This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

POSTA: A Go-to Framework for Customized Artistic Poster Generation

Haoyu Chen[1†], Xiaojie Xu[1†], Wenbo Li[2], Jingjing Ren[1], Tian Ye[1],
Songhua Liu[3], Ying-Cong Chen[1], Lei Zhu[1,4*], Xinchao Wang[3*]

[1]The Hong Kong University of Science and Technology (Guangzhou) [2]The Chinese University of Hong Kong
[3]National University of Singapore [4]The Hong Kong University of Science and Technology

Project page: https://haoyuchen.com/POSTA

[†]Equal contribution * Corresponding author



Figure 1. Generated results using our *POSTA* framework. The background, layout, and typographical designs are fully crafted from text inputs, showcasing the framework's capability to produce cohesive and visually engaging elements solely through textual guidance.

Abstract

Poster design is a critical medium for visual communication. Prior work has explored automatic poster design using deep learning techniques, but these approaches lack text accuracy, user customization, and aesthetic appeal, limiting their applicability in artistic domains such as movies and exhibitions, where both clear content delivery and visual impact are essential. To address these limitations, we present POSTA: a modular framework powered by diffusion models and multimodal large language models (MLLMs) for customized artistic poster generation. The framework consists of three modules. Background Diffusion creates a themed background based on user input. Design MLLM then generates layout and typography elements that align with and complement the background style. Finally, to enhance the poster's aesthetic appeal, ArtText Diffusion applies additional stylization to key text elements. The final result is a visually cohesive and appealing poster, with a fully modular process that allows for complete customization. To train our models, we develop the PosterArt dataset, comprising high-quality artistic posters annotated with layout, typography, and pixel-level stylized text

segmentation. Our comprehensive experimental analysis demonstrates POSTA's exceptional controllability and design diversity, outperforming existing models in both text accuracy and aesthetic quality.

1. Introduction

Poster design plays a pivotal role in visual communication, especially in advertising, education, and the arts. Designers must thoughtfully integrate background, layout, color schemes, and typography to create compelling visuals. Generative models and transformers have already found numerous applications [2–4, 6–8, 21, 22, 24, 30, 31, 34, 36, 37, 40–42, 52–59, 63, 67, 68]. While many studies have explored automatic poster design using planning and generative models [5, 9, 26, 28, 46, 64], challenges in stability, customization, and flexibility prevent their widespread use in artistic domains such as concerts, movies, and exhibitions—areas that demand both clear content delivery and striking aesthetics. We propose that an effective poster design framework for these artistic fields must meet three key criteria: **(1) Text Accuracy**: Text must be precise and clear, as it is crucial for conveying the intended message. **(2) Customization**: Design elements should be fully editable and customizable to fit specific needs. **(3) Aesthetics**: The composition should be visually harmonious, ensuring that the design elements complement the overall aesthetic.

Current methods face significant challenges in meeting these criteria, as illustrated in Figure 2, including: **(1) Inaccurate Text Generation:** Existing models struggle with generating accurate text, especially for long or complex content. They often produce misspellings or distorted characters, and even when bounding boxes are provided for text placement, the accuracy of longer text remains unreliable, limiting their practical use in real-world applications. **(2) Inflexible Control over Design Elements:** While many generative approaches can produce complete posters, their "one-shot generation" paradigm lacks the fine-grained control essential for professional design. Adjusting specific elements, such as background, text positioning, or font choices, is often cumbersome and inefficient. **(3) Insufficient Aesthetic Quality:** Existing methods tend to prioritize functionality over artistry, resulting in generated posters that lack the professional visual appeal necessary for commercial use. **(4) Lack of a Systematic Workflow:** Most current approaches focus on optimizing individual aspects of poster design, such as layout or typography, without addressing the full design process holistically. This fragmented approach makes it difficult to achieve a seamless, comprehensive workflow for poster creation.

To address these challenges, we propose POSTA, a modular framework for end-to-end intelligent poster creation. Combining generative models with Multi-Modal Large Language Models (MLLMs), POSTA offers a highly controllable and customizable solution. It overcomes key limitations through the integration of three components: **(1) Background Generation Module:** Utilizing a state-of-the-art diffusion model and several meticulously trained LoRA models [14], this module generates professional-grade backgrounds. Users also have the option to upload custom backgrounds, providing significant flexibility. **(2) Design Planning Module:** Leveraging MLLMs, this module intelligently plans the layout, text placement, and font attributes. The generated text is rendered based on these predictions, ensuring 100% accuracy. Additionally, all text elements are fully editable, allowing for efficient customization. **(3) Artistic Text Stylization Module:** To enhance stylistic diversity and aesthetic appeal, we introduce a module that employs a mask-guided inpainting model [18] to seamlessly integrate artistic text effects, ensuring visual harmony with the background. The entire workflow aligns with the pipeline of professional designers.

To support design planning and artistic text stylization, we introduce the *PosterArt* dataset, which is comprised of two parts: *PosterArt-Design*, containing high-quality artistic posters with layouts, typographic elements, and backgrounds, and *PosterArt-Text*, including pixel-wise segmentation, descriptions, and visual results of artistic text.

In summary, this work presents a systematic solution that can be seamlessly integrated into real-world design workflows, enabling designers to produce professional-quality outputs with ease. Our contributions are as follows.

- We propose a systematic, modular workflow for highly controllable and reliable poster generation at a professional level, enabling users to easily adjust every aspect of the poster—including background, layout, and text—ensuring full customization.
- We introduce effective text rendering and stylization modules that guarantee accurate text generation while enhancing the overall aesthetic appeal of the poster.
- We demonstrate the versatility of our approach across various poster creation tasks, including text-guided poster generation, reference-based generation, and poster editing. Extensive experimental results validate the superiority of our method over existing methods.
- We contribute the *PosterArt* dataset, a high-quality, professionally curated resource comprising two subsets: *PosterArt-Design*, which focuses on layout and typography, and *PosterArt-Text*, which provides artistic text stylization and segmentation.

2. Related Work

2.1. Text-Aware Image Generation

Text-aware image generative models offer advantages for text-centric creative applications such as logo and simple poster design. AnyText [46] introduced a diffusion pipeline with an auxiliary latent module and text embedding using OCR-based stroke encoding. Similarly, GlyphCon-

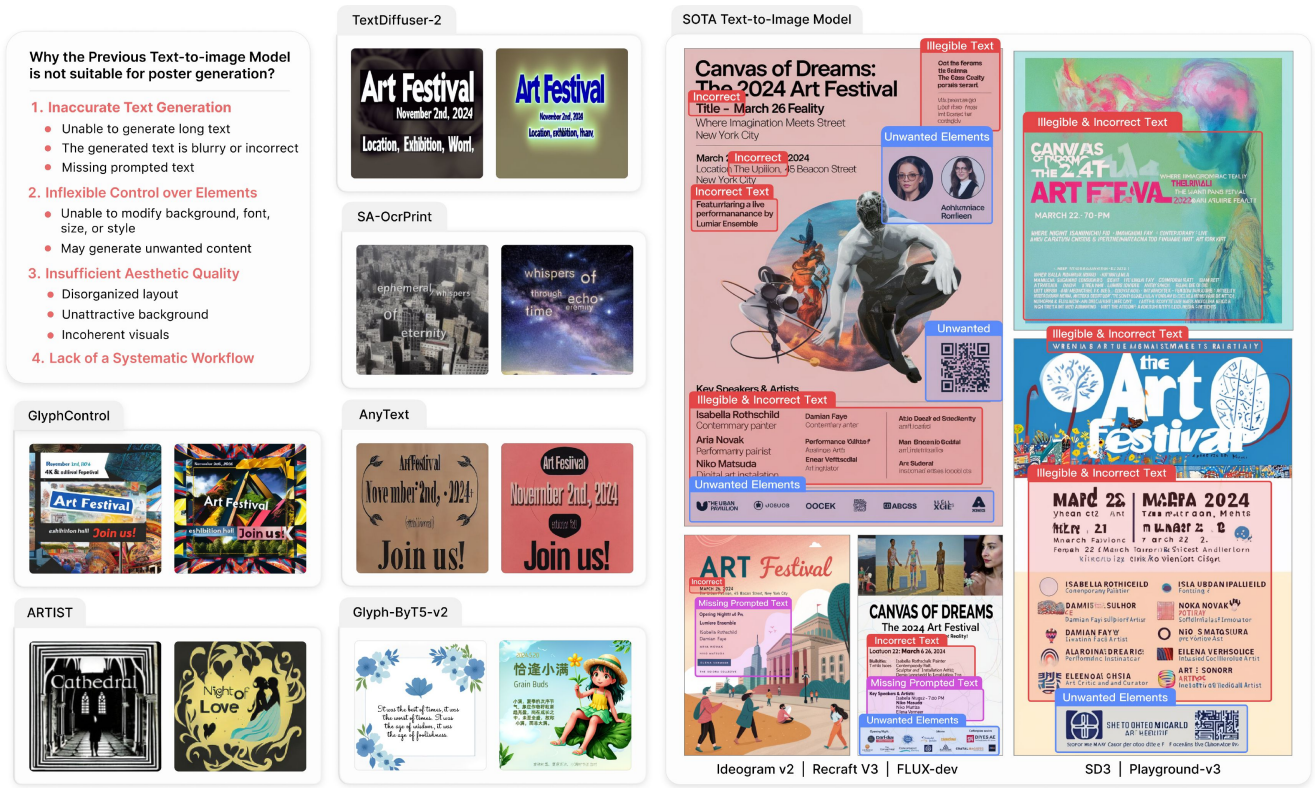


Figure 2. Our motivation stems from limitations of current methods for poster generation, which often struggle with issues like text inaccuracy, limited customization, and insufficient aesthetic quality.

trol [64] enhanced Stable Diffusion [33, 38] with glyph instructions, while Glyph-ByT5-v2 [28] expanded multilingual text rendering through step-aware preference learning. ARTIST[69] used a disentangled architecture to capture text structure features, and TextDiffuser-2 [5] integrated large language models for layout modification through conversational interaction. limiting user adjustments—crucial in professional design. Additionally, text rendering remains error-prone with complex typography. Our modular approach addresses this by ensuring text accuracy through typography prediction prior to artistic stylization.

2.2. Poster Layout Planning

Poster layout planning has advanced from simple templates to sophisticated learning-based systems. PosterLayout introduced content-aware layouts with GAN-based design sequence formation for arranging elements. RADM [23] refined it through a relation-aware diffusion model that aligns visual, textual, and geometric relationships. The integration of MLLMs [10, 44, 50, 65, 66, 71] has advanced layout generation [15, 29, 61]. PosterLLaVa [62] introduced a MLLM framework, enabling flexible layout planning based on visual and textual constraints through language specifications. These methods focus on basic layout prediction, overlooking key typographic elements like font, size, and orientation while *POSTA* predicts these attributes jointly, ensuring a more cohesive and consistent design.

2.3. Artistic Text Generation

Artistic text generation serves as a powerful tool for automating the creation of artistic visuals. Approaches evolving from traditional stroke or patch-based methods to deep learning techniques like Generative Adversarial Networks [19, 20, 25, 32, 39, 43, 60, 72]. Recently, large-scale models like Anything-to-Glyph [49] and WordArt [11] have enabled the generation of artistic text images. A detailed review is in [1]. Although these methods create vivid text effects, they lack background context—for example, vintage posters often feature worn text to convey age. Our text stylization is trained with context, ensuring text effects harmonize with the style and semantics of the background.

2.4. Modular Graphic Design

The complete poster design process encompasses background generation, layout planning, and text rendering. However, existing approaches typically address only one of these components. To our knowledge, COLE[17] is the only framework to propose a comprehensive workflow for this task. The COLE system integrates multiple fine-tuned LMMs, and Diffusion Models, each tailored for design-aware layer-wise captioning, layout planning, reasoning, and image and text generation. However, its multi-step process is complex and requires extensive intermediate checks, making it time-consuming. OpenCOLE[16] simplifies some components but is trained on datasets outside the



Figure 3. Overview of *PosterArt-Text* dataset. It contains extensive segmentation and corresponding descriptions of texts with diverse artistic styles, primarily sourced from artistic posters such as those for movies, album covers, and similar media.

artistic domain, limiting its ability to generate aesthetically pleasing results. *POSTA* streamlines layout and typographic elements, leveraging a specialized artistic text stylization module to create harmonious, visually engaging text. Powered by curated data tailored for artistic posters, our results significantly surpass those of concurrent approaches.

3. PosterArt Dataset

To support the development of our framework and enhance its ability to generate high-quality, professional poster designs, we introduce the *PosterArt* dataset. This dataset is designed specifically for poster creation tasks and is divided into two subsets: *PosterArt-Design*, which focuses on layout and typography, and *PosterArt-Text*, which is dedicated to artistic text stylization.

PosterArt-Design To facilitate the generation of visually compelling, professionally designed poster layouts and typography, we developed *PosterArt-Design*. Specifically, *PosterArt-Design* consists of over 2,000 poster backgrounds collected from the Internet, with additional aesthetically pleasing layout and typography information crafted by professional designers. For each text element, we annotated key attributes—including coordinates, font type, font size, color, alignment, and rotation angle, as shown in Figure 4. These annotations provide essential design information for fine-tuning MLLMs, enabling guided text placement and style generation. A brief user description accompanies each element to further inform the model.

In contrast to previous layout datasets, which comprise commercial ads or randomly sourced social media posters (Figure 4, bottom right) that often feature disorganized and aesthetically inferior designs lacking essential typographic elements like font type, size, and color, our dataset is curated to include only high-quality artistic posters crafted by professional designers. Each poster in our dataset exemplifies a high aesthetic standard and includes key typographic elements critical for effective design.

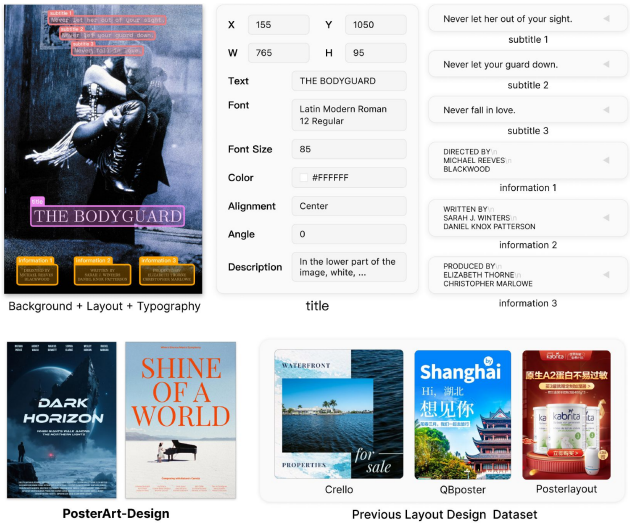


Figure 4. A sample of *PosterArt-Design* (top). *PosterArt-Design* vs. previous layout datasets (bottom). Our dataset is crafted by expert designers who carefully incorporate elements into backgrounds, including deliberate layout and typography information.

PosterArt-Text *PosterArt-Text* focuses on the artistic text regions in poster titles, as illustrated in Figure 3. This dataset aims to empower models to generate text that is both visually appealing and seamlessly integrated with the poster's background and design elements, ensuring a harmonious and cohesive appearance. We created this dataset by collecting a large number of visually compelling posters, generating descriptive captions via GPT-4V[13], and manually segmenting title regions. To enhance diversity, we used FLUX[12] to generate text with artistic effects, such as "baroque font made of coral with pearly shine" In total, the dataset comprises about 3,500 detailed text segmentations and descriptions harmonized with their backgrounds.

A notable contribution of *PosterArt-Text* is its dual functionality for both artistic text generation and segmentation tasks. Even advanced segmentation models often struggle to accurately isolate complex title fonts in posters. By pro-

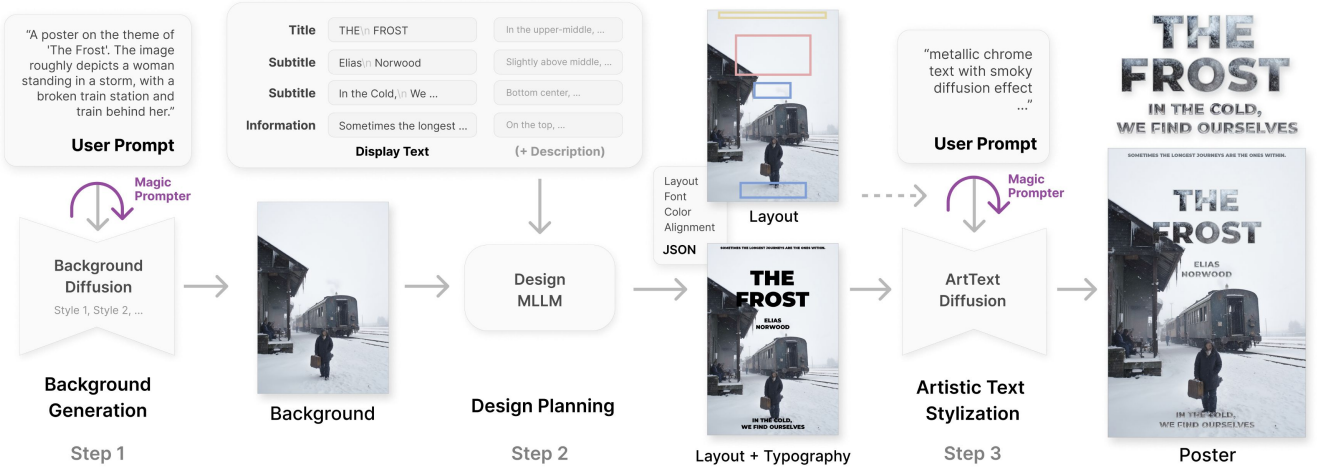


Figure 5. Our *POSTA* pipeline consists of three steps: background generation, design planning, and artistic text stylization. Background Diffusion and ArtText Diffusion are employed to generate backgrounds and text with artistic effects, while the Design MLLM predicts layout and typography information. The GPT-4V-powered Magic Prompter is used to refine prompts based on user descriptions or background images, optimizing input for the diffusion models.

viding high-quality, human-annotated segmentation masks, this dataset offers a valuable resource for improving text segmentation accuracy, especially when the text appears naturally integrated with complex or intricate backgrounds.

4. Method

Our *POSTA* pipeline, as illustrated in Figure 5, consists of three key components: Background Diffusion, Design MLLM, and ArtText Diffusion. Users provide descriptive inputs at each stage, and the framework generates the background image, layout and typography details, and the final poster featuring artistic text elements. The entire process is fully controllable, allowing for customization at each stage.

4.1. Background Generation

A high-quality art poster begins with a well-designed background. In this stage, Background Diffusion, built upon FLUX, is capable of generating a diverse range of premium background images. To improve accuracy across various artistic styles, we collect poster samples categorized into distinct styles such as minimalism, retro, and modern art. These samples are used to train multiple specialized Low-Rank Adaptation (LoRA) blocks, each responsible for controlling specific stylistic elements. To ensure optimal outputs, we integrate an MLLM-powered Magic Prompter to refine user input prompts by adding details.

4.2. Design Planning

To fundamentally address the issue of imprecise text generation while meeting the dual requirements of editability and aesthetic quality, we develop a novel poster design planning method that seamlessly integrates layout, typography generation, and text rendering. Effective text layout prediction requires a deep understanding of the user's design intent and the semantic relationship between text and images. To achieve this, we leverage an MLLM, using LLaVA[27] as the backbone of our design model. Through visual instruction tuning, the model plans the layout and typography for each text element based on user input. Text elements are categorized into three types—title, subtitle, and information—reflecting their prominence in the design. Users can specify text along with optional attributes such as position, alignment, or layout style. Additionally, users may provide high-level layout descriptions to guide the relative positioning of elements. The model outputs key attributes, including position, size, font, color, alignment, and rotation angle, as shown in Figure 4.

For text rendering, to completely resolve the issue of imprecise text generation in generative models, we directly render text based on the font details provided by the model, rather than generating text images. This approach aligns with professional design workflows. By storing text information in an editable vector format, our system enables users to fine-tune design elements such as size, position, font, and alignment after generation.

4.3. Artistic Text Stylization

To further enhance the visual impact of the text—particularly the poster title—and ensure seamless integration with the background while preserving text accuracy, we employ a mask-guided inpainting strategy. The ArtText Diffusion model is based on BrushNet [18], a ControlNet-like [70] architecture designed for the inpainting task. Trained on our high-quality *PosterArt-Text* dataset, this model is capable of understanding the textual descriptions of font styles and generating text with artistic features such as 3D effects, metallic textures, color gradients, and outlines. Guided by text prompts generated according to user input or the background style with Magic



Figure 6. The artistic texts generated by our framework are highly diverse, with each style crafted to suit a specific theme.

Prompter, the model performs local generation to ensure that the text naturally blends with the background. The blending process could be written as:

$$I_{\text{blended}} = M \odot I_1 + (1 - M) \odot I_2 \quad (1)$$

where M, I_1, and I_2 denote the mask, generated image, and original image, respectively. A Gaussian kernel is applied to the mask to achieve smooth boundaries. The entire process achieves stylized artistic text effects that harmonize with the overall design while ensuring text accuracy.

5. Application

5.1. Text-Based Poster Generation

Figure 1 demonstrates the comprehensive capabilities of our *POSTA* framework in generating diverse poster designs. Our system exhibits remarkable versatility in producing various types of visual content, including book covers, movie posters, promotional materials, and event announcements, all generated exclusively from textual descriptions.

5.2. Artistic Text Generation

One critical components of our *POSTA* framework is the ArtText Diffusion module, which enhances aesthetic appeal of the generated posters. This feature allows us to craft unique and visually engaging fonts that go beyond traditional typography, enhancing the overall expressiveness of the designs. As shown in Figure 6, our method is capable of producing a wide variety of artistic fonts, each designed to suit different stylistic themes and visual contexts.

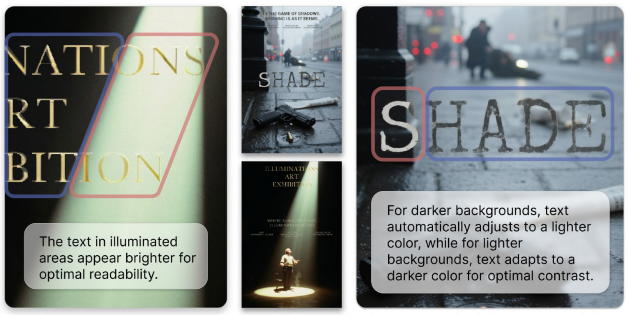


Figure 7. In various background regions, text displays distinct illumination and color effects.

Another key feature of *POSTA* is its ability to seamlessly integrate fonts into the background, ensuring that the fonts do not appear disconnected. This capability sets our framework apart from previous models, which often struggle to harmonize text with the surrounding. As shown in Figure 7, our system dynamically adjusts the font's color, brightness, and contrast to blend naturally with the background, significantly enhancing readability and visual coherence.



Figure 8. Reference-Based Poster Generation.

5.3. Reference-Based Poster Generation

Our framework also enables Reference-Based Poster Generation, as demonstrated in Figure 8. The process starts with image-to-image transfer to remove text while preserving the background. GPT-4V then extracts a detailed description of the reference, guiding the generation of a new layout and typography. Finally, the title is stylized to match the background, with the prompt generated by GPT-4V. This approach creates visually coherent posters that retain the essence of the reference while allowing full customization.

5.4. Demonstration of Editability

A key feature of our framework is its control and editability over design elements, allowing users to modify fonts, layouts, backgrounds, and styles to meet specific needs. As shown in Figure 9, the system generates multiple poster variations from the same prompt, each with unique backgrounds and typography layouts. Users can make precise edits, such as adjusting layouts, changing fonts, or refining font effects (e.g., color, weight, texture). This flexibility



Figure 9. Our framework provides full editability over all design elements-layouts, fonts, styles, and backgrounds.

ensures seamless text-background alignment, enabling iterative improvements toward professional-quality designs.

6. Experiment

6.1. Implementation Details

For Background Diffusion, we train FLUX LoRA[14] at a resolution of 1024, using approximately 50 images per style, with an embedding dimension of 64. For Design MLLM, we use the CLIP pre-trained Vision Transformer (ViT-L/14) [35] as the image encoder to convert input images into visual tokens, and Llama3 7B [45] as the language model. Training was conducted on 2 A800 GPUs. For ArtText Diffusion, we train based on SDXL's BrushNet [18, 33] at a resolution of 1216. Further implementation details are provided in the supplementary materials.

6.2. Qualitative Results and Comparisons

To evaluate the effectiveness of our *POSTA* framework, we conducted a comprehensive comparison with state-of-the-art models, including both popular end-to-end open-source and proprietary commercial models (as shown in Figure 11), along with research works on systematic poster design, namely COLE[17] and OpenCOLE[16] (as shown in Figure 10). As illustrated, *POSTA* significantly outperforms the others on accurate text rendering, user description understanding and artistic stylization, particularly in handling long text sequences.

A major flaw in all other models is their inability to accurately generate even moderately long text. In most cases, the text output is not only incomplete but degenerates into garbled characters. Models such as AnyText [46], TextDiffuser-2 [5], and Flux frequently fail to render titles correctly, resulting in unreadable text. Moreover, these models often produce unwanted content, such as irrelevant symbols or extraneous marks that the user did not specify. This behavior leads to cluttered and unprofessional results, making them unsuitable for high-quality design tasks. In contrast, *POSTA* maintains complete control over elements, ensuring that only the desired content appears, and that it is rendered with precision and clarity, even for longer text.

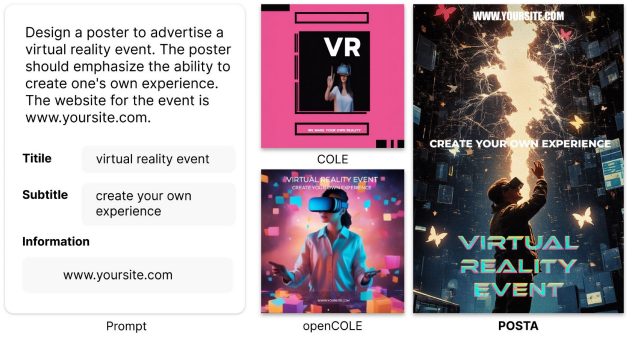


Figure 10. Comparison with COLE and OpenCOLE.

In addition to superior text accuracy, *POSTA* also excels in artistic font quality. While other models struggle to generate visually appealing and well-integrated fonts, *POSTA* produces fonts that are not only aesthetically pleasing but also harmonize with the overall design. This ensures that the text does not appear disjointed or out of place, as is often seen with other methods.

6.3. Quantitative Results and Comparisons

We conduct quantitative analysis comparing our *POSTA* framework with state-of-the-art models, including 20 poster samples generated with the same prompts using advanced general-purpose models (SD3 [9], Playground-v3 [26], FLUX-dev) and graphic design-specific models (Recraft V3[48], Ideogram v2 [47]). Sixty users with experience in AI tools and an arts background, along with GPT-4V, assess key aspects such as visual appeal, text readability, and prompt-image relevance, rating from 1 to 10.

As shown in Figure 12, general-purpose models performed moderately, with FLUX-dev leading among them. Graphic design-specific models, Recraft and Ideogram, outperform open-source models in aesthetics but have notable



Figure 11. Comparisons with state-of-the-art models, encompassing popular open-source and proprietary commercial models. Blue box: word generation errors. Red box: unclear or unrecognizable text. Yellow box: generated content that is irrelevant to the prompt.

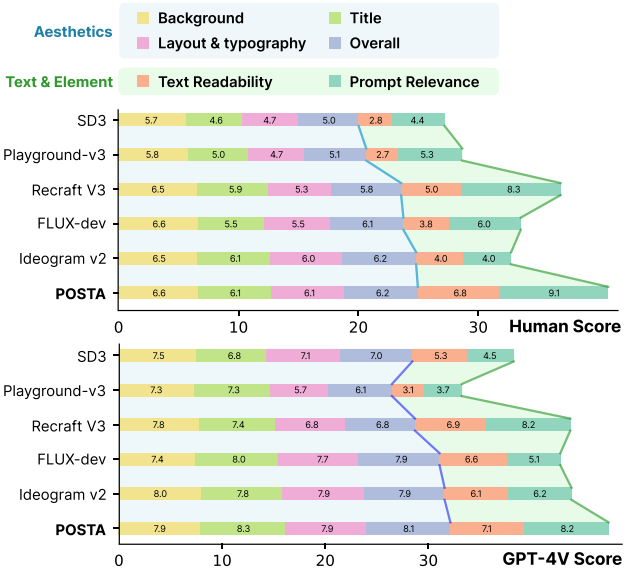


Figure 12. Human and GPT-4V evaluation.

issues. Recraft struggles with longer text sequences, often omitting specified text, while Ideogram introduces irrelevant elements that compromise design quality. In contrast, *POSTA* excels across all criteria, delivering superior aesthetics, text readability, and prompt relevance, making it the most reliable option for professional poster generation.

For text accuracy, we further employ GOT-OCR[51] to detect text on generated posters, calculating the average word-level precision and recall between detected and user-specified text. As shown in Figure 13, *POSTA* achieves the highest level of text consistency with user input, since we generate absolutely accurate text.

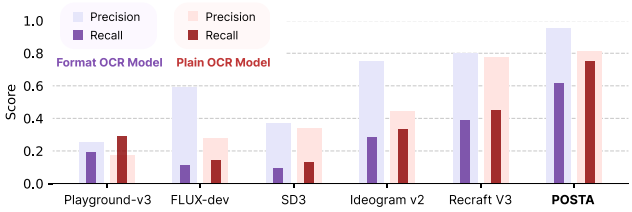


Figure 13. OCR-based comparisons, where "Format / Plain" indicates whether models are designed for electronic print or handwritten characters.

7. Conclusion

In this paper, we present *POSTA*, a framework designed to harness the capabilities of vision diffusion models and multimodal language models for customized artistic poster generation. Specifically, we employ two diffusion models—one for background generation and another for artistic text stylization—along with a multimodal language model for layout and typography planning. To train our models, we introduce *PosterArt* dataset, comprising high-quality artistic posters annotated with layout, typography, and pixel-level stylized text segmentation. Experimental results confirm that *POSTA* effectively produces highly diverse, visually compelling posters with high text accuracy.

As for limitations, our Design MLLM is currently limited to generating relatively simple layout designs and offers a restricted selection of font types, which constrains design diversity and limits the performance of the ArtText Diffusion. This limitation stems from the high standards in artistic data production. In the future, expanded datasets and more advanced model architectures will enhance the scalability.

Acknowledgements

This work is supported by the Guangdong Science and Technology Department (No. 2024ZDZX2004), and the Guangzhou-HKUST(GZ) Joint Funding Program (No. 2023A03J0671), and the Ministry of Education, Singapore, under its Academic Research Fund Tier 2 (Award Number: MOE-T2EP20122-0006).

References

[1] Yuhang Bai, Zichuan Huang, Wenshuo Gao, Shuai Yang, Jiaying Liu, et al. Intelligent artistic typography: A comprehensive review of artistic text design and generation. *APSIPA Transactions on Signal and Information Processing*, 13(1), 2024. 3

[2] Haoyu Chen, Jingjing Ren, Jinjin Gu, Hongtao Wu, Xuequan Lu, Haoming Cai, and Lei Zhu. Snow removal in video: A new dataset and a novel method. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13165–13176. IEEE, 2023. 2

[3] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Haoze Sun, Xueyi Zou, Zhensong Zhang, Youliang Yan, and Lei Zhu. Low-res leads the way: Improving generalization for super-resolution by self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25857–25867, 2024.

[4] Haoyu Chen, Wenbo Li, Jinjin Gu, Jingjing Ren, Sixiang Chen, Tian Ye, Renjing Pei, Kaiwen Zhou, Fenglong Song, and Lei Zhu. Restoreagent: Autonomous image restoration agent via multimodal large language models. *Advances in Neural Information Processing Systems*, 37: 110643–110666, 2025. 2

[5] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser-2: Unleashing the power of language models for text rendering. *arXiv preprint arXiv:2311.16465*, 2023. 2, 3, 7

[6] Sixiang Chen, Tian Ye, Jinbin Bai, Erkang Chen, Jun Shi, and Lei Zhu. Sparse sampling transformer with uncertainty-driven ranking for unified removal of raindrops and rain streaks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13106–13117, 2023. 2

[7] Sixiang Chen, Tian Ye, Yun Liu, Jinbin Bai, Haoyu Chen, Yunlong Lin, Jun Shi, and Erkang Chen. Cplformer: Cross-scale prototype learning transformer for image snow removal. In *Proceedings of the 31st ACM international conference on multimedia*, pages 4228–4239, 2023.

[8] Sixiang Chen, Tian Ye, Kai Zhang, Zhaohu Xing, Yunlong Lin, and Lei Zhu. Teaching tailored to talent: Adverse weather restoration via prompt pool and depth-anything constraint. In *European Conference on Computer Vision*, pages 95–115. Springer, 2024. 2

[9] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *Forty-first International Conference on Machine Learning*, 2024. 2, 7

[10] Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*, 2024. 3

[11] Jun-Yan He, Zhi-Qi Cheng, Chenyang Li, Jingdong Sun, Wangmeng Xiang, Xianhui Lin, Xiaoyang Kang, Zengke Jin, Yusen Hu, Bin Luo, et al. Wordart designer: User-driven artistic typography synthesis using large language models. *arXiv preprint arXiv:2310.18332*, 2023. 3

[12] Black Forest Labs. FLUX. https://github.com/black-forest labs/flux, 2024. 4

[13] OpenAI. GPT-4V. https://openai.com/index/gpt-4v-system card/, 2023. 4

[14] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 2, 7

[15] Xiwei Hu, Rui Wang, Yixiao Fang, Bin Fu, Pei Cheng, and Gang Yu. Ella: Equip diffusion models with llm for enhanced semantic alignment. *arXiv preprint arXiv:2403.05135*, 2024. 3

[16] Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. Opencole: Towards reproducible automatic graphic design generation. *arXiv preprint arXiv:2406.08232*, 2024. 3, 7

[17] Peidong Jia, Chenxuan Li, Yuhui Yuan, Zeyu Liu, Yichao Shen, Bohan Chen, Xingru Chen, Yinglin Zheng, Dong Chen, Ji Li, Xiaodong Xie, Shanghang Zhang, and Baining Guo. Cole: A hierarchical generation framework for multi-layered and editable graphic design, 2024. 3, 7

[18] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024. 2, 5, 7

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[20] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[21] Bonan Li, Yinhan Hu, Xuecheng Nie, Congying Han, Xiangjian Jiang, Tiande Guo, and Luoqi Liu. Dropkey for vision transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22700–22709, 2023. 2

[22] Bonan Li, Zicheng Zhang, Xuecheng Nie, Congying Han, Yinhan Hu, and Tiande Guo. Styo: Stylize your face in only one-shot. *arXiv preprint arXiv:2303.03231*, 2023. 2

[23] Fengheng Li, An Liu, Wei Feng, Honghe Zhu, Yaoyu Li, Zheng Zhang, Jingjing Lv, Xin Zhu, Junjie Shen, Zhangang Lin, et al. Relation-aware diffusion model for controllable poster layout generation. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, pages 1249–1258, 2023. 3

[24] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024. 2

[25] Yijun Li, Chen Fang, Jimei Yang, Zhaowen Wang, Xin Lu, and Ming-Hsuan Yang. Universal style transfer via feature transforms. *Advances in neural information processing systems*, 30, 2017. 3

[26] Bingchen Liu, Ehsan Akhgari, Alexander Visheratin, Aleks Kamko, Linmiao Xu, Shivam Shrirao, Joao Souza, Suhail Doshi, and Daiqing Li. Playground v3: Improving text-to-image alignment with deep-fusion large language models. *arXiv preprint arXiv:2409.10695*, 2024. 2, 7

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 5

[28] Zeyu Liu, Weicong Liang, Yiming Zhao, Bohan Chen, Ji Li, and Yuhui Yuan. Glyph-byt5-v2: A strong aesthetic baseline for accurate multilingual visual text rendering. *arXiv preprint arXiv:2406.10208*, 2024. 2, 3

[29] Chuwei Luo, Yufan Shen, Zhaoqing Zhu, Qi Zheng, Zhi Yu, and Cong Yao. Layoutllm: Layout instruction tuning with large language models for document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15630–15640, 2024. 3

[30] Run Luo, Zikai Song, Lintao Ma, Jinlin Wei, Wei Yang, and Min Yang. Diffusiontrack: Diffusion model for multi-object tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3991–3999, 2024. 2

[31] Xinyin Ma, Gongfan Fang, and Xinchao Wang. Deepcache: Accelerating diffusion models for free. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15762–15772, 2024. 2

[32] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 3

[33] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 3, 7

[34] Xinmin Qiu, Congying Han, ZiCheng Zhang, Bonan Li, Tiande Guo, and Xuecheng Nie. Diffbfr: Bootstrapping diffusion model for blind face restoration. In *ACM MM*, pages 7785–7795, 2023. 2

[35] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 7

[36] Jingjing Ren, Wenbo Li, Haoyu Chen, Renjing Pei, Bin Shao, Yong Guo, Long Peng, Fenglong Song, and Lei Zhu. Ultrapixel: Advancing ultra-high-resolution image synthesis to new peaks. *arXiv preprint arXiv:2407.02158*, 2024. 2

[37] Jingjing Ren, Haoyu Chen, Tian Ye, Hongtao Wu, and Lei Zhu. Triplane-smoothed video dehazing with clip-enhanced generalization. *International Journal of Computer Vision*, 133(1):475–488, 2025. 2

[38] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[39] Guoxian Song, Linjie Luo, Jing Liu, Wan-Chun Ma, Chunpong Lai, Chuanxia Zheng, and Tat-Jen Cham. Agilegan: stylizing portraits by inversion-consistent transfer learning. *ACM Transactions on Graphics (TOG)*, 40(4):1–13, 2021. 3

[40] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8791–8800, 2022. 2

[41] Haoze Sun, Wenbo Li, Jianzhuang Liu, Haoyu Chen, Renjing Pei, Xueyi Zou, Youliang Yan, and Yujiu Yang. Coser: Bridging image and language for cognitive super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25868–25878, 2024.

[42] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 3, 2024. 2

[43] Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16515–16525, 2022. 3

[44] Shengbang Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Manoj Middepogu, Sai Charitha Akula, Jihan Yang, Shusheng Yang, Adithya Iyer, Xichen Pan, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *arXiv preprint arXiv:2406.16860*, 2024. 3

[45] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 7

[46] Yuxiang Tuo, Wangmeng Xiang, Jun-Yan He, Yifeng Geng, and Xuansong Xie. Anytext: Multilingual visual text generation and editing. *arXiv preprint arXiv:2311.03054*, 2023. 2, 7

[47] Ideogram AI. Ideogram v2. https://ideogram.ai/launch, 2024. 7

[48] Recraft. Recraft v3. https://www.recraft.ai/about, 2023. 7

[49] Changshuo Wang, Lei Wu, Xiaole Liu, Xiang Li, Lei Meng, and Xiangxu Meng. Anything to glyph: Artistic font synthesis via text-to-image diffusion model. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[50] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 3

[51] Haoran Wei, Chenglong Liu, Jinyue Chen, Jia Wang, Lingyu Kong, Yanming Xu, Zheng Ge, Liang Zhao, Jianjian Sun, Yuang Peng, et al. General ocr theory: Towards ocr-2.0 via a unified end-to-end model. *arXiv preprint arXiv:2409.01704*, 2024. 8

[52] Wangyu Wu, Tianhong Dai, Xiaowei Huang, Fei Ma, and Jimin Xiao. Image augmentation with controlled diffusion for weakly-supervised semantic segmentation. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6175–6179. IEEE, 2024. 2

[53] Wangyu Wu, Xianglin Qiu, Siqi Song, Xiaowei Huang, Fei Ma, and Jimin Xiao. Prompt categories cluster for weakly supervised semantic segmentation. *arXiv preprint arXiv:2412.13823*, 2024.

[54] Wangyu Wu, Tianhong Dai, Zhenhong Chen, Xiaowei Huang, Jimin Xiao, Fei Ma, and Renrong Ouyang. Adaptive patch contrast for weakly supervised semantic segmentation. *Engineering Applications of Artificial Intelligence*, 139:109626, 2025.

[55] Zeyu Xiao, Yutong Liu, Ruisheng Gao, and Zhiwei Xiong. Cutmib: Boosting light field super-resolution via multi-view image blending. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1672–1682, 2023.

[56] Zhaomin Xiao, Yan Huang, and Eduardo Blanco. Analyzing large language models' capability in location prediction. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 951–958, Torino, Italia, 2024. ELRA and ICCL.

[57] Zeyu Xiao, Dachun Kai, Yueyi Zhang, Xiaoyan Sun, and Zhiwei Xiong. Asymmetric event-guided video super-resolution. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 2409–2418, 2024.

[58] Zeyu Xiao, Dachun Kai, Yueyi Zhang, Zheng-Jun Zha, Xiaoyan Sun, and Zhiwei Xiong. Event-adapted video super-resolution. In *European Conference on Computer Vision*, pages 217–235. Springer, 2024.

[59] Zhaomin Xiao, Zhelu Mai, Yachen Cui, Zhuoer Xu, and Jiancheng Li. Short interest trend prediction with large language models. In *Proceedings of the 2024 International Conference on Innovation in Artificial Intelligence*, page 1, New York, NY, USA, 2024. Association for Computing Machinery. 2

[60] Wenju Xu, Chengjiang Long, Ruisheng Wang, and Guanghui Wang. Drb-gan: A dynamic resblock generative adversarial network for artistic style transfer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6383–6392, 2021. 3

[61] Ling Yang, Zhaochen Yu, Chenlin Meng, Minkai Xu, Stefano Ermon, and CUI Bin. Mastering text-to-image diffusion: Recaptioning, planning, and generating with multimodal llms. In *Forty-first International Conference on Machine Learning*, 2024. 3

[62] Tao Yang, Yingmin Luo, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Posterllava: Constructing a unified multi-modal layout generator with llm. *arXiv preprint arXiv:2406.02884*, 2024. 3

[63] Xingyi Yang and Xinchao Wang. Kolmogorov-arnold transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[64] Yukang Yang, Dongnan Gui, Yuhui Yuan, Weicong Liang, Haisong Ding, Han Hu, and Kai Chen. Glyphcontrol: Glyph conditional control for visual text generation. *Advances in Neural Information Processing Systems*, 36, 2024. 2, 3

[65] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 3

[66] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 3

[67] Hanxun Yu, Wentong Li, Song Wang, Junbo Chen, and Jianke Zhu. Inst3d-lmm: Instance-aware 3d scene understanding with multi-modal instruction tuning. *arXiv preprint arXiv:2503.00513*, 2025. 2

[68] Jinran Zhang, Zhelu Mai, Zhuoer Xu, and Zhaomin Xiao. Is llama 3 good at identifying emotion? a comprehensive study. In *Proceedings of the 2024 7th International Conference on Machine Learning and Machine Intelligence (MLMI)*, page 128–132, New York, NY, USA, 2024. Association for Computing Machinery. 2

[69] Jianyi Zhang, Yufan Zhou, Jiuxiang Gu, Curtis Wigington, Tong Yu, Yiran Chen, Tong Sun, and Ruiyi Zhang. Artist: Improving the generation of text-rich images by disentanglement. *arXiv preprint arXiv:2406.12044*, 2024. 3

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 5

[71] Tao Zhang, Xiangtai Li, Hao Fei, Haobo Yuan, Shengqiong Wu, Shunping Ji, Chen Change Loy, and Shuicheng Yan. Omg-llava: Bridging image-level, object-level, pixel-level reasoning and understanding. *arXiv preprint arXiv:2406.19389*, 2024. 3

[72] Yizhe Zhang, Zhe Gan, Kai Fan, Zhi Chen, Ricardo Henao, Dinghan Shen, and Lawrence Carin. Adversarial feature matching for text generation. In *International conference on machine learning*, pages 4006–4015. PMLR, 2017. 3