Supplementary Material for: DesignDiffusion: High-Quality Text-to-Design Image Generation with Diffusion Models

Zhendong Wang[1], Jianmin Bao[2,*,†], Shuyang Gu[2], Dong Chen[2], Wengang Zhou[1,*], Houqiang Li[1]
[1] University of Science and Technology of China
[2] Microsoft Research Asia
[*]corresponding author [†]project lead

1. More Dataset Details

Our dataset consists of images sourced from Google Image Search using design-related keywords such as logo, poster, flyer, cover, sign, brochure, banner, business card, website mockup, packaging, magazine layout, advertisement, infographic, product label, menu, invitation, certificate, and presentation slide. We use the LLaVA 1.6-34B vision-language model to extract visual text and generate captions.

We apply several filters to ensure the quality of our dataset:

- **Text Length**: Images with visual text longer than 150 characters are excluded.
- **Resolution**: Images with resolutions lower than 768×768 pixels are removed.
- **Aspect Ratio**: Images with aspect ratios outside the range of 0.25 to 4.0 are excluded.
- **Aesthetic Quality**: Images with an aesthetic score lower than 4.5 are filtered out.

After filtering, our dataset includes about 1 million high-quality training images and 5,000 testing images.

Moreover, we elaborate further on the criteria that define "high quality" in our context:

- **Design-Specific Content**: The images in our dataset are specifically collected for their relevance to design purposes. This includes images featuring visual texts commonly found in design contexts. By focusing on these types of images, we ensure that the dataset is highly relevant to the task of text-to-design image generation.
- **Enhanced Image Captions**: The captions for the images are generated using the state-of-the-art, open-source large vision-language model, LLaVA 1.6-34B. This model provides detailed and contextually accurate descriptions of the images, which enhances the quality of the dataset by offering rich textual information that aligns closely with the visual content.

These criteria collectively define the "high quality" nature of our dataset, ensuring that it is both relevant and robust for the task at hand.

Table 1. Statistics of the composition of our design dataset.

Dataset	img count	img type	mean lines/img	chars/words count	unique chars/words
AnyWord-3M	3M	img w/ text	4.13	6.35M	695.2K
Ours	1M	design img	4.71	16.72M	251.1K

Our dataset is specifically curated for design image generation, setting it apart from Mario-10M [1], LAION-Glyph [6], and AnyWord-3M [5], which primarily focus on visual text rendering. The key differences include:

- **Content Composition**: Our dataset comprises images that seamlessly integrate both visual and textual elements to reflect design aesthetics. In contrast, the other datasets often focus solely on text, such as tables or isolated sentences, without the broader visual context or the integration of design elements.
- **Design Focus**: Our dataset is specifically tailored for design contexts, encompassing a wide variety of design assets, including logo, poster, flyer, cover, sign, brochure, banner, business card, website mockup, packaging, magazine layout, advertisement, infographic, product label, menu, invitation, certificate, and presentation slide. The other datasets, however, generally focus on the presence of text within images, without specifically addressing the diverse and specialized design contexts that our dataset targets. Overall, we provide detailed statistics of the composition of our design dataset compared with the existing visual text rendering dataset in Tab. 1.

2. OCR Tool Comparison

To evaluate the visual text quality in generated design images, we need an OCR tool to extract the visual text rendered in the image. We demonstrate that two popular OCR tools (PPOCR [2], EasyOCR [3]) often fail to extract visual text in design images. After deep investigations, we find that current large-scale multi-modality models do a better job than those OCR tools. Here, we give examples to show that

(a) GT text: Juliana Silva LASH BEAUTY"
(b) PPOCR: "Julinna Cilva LASH BEAUTY"
(c) EasyOCR: "Jultnm Gilwa LASH BEAUTY"
(d) LLaVA 1.6: "Juliana Silva LASH BEAUTY"

(a) GT text: "Happy Easter"
(b) PPOCR: "Eagter"
(c) EasyOCR: "dske"
(d) LLaVA 1.6: "Happy Easter"

(a) GT text: "Thank you for your purchase"
(b) PPOCR: "Jhanle you FORYOURPURCHASE"
(c) EasyOCR: "Thanke yowe"
(d) LLaVA 1.6: "Thank you for your purchase"

(a) GT text: "Happy Holidays"
(b) PPOCR: "Hapnyk Holidays"
(c) EasyOCR: "Hopnyk~Holidays"
(d) LLaVA 1.6: "Happy Holidays"

Figure 1. Examples for comparing the OCR capability for detecting text in design images. Red part denotes wrong detection. LLaVA 1.6 gets the best OCR accuracy for extracting text in design images.

LLaVA 1.6 gets a higher OCR accuracy of extracting text in design images than PPOCR [2], EasyOCR [3]), as shown in Fig. 1.

3. Image Captioning by LLaVA

In our design image dataset, we adopt LLaVA-1.6-34B [4] to describe the input image. Considering that we have obtained the text annotations of these images, we design a comprehensive prompt to encourage LLaVA-1.6 to perform image captioning:

> Please provide a concise caption for the image, detailing the text shown in the image and key image elements from both holistic and detailed perspectives. If there is text content, ensure that any extracted text content in your output is enclosed in double quotes. Do not include irrelevant information or subjective comments in your output. When given the image's text content, you should output a definite, unequivocal, correct, and objective caption that incorporates all the text information. <optional if text available> The text content in this image is <text place-holder>.

After image captioning by LLaVA-1.6-34B, we further ensure that the returned caption includes the visual text information. If not containing complete text information, we append the lost text information into the returned caption by applying a random manual-designed template. For example, 'with a text of "<text place-holder>"', or 'with words saying "<text place-holder>"', etc. We visualize some examples of our design image dataset in Fig. 5.

Figure 2. Qualitative comparisons on AnyWord-3M.

4. Effect of SP-DPO

We have shown the quantitative result of SP-DPO in the ablation studies in the main paper. Our ablation studies indicate that SP-DPO significantly enhances text quality across all text metrics. Regarding the slight decline in the FID score, this can be attributed to the specific implementation of SP-DPO, which relies heavily on the selection of winning and losing data. Currently, the winning-losing pairs are determined based on the text accuracy of synthetic images. If we incorporate image quality into the criteria for constructing the losing data, we anticipate that the FID score will improve further.

Notably, our current approach does not visually compromise image quality while enhancing text quality. In Fig. 3, we provide samples to illustrate the positive impact of SP-DPO on text quality without detracting from the overall image quality.

5. Diverse Generation

Diverse samples generated by DesignDiffusion are shown in Fig. 6. Given one prompt, we generate four images for a more clear demonstration. We observe that DiffusionDif-

Figure 3. Effect of SP-DPO. With SP-DPO, our model reduces the phenomenon of repeated text and poor text quality while keeping the overall image quality.

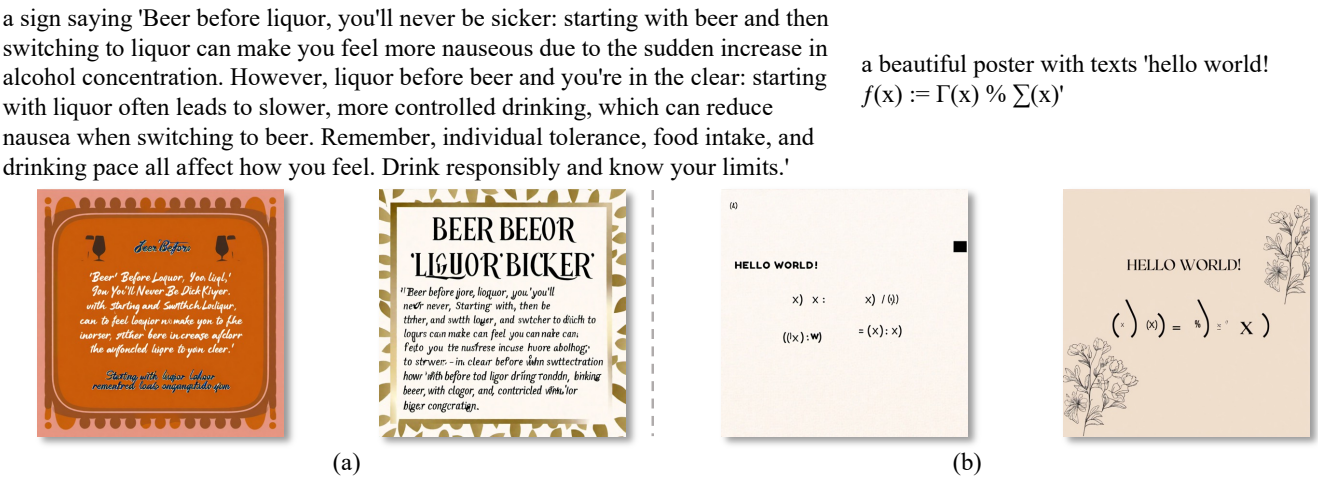

Figure 4. Failure cases. (a) Handling Very Long Visual Text. (b) Rare Formula Symbols.

Table 2. Quantitative comparison on AnyWord-3M.

Method	Training Epochs	Sen. ACC↑	NED↑	FID↓
AnyText [5]	10	0.6588	0.8568	35.87
DesignDiffusion	1.7	**0.6922**	**0.8719**	**33.98**

fusion generates design images with different image styles, text positions, and image-text corporations.

6. Comparisons on AnyWord-3M

Although our work focuses on design image generation, we further conduct experiments on the public AnyWord-3M dataset [5], training our DesignDiffusion for 40k iterations with a batch size of 128. This corresponds to only 1/6 of the training epochs used for AnyText. Tab. 2 and Fig. 2 present quantitative and qualitative comparisons, respectively. Our results demonstrate that DesignDiffusion outperforms AnyText in terms of both visual text rendering accuracy and image quality.

7. Failure Cases

In our experiments, we discovered some failure cases in which current models may also struggle to generate these images. We show two cases in Fig. 4. (a) Handling Very Long Visual Text: Our model may struggle to generate accurate visual text when the text length is exceptionally long. (b) Rare Formula Symbols: The model has difficulty generating rare formula symbols, primarily because these symbols are infrequently or never encountered in the training data. We believe that identifying and analyzing these failure cases is crucial for guiding future improvements in the task of design image generation.

References

[1] Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. Textdiffuser: Diffusion models as text painters. *arXiv preprint arXiv:2305.10855*, 2023. 1

[2] Yuning Du, Chenxia Li, Ruoyu Guo, Xiaoting Yin, Weiwei Liu, Jun Zhou, Yifan Bai, Zilin Yu, Yehua Yang, Qingqing

"Happy Orthodox Christmas" graphic with angels and a star, set against a winter forest backdrop.

The image features a graphic design with a focus on text that reads "25 reasons to invest in NEW TECH."

"Hello Winter Sale" graphic with stylized text and a deer silhouette, set against a snowy background with Christmas tree decorations.

The image features a vibrant graphic design with a central masquerade mask in pink and purple, surrounded by colorful confetti and stars. The text "MARDI GRAS MASCAREDE" is prominently displayed in a playful, handwritten font, indicating the theme of the image is related to Mardi Gras celebrations.

Man wearing headphones and looking at a tablet, with the text "technologies that will make you more productive" suggesting a focus on modern technology and efficiency.

The image features a graphic design with a theme related to curling, specifically the 2018 Winter Olympics in Pyeongchang. The central visual text reads "CURLING PYEONGCHANG 2018 8-25 FEBRUARY," indicating the event, location, and duration. The background consists of abstract shapes resembling curling stones and ice, with a color palette that includes red, blue, yellow, and white. The overall design is modern and minimalistic, with a focus on the sport of curling and the specific event details.

The image features a graphic design with a football theme, specifically related to the Super Bowl Sunday event in 2019. The central visual text reads "02 2019 SUPER BOWL SUNDAY It's Game Time", indicating the date and event. The design includes a stylized football helmet, a stadium filled with spectators, and a football field with yard lines. The overall style of the image is promotional and celebratory, designed to evoke the excitement and anticipation of the Super Bowl Sunday event.

"Super Back to School Sale 40% Off"

Figure 5. Examples of image-prompt pair from the design image dataset.

Dang, et al. Pp-ocr: A practical ultra lightweight ocr system. *arXiv preprint arXiv:2009.09941*, 2020. 1, 2

[3] JaidedAI. Github link: https://github.com/jaidedai/easyocr, 2023. 1, 2

[4] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[5] 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. 1,

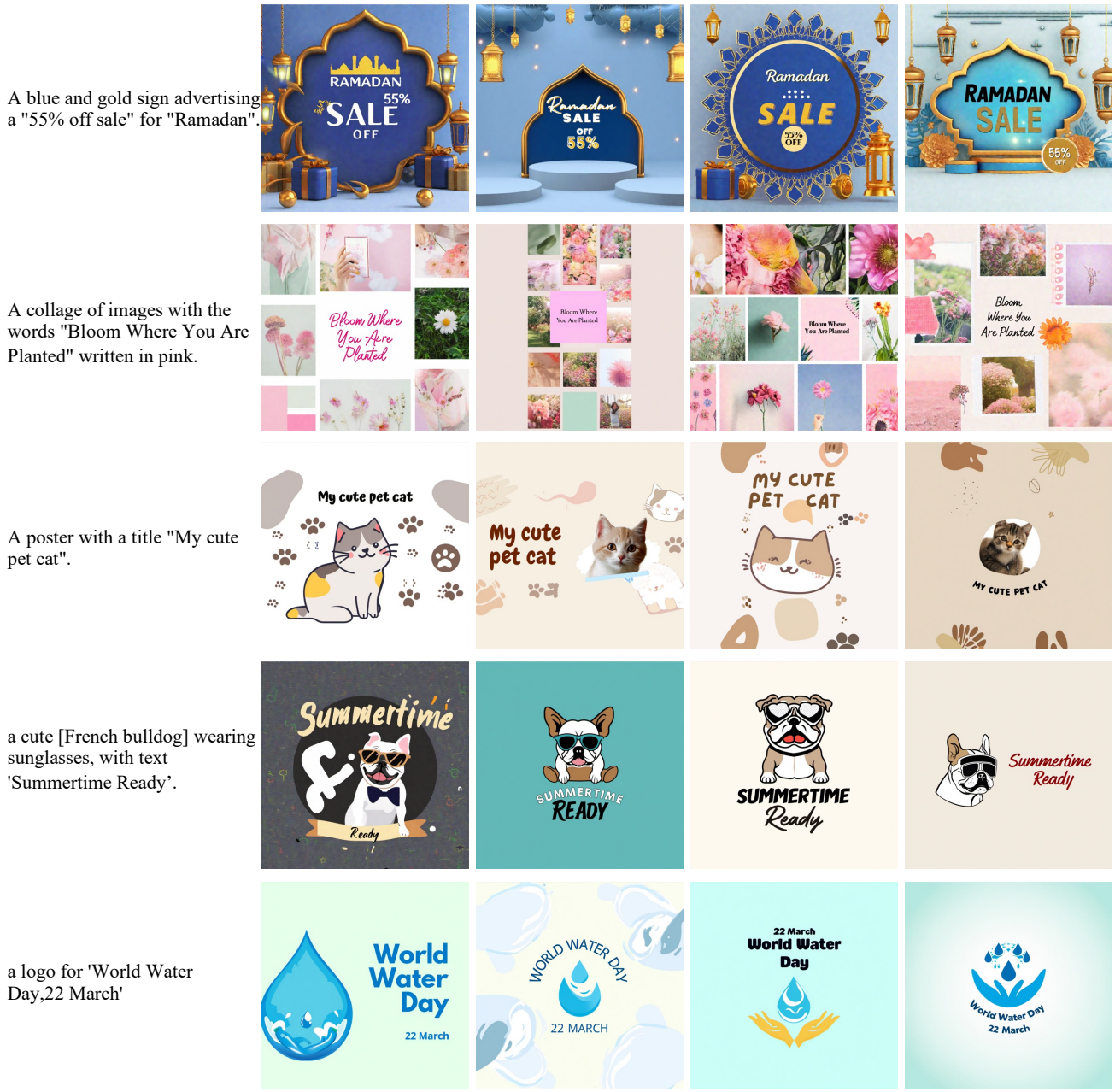

Figure 6. Diverse samples generated by DesignDiffusion. Our DesignDiffusion generates high-quality diverse design images with a one-stage framework.

[6] 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. 1