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

Supplementary Material

A. Motivation Details

As shown in Figure A and Figure B, state-of-the-art generative models, such as FLUX-dev, Recraft V3, Playground-v3, and Ideogram 2.0, demonstrate some capability in generating visually appealing titles for posters. However, they exhibit significant limitations when tasked with creating posters that include extensive textual content. These models often fail to generate essential elements such as taglines, crew lists, and detailed event descriptions, frequently resulting in **missing text**, **erroneous text**, and **unwanted elements** are common issues that compromise the coherence and completeness of the generated outputs.

In addition to these shortcomings, these models frequently generate irrelevant content that is unrelated to the given prompt, further detracting from the accuracy and usability of the final designs. Such issues not only hinder the readability and aesthetic quality of the posters but also prevent these models from aligning with the specific requirements of the task. These combined factors—missing or incorrect text and irrelevant content—render these models unsuitable for practical applications, where precision, prompt adherence, and semantic coherence are essential for generating usable and professional-quality posters.

B. Model Details

B.1. Background Diffusion

Background Diffusion is based on the pretrained FLUX.1-dev model[12]. To support various styles, we fine-tune FLUX using the LoRA [14] at a resolution of 1024. Approximately 50 images are utilized per style, with an embedding dimension of 64. The fine-tuning process is conducted on one NVIDIA RTX A800 GPU over 2000 training steps, with learning rate 0.0002. Each image is repeated 10 times, and tagged with a specific trigger word, such as *"90s Vintage Style"*. We also use LoRAs from open-source communities, such as *Dark Grundge Poster*[1] and *8mm Film*[2].

B.2. Design MLLM

We incorporate the CLIP pre-trained Vision Transformer (ViT-L/14) [35] as the image encoder to convert input images into visual tokens. For the language model, we utilize the Llama3-7B [45]. Despite their capabilities, pre-trained LLMs fail to provide accurate responses without dataset-specific fine-tuning. To address this, we adopt LoRA [14], a fine-tuning technique that efficiently modifies a limited number of parameters within the model.

Following [14], we apply LoRA to adjust the projection layers in all self-attention modules of both the vision encoder and the LLM, thereby generating our Design MLLM. We employ the Xtuner framework[3] to facilitate the training process. For our experimental setup, we configure the LoRA rank to 16. The Design MLLM undergoes training on 1 NVIDIA RTX A800 GPU, with a batch size of 32. We employ the Adam optimizer and a learning rate of 0.00002.

Template for Visual Instruction Tuning. For visual instruction tuning[27], we use the template as follows:

Input: *"Arrange <N>text elements on the background image <img>to achieve a well-structured and aesthetically pleasing poster layout. Return the result by filling in the following JSON file <input_json>. In this JSON file, each text element has 3 fixed attributes: (1) category: "title", "subtitle", or "information", indicating the prominence of this element; (2) text: the content to be displayed on the image; (3) description: the user's instructions for this element, which must be followed. Additionally, a global description specifies the relative positioning of different text elements. Based on the above information, you are required to complete the following 5 attributes for each text element: (1) bounding box: [left, top, right, bottom], with each coordinate being a continuous number between 0 and 1; (2) font type: a discrete value ranging from 0-19; (3) color: [R, G, B], with each value being an integer between 0 and 255; (4) alignment: a value of 0, -1, or 1, representing Center, Align Left, or Align Right, respectively; (5) angle: a value of 0, -1, or 1, representing Horizontal, Vertical to the left, or Vertical to the right, respectively."*

Response: *<output_json>*

Text Rendering with Layout and Typography. For each text element, we predict the font type, alignment (0, -1, and 1 represent *Center*, *Align Left*, and *Align Right*, respectively), bounding box coordinates, and angle (0, -1, and 1 indicate *Horizontal*, *Vertical to the left*, and *Vertical to the right*, respectively) to render the mask on the background. The font size is dynamically adjusted based on the bounding box dimensions using simple rules. Sample code is provided in Figure C.

In this example, the alignment type is assumed to be *Center*, and the angle is *Horizontal*. For other alignment types and angles, the *center_position* and *text_dimensions*

[1] https://civitai.com/models/709640/dark-grunge-poster-flux-dev

[2] https://civitai.com/models/725293/8mm-film-loraanalogv1flux-workflow-included

[3] https://github.com/InternLM/xtuner

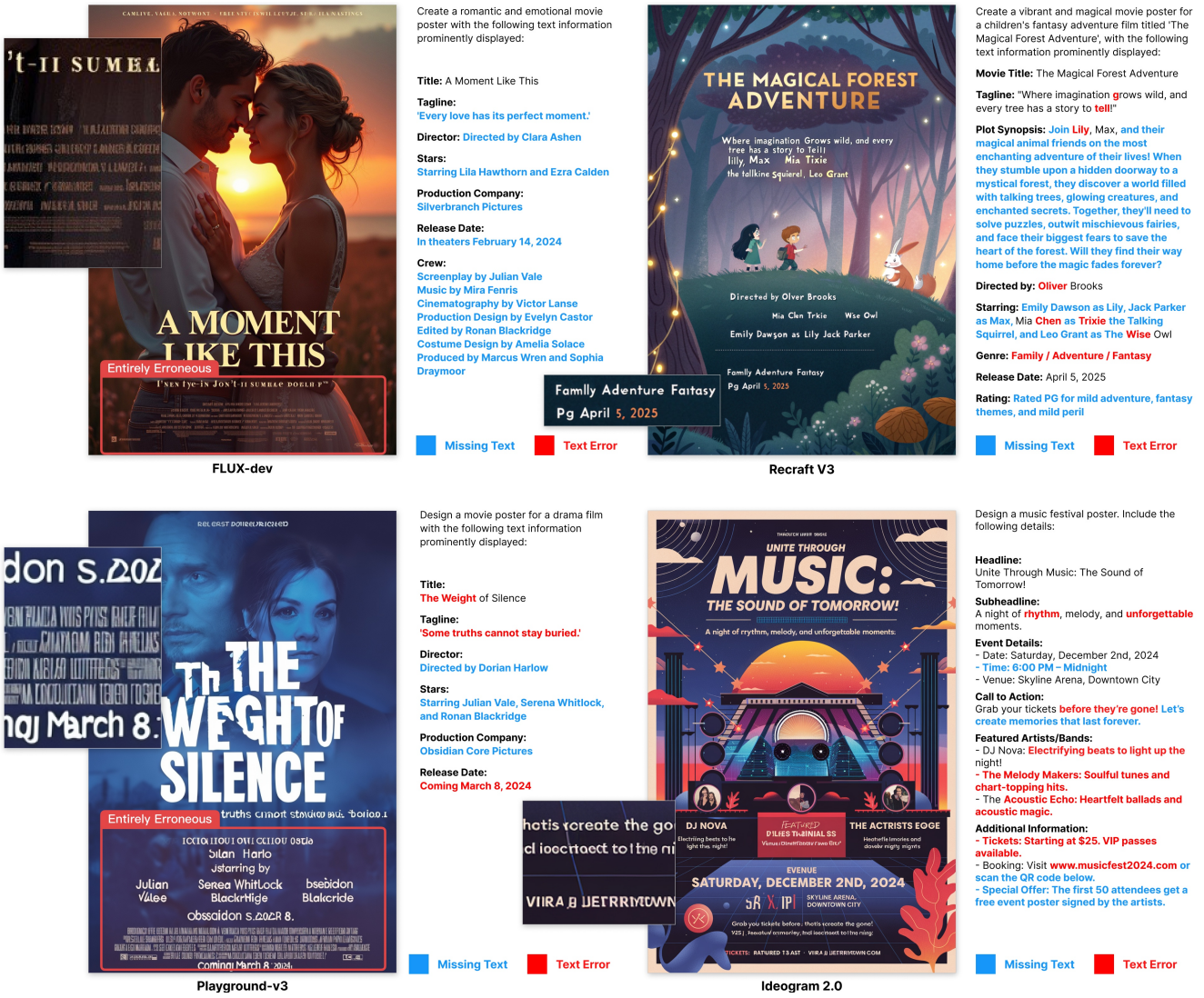

Figure A. Limitations of state-of-the-art generative models in creating posters with extensive textual content. While models such as FLUX-dev, Recraft V3, Playground-v3, and Ideogram 2.0 can generate visually appealing titles, they struggle to handle additional textual elements like taglines, crew lists, and event details. Common issues include missing text, text errors, and the generation of irrelevant content unrelated to the given prompt. These limitations make these models unsuitable for real-world applications, where accuracy, coherence, and prompt adherence are critical.

in the code differ accordingly. For text spanning multiple lines, the bounding box is divided into segments, and similar calculations are applied to each segment.

Once the font size is determined, the text is rendered on the background using the specified location and complete typographical information.

B.3. ArtText Diffusion

For ArtText Diffusion, we train our model based on SDXL's BrushNet [18, 33] at a resolution of 1216. The training process uses a learning rate of 1e-5, with a batch size of 1. To account for the small batch size, we employ gradient accumulation over 4 steps to ensure stable and effective optimization. The training is conducted on a single NVIDIA A100 GPU.

Figure B. Limitations of state-of-the-art generative models in creating posters with extensive textual content. While models such as FLUX-dev, Recraft V3, Playground-v3, and Ideogram 2.0 can generate visually appealing titles, they struggle to handle additional textual elements like taglines, crew lists, and event details. Common issues include **missing prompted text**, **illegible and incorrect Text**, and the **generation of irrelevant unwanted content** unrelated to the given prompt. These limitations make these models unsuitable for real-world applications, where accuracy, coherence, and prompt adherence are critical.

C. Prompt Details

C.1. Prompts for the Magic Prompter

The Magic Prompter is powered by GPT-4V and is utilized for both Background Diffusion and ArtText Diffusion.

Magic Prompter for Background Diffusion. For Background Diffusion, the text prompt serves as a detailed description of the desired background image. The Magic Prompter refines user input by enhancing its details and removing keywords such as *"movie poster"*, which may

```python
def adjust_font_size(font, letter_spacing, center_position,
bounding_box):
    """
    Adjusts font size until the rendered text exceeds the bounding box.

    Args:
        font (Font): Font object with properties like size, etc.
        letter_spacing (float): Space between letters.
        center_position (tuple): Center position of the text (x, y).
        bounding_box (tuple): Bounding box coordinates (x1, x2, y1, y2).

    Returns:
        int: The maximum font size that fits within the bounding box.
    """
    x1, x2, y1, y2 = bounding_box
    max_font_size = 0
    font_size = font.size  # Starting font size

    while True:
        # Calculate the rendered position and size of the text
        text_width, text_height = calculate_text_dimensions(font,
font_size, letter_spacing)
        text_x = center_position[0] - text_width / 2
        text_y = center_position[1] - text_height / 2
        text_x2 = text_x + text_width
        text_y2 = text_y + text_height

        # Check if the text exceeds the bounding box
        if text_x < x1 or text_x2 > x2 or text_y < y1 or text_y2 > y2:
            break

        # Update font size
        max_font_size = font_size
        font_size += 1

    return max_font_size
```

Figure C. Sample code for font size adjustment.

otherwise lead to unintended random text appearing in the background.

The prompt: *You are a prompt generator. Provide a detailed and precise description of the desired image based on the provided prompt, excluding keywords such as "poster", "movie poster", or "album". Ensure that no text appears within the image.*

Magic Prompter for ArtText Diffusion. For ArtText Diffusion, the text prompt provides a detailed description of the text, incorporating artistic effects that harmonize seamlessly with the background. The Magic Prompter enhances this process by directly generating prompts inferred from the background or enriching the user's input for a cohesive and visually appealing result.

The prompt for inferring from background: *You are a prompt generator. Provide a highly detailed and precise description of the artistic effects for text that integrates seamlessly with the given image, ensuring harmony with its overall aesthetic. This description should include detailed specifications of color schemes, material textures, stylistic features, and other artistic effects to achieve a cohesive and visually appealing design.*

The prompt for enhancing user input: *You are a prompt generator. Provide an enhanced and highly detailed description of the artistic effects for text based on the provided prompt. The description should encompass intricate details of color schemes, material textures, stylistic elements, and additional artistic features to deliver a refined and harmonious result.*

C.2. Prompts for GPT-4V evaluation

Evaluation on Aesthetics:

- Background: *You are a professional designer with very strict evaluation standards. Now please give a score of 1-10 based on the beauty of the background of this poster, 10 represents the best and 1 represents the worst.*
- Title: *You are a professional designer with very strict evaluation standards. Now please give a score of 1-10 based on the beauty of the title of this poster, 10 represents the best and 1 represents the worst.*
- Layout and typography: *You are a professional designer with very strict evaluation standards. Now please give a score of 1-10 based on the beauty of the layout and typography of this poster, 10 represents the best and 1 represents the worst.*
- overall: *You are a professional designer with very strict evaluation standards. Now please give a score of 1-10 based on the beauty of the overall design of this poster, 10 represents the best and 1 represents the worst.*

Evaluation on Text & Element:

- Text Readability: *You are a professional designer with very strict evaluation standards. Now please give a score of 1-10 based on the text readability of this poster, 10 represents the best and 1 represents the worst. A high score means the text is clear and readable, and there are no spelling errors.*
- Prompt Relevance: *You are a professional designer with very strict evaluation standards. Now please give a score of 1-10 based on the prompt relevance of this poster, 10 represents the best and 1 represents the worst. A high score means that the elements mentioned in the text prompt are present in the image and no irrelevant elements appear.*

D. Dataset Details

D.1. Details of PosterArt-Design

The *PosterArt-Design* dataset comprises posters with layouts and typography manually annotated by professional designers, shown in Figure D. The backgrounds of these posters are sourced from two categories: some were generated, while others were purchased, often including predefined layouts and typography. To ensure consistency with aesthetic design principles, designers were tasked with uniformly annotating the dataset, adhering to preset font and color requirements. Additionally, some background images

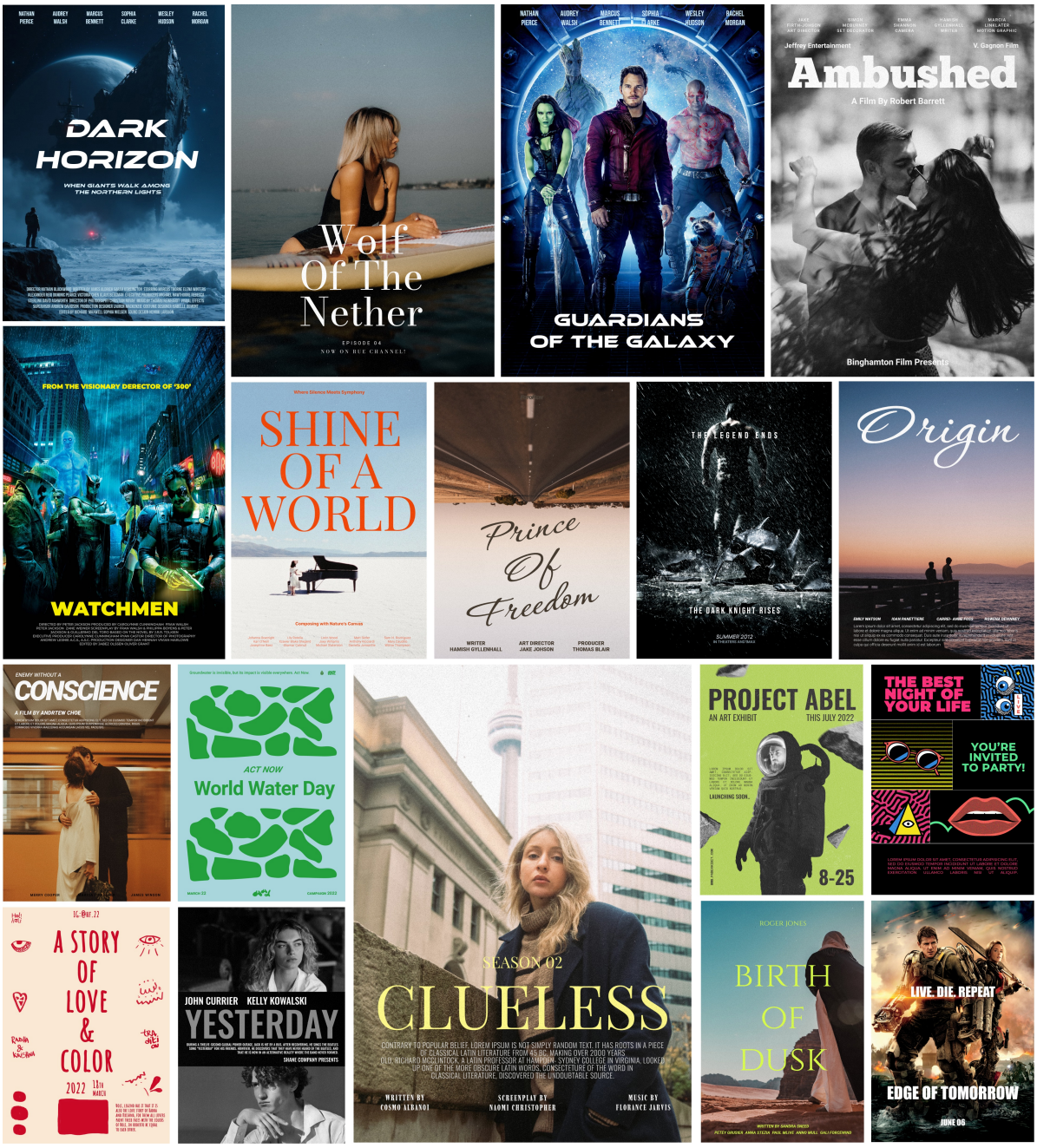

Figure D. Samples in *PosterArt-Design* dataset. All poster designs include standalone backgrounds, with all text layouts and typography saved in an editable format. Compared to previous poster datasets, our dataset offers significantly higher quality, featuring more visually appealing and refined designs. Additionally, it includes detailed information such as fonts, colors, and other stylistic attributes.

are from classic movie posters, with layouts inspired by these iconic designs.

For fonts, We carefully curates a font library consisting of 20 free-to-use, commercially licensed typefaces, ensuring both legal compliance and a wide range of stylistic options for data construction and design generation. As shown in Figure E, the font library is categorized into four major groups: *Serif*, *Sans-serif*, *Decorative*, and *Handwritten*. The *Serif* category includes timeless and elegant fonts such as *Cinzel*, *Baskerville*, and *Crimson Text*, ideal for formal and traditional designs. The *Sans-serif* group offers modern, clean aesthetics with fonts like *Bebas Neue*, *Montser-*

rat, and *Roboto*, which suit contemporary and minimalist styles. For playful and distinctive designs, the Decorative category features bold and expressive fonts such as *Luckiest Guy*, *Ethnocentric*, and *ChunkFive*. The *Handwritten* section includes fonts like *Amatic SC* and *Corinthia*, adding a personal and artistic touch.

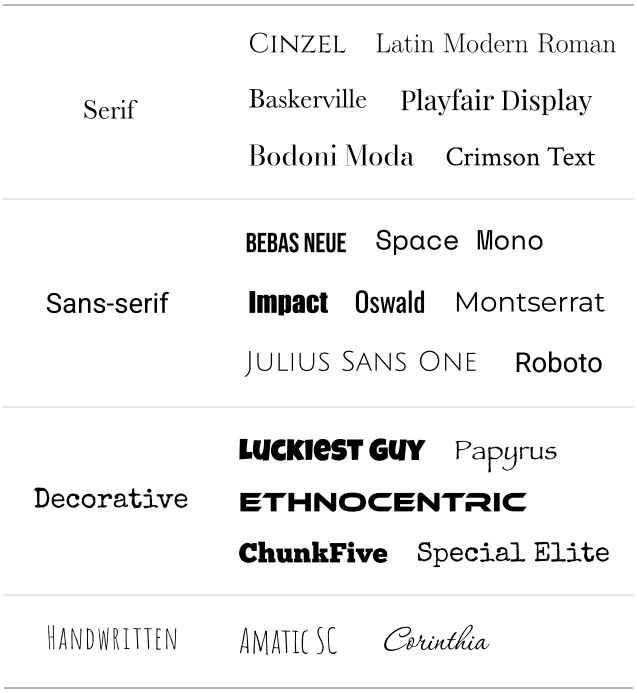

Figure E. Font Library.

D.2. Details of PosterArt-Text

The *PosterArt-Text* dataset presents a rich and diverse collection of artistic text designs, as shown in Figure F. A key feature of this dataset is that every artistic text style is paired with its corresponding poster background. The main sources of these posters include movie, exhibition, album, and other posters retrieved from the Internet. Additionally, some images with strong text effects are generated with FLUX[12]. After collection, annotators were engaged to meticulously annotate the pixel-wise segmentation.

The dataset covers a wide range of text styles, offering examples that cater to various design needs. For instance, some styles are bold and metallic (*Shocker Metal*, *Frozen*), while others are playful and colorful (*Funky Time*, *Happy Feet*). This stylistic diversity ensures that the dataset can support a broad array of creative applications, making it a valuable resource for both artistic and functional design tasks.

E. More Experiments and Results

E.1. More Results of Artistic Text Stylization

The results of our Artistic Text Stylization Module, as shown in the Figure G, demonstrate its ability to generate a wide variety of stylized text designs with rich visual diversity. Each stylization reflects a unique artistic theme, ranging from metallic and futuristic effects to natural textures and vibrant color palettes, showcasing the module's versatility and creative potential.

Our module excels in adapting to different design needs, producing artistic text with high fidelity and detailed stylization. The generated results maintain clarity and aesthetic appeal, ensuring that the text is not only visually stunning but also functional for practical applications such as posters, advertisements, and graphic design projects.

E.1.1. Comparison with the Base Model

As illustrated in Figure H, our method significantly outperforms traditional diffusion inpainting models in two key aspects:

Lack of Artistic Text Stylization in Base Models. Pretrained diffusion inpainting models lack the ability to generate stylized artistic text. They cannot interpret or apply text-specific style prompts to create visually compelling designs. In contrast, our method not only automatically generates artistic text that harmonizes with the background without relying on explicit prompts but also allows precise control via user-defined prompts. This capability ensures both flexibility and accuracy in generating text styles that align seamlessly with user requirements.

Seamless Coordination with Backgrounds. A standout feature of our approach is the natural integration of text with the background. As shown in the figure, our method automatically adapts the text color and style to achieve optimal readability and aesthetic harmony. For example, when the background is bright, our method generates darker text to enhance contrast and ensure visual clarity. This dynamic adjustment of text properties based on background context is crucial for producing professional-quality designs that feel cohesive and polished.

E.2. More Results of Poster Generation

The results of our poster generation framework, as shown in the Figure I, Figure J, Figure K, and Figure L, demonstrate its capability to produce a rich variety of visually appealing and professionally designed posters. These examples highlight the system's ability to handle diverse styles, themes, and compositions, ensuring adaptability to different creative and functional requirements.

A key strength of our method lies in its ability to generate text of varying lengths while maintaining aesthetic balance and readability. From short, impactful titles to longer, multi-line text compositions , the system consistently ensures proper alignment, spacing, and integration with background elements. This flexibility enables the generation of posters that are both visually cohesive and contextually appropriate. Furthermore, the diversity of designs spans a broad spectrum of genres and artistic styles, including science fiction, fantasy, drama, and music-inspired themes. Despite this diversity, the results consistently maintain high levels of visual harmony, with the generated text naturally blending into the background to create polished and professional outputs.

F. Limitation and Future Work

While our method demonstrates strong capabilities in generating artistic text and poster designs, there are still several limitations that highlight opportunities for future improvements:

- Our current pipeline lacks an automated evaluation module to ensure the aesthetic quality of the generated results. At present, users must manually review and select the most visually appealing outputs. This reliance on human judgment limits the automation of the process. In the future, we aim to incorporate an evaluation module, potentially leveraging agent-based architectures, to enable a more automated and intelligent selection of high-quality results.
- Our dataset, while meticulously curated and of high quality, is limited in size due to the significant time and cost associated with its creation. This constraint restricts the granularity of text layout and typography that can be achieved. Expanding the dataset in the future will allow us to support more complex and finely detailed typographic designs, further enhancing the flexibility and robustness of our approach.
- The inpainting model occasionally faces challenges posed by the limitations of the text mask shape or the complexity of the background content. In certain cases, to achieve better overall harmony with the background, the artistic diversity of the generated text may be partially sacrificed. Addressing this trade-off in future work will involve refining the inpainting process to better balance artistic variety and visual coherence with the background.

Figure F. Samples in *PosterArt-Text* dataset. The figure only showcases the extracted text from the poster backgrounds; however, every text instance in our dataset is paired with its corresponding poster background. This is a key distinction of our dataset compared to others. Having both the background and text enables training models to generate text that is naturally and harmoniously integrated with the background.

Figure G. More results of artistic text stylization. Although only standalone text is shown here, our mask-based inpainting approach ensures that the text style and the background are maximally harmonious, natural, and visually appealing.

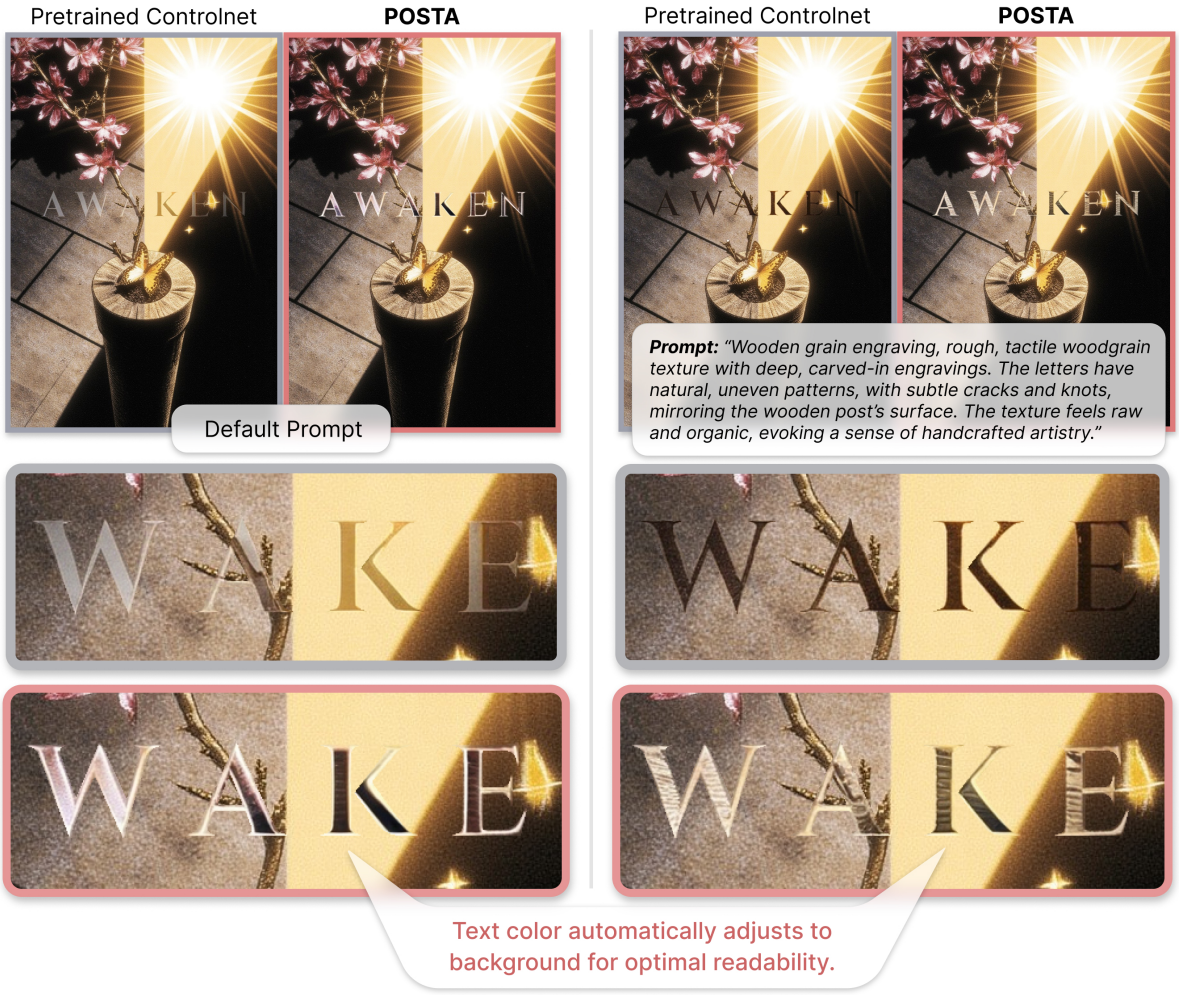

Figure H. Compared to base diffusion inpainting models, our method excels in two key aspects: (1) the ability to generate artistic stylized text that aligns with user-defined prompts or automatically harmonizes with the background, and (2) seamless coordination with backgrounds, dynamically adjusting text color and style for optimal readability and aesthetic harmony.

Figure I. More results generated by *POSTA*.

Figure J. More results generated by *POSTA*.

Figure K. More results generated by *POSTA*.

Figure L. More results generated by *POSTA*.