PosterVerse: A Full-Workflow Framework for Commercial-Grade Poster Generation with HTML-Based Scalable Typography

Junle Liu[*1,3], **Peirong Zhang**[*1], **Yuyi Zhang**[1,3], **Pengyu Yan**[1], **Hui Zhou**[2,3], **Xinyue Zhou**[2,3], **Fengjun Guo**[2,3], **Lianwen Jin**[1,3†]

[1]South China University of Technology
[2]Intsig Information Co., Ltd.
[3]INTSIG-SCUT Joint Lab on Document Analysis and Recognition
junle_liu@foxmail.com, eelwjin@scut.edu.cn

Abstract

Commercial-grade poster design demands the seamless integration of aesthetic appeal with precise, informative content delivery. Current automated poster generation systems face significant limitations, including incomplete design workflows, poor text rendering accuracy, and insufficient flexibility for commercial applications. To address these challenges, we propose **PosterVerse**, a full-workflow, commercial-grade poster generation method that seamlessly automates the entire design process while delivering high-density and scalable text rendering. PosterVerse replicates professional design through three key stages: (1) blueprint creation using fine-tuned LLMs to extract key design elements from user requirements, (2) graphical background generation via customized diffusion models to create visually appealing imagery, and (3) unified layout-text rendering with an MLLM-powered HTML engine to guarantee high text accuracy and flexible customization. In addition, we introduce **PosterDNA**, a commercial-grade, HTML-based dataset tailored for training and validating poster design models. To the best of our knowledge, PosterDNA is the first Chinese poster generation dataset to introduce HTML typography files, enabling scalable text rendering and fundamentally solving the challenges of rendering small and high-density text. Experimental results demonstrate that PosterVerse consistently produces commercial-grade posters with appealing visuals, accurate text alignment, and customizable layouts, making it a promising solution for automating commercial poster design.

Code — https://github.com/wuhaer/PosterVerse

Introduction

Poster design plays a crucial role in business, culture, and marketing. An effective poster must distill complex information into clear, compelling messages while balancing informational richness with focused messaging. This requires sophisticated orchestration of font, color, and layout skills that traditionally demand extensive design expertise and time-intensive manual processes. With the rapid development of AI-generated content (AIGC) technologies (Gemini et al. 2023; Black Forest Labs 2024; OpenAI 2025), generative models have become key tools for commercial creation (Lin et al. 2023b; Gao et al. 2025b), making their application in automatic design an inevitable trend.

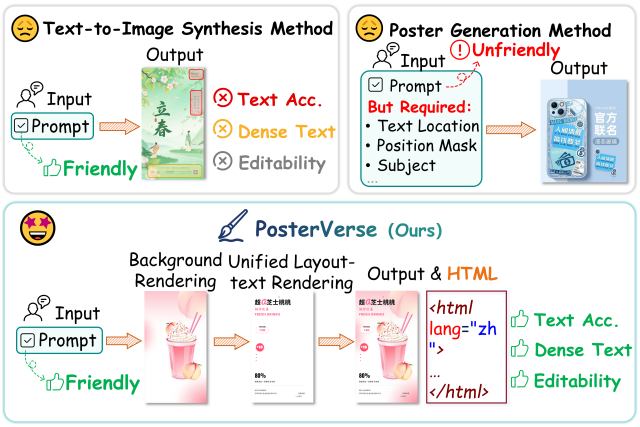

Figure 1: Comparison between existing poster generation methods (top) and the proposed PosterVerse (bottom).

However, existing approaches suffer from several limitations. **(1) Lack of full workflow solutions.** Poster generation involves multiple stages, including background graphic design, layout planning, and font rendering. Yet, many methods (Seol, Kim, and Yoo 2024; Li et al. 2023a,b; Hsu et al. 2023) only focus on partial stages such as generating layout or typography alone, failing to provide full-workflow poster design systems. **(2) Poor flexibility and accessibility.** While some works (Gao et al. 2025b; Peng et al. 2025) demonstrate promising visual effects, they heavily rely on additional inputs beyond textual prompts, such as positional masks, text bounding boxes, and graphical subjects. This greatly hampers their flexibility compared to prompt-only systems, especially for non-technical users. **(3) Inaccurate text rendering.** Despite the stunning aesthetic creation abilities of models like GPT-4o (OpenAI 2025) and Gemini (Gemini et al. 2023), they typically struggle with generating accurate text, particularly evident in Chinese characters and dense, small-scale characters (Zhang et al. 2025b). The generated texts are usually illegible or semantically meaningless. **(4)**

[*]These authors contributed equally.
[†]Corresponding Author.

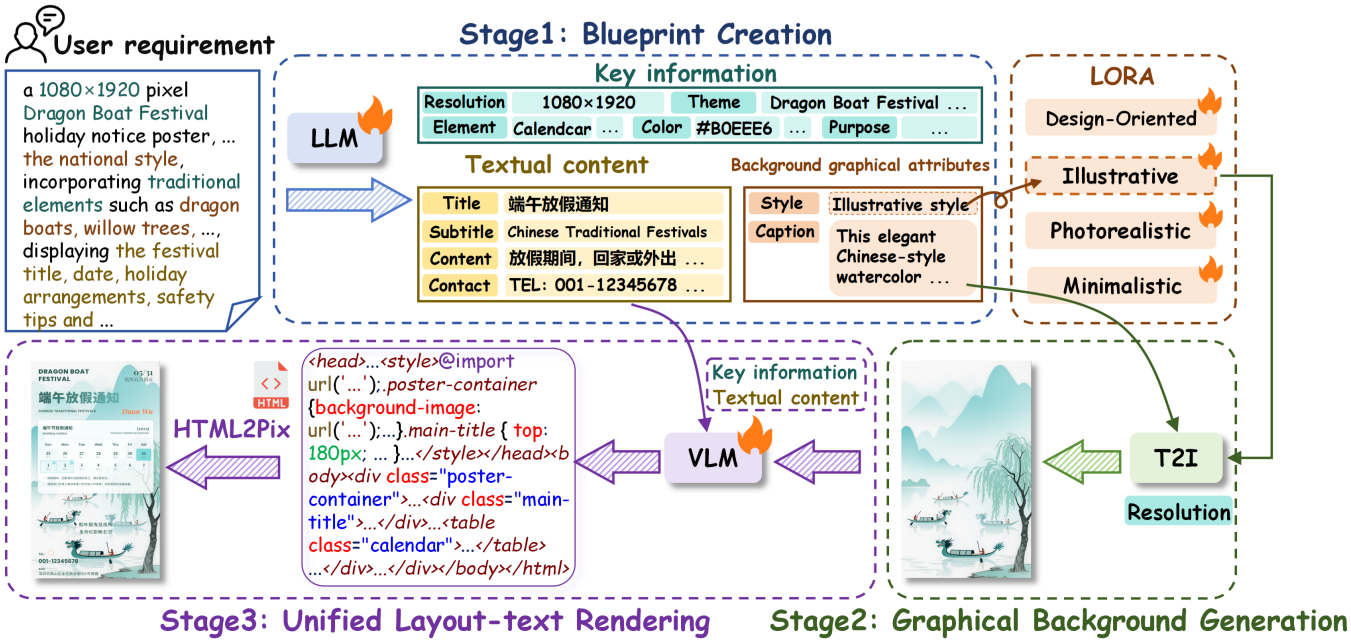

Figure 2: The overview of PosterVerse: A full-workflow method integrating blueprint creation, graphical background generation, and unified layout-text rendering to produce commercial-grade, aesthetically appealing, and text-rich posters.

Insufficient understanding of user requirements. Existing text-to-image models (Black Forest Labs 2024; Stability AI 2024) are usually constrained by input token limits or primitive understanding capabilities of text encoders such as T5 (Raffel et al. 2020), limiting them to fully comprehend user needs. **(5) Gap to commercial utility.** Current poster design methods (Chen et al. 2025b; Wang et al. 2024) mainly prioritize artistic expression over user requirement adherence. Although aesthetically appealing, the generated posters are commercially impractical and require substantial manual post-processing. Additionally, most of them generate non-editable static images. This prevents post-adjustments to poster content like text, fonts, or layout, making them ill-suited for the dynamic nature of commercial scenarios where requirements frequently evolve.

To bridge these gaps, we propose **PosterVerse**, a full-workflow, prompt-driven poster generation framework, featuring scalable text rendering to address small, dense text synthesis and natively editable output for flexible post-editing. Specifically, PosterVerse mimics the process of professional designers creating posters by dividing the poster design process into three stages: **(1) Blueprint creation.** In this stage, we utilize a fine-tuned Large Language Model (LLM) to interpret and expand upon the user's requirement. Irrespective of the initial input's level of detail, this process transforms the request into a comprehensive design specification. We also extract key design elements such as themes, styles, colors, and text content, serving as foundations for the subsequent stages. **(2) Graphical background generation.** Building upon Flux.1-dev fine-tuned by LoRA, this stage generates high-quality backgrounds that strictly adhere to the design blueprint from the first stage. To afford users greater creative control, PosterVerse also supports direct upload of custom background images as an alternative. **(3) Unified layout-text rendering.** In this stage, a multimodal LLM (MLLM) is used to synthesize the final layout, integrating specified text and design elements with the background graphic. The output is a complete HTML document that ensures perfect typographic accuracy, addressing a known weakness in many generative models. Moreover, HTML enables efficient and customizable post-edits, accommodating dynamic design scenarios that require frequent modifications such as commercial.

Furthermore, to address the critical lack of commercial-grade dataset, we present **PosterDNA**, an HTML-based poster generation dataset with fine-grained specifications. Developed in collaboration with professional designers for high quality and practical relevance, PosterDNA comprises a diverse collection of poster samples characterized by complex layouts and dense textual designs. Each entry is a structured tuple of "requirements-graphic-layout-poster", specifically engineered to support the modular training and validation of our PosterVerse. It pioneers in introducing HTML-based typography files, standing as the first Chinese poster design dataset that addresses small, dense text rendering and potentially inspiring future works.

Overall, our contributions can be summarized as follows:

- We propose **PosterDNA**, the first commercial-grade and text-dense poster generation dataset with fine-grained HTML-based specifications, designed to support modular training and validation with high-quality samples.
- We propose **PosterVerse**, a full-workflow method that integrates blueprint creation, graphical background generation, and unified layout-text rendering, enabling the

creation of posters with aesthetically sophisticated layouts and text-dense designs for commercial-grade use.
- PosterVerse allows users to generate commercial-grade posters solely from textual prompts, while maintaining editability capabilities for further customization.
- Extensive experiments demonstrate that PosterVerse can generate visually appealing posters with aesthetic designs, precise text, and well-crafted layouts, meeting the standards of commercial-grade posters.

Related Work

Visual Text Image Synthesis In recent years, text-to-image generation has proliferated due to its unprecedented controllability and high fidelity (Reed et al. 2016; Nonghai Zhang 2024). However, visual text synthesis remains challenging, requiring models to accurately render font structures while maintaining visual aesthetics (Zhang et al. 2025b). Early approaches improve text encoders by scaling them up (Balaji et al. 2022; Lab 2023) or re-aligning them with visual features (Zhao and Lian 2024). Subsequently, for enhanced controllability and accuracy, researchers focus on conditioning diffusion models (Ho, Jain, and Abbeel 2020; Rombach et al. 2022) with various prior information, which can be categorized into three types. The first type employs glyph images rendered on white backgrounds as conditions (Ma et al. 2023; Yang et al. 2023). The second type combines a position mask and a rendered glyph (sometimes not) as input, using binary masks to refine text positions (Chen et al. 2023; Tuo, Geng, and Bo 2024; Zhang et al. 2025a, 2024; Xie et al. 2025). In contrast to glyph or position masks, the third type extracts layout information for multiple text instances. TextDiffuser-2 (Chen et al. 2024) uses one LLM to generate language-like text layout and another to encode it as diffusion inputs. Lakhanpal et al. (Lakhanpal et al. 2025) propose a training-free framework, using a frozen layout generator for iterative refinement.

Layout Planning Layout planning is a crucial aspect to maintain natural text-background integration and visual aesthetics in text image synthesis. Preliminary studies focus on conditional layout generation using Transformer (Gupta et al. 2021) and sequential Diffusion models (Hui et al. 2023), producing bounding-box layouts without visual content. With the rise of LLM, their semantic and logical reasoning abilities are explored for layout planning (Lin et al. 2023a; Zhang et al. 2025c). Researchers then shift to content-aware layout generation, feeding background images and text prompts to models that place text appropriately without obscuring main content (Horita et al. 2024; Seol, Kim, and Yoo 2024; Hsu and Peng 2025). Recently, leveraging MLLMs' understanding capabilities, some works input background images and user requirements to generate typography JSON files that plan layouts with content-awareness and render textual content in tandem for better text arrangement and visual coherence (Yang et al. 2024b; Jia et al. 2023).

Poster Generation Building upon advances in visual text synthesis and layout planning, automatic poster generation emerges as a specialized task that creates infographics with rich text and high-quality artistic presentation, primarily for advertising and marketing campaigns. Current poster generation methods can be grouped into two categories. The first category is semi-automatic, relying on pre-given conditions like positional masks (Tuo, Geng, and Bo 2024), user-specified subjects (Gao et al. 2025b), and text bounding boxes (Peng et al. 2025). This heavy reliance on conditions greatly hampers their flexibility and user-friendliness. Conversely, the second type is fully-automatic and condition-free, leveraging only textual prompts to automatically plan layouts and generate visual content (Wang et al. 2024; Chen et al. 2025b; Wang et al. 2025; Chen et al. 2025a). This approach offers enhanced accessibility, enabling users to create commercial-grade posters through natural language descriptions. Despite the enhanced automation and visual quality, existing methods typically fall short in generating large-amount, high-density, and small-scale text, outputting illegible or misplaced characters. Also, they typically generate static posters, prohibiting post-editing. This motivates us to develop PosterVerse, a prompt-driven poster generation framework that novelly generates editable HTML-based typography file for poster design, enabling scalable text rendering and flexible post-generation customizability.

Method

Overall Architecture The framework of PosterVerse is demonstrated in Fig. 2. It takes only the user requirement as input, and then designs a complete poster following three stages: blueprint creation, graphical background generation, and unified layout-text rendering. This architecture mirrors the workflow of professional designers while maintaining complete automation.

Blueprint Creation User requirements for poster design are typically expressed through natural language, which tends to be ambiguous and lacks specificity. Designers should interpret these vague descriptions to understand the user's intentions, regardless of whether they are detailed or brief. Inspired by this, we design a Detail-Insensitive Requirement Parsing (DIPR) mechanism for the first stage of blueprint creation. We established three different user requirement levels (basic, medium, detailed) and trained the model to transform them into consistently comprehensive generation blueprints. The output blueprint is formatted as JSON and includes three parts: *textual content* (e.g., title, subtitle, main content, and contact information), *background graphical attributes* (e.g., style and image captions), and *key extracted parameters* (e.g., resolution, theme, elements, color, and purpose). During DIPR training, we fine-tuned Qwen2.5-14B (Yang et al. 2024a) using randomly selected detail levels as input. The model is supervised by the same ground-truths of the blueprint information, thus developing insensitiveness to the richness of user input. The generated blueprints are used for training in subsequent stages.

Graphical Background Generation The second stage of PosterVerse generates a graphical background for the poster. Background generation plays a crucial role in defining the overall aesthetic and tone of the poster. Motivated by professional designers who tailor their artistic styles to match

project requirements, we classify poster backgrounds into four styles: *Illustrative*, *Design-Oriented*, *Minimalistic*, and *Photorealistic*. We fine-tuned Flux.1-dev (Black Forest Labs 2024) using LoRA (Hu et al. 2022) to obtain a specialized T2I model for each background type, respectively.

To further improve the quality of the output images, we integrated two core techniques into the training process. First, we implemented a *resolution-based data bucketing strategy*, grouping training images by resolution and aspect ratio. This ensured that each batch contained visually consistent samples, preserving artistic composition and avoiding instability caused by mixing images with varying resolutions. Second, we introduced a *dynamic prompt sampling mechanism*. Instead of using single prompts, we set up three hierarchical prompt levels, where basic prompts describe core visual elements and themes; medium prompts add artistic style; and detailed prompts precisely specify color, composition, and underlying meaning. Note that these three levels differ from those in the blueprint creation stage (the first stage) and are designed exclusively for the training of this stage. This hierarchical approach enables the generation model to adapt to diverse textual descriptions, significantly improving the diversity and accuracy of generated outputs. To train the four T2I models, we construct a tailored dataset that pairs hierarchical prompts and background images. For inference, the background graphical attributes generated from the first stage are fed into the model for background generation.

Unified Layout-Text Rendering In the third stage, PosterVerse consolidates the foundational outputs from previous stages for complete poster generation, as depicted in the bottom left panel of Fig. 2. Unlike previous models that output static posters (Yang et al. 2024b; Gao et al. 2025b), we innovatively choose HTML as the output format due to the following merits. (1) HTML's inherent text scalability perfectly addresses small-scale, high-density text synthesis that previous models struggle with. (2) HTML enables flexible post-editing of fonts, text, and layouts for frequently changing requirements; (3) HTML simultaneously covers layout and text rendering, avoiding the complexity of separate planning and generation. Specifically, we fine-tune Qwen2.5-VL-7B (Bai et al. 2025) to generate HTML files using textual content and parameters from the first stage plus background images from the second stage. The model is tasked with layout planning, graphical rendering, and text rendering as per the given inputs, producing a cohesive and aesthetically pleasing design. The dataset used for training is described in the next section. Finally, the generated HTML file can be rendered in a web browser, ensuring 100% text fidelity and high-quality visual presentation. PosterVerse provides users with both editable HTML files and final image assets suitable for various distribution purposes.

PosterDNA Dataset

Currently, while some layout generation datasets (Zhou et al. 2022; Hsu et al. 2023) have been published, they largely suffer from small scale and limited diversity. Also, none of them has explored a flexible, editable poster format that supports post-generation customization. To fill this gap, we present PosterDNA, consisting of three specialized subsets: blueprint-creation (57,000 samples), graphic generation (100,000 samples), and unified text-layout creation (9,000 samples), responsible for PosterVerse's three-stage training and evaluation. An intuitive description of data construction is demonstrated in Fig. 3. We elaborate on the three training subsets, followed by the testing methodology.

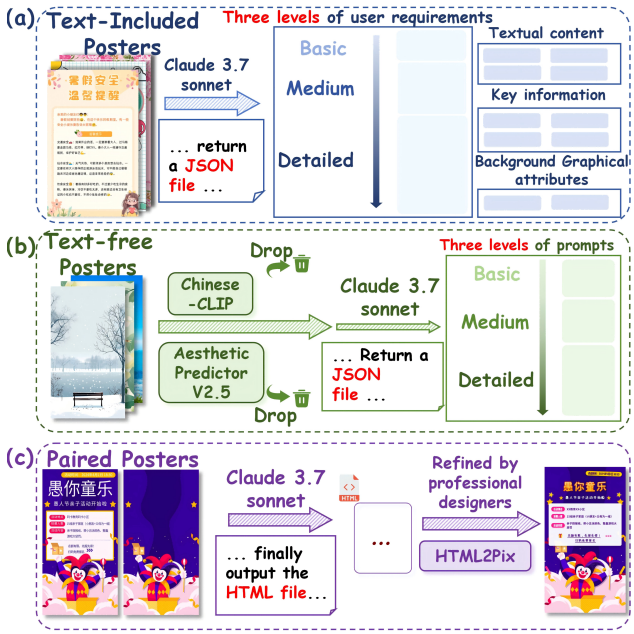

Figure 3: Overview of the three core components in the PosterDNA dataset generation pipeline.

Blueprint Creation Subset The first stage of PosterVerse transforms user requirements into a consistently detailed blueprint regardless of the input's detail level. Hence, we curated 57,000 high-quality posters featuring rich, dense textual content to form the blueprint creation subset, as illustrated in the top panel of Fig. 3. To begin with, we employed Claude 3.7 Sonnet (Anthropic 2024) to reverse-engineer three different detail levels (basic, medium, detailed) of user requirements used to generate these posters. Subsequently, we use Claude to extract refined requirement clues for each poster, including *Key Information*, *Background Graphical Attributes*, and *Textual Content*. *Key Information* includes poster's theme, dominant colors (hex codes), intended purpose (e.g., event promotion), and key visual elements (such as icons, objects, and decorations). *Background Graphical Attributes* includes the poster's design style (e.g., Design-Oriented style) and graphical captions, in which the caption is the output portion that should be insensitive to the input requirements' detail levels. *Textual Content* contains titles, subtitles, main body text, and contact information. During training, the three levels of user requirements are addressed to the Qwen2.5-14B for fine-tuning, with *Textual Content*, *Key Information*, and *Background Graphical Attributes* as supervision labels.

Method	CR ↑	F1 ↑	FID ↓	Ove. ↓	User Study ↑		GPT-4o Evaluations ↑				
					Vote	Ave.	Ave.	PA.	TA.	IQ.	LC.
Text-to-Image (T2I) Models											
Kolors *(Open)*	4.59%	2.30%	123.41	0.0138	0%	2.26	5.59	5.64	3.84	6.92	5.95
Cogview4 *(Open)*	31.13%	21.01%	78.20	0.0140	0%	4.67	6.03	6.42	5.56	5.81	6.33
Ideogramv3 *(Close)*	30.57%	19.71%	97.40	0.0167	1%	4.81	6.53	6.59	6.10	6.87	6.57
Klingv2 *(Close)*	35.27%	26.81%	71.64	0.0118	1%	5.13	6.21	6.33	5.49	6.72	6.30
Jimeng2.1 *(Close)*	33.25%	23.01%	68.70	0.0112	0%	5.53	6.25	6.29	5.64	6.77	6.31
Unified Generative Models											
Seedream3.0 *(Close)*	49.91%	39.66%	83.20	0.0103	2%	6.12	7.82	7.99	7.55	**8.03**	7.71
Gemini2.0 *(Close)*	38.22%	28.46%	74.39	0.0086	1%	4.53	6.22	6.49	5.69	6.53	6.18
GPT-4o *(Close)*	49.73%	48.49%	89.39	0.0106	24%	6.30	7.87	7.93	7.92	7.89	**7.73**
Specialized Poster Generation Models											
Anytext2 *(Open)*	32.57%	26.46%	87.68	0.0105	0%	2.51	3.78	3.78	3.30	4.27	3.77
PosterMaker *(Open)*	27.25%	25.09%	78.01	0.0098	0%	3.08	4.74	4.82	3.95	5.44	4.75
Bizgen *(Open)*	14.67%	13.32%	101.86	0.0094	0%	2.05	3.46	3.06	3.19	4.25	3.34
PosterVerse (Ours)	**92.33%**	**78.58%**	**62.54**	**0.0027**	**71%**	**6.85**	**8.02**	**8.19**	**8.51**	7.66	7.72

Table 1: Comparison of PosterVerse with existing methods. 'Ave.', 'PA.', 'TA.', 'IQ.', and 'LC.' indicate Average, Prompt Adherence, Text Accuracy, Image Quality, and Layout & Composition, respectively. *Open* and *Close* denote open-source and closed-source. The inputs for all methods are aligned with user requirements at the detailed level on the testing set.

Graphic Generation Subset In the second stage, PosterVerse fine-tuned Flux.1-dev using LoRA to obtain four specialized instances for generating distinct background types. To support this training process, we constructed the graphic generation subset (middle of Fig. 3), constituting 100,000 text-free poster background graphics with diverse styles and designs. To retain only high-quality samples, we designed a multi-stage pipeline to verify image resolution and file format, as well as perform deduplication using a pretrained Chinese-CLIP (Yang et al. 2022) model and aesthetic filtering with Aesthetic Predictor V2.5. Following quality filtering, we exploited Claude to generate prompts for each background, forming a prompt-image pair for model training. Corresponding to the dynamic caption sampling mechanism, we instruct Claude to generate three prompts with hierarchical details. Note that only one randomly selected prompt constitutes the input during training.

Unified Layout-Text Rendering Subset The third stage of PosterVerse performs unified layout planning for visual elements and text while rendering typography. This stage requires the refined requirement specifications and a pre-given background image as input and delivers the HTML output file. Hence, we select 9,000 posters from the blueprint creation set to construct a unified text-layout rendering subset, as shown in the bottom panel of Fig. 3. We removed all text in the posters to obtain a text-free version, then paired the original text-included and text-free versions as input to Claude 3.7 Sonnet for generating corresponding HTML representations that capture the original layouts. Each HTML file underwent manual review by professional designers, including correcting positional errors and extracting text errors, adjusting element positions for better aesthetics, etc. These HTML files provide both structural graphical layouts and support diverse text rendering effects, enabling unified text-layout creation. For training, the text-free background along with Key Information and Textual Content extracted from the first stage are fed into the model, and the manually labeled HTML files serve as the supervision.

Additionally, we collected 1000 samples external to the training data to form the test set. To create the ground truth for each sample, they consistently went through the processing of the blueprint creation subset and unified layout-text rendering subsets. The ground-truths include "basic-medium-detailed" user requirements, key information, textual content, paired posters, and corresponding HTML files.

Constructing PosterDNA has consumed four months of manual effort, encompassing workflow design, data curation, and meticulous correction. Notably, the manual correction phase was the most intensive, accounting for approximately 80% of the total labor. PosterDNA is the first Chinese poster generation dataset that pioneers in equipping HTML typography files for scalable text rendering, thus fundamentally addressing the challenge of small, large-amount, and high-density text rendering. It not only paves the way for commercial-grade poster design in text-rich cases but also contributes to the development of more dedicated methods.

Experiments

Implementation Details

Model Implementation For blueprint creation, we fine-tune the Qwen-2.5-14B model using full-parameter SFT for 15 epochs at a 1e-5 learning rate on 8 H800 GPUs, completing in 30 hours. For graphical background generation, we fine-tune Flux.1 dev with LoRA (rank 64) at a 5e-4 learning rate for 50 epochs. For unified layout-text rendering, we fine-tune the Qwen2.5-VL-7B model at a 1e-5 learning rate for 50 epochs in 30 hours on 8 H800 GPUs. More details are included in the supplementary materials.

Evaluation We assess our framework using objective quantitative metrics and subjective evaluation from both AI

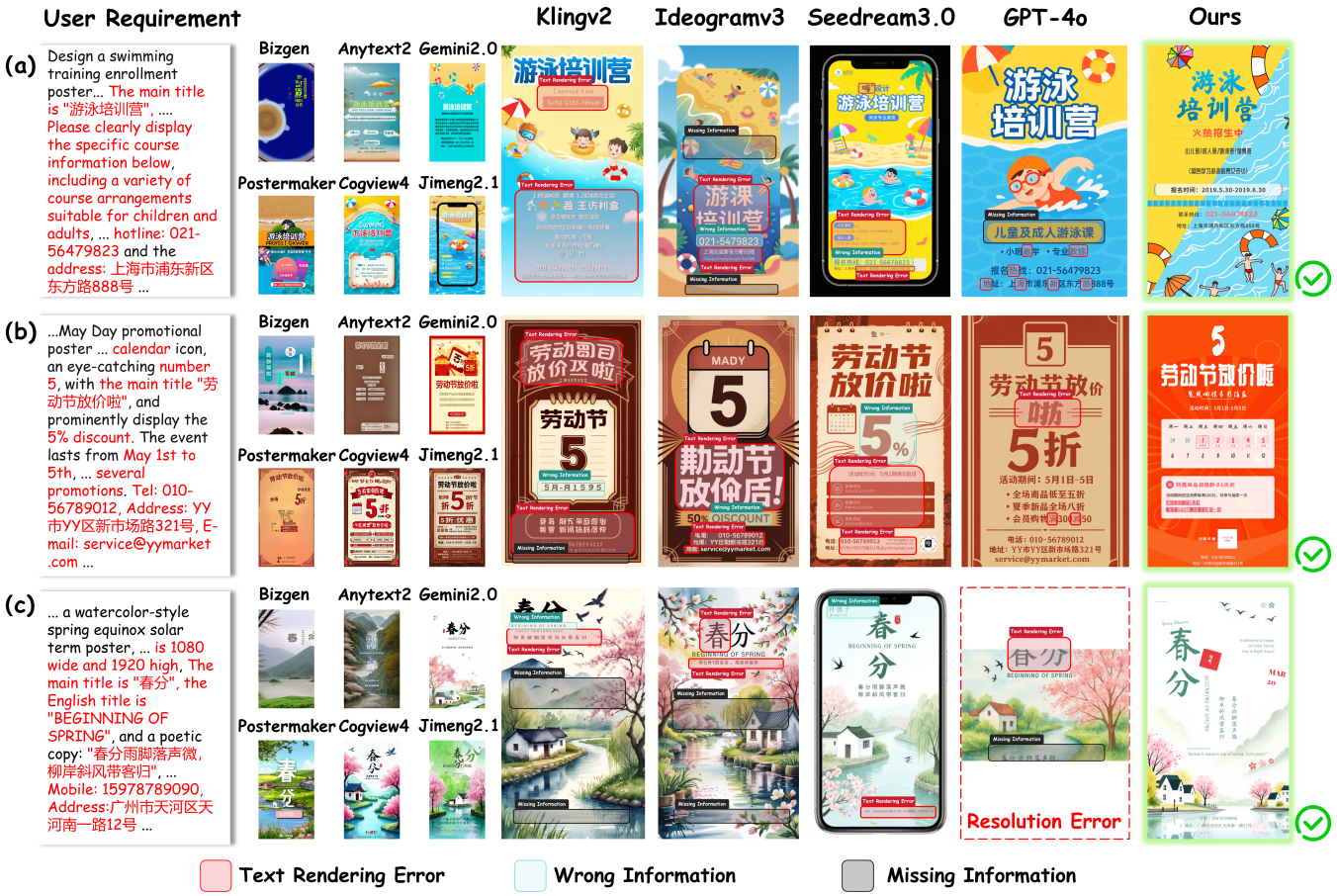

Figure 4: Visual comparison of PosterVerse with state-of-the-art models. The inputs for all methods are aligned with user requirements at the detailed level on the testing set.

and human users. As for objective analysis, we measure text generation accuracy using Correct Rate (CR) and F1 scores (via PPOCRv5 (Cui et al. 2025)) following (Zhang et al. 2025d). We also assess layout fidelity with overlap metrics (Hsu et al. 2023) and quantify perceptual similarity using FID (Heusel et al. 2017). For subjective evaluation, we utilize GPT-4o to provide a detailed, four-dimensional rating (1-10) of prompt adherence, text accuracy, image quality, and layout & composition. Furthermore, to evaluate real-world user experience, we conduct a human study where participants use the same four-dimensional rubric and also vote for their preferred model outputs. This dual approach to qualitative feedback ensures our evaluation is both comprehensive and truly reflective of the user experience.

Comparison with Existing Methods

We compare our method with 11 representative methods spanning three paradigms, including Text-to-Image models (Kolors (Team, Kolors 2024), Cogview4 (Zheng et al. 2024), Ideogramv3 (Ideogram AI 2025), Klingv2 (Kling AI 2025), and Jimengv2,1 (Hu et al. 2025)), unified generative models (Seedream3.0 (Gao et al. 2025a), Gemini2.0-Flash-Gen (Gemini et al. 2023), and GPT-4o (OpenAI 2025)), and specialized poster generation models (Anytext2 (Tuo, Geng, and Bo 2024), PosterMaker (Gao et al. 2025b), and Bizgen (Peng et al. 2025)).

Quantitative Comparison Quantitative comparison results between PosterVerse and baseline methods are presented in Tab. 1, where consistent user requirements are used as inputs for fairness. On multiple objective metrics, PosterVerse achieves the best performance with a CR score of 92.33% and an F1 score of 78.58%, surpassing existing models by at least 42.42% in CR and 30.09% in F1 score. This not only demonstrates its effectiveness in producing accurate and readable text content but also highlights its ability to align closely with user requirements for copywriting. Furthermore, PosterVerse achieves an FID score of 62.54, highlighting the ability of PosterVerse to generate images with high perceptual similarity to poster visuals. In contrast, other models often exhibit text rendering errors and subpar background quality, resulting in lower FID scores. In addition, PosterVerse achieves the best Overlap score of 0.0027, reflecting its superior layout quality.

Regarding subjective assessment, PosterVerse demonstrates exceptional performance across GPT-4o's four evaluation dimensions, achieving the highest overall average

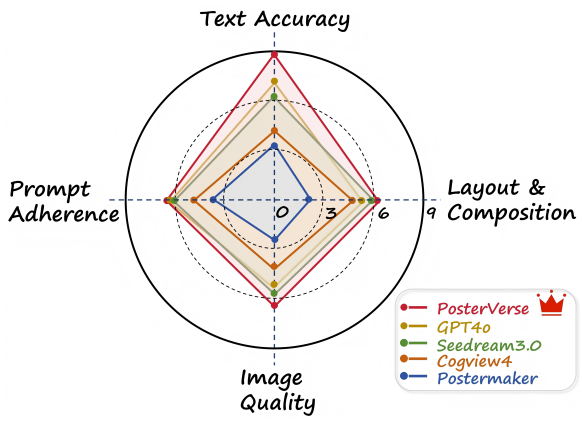

Figure 5: Radar chart of the user study in four dimensions.

score, particularly excelling in Prompt Adherence and Text Accuracy. Furthermore, we invited 30 poster design users to conduct a user study comparing the poster generation performance of different models. As shown in Fig. 5, partial results of the user study across the four key dimensions (aligned with the GPT-4o evaluation) are presented, with the corresponding average scores listed in the 7th column of Tab. 1. PosterVerse demonstrated comparable performance to GPT-4o and Seedream3.0 across most dimensions, showing a significant advantage in textual accuracy. For the voting user study, users were asked to compare the posters generated by the models under the same input requirements and select the one with the best overall impression and practicality. As shown in the 6th column of Tab. 1, PosterVerse received 71% of the votes, significantly outperforming the other methods.

Qualitative Comparison We present the visualizations of PosterVerse and the existing methods on the testing set. As illustrated in Fig. 4, we identify three main types of errors in existing methods. (1) The areas marked with red boxes highlight *text rendering errors*. Models like Anytext2 and Cogview4 struggle to generate completely accurate text regardless of font size, while Seedream3.0 and GPT-4o perform better but still face significant challenges when handling dense text and small fonts. In contrast, PosterVerse is capable of accurate text rendering. (2) The areas marked with green boxes indicate instances where the rendered text contains *wrong information*. For example, as shown in Fig. 4 (a), while the user specified the phone number on the poster as "021-56479823, Ideogramv3 incorrectly rendered it as "021-5479823". (3) The areas marked with gray boxes indicate cases where *missing information* occurs. For instance, while the user requested the poster to include a phone number and physical address, Klingv2 missed rendering the required information. Moreover, GPT-4o often misunderstands user requirements regarding resolution, such as generating a landscape poster when a portrait poster was requested. In contrast, PosterVerse not only effectively aligns with user requirements but also supplements unclear user inputs, providing a more comprehensive and accurate output.

Extended results for English and other languages are included in the supplementary materials.

Method	Basic	Medium	Detailed	Ave. ↑
Seedream 3.0	✓			4.92
		✓		5.33
			✓	6.12
GPT-4o	✓			4.52
		✓		6.45
			✓	6.30
PosterVerse (Ours)	✓			**6.76**
		✓		**6.53**
			✓	**6.85**

Table 2: Results of the user study demonstrating the effectiveness of the DIPR mechanism.

#Line	Basic	Medium	Detailed	FID ↓	IS ↑
1	✗	✗	✓	136.72	62.39
2	✓	✓	✓	**62.54**	**77.85**

Table 3: Ablation study on the effectiveness of the dynamic prompt sampling mechanism.

Ablation Study

We conducted an ablation study through human evaluation (with the same setting as before) to investigate DIPR mechanism's effectiveness. As shown in Tab. 2, we observe that the two representative baseline models are highly sensitive to input detail levels. When given basic requirements, their average ratings are low, indicating their deficiency in generating high-quality posters with brief requirements. In contrast, PosterVerse maintains consistently high performance across all input detail levels, validating DIPR's effectiveness.

Additionally, to further validate that the dynamic prompt sampling mechanism during the second stage of PosterVerse training can enhance the effectiveness of graphical background generation, we conducted a comparison with models trained using only the detailed-level prompt. As shown in Tab. 3, both the FID and CLIP-IS metrics are significantly better when using hierarchical prompts compared to using only the detailed-level prompt.

Conclusion

In this paper, we present PosterVerse, a full-workflow method that seamlessly combines blueprint creation, graphical background generation, and unified layout-text rendering, enabling commercial-grade posters with sophisticated layouts and text-dense designs. Additionally, we introduce PosterDNA, the first high-quality, text-dense poster generation dataset with fine-grained HTML-based specifications, tailored for modular training and validation. Extensive experiments demonstrate PosterVerse's superior performance, significantly outperforming existing methods. The PosterVerse's ability to generate commercial-grade posters directly from natural language prompts, combined with its scalable and editable output format, establishes a new paradigm for automated commercial design and provides a promising solution for marketing and creative industries.

Acknowledgements

This research is supported in part by the National Natural Science Foundation of China (Grant No.:62476093).

References

Anthropic. 2024. Claude Sonnet. Https://www.anthropic.com/claude/sonnet.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Balaji, Y.; Nah, S.; Huang, X.; Vahdat, A.; Song, J.; Zhang, Q.; Kreis, K.; Aittala, M.; Aila, T.; Laine, S.; et al. 2022. ediff-i: Text-to-Image Diffusion Models with an Ensemble of Expert Denoisers. *arXiv preprint arXiv:2211.01324*.

Black Forest Labs. 2024. Flux. Https://github.com/black-forest-labs/flux.

Chen, H.; Xu, X.; Li, W.; Ren, J.; Ye, T.; Liu, S.; Chen, Y.-C.; Zhu, L.; and Wang, X. 2025a. POSTA: A Go-to Framework for Customized Artistic Poster Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 28694–28704.

Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2023. TextDiffuser: Diffusion Models as Text Painters. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 9353–9387.

Chen, J.; Huang, Y.; Lv, T.; Cui, L.; Chen, Q.; and Wei, F. 2024. TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering. In *European Conference on Computer Vision (ECCV)*, 386–402. Cham.

Chen, S.; Lai, J.; Gao, J.; Ye, T.; Chen, H.; Shi, H.; Shao, S.; Lin, Y.; Fei, S.; Xing, Z.; et al. 2025b. PosterCraft: Rethinking High-Quality Aesthetic Poster Generation in a Unified Framework. *arXiv preprint arXiv:2506.10741*.

Cui, C.; Sun, T.; Lin, M.; Gao, T.; Zhang, Y.; Liu, J.; Wang, X.; Zhang, Z.; Zhou, C.; Liu, H.; Zhang, Y.; Lv, W.; Huang, K.; Zhang, Y.; Zhang, J.; Zhang, J.; Liu, Y.; Yu, D.; and Ma, Y. 2025. PaddleOCR 3.0 Technical Report. arXiv:2507.05595.

Gao, Y.; Gong, L.; Guo, Q.; Hou, X.; Lai, Z.; Li, F.; Li, L.; Lian, X.; Liao, C.; Liu, L.; et al. 2025a. Seedream 3.0 Technical Report. *arXiv preprint arXiv:2504.11346*.

Gao, Y.; Lin, Z.; Liu, C.; Zhou, M.; Ge, T.; Zheng, B.; and Xie, H. 2025b. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 8083–8093.

Gemini, T.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Gupta, K.; Lazarow, J.; Achille, A.; Davis, L. S.; Mahadevan, V.; and Shrivastava, A. 2021. LayoutTransformer: Layout Generation and Completion With Self-Attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 1004–1014.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *NeurIPS*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 6840–6851.

Horita, D.; Inoue, N.; Kikuchi, K.; Yamaguchi, K.; and Aizawa, K. 2024. Retrieval-Augmented Layout Transformer for Content-Aware Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 67–76.

Hsu, H. Y.; He, X.; Peng, Y.; Kong, H.; and Zhang, Q. 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.

Hsu, H. Y.; and Peng, Y. 2025. PosterO: Structuring Layout Trees to Enable Language Models in Generalized Content-Aware Layout Generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8117–8127.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. *International Conference on Learning Representations (ICLR)*, 1(2): 3.

Hu, X.; Chen, H.; Qi, Z.; Zhang, H.; Hong, D.; Shao, J.; and Wu, X. 2025. DreamPoster: A Unified Framework for Image-Conditioned Generative Poster Design. *arXiv preprint arXiv:2507.04218*.

Hui, M.; Zhang, Z.; Zhang, X.; Xie, W.; Wang, Y.; and Lu, Y. 2023. Unifying Layout Generation With a Decoupled Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 1942–1951.

Ideogram AI. 2025. Ideogram v3. Https://ideogram.ai/launch.

Jia, P.; Li, C.; Yuan, Y.; Liu, Z.; Shen, Y.; Chen, B.; Chen, X.; Zheng, Y.; Chen, D.; Li, J.; et al. 2023. Cole: A Hierarchical Generation Framework for Multi-Layered and Editable Graphic Design. *arXiv preprint arXiv:2311.16974*.

Kling AI. 2025. Klingv2.

Lab, D. 2023. DeepFloyd-IF.

Lakhanpal, S.; Chopra, S.; Jain, V.; Chadha, A.; and Luo, M. 2025. Refining Text-to-Image Generation: Towards Accurate Training-Free Glyph-Enhanced Image Generation. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 4372–4381.

Li, F.; Liu, A.; Feng, W.; Zhu, H.; Li, Y.; Zhang, Z.; Lv, J.; Zhu, X.; Shen, J.; Lin, Z.; et al. 2023a. Relation-aware diffusion model for controllable poster layout generation. In *CIKM*, 1249–1258.

Li, Z.; Li, F.; Feng, W.; Zhu, H.; Li, Y.; Zhang, Z.; Lv, J.; Shen, J.; Lin, Z.; Shao, J.; et al. 2023b. Planning and Rendering: Towards Product Poster Generation with Diffusion Models. *arXiv preprint arXiv:2312.08822*.

Lin, J.; Guo, J.; Sun, S.; Yang, Z.; Lou, J.-G.; and Zhang, D. 2023a. LayoutPrompter: Awaken the Design Ability of Large Language Models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 43852–43879.

Lin, J.; Zhou, M.; Ma, Y.; Gao, Y.; Fei, C.; Chen, Y.; Yu, Z.; and Ge, T. 2023b. AutoPoster: A highly Automatic and Content-Aware Design System for Advertising Poster Generation. In *Proceedings of the 31st ACM International Conference on Multimedia (MM)*, 1250–1260.

Ma, J.; Zhao, M.; Chen, C.; Wang, R.; Niu, D.; Lu, H.; and Lin, X. 2023. Glyphdraw: Seamlessly Rendering Text with Intricate Spatial Structures in Text-to-Image Generation. *arXiv preprint arXiv:2303.17870*.

Nonghai Zhang, H. T. 2024. Text-to-Image Synthesis: A Decade Survey. *arXiv preprint arXiv:2411.16164*.

OpenAI. 2025. Introducing GPT-4o Image Generation. https://openai.com/index/introducing-4o-image-generation/.

Peng, Y.; Xiao, S.; Wu, K.; Liao, Q.; Chen, B.; Lin, K.; Huang, D.; Li, J.; and Yuan, Y. 2025. Bizgen: Advancing Article-Level Visual Text Rendering for Infographics Generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, 23615–23624.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Reed, S.; Akata, Z.; Yan, X.; Logeswaran, L.; Schiele, B.; and Lee, H. 2016. Generative Adversarial Text to Image Synthesis. In *Proceedings of The 33rd International Conference on Machine Learning (ICML)*, volume 48, 1060–1069.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 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.

Seol, J.; Kim, S.; and Yoo, J. 2024. PosterLlama: Bridging Design Ability of Language Model to Content-Aware Layout Generation. In *European Conference on Computer Vision (ECCV)*, 451–468.

Stability AI. 2024. Stable Diffusion 3.5. Https://github.com/Stability-AI/sd3.5.

Team, Kolors. 2024. Kolors: Effective Training of Diffusion Model for Photorealistic Text-to-Image Synthesis. *arXiv preprint*.

Tuo, Y.; Geng, Y.; and Bo, L. 2024. Anytext2: Visual Text Generation and Editing with Customizable Attributes. *arXiv preprint arXiv:2411.15245*.

Wang, S.; Ge, Y.; Chen, L.; Zhou, H.; Wang, Q.; Cheng, X.; and Yuan, L. 2024. Prompt2poster: Automatically Artistic Chinese Poster Creation from Prompt Only. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM)*, 10716–10724.

Wang, Z.; Bao, J.; Gu, S.; Chen, D.; Zhou, W.; and Li, H. 2025. DesignDiffusion: High-Quality Text-to-Design Image Generation with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 20906–20915.

Xie, Y.; Zhang, J.; Chen, P.; Wang, Z.; Wang, W.; Gao, L.; Li, P.; Sun, H.; Zhang, Q.; Qiao, Q.; et al. 2025. TextFlux: An OCR-Free DiT Model for High-Fidelity Multilingual Scene Text Synthesis. *arXiv preprint arXiv:2505.17778*.

Yang, A.; Pan, J.; Lin, J.; Men, R.; Zhang, Y.; Zhou, J.; and Zhou, C. 2022. Chinese CLIP: Contrastive Vision-Language Pretraining in Chinese. *arXiv preprint arXiv:2211.01335*.

Yang, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Li, C.; Liu, D.; Huang, F.; Wei, H.; et al. 2024a. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

Yang, T.; Luo, Y.; Qi, Z.; Wu, Y.; Shan, Y.; and Chen, C. W. 2024b. PosterLLaVa: Constructing a Unified Multi-modal Layout Generator with LLM. arXiv:2406.02884.

Yang, Y.; Gui, D.; YUAN, Y.; Liang, W.; Ding, H.; Hu, H.; and Chen, K. 2023. GlyphControl: Glyph Conditional Control for Visual Text Generation. In *NeurIPS*, volume 36, 44050–44066.

Zhang, J.; Zhou, Y.; Gu, J.; Wigington, C.; Yu, T.; Chen, Y.; Sun, T.; and Zhang, R. 2025a. ARTIST: Improving the Generation of Text-Rich Images with Disentangled Diffusion Models and Large Language Models. In *IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 1268–1278.

Zhang, L.; Chen, X.; Wang, Y.; Lu, Y.; and Qiao, Y. 2024. Brush Your Text: Synthesize Any Scene Text on Images via Diffusion Model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7): 7215–7223.

Zhang, P.; Xu, H.; Zhang, J.; Xu, G.; Zheng, X.; Yang, Z.; Liu, J.; Zhang, Y.; and Jin, L. 2025b. Aesthetics is Cheap, Show me the Text: An Empirical Evaluation of State-of-the-Art Generative Models for OCR. *arXiv preprint arXiv:2507.15085*.

Zhang, P.; Zhang, J.; Cao, J.; Li, H.; and Jin, L. 2025c. Smaller But Better: Unifying Layout Generation with Smaller Large Language Models. *International Journal of Computer Vision (IJCV)*, 133: 3891–3917.

Zhang, Y.; Zhu, Y.; Peng, D.; Zhang, P.; Yang, Z.; Yang, Z.; Yao, C.; and Jin, L. 2025d. Hiercode: A lightweight hierarchical codebook for zero-shot Chinese text recognition. *Pattern Recognition*, 158: 110963.

Zhao, Y.; and Lian, Z. 2024. UDiffText: A Unified Framework for High-Quality Text Synthesis in Arbitrary Images via Character-Aware Diffusion Models. In *European Conference on Computer Vision (ECCV)*, 217–233.

Zheng, W.; Teng, J.; Yang, Z.; Wang, W.; Chen, J.; Gu, X.; Dong, Y.; Ding, M.; and Tang, J. 2024. Cogview3: Finer and Faster Text-to-Image Generation via Relay Diffusion. In *European Conference on Computer Vision (ECCV)*, 1–22.

Zhou, M.; Xu, C.; Ma, Y.; Ge, T.; Jiang, Y.; and Xu, W. 2022. Composition-Aware Graphic Layout GAN for Visual-Textual Presentation Designs. *arXiv preprint arXiv:2205.00303*.