Prompt2Poster: Automatically Artistic Chinese Poster Creation from Prompt Only

Anonymous Author(s)

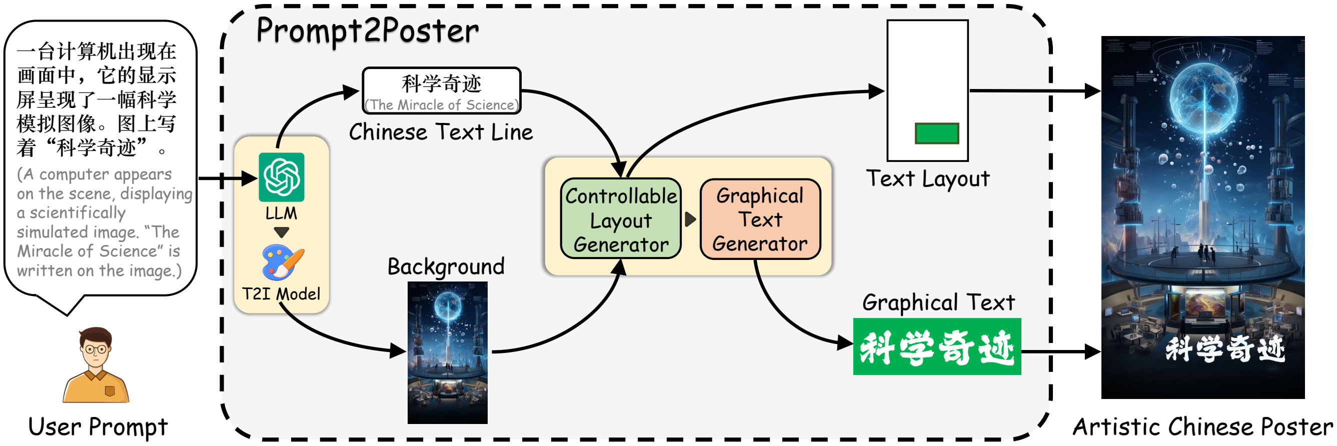

一台计算机出现在画面中，它的显示屏呈现了一幅科学模拟图像。图上写着“科学奇迹”。
(A computer appears on the scene, displaying a scientifically simulated image. "The Miracle of Science" is written on the image.)
User Prompt
Prompt2Poster
LLM
T2I Model
科学奇迹
(The Miracle of Science)
Chinese Text Line
Background
Controllable Layout Generator
Graphical Text Generator
Text Layout
Graphical Text
科学奇迹
科学奇迹
Artistic Chinese Poster

Figure 1: Overview. Prompt2Poster creates artistic Chinese posters with harmonious visual effects and stylized graphical text from a given prompt only. Prompt2Poster first extracts user intention and generates the aligned background with large models, and then creates layouts and graphical texts with designed modules, leading to correct and pleasing visual results.

ABSTRACT

As a critical component in graphic design, artistic posters are widely applied in the advertising and entertainment industry, thus the automatic poster creation from user-provided prompts has become increasingly desired recently. Although existing Text2Image methods create impressive images aligned with given prompts, they fail to generate ideal artistic posters, especially posters with Chinese texts. To create desired artistic Chinese posters including an aligned background, reasonable layouts, and stylized graphical texts from given prompts only, we propose an automatic poster creation framework, named **Prompt2Poster**. Our framework first utilizes the capacity of the powerful Large Language Model (LLM) to extract user intention from provided prompts and generate the aligned background. For the harmonious layout and graphical text generation, we propose **Controllable Layout Generator (CLG)** and **Graphical Text Generator (GTG)** modules that both leverage sufficient multi-modal information, leading to accurate and pleasurable visual results. Comprehensive experiments demonstrate that our Prompt2Poster achieves superior performance especially on text quality and visual harmony than existing poster creation methods. Our codes will be released after the paper review.

CCS CONCEPTS

• **Computing methodologies** → *Computer vision tasks*; *Natural language processing*.

KEYWORDS

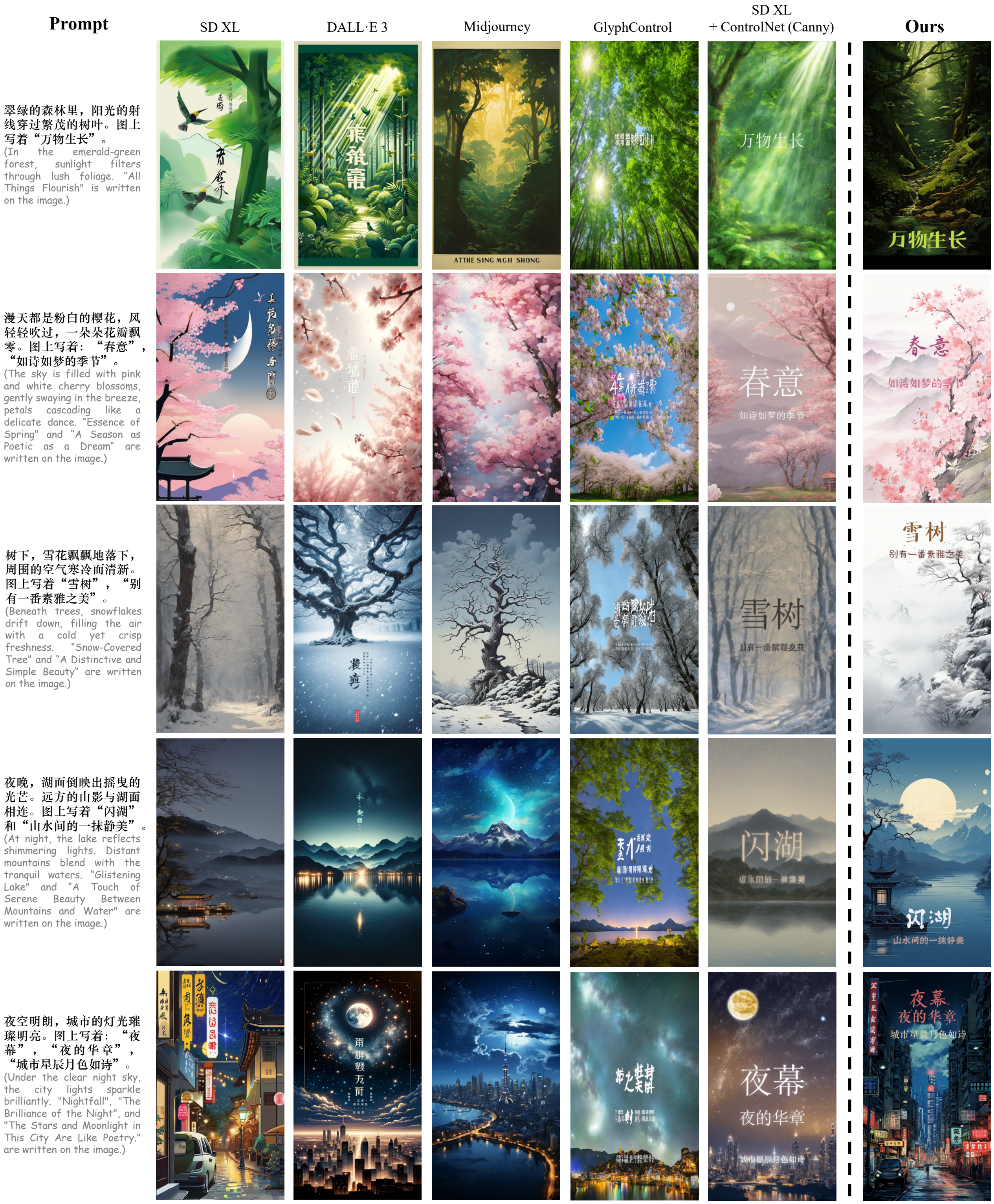
Text to Image Generation, Layout Generation, Stylized Font Generation, Large Language Models

1 INTRODUCTION

Designing and integrating visually appealing backgrounds and Chinese graphical texts for artistic Chinese posters represent a critical task in visual information presentation. Traditional Chinese poster creation requires professional designers to invest a significant amount of time. In recent years, some methods [10, 22, 31, 32] have been proposed to partly automate the poster creation through multistep division. However, these methods involve manual operations in some steps, such as selecting existing backgrounds or element attributes, still leading to the necessity of expensive professional labor. Intuitively, obtaining a poster only from a user-provided prompt describing the desired artistic effect would align naturally with human interaction instincts. This significantly streamlines the design process, enhancing creation efficiency, and saving extensive manual labor. Thus, such a prompt-guided automatic poster creation framework has gradually caught greater attention for high interaction, efficiency, and economy.

With the proposal and subsequent developments of Denoising Diffusion Probabilistic Model [3, 12, 25–28], Large Text-to-Image (T2I) Diffusion Models have demonstrated remarkable performance in generating aesthetically pleasing and style-diverse images by user prompts. Some fine-tuning or plus versions of these T2I Diffusion models [3, 12, 25–28], with linguistic and optionally graphical control through LoRA [14] and ControlNet [37], have achieved state-of-the-art for prompt-guided poster generation.

Despite performing excellently in prompt-guided poster generation, limited by the diffusion framework [3, 12, 25–28], these methods face a critical problem of data distribution mixing. For an artistic Chinese poster, the background and Chinese graphical texts



翠绿的森林里，阳光的射线穿过繁茂的树叶。图上写着“万物生长”。
(In the emerald-green forest, sunlight filters through lush foliage. “All Things Flourish” is written on the image.)

漫天都是粉白的樱花，风轻轻吹过，一朵朵花瓣飘零。图上写着：“春意”，“如诗如梦的季节”。
(The sky is filled with pink and white cherry blossoms, gently swaying in the breeze, petals cascading like a delicate dance. “Essence of Spring" and “A Season as Poetic as a Dream“ are written on the image.)

树下，雪花飘飘地落下，周围的空气寒冷而清新。图上写着“雪树”，“别有一番素雅之美”。
(Beneath trees, snowflakes drift down, filling the air with a cold yet crisp freshness. “Snow-Covered Tree" and “A Distinctive and Simple Beauty“ are written on the image.)

夜晚，湖面倒映出摇曳的光芒。远方的山影与湖面相连。图上写着“闪湖”和“山水间的一抹静美”。
(At night, the lake reflects shimmering lights. Distant mountains blend with the tranquil waters. “Glistening Lake” and “A Touch of Serene Beauty Between Mountains and Water” are written on the image.)

夜空明朗，城市的灯光璀璨明亮。图上写着：“夜幕”，“夜的华章”，“城市星辰月色如诗”。
(Under the clear night sky, the city lights sparkle brilliantly. "Nightfall", "The Brilliance of the Night", and "The Stars and Moonlight in This City Are Like Poetry.” are written on the image.)

Figure 2: Artistic Chinese posters generated by our Prompt2Poster and other poster generation methods, including SD XL [25], DALL·E 3 [1], Midjourney, GlyphControl [36], and SD XL + ControlNet (Canny) [37]. The leftmost column represents the prompts, while the remaining six columns display the images generated by different methods. Visual results demonstrate that Prompt2Poster achieves more accurate and pleasurable artistic Chinese poster creation than the compared methods.

are extremely different in terms of data distribution. Most of these prompt-guided poster generation methods try to obtain a poster containing the two different distributions from a single distribution, largely weakening their data fitting for the Chinese graphical texts. Thus, they fail to promise the visual correctness and style diversity of Chinese graphical texts on the posters, which are the keys to artistic Chinese poster creation.

To guarantee the fitting of the two different data distributions, we propose a novel poster generation framework, named **Prompt2-Poster**, shown in Fig. 1. According to Fig. 1, although only fed into linguistic control from the user prompt, inner multi-modal information is fully utilized to catch the two different data distributions. Specifically, a Large Language Model (LLM) [20] first extracts various information from the user-provided prompt. Then, a pre-trained Large Text-to-Image (T2I) Diffusion Model with high fitting for the poster background data distribution, generates a high-quality background based on the extracted information. After that, a novel **Controllable Layout Generator (CLG)**, is proposed to harmoniously decide the placement for Chinese graphical texts on the background, based on both the visual and linguistic information, coming from the LLM and T2I model, respectively. The output text layout from CLG contains the placement for Chinese graphical texts. This geometrical text layout, along with previous visual and linguistic information is further sent to a novel proposed **Graphical Text Generator (GTG)**. GTG utilizes the three modal information and continuous spaces to generate correct and style-diverse Chinese graphical texts from the graphic data distribution. A high-quality artistic Chinese poster is conveniently finished after integrating stylized Chinese graphical texts with the background image based on the text layout. As shown in Fig. 2, our Prompt2Poster has a much stronger ability to generate correct and diverse Chinese graphical texts on the poster with a specific prompt, for treating the two different distributions separately. Extensive experiments show the artistic superiority of our Prompt2Poster, CLG, and GTG for artistic Chinese poster generation.

Our contribution can be summarized as follows:

- We introduce Prompt2Poster, a unique framework designed specifically for the automatic creation of artistic Chinese posters. This system fully leverages multi-modal information to accommodate different data distributions.
- A Controllable Text Layout Generator, named CLG, is proposed to produce text layout according to both the background image and Chinese texts, promising harmony between background and Chinese graphical texts.
- We've developed an all-new Graphical Text Generator (GTG) that operates initially in the continuous graphic feature space, aside from predicting color in pixel space. This technique not only maintains text correctness but also enhances the artistic style diversity of Chinese graphical texts.
- Through extensive experimentation, we have been able to demonstrate the effectiveness of our Prompt2Poster, along with its integral modules CLG and GTG, in the creation of artistic Chinese posters. The results highlight the capability of our system to reliably produce high-quality posters, reinforcing the value and potential of our approach.

2 RELATED WORK

Prompt-Guided Poster Generation. With the rapid development of Large Text-to-Image (T2I) Diffusion Models [3, 12, 25–28], T2I-Diffusion-based poster systems have been widely applied to prompt-guided poster generation for their huge interaction, efficiency and economy. By adding linguistic and optional graphical control by LoRA [14] and ControlNet [37], they return an attractive poster from the noise. However, generating visually correct and various Chinese graphical texts on posters remains a significant challenge for general T2I methods. Recent poster-specific generation methods including AutoPoster [22], Desigen[33], and Visual Layout Composer[29] are proposed, which achieve remarkable creation results for artistic posters. However, their methods most focus on component layout planning thus requiring visual elements as input necessarily, resulting in limitations when users desire to create posters from scratch. We highlight that our Prompt2Poster is a fully automatic and generative framework for artistic Chinese poster creation, which achieves correct and pleasing Chinese posters from given prompts only.

Controllable Text Layout Generation. The core of the previous controllable text layout generation models [5, 10, 13, 31, 38] is to predict the placement of graphical texts which is visually harmonious and do not cover the semantic area of the background. Traditional methods [10, 31] take template or optimization strategies while deep-learning-based methods [5, 13, 38] utilize Condition-GAN [9] to achieve better results. However, only receiving the background image, all previous advanced learning-based methods [5, 13, 38] are not able to consider Chinese texts, leading to their complete incompatibility with our Prompt2Poster.

Stylized Graphical Text Generation. Following most poster systems [10, 22, 31, 32], in our Prompt2Poster, the stylized attributes of Chinese graphical texts contain the font and color. For font design, previous works primarily concentrate on selecting the most suitable font from a feasible discrete font space, by heuristic method [6] or network learning [22, 31]. The selection in discrete space limits the font variation of Chinese graphical texts. In the realm of color design, some works [15, 31] predict colors based on color rules [16, 30], while the recent TextPainter [8] utilizes a color style encoder to extract color features and use them to reconstruct a visual graphical text.

3 METHOD

In this section, we first introduce the overview of our automatic artistic Chinese poster creation framework **Prompt2-Poster** in Sec. 3.1. We then describe the design details of two proposed modules including Controllable Layout Generator (CLG) and Graphical Text Generator (GTG) in the following Sec. 3.2 and Sec. 3.3, respectively. Finally, the loss function and optimization details are given in Sec. 3.4. The overview illustration is shown in Fig .1.

3.1 Overview of Prompt2Poster

Given a user prompt $\boldsymbol{y}$, Prompt2Poster aims to automatically create an artistic Chinese poster that satisfies all requirements described by the $\boldsymbol{y}$. Concretely, Prompt2Poster firstly processes the prompt $\boldsymbol{y}$ through a Large Language Model (LLM) [20] to extract the style and content information $\boldsymbol{s}$ of the desired poster background, as well as the Chinese text, set $\boldsymbol{T}$:

$$\boldsymbol{s}, \boldsymbol{T} = \mathrm{LLM}(\boldsymbol{y}). \tag{1}$$

Here, LLM$(\cdot)$ can be various powerful models, such as GPT series[4, 23] and PaLM series [7, 19, 24]. $\boldsymbol{T}$ is the text set containing M lines of Chinese texts which are expected to appear on the poster graphically:

$$\boldsymbol{T} = [\boldsymbol{t}_1, \boldsymbol{t}_2, ..., \boldsymbol{t}_M], \quad \boldsymbol{t}_i = [\boldsymbol{c}_1^i, \boldsymbol{c}_2^i, ..., \boldsymbol{c}_{N_i}^i], \tag{2}$$

where $\boldsymbol{t}_i$ is the i-th Chinese text line consisting of N_i Chinese characters defined by the user, and $\boldsymbol{c}_j^i$ is the j-th character in the i-th text line.

According to the extracted background information $\boldsymbol{s}$, Prompt2-Poster leverages the Large Text-to-Image (T2I) Diffusion Model to create the poster background $\boldsymbol{I}$ with high quality and diversity:

$$\boldsymbol{I} = \mathrm{T2I}(\boldsymbol{s}). \tag{3}$$

We underline that T2I$(\cdot)$ can be various powerful Text-to-Image models, such as Stable Diffusion series [25, 27], DALL·E series [1, 26], or Midjourney, which are high fitting of the data distribution for poster backgrounds.

After obtaining the background $\boldsymbol{I}$, Prompt2Poster generates harmonious and elegant text layout $\boldsymbol{L} = [\boldsymbol{l}_1, \boldsymbol{l}_2, ..., \boldsymbol{l}_M]$ based on the $\boldsymbol{I}$ and $\boldsymbol{T}$. Within this, $\boldsymbol{l}_i = [x_i, y_i, w_i, h_i]$ is the i-th layout element, containing the geometrical placement of the i-th Chinese text line $\boldsymbol{t}_i$. Specifically, $[x_i, y_i]$ is the center location of the i-th Chinese graphical text and $[w_i, h_i]$ is the size. The process of text layout generation can be formulated as:

$$\boldsymbol{L} = \mathrm{CLG}(\boldsymbol{I}, \boldsymbol{T}), \tag{4}$$

where CLG$(\cdot)$ is our designed Controllable Text Layout Generator, which will be further introduced in Sec. 3.2.

With $\boldsymbol{I}, \boldsymbol{T}$ and $\boldsymbol{L}$, we propose a Stylized Graphical Text Generator to create precise and diverse stylized Chinese graphical texts, different from the poster background in terms of data distribution. Each Chinese text line is graphically generated, separately:

$$[\boldsymbol{g}_i, \boldsymbol{o}_i] = \mathrm{GTG}(\boldsymbol{I}, \boldsymbol{t}_i, \boldsymbol{l}_i), \tag{5}$$

where GTG$(\cdot)$ is our Graphical Text Generator introduced in Sec. 3.3, and $\boldsymbol{g}_i, \boldsymbol{o}_i$ are the font and color of the i-th Chinese graphical text line, respectively.

Finally, according to the text layout $\boldsymbol{L}$, Prompt2Poster integrates the background image $\boldsymbol{I}$ and Chinese graphical texts together through geometrical pasting. In our Prompt2Poster, although starting from only the linguistic modal $\boldsymbol{y}$, visual and geometrical information $\boldsymbol{I}$ and $\boldsymbol{L}$ also play an important role in the poster generation, which fully utilizes the inner multi-modal information. Through separately generating backgrounds and Chinese graphical texts with different data distributions, Prompt2Poster can obtain overall harmonious and visually appealing artistic Chinese posters with satisfying graphical texts.

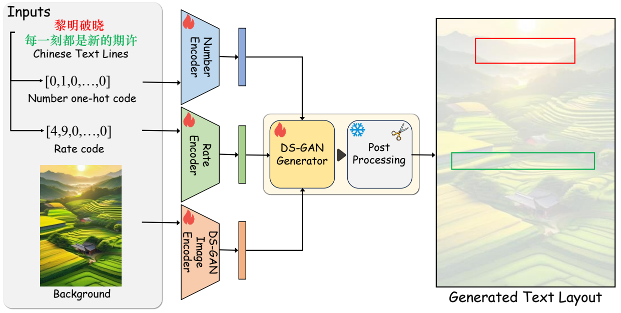


Figure 3: The overview of our Controllable Layout Generator. Trainable NE, RE, and fixed PC are added based on the DS-GAN Generator and DS-GAN Image Encoder.

3.2 Controllable Layout Generator

To guarantee harmony between the background and Chinese graphical texts, how to place the graphical texts on the background elegantly is extremely critical. Serving as a module in our Prompt2-Poster, our Controllable Layout Generator (CLG) determines the placement of Chinese graphical texts by predicting their center locations and sizes on the background image. Previous state-of-the-art controllable text layout models [5, 13, 38] are all designed for generating elements only according to the background $\boldsymbol{I}$. To make the text layout generation controlled by both the background and Chinese texts, further acceptance of linguistic information in our CLG is necessary. This is achieved by adding trainable **Number Encoder** (NE), **Rate Encoder** (RE), and fixed **Post-processing Cropping module** (PC) based on an advanced layout model DS-GAN. Besides considering background $\boldsymbol{I}$, CLG promises the text layout $\boldsymbol{L}$ completely matches the Chinese text set $\boldsymbol{T}$ in terms of element number and text length [13].

The generation process of CLG is shown in Fig. 3. A structurally similar discriminator is also designed for training, and more training details are introduced in the supplementary material. Specifically, for a text set $\boldsymbol{T}$ containing M lines of Chinese texts, NE receives an x dimension one-hot layout number vector $\boldsymbol{H} = (h_1, h_2, ..., h_x)$ and RE takes a same dimension rate vector $\boldsymbol{R} = (r_1, r_2, ..., r_x)$. x is a predefined large number that $M \leq x$. The value of $h_i (1 \leq i \leq x)$ is 1 if $i = M$ else 0, and $r_i = N_i$. Only the first M elements in $\boldsymbol{R}$ have corresponding rate values, and other elements are all zero. Through number control from NE, when there are M lines of Chinese texts in $\boldsymbol{T}$, CLG is endowed with the ability to produce text layout $\boldsymbol{L}$ with M elements exactly. As each graphical character in the same text captures the same geometrical square space [11], introduced in Sec. 3.3, the size rate of the i-th element w_i/h_i in $\boldsymbol{L}$ must be equal to i_N. With RE, the origin DS-GAN generator can directly generate an estimated text layout $\boldsymbol{L}^*$, satisfying w_i/h_i very close to i_N. Then, PC transforms the $\boldsymbol{L}^*$ to $\boldsymbol{L}$ whose size rate of element completely matches the text length $(w_i/h_i = N_i)$, which can be expressed as:

$$\mathrm{PC}(\boldsymbol{L}^*) = \boldsymbol{L} = (\boldsymbol{l}_1, \boldsymbol{l}_2, ...\boldsymbol{l}_M), \tag{6}$$

$$l_i^t = \begin{cases} (x_i^*, y_i^*, h_i^*, h_i^* \times N_i), \text{if } w_i^*/h_i^* > N_i, \\ (x_i^*, y_i^*, w_i^*, w_i^*/N_i), \text{else}, \end{cases} \tag{7}$$

where x_i^*, y_i^* and w_i^*, h_i^* is the estimated center location and size of the i-th Chinese graphical text. Upon processing by PC, the text layout $\boldsymbol{L}$ thoroughly takes into account both the background image and Chinese text lines. This approach ensures the overall aesthetics and harmony of the final poster.

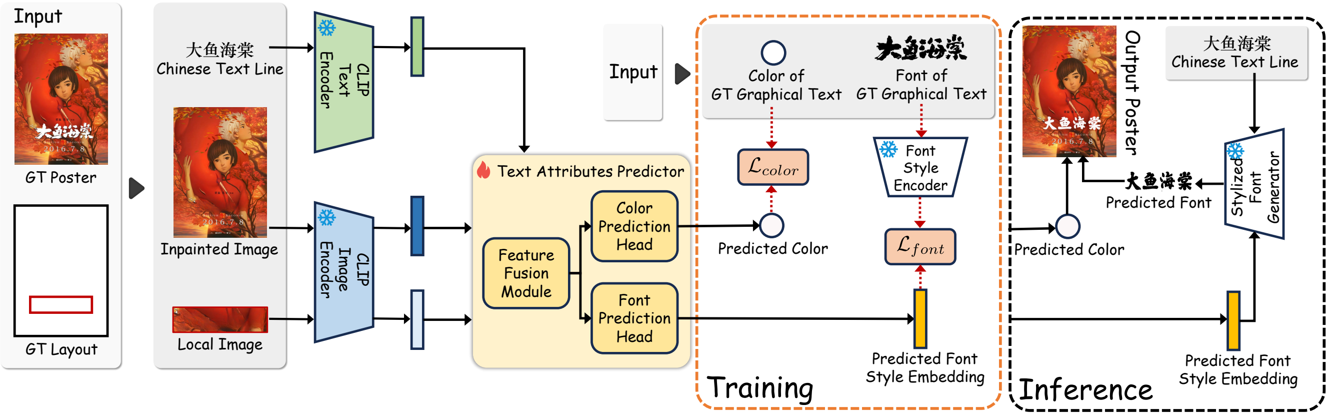
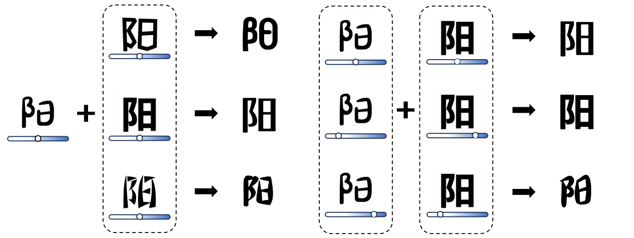


Figure 4: The overview of our Graphical Text Generator. The CLIP-based Text Attributes Predictor predict the color and font style embedding which is then used by the Stylized Font Generator to create the font. These elements collectively determine the Chinese graphical texts on the output poster.



(a) Unlike pairs but same weight. (b) Same pair but unlike weights.

Figure 5: Interpolation of style embeddings. For the Chinese character, each graphical version corresponds to a style embedding in the feature space of the font style encoder in Diff-Font. So adjusting the weights of two styles can result in unlike graphical versions.

3.3 Graphical Text Generator

For an artistic Chinese poster, the correctness and style diversity of Chinese graphical texts are extremely crucial. The most critical weakness of previous stylized models is that they predict font in discrete space, leading to loss of font diversity [6, 22, 31]. Our proposed Graphical Text Generator (GTG) generates stylized Chinese graphical texts with diverse fonts and colors while guaranteeing their correctness. Thus, as the main improvement, we utilize the continuous feature space of Diff-Font [11] $\mathcal{G}_f$ to generate Chinese graphical texts character-wise, and each graphical Chinese character in the same text line captures the same geometrical square space. Specifically, $\mathcal{G}_f$ takes a font style encoder [34] $\mathcal{E}_f$ to create a continuous feature style space. Each style embedding in the space determines a specific font of the Chinese graphical character. For the same graphical character, any interpolation of two style embeddings brings a final font style in between, shown in Fig. 5.

Therefore, we can acquire Chinese graphical characters with unlimited font styles theoretically, by first predicting style embeddings $\boldsymbol{e}$ and then getting the fonts with $\mathcal{G}_f$ based on $\boldsymbol{e}$:

$$font = \mathcal{G}_f(\boldsymbol{e})\,. \tag{8}$$

This highly enriches the font diversity compared with directly selecting limited font attributes in discrete space [6, 22, 31]. Following previous works [15, 31], the color prediction is directly in continuous pixel space. Noticing that $\mathcal{G}_f$ is a state-of-the-art font generation model, using which in our GTG naturally guarantees high correctness of generated Chinese graphical texts.

As shown in Fig. 4, the prediction of font attribute is achieved through modal alignment with $\mathcal{E}_f$, and the prediction of color attribute is accomplished through training from scratch. Specifically, the generation of the i-th font $\boldsymbol{g}_i$ is determined by the Chinese text line $\boldsymbol{t}_i$, background image $\boldsymbol{I}$, and layout $\boldsymbol{l}_i$. Since the local area has a strong influence on the style, we further obtain the local background $\boldsymbol{I}_i$ based on $\boldsymbol{l}_i$. Then, the pre-trained Text Encoder in Chinese CLIP [35] is used to encode $\boldsymbol{t}_i$ to f_i^t, and Image Encoder in the same space is utilized to encode $\boldsymbol{I}$ and $\boldsymbol{I}_i$ to f_i^I and f_i^l. Subsequently, we concatenate all the encoded features as the input feature $\boldsymbol{f}_i = \mathrm{concat}(f_i^t, f_i^I, f_i^l)$ for prediction.

To predict the corresponding style embedding $\boldsymbol{e}_i$ and color value $\boldsymbol{o}_i$ from concatenated feature $\boldsymbol{f}_i$. We propose a Text Attributes Predictor (TAP), which incorporates a Feature Fusion Module (FFM) consisting of multiple MLPs, a linear Font Prediction Head ($\mathcal{H}_e$) and a linear Color Prediction Head ($\mathcal{H}_o$) for predicting the style embedding $\boldsymbol{e}_i$ and color value $\boldsymbol{o}_i$ respectively:

$$[\boldsymbol{e}_i, \boldsymbol{o}_i] = \mathrm{TAP}(\boldsymbol{f}_i) = \left[\mathcal{H}_e(\mathrm{FFM}(\boldsymbol{f}_i)), \mathcal{H}_o(\mathrm{FFM}(\boldsymbol{f}_i))\right]. \tag{9}$$

The font of the i-th Chinese graphical text $\boldsymbol{g}_i$ is then obtained through $\mathcal{G}_f$ character-wise, based on their shared style embedding.

3.4 Loss Function and Optimization

We underline that the trainable module in Prompt2Poster is the Controllable Layout Generator (CLG) and the Graphical Text Generator (GTG). Both generators are supervised separately under corresponding losses, and combined with other proposed components as Prompt2Poster framework after training for automatic artistic Chinese poster creation.

The training of CLG is inherited from the basic DS-GAN [13], with a composition loss consisting of an adversarial component, a reconstruction component, and general losses including NLL, L1, and IoU. For the training of GTG, two losses are introduced for supervising the font style and font color respectively. Noticing that the predictions for i-th line Chinese texts are style embedding $\boldsymbol{e}_i$ and color value $\boldsymbol{o}_i$. Concretely, we leverage the Kullback-Leibler (KL) divergence as the font loss $\mathcal{L}_{font}$ to distill the style distribution of real fonts into CTG, and $\mathcal{L}_{font}$ can be formulated as:

$$\begin{aligned}\mathcal{L}_{font} &= D_{KL}\left(\sigma(\boldsymbol{e}_i/t)\,||\,\sigma(\boldsymbol{e}_i^{gt}/t)\right)\\ &= \int \sigma(\boldsymbol{e}_i(x)/t)\log\left(\frac{\sigma(\boldsymbol{e}_i(x)/t)}{\sigma(\boldsymbol{e}_i^{gt}(x)/t)}\right)dx,\end{aligned} \tag{10}$$

where $\boldsymbol{e}_i^{gt}$ is the ground-truth style embedding of i-th line, $\sigma(\cdot)$ is the softmax function and t serves as the temperature factor to scale the feature distributions. In addition, we employ Mean Squared Error (MSE) as color value loss $\mathcal{L}_{color} = ||\boldsymbol{o}_i - \boldsymbol{o}_i^{gt}||^2$ to supervise the value between $\boldsymbol{o}_i$ and ground-truth $\boldsymbol{o}_i^{gt}$ directly.

Due to the significant difference in the scales of $\mathcal{L}_{font}$ and $\mathcal{L}_{color}$, we use uncertainty to balance multiple loss functions [17, 21], thus GTG is optimized as:

$$\mathcal{L}_{GTG} = \frac{1}{2\gamma_1^2}\mathcal{L}_{font} + \frac{1}{2\gamma_2^2}\mathcal{L}_{color} + \mu\left[\log\left(1+\gamma_1\right) + \log\left(1+\gamma_2\right)\right], \tag{11}$$

where γ_1 and γ_2 are learnable parameters, while μ is the weight to scale the regularization term.

4 EXPERIMENTS

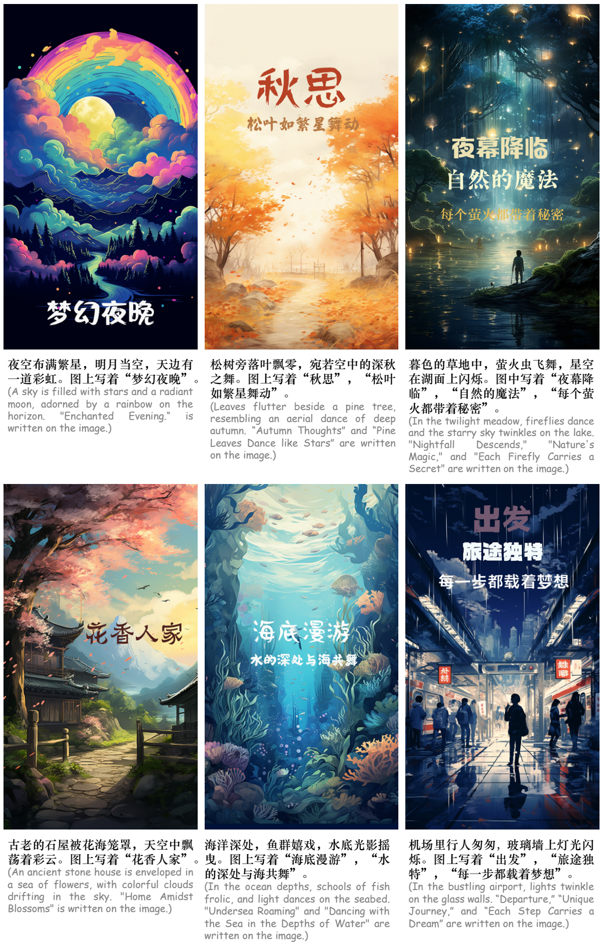
In this section, we present both quantitative and qualitative results to demonstrate proposed Prompt2Poster achieves superior artistic Chinese poster creation performance. Furthermore, extensive ablations are provided to certify the effectiveness of designed components in our Controllable Layout Generator and Graphical Text Generator. The implementation details are omitted due to the length limitation and can be found in our Appendix.

4.1 Comparison

We compared our Prompt2Poster with other five advanced prompt-guided poster generation methods, including SD XL [25], DALL·E 3 [1], Midjourney, GlyphControl [36], and SD XL + ControlNet (Canny) [37] for artistic Chinese poster creation. Following previous works [10, 22, 31, 32], we evaluate an artistic Chinese poster produced by the above methods from three aspects, including text quality, visual harmony, and overall attraction.

They are mainly determined by the visual correctness and style diversity of Chinese graphical texts, the matching between the background and Chinese graphical texts, and the overall artistic attraction to humans, respectively.

Since these artistic metrics are subjective, we conduct a user study to provide an intuitive quantification. Specifically, 41 people of various ages, gender, and occupation are invited to evaluate 30 groups' posters. Each group contains 6 posters which are generated by the above six methods from the same prompt. For the 6 posters in the same group, each participant compares and scores them in terms of the above three metrics respectively. It means that a poster will receive three different scores from a participant, corresponding to the three metrics respectively. We set all the scores ranging from 0 to 10, and a higher score means the poster is better in the degree of a specific metric. Under the setting, we collect a total of 22140 ratings, bringing sufficient data support for the comparison. According to [10, 22, 31, 32], for each metric, instead of directly calculating the average scores for different methods, we statistic the times that the methods receive the highest score. The results are visually displayed



夜空布满繁星，明月当空，天边有一道彩虹。图上写着“梦幻夜晚”。
(A sky is filled with stars and a radiant moon, adorned by a rainbow on the horizon. "Enchanted Evening." is written on the image.)

松树旁落叶飘零，宛若空中的深秋之舞。图上写着“秋思”，“松叶如繁星舞动”。
(Leaves flutter beside a pine tree, resembling an aerial dance of deep autumn. “Autumn Thoughts” and “Pine Leaves Dance like Stars” are written on the image.)

暮色的草地中，萤火虫飞舞，星空在湖面上闪烁。图中写着“夜幕降临”，“自然的魔法”，“每个萤火都带着秘密”。
(In the twilight meadow, fireflies dance and the starry sky twinkles on the lake. "Nightfall Descends," "Nature's Magic," and "Each Firefly Carries a Secret" are written on the image.)

古老的石屋被花海笼罩，天空中飘荡着彩云。图上写着“花香人家”。
(An ancient stone house is enveloped in a sea of flowers, with colorful clouds drifting in the sky. "Home Amidst Blossoms" is written on the image.)

海洋深处，鱼群嬉戏，水底光影摇曳。图上写着“海底漫游”，“水的深处与海共舞”。
(In the ocean depths, schools of fish frolic, and light dances on the seabed. "Undersea Roaming" and "Dancing with the Sea in the Depths of Water" are written on the image.)

机场里行人匆匆，玻璃墙上灯光闪烁。图上写着“出发”，“旅途独特”，“每一步都载着梦想”。
(In the bustling airport, lights twinkle on the glass walls. “Departure,” “Unique Journey,” and “Each Step Carries a Dream” are written on the image.)

Figure 6: Prompt2Poster can generate artistic posters with Chinese texts in different row numbers.

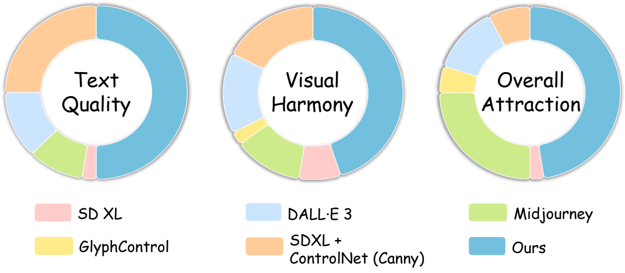


Figure 7: Pie chart results of user studies. Prompt2Poster achieves superior human preference in different aspects. Detailed values are provided in our Appendix.

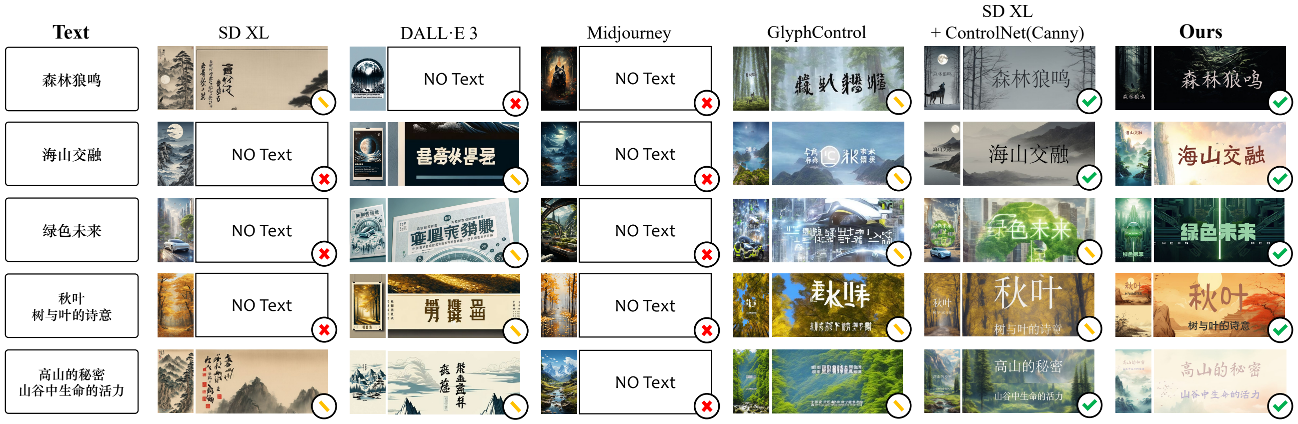


Figure 8: Correctness comparison for the six methods of generating Chinese graphical texts on the posters. ✓ indicates no errors in the poster, ✎ indicates the presence of errors, and ⊗ indicates the failure to generate any required Chinese graphical texts. Except for our method, errors appear on all other methods and some of them even fail to generate graphical texts.

through three pie charts, shown in Fig. 7. According to Fig. 7, for each metric, our Prompt2Poster all has a much higher frequency of receiving the highest score compared with any other methods. Concretely, for the occurrence percentage of achieving the best text quality and the best overall attraction, we are both nearly twice as much as the second-best method. For the occurrence percentage of achieving the best visual harmony, we are even close to three times as much as the second-best method. These indicate the great power of our Prompt2Poster for artistic Chinese poster generation. More details about the design of the user study are listed in the supplementary materials. Unlike other kinds of poster creation, the correctness of Chinese graphical texts is the foundation of artistic Chinese poster creation. Thus, a further comparison between our Prompt2Poster and other methods in terms of the correctness of Chinese graphical texts is conducted. Given some user prompts, all the methods generate corresponding posters and we only focus on the correctness of Chinese graphical texts, shown in Fig. 8.

According to Fig. 8, except for SD XL + ControlNet (Canny) and ours, the other four methods all fail to generate satisfying artistic Chinese posters in terms of correctness. Despite SD XL + ControlNet (Canny) having a better performance than the above four methods, it's still inferior to our Prompt2Poster which has a stronger correctness guarantee. These demonstrate that separately generating backgrounds and graphical texts from different distributions bring better fitting for the graphical data, enhancing the graphical correctness. As shown in Fig. 6, Prompt2Poster can generate posters with Chinese texts in different row numbers, incorporating elegant Chinese graphical texts. This is realized by a multi-modal framework design, which fully utilizes inner multi-modal information and completely exploits the potential of each module in the framework.

4.2 Ablation Study

Controllable Layout Generator. To verify the design effectiveness of our CLG, we also propose another two controllable text layout models, by removing modules in CLG. Specifically, we first remove module RE to get CLG (w/o RE), then SR-CLG is obtained by removing RE and NE. Following [38], we utilize a self-regression strategy (SR) for SR-CLG to control the element number in the text Layout. PC is retained since it promises the text layout to finally match the length of the Chinese text. Training details for all these models are introduced in the supplementary materials.

	Components		Metrics (↓)			
Methods	NE	RE	Ove	Ali	Occ	Rea
SR-CLG			1.027	**0.006**	0.127	0.304
CLG (w/o RE)	✓		0.017	0.010	0.115	**0.301**
CLG	✓	✓	**0.004**	0.008	**0.108**	0.302

Table 1: Ablations for controllable layout generator with different components, where bold is the best performance.

To evaluate the text layout generated by the three models, We randomly selected 1500 background images and corresponding Chinese text lines as the input for them to generate specific text layouts. Following [13, 38], we compare the three models in terms of four numerical indexes, including Ove [18], Ali [2], Occ [13], and Rea [38]. These indexes measure the alignment, occlusion, and visual harmony of the text layout. For each index, the lower the value is, the better the layout is. The results are shown in Table. 1.

According to the Table. 1, taking self-regression strategy [38] without NE has a much higher Ove, which means the produced layout elements probably overlap together. This reflects the significance of the NE. Besides, the differences between Ali, Occ, and Rea among the three models are not obvious, indicating that adding modules for text control does not weaken other abilities. For a more intuitive visual comparison, we also run the original DS-GAN [13] by only providing the background images. Related results are shown in Fig. 9. According to Fig. 9, origin DS-GAN [13] generates text layouts with complete mistakes of number and size, leading to full incompatibility in our Prompt2Poster. Although promising the correct number of elements in $\boldsymbol{T}$, SR-CLG faces serious element overlap, resulting in unacceptable results. Lacking RE, CLG (w/o RE) cannot estimate the final text layout in advance. This leads to significant cropping by PC and reduces space utilization. However,

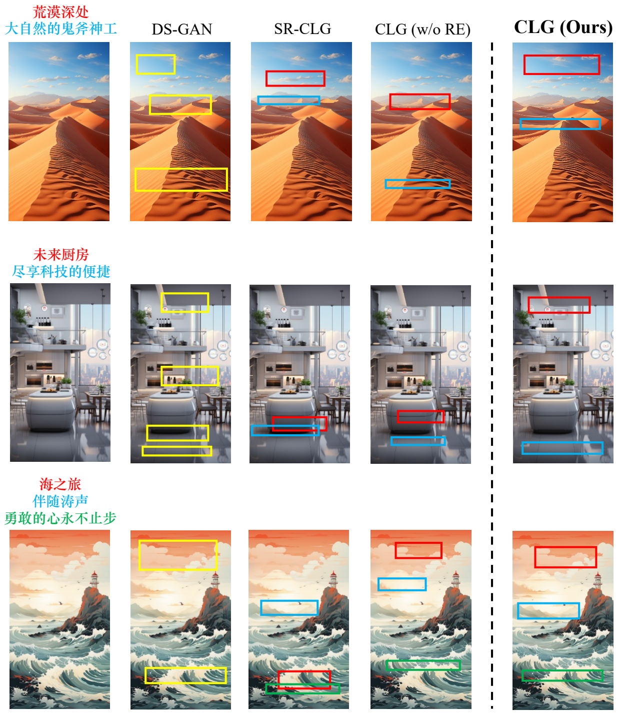
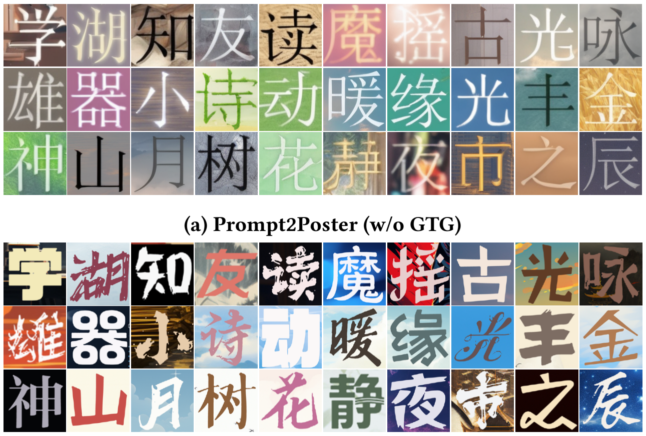


Figure 9: Visual comparison among four layout models. The leftmost column presents input Chinese texts and background, and other columns show the text layouts generated by corresponding models respectively.

with the addition of NE and RE, our CLG can produce a text layout that harmonizes with background and aligns texts accurately.

Graphical Text Generator. Besides keeping the visual correctness, the main purpose of utilizing GTG in our Prompt2Poster is to bring a huge style diversity of Chinese graphical texts, especially fonts. To verify that using GTG brings Prompt2Poster strongly stylized ability, we remove this module and obtain another system, named Prompt2Poster (w/o GTG). Specifically, following [37], we only exploit graphical templates to control the generation of Chinese graphical texts. The templates are created with the help of our CLG. More details are shown in the supplementary materials. We compared the style diversity of the Chinese graphical texts for Prompt2Poster (w/o GTG) and Prompt2Poster for each character, illustrated in Fig. 10. As Fig. 10 shows, without GTG, the styles of those generated Chinese graphical characters are much more monotonous. Particularly, although on different backgrounds, they seem to be very similar in terms of font style. With GTG, in terms of the graphical text style, our Prompt2Poster produces more diverse results and the fonts are much more elegant and varied.

In addition, it's worth mentioning that three modal inputs, including vision, geometry, and linguistics, all influence the output of our GTG. As shown in Fig. 11, any changes to the background image, layout, and Chinese text will bring variations in the generated Chinese graphical text. This full utilization of multi-modal information largely enhances the style diversity and harmony of the generated Chinese graphical texts.



(a) Prompt2Poster (w/o GTG)

(b) Prompt2Poster

Figure 10: Comparison of diversity in Chinese graphical texts between Prompt2Poster (w/o GTG) and Prompt2Poster.

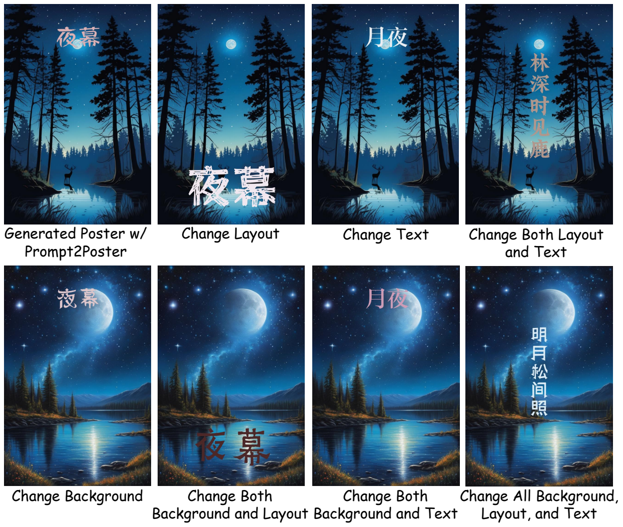

夜幕
夜幕
月夜
林深时见鹿
Generated Poster w/ Prompt2Poster
Change Layout
Change Text
Change Both Layout and Text
夜幕
夜幕
月夜
明月松间照
Change Background
Change Both Background and Layout
Change Both Background and Text
Change All Background, Layout, and Text

Figure 11: Visualize ablations of different inputs of Graphical Text Generator (GTG), which demonstrate that GTG creates various graphical Chinese styles when inputs change.

5 CONCLUSION

In this work, we propose an automatic prompt-guided poster creation framework named Prompt2Poster for creating desired artistic Chinese posters including an aligned background, reasonable layouts, and stylized graphical texts. Prompt2Poster first extracts the user intention and generates the aligned background supported by powerful large models, and then creates visually pleasurable content with our carefully designed modules including Controllable Layout Generator and Graphical Text Generator, leading to harmonious layout and accurate graphical texts. Extensive experiments demonstrate the superiority of our Prompt2Poster over other prompt-only poster generation methods. We believe our fully automatic poster creation framework will inspire potential research and interesting applications in the graphic design area.

REFERENCES

[1] 2023. DALL-E 3. https://openai.com/dall-e-3.

[2] Diego Martin Arroyo, Janis Postels, and Federico Tombari. 2021. Variational transformer networks for layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 13642–13652.

[3] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. 2022. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324* (2022).

[4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 1877–1901.

[5] Yunning Cao, Ye Ma, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. 2022. Geometry aligned variational transformer for image-conditioned layout generation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 1561–1571.

[6] Saemi Choi, Kiyoharu Aizawa, and Nicu Sebe. 2018. Fontmatcher: font image paring for harmonious digital graphic design. In *23rd International Conference on Intelligent User Interfaces*. 37–41.

[7] Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. 2023. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. *arXiv preprint arXiv:2305.15393* (2023).

[8] Yifan Gao, Jinpeng Lin, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. 2023. TextPainter: Multimodal Text Image Generation withVisual-harmony and Text-comprehension for Poster Design. *arXiv preprint arXiv:2308.04733* (2023).

[9] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. *Advances in Neural Information Processing Systems (NeurIPS)* 27 (2014).

[10] Shunan Guo, Zhuochen Jin, Fuling Sun, Jingwen Li, Zhaorui Li, Yang Shi, and Nan Cao. 2021. Vinci: an intelligent graphic design system for generating advertising posters. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–17.

[11] Haibin He, Xinyuan Chen, Chaoyue Wang, Juhua Liu, Bo Du, Dacheng Tao, and Yu Qiao. 2022. Diff-Font: Diffusion Model for Robust One-Shot Font Generation. *arXiv preprint arXiv:2212.05895* (2022).

[12] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems (NeurIPS)* 33 (2020), 6840–6851.

[13] Hsiao Yuan Hsu, Xiangteng He, Yuxin Peng, Hao Kong, and Qing Zhang. 2023. PosterLayout: A New Benchmark and Approach for Content-aware Visual-Textual Presentation Layout. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6018–6026.

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).

[15] Ali Jahanian, Jerry Liu, Qian Lin, Daniel Tretter, Eamonn O'Brien-Strain, Seungyon Claire Lee, Nic Lyons, and Jan Allebach. 2013. Recommendation system for automatic design of magazine covers. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. 95–106.

[16] Dorothea Jameson and Leo M Hurvich. 1964. Theory of brightness and color contrast in human vision. *Vision research* 4, 1-2 (1964), 135–154.

[17] Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-task learning using uncertainty to weigh losses for scene geometry and semantics. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 7482–7491.

[18] Jianan Li, Jimei Yang, Jianming Zhang, Chang Liu, Christina Wang, and Tingfa Xu. 2020. Attribute-conditioned layout gan for automatic graphic design. *IEEE Transactions on Visualization and Computer Graphics* 27, 10 (2020), 4039–4048.

[19] Long Lian, Boyi Li, Adam Yala, and Trevor Darrell. 2023. LLM-grounded Diffusion: Enhancing Prompt Understanding of Text-to-Image Diffusion Models with Large Language Models. *arXiv preprint arXiv:2305.13655* (2023).

[20] Long Lian, Baifeng Shi, Adam Yala, Trevor Darrell, and Boyi Li. 2023. LLM-grounded Video Diffusion Models. *arXiv preprint arXiv:2309.17444* (2023).

[21] Lukas Liebel and Marco Körner. 2018. Auxiliary tasks in multi-task learning. *arXiv preprint arXiv:1805.06334* (2018).

[22] Jinpeng Lin, Min Zhou, Ye Ma, Yifan Gao, Chenxi Fei, Yangjian Chen, Zhang Yu, and Tiezheng Ge. 2023. AutoPoster: A Highly Automatic and Content-aware Design System for Advertising Poster Generation. *arXiv preprint arXiv:2308.01095* (2023).

[23] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[24] Quynh Phung, Songwei Ge, and Jia-Bin Huang. 2023. Grounded Text-to-Image Synthesis with Attention Refocusing. *arXiv preprint arXiv:2306.05427* (2023).

[25] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

[26] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125* 1, 2 (2022), 3.

[27] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10684–10695.

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems (NeurIPS)* 35 (2022), 36479–36494.

[29] Mohammad Amin Shabani, Zhaowen Wang, Difan Liu, Nanxuan Zhao, Jimei Yang, and Yasutaka Furukawa. [n. d.]. Visual Layout Composer: Image-Vector Dual Diffusion Model for Design Layout Generation. ([n. d.]).

[30] Masataka Tokumaru, Noriaki Muranaka, and Shigeru Imanishi. 2002. Color design support system considering color harmony. In *2002 IEEE world congress on computational intelligence. 2002 IEEE international conference on fuzzy systems. FUZZ-IEEE'02. Proceedings (Cat. No. 02CH37291)*, Vol. 1. IEEE, 378–383.

[31] Praneetha Vaddamanu, Vinay Aggarwal, Bhanu Prakash Reddy Guda, Balaji Vasan Srinivasan, and Niyati Chhaya. 2022. Harmonized Banner Creation from Multimodal Design Assets. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*. 1–7.

[32] Sreekanth Vempati, Korah T Malayil, V Sruthi, and R Sandeep. 2020. Enabling hyper-personalisation: Automated ad creative generation and ranking for fashion e-commerce. In *Fashion Recommender Systems*. Springer, 25–48.

[33] Haohan Weng, Danqing Huang, Yu Qiao, Zheng Hu, Chin-Yew Lin, Tong Zhang, and CL Chen. 2024. Desigen: A Pipeline for Controllable Design Template Generation. *arXiv preprint arXiv:2403.09093* (2024).

[34] Yangchen Xie, Xinyuan Chen, Li Sun, and Yue Lu. 2021. Dg-font: Deformable generative networks for unsupervised font generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5130–5140.

[35] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *arXiv preprint arXiv:2211.01335* (2022).

[36] Yukang Yang, Dongnan Gui, Yuhui Yuan, Haisong Ding, Han Hu, and Kai Chen. 2023. GlyphControl: Glyph Conditional Control for Visual Text Generation. *arXiv preprint arXiv:2305.18259* (2023).

[37] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

[38] Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. 2022. Composition-aware graphic layout GAN for visual-textual presentation designs. *arXiv preprint arXiv:2205.00303* (2022).