HDLayout: Hierarchical and Directional Layout Planning for Arbitrary Shaped Visual Text Generation

Tonghui Feng[1] Chunsheng Yan[2] Qianru Wang[1] Jiangtao Cui[1] Xiaotian Qiao[1, 2 *]

[1]School of Computer Science and Technology, Xidian University, China [2]Guangzhou Institute of Technology, Xidian University, China

AAAI-25 / IAAI-25 / EAAI-25
FEBRUARY 25 – MARCH 4, 2025 | PHILADELPHIA, USA

Introduction

Problem

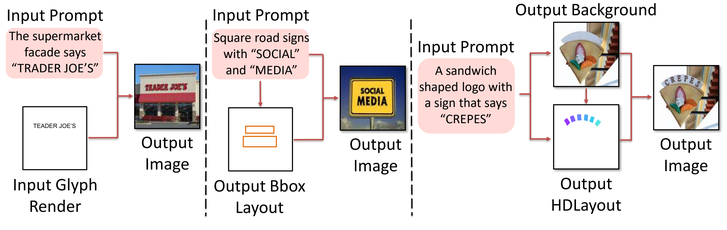
- Arbitrary Shaped Visual Text Generation
  - Real-world visual text often appears in various shapes and layouts (e.g., curved, multi-oriented).

Motivation

- Multi-Granularity Perception
  - Text in scenes inherently exhibits correlations across different granularities (character, word, line).
- Flexibility and User Accessibility
  - Design a framework that allows users to generate diverse and coherent visual text directly from text prompts.

Related Work

- GlyphControl[3] (left), TextDiffuser[4] (middle), Ours (right).

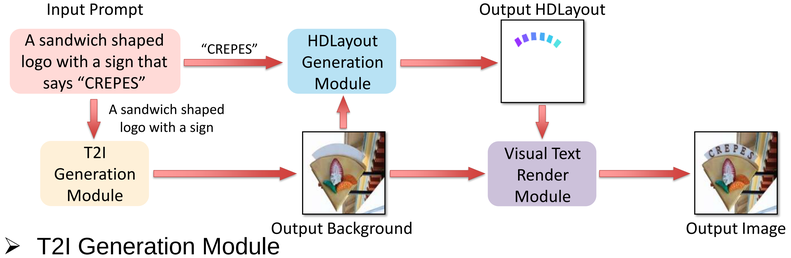


Contribution

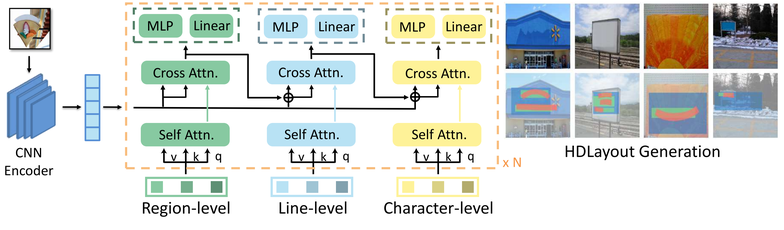
- A hierarchical and directional layout representation to model the unique characteristics of visual text.
- A new separation and composition framework to generate both textual and visual information.
- A new HDLayout3k dataset with diverse and arbitrarily shaped text layout.

Approach

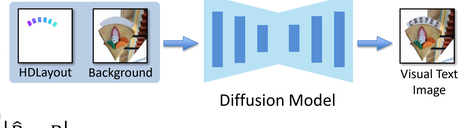
Visual Text Generation Architecture



- T2I Generation Module
  - Separating description and keywords, and generating initial visual content.
- HDLayout Generation Module
  - Predicting fine-grained visual text structures.



- Text Rendering Module
  - Visual text image generation.



Training

- Bezier Point Loss $L_{L1} = \frac{1}{N}\sum_{i=1}^{N}\left|\hat{B} - B\right|$
- Bounding Box Loss $L_{bbox} = c_1 L_{L1} + c_2 L_{GIoU} + c_3 L_{ol}$

$$L_{ol} = \frac{2}{M(M-1)}\sum_{i=1}^{M}\sum_{j=i+1}^{M}\mathrm{IoU}(B_i, B_j) \qquad L_{GIoU} = \frac{1}{N}\sum_{i=1}^{N}\left(1 - \mathrm{GIoU}(\hat{B}, B)\right)$$

- Confidence Loss $L_{conf} = -p \cdot \log(q)$

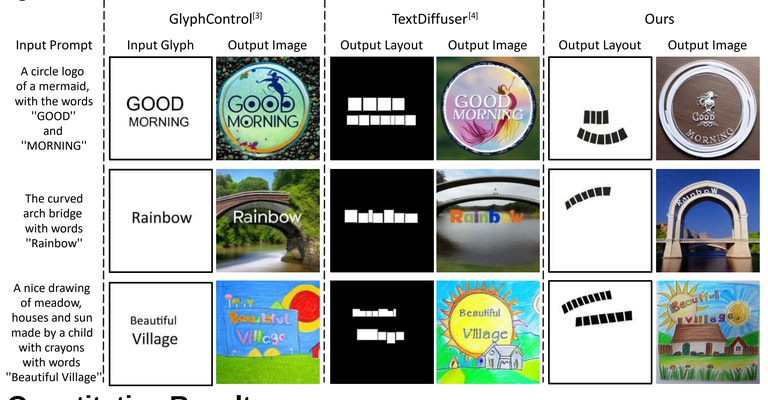
Dataset

HDLayout3k

- We construct the HDLayout3k dataset, which consists of 2,749 training samples and 813 test samples.

Evaluation

Qualitative Results



Quantitative Results

Metric	SD-XL[1]	ControlNet[2]	GlyphControl[3]	TextDiffuser[4]	Ours
Image FID ↓	102.128	89.802	83.402	82.068	**78.027**
Layout FID ↓	-	-	-	27.135	**12.343**

References

[1] Podell, D , et al. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In ICLR 2024.
[2] Zhang, L , et al. Adding conditional control to text-to-image diffusion models. In ICCV 2023.
[3] Yang, Y , et al. GlyphControl: Glyph Conditional Control for Visual Text Generation. In NeurIPS 2024.
[4] Chen, J , et al. Textdiffuser: Diffusion models as text painters. In NeurIPS 2024.