LayoutDETR: Detection Transformer Is a Good Multimodal Layout Designer

Ning Yu Chia-Chih Chen Zeyuan Chen Rui Meng
Gang Wu Paul Josel Juan Carlos Niebles Caiming Xiong Ran Xu
Salesforce Research

https://ningyu1991.github.io/projects/LayoutDETR.html

Motivations

- Graphic designs are at the foundation of communication between marketers and target audience.
- Graphic designs require designers' thoughtful understanding of multimodal inputs:
  - Background images
  - Multiple foreground texts
  - Multiple foreground product images
- Graphic designs require reasonable and aesthetically appealing compositions.
- Manual graphic designs are skill-demanding, time-consuming, and not scalable.

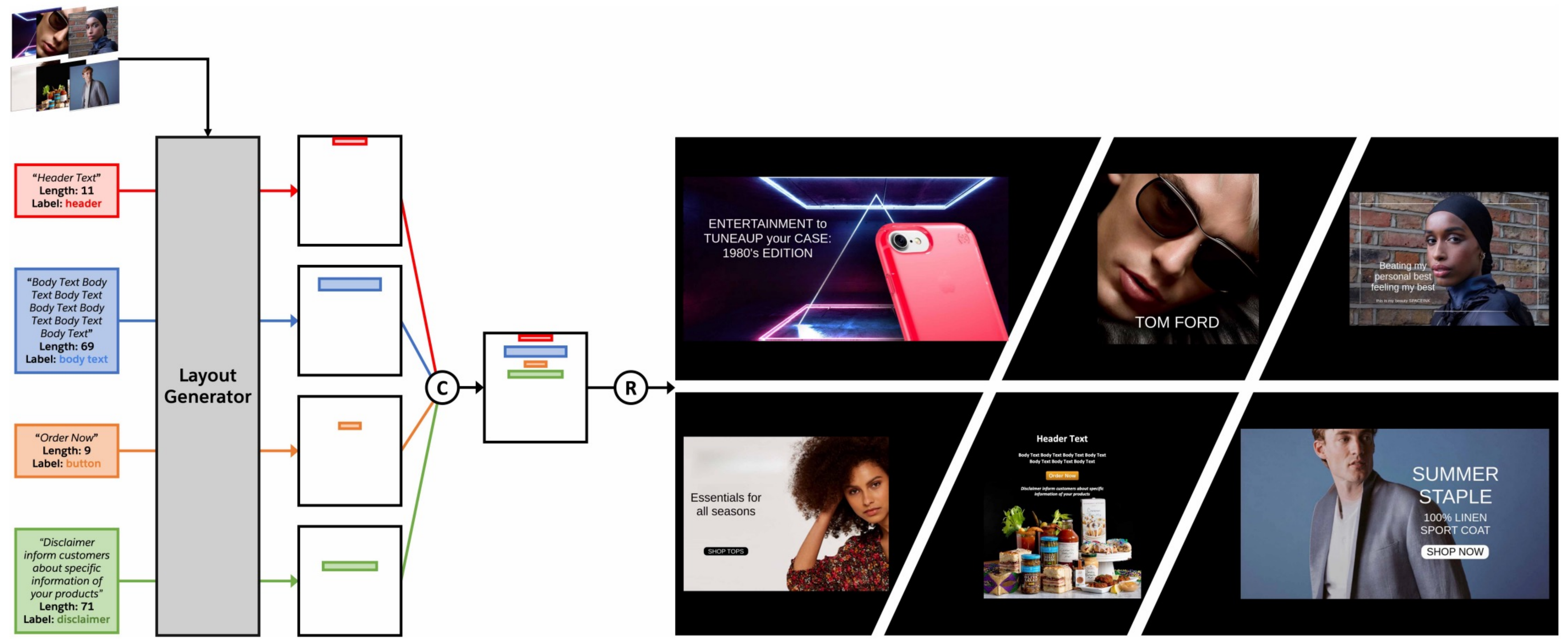
Goal

- Automate the multimodal layout design process by learning a layout generator.

Contributions

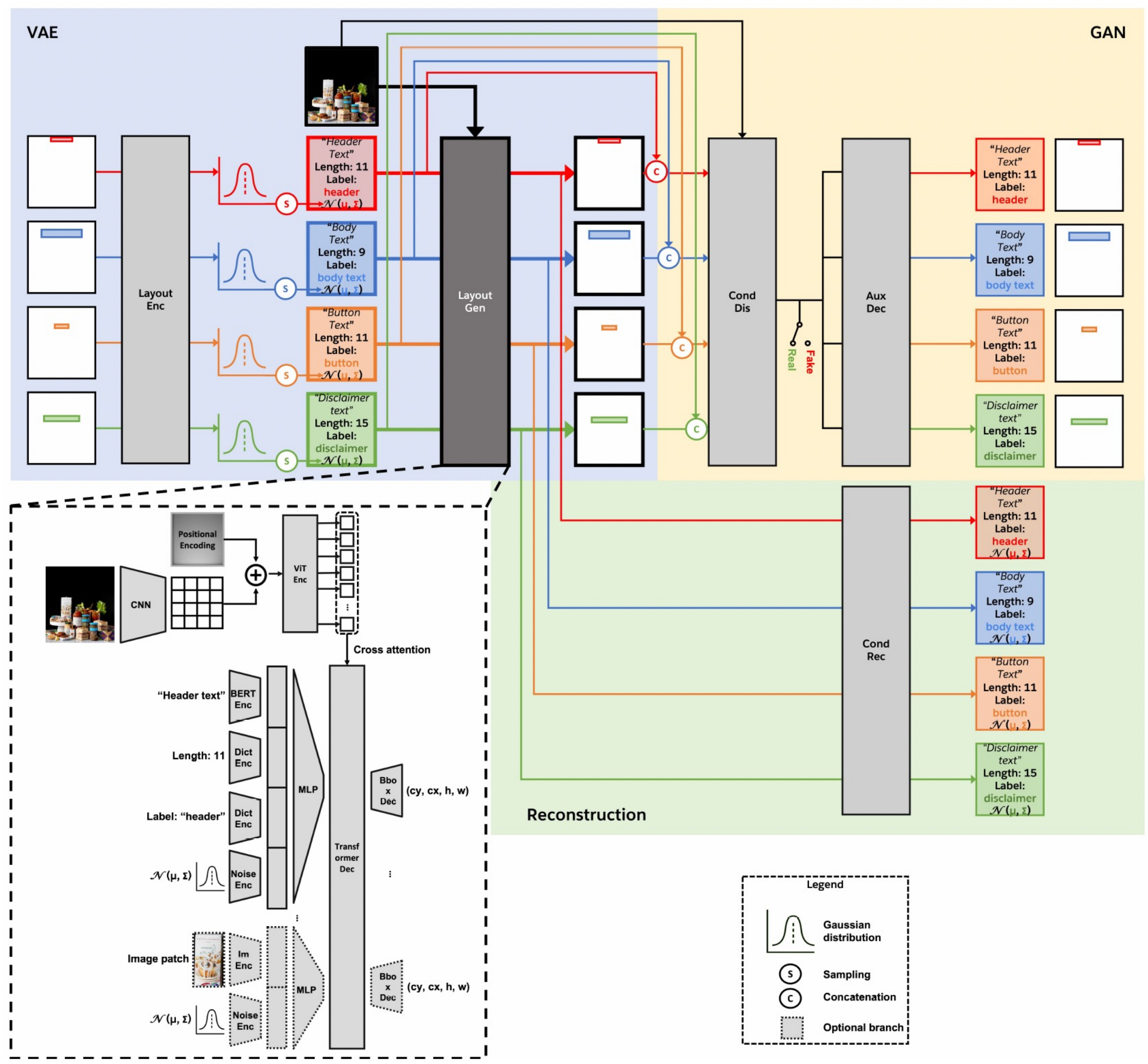
- **Method**: We bridge layout generation and visual detection into one framework that solves multimodal graphic layout design.
- **Dataset**: A large-scale multimodal ad banner dataset with 7,196 samples.
- **State-of-the-art performance** in six evaluation metrics, which measure the realism, accuracy, and regularity of generated layouts
- **Graphical system and user study**: Scales up layout generation and facilitates user studies. Users prefer our designs by significant margins.

Pipeline


"Header Text"
Length: 11
Label: header
"Body Text Body Text Body Text Body Text Body Text Body Text"
Length: 69
Label: body text
"Order Now"
Length: 9
Label: button
"Disclaimer inform customers about specific information of your products"
Length: 71
Label: disclaimer
Layout Generator
C
R
ENTERTAINMENT to TUNEAUP your CASE: 1980's EDITION
TOM FORD
Header Text
Essentials for all seasons
SUMMER STAPLE
100% LINEN SPORT COAT
SHOP NOW

Framework

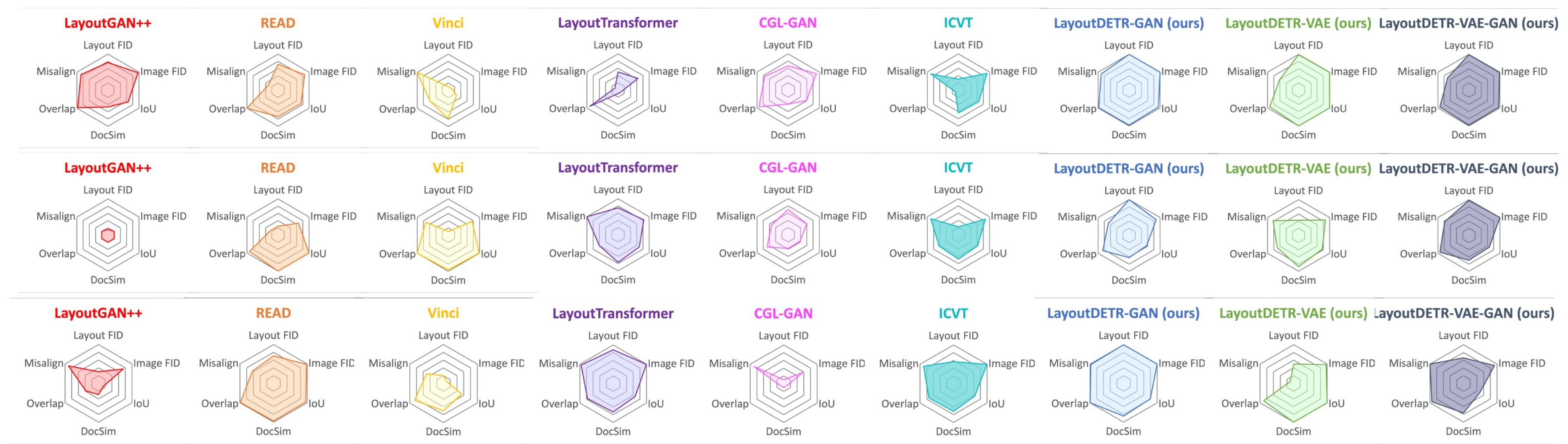
- GAN+VAE
- Unconditional/condition discriminators
- Auxiliar reconstructor
- DETR-based multimodal architecture
- BERT text encoder; ViT image encoder



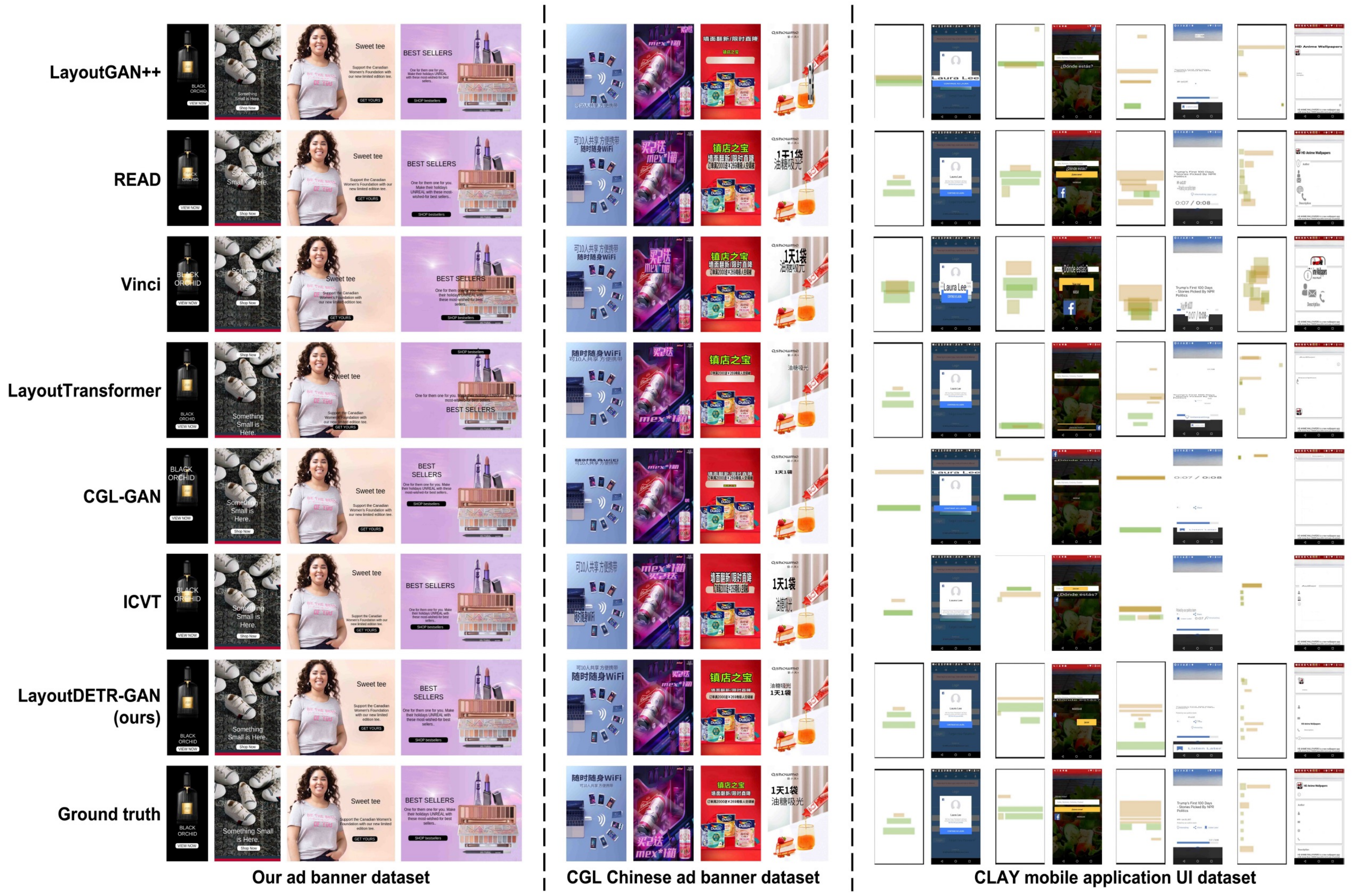
Ablation study

Method	Realism Layout FID ⇓	 Layout KID ($\times10^{-3}$) ⇓	 Image FID ⇓	 Image KID ($\times10^{-5}$) ⇓	Accuracy IoU ⇑	 DocSim ⇑	Regularity Overlap ⇓	 Misalign ($\times10^{-2}$) ⇓
Random layout on real bg	$58.21_{\pm4.04}$	$525.93_{\pm45.08}$	$51.01_{\pm0.41}$	$582.47_{\pm7.53}$	–	–	–	–
Conditional LayoutGAN++	$11.33_{\pm0.10}$	$44.77_{\pm0.36}$	$36.06_{\pm0.02}$	$115.16_{\pm3.37}$	$0.111_{\pm0.001}$	$0.121_{\pm0.001}$	$0.374_{\pm0.006}$	$2.194_{\pm0.058}$
+ Aux. Dec. (Eq. 4-7)	$4.25_{\pm0.01}$	$16.62_{\pm0.05}$	$28.40_{\pm0.06}$	$58.5_{\pm1.45}$	$0.163_{\pm0.002}$	$0.130_{\pm0.001}$	$0.104_{\pm0.003}$	$0.759_{\pm0.021}$
+ Gen. Rec. (Eq. 11)	$3.27_{\pm0.01}$	$11.80_{\pm0.04}$	$29.56_{\pm0.06}$	$11.29_{\pm0.20}$	$0.186_{\pm0.002}$	$\underline{0.148}_{\pm0.001}$	$0.125_{\pm0.003}$	$0.853_{\pm0.016}$
+ Uncond. Dis. D^u	$3.70_{\pm0.05}$	$16.23_{\pm0.08}$	$29.21_{\pm0.08}$	$25.09_{\pm0.02}$	$0.177_{\pm0.002}$	$0.140_{\pm0.001}$	$\underline{0.103}_{\pm0.003}$	$\underline{0.681}_{\pm0.011}$
+ gIoU loss (Eq. 10)	$\underline{3.23}_{\pm0.01}$	$11.60_{\pm0.02}$	$\underline{28.20}_{\pm0.04}$	$\underline{10.51}_{\pm0.09}$	$0.182_{\pm0.002}$	$0.138_{\pm0.001}$	$0.106_{\pm0.003}$	$0.721_{\pm0.011}$
+ Overlap & Misalign loss ≐ LayoutDETR-GAN (ours)	$\mathbf{3.19}_{\pm0.01}$	$\mathbf{5.62}_{\pm0.01}$	$\mathbf{27.35}_{\pm0.04}$	$\mathbf{8.31}_{\pm0.80}$	$\mathbf{0.208}_{\pm0.002}$	$\mathbf{0.151}_{\pm0.000}$	$\mathbf{0.101}_{\pm0.003}$	$\mathbf{0.646}_{\pm0.011}$
- Text length embeddings	$3.24_{\pm0.01}$	$\underline{9.25}_{\pm0.05}$	$28.65_{\pm0.03}$	$11.42_{\pm0.35}$	$\underline{0.191}_{\pm0.002}$	$0.144_{\pm0.001}$	$0.117_{\pm0.003}$	$0.807_{\pm0.012}$
- Text class embeddings	$25.17_{\pm0.54}$	$171.88_{\pm5.17}$	$29.25_{\pm0.25}$	$139.16_{\pm4.44}$	$0.166_{\pm0.002}$	$0.132_{\pm0.001}$	$0.110_{\pm0.001}$	$0.000_{\pm0.000}$

Quantitative comparisons



Qualitative comparisons



User study

Method	READ	Vinci	LayoutTransformer	CGL-GAN	ICVT	LayoutDETR-GAN (ours)
LayoutGAN++	$49.8\%_{p=0.4}$	$45.6\%_{p=3e-3}$	$44.4\%_{p=3e-4}$	$53.9\%_{p=0.01}$	$47.1\%_{p=0.04}$	$55.7\%_{p=2e-4}$
READ	–	$45.1\%_{p=1e-3}$	$44.5\%_{p=3e-4}$	$53.8\%_{p=0.01}$	$53.0\%_{p=0.04}$	$54.2\%_{p=5e-3}$
Vinci	–	–	$51.7\%_{p=0.2}$	$55.8\%_{p=2e-4}$	$56.9\%_{p=1e-5}$	$62.6\%_{p=3e-15}$
LayoutTransformer	–	–	–	$57.1\%_{p=8e-6}$	$56.0\%_{p=2e-4}$	$63.5\%_{p=2e-17}$
CGL-GAN	–	–	–	–	$48.9\%_{p=0.2}$	$54.7\%_{p=3e-3}$
ICVT	–	–	–	–	–	$55.4\%_{p=6e-4}$