





GlyphControl: Glyph Conditional Control for Visual Text Generation

Yukang Yang^{1†}, Dongnan Gui^{2†}, Yuhui Yuan^{3†‡}, Weicong Liang³, Haisong Ding³, Han Hu³, Kai Chen³

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

[†]Core Contribution ^aInterns at Microsoft Research Asia [‡]Corresponding Author: <u>yuhui.yuan@microsoft.com</u>

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

A photo of a cute squirrel holding a sign that says "Please Protect Environment", 4k, dslr.

A photo of a road sign with text "Mystery" at the beginning of a mystic forest.

A photo of a modern food street with text "China Town" at the gate.

A photo of a women's handbag made of feathers with the text "Hello World" engraved on it.



Attempts: Modifying Text Encoder (more than CLIP)

- Large Language Models like T5 used in Imagen¹ and IF
- Character-aware Language Models like ByT5²

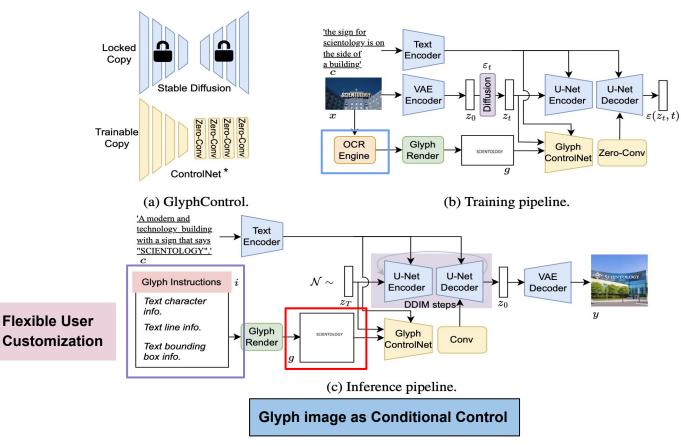
layout errors such as missing or merged glyphs exist

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



*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

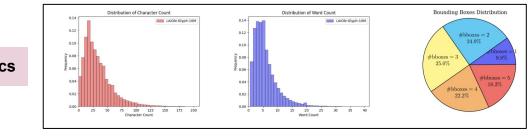
- → Text character information: the *text* placed at the same area (single words, sentences, or phrases)
- → Text line information: assign words to multiple lines by adjusting the *number of rows*
- → **Text box** information:
 - Font size: the *width* property of the text bounding box.
 - Location of 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.



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

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:
 - \circ exact match accuracy Acc
 - capitalization-insensitive exact match accuracy 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	$\mathbf{Acc}(\%)\uparrow$	$\hat{\mathbf{Acc}}(\%)\uparrow$	$\mathbf{LD}\downarrow$
Stable Diffusion v2.0	865M	CLIP(354M)	LAION 1.2B	0/0	3/2	4.25/5.01
SDXL 1.0	$5.8\mathbf{B}$	CLIP & OpenCLIP(817M)	Internal Dataset (>100M)	0.3/0.5	13/8	6.26/6.30
DeepFloyd (IF-I-M)	$2.1\mathbf{B}$	T5-XXL(4.8B)	LAION 1.2B	0.3/0.1	18/11	2.44/3.86
DeepFloyd (IF-I-L)	$2.6\mathbf{B}$	T5-XXL(4.8B)	LAION 1.2B	0.3/0.7	26/17	1.97/3.37
DeepFloyd (IF-I-XL)	6.0 B	T5-XXL(4.8B)	LAION 1.2B	0.6/1	33/21	1.63/3.09
GlyphControl	1.3 B	CLIP(354M)	LAION-Glyph-100K	30/19	37/24	1.77/2.58
GlyphControl	$1.3\mathbf{B}$	CLIP(354M)	LAION-Glyph-1M	40/26	45/30	1.59/2.47
GlyphControl	1.3 B	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





- Ablation Studies
 - Ablation on Font Size.

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

• Ablation on a Large Amount of Small Text.

-

Che I					AUTOR MINE	AMR IF ROOK			
opt of the local test	NEWYORK, SATURDAY	FEBRUARYS, 1987	These Segments in the second	1000 miles	congression and an owner				
hast Mallers	70 1 1 1 1 1 1 1 1 1 1	No. 4 Mar. 6		1111		with one has been as a set	5000	DANIGANES	
bert Mallary,	79, Junk Artist Bel				Later of the second sec	Internet specific all publications and units of an internet an exercision audient of the part, starting from particulation, so publication and services of the service from particulation of the service of the service of the service from particulation of the service of the service of the service from particulation of the service of the service of the service from particulation of the service of the service of the service from particulation of the servi	EGGS	PANCAKES	WAFFIES
ent Mallan, o just artist also rituded to the blockweing of our analytics in the block cliffshi		A pioneer in the use of silutrics and	to a difference of the first protocol and the data in the first of the first of the data in the second seco	and the second	I pro migos en bring ecoliti eroma non ingel andre prof delenar en con contente en	No se chalaith ais é nais anns an chadh Ba Bealaí at	hithtone ciy lipsthic rift Delicips Faed	Athtone cig (gathe r/f)	hebitone cir jestiho rifij
car and but in the bit offers and the and and and the appli- te the part of plantics and can recte all, deal at Mandag at		of prastics and computers who also	compared particular as where	Augend of a second seco	k dite velleranget gelok angehilt neregelikasis nar ange Allgibuk k bigdeleragan hak alt iP 1815 ur synt, it, alleri, anona saring	-		Delicious Food htthicner cjr. (gathic rjif)	helicians front helicians front helicians front
s Bulansen hogelar is North m. Ries. He was til antifisiot it.		favored trash.	work that is considered une-of the first computer designed scotphares. Attempt for continued making as another and the scottinued making as		ur a gan, a labol anana anina Ignar ahar sa tatu ngata A Labi ana at'gan ak tatu ngata A Labi ana at'gan ak tatu ta	niter invasid al entre a santije stragte folger de Transferante	hightone cy (gethorigh) Delicion Food	Public Fred	
	Red monthletimest for angel of Alexandri Copyrighterior and the in- munitories of Page Art	ernt ausgebornung och schlatter ausgeborn för pasströge at den utbare barkers in New Task in 1954 anlete baseling att	pline anogenite this new modern antiog and lensing, and relating computer programs and relation attest compares. An advantage of	beingut it	alor as should have be	Interpret Recting and Annual and Annual Interpret Recting and and and an and a PAT an Abrican Record of Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Annual Annua	Million of Mathe (8)	Laren ipsun se jrhter	Mathcow car spathic right
fork particly show that, once has	Chinese process in Beckeley, Said	in the start of the state in solution	attent and pares. An addition of trace executively and the impactor prophers and heat of the Markoff Alexe Interna in Annual (1999)	Alter setting and the set of the se	na anno 15 degat na altate a gus bhanga Aguan bhar aiste natarea airte	Remark of Plant & Hold State Control of State			
carls 1860, he was a priori	of the Maxim muchles is a feet-	at the princip of Rea Wester, he begue making chick of and and more wood with prigrate its	100 petition opposite hat your		nienze Antoninsker son to opristikas aslagischergan ^{er} fersker bise Samuresma ^{nt} skingsbase aslasi aslagikas nytef	Representation of the second s	Debous Food	Deboui Ford	Decision Food Decision Food
el esperant, activito cont estis che magnete, schador Acason d'Abden art 180 Acasoniger contenta and contenta esperante dest	Monitor The In study with Inter The monte Orseon and Inside Alfans 5 product on the set of the se	moved in New York in teach of the Institution in Examples in 2021, and seen logarithe presenting mere-	Pradiffor to he on Writer's Building to he on Writer's Building of the second ly in		the behave set of a sectory to the set of a sectory of the sectory	10 Michael Stationaria en engenerational ante estationaria estatuti estationaria estationaria est estationaria estationaria estationaria estationaria filmania estationaria estationaria estationaria estationaria filmania estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria estationaria est	Deiccus food10	pervisiroou	of the second se
		Harrison works at the idland barran Indexes at 1800 and 1800 parts	a son Alaman, area of New Racheller Seat development Workels business of		stream print, support with	In particular, for approximation of the same prime of any state of the same of the same advances of the order and part in prime science before advances of the	COFFEE	ILLICE	MODE
el Marikovi () (Ner-Mile- Lerrar lamon, a seña			of the back (off a sinterious half				Delicos Fod 10	Delius Ford 10 Delives Food	MORF Delcicus Eq. 10
rtuna Tapo, M. Balary	Condition (All to ends being a	us al polycler scattered is take,	Contra d				Delcos Foet 10 Delcos Foet 10 Delcos Foet 10	Delivis Food 10	Delcicus F.od10
Ehe a	New y	ork,T	imes.					<mark>A Menc</mark>	<u>u</u>
The a	New U	ork C	imes.		And many the second sec		FGESS Market of the second sec		WAFFELES WAR

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



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



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



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



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!