基于扩散模型的文本图像生成进展

分享人：高逸凡

2024.01.04

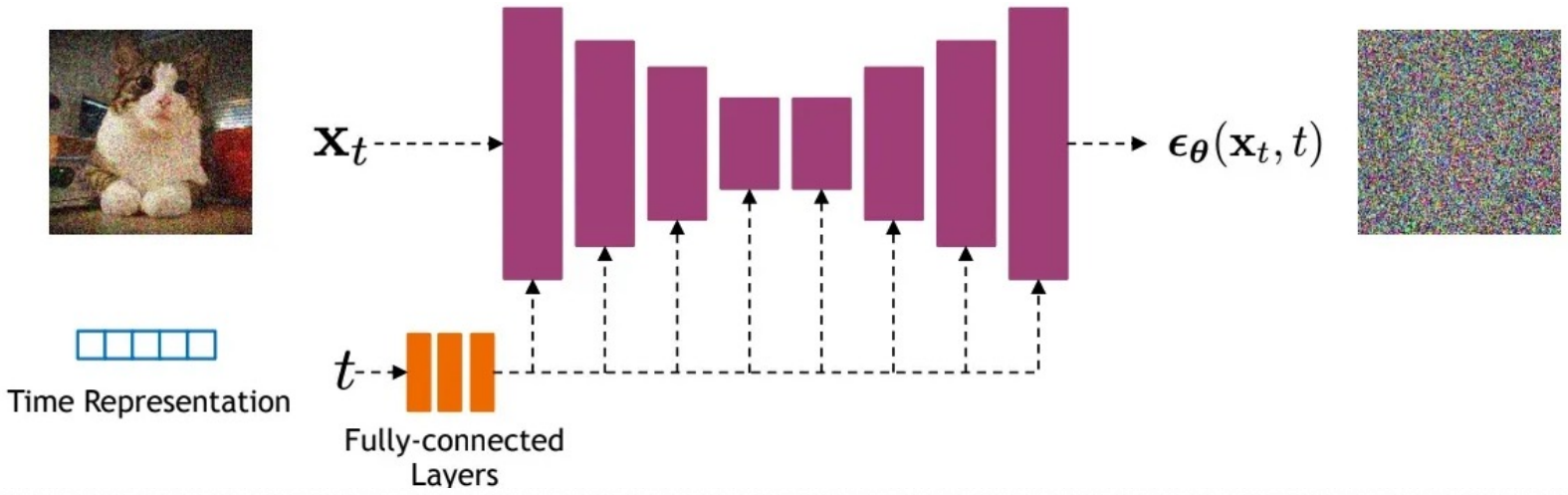
目录

- 作者介绍
- 研究动机
- 本文方法
- 实验效果
- 实验结果

- 作者介绍
- 研究背景
- 研究动机
- 本文方法
- 实验结果

研究背景

- 扩散模型(Diffusion Model)
  - DDPM



Algorithm 1 Training

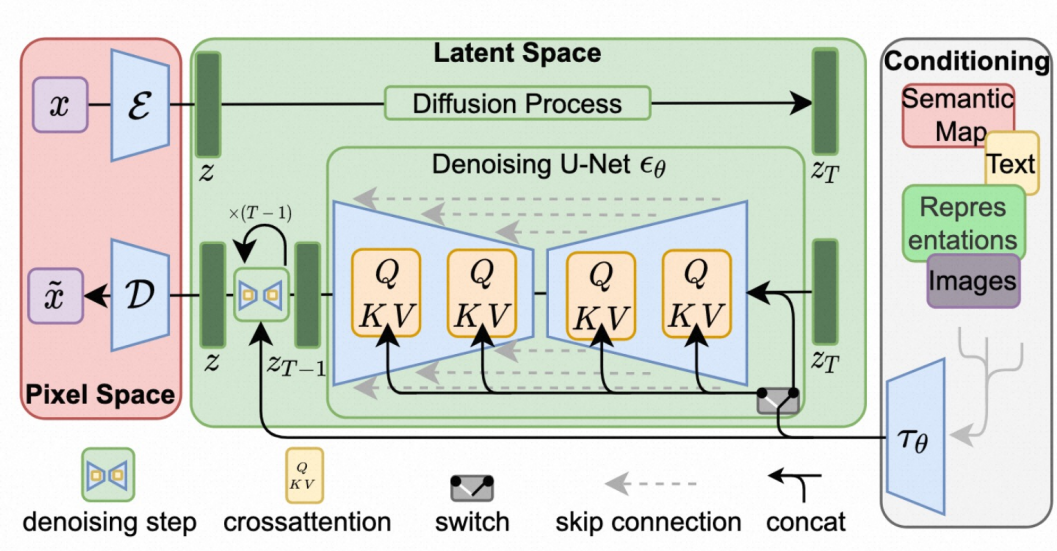
1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

研究背景

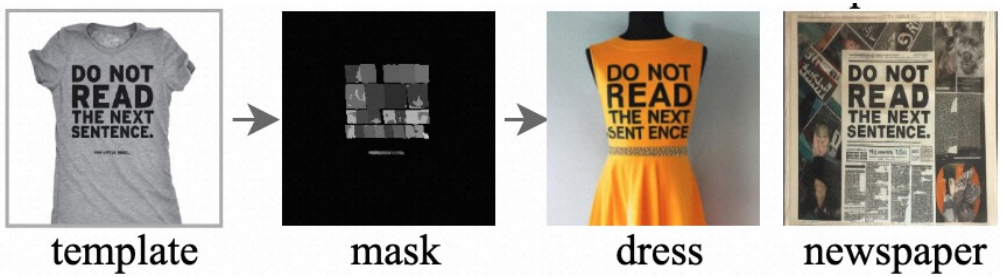
- 隐扩散模型(Latent Diffusion Model)
  - Latent Space：高分辨率生成
  - Cross Attention：可控生成（eg. T2I）



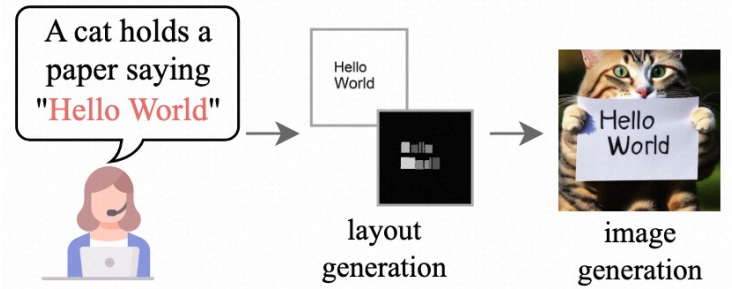
$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t)\|_2^2\right].$$

研究背景

- 文本图像生成任务
  - Text-to-Image
    - w/ Layout

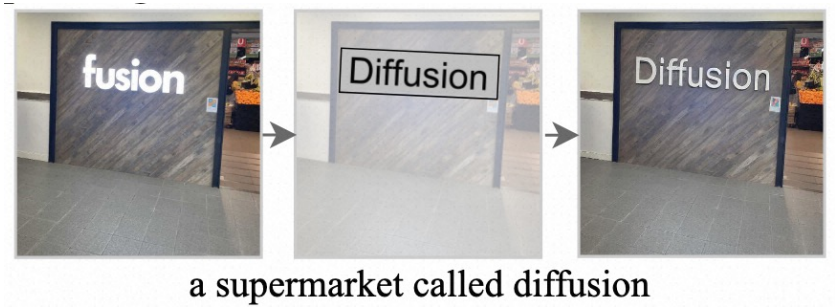


    - w/o Layout

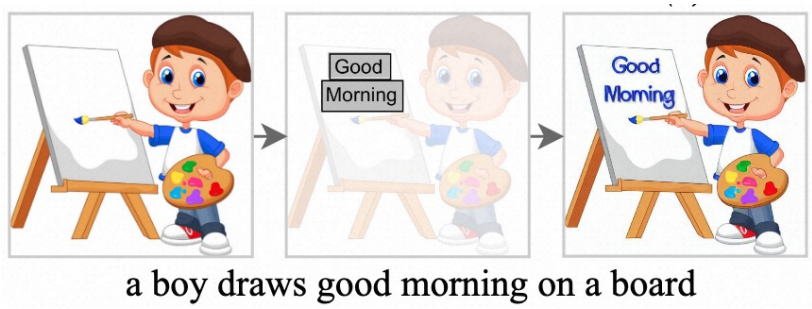


研究背景

- 文本图像生成任务
  - Text Inpainting
    - Text Editing



    - Text Generation



研究趋势

https://github.com/yeungchenwa/Recommendations-Diffusion-Text-Image

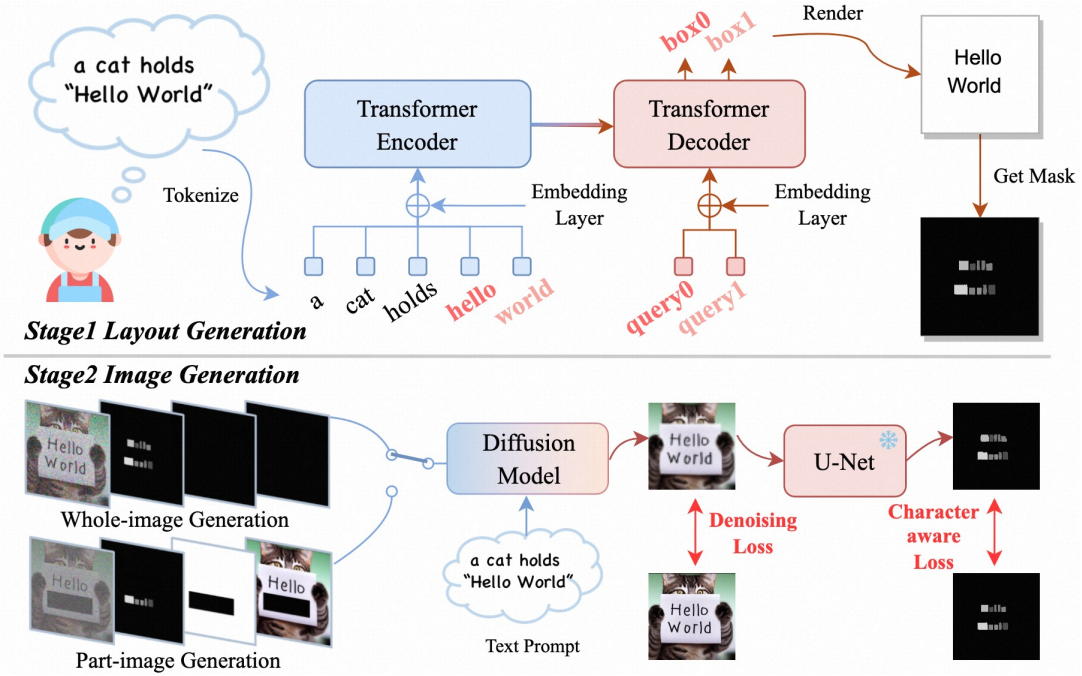
- 研究任务
  - Multi-task or single-task
- 微调方式
  - Full FT，PEFT，ControlNet，Training-free et al.
- 数据
  - T2I datasets with OCR annotations and OCR datasets
- 研究内容
  - text information
  - glyph information
  - layout information

- 作者介绍
- 背景介绍
- 研究动机
- 方法介绍
- 实验结果

TextDiffuser: Diffusion Models as Text Painters

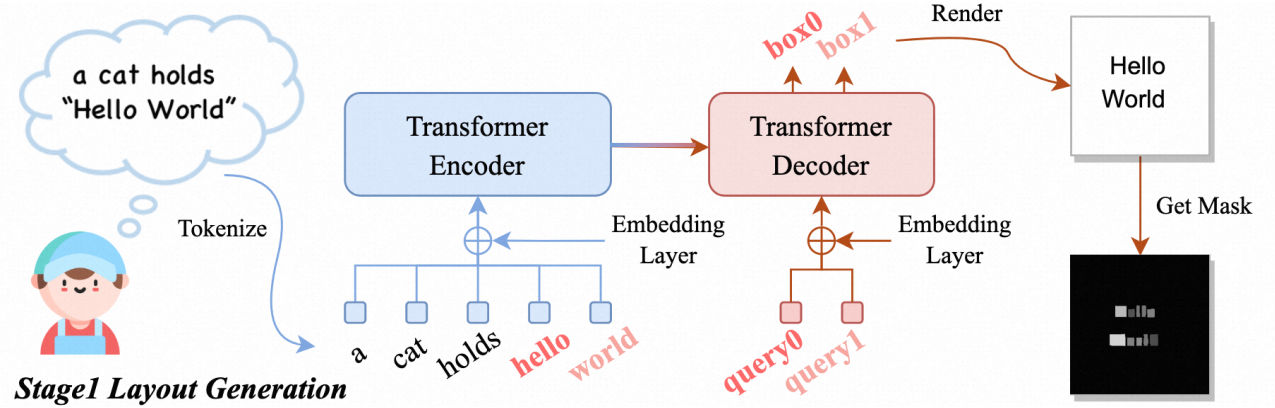
NIPS'23 MSRA

- Task：Text Inpainting & Text-to-Image
- Innovation：Unify Layout and Image Generation
- Language：English



TextDiffuser: Diffusion Models as Text Painters

- Method: Layout Generation



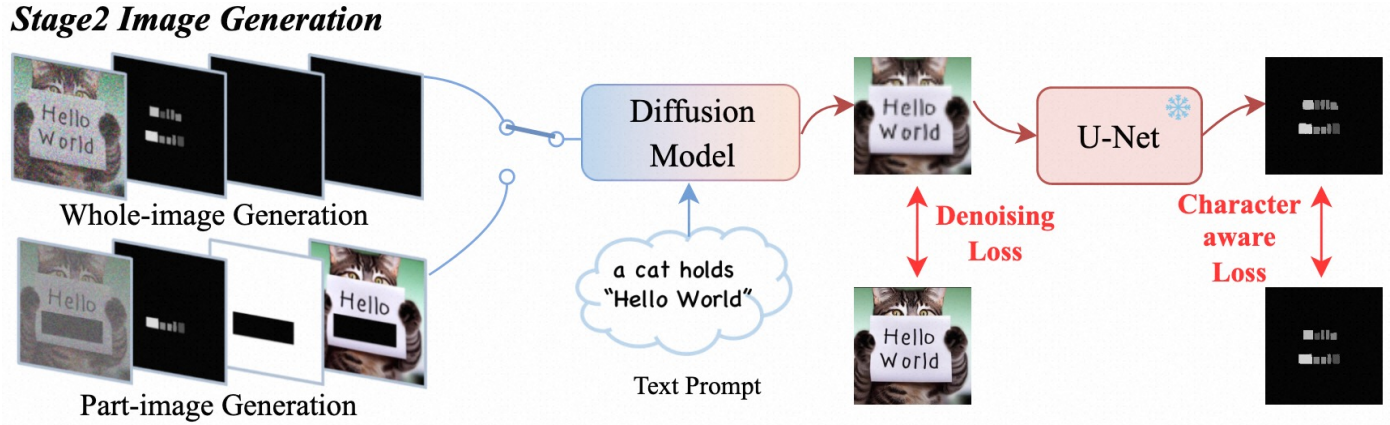
- 1.Text-to-Layout Transformer
- 2.Keyword embedding and width embedding

$$\text{Embedding}(\mathcal{P}) = \text{CLIP}(\mathcal{P}) + \text{Pos}(\mathcal{P}) + \text{Key}(\mathcal{P}) + \text{Width}(\mathcal{P}).$$

- 3.positional embedding as decoder query
- 4.get character-level segmentation

TextDiffuser: Diffusion Models as Text Painters

- Method: Image Generation



  - 1.Input: latent features, character-level segmentation, mask, mask features
  - 2.Loss: denoising loss & character aware loss(pretrain)

$$l = l_{denoising} + \lambda_{char} * l_{char}$$

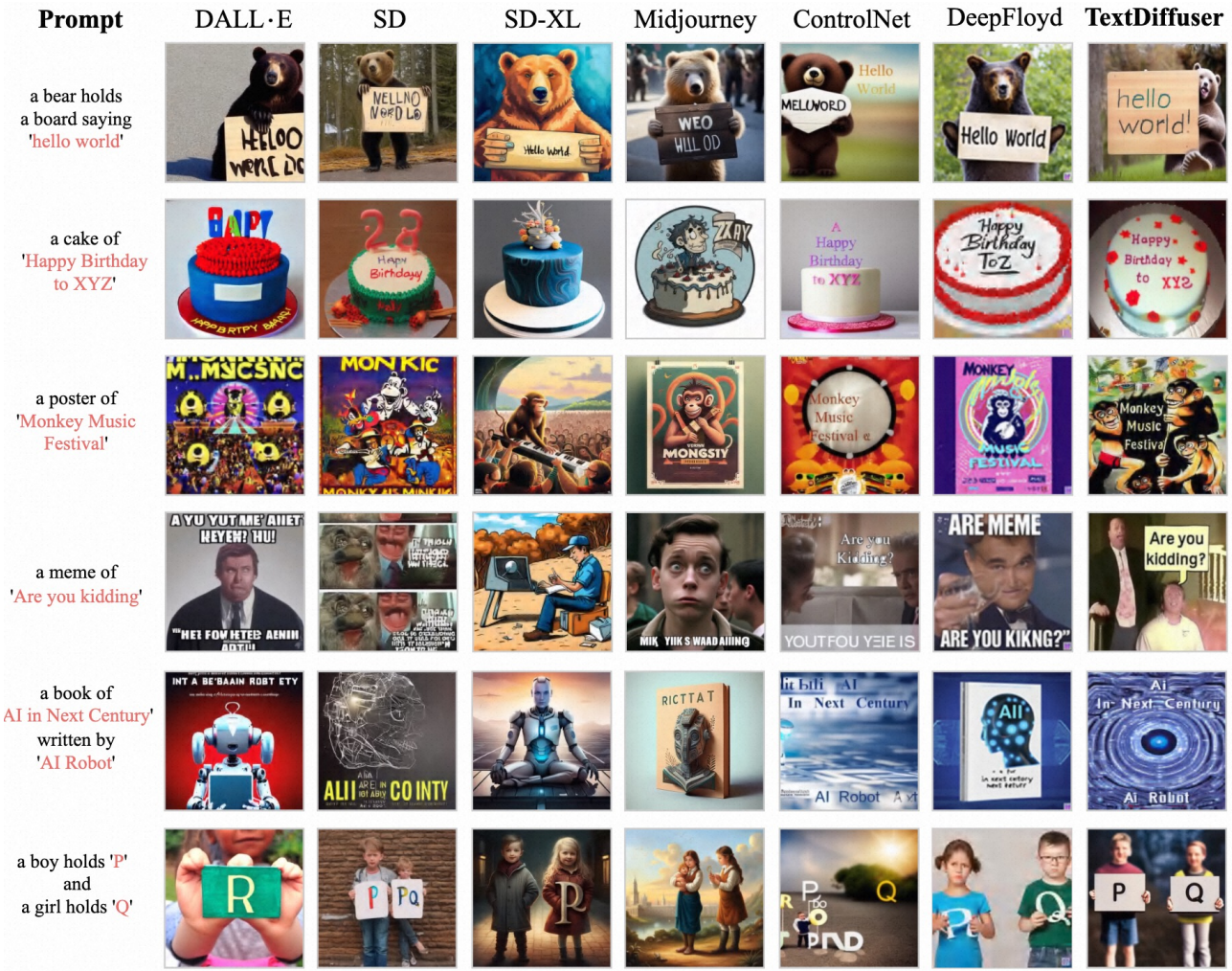
  - 3.train a segmentation model for character aware loss

TextDiffuser: Diffusion Models as Text Painters

- Detail
  - Dataset
    - MARIO-10M
      - OCR for annotations, 10M for training, 61K for testing
      - From LAION-400M, TMDB, OpenLibrary
    - MARIO-Eval Benchmark
      - 5,414 prompts: MARIO-10M test set, DrawBenchText, , DrawTextCreative, ChineseDrawText
  - Implementation
    - Model: runwayml/stable-diffusion-v1-5
    - Resolution: 512*512
    - Training: Full Fine-tuning and 8 Tesla V100 GPUs for 2 day
    - Alphabet: 95 characters

TextDiffuser: Diffusion Models as Text Painters

- Visualizations



TextDiffuser: Diffusion Models as Text Painters

- Visualizations

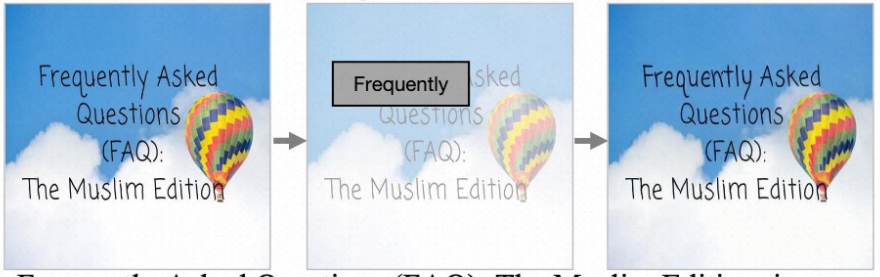
Interesting terrace party

How to make a newspaper

FIMI X8 SE Range Test in Country

She is teaching Chinese lesson.

A man wears a cloth containing handsome boy

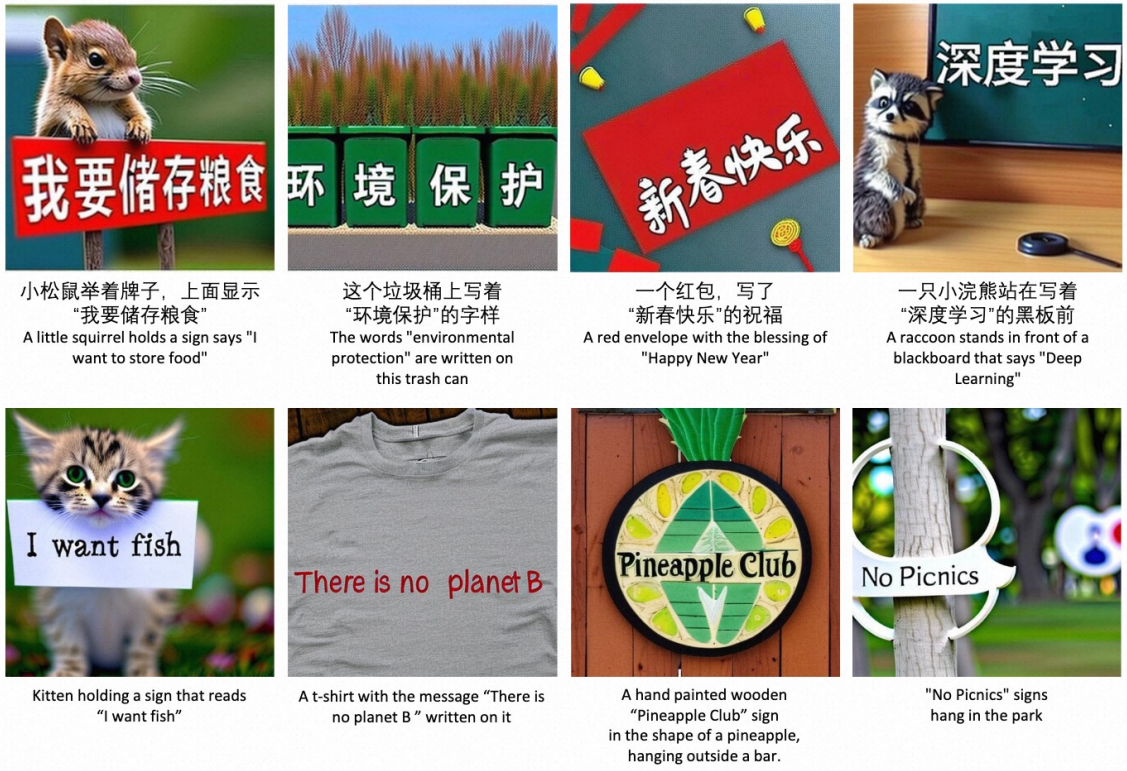


Frequently Asked Questions (FAQ): The Muslim Edition image

GlyphDraw: Seamlessly Rendering Text with Intricate Spatial Structures in Text-to-Image Generation

NIPS'23 Submission by OPPO Research

- Task：Text-to-Image
- Innovation：First diffusion-based method for Chinese
- Language：Chinese&English



小松鼠举着牌子，上面显示
“我要储存粮食”
A little squirrel holds a sign says "I want to store food"

这个垃圾桶上写着
“环境保护”的字样
The words "environmental protection" are written on this trash can

一个红包，写了
“新春快乐”的祝福
A red envelope with the blessing of "Happy New Year"

一只小浣熊站在写着
“深度学习”的黑板前
A raccoon stands in front of a blackboard that says "Deep Learning"

Kitten holding a sign that reads “I want fish”

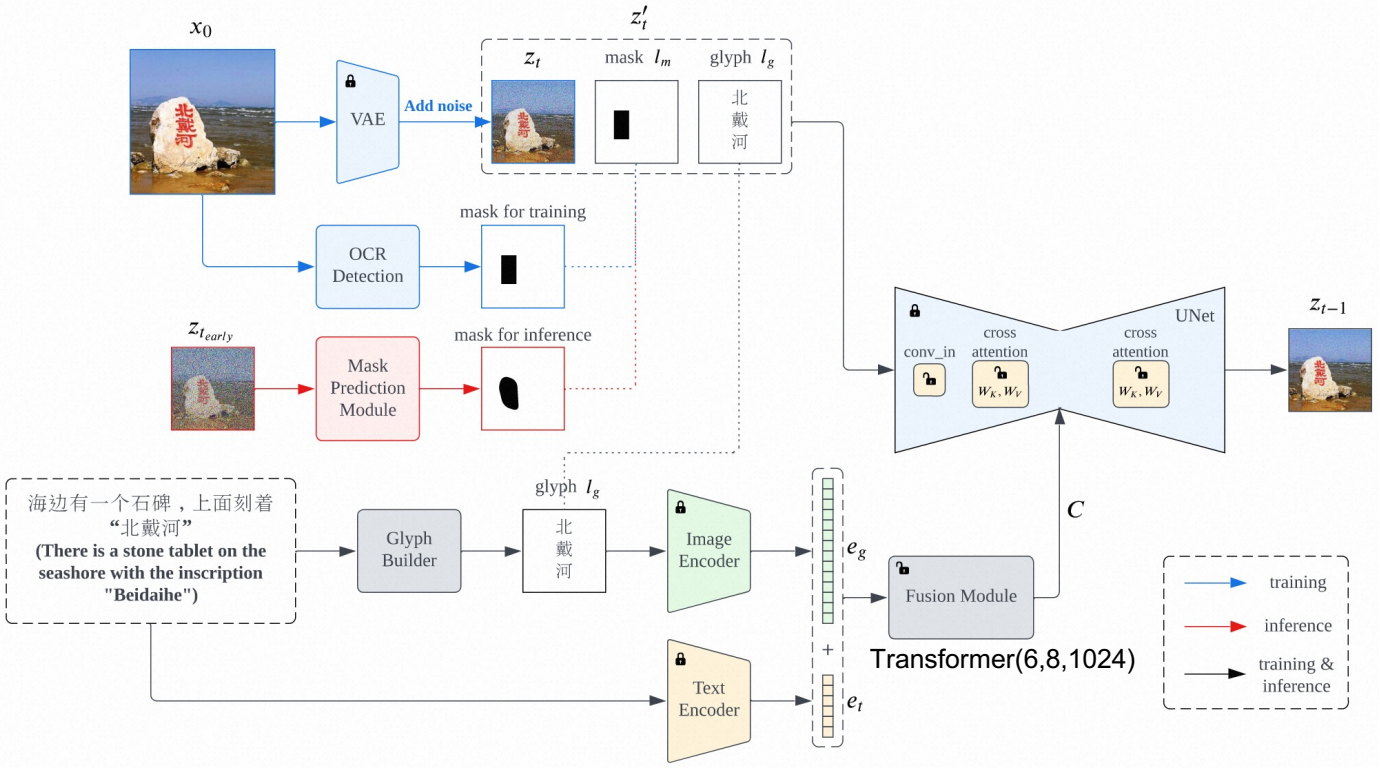
A t-shirt with the message “There is no planet B ” written on it

A hand painted wooden “Pineapple Club” sign in the shape of a pineapple, hanging outside a bar.

"No Picnics" signs hang in the park

GlyphDraw: Seamlessly Rendering Text with Intricate Spatial Structures in Text-to-Image Generation

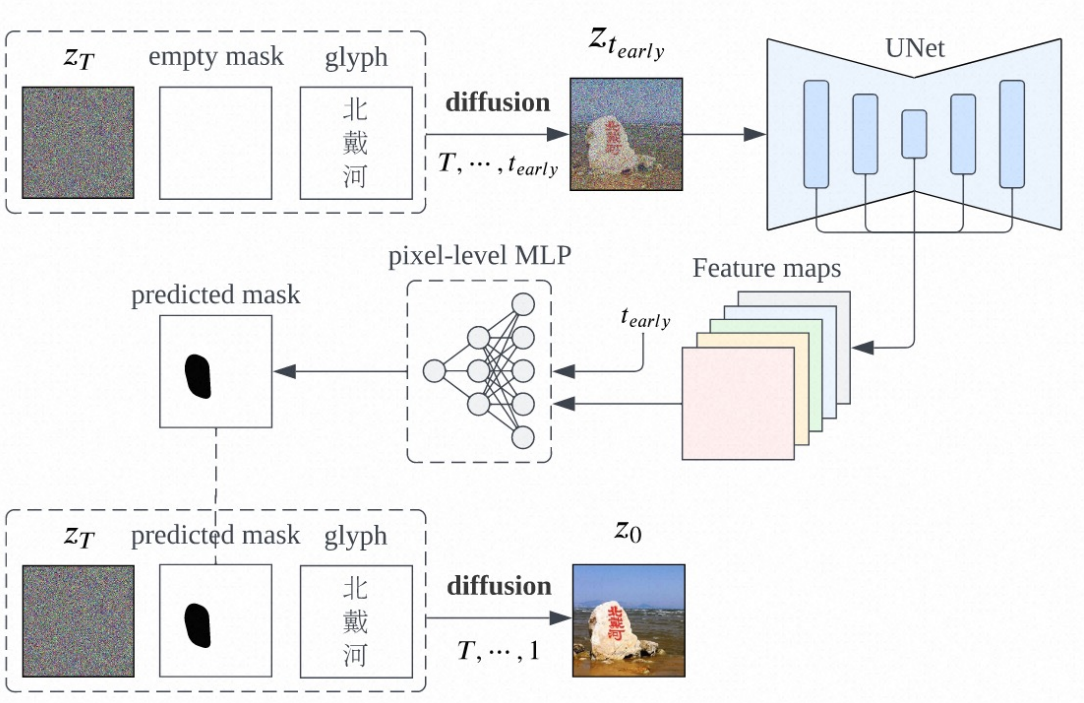
- Method: Training



- Input: latent feature, mask, glyph
- Loss: $\mathcal{L}_{GD}=\mathbb{E}_{\mathcal{E}(x_0),y,l_g,l_m,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon-\epsilon_\theta(z_t,t,y,l_g,l_m)\|_2^2+\alpha\|\left[\epsilon-\epsilon_\theta(z_t,t,y,l_g,l_m)\right]*(1-l_m)\|_2^2\right]$,
- Learnable Parameters: conv_in, cross-attention, fusion Module

GlyphDraw: Seamlessly Rendering Text with Intricate Spatial Structures in Text-to-Image Generation

- Method: inference



- 1.predict mask in early step by the MLP
- 2. regenerate by our Glyphdraw model in early step
- 3. continue by SD without glyph and mask in remaining steps

GlyphDraw: Seamlessly Rendering Text with Intricate Spatial Structures in Text-to-Image Generation

- Detail
  - Dataset
    - From LAION, Zero23M, Wukong (792K for Chinese, 1.9M for English)
    - OCR by PP-OCRv2
    - Regenerate Captions by BLIP-2, embed the characters by quotation marks into the captions
  - Implement
    - SD: stabilityai/stable-diffusion-2-base
      - Pretrain by only fine-tuing the text encoder.
      - Training at the Chinese part of Laion-5B, Noah-Wukong, self-crawled data (100M)
      - 80 A100GPUs for 160K steps
    - CLIP: Chinese-CLIP and OpenCLIP-ViT/H
    - Resolution: 512*512
    - 0.1B parameters are trainable，24 A100 GPUs for 20 epochs

GlyphDraw: Seamlessly Rendering Text with Intricate Spatial Structures in Text-to-Image Generation

- Quantitative experiments

Category	Method	Chinese Model		English Model	
		Accuracy (↑ %)	FID (↓)	Accuracy(↑ %)	FID (↓)
baselines	Stable Diffusion [29]	0.00±0.00	15.87	0.14	14.69
	Stable Diffusion + Fine-tuning	0.00±0.00	20.09	15.01±0.64	20.24
	ControlNetDraw	0.00±0.00	15.91	7.18±0.41	14.63
	Imagen [32]	-	-	44.54±0.68	-
component ablations	Glyphdraw w/o loss weighting	68.60±1.09	17.27	64.96±0.32	16.92
	Glyphdraw w/o location mask	66.15±0.89	17.68	65.79±0.70	17.85
	Glyphdraw w/ random mask	60.23±0.18	16.78	60.55±1.06	16.09
	Glyphdraw w/o glyph latent	69.88±0.52	17.36	65.21±0.57	16.58
training parameter ablations	Glyphdraw + all UNet layers	74.21±0.83	19.01	77.29±0.82	19.78
	Glyphdraw + all UNet attention layers	72.24±0.74	17.95	76.74±0.83	18.01
	Glyphdraw (only $W_k^{(i)}$ and $W_v^{(i)}$)	74.00±1.13	16.89	75.23±1.06	16.02

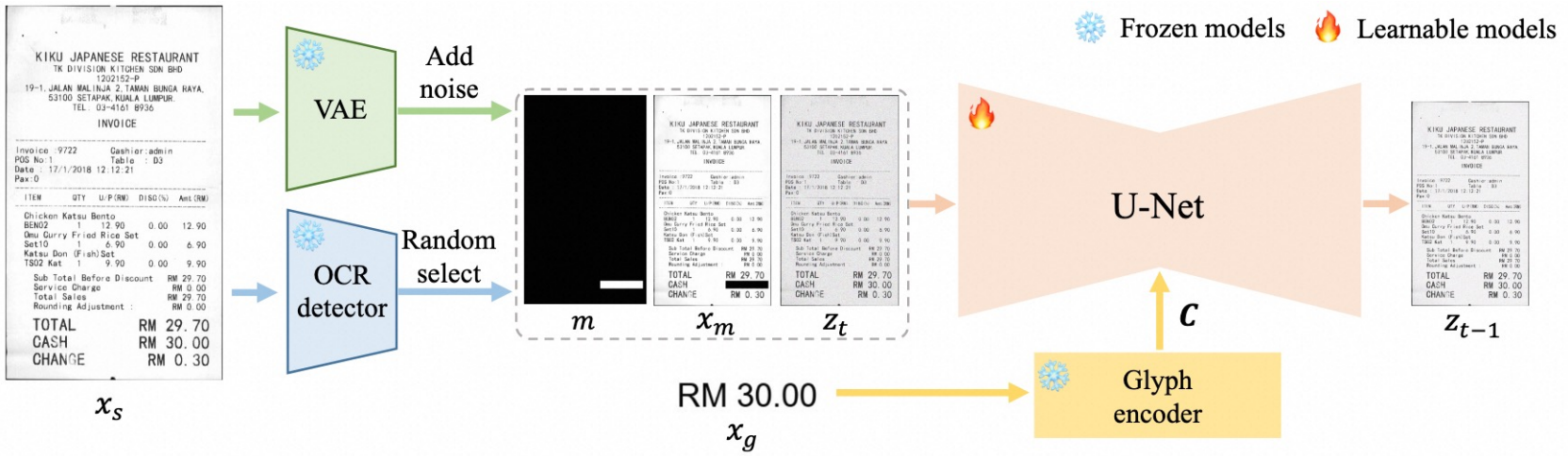
DiffUTE: Universal Text Editing Diffusion Model

NIPS'23 Ant Group

- Task：Text Inpainting
- Innovation：
  - First diffusion-based method for Chinese text inpainting
  - A progressive training strategy for fine-tuning VAE
- Language：Chinese&English

DiffUTE: Universal Text Editing Diffusion Model

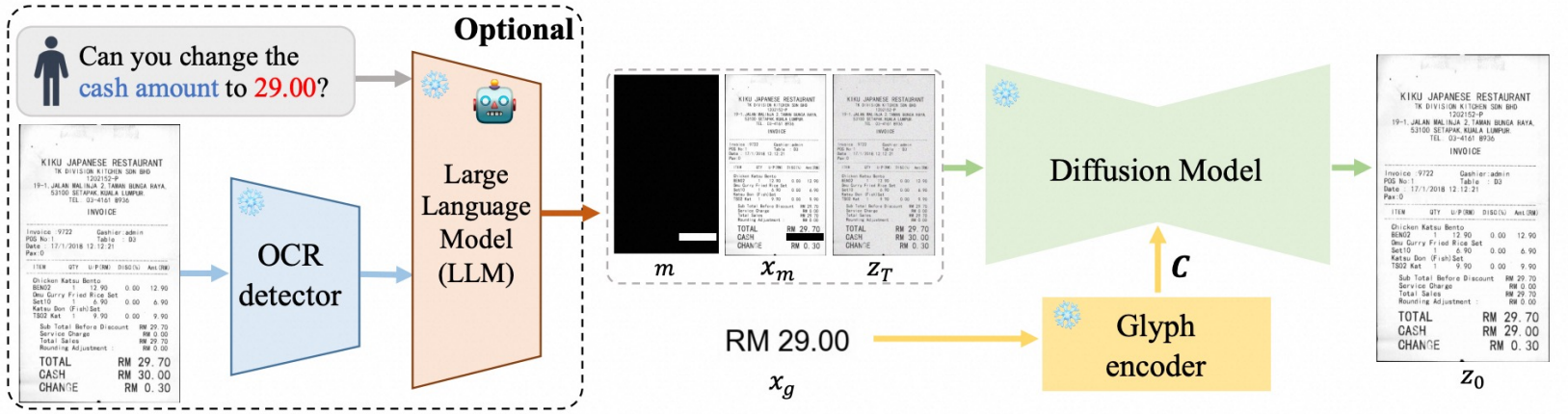
- Method: Training



- Input: mask, masked features, latent features
- Glyph encoder is from OCR, and size of output is 577*1024
- Random crop small sizes and Resize to train VAE in 3 stages

DiffUTE: Universal Text Editing Diffusion Model

- Method: Inference



- Interactive Scene Text Editing with LLM:
  - ChatGLM fine-tuned by OCR data
  - Users query LLM
  - LLM return the target text and its corresponding bounding box

DiffUTE: Universal Text Editing Diffusion Model

- Details
  - Dataset
    - Training set: 5M image from web-crawled data and publicly available text image datasets including: Xfund, PubLayNet, ICDAR series
    - Test set: 1000 from ArT, TextOCR, ICDAR13
  - Implement
    - SD: stabilityai/stable-diffusion-2-inpainting
    - Glyph encoder: TrOCR
    - VAE:

Module	Batch size	Crop Image Size	Iterations	Learning Rate
VAE	48	64 128 256 512	0–8w 8w–16w 16w–24w 24w–32w	5e-6
UNet	256	256	0–10w	1e-5

    - Metric: OCR accuracy and human evaluation of the correctness

GlyphControl: Glyph Conditional Control for Visual Text Generation

NIPS'23 MSRA

- Task：Text-to-Image
- Innovation：Utilize ControlNet
- Language：English

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".

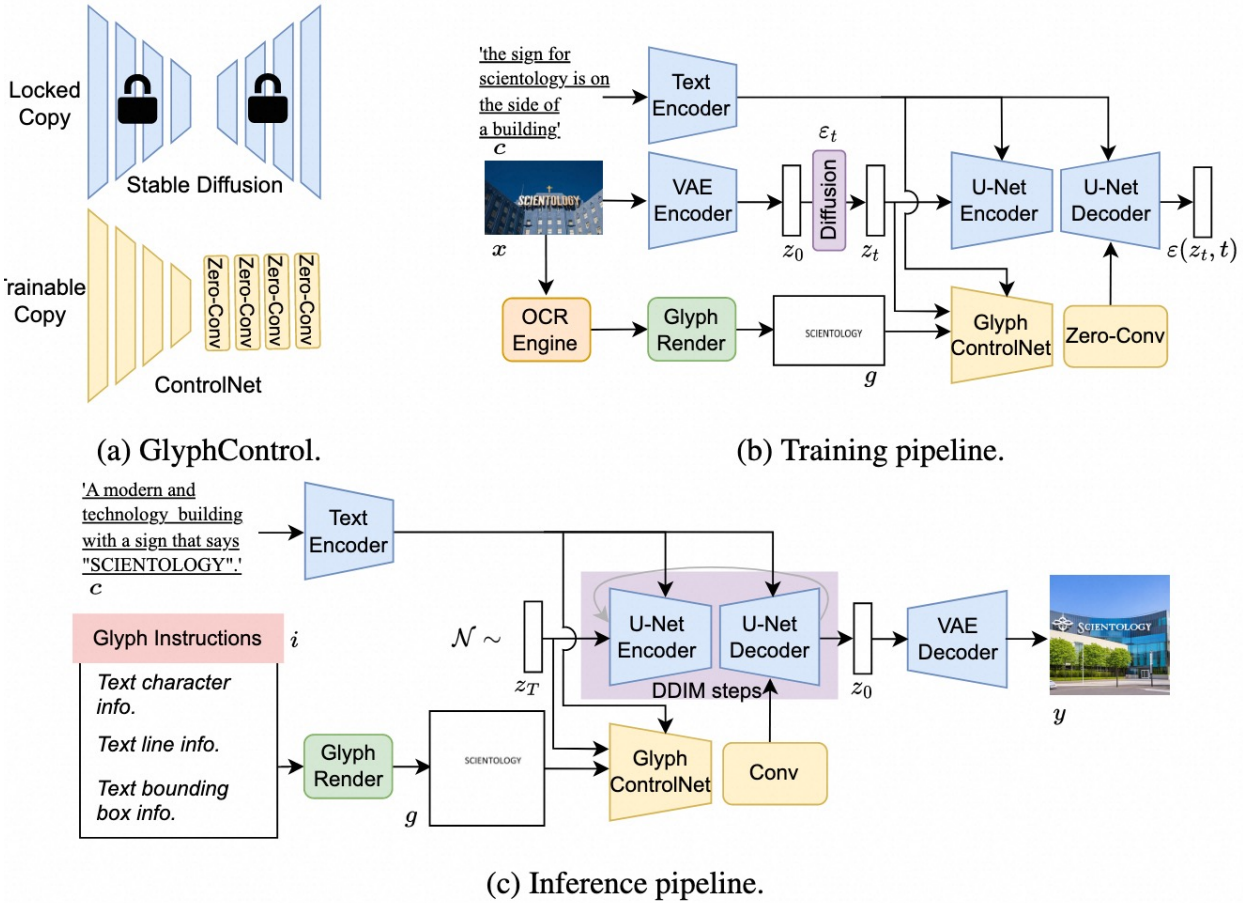
A menu of a fast food restaurant that contains "Sandwich Combo", "French Fries", and "Pepsi".

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.

GlyphControl: Glyph Conditional Control for Visual Text Generation

- Method: Training and Inference
  - Glyph image as the input of ControlNet



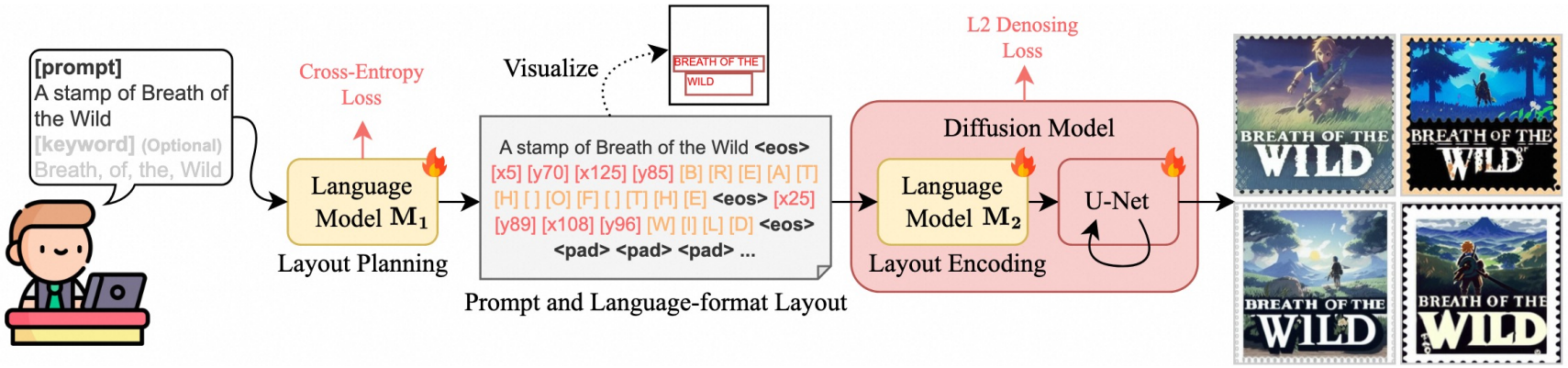
GlyphControl: Glyph Conditional Control for Visual Text Generation

- Detail
  - Dataset
    - LAION-Glyph(10M/1M/100K) from LAION-2B-en by PP-OCRv3.
    - aesthetic score higher than 4.5
    - remove images that have total OCR areas less than 5% of the whole image area or contain more than 5 bounding boxes
  - Evaluation Benchmark
    - SimpleBench: 'A sign that says "<word>". '
    - CreativeBench: 'Little panda holding a sign that says "<word>".'
  - Implement
    - SD: stable-diffusion 2.0-base model

TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering

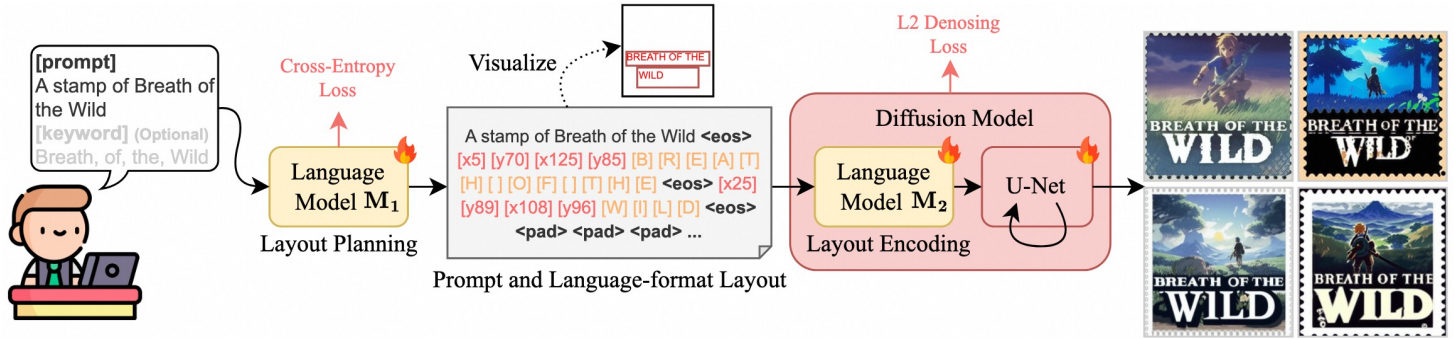
CVPR'24 Submission by MSRA

- Task：Text-to-Image & Text Inpainting
- Innovation：
  - LLM for Layout Gneration
  - Layout as Text Prompt
- Language：English



TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering

- Method:



  - Layout Planning (LLM)
    - Model: finetune the vicuna-7b-v1.5 at 5k sample from MARIO-10M
    - Language-format Layout : “textline(keywords) x0, y0, x1, y1”
  - Language model for layout encoding (CLIP)
    - hybrid-granularity tokenization
      - BPE for the whole prompt
      - character tokens for the keywords: WILD -> "[W]", "[I]", "[L]", "[D]”
      - coordinate tokens for the postions: (50, 70) -> “[x5]”, “[y70]”
      - Padding with “⟨eos⟩” to the maximum length L

TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering

- Details
  - Dataset:
    - MARIO-10M and MARIO-Eval
  - Implement
    - SD: runwayml/stable-diffusion-v1-5
    - LLM: finetune the vicuna-7b-v1.5 based on the FastChat framework
    - Maximum length L is 128

How to represent the position of text lines?

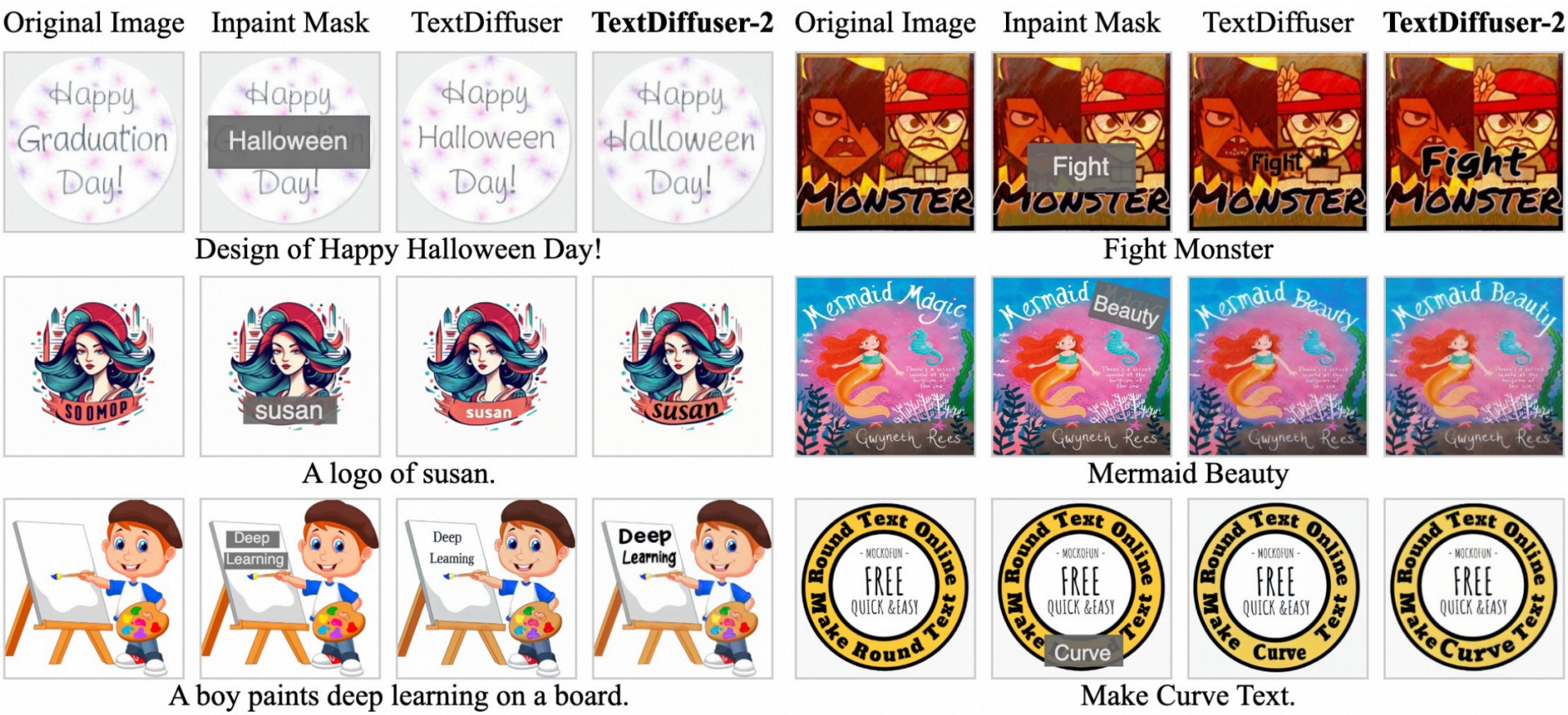
#Data	Acc↑	Pre↑	Rec↑	F↑	IOU↓
0k-2shot	49.65	84.18	69.69	76.25	19.69
2.5k	61.10	82.20	85.18	83.67	**3.21**
5k	**64.85**	84.98	**86.38**	**85.67**	3.25
10k	**64.85**	84.38	86.23	85.29	4.27
50k	63.72	**85.32**	85.78	85.55	3.68
100k	62.87	85.26	85.98	85.62	4.31

How to represent the position of text lines?

Representation	Acc↑	Pre↑	Rec↑	F↑
Center (Char)	35.19	61.75	62.71	62.23
LT (Char)	28.32	54.94	55.64	55.29
LT+RB (Subword)	15.48	41.74	42.53	42.13
LT+RB (Char)	**57.58**	**74.02**	**76.14**	**75.06**

TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering

- Visualizations



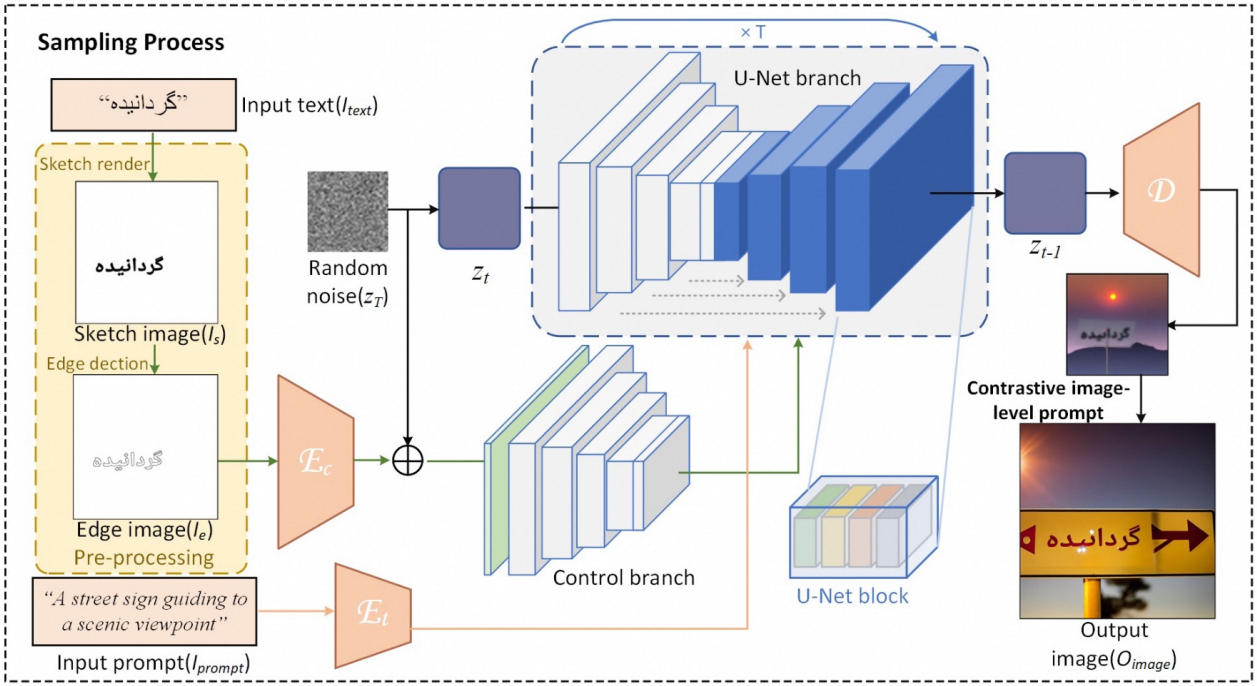
Brush Your Text: Synthesize Any Scene Text on Images via Diffusion Model

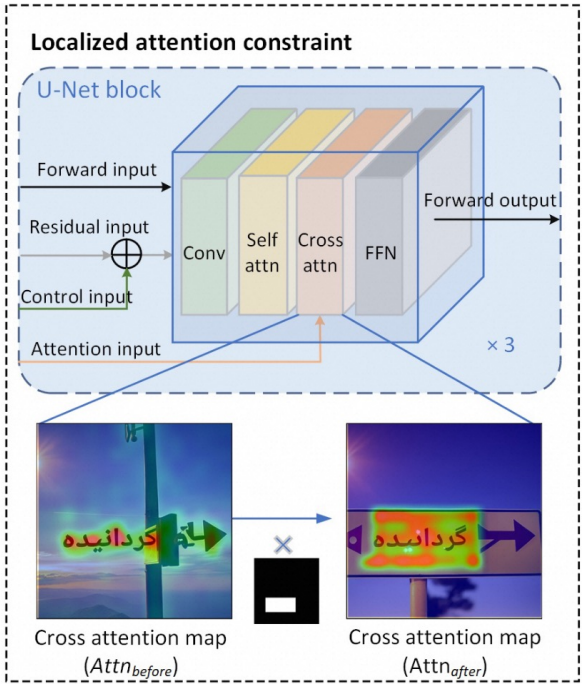
AAAI'24 SH AI Lab

- Task：Text-to-Image
- Innovation：
  - Training-Free
  - Any language
- Language：Any

Brush Your Text: Synthesize Any Scene Text on Images via Diffusion Model

- Method

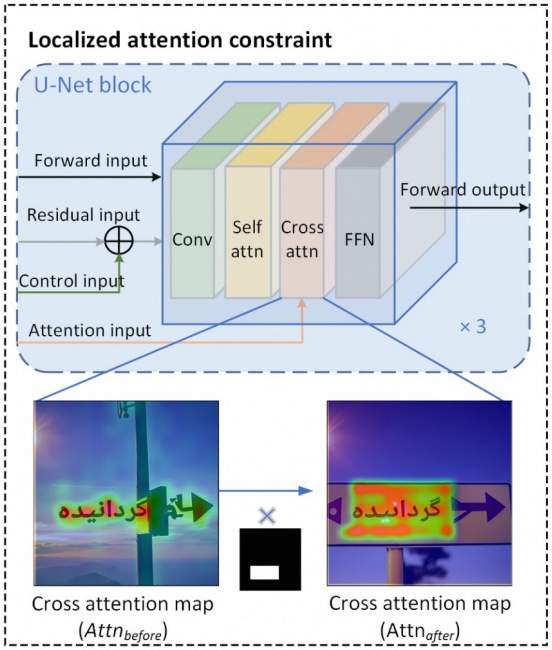


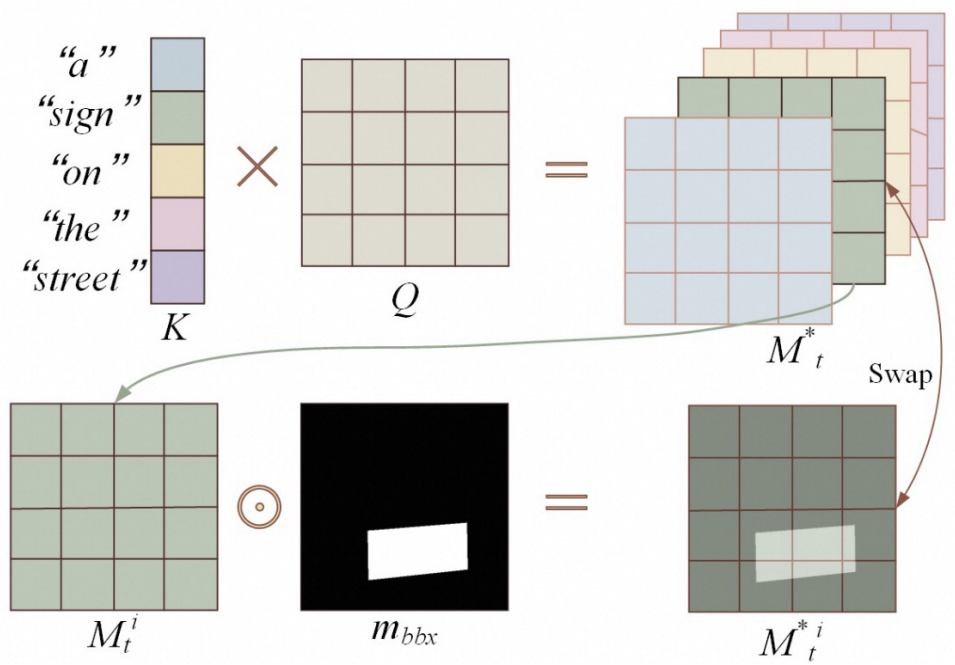


- The canny edge of glyph image is the input of ControlNet
- Localized Attention Constraint
- Contrastive Image-level Prompts

Brush Your Text: Synthesize Any Scene Text on Images via Diffusion Model

- Method: Localized Attention Constraint





- Like prompt2prompt

$$M_t^* = \{\lambda \times M_t^i \odot m_{bbx} | \forall i \in I\}$$

Brush Your Text: Synthesize Any Scene Text on Images via Diffusion Model

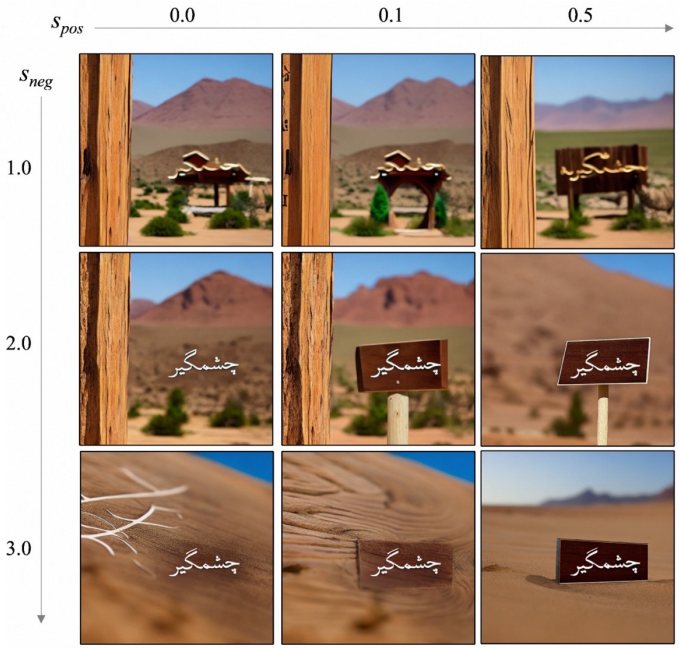
- Contrastive Image-level Prompts

$$z_{t-1} = \tilde{\epsilon}(z_t, I_e, I_{prompt})$$
$$= \epsilon(z_t, \varnothing, \varnothing) + s_{cfg}(\epsilon(z_t, \varnothing, I_{prompt}) - \epsilon(z_t, \varnothing, \varnothing))$$
$$+ s_{neg}(\epsilon^{'}(z_t, I_e, I_{prompt}) - \epsilon(z_t, \varnothing, I_{prompt})),$$
$$\epsilon^{'}(z_t, I_e, I_{prompt}) = \epsilon(z_t, I_e, I_{prompt})$$
$$+ s_{pos}(\epsilon(z_t, I_e^{'}, I_{prompt}) - \epsilon(z_t, I_e, I_{prompt})),$$

sketch image

original edge image incorporating the depiction of a bounding box

purely white



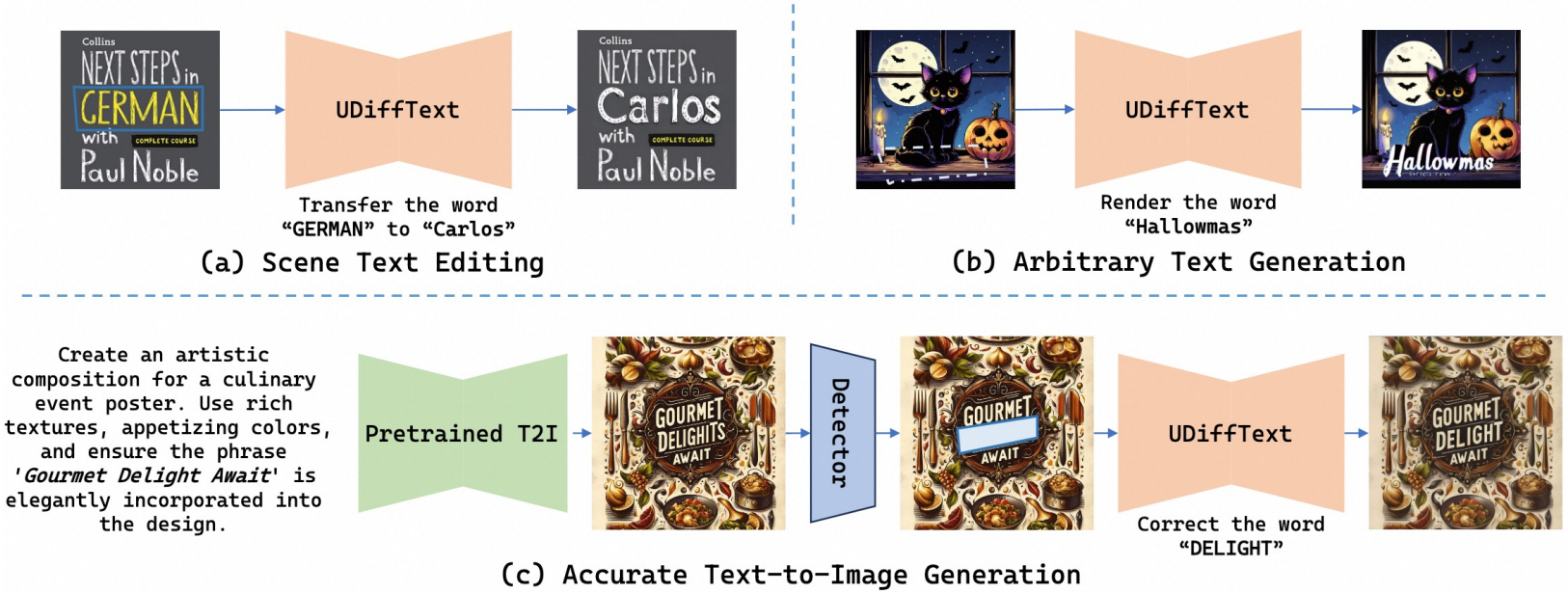
Brush Your Text: Synthesize Any Scene Text on Images via Diffusion Model

- Detail
  - SD: runwayml/stable-diffusion-v1-5
  - ContriNet: lllyasviel/sd-controlnet-canny
  - The wordlist for localized attention constraint includes "sign", "billboard", "label", "promotions", "notice", "marquee", "board", "blackboard", "slogan", "whiteboard", and "logo"

UDiffText: A Unified Framework for High-quality Text Synthesis in Arbitrary Images via Character-aware Diffusion Models

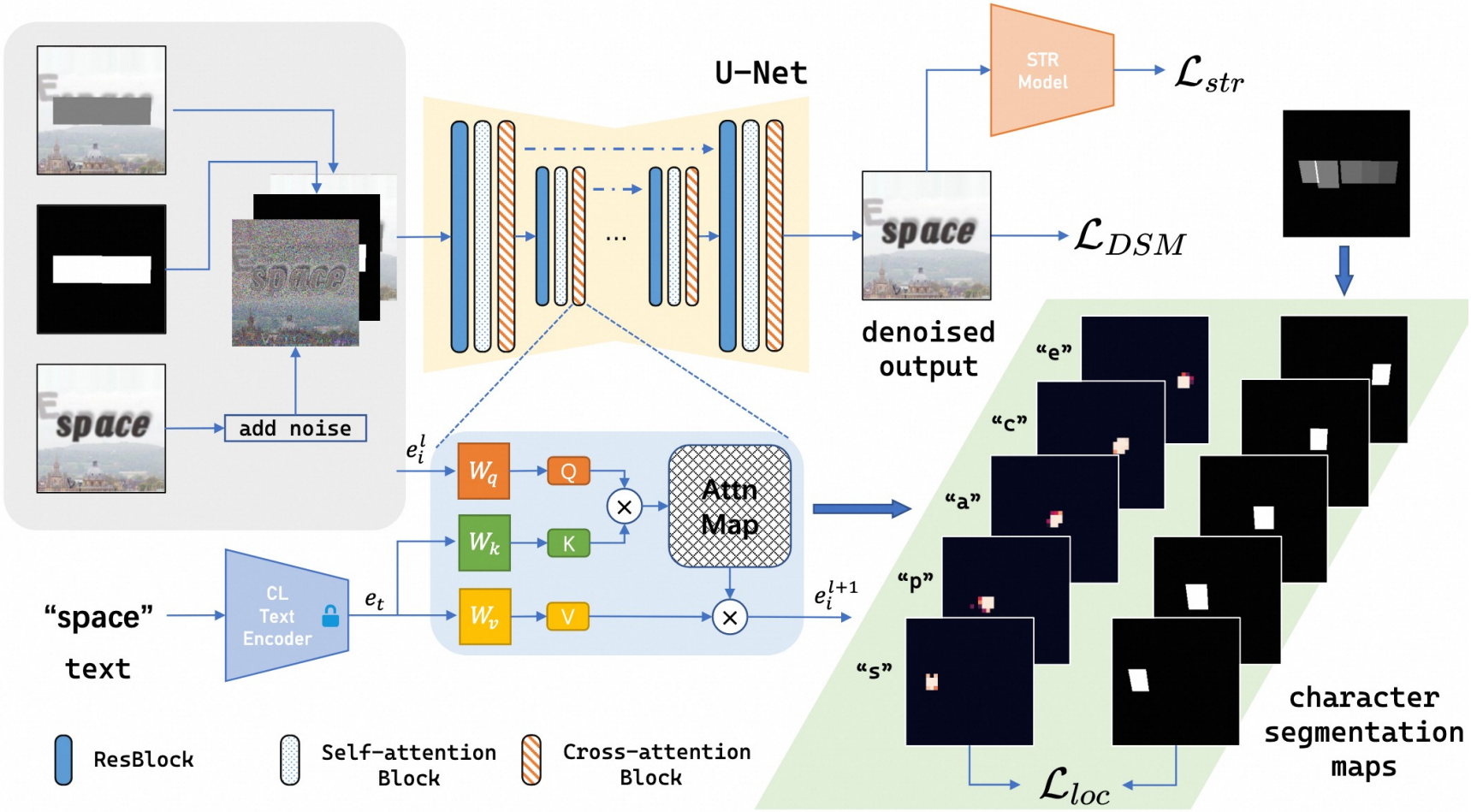
CVPR'24 Submission by PKU

- Task：Text Inpainting
- Innovation：
  - a character-level text encoder
  - a segmentation map supervision for the attention map
- Language：English



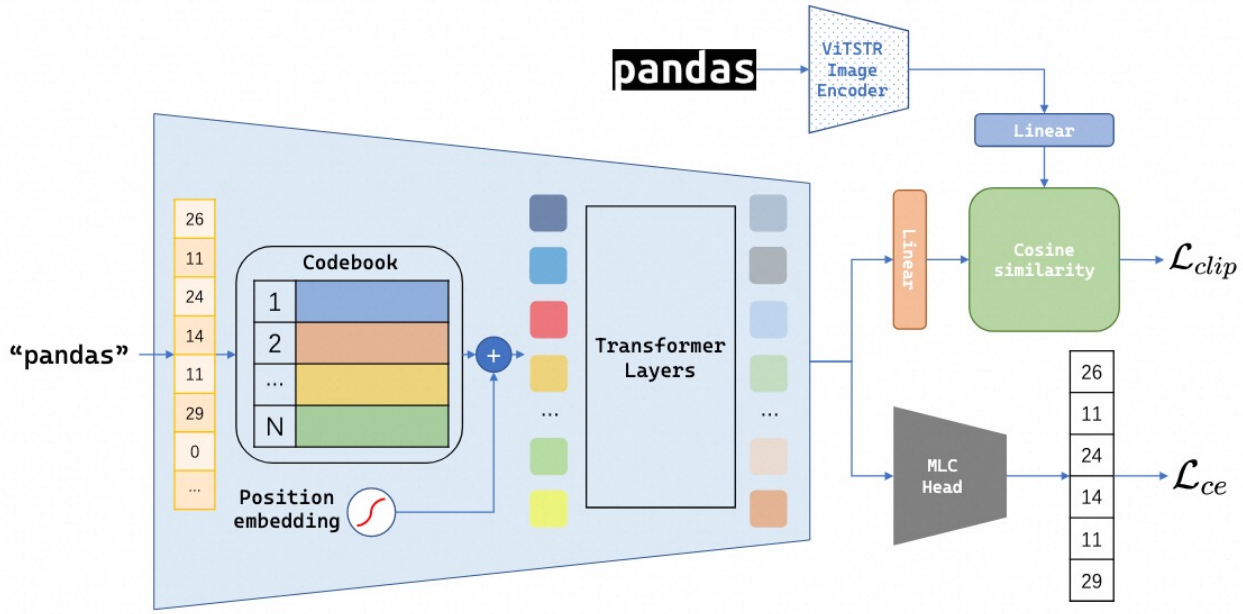
UDiffText: A Unified Framework for High-quality Text Synthesis in Arbitrary Images via Character-aware Diffusion Models

- Method



UDiffText: A Unified Framework for High-quality Text Synthesis in Arbitrary Images via Character-aware Diffusion Models

- Method: Character-level Text Encoder



$$\mathbf{e}_{text} = \mathcal{E}_{text}(\mathcal{T}), \quad \mathbf{e}_{image} = \mathcal{E}_{image}(\mathcal{I}_{\mathcal{T}}), \tag{1}$$

$$\mathcal{L}_{clip} = -\boldsymbol{CS}(W_t \mathbf{e}_{text}, W_i \mathbf{e}_{image}), \tag{2}$$

$$\mathcal{L}_{ce} = \boldsymbol{CE}(\mathcal{H}_{MLC}(\mathbf{e}_{text}), Ids), \tag{3}$$

$$\mathcal{L} = \mathcal{L}_{clip} + \lambda_{ce}\mathcal{L}_{ce}. \tag{4}$$

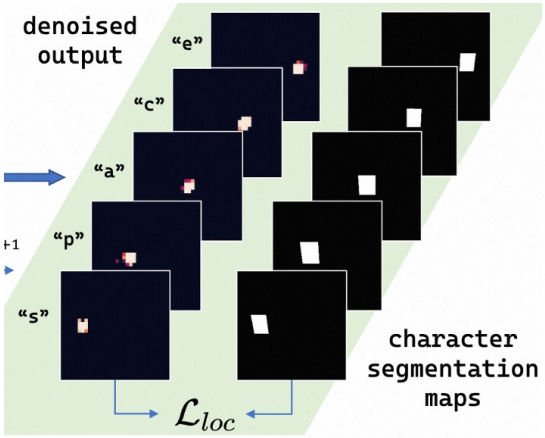
UDiffText: A Unified Framework for High-quality Text Synthesis in Arbitrary Images via Character-aware Diffusion Models

- Method: Local Attention Loss
  - Calculate the attention map

$$\mathcal{Q}_i = W_i^Q \mathbf{e}_{image},\ \mathcal{K}_i = W_i^K \mathbf{e}_{text},\ \mathcal{V}_i = W_i^V \mathbf{e}_{text}, \quad (6)$$

$$\mathcal{A}_i = \boldsymbol{softmax}\left(\mathcal{Q}_i \mathcal{K}_i^T / \sqrt{d}\right) \mathcal{V}_i. \quad (7)$$

  - Compute the loss attention loss



$$\mathcal{L}_{loc} = \frac{1}{C}\sum_{i=1}^{C}\left\{\frac{1}{L}\sum_{j=1}^{L}\left(\boldsymbol{max}\left(\mathbb{G}\left(\mathbf{A}_i^j\right)\odot\left(\boldsymbol{J}-\mathbf{S}^j\right)\right)\right) - \frac{1}{L}\sum_{j=1}^{L}\left(\boldsymbol{max}\left(\mathbb{G}\left(\mathbf{A}_i^j\right)\odot\mathbf{S}^j\right)\right)\right\},$$

UDiffText: A Unified Framework for High-quality Text Synthesis in Arbitrary Images via Character-aware Diffusion Models

- Dataset
  - Training with the character-level segmentation maps:
    - SynthText in the Wild: 800K images
    - LAION-OCR: 9M from TextPainter(The authors of TextPainter trained a character-level segmentation model)
  - Evaluation
    - ICDAR13 (233)，TextSeg (340)，LAION-OCR Evaluation
- Implement
  - SD: Stable Diffusion (v2.0) inpainting version
  - cross-attention layers) are updated during training
  - on the SynthText dataset for 100k steps and then on the LAION-OCR dataset for an additional 100k steps.

ANYTEXT: MULTILINGUAL VISUAL TEXT GENERA-TION AND EDITING

ICLR'24 Submission by Alibaba

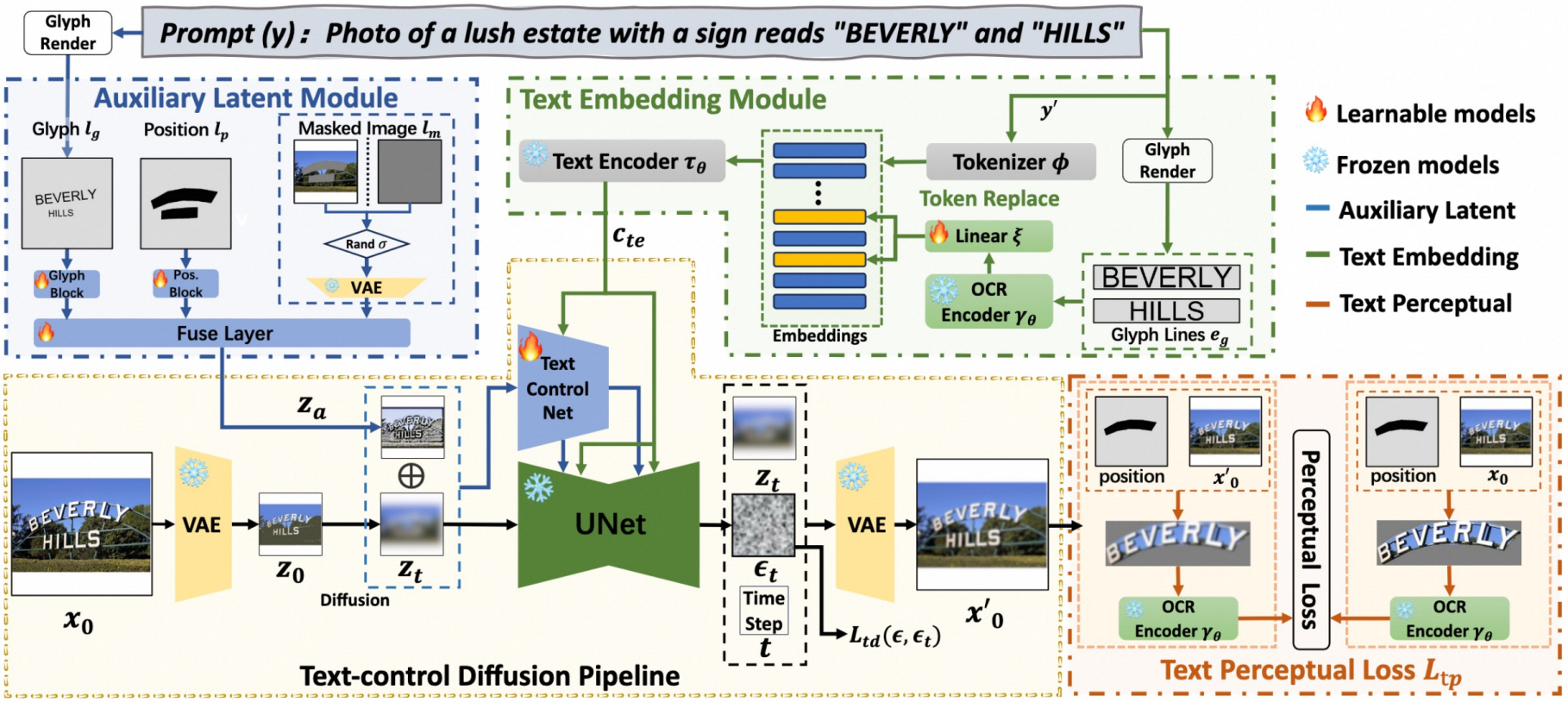
- Task：Text-to-Imgae, Text Inpainting
- Innovation： almighty
- Language：Multi-Lingual

Table 1: Comparison of AnyText with other competitors based on functionality.

Functionality	Multi-line	Deformed regions	Multi-lingual	Text editing	Plug-and-play
GlyphDraw	×	×	×	×	×
TextDiffuser	✓	×	×	✓	×
GlyphControl	✓	×	×	×	✓
AnyText	✓	✓	✓	✓	✓

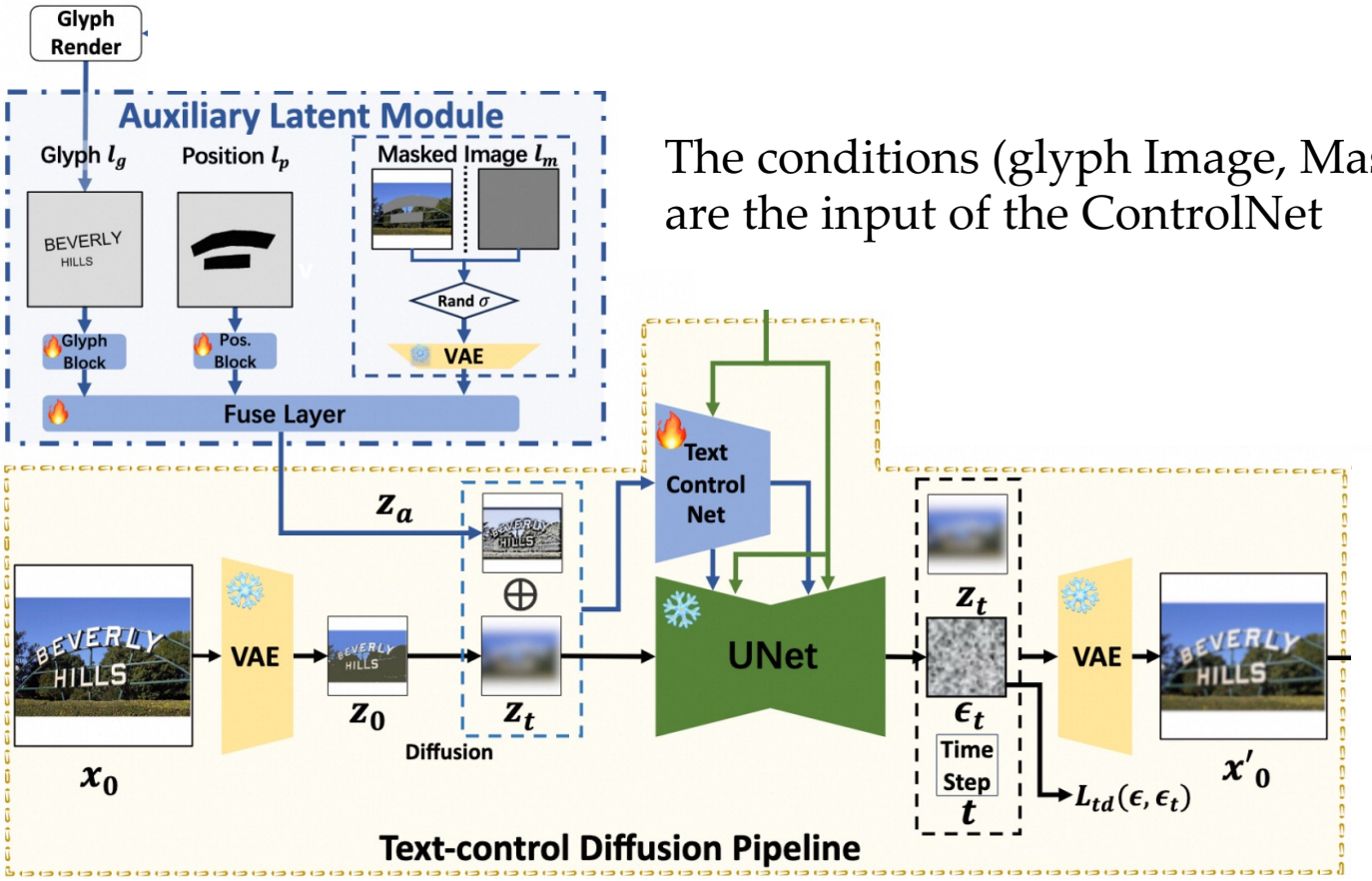
ANYTEXT: MULTILINGUAL VISUAL TEXT GENERA-TION AND EDITING

- Method



ANYTEXT: MULTILINGUAL VISUAL TEXT GENERA-TION AND EDITING

- Auxiliary latent module

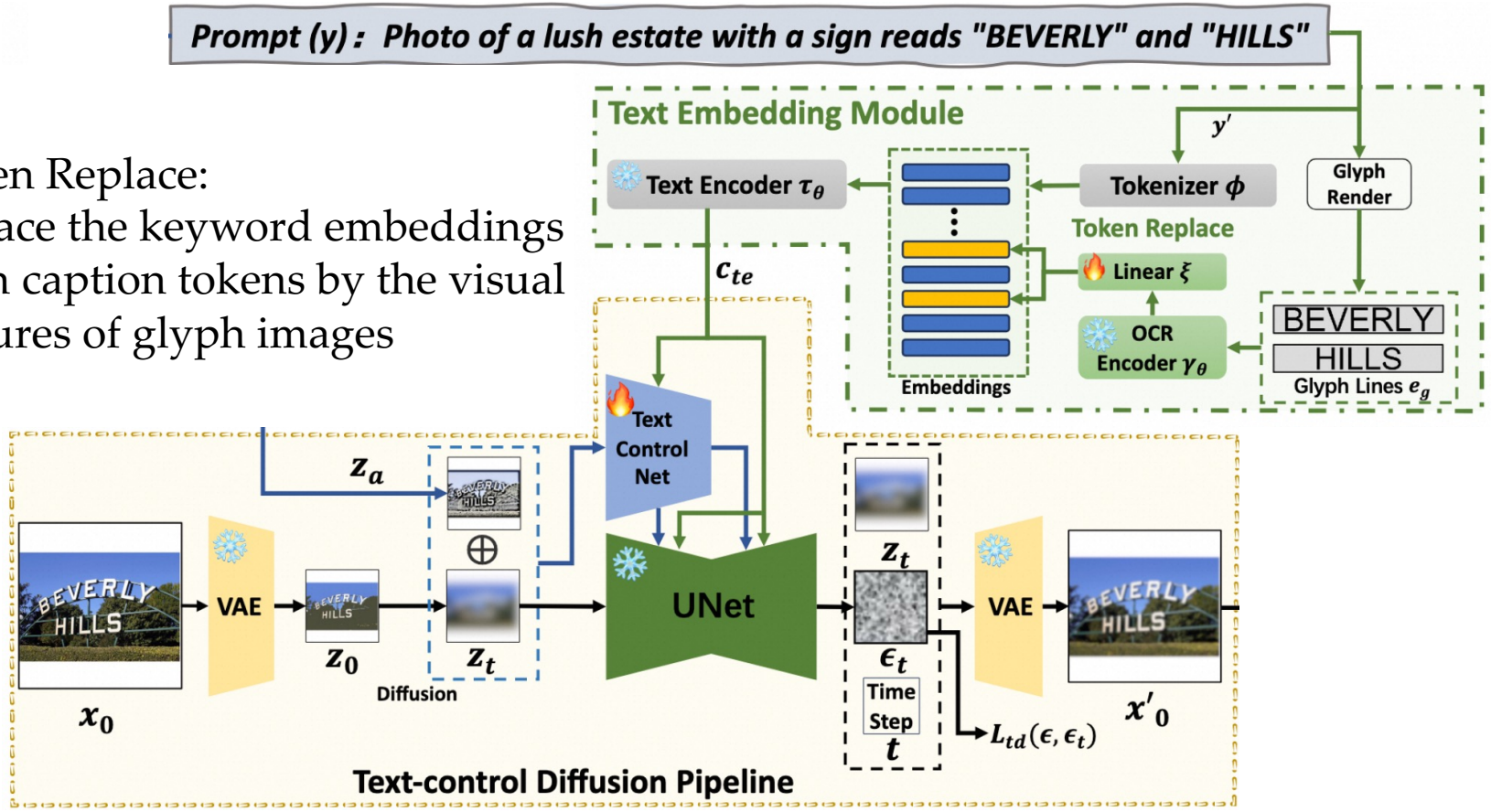


The conditions (glyph Image, Mask, Masked feature) are the input of the ControlNet

ANYTEXT: MULTILINGUAL VISUAL TEXT GENERA-TION AND EDITING

- Text Embedding Module

Token Replace:
replace the keyword embeddings from caption tokens by the visual features of glyph images

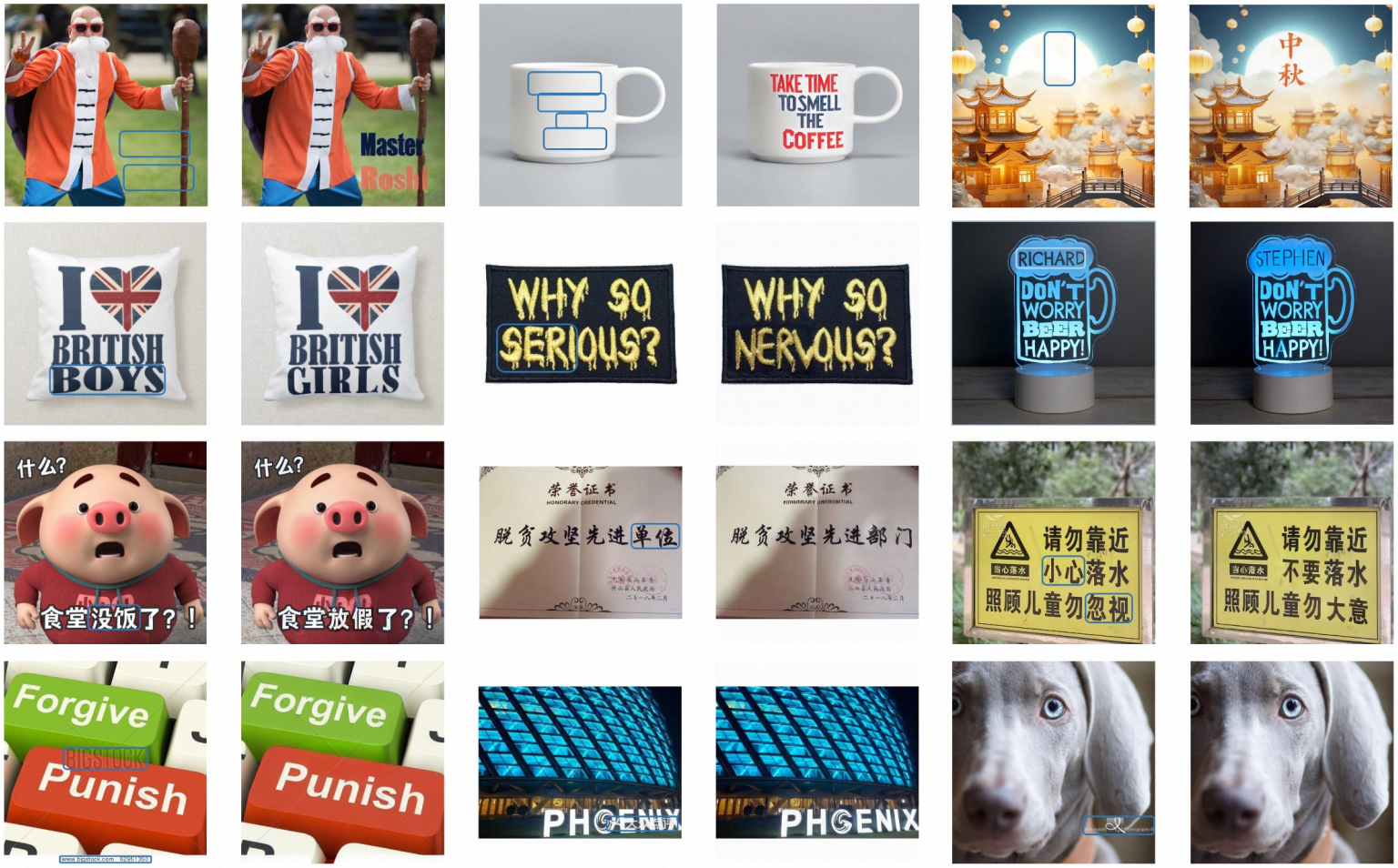


ANYTEXT: MULTILINGUAL VISUAL TEXT GENERA-TION AND EDITING

- Details
  - Dataset
    - AnyWord-3M
      - NoahWukong, LAION-400M, other OCR datasets
      - Annotation: PP-OCRv3, BLIP-2
      - 1.6M Chinese, 1.39M English, 10k other languages
    - AnyText-benchmark
      - 1000 image from NoahWukong and LAION-400M
  - Implement
    - SD: runwayml/stable-diffusion-v1-5 & lllyasviel/ControlNet

ANYTEXT: MULTILINGUAL VISUAL TEXT GENERA-TION AND EDITING

- Visualizations

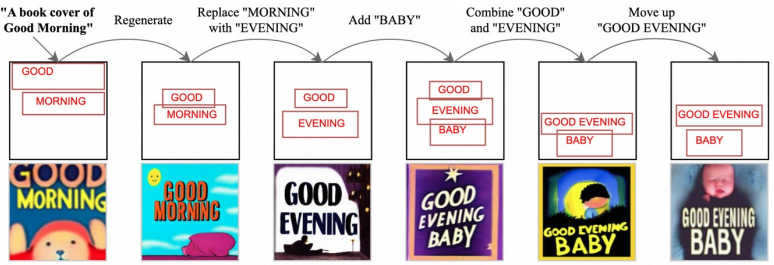


Limitations

- Small size font
  - From GlyphControl

Font Size	Acc(%)↑	Âcc(%)↑	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

- Consistency



- Multi-Lingual(eg.Chinese)
- Multi-Size(eg. our 513*750 poster)
- User controlled and editable

- AnyText

- AnyText

- AnyText

数据集

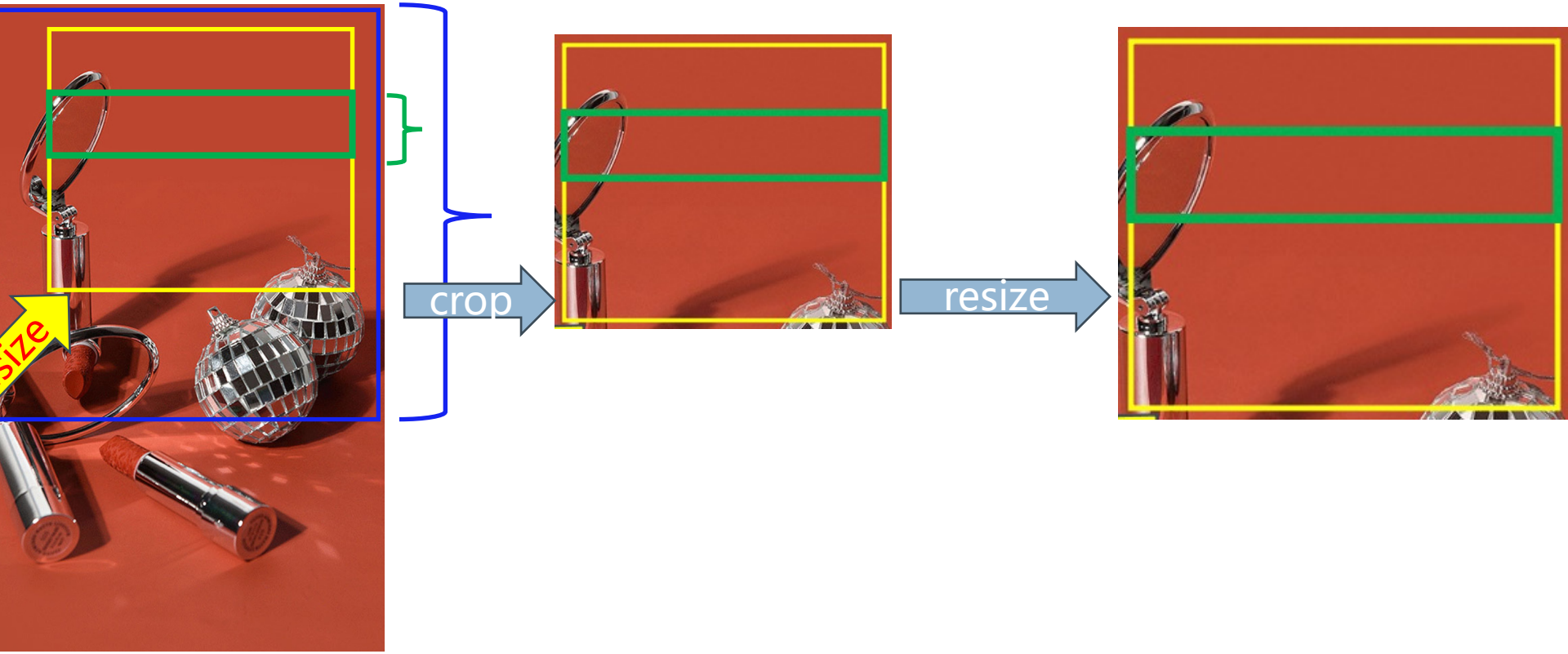
名称	类型	地址	备注
CDLA: A Chinese document layout analysis dataset	doc	https://github.com/buptlihang/CDLA	cn
XFUND: A Multilingual Form Understanding Benchmark	doc	https://github.com/docanalysis/XFUND	ml
Publaynet: largest dataset ever for document layout analysis.	doc	https://github.com/ibm-aur-nlp/PubLayNet	
ICDAR 2019 Robust Reading Challenge on Reading Chinese Text on Signboard	scene	https://rrc.cvc.uab.es/?ch=12&com=introduction	cn
ICDAR2019 Robust Reading Challenge on Multi-lingual Scene Text Detection and Recognition	scene	https://rrc.cvc.uab.es/?ch=15	ml
ICDAR 2015 Competition on Robust Reading	scene	https://rrc.cvc.uab.es/files/Robust-Reading-Competition-Karatzas.pdf	
Icdar2019 robust reading challenge on large-scale street view text with partial labeling	scene	https://rrc.cvc.uab.es/?ch=16	
ICDAR2017 Competition on Reading Chinese Text in the Wild	scene	https://rctw.vlrlab.net/dataset	cn
Icpr 2018 challenge on multi-type web images	web	https://tianchi.aliyun.com/dataset/137084	
A large-scale scene text dataset based on mscoco	scene	https://bgshih.github.io/cocotext/#h2-explorer	en
Icdar2019 robust reading challenge on arbitrary-shaped text	scene	https://rrc.cvc.uab.es/?ch=14	
Wukong: 100 Million Large-scale Chinese Cross-modal Pre-training Dataset and A Foundation Framework	all	https://wukong-dataset.github.io/wukong-dataset/	cn
LAION-400M: open dataset of clip-filtered 400 million image-text pairs	all	https://laion.ai/blog/laion-400-open-dataset/	en

数据集

- 文档
  - Publaynet: largest dataset ever for document layout analysis.
    - https://github.com/ibm-aur-nlp/PubLayNet

PubLayNet is a large dataset of document images, of which the layout is annotated with both bounding boxes and polygonal segmentations. The source of the documents is PubMed Central Open Access Subset (commercial use collection). The annotations are automatically generated by matching the PDF format and the XML format of the articles in the PubMed Central Open Access Subset. More details are available in our paper "PubLayNet: largest dataset ever for document layout analysis.".

数据集



Thank for your attention !