but also creates a visual rhythm and focus, making WordArt more memorable and impactful.

4.4 MetaDesigner Dataset

To contribute to the continuous growth and innovation within the WordArt community, our team has developed the Creative and Diverse WordArt Dataset, leveraging the advanced capabilities of the MetaDesigner framework. Detailed in Figure 8, this dataset is a rich collection of 5000 multilingual images, showcasing a wide array of artistic expressions across languages including English, Chinese, Japanese, and Korean:
Multilingual Diversity: The inclusion of multiple languages in the dataset not only celebrates linguistic diversity but also provides a broad spectrum of artistic styles and cultural nuances, enabling more inclusive and wide-ranging research and development in the field of WordArt.

5 Conclusion

This research presents MetaDesigner, an advanced framework engineered to create artistic typography by harnessing the capabilities of Large Language Models (LLMs) such as GPT-4. Central to MetaDesigner are three fundamental agents: the Pipeline, Glyph, and Texture Agents. These agents work in synergy to craft WordArt that not only adheres to but also enhances user preferences, showcasing the framework's ability to tailor outputs to individual design requirements. MetaDesigner stands at the intersection of generative artificial intelligence and artistic typography, offering a novel approach that leverages the best of both worlds. Through its multi-agent system, MetaDesigner facilitates the generation of WordArt across a spectrum of applications, from promotional posters to bespoke jewelry designs, demonstrating its versatility and effectiveness.

Empirical validations of the framework have underscored its efficacy and innovation in the field. These studies highlight MetaDesigner's capacity to produce WordArt that resonates with a broad audience, catering to diverse design needs while maintaining a high standard of aesthetic appeal and relevance. By integrating cutting-edge AI technology with the nuanced demands of artistic typography, MetaDesigner not only

contributes to the advancement of digital art creation but also opens new avenues for exploration and application in various domains. In summary, MetaDesigner exemplifies the potential of combining generative AI with artistic creativity to meet and exceed the evolving expectations of digital art synthesis. Its success in empirical studies across different domains reaffirms the framework's value as a pivotal tool for artists, designers, and anyone looking to explore the vast possibilities of WordArt generation.

References

1. Amar, J., Droulers, O., Legohérel, P.: Typography in destination advertising: An exploratory study and research perspectives. Tourism Management 63, 77–86 (2017). https://doi.org/https://doi.org/10.1016/j.tourman.2017.06.002, https://www.sciencedirect.com/science/article/pii/S0261517717301243
2. Balaji, Y., Nah, S., Huang, X., Vahdat, A., Song, J., Kreis, K., Aittala, M., Aila, T., Laine, S., Catanzaro, B., Karras, T., Liu, M.: ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint abs/2211.01324 (2022)
3. Berio, D., Leymarie, F.F., Asente, P., Echevarria, J.: Strokestyles: Stroke-based segmentation and stylization of fonts. ACM Trans. Graph. 41(3), 28:1–28:21 (2022)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. arXiv preprint abs/2211.09800 (2022)
5. Chang, H., Zhang, H., Barber, J., Maschinot, A., Lezama, J., Jiang, L., Yang, M., Murphy, K., Freeman, W.T., Rubinstein, M., Li, Y., Krishnan, D.: Muse: Text-to-image generation via masked generative transformers. arXiv preprint abs/2301.00704 (2023)
6. Chen, H., Xu, Z., Gu, Z., Lan, J., Zheng, X., Li, Y., Meng, C., Zhu, H., Wang, W.: Diffute: Universal text editing diffusion model. arXiv preprint abs/2305.10825 (2023)
7. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser-2: Unleashing the power of language models for text rendering. CoRR abs/2311.16465 (2023)
8. Chen, J., Huang, Y., Lv, T., Cui, L., Chen, Q., Wei, F.: Textdiffuser: Diffusion models as text painters. arXiv preprint abs/2305.10855 (2023)
9. Chen, Z., Fang, S., Liu, W., He, Q., Huang, M., Zhang, Y., Mao, Z.: Dreamidentity: Improved editability for efficient face-identity preserved image generation. arXiv preprint abs/2307.00300 (2023)
10. Cheng, Z.Q., Liu, Y., Wu, X., Hua, X.S.: Video ecommerce: Towards online video advertising. In: Proceedings of the 24th ACM international conference on Multimedia. pp. 1365–1374 (2016)
11. Cheng, Z.Q., Wu, X., Liu, Y., Hua, X.S.: Video ecommerce++: Toward large scale online video advertising. IEEE transactions on multimedia 19(6), 1170–1183 (2017)
12. Cheng, Z.Q., Wu, X., Liu, Y., Hua, X.S.: Video2shop: Exact matching clothes in videos to online shopping images. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4048–4056 (2017)
13. David Turner, Robert Wilhelm, Werner Lemberg: FreeType 2 (1996), https://freetype.org/index.html
14. Dhariwal, P., Nichol, A.Q.: Diffusion models beat gans on image synthesis. In: NeurIPS. pp. 8780–8794 (2021)
15. Frans, K., Soros, L.B., Witkowski, O.: Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. In: NeurIPS (2022)

16. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. In: ICLR (2023)
17. Gupta, T., Kembhavi, A.: Visual programming: Compositional visual reasoning without training. In: CVPR. pp. 14953–14962 (2023)
18. He, J., Cheng, Z., Li, C., Sun, J., Xiang, W., Lin, X., Kang, X., Jin, Z., Hu, Y., Luo, B., Geng, Y., Xie, X.: Wordart designer: User-driven artistic typography synthesis using large language models. In: EMNLP. pp. 223–232 (2023)
19. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) NeurIPS (2020)
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net (2022)
21. Huang, L., Chen, D., Liu, Y., Shen, Y., Zhao, D., Zhou, J.: Composer: Creative and controllable image synthesis with composable conditions. arXiv preprint abs/2302.09778 (2023)
22. Iluz, S., Vinker, Y., Hertz, A., Berio, D., Cohen-Or, D., Shamir, A.: Word-as-image for semantic typography. SIGGRAPH (2023)
23. Lab, D.: Deepfloyd if. https://github.com/deep-floyd/IF (2023)
24. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning (2023)
25. Liu, R., Garrette, D., Saharia, C., Chan, W., Roberts, A., Narang, S., Blok, I., Mical, R., Norouzi, M., Constant, N.: Character-aware models improve visual text rendering. In: ACL. pp. 16270–16297 (2023)
26. Ma, J., Zhao, M., Chen, C., Wang, R., Niu, D., Lu, H., Lin, X.: Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. arXiv preprint abs/2303.17870 (2023)
27. Ma, Y., Yang, H., Wang, W., Fu, J., Liu, J.: Unified multi-modal latent diffusion for joint subject and text conditional image generation. arXiv preprint abs/2303.09319 (2023)
28. Meng, C., He, Y., Song, Y., Song, J., Wu, J., Zhu, J., Ermon, S.: Sdedit: Guided image synthesis and editing with stochastic differential equations. In: ICLR (2022)
29. Mou, C., Wang, X., Xie, L., Zhang, J., Qi, Z., Shan, Y., Qie, X.: T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint abs/2302.08453 (2023)
30. Nichol, A.Q., Dhariwal, P.: Improved denoising diffusion probabilistic models. In: Meila, M., Zhang, T. (eds.) ICML. vol. 139, pp. 8162–8171 (2021)
31. Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., Rombach, R.: SDXL: improving latent diffusion models for high-resolution image synthesis. CoRR abs/2307.01952 (2023)
32. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with CLIP latents. arXiv preprint abs/2204.06125 (2022)
33. Ramesh, A., Pavlov, M., Goh, G., Gray, S., Voss, C., Radford, A., Chen, M., Sutskever, I.: Zero-shot text-to-image generation. In: ICML. vol. 139, pp. 8821–8831. PMLR (2021)
34. Rodriguez, J.A., Vázquez, D., Laradji, I.H., Pedersoli, M., Rodríguez, P.: OCR-VQGAN: taming text-within-image generation. In: WACV. pp. 3678–3687 (2023)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (June 2022)
36. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, S.K.S., Lopes, R.G., Ayan, B.K., Salimans, T., Ho, J., Fleet, D.J., Norouzi, M.: Photorealistic text-to-image diffusion models with deep language understanding. In: NeurIPS (2022)

37. Shi, J., Xiong, W., Lin, Z., Jung, H.J.: Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint abs/2304.03411 (2023)
38. Song, Y., Sohl-Dickstein, J., Kingma, D.P., Kumar, A., Ermon, S., Poole, B.: Score-based generative modeling through stochastic differential equations. In: ICLR (2021)
39. Sun, G.L., Cheng, Z.Q., Wu, X., Peng, Q.: Personalized clothing recommendation combining user social circle and fashion style consistency. Multimedia Tools and Applications 77, 17731–17754 (2018)
40. Tanveer, M., Wang, Y., Mahdavi-Amiri, A., Zhang, H.: Ds-fusion: Artistic typography via discriminated and stylized diffusion. arXiv preprint abs/2303.09604 (2023)
41. Tendulkar, P., Krishna, K., Selvaraju, R.R., Parikh, D.: Trick or treat: Thematic reinforcement for artistic typography. In: Grace, K., Cook, M., Ventura, D., Maher, M.L. (eds.) Proceedings of the Tenth International Conference on Computational Creativity, ICCC. pp. 188–195
42. Tuo, Y., Xiang, W., He, J., Geng, Y., Xie, X.: Anytext: Multilingual visual text generation and editing. CoRR abs/2311.03054 (2023)
43. Vungthong, S., Djonov, E., Torr, J.: Images as a resource for supporting vocabulary learning: A multimodal analysis of thai efl tablet apps for primary school children. TESOL Quarterly 51(1), 32–58 (2017)
44. Wang, L., Ma, C., Feng, X., Zhang, Z., Yang, H., Zhang, J., Chen, Z., Tang, J., Chen, X., Lin, Y., Zhao, W.X., Wei, Z., Wen, J.: A survey on large language model based autonomous agents. arXiv preprint abs/2308.11432 (2023)
45. Wei, Y., Zhang, Y., Ji, Z., Bai, J., Zhang, L., Zuo, W.: ELITE: encoding visual concepts into textual embeddings for customized text-to-image generation. arXiv preprint abs/2302.13848 (2023)
46. Xi, Z., Chen, W., Guo, X., He, W., Ding, Y., Hong, B., Zhang, M., Wang, J., Jin, S., Zhou, E., Zheng, R., Fan, X., Wang, X., Xiong, L., Zhou, Y., Wang, W., Jiang, C., Zou, Y., Liu, X., Yin, Z., Dou, S., Weng, R., Cheng, W., Zhang, Q., Qin, W., Zheng, Y., Qiu, X., Huan, X., Gui, T.: The rise and potential of large language model based agents: A survey. arXiv preprint abs/2309.07864 (2023)
47. Yang, Y., Gui, D., Yuan, Y., Ding, H., Hu, H., Chen, K.: Glyphcontrol: Glyph conditional control for visual text generation. arXiv preprint abs/2305.18259 (2023)
48. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T., Cao, Y., Narasimhan, K.: Tree of thoughts: Deliberate problem solving with large language models. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems (2023)
49. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K.R., Cao, Y.: React: Synergizing reasoning and acting in language models. In: The International Conference on Learning Representations, ICLR. OpenReview.net (2023)
50. Zhang, J., Wang, Y., Xiao, W., Luo, Z.: Synthesizing ornamental typefaces. Comput. Graph. Forum 36(1), 64–75 (2017)
51. Zhang, L., Agrawala, M.: Adding conditional control to text-to-image diffusion models. arXiv preprint abs/2302.05543 (2023)