MetaDesigner ADVANCING ARTISTIC TYPOGRAPHY THROUGH AI-DRIVEN, USER-CENTRIC, AND MULTI-LINGUAL WORDART SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

MetaDesigner revolutionizes artistic typography synthesis by leveraging the strengths of Large Language Models (LLMs) to drive a design paradigm centered around user engagement. At the core of this framework lies a *multi-agent system* comprising the *Pipeline*, *Glyph*, and *Texture* agents, which collectively enable the creation of customized WordArt, ranging from semantic enhancements to the imposition of complex textures. MetaDesigner incorporates a comprehensive *feedback mechanism* that harnesses insights from multimodal models and user evaluations to refine and enhance the design process iteratively. Through this feedback loop, the system adeptly tunes hyperparameters to align with user-defined stylistic and thematic preferences, generating WordArt that not only meets but exceeds user expectations in terms of visual appeal and contextual relevance. Empirical validations highlight *MetaDesigner's* capability to effectively serve diverse WordArt applications, consistently producing aesthetically appealing and context-sensitive results. More details are available at `https://shorturl.at/Hdl37`.

1 INTRODUCTION

Typography, a fusion of linguistic expression and visual design, plays a key role in various domains, including advertising Cheng et al. (2016; 2017a;b); Sun et al. (2018), education Vungthong et al. (2017), and tourism Amar et al. (2017). As a medium for communication, artistic expression, and innovation, typography requires a deep understanding of aesthetics and design principles, posing significant challenges for non-professionals. The allure of typography lies not only in its ability to convey information but also in its capacity to evoke emotions and captivate audiences through imaginative and visually appealing designs.

Generative models have revolutionized typographic design by adapting to diverse aesthetic preferences. However, integrating these models to meet complex typesetting requirements presents notable challenges. (1) *The subjective nature of artistic typography*, influenced by individual and cultural factors, complicates the development of generative models that resonate widely. (2) *The scarcity of comprehensive*, annotated resources in artistic typography hinders models' ability to capture and generate diverse typographic styles. Despite significant advancements in text-to-image synthesis, such as denoising diffusion probabilistic models Ho et al. (2020); Ramesh et al. (2021); Song et al. (2021), they remain insufficient in addressing these specific challenges.

Figure 1: Overview of MetaDesigner, showcasing the interplay between the *Pipeline*, *Glyph*, and *Texture* agents, orchestrated to tailor WordArt in alignment with user preferences.

To tackle these challenges head-on, we introduce *MetaDesigner*, a multi-agent system designed to generate artistic typographies aligned with user preferences. MetaDesigner integrates four dedicated intelligent agents: *Pipeline Designer*, *Glyph Designer*, *Texture Designer*, and *Q&A Evaluation Agent*, each playing a crucial role in generating personalized artworks and ensuring a comprehensive, user-centric design process (see Sec. 3). The *Pipeline Designer* orchestrates the system, translating user prompts into structured tasks for other agents (see Sec. 3.1). The *Glyph Designer* creates diverse glyph types, including regular, traditional, and semantic glyphs, customized according to the context of the artwork (see Sec. 3.2). The *Texture Designer* enhances glyphs with various texture styles using LoRA model matching (see Sec. 3.3) on the provided tree structure. Finally, the *Q&A Evaluation Agent* refines the output through an iterative process that incorporates feedback (see Sec. 3.4). Overall, MetaDesigner democratizes artistic text creation by focusing on three key aspects.

- MetaDesigner introduces a multi-agent system that integrates evaluation and optimization modules to facilitate the discovery and realization of personalized artistic text styles. The system's meticulous hyperparameter selection process ensures the generation of artworks that align with individual preferences. The accessibility and popularity of the model are demonstrated by the over 400,000 visits it has received on ModelScope[1].
- The glyph design agent in MetaDesigner enables extensive glyph transformations by leveraging font libraries and semantic translation techniques. The system utilizes a hierarchical model tree containing 68 LoRA models, covering various design aspects to ensure rich diversity in the generated output. This advanced glyph design capability significantly expands the scope of artistic expression in typographic design.
- MetaDesigner showcases the authors' commitment to supporting the artistic typography research community by developing a carefully curated dataset of artistic texts. The dataset contains over 5,000 multilingual images, spanning English, Chinese, Japanese, and Korean, covering a wide range of artistic styles and cultural elements. This extensive dataset serves as a valuable resource for researchers and practitioners to explore new techniques, evaluate existing methods, and push the boundaries of artistic text generation.

2 Related Work

Text-to-Image Synthesis. Denoising diffusion probabilistic models have revolutionized text-to-image synthesis Ho et al. (2020); Song et al. (2021); Dhariwal & Nichol (2021); Nichol & Dhariwal (2021); Saharia et al. (2022); Ramesh et al. (2022); Rombach et al. (2022); Chang et al. (2023), enabling interactive image editing Meng et al. (2022); Gal et al. (2023); Brooks et al. (2022) and multi-condition controllable synthesis Zhang & Agrawala (2023); Mou et al. (2023); Huang et al. (2023). Methods like ELITE Wei et al. (2023), UMM-Diffusion Ma et al. (2023b), and InstantBooth Shi et al. (2023) leverage CLIP image encoders to embed visual concepts into text embeddings.

Visual Text Generation. Generating legible text within images is challenging Rombach et al. (2022). *GlyphDraw* Ma et al. (2023a) and *GlyphControl* Yang et al. (2023) focus on rendering and aligning characters, while *TextDiffuser* Chen et al. (2023b) introduces character-level segmentation masks. Large-scale language models are crucial for visually accurate text generation Saharia et al. (2022); Balaji et al. (2022); Lab (2023). However, character-blind text encoders struggle with non-Latin scripts Liu et al. (2023). *GlyphDraw* addresses Chinese text rendering by fine-tuning the text encoder and using CLIP for glyph embeddings Ma et al. (2023a), while *DiffUTE* employs a pre-trained image encoder for glyph extraction in image editing Chen et al. (2023a).

WordArt Synthesis. Synthesizing WordArt Tanveer et al. (2023); Iluz et al. (2023); Berio et al. (2022); Tendulkar et al.; Zhang et al. (2017) requires integrating semantics with artistic and legible text representation. Early attempts Tendulkar et al. explored substituting letters with semantically similar icons. Recent advancements leverage large generative models, with *Word-As-Image* Iluz et al. (2023) introducing artistic typography for Latin text and *DS-Fusion* Tanveer et al. (2023) synthesizing intricate text forms, including hieroglyphics.

MetaDesigner distinguishes itself by integrating specialized agents—*Pipeline Designer*, *Glyph Designer*, *Texture Designer*, and *Q&A Evaluation* Agent—into an *interactive system* that iteratively

[1]The anonymity rule might have been violated, so the model link is not provided.

refines WordArt based on user feedback. This adaptive approach addresses a broader spectrum of user preferences and creative demands compared to previous methods.

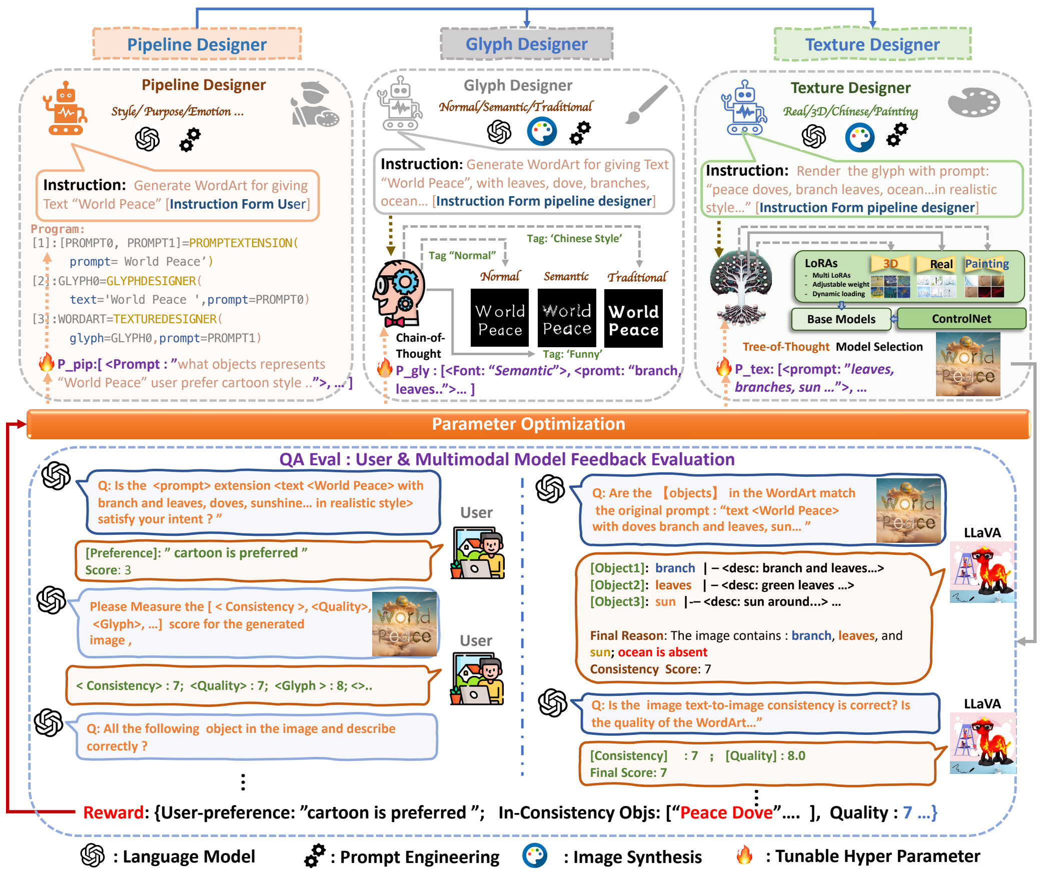

Figure 2: MetaDesigner Architectural Overview: This schematic illustrates the structured integration of MetaDesigner's three core intelligent agents—*Pipeline Designer*, *Glyph Designer*, and *Texture Designer*—each contributing to personalized WordArt creation. The system also incorporates a *Q&A Evaluation* agent that collaboratively enhances output quality. The diagram highlights the interactive, iterative process through which textual inputs are transformed into visually compelling, user-preference-driven artistic typography.

3 MetaDesigner Framework

The MetaDesigner framework is an interactive multi-agent system for synthesizing WordArt that aligns with user preferences. It integrates four specialized intelligent agents: Pipeline Designer, Glyph Designer, Texture Designer, and Q&A Evaluation Agent, each playing a crucial role in generating personalized WordArt. The WordArt image $\hat{I}$ synthesized by MetaDesigner, denoted by Ψ, is mathematically represented as:

$$\hat{I} = \Psi(s^{\text{user}}, \phi, \mathcal{P}, \mathcal{M}) \tag{1}$$

where s^{user} signifies a user's prompt, encapsulating their preferences and input. The term ϕ represents the collective functionality of the involved agents: $\phi = \{\phi^{\text{pip}}, \phi^{\text{gly}}, \phi^{\text{tex}}, \phi^{\text{qa}}\}$, corresponding to the Pipeline Designer, Glyph Designer, Texture Designer, and Q&A Evaluation Agent, respectively. $\mathcal{M}$ is the library of models utilized by the Texture Designer, while $\mathcal{P}$ denotes a set of learnable hyperparameters designed to fine-tune the system's outputs to closely match user preferences through interactive and context-aware learning. $\mathcal{P} = \{\mathcal{P}^{\text{pip}}, \mathcal{P}^{\text{gly}}, \mathcal{P}^{\text{tex}}, \mathcal{P}^{\text{qa}}\}$, with each subset specifically allocated to its corresponding agent. $\mathcal{I}$ symbolizes the WordArt image crafted in the preceding iteration, highlighting the capacity for iterative learning and adaptation to user preferences:

$$\begin{aligned} \hat{I} &= \Psi(s^{\text{user}}, \phi, \mathcal{P}, \mathcal{M}) \\ &= \phi^{\text{pip}}(s^{\text{user}}) \cdot \phi^{\text{tex}}(\phi^{\text{gly}}(s^{\text{gly}}), s^{\text{tex}}, \mathcal{M}) \cdot \phi^{\text{qa}}(I_{\text{prev}}, \mathcal{P}^{\text{qa}}) \end{aligned} \tag{2}$$

where $\mathbf{S} = \phi^{\text{pip}}(s^{\text{user}})$ generates the extension prompt, further decomposed into $\mathbf{S} = \{s^{\text{gly}}, s^{\text{tex}}\}$, which are tailored prompts directing the Glyph and Texture Designers. ϕ^{qa} evaluates and adjusts the parameters based on the feedback from the Q&A Evaluation Agent. The following sections introduce each agent module in detail.

3.1 Pipeline Designer Agent

The Pipeline Designer transforms visual tasks into a structured coding style using visual programming techniques, enabling precise coordination among synthesis agents. This approach enhances the system's accessibility and usability for end-users by generating a series of questions that guide the GPT-4 model through a Chain-of-Thought (CoT) process, which breaks down complex tasks into manageable steps and adapts the workflow based on iterative feedback.

Visual Programming. As demonstrated in previous work and Fig. 2, we leverage GPT-4's in-context learning ability to generate visual programs from natural language instructions, without the need for fine-tuning. The in-context examples consist of manually written programs that can typically be constructed without an accompanying image. Each program step comprises a module name, input argument names and values, and an output variable name.

Prompt Extension. To tailor the system to user preferences based on image style, application domain, and background, the Pipeline Designer employs a series of questions to guide GPT-4 through a CoT reasoning process. Given a user prompt s^{user}, the Pipeline Designer extends it using the GPT-4 to create an enriched prompt $\mathbf{S}$:

$$\mathbf{S} = \phi^{\text{pip}}(s^{\text{user}}) = \{s^{\text{gly}}, s^{\text{tex}}\} \tag{3}$$

where ϕ^{pip} represents the Pipeline Designer function, and s^{gly} and s^{tex} are the specific prompts for the Glyph Designer and Texture Designer, respectively.

Feedback Integration. The Pipeline Designer incorporates feedback to refine the workflow and align the synthesis process with user preferences. This feedback loop is formalized as:

$$\mathcal{P}^{\text{pip}}_{\text{new}} = \mathcal{F}(G|s^{\text{update}}) \tag{4}$$

where $\mathcal{F}$ is the feedback function, G represents the aggregated feedback, and s^{update} is the update signal for the Pipeline Designer's hyperparameters $\mathcal{P}^{\text{pip}}$. The overall process of the Pipeline Designer is summarized as:

$$\mathbf{S}, \mathcal{P}^{\text{pip}}_{\text{new}} = \phi^{\text{pip}}(s^{\text{user}}, \mathcal{P}^{\text{pip}}, \mathcal{F}(G|s^{\text{update}})) \tag{5}$$

This formulation captures the interplay between user input, prompt extension, and feedback integration. By integrating user preferences, extending prompts using GPT-4 model, and iteratively refining the synthesis process through a structured feedback loop, the Pipeline Designer ensures the generation of high-quality, personalized WordArt.

3.2 Glyph Designer Agent

The Glyph Designer is a key component of MetaDesigner, supporting the generation of three glyph types: *Normal*, *Traditional*, and *Semantic*. Normal and traditional glyphs suit formal contexts like weddings and galas, while semantic glyphs are tailored for humorous and creative applications. The GPT-4 selects the appropriate glyph type based on the given prompts.

Normal & Traditional Glyph. For formal scenarios, the Glyph Designer renders normal and traditional glyphs using the FreeType font library:

$$G_n, G_t = \phi_n(s^{\text{gly}}), \phi_t(s^{\text{gly}}) \tag{6}$$

where ϕ_n and ϕ_t are the rendering functions for normal and traditional glyphs, respectively, and s^{gly} is the input prompt.

Semantic Glyph Transformation. The Glyph Designer enhances WordArt typography through semantic glyph transformation, optimizing vector graphs to approximate target objects using differentiable rasterization and a depth-to-image SD model:

$$G_s = \phi_s(s^{\text{sem}}, \mathcal{M}) \tag{7}$$

where ϕ_s is the semantic glyph transformation function, s^{sem} is the input prompt, and $\mathcal{M}$ is the model library. The GPT-4 ensures readability by integrating context for full paragraph recognition.

The transformed glyph image I_{sem} is created from the trainable parameters θ of the SVG-format character input using DiffVG $\phi(\cdot)$. An optimized and cropped character segment x yields an enhanced image batch X_{aug}. The semantic concept S and X_{aug} are input into a vision-language model to compute the loss for parameter optimization, applying the SDS loss in the latent space code z.

Glyph Style Selection. The choice of glyph type depends on the context, with normal and traditional glyphs used in formal settings and semantic glyphs in creative and humorous contexts. Glyphs play a crucial role in visual communication, influencing readability and the reader's emotional response.

The overall glyph generation process in the Glyph Designer can be summarized as:

$$G = \phi(s^{\text{gly}}, \mathcal{P}^{\text{gly}}, \mathcal{M}) = \begin{cases} \phi_n(s^{\text{gly}}) & \text{if formal context} \\ \phi_t(s^{\text{gly}}) & \text{if traditional context} \\ \phi_s(s^{\text{sem}}, \mathcal{M}) & \text{if creative context} \end{cases} \tag{8}$$

where G is the generated glyph, $\mathcal{P}^{\text{gly}}$ represents the glyph design hyperparameters, and the glyph type selection (ϕ_n, ϕ_t, ϕ_s) depends on the context derived from the user prompt. The Glyph Designer dynamically selects and renders contextually appropriate glyph styles using advanced techniques and iterative optimization.

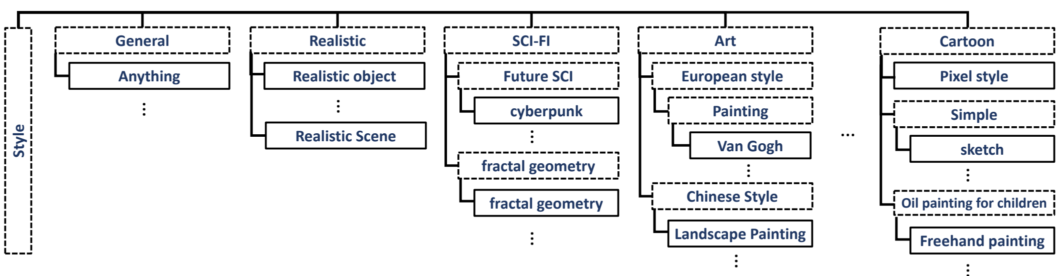

Figure 3: The hierarchical model tree consists of multi-level sub-categories to facilitate finer granularity in **ToT** model selection.

3.3 TEXTURE DESIGNER AGENT

The Texture Designer overcomes single-model limitations by integrating controllable image synthesis with the Tree-of-Thought (ToT) framework, catering to diverse user demands while infusing creativity and customization into the design process. The agent is formulated as:

$$\begin{aligned} I_{\text{tex}} &= \psi^{\text{tex}}(I_{\text{gly}}, s^{\text{tex}}, \mathcal{F}, \mathcal{M}) \\ &= \mathbf{TexR}(I_{\text{gly}}, s^{\text{tex}}, C, \mathbf{ToTSel}(s^{\text{tex}}, \mathcal{F}, \mathcal{M})) \end{aligned} \tag{9}$$

where I_{tex} is the final textured image, **TexR** is the controllable synthesis mechanism, **ToTSel** is the ToT-informed model selection function, I_{gly} is the input glyph image, s^{tex} is the texture prompt, C are control conditions, $\mathcal{F}$ guides the ToT process, and $\mathcal{M}$ is the model library. Next we only give a formulaic description, more details are in the supplementary materials.

(1) Controllable Synthesis. The ControlNet enhances the adaptability and diversity of texture styles by manipulating parameters such as Canny Edge, Depth, Scribble, and original font images:

$$I_{\text{tex}} = \mathbf{TexR}(I_{\text{gly}}, s^{\text{tex}}, C, \mathcal{W}) \tag{10}$$

where $\mathcal{W}$ represents the selected model weights, allowing for fine control over texture characteristics.

(2) Tree-of-Thought Selection. The ToT strategy explores various reasoning paths to ensure uniqueness and artistic integrity. It begins with *Thought Decomposition and Generation*, breaking down prompt s^{tex} into conceptual pathways $\{z_1, z_2, \ldots, z_n\}$ guided by $\mathcal{F}(s^{\text{tex}})$. *State Evaluation and Model Search* then assesses each pathway's feasibility using heuristic $V(z_i)$ and searches for the best-fitting model $\mathcal{M}_{\text{best}}$:

$$\mathcal{M}_{\text{best}} = \arg\max_{\mathcal{M} \in \mathcal{M}_{\text{lib}}} \sum_{i=1}^{n} V(z_i | \mathcal{M}) \tag{11}$$

ensuring the selected model aligns with the user's creative vision.

(3) Model Library Integration. The 68 LoRA model library, organized into categories like "General," "Realistic," "SCI-FI," "Art," "Design," and "Cartoon," facilitates refined model choices as $\mathcal{M} = \{\mathcal{M}_{\text{gen}}, \mathcal{M}_{\text{real}}, \mathcal{M}_{\text{sci}}, \mathcal{M}_{\text{art}}, \mathcal{M}_{\text{des}}, \mathcal{M}_{\text{car}}\}$. To meet diverse requirements, the Texture Designer merges LoRA models through weighted combinations:

$$\mathcal{W}_{\text{fusion}} = \sum_i \alpha_i \mathcal{W}_i \tag{12}$$

with α_i representing the weights assigned to each LoRA model $\mathcal{W}_i$. The unified framework in Eq. 9 captures the transformation of glyph images into textural masterpieces, guided by user inputs and control conditions. This sophisticated and iterative process ensures each texture is uniquely tailored to the user's vision, resulting in a highly personalized and expressive output.

3.4 Q&A Evaluation Agent

We have designed a feedback mechanism for hyperparameter tuning that focuses on four key aspects: text-to-image consistency, image quality, glyph feedback, and user preference. The LLaVA model evaluates text-to-image consistency and image quality, while user studies provide feedback in a Q&A format. Evaluation prompts are generated based on the goals of each synthesized image, and the LLaVA model analyzes these prompts against the generated images, producing feedback metrics: $G_m = \{g_m^{\text{cos}}, g_m^{\text{qua}}, \mathcal{L}_m\}$. GPT-4 summarizes this feedback, integrating the metrics into coherent scores and providing rationales for hyperparameter adjustments.

MetaDesigner also gathers user feedback structured around GPT-4-generated questions to capture user preferences (g_u^{pref}) and perceptions of glyph style (g_u^{gly}): $G_u = \{g_u^{\text{cos}}, g_u^{\text{qua}}, g_u^{\text{gly}}, g_u^{\text{pref}}, \mathcal{L}_u\}$. Although optional, user feedback is crucial for aligning outputs with user expectations. The final feedback, $G = \text{Merge}(G_m, G_u)$, combines insights from multi-modal evaluations and user responses for hyperparameter optimization, with user input prioritized through a voting strategy.

Formally, the optimization strategy focuses on the objective function, capturing the essence of feedback integration: $\mathcal{L} = \mathcal{L}_m + \mathcal{L}_u$, where $\mathcal{L}$ combines contributions from the LLaVA model ($\mathcal{L}_m$) and user feedback ($\mathcal{L}_u$). The goal is to optimize the hyperparameters $\mathcal{P}^{\text{gly}}$ and $\mathcal{P}^{\text{tex}}$ for the synthesized image $\hat{I}$ as:

$$\mathcal{L}_m = \underset{\{\mathcal{P}^{\text{gly}}, \mathcal{P}^{\text{tex}}\}}{\text{argmax}} \; \mathcal{H}(s^{\text{eval}}, \hat{I}) \tag{13}$$

where s^{eval} is the evaluative prompt, and $\mathcal{H}$ denotes the heuristic from the LLaVA model's analysis. This process guides the search for optimal hyperparameters, maximizing the quality and relevance of the generated WordArt.

As shown in Alg. 1 and Fig. 4, the adaptive feedback integration and tuning process dynamically adjusts the hyperparameter set $\mathcal{P}$, including pipeline, glyph, and texture design parameters, based on aggregated feedback. User preferences and text feedback primarily influence glyph design parameters ($\mathcal{P}^{\text{gly}}$) and the pipeline strategy, while objective feedback on text-to-image consistency and image quality informs texture design parameters ($\mathcal{P}^{\text{tex}}$). This iterative loop, where $\hat{\mathcal{P}} = \mathcal{F}(G|s^{\text{update}})$ denotes the updated hyperparameter set derived from synthesized feedback $G = \text{Merge}(G_m, G_u)$, ensures responsiveness to both quality metrics and user preferences, fostering continuous improvement and customization. More details are in the supplementary materials.

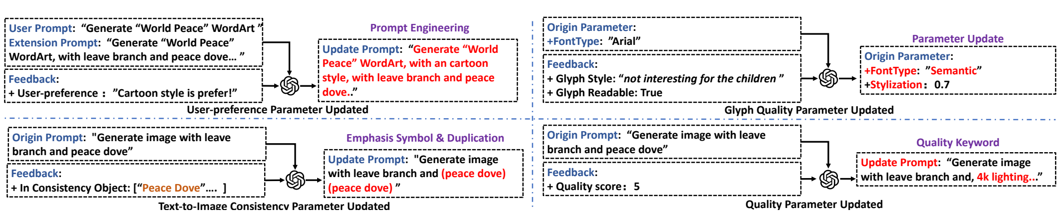

Figure 4: Illustration of the feedback loop for hyperparameter tuning, showcasing the integration of user preference, glyph quality, text-to-image consistency, and image quality evaluations.

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.* denotes the result without evaluation feedback.

Evaluation Dimension	SDXL	SDXL-Augmented	TextDiffuser	TextDiffuser-2	Anytext	DALLE 3	Ours*	Ours
Text Accuracy	3.6	3.6	44.1	42.7	82.3	33.2	88.2	**96.8**
Aesthetics	0.5	5.9	0.9	0.5	0.9	10.9	5.0	**75.4**
Creativity	0.9	1.8	0.5	0.9	0.9	7.3	10.5	**77.2**

Table 2: **Quantitative comparison**. We compare our method with SOTA methods on the SSIM and LPIPS metrics. $\mathcal{P}$ and $\mathcal{D}$ respectively indicate ground truth images sourced from Promeai and design-related websites. **Boldface** and <u>underlining</u> denote the best and second-best results.* denotes the result without evaluation feedback.

GT	Metrics	Text	SDXL	SDXL-Aug.	TextDiffuser	TextDiffuser-2	Anytext	DALLE 3	Ours*	Ours
$\mathcal{P}$	SSIM↑	Eng.	0.1254	0.1381	<u>0.1860</u>	0.1641	0.1324	0.0834	0.1730	**0.2397**
		CJK	0.1853	0.2092	0.1747	0.2037	0.1021	0.1401	<u>0.2269</u>	**0.2643**
		All	0.1553	0.1736	0.1803	0.1839	0.1172	0.1117	<u>0.2000</u>	**0.2520**
	LPIPS↓	Eng.	0.7491	0.7684	0.7652	0.7441	0.7453	0.7653	<u>0.6960</u>	**0.6910**
		CJK	0.7712	0.7307	0.7970	0.7687	0.7601	0.7693	<u>0.6937</u>	**0.6846**
		All	0.7602	0.7496	0.7811	0.7564	0.7527	0.7673	<u>0.6949</u>	**0.6878**
$\mathcal{D}$	SSIM↑	Eng.	0.1802	<u>0.2439</u>	0.2342	0.2036	0.1669	0.1413	0.1913	**0.3119**
		CJK	0.1951	0.2093	0.1846	0.1987	0.1073	0.1542	<u>0.2184</u>	**0.2539**
		All	0.1877	<u>0.2266</u>	0.2094	0.2012	0.1371	0.1478	0.2048	**0.2829**
	LPIPS↓	Eng.	0.7993	0.8157	0.8312	0.8366	0.8495	**0.7650**	0.8169	<u>0.7950</u>
		CJK	0.8023	0.7964	0.8249	0.8429	0.8437	**0.7872**	0.7912	<u>0.7880</u>
		All	0.8008	0.8060	0.8280	0.8397	0.8466	**0.7761**	0.8040	<u>0.7915</u>

4 EXPERIMENTS

To evaluate the effectiveness of MetaDesigner, we compiled a set of 150 prompts to synthesize WordArt, categorized into five themes: "cartoon," "design," "reality," "sci-fi," and "traditional culture." These prompts were carefully curated to cover a wide range of artistic styles and cultural elements. For the user study, a subset of 20 prompts was selected, encompassing English, Chinese, Japanese, and Korean, to evaluate the system's multilingual capabilities.

4.1 COMPARISON WITH SOTA METHODS

We conducted comparative experiments against contemporary state-of-the-art (SOTA) techniques, including Stable Diffusion XL (SD-XL) Podell et al. (2023), TextDiffuser Frans et al. (2022), TextDiffuser-2 Chen et al. (2023c), Anytext Tuo et al. (2023), and DALL-E3. These methods were chosen to represent a diverse range of approaches in text-to-image synthesis. The results, depicted in Figure 5, were evaluated in three key aspects:

WordArt Synthesis Success. As illustrated in Figure 5, SD-XL fails to accurately render text, exhibiting incomplete support for English. The TextDiffuser series performs adequately in English but falls short in Chinese, Korean, and Japanese. Anytext delivers superior results but does not successfully handle Korean and Japanese. DALL-E3 is limited to English support only. In contrast, MetaDesigner demonstrates robust performance across all tested languages, successfully generating high-quality WordArt. More details are in the supplementary material.

Quality and Diversity. The TextDiffuser series generates similar WordArt, restricted by the SD 1.5 model and limited training data distribution. Anytext has more diversity but lower quality. DALL-E3 has better quality but a cinematic and 3D style. MetaDesigner achieves the best diversity in styles, including realistic, cartoon, and 3D, while maintaining high-quality output across all styles.

Creativity and Reasonability. SD-XL produces stylistic text unrelated to the meaning. The TextDiffuser series incorporates relevant elements like leaves and a peace dove. Anytext displays a logo design that fails to align with "World Peace." DALL-E3 produces a cinematic-tone 3D WordArt with elements somewhat aligning with the theme but lacking clarity. MetaDesigner intuitively grasps the essence of "World Peace," crafting WordArt with the sun, branch leaves, a peace dove, and additional resonating elements, demonstrating its ability to generate creative and semantically meaningful designs. More details are in the supplementary material.

Figure 5: 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 are generated with more keywords: "Sun, Peace Dove, leaves, cloud".

Figure 6: Comparison of single letter WordArt. Part of the comparison results are copied from DS-Fusion Tanveer et al. (2023). The prompts are the "Dragon", and "Plant", respectively.

Quantitative Analysis. Given the challenges OCR models face with artistic words, we performed a user study to evaluate our approach and SOTA techniques based on "Text accuracy" and "Aesthetics & Creativity". A total of 11 participants were involved in the study, ensuring a diverse range of perspectives. As shown in Table 1, our approach demonstrated significant advantages in both dimensions, highlighting its superiority in generating accurate and visually appealing WordArt. Table 2 presents quantitative results using SSIM and LPIPS metrics, further reaffirming the excellence of our approach in terms of readability and aesthetics.

Letter-Level Comparison. To evaluate MetaDesigner's performance at a more granular level, we benchmarked it against Google search, Stable Diffusion Rombach et al. (2022), DALL-E3, and DS-Fusion Tanveer et al. (2023). As shown in Figure 6, Stable Diffusion struggles to generate eligible WordArts comprehensively. DS-Fusion produces superior outcomes with shape deformation and legibility but lacks variety. DALL-E3 delivers high-quality renderings focused on textural representation. MetaDesigner generates WordArt with greater stylistic creativity and maintains high detail quality, demonstrating its ability to create visually striking and diverse designs at the letter level. More examples are in the supplementary material.

4.2 Effect of Tree-of-Thought

To verify the effectiveness of the Tree-of-Thought (ToT) scheme, we conducted a quantitative analysis comparing the ControlNet and the ToT-LoRA+ControlNet approaches. GPT-4, a state-of-the-art language model, was used to measure the "Relevance," "Quality," and "Style" of the synthesized WordArt. The results, illustrated in Figure 7, show significant improvements achieved by the ToT-LoRA in all metrics, highlighting its ability to generate WordArt that is highly relevant to the given prompts, visually appealing, and stylistically diverse.

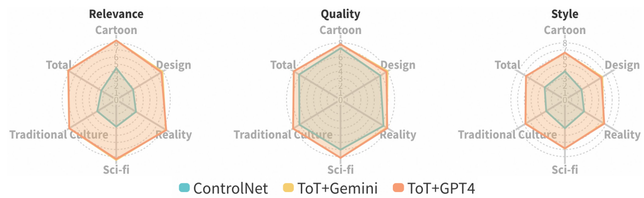

Figure 7: The evaluation of the synthesis WordArt. From left to right are the “Relevance", “Quality", and “Style" scores generated by the ChatGPT-4 in the sub-categories. (The performance of ToT+Gemini and ToT+GPT4 are very close)

A case study, presented in Figure 8, demonstrates that the ToT-LoRA scheme significantly outperforms ControlNet in WordArt style and text-to-image consistency. Specifically, ToT-LoRA+ControlNet excels in rendering complex textures and maintaining coherence with the given prompts, such as "ethnic customs, New Year, dumplings, steamed bread, kitchen, mother, little boy" and "Renaissance style, Castle." The results show clearer thematic representation and enhanced visual appeal, further validating the effectiveness of our approach. More examples are in the supplementary material.

Figure 8: WordArt Texture rendering

4.3 Effect of Optimization

To illustrate the impact of the optimization process, we conducted a detailed case study, presented in Figure 9. The LLaVA system is employed to identify objects mentioned in the prompt but absent in the generated WordArt. This information is then used to update the generation process, incorporating the omitted elements. Techniques such as symbol enhancement, word sequencing adjustments, and keyword repetition are employed to augment the WordArt generation process, ensuring that the final output accurately reflects the input prompt. As shown in Figure 9, the optimization process proceeds through several steps to achieve better alignment with the prompts. In Step 1, the initial generation may miss key elements such as "little girl" in the prompt "old man, cake, candles, little girl." In Step 2, iterative refinement introduces missing elements, enhancing text-to-image consistency. Similarly, for the prompt "kitchen, girl, steamed bread, a plate of fruit," the optimization process adds missing objects and refines their depiction over multiple steps. This optimization process plays a crucial role in enhancing the semantic consistency and visual quality of the generated WordArt.

Figure 9: The examples are the optimization of the text-to-image consistency.

Figure 10: A variety of exquisite examples are provided. Our method supports multiple languages (Chinese, English, Japanese, Korean, Arabic numerals, etc.) and accommodates different word counts. The generated images can be applied to scenarios such as e-commerce marketing, film and entertainment, and personalized social networking.

4.4 WordArt Dataset

To further advance the WordArt community and support ongoing research in this domain, we have meticulously curated a Creative and Diverse WordArt Dataset using our approach. This comprehensive dataset encompasses a total of 5,000 multilingual images, including representations from English, Chinese, Japanese, and Korean, spanning a wide array of artistic styles and cultural elements. By providing such a rich and varied resource, we aim to empower researchers and practitioners in the field of artistic typography, enabling them to explore novel techniques, evaluate existing methods, and push the boundaries of WordArt generation.

5 Conclusion

In this work, we introduce *MetaDesigner*, a revolutionary framework for synthesizing artistic typography by leveraging the strengths of Large Language Models (LLMs) to drive a user-centric design paradigm. At the core of MetaDesigner is a multi-agent system comprising the *Pipeline*, *Glyph*, and *Texture* agents, which collectively enable customized WordArt creations, from semantic enhancements to the imposition of complex textures. By merging generative AI with artistic typography, MetaDesigner represents a significant advancement in the field, democratizing typography design and enriching the digital art domain. The system's ability to understand user preferences and generate contextually relevant WordArt sets it apart from existing solutions, making it a valuable tool for both professionals and enthusiasts. Future work will focus on expanding the WordArt Dataset, refining the multi-agent system, and exploring new applications and enhancements in artistic typography synthesis. More details at `https://shorturl.at/Hdl37`, enabling researchers and practitioners to explore and build upon the MetaDesigner framework.

REFERENCES

Jennifer Amar, Olivier Droulers, and Patrick Legohérel. Typography in destination advertising: An exploratory study and research perspectives. *Tourism Management*, 63:77–86, 2017. ISSN 0261-5177. doi: https://doi.org/10.1016/j.tourman.2017.06.002. URL `https://www.sciencedirect.com/science/article/pii/S0261517717301243`.

Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint*, abs/2211.01324, 2022.

Daniel Berio, Frederic Fol Leymarie, Paul Asente, and Jose Echevarria. Strokestyles: Stroke-based segmentation and stylization of fonts. *ACM Trans. Graph.*, 41(3):28:1–28:21, 2022.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint*, abs/2211.09800, 2022.

Huiwen Chang, Han Zhang, Jarred Barber, Aaron Maschinot, José Lezama, Lu Jiang, Ming-Hsuan Yang, Kevin Murphy, William T. Freeman, Michael Rubinstein, Yuanzhen Li, and Dilip Krishnan. Muse: Text-to-image generation via masked generative transformers. *arXiv preprint*, abs/2301.00704, 2023.

Haoxing Chen, Zhuoer Xu, Zhangxuan Gu, Jun Lan, Xing Zheng, Yaohui Li, Changhua Meng, Huijia Zhu, and Weiqiang Wang. Diffute: Universal text editing diffusion model. *arXiv preprint*, abs/2305.10825, 2023a.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *arXiv preprint*, abs/2305.10855, 2023b.

Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. *CoRR*, abs/2311.16465, 2023c.

Zhi-Qi Cheng, Yang Liu, Xiao Wu, and Xian-Sheng Hua. Video ecommerce: Towards online video advertising. In *Proceedings of the 24th ACM international conference on Multimedia*, pp. 1365–1374, 2016.

Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. Video ecommerce++: Toward large scale online video advertising. *IEEE transactions on multimedia*, 19(6):1170–1183, 2017a.

Zhi-Qi Cheng, Xiao Wu, Yang Liu, and Xian-Sheng Hua. 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, 2017b.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pp. 8780–8794, 2021.

Kevin Frans, Lisa B. Soros, and Olaf Witkowski. Clipdraw: Exploring text-to-drawing synthesis through language-image encoders. In *NeurIPS*, 2022.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *ICLR*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *NeurIPS*, 2020.

Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint*, abs/2302.09778, 2023.

Shir Iluz, Yael Vinker, Amir Hertz, Daniel Berio, Daniel Cohen-Or, and Ariel Shamir. Word-as-image for semantic typography. *SIGGRAPH*, 2023.

DeepFloyd Lab. Deepfloyd if. *https://github.com/deep-floyd/IF*, 2023.

Rosanne Liu, Dan Garrette, Chitwan Saharia, William Chan, Adam Roberts, Sharan Narang, Irina Blok, RJ Mical, Mohammad Norouzi, and Noah Constant. Character-aware models improve visual text rendering. In *ACL*, pp. 16270–16297, 2023.

Jian Ma, Mingjun Zhao, Chen Chen, Ruichen Wang, Di Niu, Haonan Lu, and Xiaodong Lin. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *arXiv preprint*, abs/2303.17870, 2023a.

Yiyang Ma, Huan Yang, Wenjing Wang, Jianlong Fu, and Jiaying Liu. Unified multi-modal latent diffusion for joint subject and text conditional image generation. *arXiv preprint*, abs/2303.09319, 2023b.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. *arXiv preprint*, abs/2302.08453, 2023.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang (eds.), *ICML*, volume 139, pp. 8162–8171, 2021.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, volume 139, pp. 8821–8831. PMLR, 2021.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint*, abs/2204.06125, 2022.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10684–10695, June 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS*, 2022.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint*, abs/2304.03411, 2023.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Guang-Lu Sun, Zhi-Qi Cheng, Xiao Wu, and Qiang Peng. Personalized clothing recommendation combining user social circle and fashion style consistency. *Multimedia Tools and Applications*, 77:17731–17754, 2018.

Maham Tanveer, Yizhi Wang, Ali Mahdavi-Amiri, and Hao Zhang. Ds-fusion: Artistic typography via discriminated and stylized diffusion. *arXiv preprint*, abs/2303.09604, 2023.

Purva Tendulkar, Kalpesh Krishna, Ramprasaath R. Selvaraju, and Devi Parikh. Trick or treat : Thematic reinforcement for artistic typography. In Kazjon Grace, Michael Cook, Dan Ventura, and Mary Lou Maher (eds.), *Proceedings of the Tenth International Conference on Computational Creativity, ICCC*, pp. 188–195.

Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *CoRR*, abs/2311.03054, 2023.

Sompatu Vungthong, Emilia Djonov, and Jane Torr. 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.

Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint*, abs/2302.13848, 2023.

Yukang Yang, Dongnan Gui, Yuhui Yuan, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *arXiv preprint*, abs/2305.18259, 2023.

Junsong Zhang, Yu Wang, Weiyi Xiao, and Zhenshan Luo. Synthesizing ornamental typefaces. *Comput. Graph. Forum*, 36(1):64–75, 2017.

Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint*, abs/2302.05543, 2023.

ABSTRACT

This appendix provides more details of our work that cannot be expanded in the main text. The main content can be summarized as follows:

1. More Details of ToT Model Selection, appendix A.
   (a) Tree-of-Though Model Selection Prompt, appendix A.1.
   (b) Tree-of-Though Model Selection Case , appendix A.2.
2. More Details of Image Evaluation appendix B.
   (a) ToT-LoRA Evaluation Prompt, appendix B.1.
   (b) LLaVA Evaluation Prompt , appendix B.2.
3. More Details of Dataset appendix C.
4. More Comparison Results with SOTA methods fig. 17.
5. Mixing/Seprating of Glyph and Texture Designer fig. 18.

A MORE DETAILS OF TOT MODEL SELECTION

A.1 TOT MODEL SELECTION TEMPLATE

You know the cultures of different countries and different types of art styles. Identify and analyze aspects of The characteristics of vocabulary, theme, content, Implicit culture, and Reader perception. Then from each aspect of the analysis, you must select an element from the input list that best matches the input prompt. All analyses will be written down and then shared with the group. You then analyze all analyses and output the consensus-selected element or your best guess-matched element. The final selected word MUST be the same as the TEMPLATE:

TEMPLATE:

Selected: [the selected word]

Input list: {search_list}

Input prompt: {input}

A.2 TOT MODEL SELECTION CASE:

Input Prompt: A girl, a boy, in a church

- **Thoughts-1st: Analyzing the given prompt with keywords for selection: "General", "Traditional Art", "Design", "Cartoon", "SCI-FI", "Realistic"**
  - **Vocabulary:** *The vocabulary used in the prompt is quite general and straightforward. It mentions a girl, a boy, and a church, which are common words and do not necessarily indicate a specific art style or culture.*
  - **Theme:** *The theme seems to be centered around a girl and a boy in a church. This could suggest a religious or spiritual theme, which is often depicted in traditional art.*
  - **Content:** *The content is quite simple and does not provide much detail. However, the mention of a church could suggest a religious or spiritual context, which is often associated with traditional art.*
  - **Implicit Culture:** *The mention of a church could suggest a culture where Christianity is prevalent. Many traditional art pieces are influenced by religious themes, especially Christianity.*
  - **Reader Perception:** *The reader might perceive this as a simple scene of a girl and a boy in a church. Depending on their cultural background and personal experiences, they might associate this with traditional art.*
  - **Summarization:** *After analyzing all aspects, the element from the input list that best matches the input prompt is 'Traditional Art'.*
  - **Selected:** *Traditional Art*
- **Thoughts-2nd: Analyzing the given prompt with keywords: "European Style", "Chinese Style" for selection:**

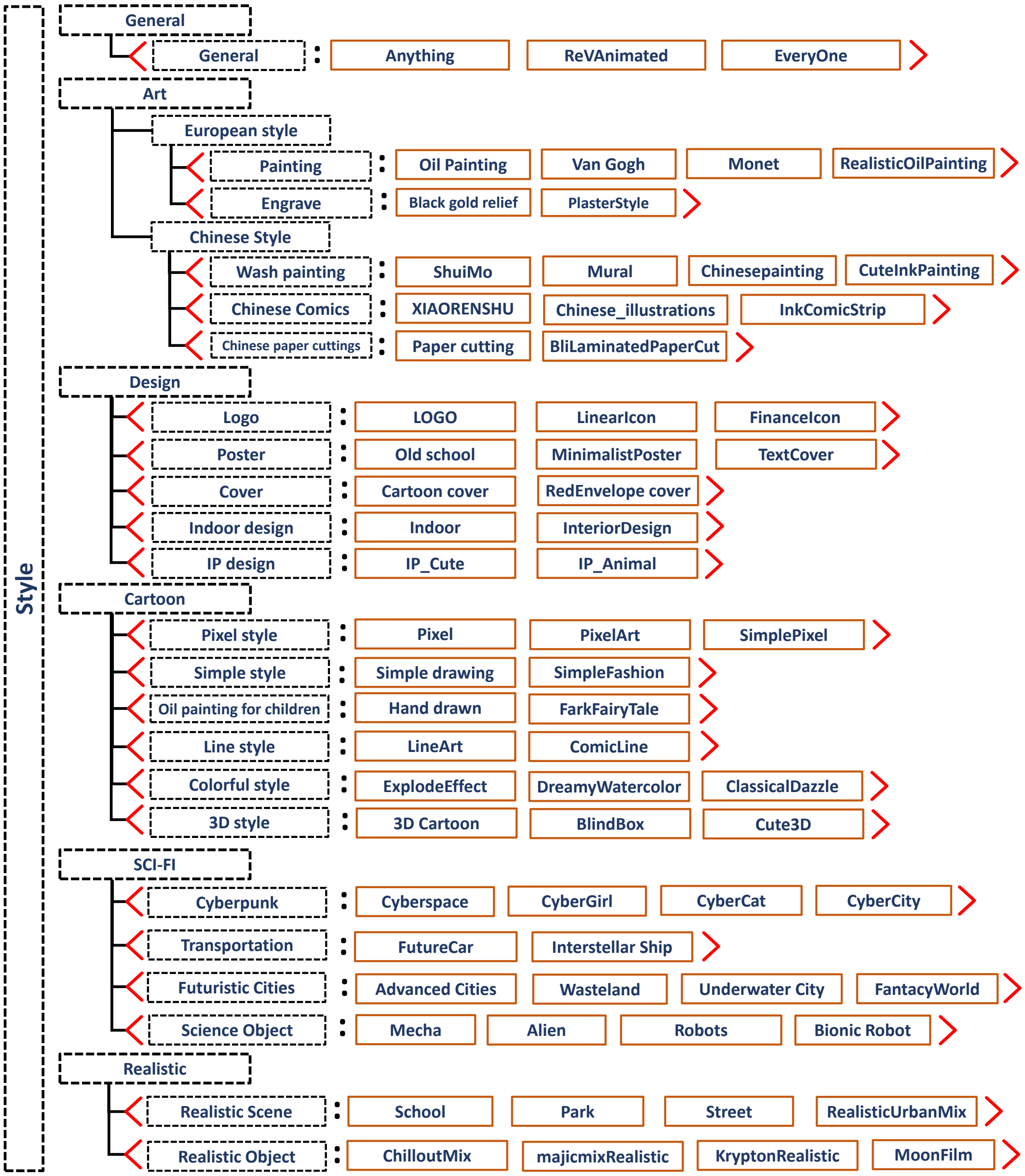

Figure 11: The hierarchical model tree consists of multi-level sub-categories to facilitate finer granularity in **ToT** model selection.

- **Vocabulary:** *The words used are simple and straightforward. They refer to two individuals, a girl and a boy, and a location, a church. The vocabulary does not directly indicate a specific culture or art style.*
- **Theme:** *The theme seems to be about two individuals in a religious setting. This could be interpreted in many ways, but it does not directly indicate a specific culture or art style.*
- **Content:** *The content is minimal and does not provide much context. However, the mention of a church might suggest a more Western or European context, as churches are more commonly associated with Western religions.*
- **Implicit Culture:** *The implicit culture could be inferred from the mention of a church. Churches are more commonly found in European cultures, suggesting a European style.*
- **Reader Perception:** *The reader might perceive this as a scene from a European setting due to the mention of a church. However, this is subjective and could vary depending on the reader's background and experiences.*

- **Summarization:** *After analyzing all aspects, the element from the input list that best matches the input prompt is 'European Style'.*
- **Selected:** *European style*

- **Thoughts-3rd: Analyzing the given prompt with keywords: "Painting", "Engrave" for selection:**
  - **Vocabulary:** *The input prompt uses simple, everyday language. The words "girl," "boy," and "church" are common and straightforward, suggesting a scene that is easy to visualize.*
  - **Theme:** *The theme seems to be centered around a religious or spiritual setting, given the mention of a church. It could also suggest a sense of community or family, as it involves a girl and a boy.*
  - **Content:** *The content is quite minimalistic, focusing on two characters within a specific setting. There's no explicit action or event mentioned, leaving much to the reader's interpretation.*
  - **Implicit Culture:** *The mention of a church suggests a culture where Christianity is prevalent. It could also imply a traditional or conservative society where church attendance is common.*
  - **Reader Perception:** *The reader might perceive this as a peaceful, serene scene. Depending on their personal experiences and cultural background, they might also associate it with feelings of spirituality, community, or nostalgia.*
  - **Summarization:** *all these aspects, the scene seems more suited to a 'Painting' rather than an 'Engrave'. A painting can capture the subtleties of the scene, the characters, and the setting in a more vivid and colorful way.*
  - **Selected:** *Painting*

- **Thoughts-3rd: Analyzing the given prompt with keywords: "Oil Painting", "Van Gogh", "Monet", "RealisticOilPainting" for selection:**
  - **Vocabulary:** *The vocabulary used in the prompt is simple and straightforward. It describes a scene involving two individuals in a specific location. There is no specific artistic term used that could directly link to any of the elements in the input list.*
  - **Theme:** *The theme seems to be a common scene or event, possibly a religious or spiritual one given the setting is a church. This could potentially link to 'Oil Painting' or 'RealisticOilPainting' as these art styles often depict everyday life and common scenes.*
  - **Content:** *The content is a depiction of two individuals in a church. This could be a scene from everyday life, a religious event, or a spiritual gathering. This could potentially link to 'Oil Painting' or 'RealisticOilPainting' as these art styles often depict such scenes.*
  - **Implicit Culture:** *The culture implied here could be a Christian or Western culture, given the setting of a church. This could potentially link to 'Van Gogh' or 'Monet' as these artists were from the Western world and often depicted scenes from their culture.*
  - **Reader Perception:** *The reader might perceive this as a simple, everyday scene or a depiction of a religious event. This could potentially link to 'Oil Painting' or 'RealisticOilPainting' as these art styles often depict such scenes.*
  - **Summarization:** *After analyzing all aspects, the element that best matches the input prompt from the input list is 'Oil Painting'. This is because the prompt describes a scene that could easily be depicted in this art style, and it does not specifically link to any particular artist or a more specific style of oil painting.*
  - **Selected:** *Oil Painting*

B More Details of Image Evaluation

B.1 ToT-LoRA Evaluation Prompt:

"You are now a scoring judge for a generative model. I will provide some images along with prompts generated by the model. Now, assess the relevance of the image content to the prompts and rate the overall quality of these generated images on a scale of 1-10. Reply with the Prompt Relevance scores for the prompt word and image relevance, the Image Quality, and Style and Prompt Match score between the image style and prompt word. Provide only numerical scores without any explanations."

Algorithm 1 Hyperparameter Tuning

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

Figure 12: Effect of Merging LoRAs with Different Weights. The upper case is generated by fusing the **Chinese Bronzes** and **Sketch Style** LoRAs; the rest is by fusing the **Comic** and **Croissant Style** LoRAs.

Figure 13: The Comparison of WordArt texture rendering on the glyph “World Peace" (ControlNet vs. ToT-GPT4-LoRA+ControlNet).

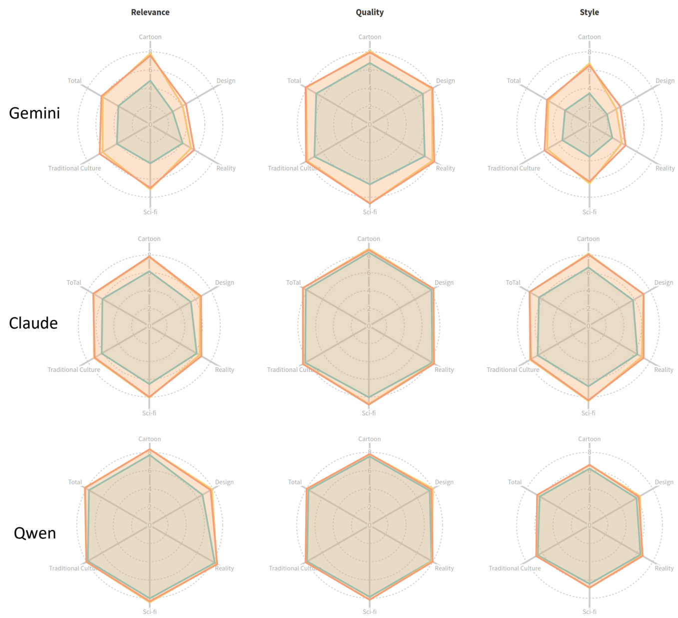

Figure 14: The Comparison of WordArt texture rendering on the glyph “World Peace" (ControlNet vs. ToT-GPT4-LoRA+ControlNet).

B.2 LLaVA Evaluation Prompt:

You have to decouple the <Input sentence> into individual, Real world Targets. The target must be something visible and tangible in the real world. The decoupled Targets MUST be the same as the TEMPLATE:

TEMPLATE:

Targets:{target,target}

<Input sentence>: {input}

Subsequently, we employ an iterative querying manner, systematically verifying the presence of each object via the LLaVA:

Is {target} exist in the photo? Just answer Yes or No.

C More Details of WordArt Dataset

To evaluate the capabilities of our proposed MetaDesigner framework, we assembled a comprehensive dataset comprising 150 distinct prompts specifically tailored for synthesizing WordArt. These prompts are categorized into five broad themes, namely "cartoon", "design", "reality", "sci-fi", and "traditional culture", to cover a wide spectrum of artistic styles and thematic elements. This thematic diversity ensures that our dataset not only supports but also challenges the MetaDesigner in generating WordArt that is both aesthetically pleasing and thematically appropriate across various artistic genres.

Linguistic Diversity and Selection Criteria. In our user study, a carefully curated subset of 20 prompts was deployed to generate WordArt, leveraging the proposed comparison methods. This subset was chosen to represent a broad linguistic palette, including English, Chinese, Japanese, and Korean, to test the MetaDesigner's capacity to handle multilingual input effectively. The selection criteria for these prompts were based on their ability to represent the thematic and linguistic diversity of the larger dataset, as well as their potential relevance to real-world applications in e-commerce, education, and digital content creation.

Statistics and Analysis. Statistically, the dataset is balanced across the five themes, with each theme comprising 30 prompts. This ensures an equitable representation of styles and thematic content, facilitating a comprehensive evaluation of the MetaDesigner's performance. The prompts were

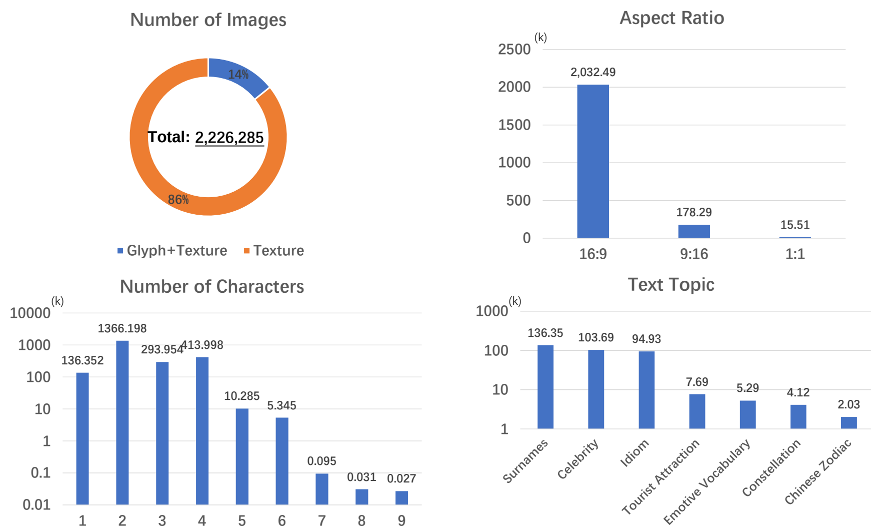

Figure 15: Numerical analyses on our WordArt dataset. The x-axis of the histogram denotes the number of images, measured in thousands, whereas the y-axis represents the specific categories.

Figure 16: Numerical analyses on user preference. The x-axis of the histogram denotes the number of images, measured in thousands, whereas the y-axis represents the specific categories. "0" in "Number of Characters" part denotes that the user uploaded a mask image instead of inputting texts.

selected based on criteria such as cultural significance, popularity in contemporary digital art, and relevance to potential commercial and educational applications.

Applicability and Use Cases. The versatility of our dataset is further demonstrated through an extensive selection of cases, highlighted in Figure 10. These examples include but are not limited to, cartoon text, Chinese surnames, and Chinese solar terms, showcasing the dataset's ability to support a wide range of artistic and linguistic expressions. Such diversity not only validates the MetaDesigner's efficacy in generating contextually and culturally relevant WordArt but also underscores the dataset's broad applicability across various domains. From e-commerce promotions leveraging customized WordArt to enhance consumer engagement, to educational settings where multilingual WordArt facilitates language learning and cultural education, the dataset serves as a foundational resource for exploring the intersection of AI, art, and language.

WordArt Dataset Analysis. We have conducted a series of numerical analyses on our WordArt dataset, as illustrated in Figure 15. We have generated over **two million** images in our "Image Plaza", encompassing various image ratios, word counts, and topics. Images in the dataset can be retrieved on our WordArt space[2].

We have also conducted numerical analyses on user preferences, as depicted in Figure 16. To date, users have created more than one million images on our space. The majority of users show a preference for an image aspect ratio of 16:9, typically input text consisting of no more than five characters. Furthermore, users exhibit a marked preference for images that feature an alpha channel.

Figure 17: More exemplar results of the comparison methods.

[2]The anonymity rule might have been violated, so the link is not provided

Table 3: **VLM evaluation**. We present the results from VLM evaluation conducted in two dimensions, “Aesthetics” and “Creativity”. All values range from 0 to 10 and higher is better. Ours* denotes our methods without agents.

Evaluation Dimension	SDXL	SDXL-Augmented	TextDiffuser	TextDiffuser-2	Anytext	DALLE 3	Ours*	Ours
Aesthetics	7.9	8.1	6.1	6.4	7.0	8.5	8.3	**8.7**
Creativity	7.5	7.7	5.8	6.0	6.1	7.6	7.7	**7.9**

Figure 18: Visual Comparison Between mixing/separating of Glyph Designer and Texture Designer.