# 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.

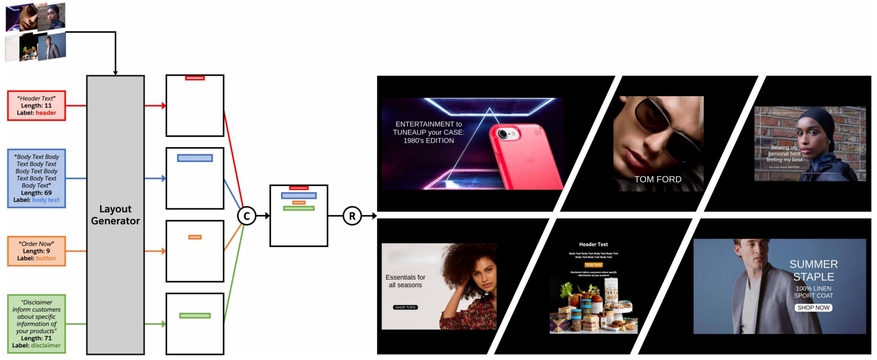
## Goal

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

## Contributions

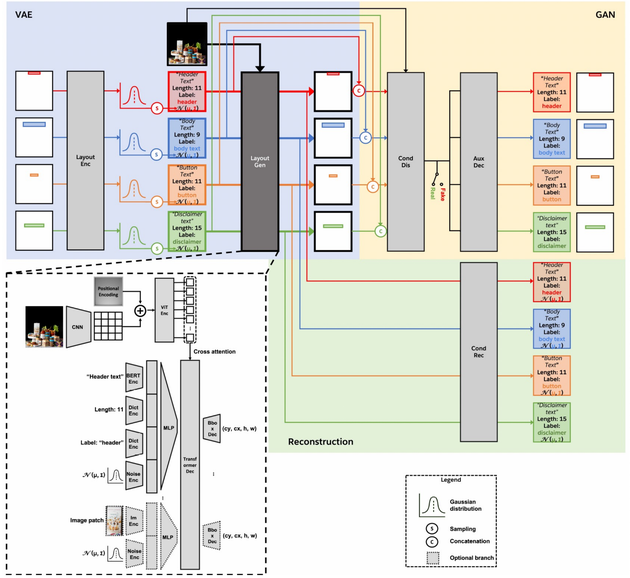
- **Method**: We bridge layout generation and visual detection into one framework that solves multimodal graphic layout design.
- **Dataset**: A large-scale multimodal ad banner dataset with 7,196 samples.
- **State-of-the-art performance** in six evaluation metrics, which measure the realism, accuracy, 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




## Framework

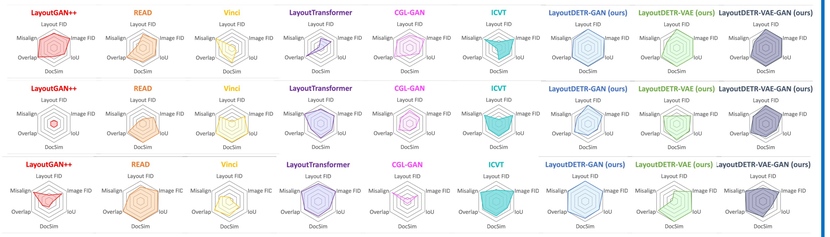
- GAN+VAE
- Unconditional/condition discriminators
- Auxiliar reconstructor
- DETR-based multimodal architecture
- BERT text encoder; ViT image encoder




## Ablation study

| Method | Realism: Layout FID ⇓ | Realism: Layout KID ($\times 10^{-3}$) ⇓ | Realism: Image FID ⇓ | Realism: Image KID ($\times 10^{-5}$) ⇓ | Accuracy: IoU ⇑ | Accuracy: DocSim ⇑ | Regularity: Overlap ⇓ | Regularity: Misalign ($\times 10^{-2}$) ⇓ |
|---|---|---|---|---|---|---|---|---|
| Random layout on real bg | $58.21_{\pm 4.04}$ | $525.93_{\pm 45.08}$ | $51.01_{\pm 0.41}$ | $582.47_{\pm 7.53}$ | – | – | – | – |
| Conditional LayoutGAN++ | $11.33_{\pm 0.10}$ | $44.77_{\pm 0.36}$ | $36.06_{\pm 0.02}$ | $115.16_{\pm 3.37}$ | $0.111_{\pm 0.001}$ | $0.121_{\pm 0.001}$ | $0.374_{\pm 0.006}$ | $2.194_{\pm 0.058}$ |
| + Aux. Dec. (Eq. 4-7) | $4.25_{\pm 0.01}$ | $16.62_{\pm 0.05}$ | $28.40_{\pm 0.06}$ | $58.5_{\pm 1.45}$ | $0.163_{\pm 0.002}$ | $0.130_{\pm 0.001}$ | $0.104_{\pm 0.003}$ | $0.759_{\pm 0.021}$ |
| + Gen. Rec. (Eq. 11) | $3.27_{\pm 0.01}$ | $11.80_{\pm 0.04}$ | $29.56_{\pm 0.06}$ | $11.29_{\pm 0.20}$ | $0.186_{\pm 0.002}$ | $\underline{0.148}_{\pm 0.001}$ | $0.125_{\pm 0.003}$ | $0.853_{\pm 0.016}$ |
| + Uncond. Dis. $D^u$ | $3.70_{\pm 0.05}$ | $16.23_{\pm 0.08}$ | $29.21_{\pm 0.08}$ | $25.09_{\pm 0.02}$ | $0.177_{\pm 0.002}$ | $0.140_{\pm 0.001}$ | $\underline{0.103}_{\pm 0.003}$ | $\underline{0.681}_{\pm 0.011}$ |
| + gIoU loss (Eq. 10) | $\underline{3.23}_{\pm 0.01}$ | $11.60_{\pm 0.02}$ | $\underline{28.20}_{\pm 0.04}$ | $\underline{10.51}_{\pm 0.09}$ | $0.182_{\pm 0.002}$ | $0.138_{\pm 0.001}$ | $0.106_{\pm 0.003}$ | $0.721_{\pm 0.011}$ |
| + Overlap & Misalign loss ≙ LayoutDETR-GAN (ours) | $\mathbf{3.19}_{\pm 0.01}$ | $\mathbf{5.62}_{\pm 0.01}$ | $\mathbf{27.35}_{\pm 0.04}$ | $\mathbf{8.31}_{\pm 0.80}$ | $\mathbf{0.208}_{\pm 0.002}$ | $\mathbf{0.151}_{\pm 0.000}$ | $\mathbf{0.101}_{\pm 0.003}$ | $\mathbf{0.646}_{\pm 0.011}$ |
| - Text length embeddings | $3.24_{\pm 0.01}$ | $\underline{9.25}_{\pm 0.05}$ | $28.65_{\pm 0.03}$ | $11.42_{\pm 0.35}$ | $\underline{0.191}_{\pm 0.002}$ | $0.144_{\pm 0.001}$ | $0.117_{\pm 0.003}$ | $0.807_{\pm 0.012}$ |
| - Text class embeddings | $25.17_{\pm 0.54}$ | $171.88_{\pm 5.17}$ | $29.25_{\pm 0.25}$ | $139.16_{\pm 4.44}$ | $0.166_{\pm 0.002}$ | $0.132_{\pm 0.001}$ | $0.110_{\pm 0.001}$ | $0.000_{\pm 0.000}$ |

## Quantitative comparisons




## Quantitative 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 | – | – | – | – | – | $55.4\%_{p=6e-4}$ |