Glyph-ByT5: A Customized Text Encoder for Accurate Visual Text Rendering

作者

Poject Leader:

Yuhui Yuan https://scholar.google.com/citations?hl=zh-CN&user=PzyvzksAAAAJ

原来做segmentation，现在做Gen AI+Design

Yuhui Yuan

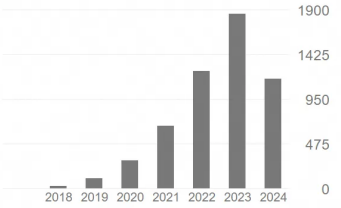
关注

Microsoft Research Asia
在 microsoft.com 的电子邮件经过验证 - 首页
Generative AI + Design Computer Vision

标题	引用次数	年份
FontStudio: Shape-Adaptive Diffusion Model for Coherent and Consistent Font Effect Generation X Mu, L Chen, B Chen, S Gu, J Bao, D Chen, J Li, Y Yuan arXiv preprint arXiv:2406.08392		2024
Step-aware Preference Optimization: Aligning Preference with Denoising Performance at Each Step Z Liang, Y Yuan, S Gu, B Chen, T Hang, J Li, L Zheng arXiv preprint arXiv:2406.04314		2024
DesignEdit: Multi-Layered Latent Decomposition and Fusion for Unified & Accurate Image Editing Y Jia, Y Yuan, A Cheng, C Wang, J Li, H Jia, S Zhang arXiv preprint arXiv:2403.14487		2024
Glyph-byt5: A customized text encoder for accurate visual text rendering Z Liu, W Liang, Z Liang, C Luo, J Li, G Huang, Y Yuan arXiv preprint arXiv:2403.09622	3	2024

引用次数

	总计	2019 年至今
引用	5435	5390
h 指数	15	15
i10 指数	19	19



开放获取的出版物数量 查看全部

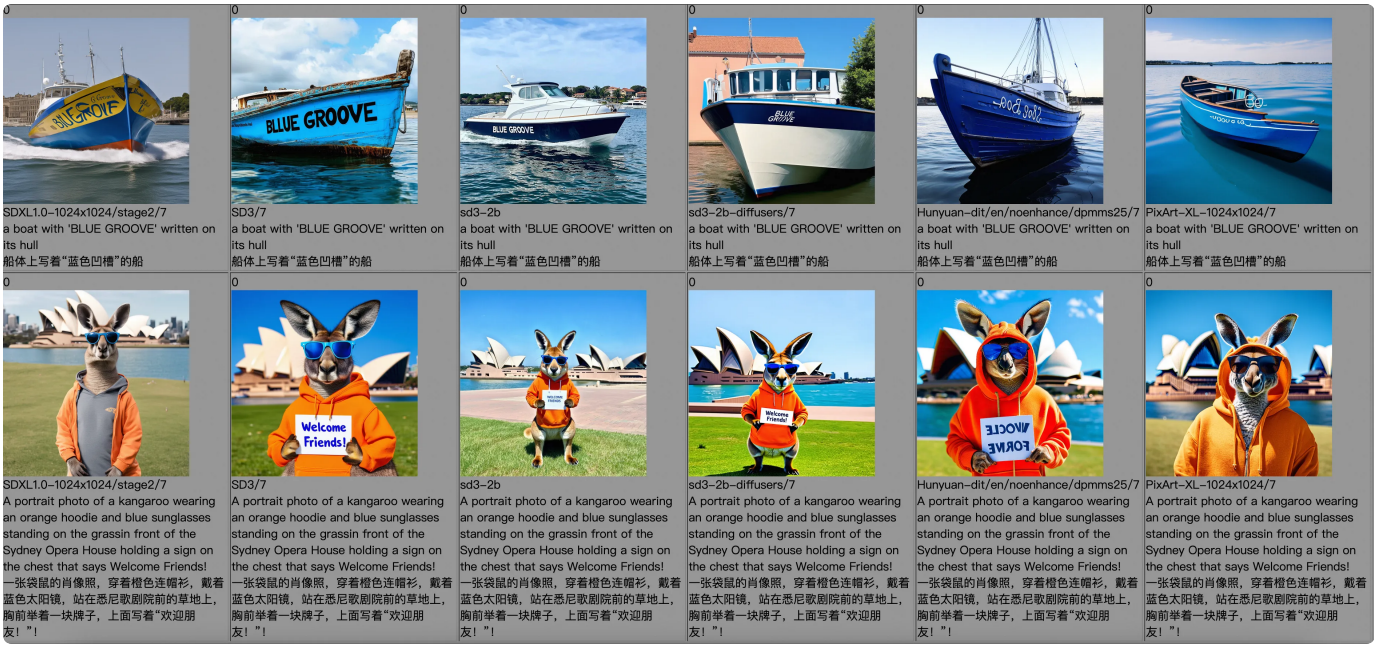
0 篇文章 11 篇文章

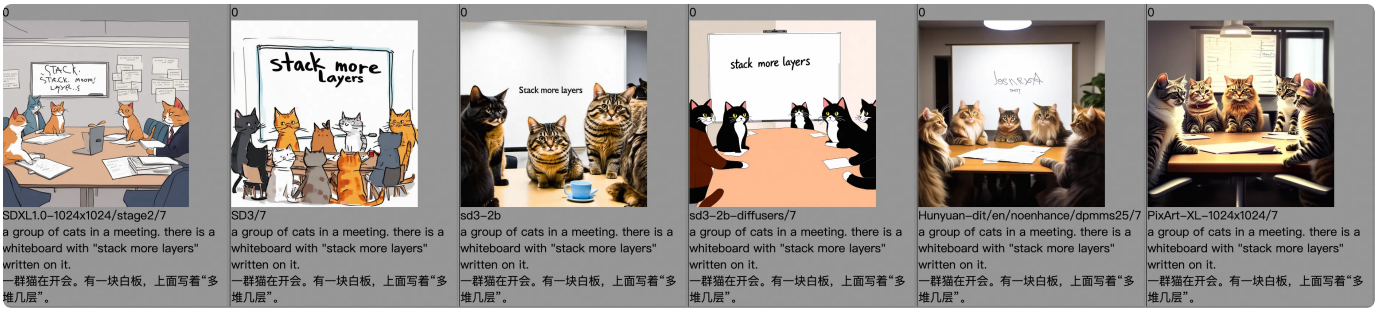
无法查看的文章 可查看的文章

根据资助方的强制性开放获取政策

背景

当前文生图模型难以在图像上准确的文字（2024.06.19）





任务定义

- 可控文生图->文本内容、文本位置可控（视觉文本渲染）
- 输入：global text prompt、[(content_1, position_1), ... , (content_n, position_n)]
- 输入：image

论文效果

研究动机

作者认为现有text2image模型无法生成准确的视觉文本，是因为text encoder的限制

- text encoder(CLIP)预训练过程对齐整张图的视觉信号与text prompt

Table 1: Comparison of *SDXL* and older *Stable Diffusion* models.

Model	*SDXL*	SD 1.4/1.5	SD 2.0/2.1
# of UNet params	2.6B	860M	865M
Transformer blocks	[0, 2, 10]	[1, 1, 1, 1]	[1, 1, 1, 1]
Channel mult.	[1, 2, 4]	[1, 2, 4, 4]	[1, 2, 4, 4]
Text encoder	CLIP ViT-L & OpenCLIP ViT-bigG	CLIP ViT-L	OpenCLIP ViT-H
Context dim.	2048	768	1024
Pooled text emb.	OpenCLIP ViT-bigG	N/A	N/A

- text encoder无法提取字符级文字特征

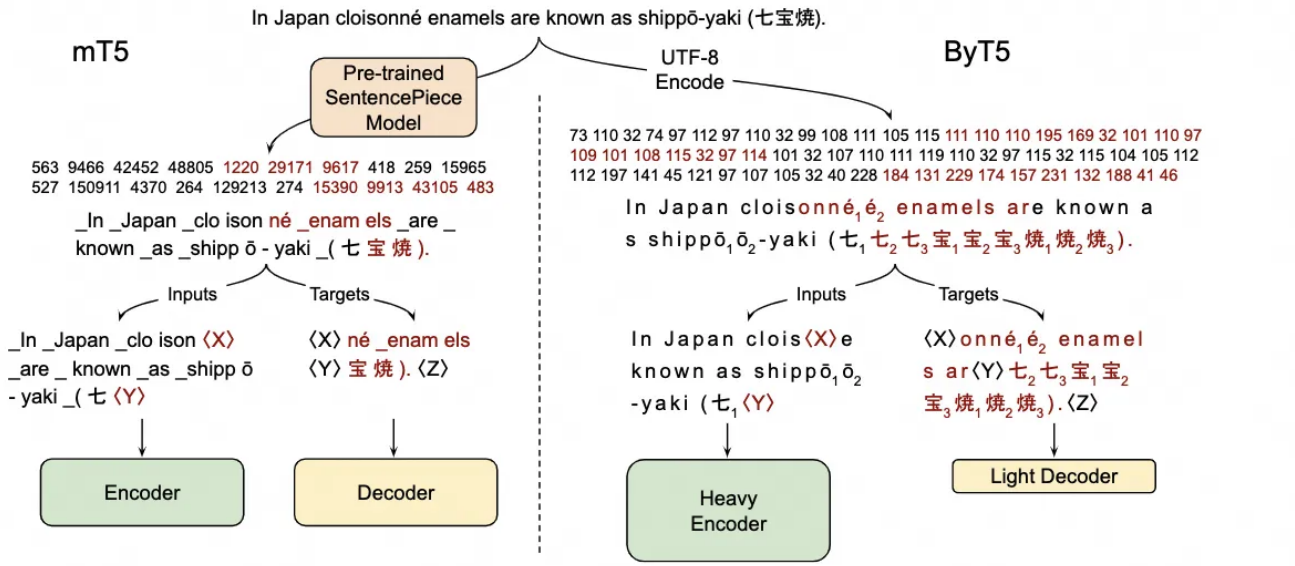


Figure 1: Pre-training example creation and network architecture of mT5 (Xue et al., 2021) vs. ByT5 (this work). **mT5**: Text is split into SentencePiece tokens, spans of ~3 tokens are masked (red), and the encoder/decoder transformer stacks have equal depth. **ByT5**: Text is processed as UTF-8 bytes, spans of ~20 bytes are masked, and the encoder is 3× deeper than the decoder. ⟨X⟩, ⟨Y⟩, and ⟨Z⟩ represent sentinel tokens.

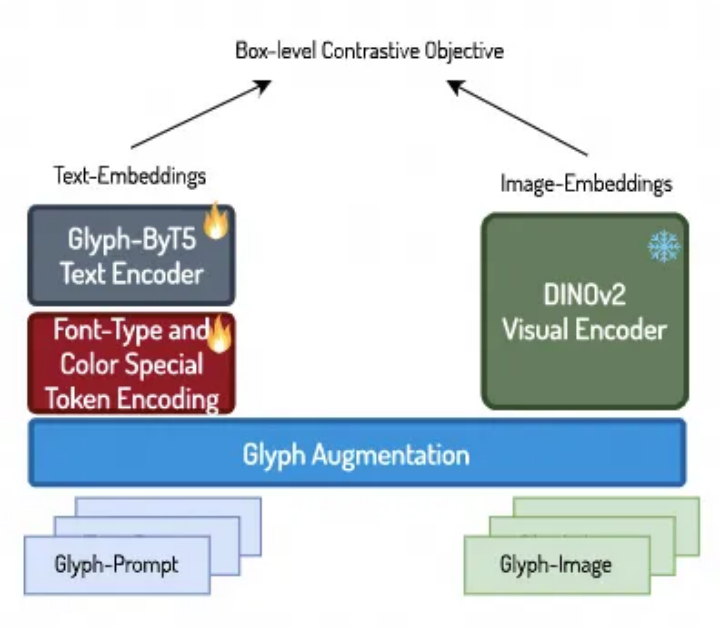
论文方法

需要更强的text encoder！编码字符级特征且与对应glyph image对齐

总体流程：

1. 利用预训练的ByT5作为text encoder
2. 利用glyph image与对应text进行对比学习再训练text encoder，再得到Glyph-ByT5
3. 将Glyph-ByT5通过Region-wise Multi-head Cross-Attention的方式合并到SDXL中，在合成数据上微调，得到Glyph-SDXL
4. 将Glyph-SDXL在场景文本图像上微调，增强Scene-text Generation能力

对比学习（Glyph-Alignment Pre-training）



- **数据**
  - 1 million pairs of synthetic data
  - text content 从语料库采样并随机替换word、512种font、100种颜色
  - 参考：Cole: A hierarchical generation framework for graphic design (https://arxiv.org/abs/2311.16974)
- **glyph-prompt**
  - 文本内容+文本颜色+文字字体
  - {Text "oh! the places you'll go!" in [font-color-39], [font-type-90]. Text "Happy Graduation Kim!" in [font-color-19] [font-type181]}
- **glyph-Image**
  - 句子级文本 

段落级文本（>100）



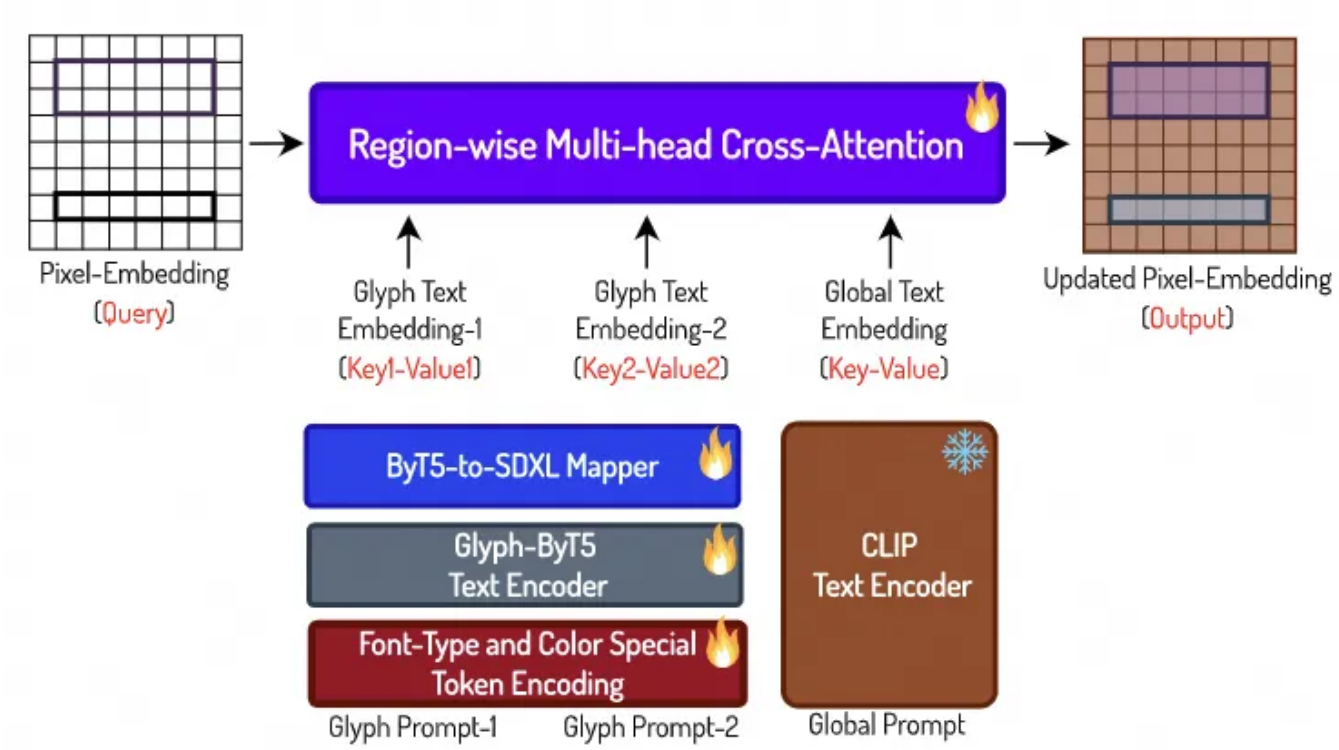
- **文本编码器：**文本内容采用ByT5，字体和颜色采用特殊的token

- **视觉编码器**：DINOv2（效果最好）
- **Box-level Contrastive Loss**：
  - 一个batch内，任意文本和其对区域的文本图像作为正样本，其他都是负样本
  - 正样本：内容、颜色、字体、位置相同
  - 负样本：内容、颜色、字体、位置随机不同
  - $$\mathcal{L}_{\text{box}} = -\frac{1}{2\sum_{i=1}^{|\mathcal{N}|}|\mathcal{B}_i|}\sum_{i=1}^{|\mathcal{N}|}\sum_{j=1}^{|\mathcal{B}_i|}(\log\frac{e^{t\mathbf{x}_i^j\cdot\mathbf{y}_i^j}}{Z_x}+\log\frac{e^{t\mathbf{x}_i^j\cdot\mathbf{y}_i^j}}{Z_y}), \quad (1)$$
- **Hard-negative Contrastive Loss based on Glyph Augmentation**：
  - 数据增强后的样本对作为负样本，同时在本文和图像上做
  - 正样本：内容、颜色、字体、位置相同
  - 负样本：颜色、字体、位置相同，内容不同
  - 

    Figure 2. Illustrating the scheme of glyph augmentation. (a) original glyph. (b) character replacement (Happy → Hdppy). (c) character repeat (Happy → Happppppy). (d) character drop (Happy → Hapy). (e) character add (Graduation → Gradumation). (f) word replacement (Graduation → Gauatikn). (g) word repeat (Kim → Kim Kim). (h) word drop (Happy Graduation Kim → Graduation).
  - $$\mathcal{L}_{\text{hard}} = -\frac{1}{2\sum_{i=1}^{|\mathcal{N}|}|\mathcal{B}_i|}\sum_{i=1}^{|\mathcal{N}|}\sum_{j=1}^{|\mathcal{B}_i|}(\log\frac{e^{t\mathbf{x}_i^j\cdot\mathbf{y}_i^j}}{Z_x^{\text{aug}}}+\log\frac{e^{t\mathbf{x}_i^j\cdot\mathbf{y}_i^j}}{Z_y^{\text{aug}}}), \quad (2)$$

将Glyph-ByT5通过Region-wise Multi-head Cross-Attention的方式合并到SDXL中

- 在sd的cross-attention中，visual特征是query，text特征是key&value，输出序列维度visual特征一致
- glyph text prompt->想要渲染的文字，global prompt->背景&风格
- 每个glyph text prompt特征只于对应位置的visual特征计算注意力
- global prompt特征只和背景visual特征计算注意力
- 最后将视visual特征再进行融合



(b) Region-wise Multi-Text-Encoder Fusion

数据

build a **high-quality visual design image dataset** with **dense paragraph-level** visual text rendered on each image by crawling from a lot of graphic design websites following COLE

- 打标：LLaVA based on Llama2-13B
- 数据清理：去除和glyph alignment pre-training中一样的typography
- 数据量：1M

微调

CLIP text冻住，Unet进行Lora，cross-attention训练，引入ByT5-to-SDXL mapper (transformer block)

Design-to-Scene Alignment

由于上一阶段使用合成数据会导致Scene-text Generation下降，所以需要进行微调

build a hybrid design-to-scene alignment dataset

- 4,000 scene-text and design text images from TextSeg
- 4,000 synthetic images generated using SDXL
- 4,000 design images

fine-tune 2 epochs

实验

定性结果

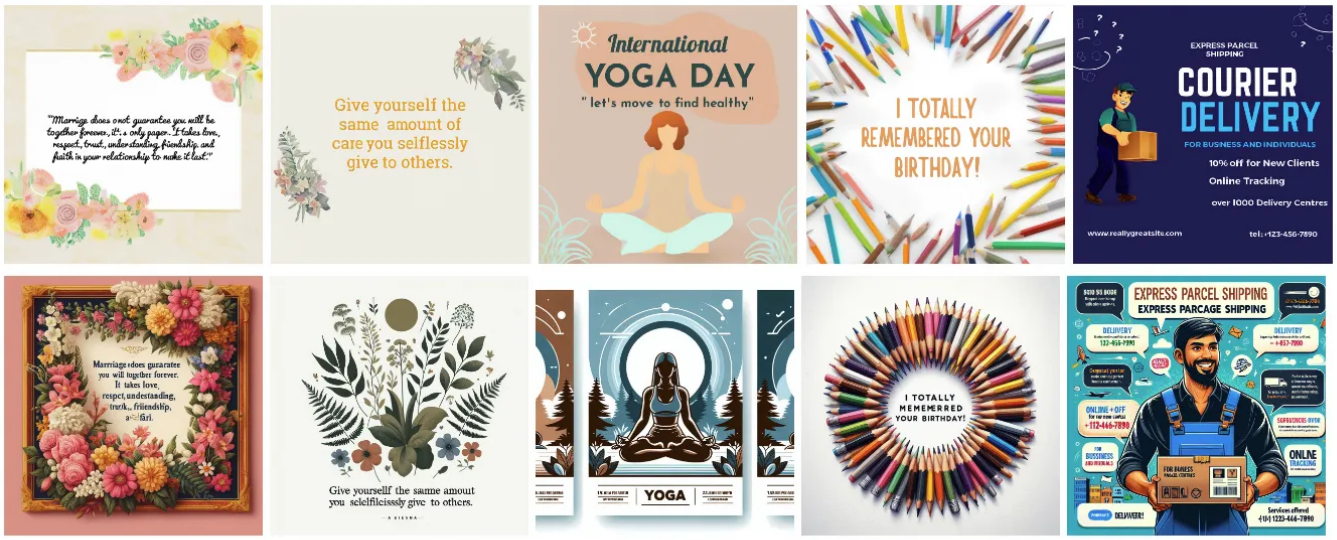


Figure 5. Qualitative comparison results. We show the results generated with our Glyph-SDXL and DALL·E3 in the first row and second row, respectively.

定量对比

Method	SimpleBench			CreativeBench			MARIO-Eval			
	Recall	Case-Recall	Edit-Dis.	Recall	Case-Recall	Edit-Dis.	Accuracy [IMG]	Precision	Recall	F-measure
DeepFloyd IF [14]	0.6	33	1.63	1	21	3.09	2.6	14.5	22.5	17.6
GlyphControl [29]	42	48	1.43	28	34	2.40	-	-	-	-
TextDiffuser [6]	-	-	-	-	-	-	56.1	78.5	78.0	78.2
TextDiffuser-2 [7]	-	-	-	-	-	-	57.6	74.0	76.1	75.1
Glyph-SDXL	**93.56**	93.62	0.09	**92.00**	**92.06**	0.16	**74.8**	**88.2**	**92.6**	**90.4**
Glyph-SDXL-Scene	92.69	**95.88**	**0.05**	88.81	91.38	**0.15**	66.5	83.9	89.0	86.4

消融实验

Method	#Params	Char-aware	Glyph-align	Precision (%)			
				≤20 chars	≤20-50 chars	≤50-100 chars	≥100 chars
SDXL (CLIP & OpenCLIP)	817M	✗	✗	21.72	20.98	18.23	19.17
+ T5-L	+ 394M	✗	✗	48.46	44.89	34.59	26.09
+ ByT5-S	+ 292M	✓	✗	60.52	52.79	50.11	42.05
+ Glyph-ByT5-S	+ 292M	✓	✓	92.58	90.38	87.16	83.17
+ Glyph-ByT5-S^{1M}	+ 292M	✓	✓	**93.89**	**93.67**	**91.45**	**89.17**
DeepFloyd-IF (T5-XXL)	4.3B	✗	✗	17.63	17.17	16.42	13.05
DALL·E3	Unknown	✗	✗	23.23	21.59	20.1	15.81

视觉编码器的影响

Visual encoder	Precision (%)			
	≤20 chars	≤20-50 chars	≤50-100 chars	≥100 chars
DINOv2 ViT-B/14 + reg	**84.54**	**84.56**	**79.89**	**73.29**
CLIP ViT-B/16	77.17	74.78	74.94	66.34
ViTSTR	79.29	78.2	75.35	68.49
CLIP4STR ViT-B/16	80.38	79.12	77.08	69.24

NOTE：带STR结尾的是在文本识别任务上进行过预训练的

数据增强的影响

Glyph aug. ratio	Precision (%)			
	≤20 chars	≤20-50 chars	≤50-100 chars	≥100 chars
None	78.93	78.35	74.0	65.40
1:8	81.15	80.45	77.03	70.03
1:16	**84.54**	84.56	**79.89**	**73.29**
1:32	83.24	**85.02**	78.92	72.16

损失函数的影响

Loss design	Precision (%)			
	≤20 chars	≤20-50 chars	≤50-100 chars	≥100 chars
IL-CL	83.13	81.83	77.15	69.42
BL-CL	**84.54**	**84.56**	**79.89**	**73.29**
IL-CL + BL-CL	83.86	82.08	78.07	70.54

glyph映射器的影响

ByT5-to-SDXL mapper	Precision (%)			
	≤20 chars	≤20-50 chars	≤50-100 chars	≥100 chars
w/o mapper	80.22	78.48	72.91	65.02
w/ mapper	**84.54**	**84.56**	**79.89**	**73.29**

text encoder参数量

Text encoder	#Params	Precision (%)			
		≤20 chars	≤20-50 chars	≤50-100 chars	≥100 chars
Glyph-ByT5-S	292M	84.54	84.56	**79.89**	73.29
Glyph-ByT5-B	510M	**87.10**	**84.93**	78.72	72.81
Glyph-ByT5-L	963M	87.07	82.87	79.12	**73.72**

对比学习预训练数据量

# Glyph Image-Text	Precision (%)			
	≤20 chars	≤20-50 chars	≤50-100 chars	≥100 chars
100K	85.6	85.02	81.2	74.58
500K	91.11	93.35	85.43	82.83
1M	**93.54**	**93.96**	**91.0**	**89.96**

注意力可视化