TextLap: Customizing Language Models for Text-to-Layout Planning

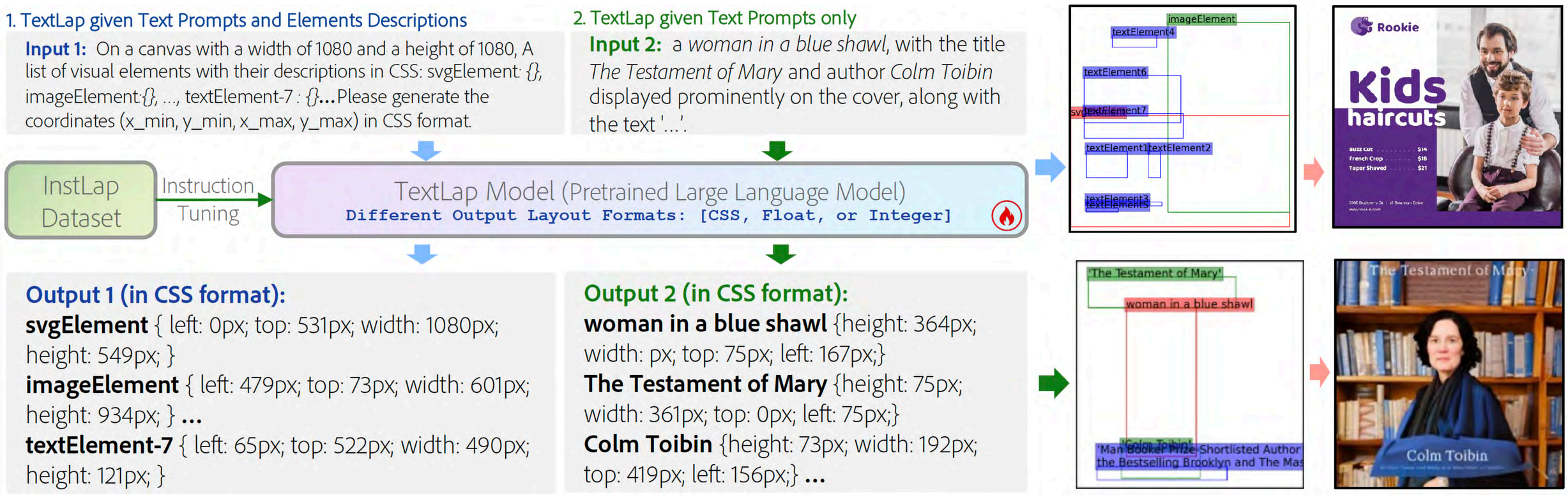
Jian Chen[1], Ruiyi Zhang[2], Yufan Zhou[2], Jennifer Healey[2],
Jiuxiang Gu[2], Zhiqiang Xu[3], Changyou Chen[1]
1. University at Buffalo, 2. Adobe Research, 3. MBZUAI

Introduction

We customize large language models (LLMs) for text-guided layout planning, introducing a new dataset, InsLap, derived from natural image and graphic design sources. Using InsLap, we train our TextLap model, which empowers users to create and adjust layouts with simple text instructions, enhancing the efficiency and accessibility of graphic design tasks.

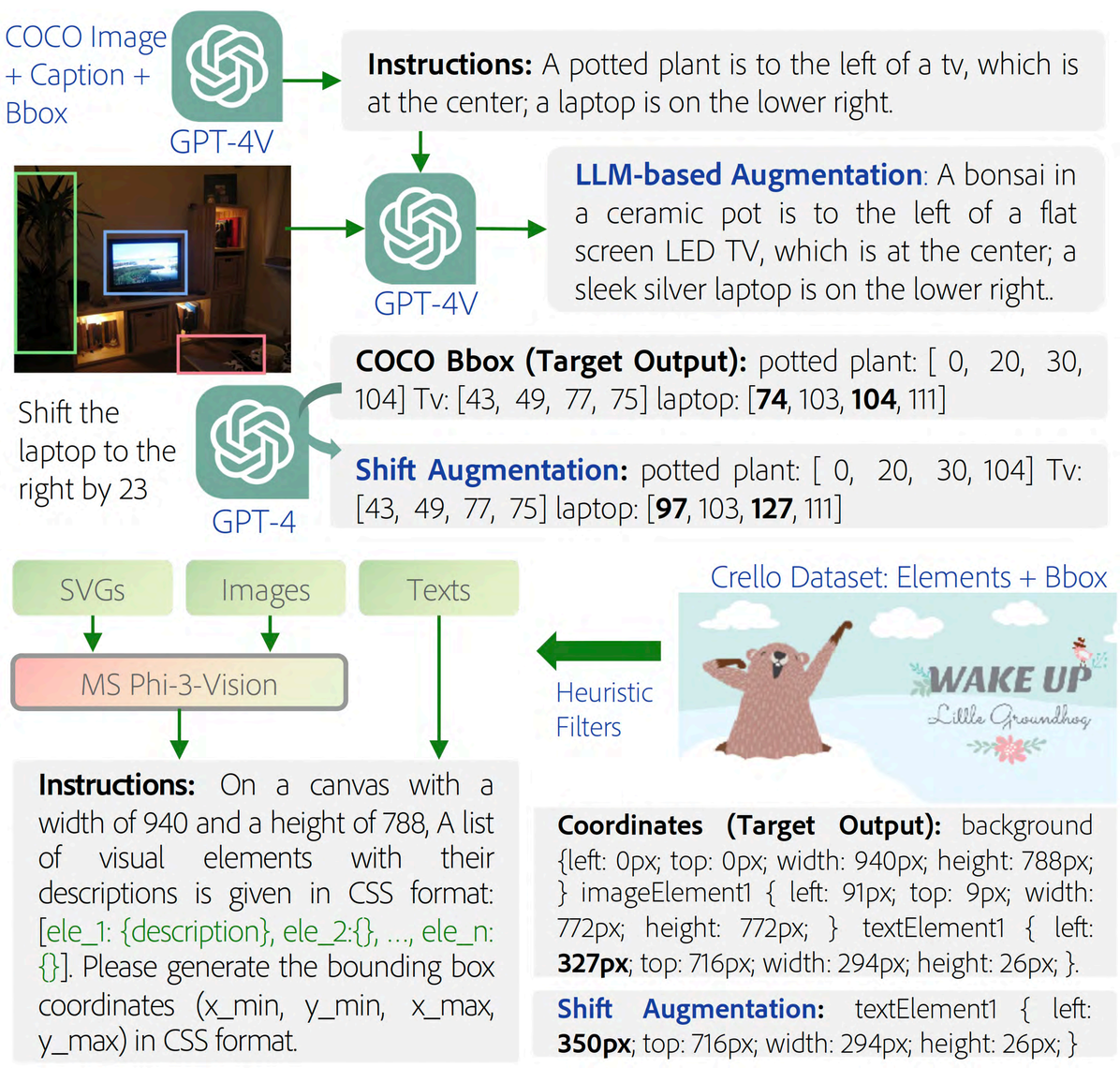
TextLap Model

TextLap is a 7B parameter language model that takes only text input to generate structured layouts. It learns layout distributions through supervised fine-tuning with text-based coordinate formats, including integer, float, CSS, and JSON. The generated layouts can serve as an intermediate step to enhance spatial accuracy in text-to-image tasks or to automatically arrange graphic elements.



InstLap Dataset

- **Image Layouts for Visual Objects and Text**
  The InstLap dataset includes text-to-layout pairs derived from MS-COCO for visual objects, combined with text from the TextDiffuser2 training data, sourced from the MARIO-10M dataset. This subset standardizes and filters images to retain only essential objects, ensuring clarity and simplifying layout complexity. Text descriptions and spatial relationships are generated using language models, aligning bounding boxes with captions to enhance LLMs' understanding of 2D layouts.
- **Graphic Layouts from Crello**
  InstLap also incorporates graphic layouts from the Crello dataset for applications like posters and advertisements. This subset filters out overly complex designs and merges small, non-essential elements into the background. Each layout is annotated with captions and spatial instructions, enabling LLMs to generate content-aware and visually coherent graphic layouts.



Experiments

We tested TextLap with different coordinate formats—integer, float, CSS, and JSON—and compared its performance to GPT-4 and TextDiffuser2. Results show that CSS and JSON (float) formats performed best, thanks to the model's pretrained coding abilities for precise layout generation. Additionally, using generated layouts with a layout-to-image model improved GPT-4o preference scores over Stable Diffusion 1.5 (a text-to-image model) while maintaining a similar FID, demonstrating the effectiveness of TextLap's structured layouts in enhancing spatial accuracy.

Methods	MaxIoU (suc) ↑	MaxIoU ↑	Fail % ↓
GPT-4	0.231	0.206	10.778
GPT-4 (rCSS)	0.252	0.241	4.192
TextDiffuser-2	0.166	0.165	0.80
TextLap-Float	0.209	0.096	54.09
TextLap-CSS	0.211	0.211	0.00

Results on text layout generation on the InstLap-Bench.

	MaxIoU↑	Precision↑	Recall↑	F-score↑
GPT-4 (CSS)	0.190	0.874	0.257	0.364
GPT-4 (rCSS)	0.209	0.934,	0.296,	0.406
GPT-4 (Float)	0.457	0.980	0.980	0.980
GPT-4 (rFloat)	0.440	0.984	0.984	0.984
TextLap-CSS	0.407	0.998	0.986	0.990
TextLap-Float	0.535	1.000	1.000	1.000

Results on automatic graphic design on Crello.

Method	Image FID	GPT-Preference
SD 1.5	58.09	-
InstDiff (TextLap)	58.92	72%
InstDiff (True Layout)	58.39	76%

Text-to-image generation results on the COCO dataset. Layout-guided models are compared to the layout-free baseline, Stable Diffusion (SD) 1.5.