Reinforcement Learning finetuned Vision-Code Transformer for UI-to-Code Generation

Davit Soselia, Khalid Saifullah, Tianyi Zhou

University of Maryland, College Park

Generating code from screenshots

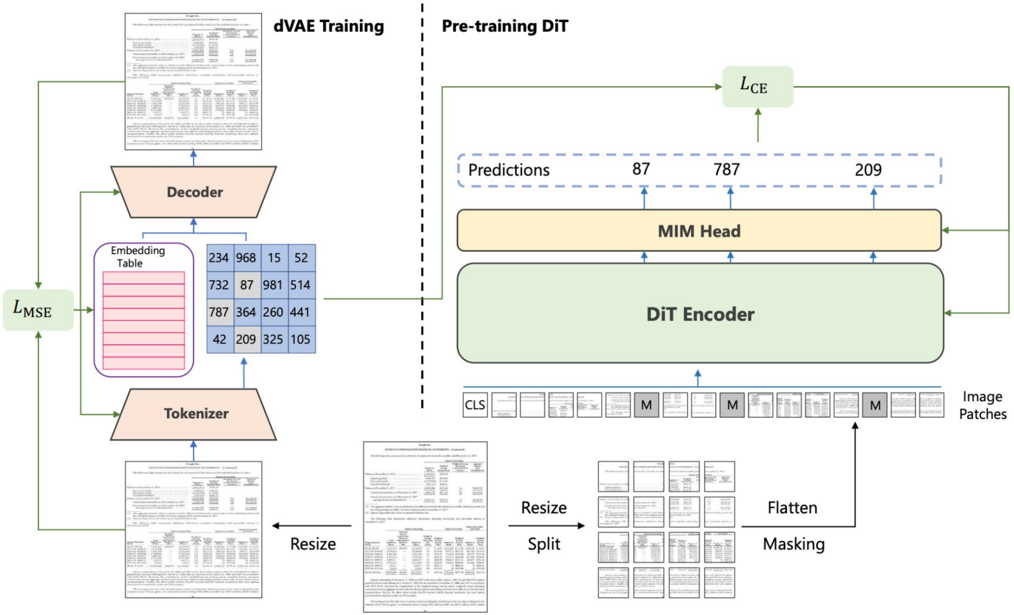
- Labor-intensive and time-consuming process
- Automation prone to errors
- Text similarity ≠ visual similarity

Related

- Pix2code
- Sketch2code
- Pix2Struct

Related

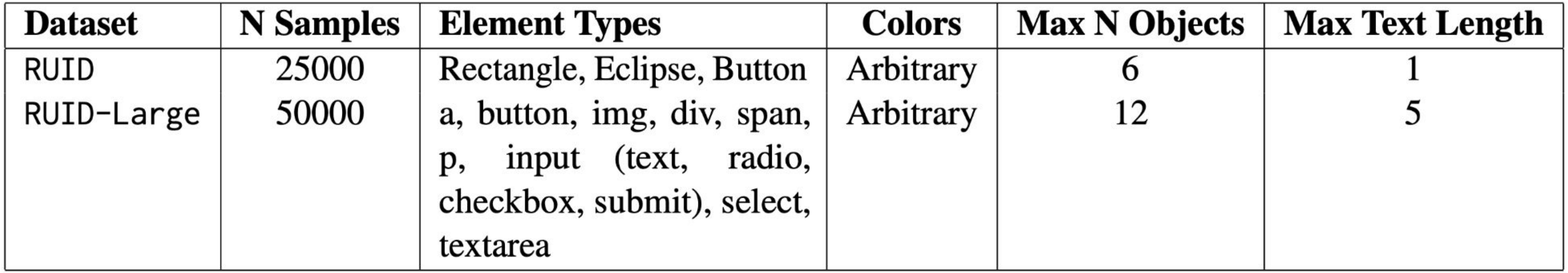
- Pix2code
- Sketch2code
- Pix2Struct



234	968	15	52
732	87	981	514
787	364	260	441
42	209	325	105

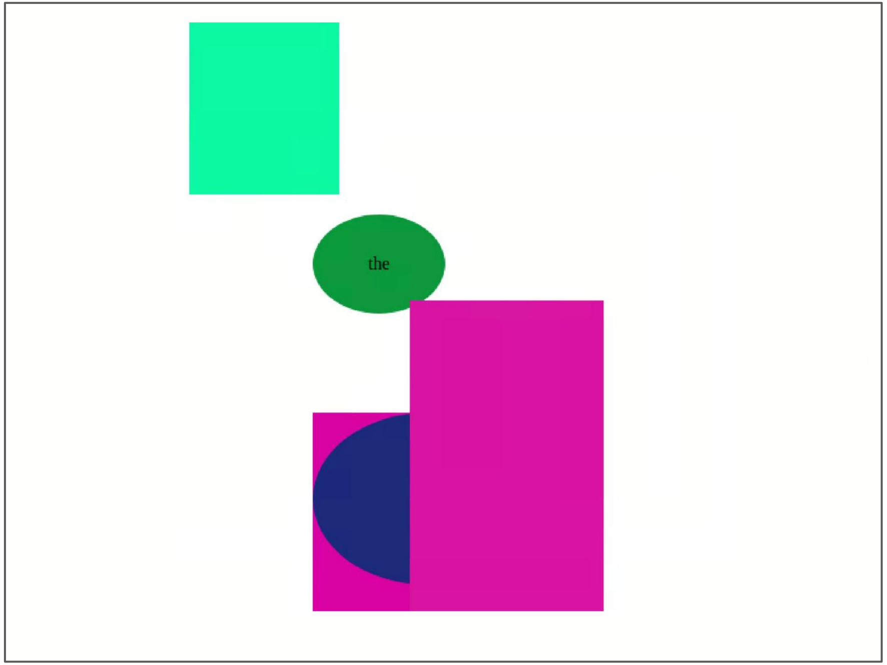
Figure 2: The model architecture of DiT with MIM pre-training.

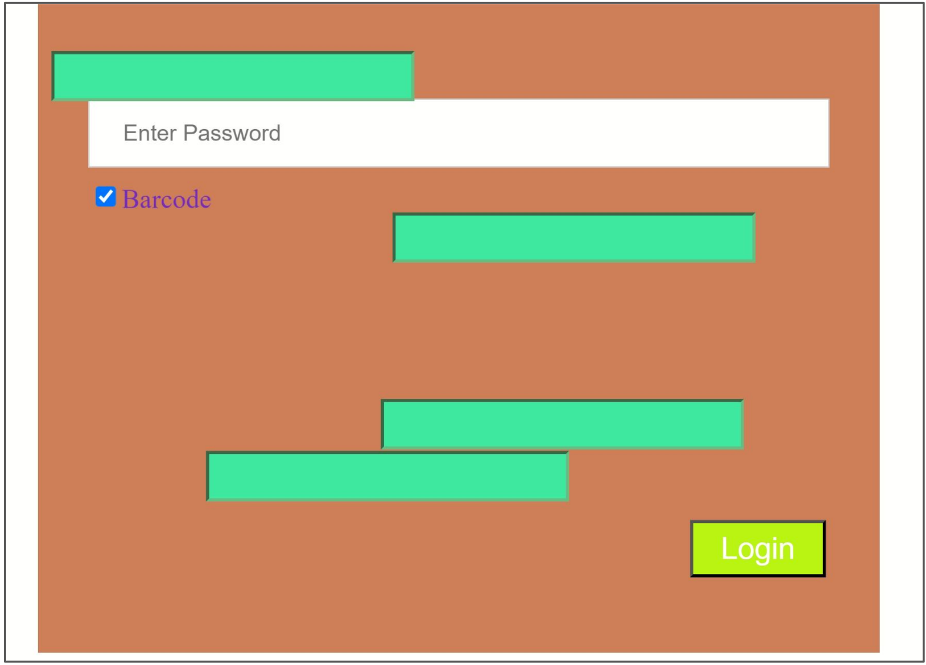
Datasets



Dataset	N Samples	Element Types	Colors	Max N Objects	Max Text Length
RUID	25000	Rectangle, Eclipse, Button	Arbitrary	6	1
RUID-Large	50000	a, button, img, div, span, p, input (text, radio, checkbox, submit), select, textarea	Arbitrary	12	5

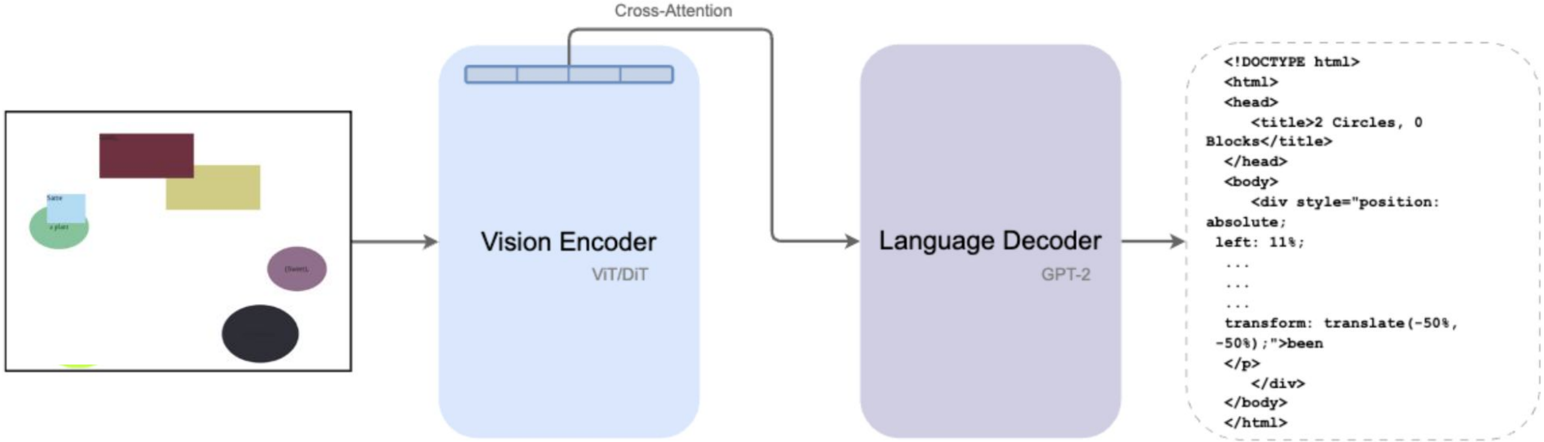
Samples




Enter Password

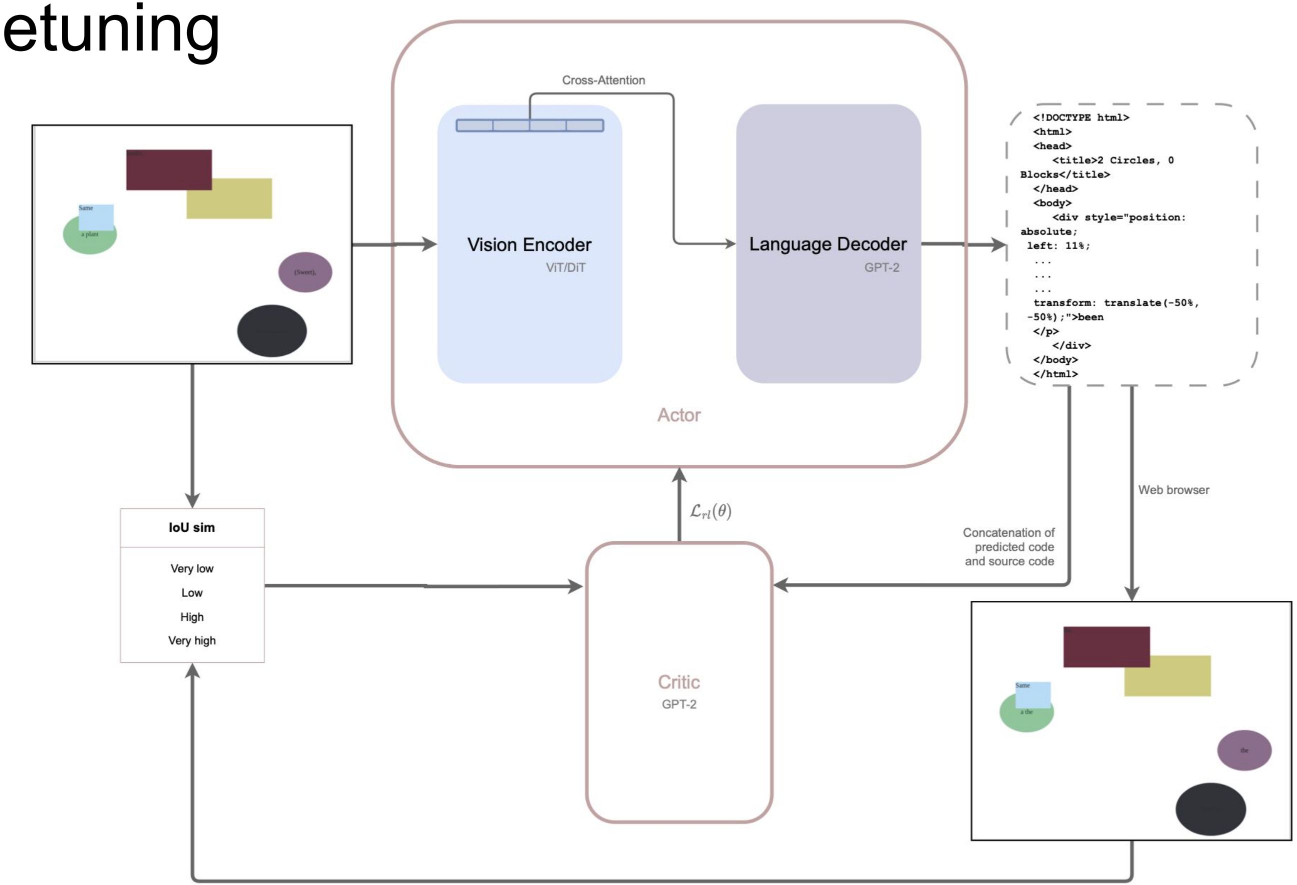
☑ Barcode

Baseline



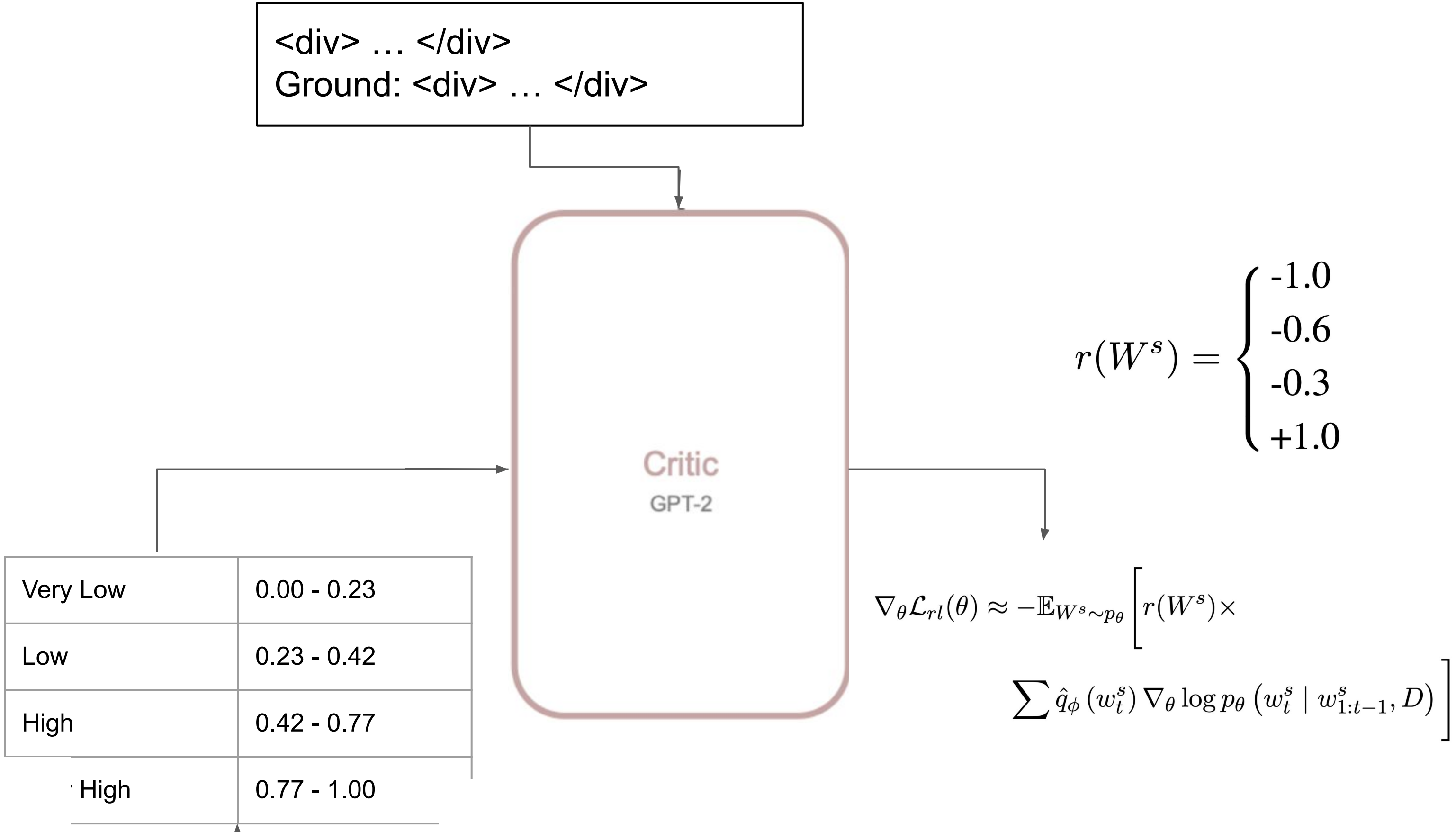
```
  <!DOCTYPE html>
  <html>
  <head>
     <title>2 Circles, 0
Blocks</title>
  </head>
  <body>
     <div style="position:
absolute;
 left: 11%;
  ...
  ...
  ...
  transform: translate(-50%,
 -50%);">been
  </p>
     </div>
  </body>
  </html>
```

RL Finetuning



```
  <!DOCTYPE html>
  <html>
  <head>
     <title>2 Circles, 0
Blocks</title>
  </head>
  <body>
     <div style="position:
absolute;
 left: 11%;
  ...
  ...
  ...
  transform: translate(-50%,
 -50%);">been
  </p>
     </div>
  </body>
  </html>
```

Critic



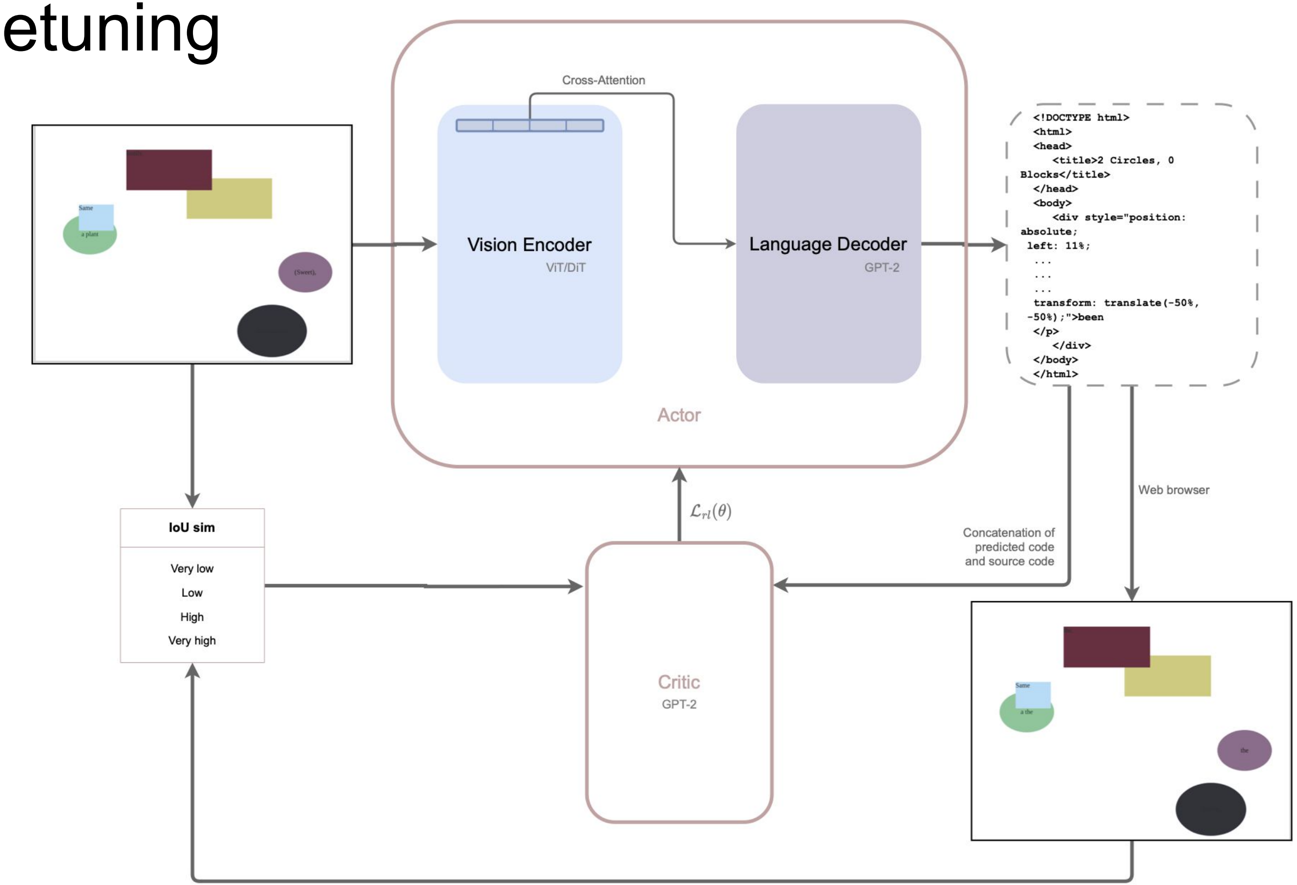
<div> … </div>
Ground: <div> … </div>

Very Low	0.00 - 0.23
Low	0.23 - 0.42
High	0.42 - 0.77
High	0.77 - 1.00

$$r(W^s) = \begin{cases} -1.0 \\ -0.6 \\ -0.3 \\ +1.0 \end{cases}$$

$$\nabla_\theta \mathcal{L}_{rl}(\theta) \approx -\mathbb{E}_{W^s \sim p_\theta}\left[r(W^s) \times \sum \hat{q}_\phi\left(w_t^s\right) \nabla_\theta \log p_\theta\left(w_t^s \mid w_{1:t-1}^s, D\right)\right]$$

RL Finetuning

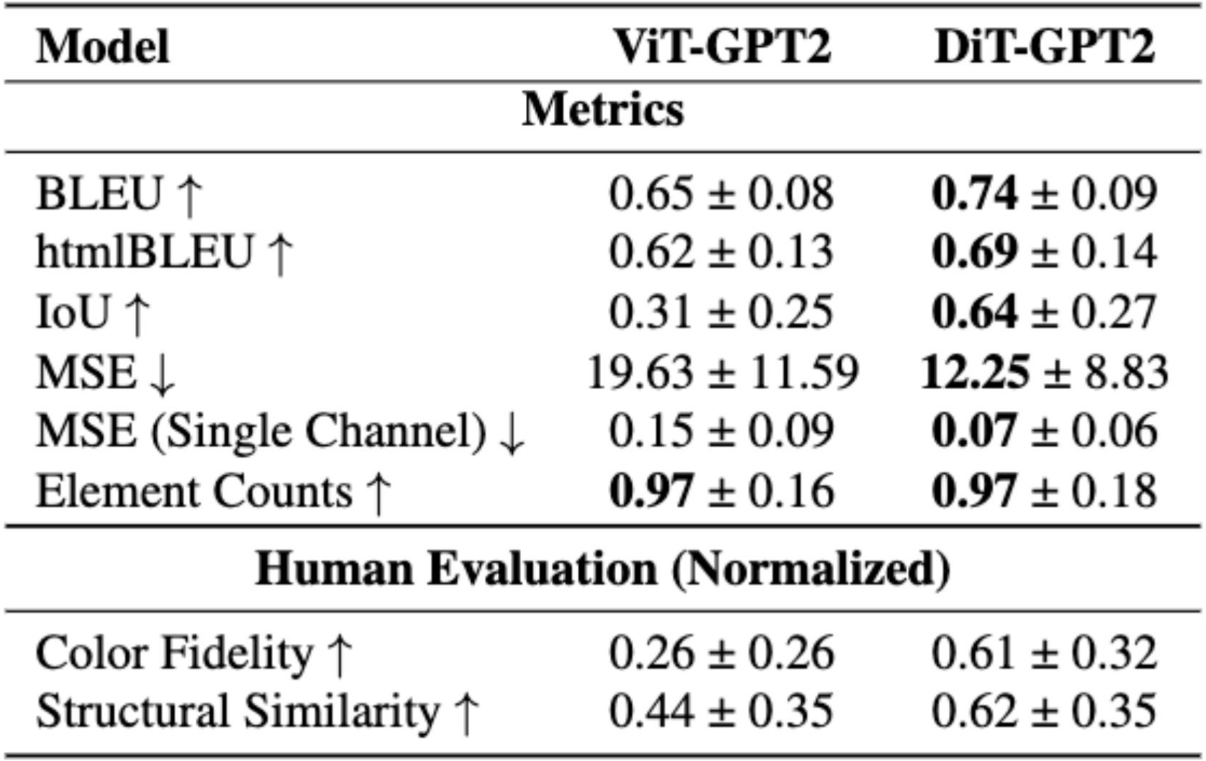


```
  <!DOCTYPE html>
  <html>
  <head>
     <title>2 Circles, 0
Blocks</title>
  </head>
  <body>
     <div style="position:
absolute;
 left: 11%;
  ...
  ...
  ...
  transform: translate(-50%,
 -50%);">been
  </p>
     </div>
  </body>
  </html>
```

CodeBLEU → htmlBLEU

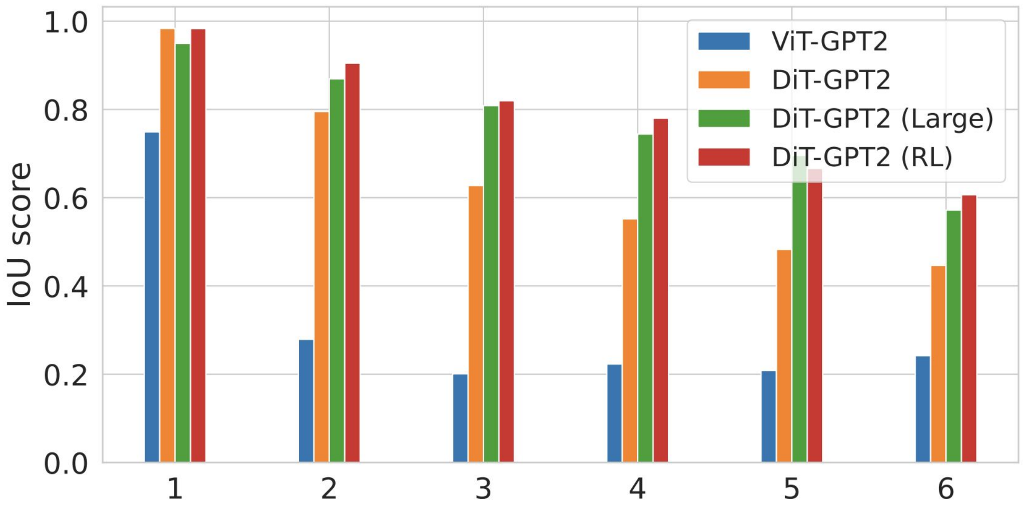
- Syntactic AST match - > Syntactic DOM Tree Match
- Semantic data-flow match -> Semantic Attribute Match
- Weighted tag matching - Calculated by seeing effect of tag error on MSE

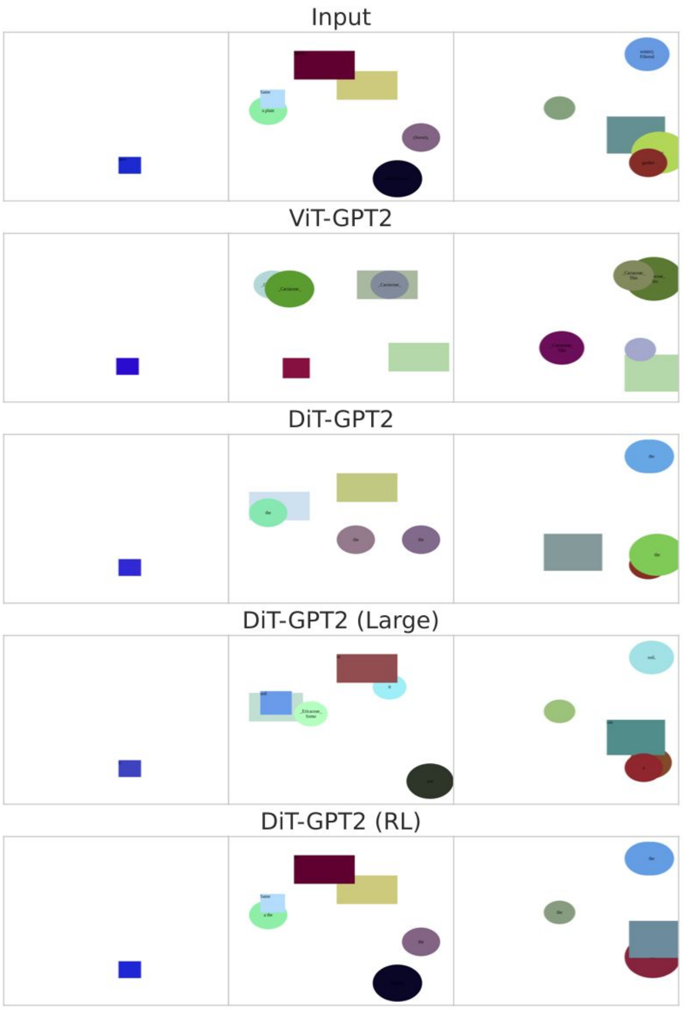
Results - Baselines



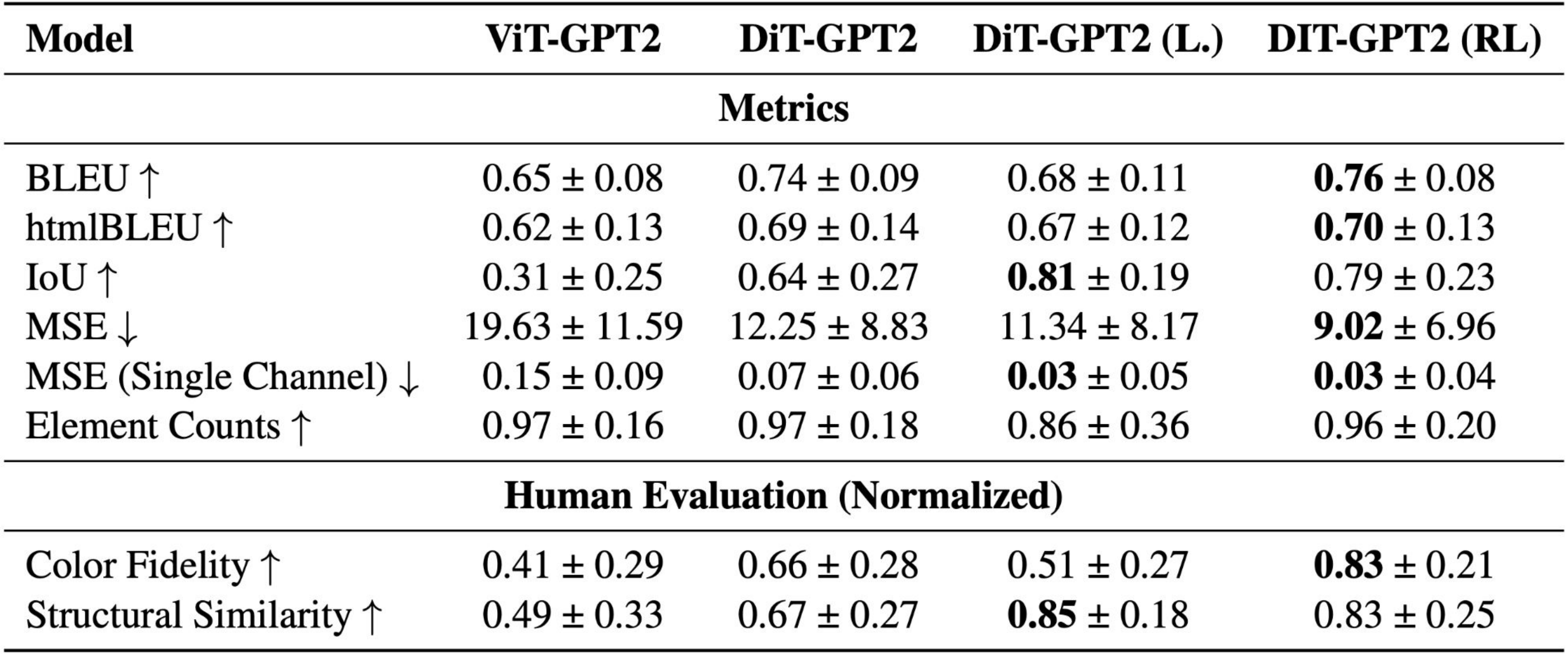
Model	**ViT-GPT2**	**DiT-GPT2**
Metrics		
BLEU ↑	0.65 ± 0.08	**0.74** ± 0.09
htmlBLEU ↑	0.62 ± 0.13	**0.69** ± 0.14
IoU ↑	0.31 ± 0.25	**0.64** ± 0.27
MSE ↓	19.63 ± 11.59	**12.25** ± 8.83
MSE (Single Channel) ↓	0.15 ± 0.09	**0.07** ± 0.06
Element Counts ↑	**0.97** ± 0.16	**0.97** ± 0.18
Human Evaluation (Normalized)		
Color Fidelity ↑	0.26 ± 0.26	0.61 ± 0.32
Structural Similarity ↑	0.44 ± 0.35	0.62 ± 0.35

Results


IoU score
1.0
0.8
0.6
0.4
0.2
0.0
1
2
3
4
5
6
ViT-GPT2
DiT-GPT2
DiT-GPT2 (Large)
DiT-GPT2 (RL)


Input
ViT-GPT2
DiT-GPT2
DiT-GPT2 (Large)
DiT-GPT2 (RL)

Results



Model	ViT-GPT2	DiT-GPT2	DiT-GPT2 (L.)	DIT-GPT2 (RL)
Metrics				
BLEU ↑	0.65 ± 0.08	0.74 ± 0.09	0.68 ± 0.11	**0.76** ± 0.08
htmlBLEU ↑	0.62 ± 0.13	0.69 ± 0.14	0.67 ± 0.12	**0.70** ± 0.13
IoU ↑	0.31 ± 0.25	0.64 ± 0.27	**0.81** ± 0.19	0.79 ± 0.23
MSE ↓	19.63 ± 11.59	12.25 ± 8.83	11.34 ± 8.17	**9.02** ± 6.96
MSE (Single Channel) ↓	0.15 ± 0.09	0.07 ± 0.06	**0.03** ± 0.05	**0.03** ± 0.04
Element Counts ↑	0.97 ± 0.16	0.97 ± 0.18	0.86 ± 0.36	0.96 ± 0.20
Human Evaluation (Normalized)				
Color Fidelity ↑	0.41 ± 0.29	0.66 ± 0.28	0.51 ± 0.27	**0.83** ± 0.21
Structural Similarity ↑	0.49 ± 0.33	0.67 ± 0.27	**0.85** ± 0.18	0.83 ± 0.25

Open Questions

- RL methodology choice
- Human Feedback
- Complex datasets
- Model Selection

Thank you!
Any Questions?