GlyphControl: Glyph Conditional Control for Visual Text Generation

Yukang Yang[1†♮], Dongnan Gui[2†♮], Yuhui Yuan[3†‡], Weicong Liang[3♮], Haisong Ding[3], Han Hu[3], Kai Chen[3]

1. Princeton University
2. University of Science and Technology of China
3.Microsoft Research Asia

[†]Core Contribution [♮]Interns at Microsoft Research Asia
[‡]Corresponding Author: yuhui.yuan@microsoft.com

https://github.com/AIGText/GlyphControl-release

❖ Motivation

- Current superior diffusion-based text-to-image generation methods still **lack the ability** to produce **legible and readable visual text** in generated images

Attempts: Modifying **Text Encoder** (more than CLIP)

- ➢ Large Language Models like T5 used in Imagen[1] and IF
- ➢ Character-aware Language Models like ByT5[2]

layout errors such as missing or merged glyphs exist

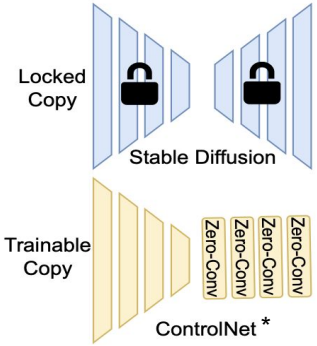
textual input prompts alone would not be sufficient for accurate visual text rendering

⇩

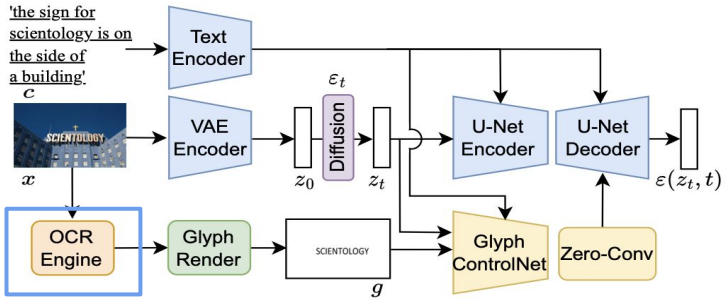
incorporate text **glyph** information

1. Chitwan Saharia, et al.Photorealistic text-to-image diffusion models with deep language understanding. In *NeurIPS 2022.*
2. Rosanne Liu, et al. Character-aware models improve visual text rendering. In *ACL* 2022.

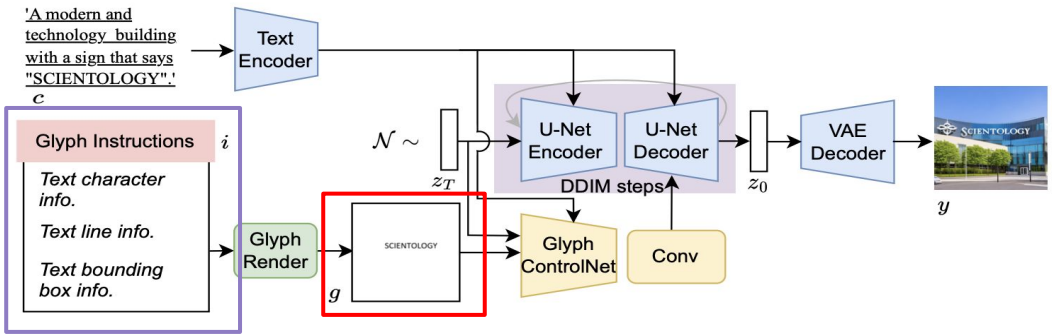
❖ Approach: **GlyphControl**



(a) GlyphControl.



(b) Training pipeline.



(c) Inference pipeline.

Flexible User Customization

Glyph image as Conditional Control

*Lvmin Zhang, et al. Adding conditional control to text-to-image diffusion models. In *CVPR 2023.*

❑ Glyph Instructions

multiple groups of text at **different** locations

➔ **Text character** information: the ***text*** placed at the same area (single words, sentences, or phrases)

➔ **Text line** information: assign words to multiple lines by adjusting the ***number of rows***

➔ **Text box** information:

- Font size: the ***width*** property of the text bounding box.
- Location of the text: the ***coordinates*** property of the top left corner.
- Rotation of the text: the ***yaw rotation angle*** property of the text box
- ***width-height ratio*** (optional): to precisely control the height of the text box.

A portrait of a parrot holding a sign with text "Hello World".

A storefront with "GlyphControl" written on it, centered.

A hand-painted wooden "Free Beer" sign hanging out of a bar.

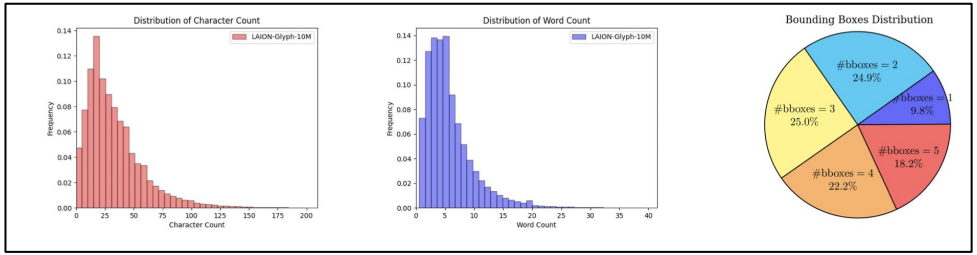
A fancy violet T-shirt decorated with flowers while the message [X] are written on it.

❖ **LAION-Glyph** Benchmark

- Extracted From an extensive multi-model dataset **LAION-2B-en**
- Selecting examples with abundant visual text context using **PP-OCR engine**.
- **Filtering:** aesthetic score>4.5, total OCR areas >5% of the whole image area, 1-5 OCR text bounding boxes
- generate new captions using the **BLIP-2** model

Different Scales: **LAION-Glyph-100K**, **LAION-Glyph-1M**, and **LAION-Glyph-10M**

Statistics



Samples

❖ Main Results

❑ Evaluation Benchmarks

- **SimpleBench:** The format of prompts remains the same: *'A sign that says "<word>".'*
- **CreativeBench:** Diverse text prompts adapted from GlyphDraw*.

  e.g.,: *'Little panda holding a sign that says "<word>".'* or *'A photographer wears a t-shirt with the word "<word>" printed on it.'*

** ***<word>*** ** **:**

single-word candidates from Wikipedia; selecting 100 words from each **frequency bucket**; in total **400** words.

❑ Evaluation Metrics

- **OCR accuracy**:
  - exact match accuracy **Acc**
  - **capitalization-insensitive** exact match accuracy $\widehat{\mathbf{Acc}}$
  - average Levenshtein distance **LD**
- **CLIP Score**: Image-prompt Alignment
- **FID**: Image Quality

*Jian Ma, et al. Glyphdraw: Learning to draw chinese characters in image synthesis models coherently. *arXiv preprint arXiv:2303.17870*, 2023.

❑ Quantitative Comparison

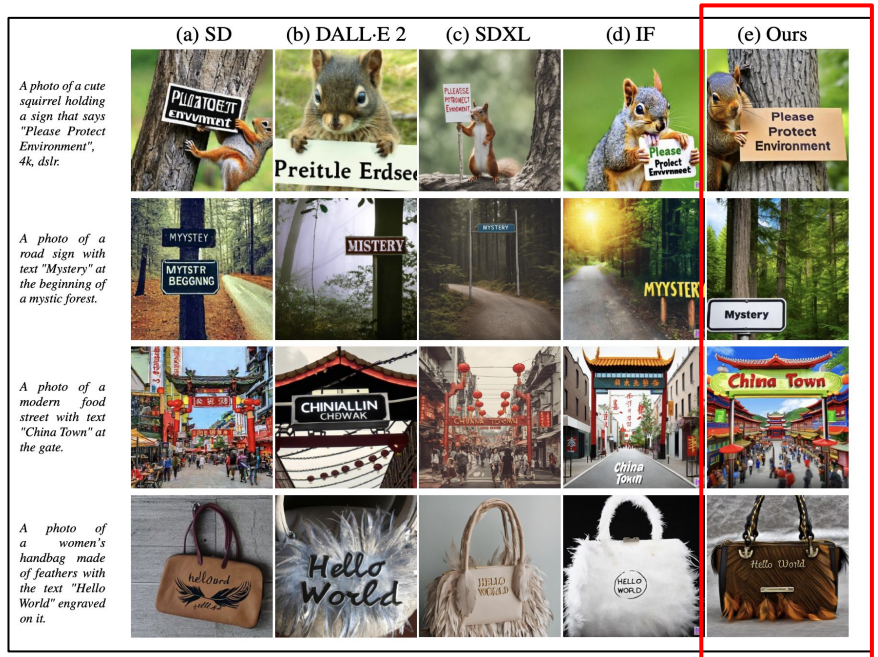
OCR Accuracy

Method	#Params	Text Encoder	Training Dataset	Acc(%)↑	Âcc(%)↑	LD ↓
Stable Diffusion v2.0	865M	CLIP(354M)	LAION 1.2B	0/0	3/2	4.25/5.01
SDXL 1.0	5.8B	CLIP & OpenCLIP(817M)	Internal Dataset (>100M)	0.3/0.5	13/8	6.26/6.30
DeepFloyd (IF-I-M)	2.1B	T5-XXL(4.8B)	LAION 1.2B	0.3/0.1	18/11	2.44/3.86
DeepFloyd (IF-I-L)	2.6B	T5-XXL(4.8B)	LAION 1.2B	0.3/0.7	26/17	1.97/3.37
DeepFloyd (IF-I-XL)	6.0B	T5-XXL(4.8B)	LAION 1.2B	0.6/1	33/21	1.63/3.09
GlyphControl	1.3B	CLIP(354M)	LAION-Glyph-100K	30/19	37/24	1.77/2.58
GlyphControl	1.3B	CLIP(354M)	LAION-Glyph-1M	40/26	45/30	1.59/2.47
GlyphControl	1.3B	CLIP(354M)	LAION-Glyph-10M	**42/28**	**48/34**	**1.43/2.40**

FID & CLIP Score

Method	Stable Diffusion v2.0	SDXL 1.0	DeepFloyd (IF-I-M)	DeepFloyd (IF-I-L)	DeepFloyd (IF-I-XL)	GlyphControl-100K	GlyphControl-1M	GlyphControl-10M
CLIP Score↑	31.6/33.8	31.9/33.3	32.8/34.3	33.1/34.9	33.5/35.2	33.7/36.2	33.4/36.0	**33.9/36.2**
FID-10K-LAION-Glyph ↓	34.03	44.77	23.37	30.97	26.58	**22.04**	22.19	22.22

❑ Visualization

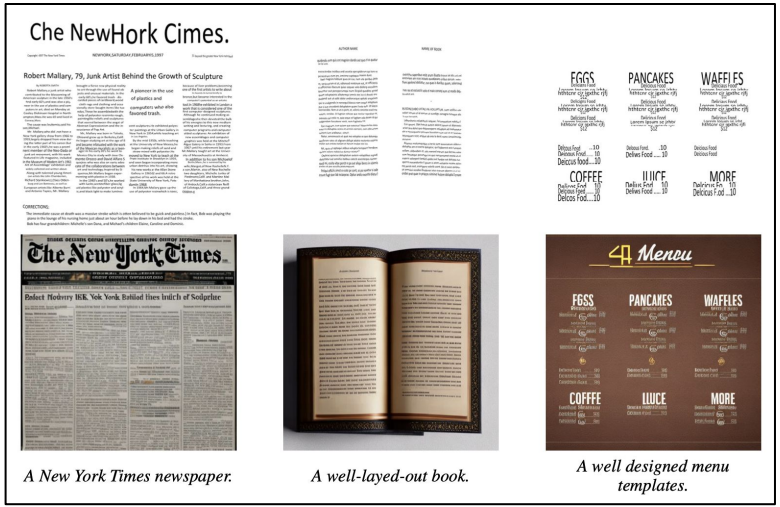


❖ Ablation Studies

- Ablation on Font Size.

Font Size	Acc(%)↑	$\hat{\text{Acc}}$(%)↑	LD ↓	CLIP Score↑
Small	5/4	10/7	4.85/5.51	31.7/33.4
Medium	**30** / 19	**37/24**	**1.77** / 2.58	**33.7/36.2**
Large	23 / **20**	27/23	1.94 / **2.37**	33.1/35.7

- Ablation on a Large Amount of Small Text.



A New York Times newspaper. *A well-layed-out book.* *A well designed menu templates.*

❖ Limitations and Future Work

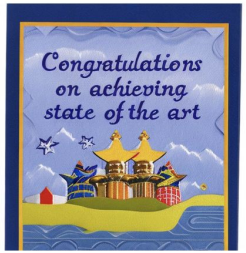
- Lack of the ability to **control font style and text color**
- Sub-optimal performance when generating **a large amount of small text**
- BLIP-2 **captions** still have troubles in consistently representing both image content and OCR text information

- Applying the GlyphControl framework to **high-resolution text-to-image methods** (e.g. SDXL at 1024×1024)
- Exploring further possibilities for **local editing** of visual text within generated images
- **Improving OCR accuracy** of visual text while **keeping the diversity and creativity** of generated images
- ……

❖ Conclusion

*a remarkably **simple yet highly effective** approach for generating **legible and well-formed visual text**.*

Newspaper with the headline "Aliens Found in Space" and "Monster Attacks Mars".



A decorative greeting card that reads "Congratulations on achieving state of the art".

Dslr portrait of a robot holds a sign that says "StrongAI will Empower The World".

A menu of a fast food restaurant that contains "Sandwich Combo", "French Fries", and "Pepsi".

A sign in front of a beautiful village that says "Bear Infested Be Careful".

A sign "OpenSource" facing another sign "CloseSource". They point to two completely different paths.

- Employing **Glyph ControlNet**, which encodes text shape information based on rendered glyph images, as additional conditional control
- Establishing the large-scale visual text generation benchmark dataset **LAION-Glyph** for training

our approach consistently outperforms recent text-to-image models such as the DeepFloyd IF in terms of **OCR accuracy, FID, and CLIP score**.

a valuable foundation for future research in developing robust visual text generation models…

Thank you!