Supplementary Materials: Prompt2Poster: Automatically Artistic Chinese Poster Creation from Prompt Only

Anonymous Authors

In this supplementary, we introduce more details of the Prompt2Poster, including the Controllable Layout Generator, the Graphical Text Generator, and the training details in Sup. A. More experimental results including the quantitative user study, ablation for the proposed Graphical Text Generator module, and more visual results are provided in Sup. B.

A IMPLEMENT DETAILS

A.1 Controllable Layout Generator

This subsection gives the detailed structural introduction of our Controllable Layout Generator (CLG) and the corresponding discriminator for training. The whole framework contains six modules. They are the Number Encoder (NE), the Rate Encoder (RE), the DS-GAN Image Encoder, the DS-GAN layout prediction backbone, the DS-GAN discrimination backbone, and the post-processing cropping module (PC). When training, Our CLG, and the discriminator, share the same NE, RE, and Image Encoder, but include the layout predict backbone and the discriminated backbone, respectively. The NE and the RE are trainable two-layer MLP networks with the same structure. The DS-GAN Image Encoder includes a pretrain salience-detection network and a ResNet50-based feature encoder [4]. When a background image is sent to the Image Encoder, the saliency-detection network first extracts the gray saliency map for the background. Then, the saliency map is concatenated with the background and fed into the ResNet50-based image feature encoder to output the image features. The DS-GAN layout prediction backbone, and the DS-GAN discrimination backbone, are both LSTM-based networks [4]. The former in the CLG receives the number, rate, and image features, and outputs the text layout. The latter in the discriminator receives the number, rate, image features, and truth or false text layout, and outputs the authenticity of the text layout. The PC is removed in training as it is a fixed algorithm.

A.2 Graphical Text Generator

In this section, we analyze the distinctions between the Stylized Font Generator and Diff-Font [3]. The structure of the Stylized Font Generator used in GTG is the same as Diff-Font [3]. However, as shown in Fig. 1, the difference in their forward process is that Diff-Font generates a particular font similar in style from a character example, while the Stylized Font Generator obtains the specified font based on the font style embedding predicted by the Text Attributes Predictor. In other words, in terms of font prediction, we have made GTG learn the mapping from the background image, layout, and text to a character example which is obviously the average of the graphical texts in the ground truth.

A.3 Training Details

Controllable Layout Generator. For training, since our CLG is an extension from the DS-GAN, we followed nearly all the settings introduced in [4], including the learning rates for the generator and the discriminator, the loss function, the training epoch, the batch size, and so on. The only difference was that we adjusted the original training dataset. Origin DS-GAN was trained on the PKU PosterLayout [4] training dataset with 9973 background-layout pairs. The elements of the layouts in the dataset included text elements, logo elements, and underline elements. Since only text layout was considered in our task, we ignored all the elements in the layouts except for the text elements and obtained the text layouts. When training, given a background-layout pair, the one hot vector input to the NE was acquired by calculating the total number of text elements in a layout. For the rate vector input to the RE, its i-th element was the size rate of the i-th text layout element, expressed as w_i/h_i. It's worth mentioning that the other two proposed weaker text-controllable layout models, CLG (w/o RE) and SR-CLG remained the same inner structures and training settings except for removing corresponding modules. The training for each text layout model was conducted on a single Nvidia-A100 80GB GPU within one day.

Graphical Text Generator. For training, we selected the *ViT-L-14-336* [2] model from *Chinese-CLIP* [7] as our pre-trained model. For the loss function $\mathcal{L}_{font}$, the temperature was set to 0.1. Meanwhile, for the total loss function $\mathcal{L}_{total}$, the regularization term μ was assigned a value of 0.1. During training, we utilized the AdamW optimizer with a weight decay of 5e-4. Additionally, we employed a cosine learning rate decay with a warm-up strategy, where the initial learning rate was set to 1e-5, and the maximum learning rate was set to 1e-2. The training batch size was set to 256, and a dropout strategy with a rate of 0.5 was also applied. The entire training process spanned 600 epochs, with a warm-up period of 80 epochs. Training was conducted over the course of one day using two Nvidia-A100 80GB GPUs. The training dataset of GTG comprises 3,170 posters from the TextLogo3K [6] training set and 5,276

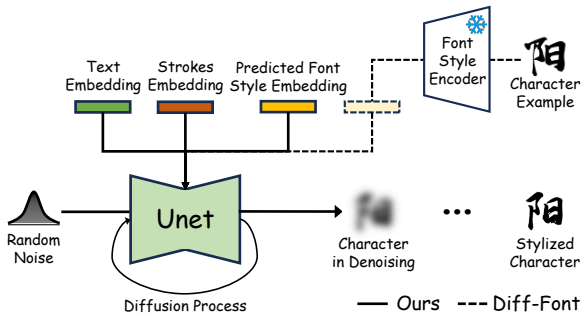

Figure 1: Our GTG's Stylized Font Generator differs from Diff Font's Font Generator. We employ font style embeddings predicted by multimodal perception. This replaces the need for manually specifying example images that are encoded by a font style encoder.

Metrics	SD XL	DALL·E 3	Midjourney	GlyphControl	SDXL + ControlNet (Canny)	Ours
Text Quality	4.35%	13.04%	10.87%	2.17%	23.91%	**45.65%**
Visual Harmony	8.70%	15.22%	13.04%	4.35%	17.39%	**41.30%**
Overall Attention	4.35%	13.04%	23.91%	6.52%	8.70%	**43.48%**

Table 1: The detailed scores of each model in the user study, where the percentage represents the proportion of achieving the highest score rated by users. In the event of a tie for the highest score, all tied scores are considered to be the highest. The data is rounded to the nearest hundredth using standard rounding rules.

Figure 2: Samples in the PKU PosterLayout dataset.

Figure 3: Image samples in the dataset used for training Graphical Text Generator.

from our custom collection, totaling 8,446 posters. We extracted stylized fonts using a Chinese OCR character recognition model and the Canny edge detection model as our font Ground Truth. Simultaneously, the average color of each text line was established as the Ground Truth for font color in GTG's training.

B MORE EXPERIMENTS

B.1 User Study

We have designed 30 groups of prompts in distinct styles to demonstrate the outstanding generalization and performance of Prompt2Poster. Our prompts encompass a variety of common poster scenarios, such as frequently occurring solar terms, promotional notices, and sports, among others.

The same group of prompts is inputted into Prompt2Poster, SD XL [5], Midjourney, DALL·E 3 [1], GlyphControl [8], and SD XL + ControlNet (Canny) [9] to generate posters. Notably, the Midjourney used is the latest version available at that time, which is version 5.2. As the official GlyphControl code cannot render Chinese graphical texts, we simply changed the font library to generate Chinese graphical texts. We required participants to rate each poster on text quality, visual harmony, and overall attraction. The model that produced the highest-scoring poster in each group of six was deemed the winner. To ensure fairness, the model results presented to the participants were anonymized and randomized.

The detailed scoring breakdown is shown in Table. 1. In terms of text quality, we achieved a favorability rate of 45.65% from users. It's also noteworthy that SD XL + ControlNet (Canny) performed well. This is because the text templates used by ControlNet were manually rendered by us, ensuring a high accuracy rate at the outset. Nevertheless, Prompt2Poster still outperforms it by a significant margin. In the field of visual harmony, we achieved the favor of 41.30% of users. Despite being the latest state-of-the-art model, DALL·E 3 is still inferior to us in this regard. In terms of overall attraction, we also achieved the highest score of 43.48%. It can be seen that Midjourney, as a proprietary large model that requires payment, achieved commendable results. However, it struggles to generate readable Chinese graphical characters, which inherently limits its overall aesthetic appeal.

B.2 Ablation for GTG

To evaluate the contribution of our proposed Graphical Text Generator (GTG) to poster production, we removed it from the original Prompt2Poster, creating a variant named Prompt2Poster (w/o GTG). This version generates the poster based on the same prompt with ControlNet (Canny), which utilizes templates that incorporate the same layout as predicted by CLG but exclusively uses a standard Chinese Song typeface for graphical texts. Visual results demonstrate the effectiveness of our proposed GTG.

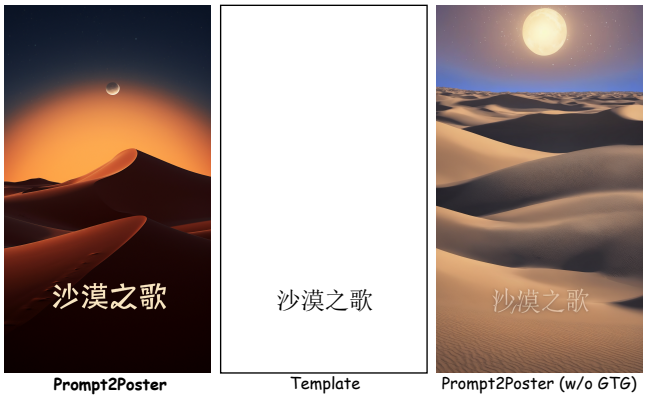

Figure 4: The templates for Prompt2Poster (w/o GTG) are created with the help of CLG and graphical texts in the Chinese Song typeface. Using the templates, Prompt2Poster (w/o GTG) generates posters through the ControlNet (Canny).

Figure 5: Prompt2Poster is capable of generating images at various resolutions.

B.3 Visual Results

Variable Resolution Image Generation. As shown in Fig. 5, Prompt2Poster can generate images in various visual styles and resolutions with different aspect ratios. This capability enables Prompt2Poster to better meet user prompts, offering flexibility in design.

More Results. The additional visual results generated by Prompt2-Poster are illustrated as follows.

REFERENCES

[1] 2023. DALL-E 3. https://openai.com/dall-e-3.
[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
[3] 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).
[4] 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.
[5] 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).
[6] Yizhi Wang, Gu Pu, Wenhan Luo, Pengfei Wang, Yexin ans Xiong, Hongwen Kang, Zhonghao Wang, and Zhouhui Lian. 2022. Aesthetic Text Logo Synthesis via Content-aware Layout Inferring. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
[7] 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).
[8] 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).
[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.

一片静谧的湖面，周围环绕着山脉，天空斜阳洒满余晖。图上写着“湖光山色”。
(A tranquil lake, surrounded by mountains, basks in the slanting sun's afterglow. "Lake and Mountain Hue" is written on the image.)

一座旧时钟，钟面上反射着城市的灯光，象征着过去和现在的交织。图上写着“时间的指针”。
(An old clock, its face reflecting city lights, symbolizes the interweaving of past and present. "Hands of Time" is written on the image.)

精致钢琴键盘上布满优雅的玫瑰花瓣，背后是闪烁着温馨洋溢的烛光。图上写着“旋律之恋”。
(An elegant piano keyboard covered in rose petals, with warm, flickering candlelight in the background. "Melody of Love" is written on the image.)

月光下，沉静的湖面上水波轻轻荡漾，宛如星光在水面舞动。图上写着“月光之舞”。
(Under the moonlight, gentle ripples dance across the calm lake surface, resembling stars twinkling on water. "Dance of the Moonlight" is written on the image.)

在城市街头，行人熙熙攘攘，大厦高耸入云。图上写着“都市生活”。
(On a city street, crowds swarm and skyscrapers soar into the clouds. "Urban Life" is written on the image.)

在密布的星空下，天空中有几颗星星闪烁着光芒，映着少女的裙摆。图上写着“星空奥秘”。
(Under a dense starry sky, a few stars twinkle, reflecting on a young girl's skirt. "Mysteries of the Starry Sky" is written on the image.)

朝霞弥漫的早晨，群山掩映，田野上，农夫们挥汗如雨。图上写着“朝气蓬勃”，“每一日都是新的希望”。
(In a morning filled with rosy dawn, mountains loom over fields where farmers toil profusely. "Vibrant Morning" and "Every Day Brings New Hope" are written on the image.)

在静谧的山间，远处是挺拔优雅的松树矗立，近处是一道道清澈的溪流。图上写着“山水精神”，“天然之韵生活宁静”。
(In the tranquil mountains, tall, elegant pines stand in the distance, with clear streams flowing nearby. "Spirit of the Landscape" and "The Rhythm of Nature Brings Peace" are written on the image.)

山谷中，雾气缭绕，树木点缀其中，轮廓模糊却又生动。远处的山峰与薄雾构成美丽的画面。图上写着“雾绕”，“山与雾的诗意”。
(In the valley, mist swirls among the trees, their outlines blurred yet vivid. Distant peaks emerge through the thin mist, creating a picturesque scene. "Enshrouded in Mist" and "The Poetic Interplay of Mountains and Fog" are written on the image.)

金色的阳光洒在青翠的山野，形成一片生机盎然的生命海洋。图上写着“黎明破晓”，“每一刻都是新的期许”。
(Golden sunlight bathes the verdant hills, creating a vibrant sea of life. "Dawn Breaks" and "Every Moment Holds New Expectations" are written on the image.)

阳光照射在静谧的湖面，波光粼粼，形成了一片丽影。图上写着“湖影”，“光与水的交响曲”。
(Sunlight shines on the tranquil lake surface, creating a dazzling display of light. "Lake Reflection" and "A Symphony of Light and Water" are written on the image.)

在书房里，整齐排列着各种书籍，灯光下的书页影显深邃的智慧光环。图上写着“知识的海洋”，“寂静中的智慧之灯”
(In the study, various books are neatly arranged, with the light illuminating the pages, highlighting a profound aura of wisdom. "Ocean of Knowledge" and "Lamp of Wisdom in Silence" are written on the image.)

古典的图书馆里，窗外是远方的浩渺星空和明亮的灯塔，室内的灯火为探索者照亮了前行的路。图上写着“星海图书之堂”，“追求智慧的灯塔”。
(In a classical library, the vast starry sky and a bright lighthouse are visible through the window, while indoor lights illuminate the path for explorers. "Hall of the Starry Sea Library" and "Lighthouse of Wisdom Pursuit" are written on the image.)

日落时分，夕阳正好落在山顶，金色的光芒洒满山头。图上写着“山影”，“日光与地面的对话”。
(At sunset, the sun perfectly sets atop a mountain, casting golden light across the peak. "Mountain Shadow" and "Dialogue Between Sunlight and Earth" are written on the image.)

太阳从地平线苏醒，阳光射出，照亮整个天空。图中写着“日出”，“光与影的交响”。
(The sun awakens from the horizon, casting rays that illuminate the entire sky. "Sunrise" and "Symphony of Light and Shadow" are written on the image.)

在浩渺的星空下，一轮皎洁的明月高挂，守护着几株起伏的松树。图上写着“月光的赞歌”，“梦想的翼展”。
(Beneath a vast starry sky, a bright moon hangs high, guarding undulating pines. "Ode to the Moonlight" and "Wingspan of Dreams" are written on the image.)

优雅的咖啡馆内，曲线美的明窗外是热闹的城市街头。内部装潢别致，吧台上摆放着老式的咖啡机和精致的甜点。图上写着“浓情咖啡馆”，“追求美味启发灵感”，“加入我们共享生活的色彩”。
(Inside an elegant café, curved windows overlook a bustling city street. The chic interior features an antique coffee machine and exquisite desserts on the bar. "Rich Aroma Café," "Pursuing Flavor that Inspires," and "Join Us to Share Life's Colors" are written on the image.)

静谧的森林中，晨露滋润，古老的树木诉说着岁月的秘密。树下，有鸟在欢快歌唱。图上写着“森之语”，“生命之声”，“自然的节奏慢慢跳动”。
(In the tranquil forest, morning dew nourishes ancient trees that whisper the secrets of time. Underneath, birds sing joyfully. "Whispers of the Forest," "Voices of Life," and "Nature's Rhythm Beats Slowly" are written on the image.)