Retrieval-Augmented Layout Transformer for Content-Aware Layout Generation





Daichi Horita | **Naoto Inoue** | **Kotaro Kikuchi** | **Kota Yamaguchi** | **Kiyoharu Aizawa**

UTokyo

CyberAgent

CVPR
JUNE 17-21, 2024
SEATTLE, WA

Concept

Content-aware layout generation

Input image

Output layouts

RALF

Retrieval-Augmented Layout Transformer

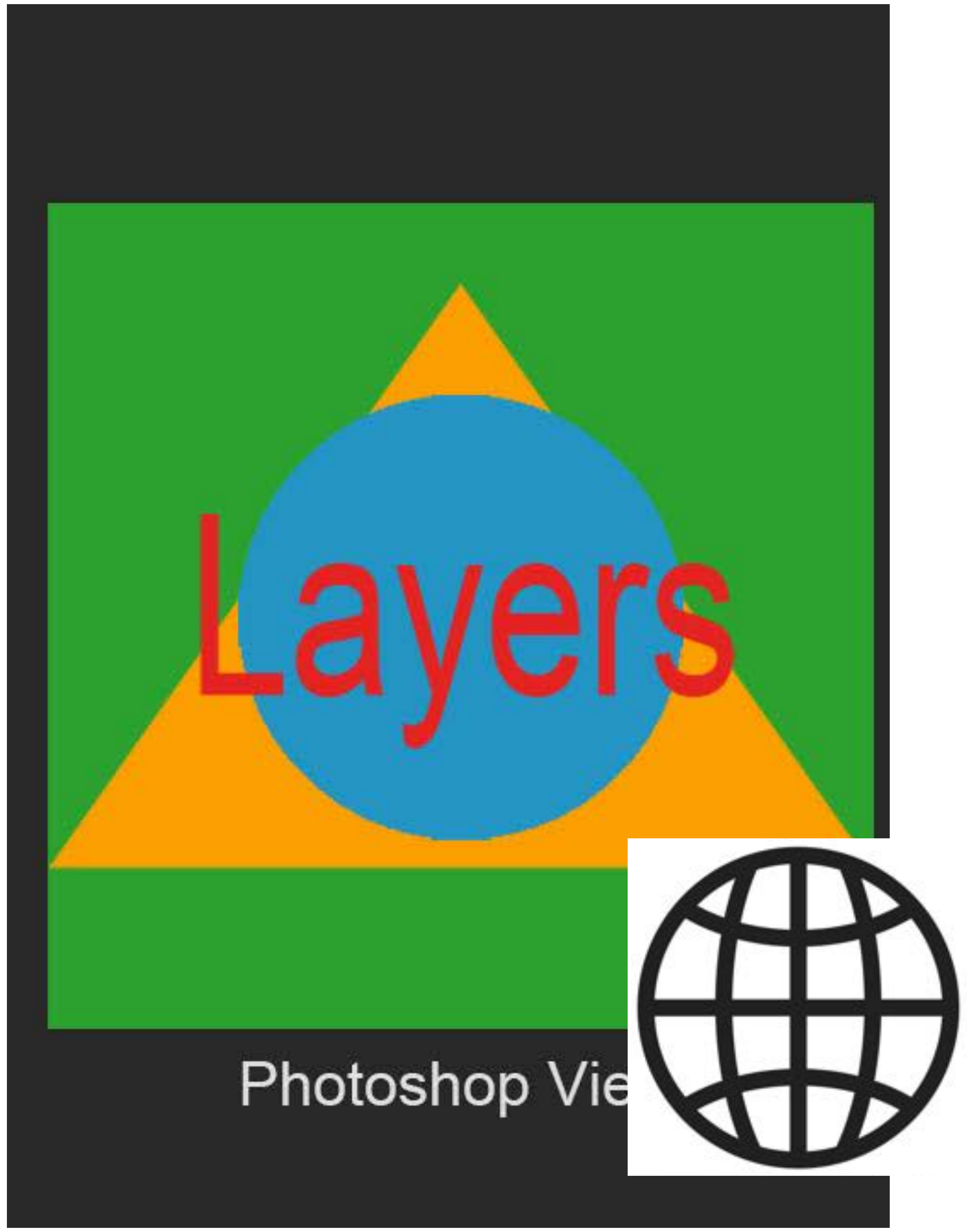
1) **Retrieve nearest neighbor layouts** based on the input image
2) **use them as a reference to augment** the generation process.

Challenges

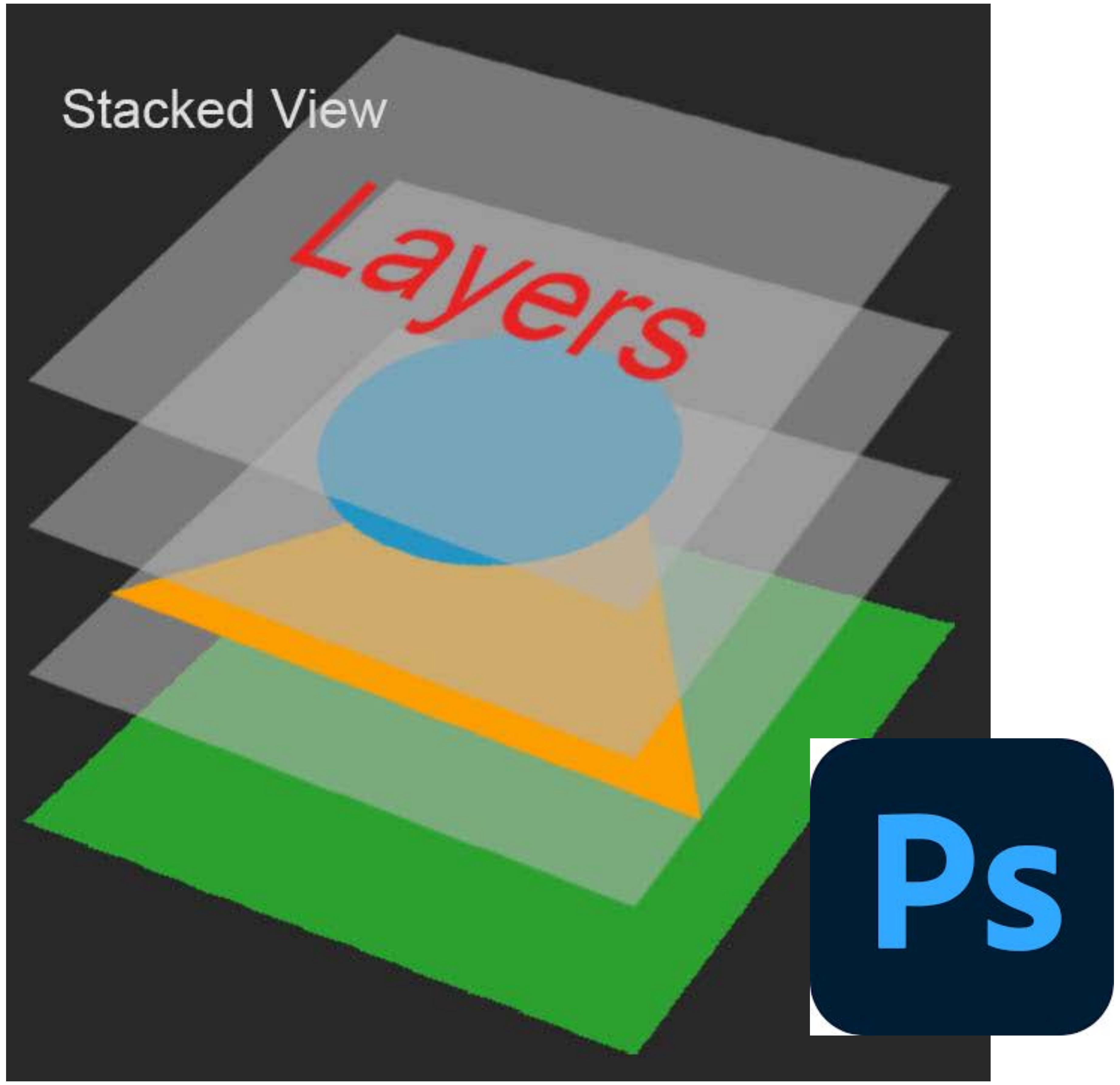
- Data Scarcity & Training Efficiency
- Content-Layout Harmonization
- Controllability to User-Specified Constraints

Challenges

Data Scarcity & Training Efficiency



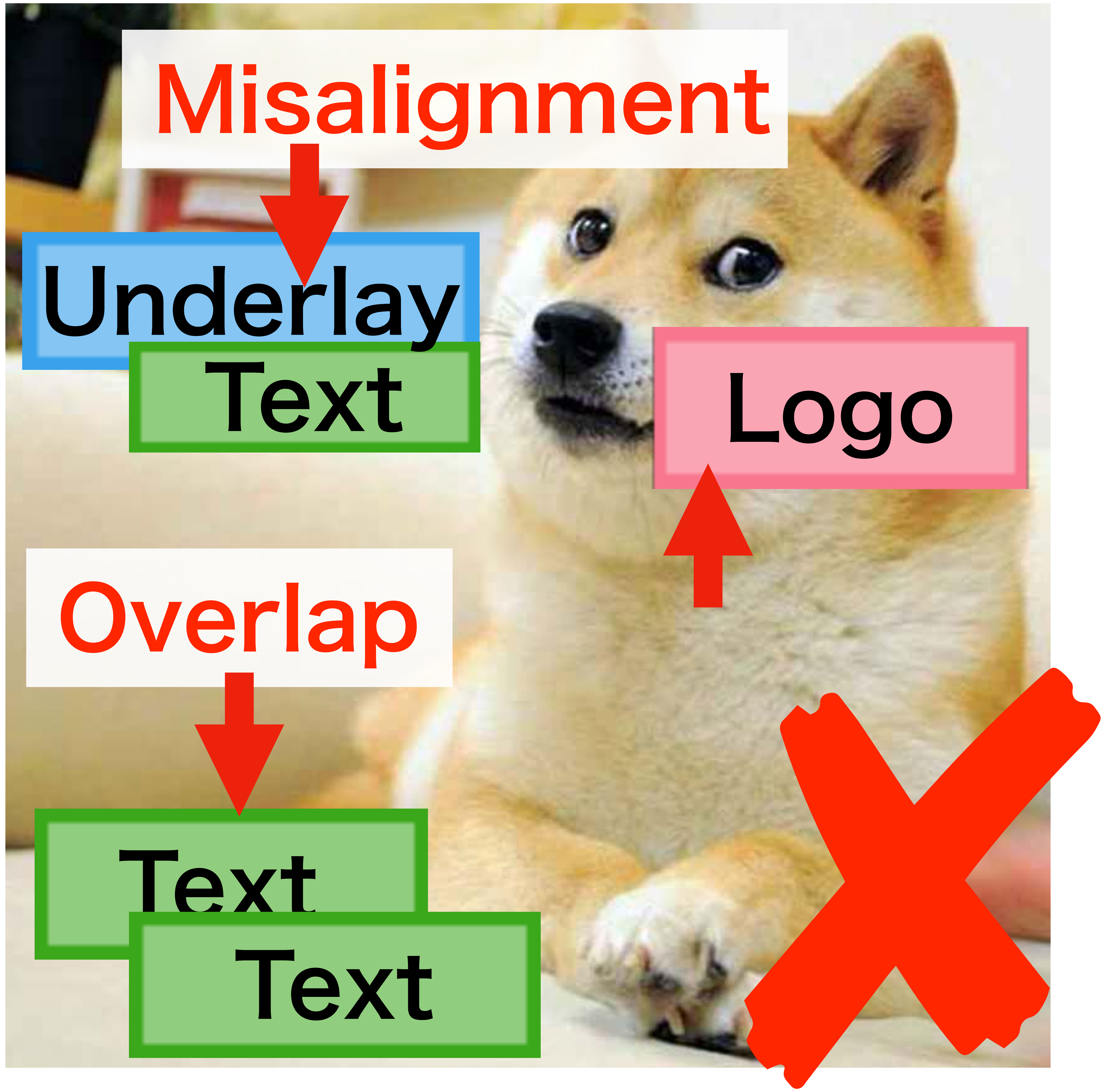
Web

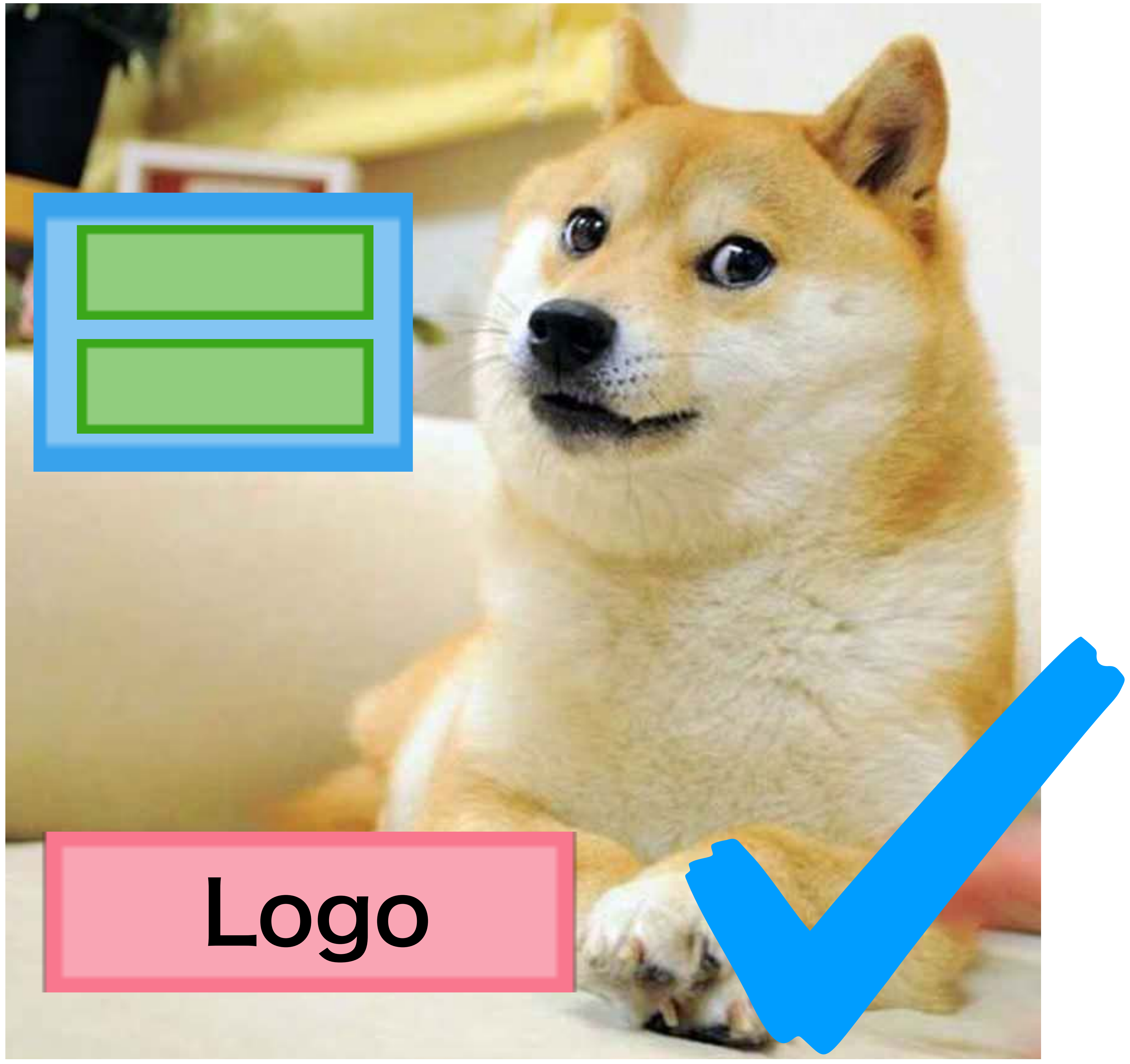


Authoring tool

Challenges

Content-Layout Harmonization





Challenges

Controllability to User-Specified Constraints

Category → Size +Position

“Logo, Text x2, Underlay”

Relationship

“Logo top on Text”

Contributions

- Data Scarcity & Training Efficiency

✓ **Retrieval augmentation effectively addresses the data scarcity problem**

- Content-Layout Harmonization

✓ **Propose RALF**

- Controllability to User-Specified Constraints

✓ **Show RALF outperforms the baselines on unconditional & conditional tasks**

Why Retrieve & Generate?

Just retrieve

GT Input Top1 Output

Why Retrieve & Generate?

Retrieve & Generate

GT

Input

Retrieve

Output

Preliminaries

Representation of layout

$$Z = (bos,$$

① $\boxed{c_1, x_1, y_1, w_1, h_1,}$

② $\boxed{c_2, x_2, y_2, w_2, h_2,}$

$$\ldots, eos)$$

Sorted by raster scan order

Preliminaries

Tokenization of layout

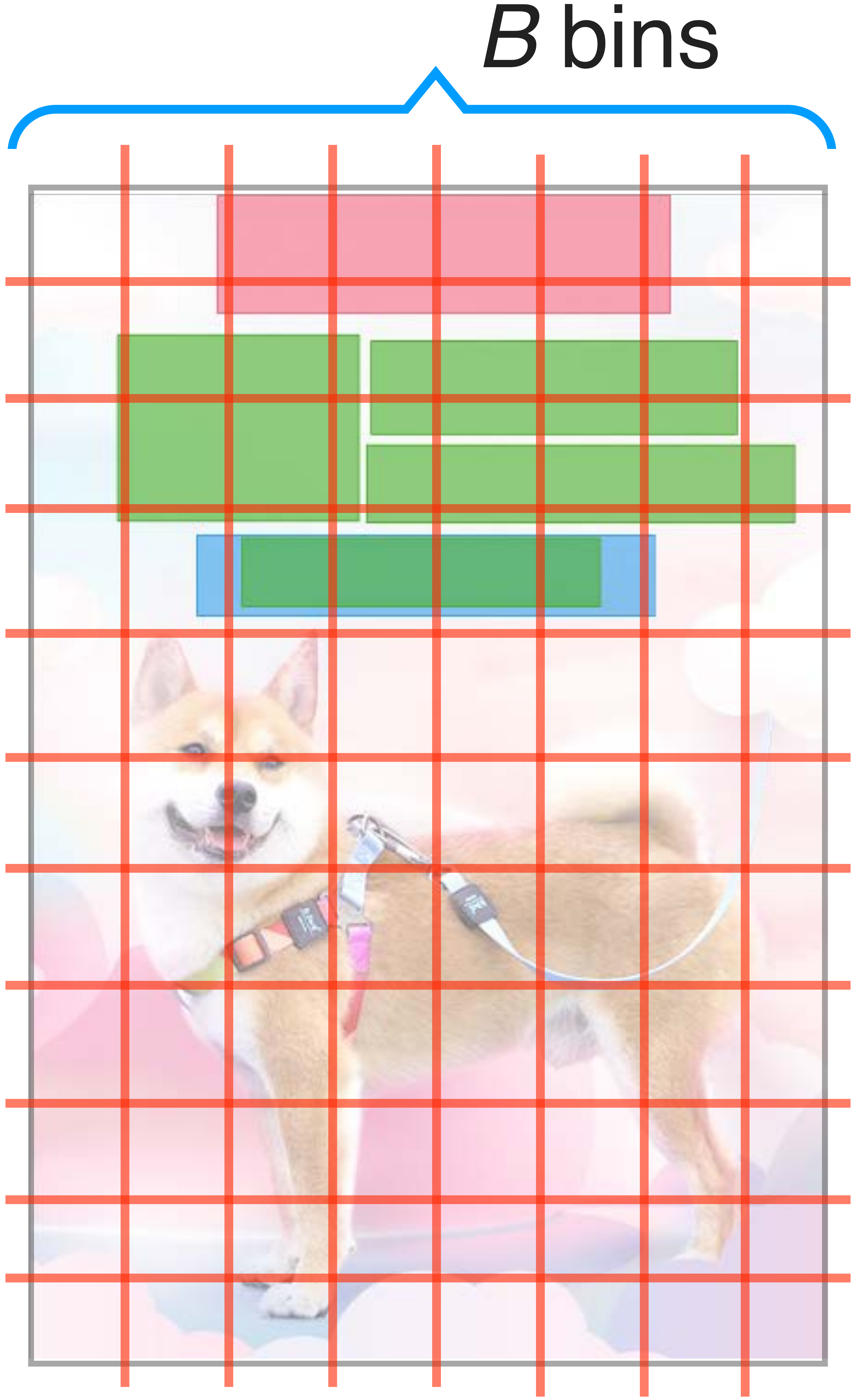
- Quantize bbox $\boldsymbol{b}_i$:

$$[x_i, y_i, w_i, h_i]^T \in \{1, \ldots B\}^4$$

- Autoregressive modeling:

$$P_\theta(Z|I,S) = \prod_{t=2}^{5T+2} P_\theta(Z_t|Z_{<t}, I, S)$$

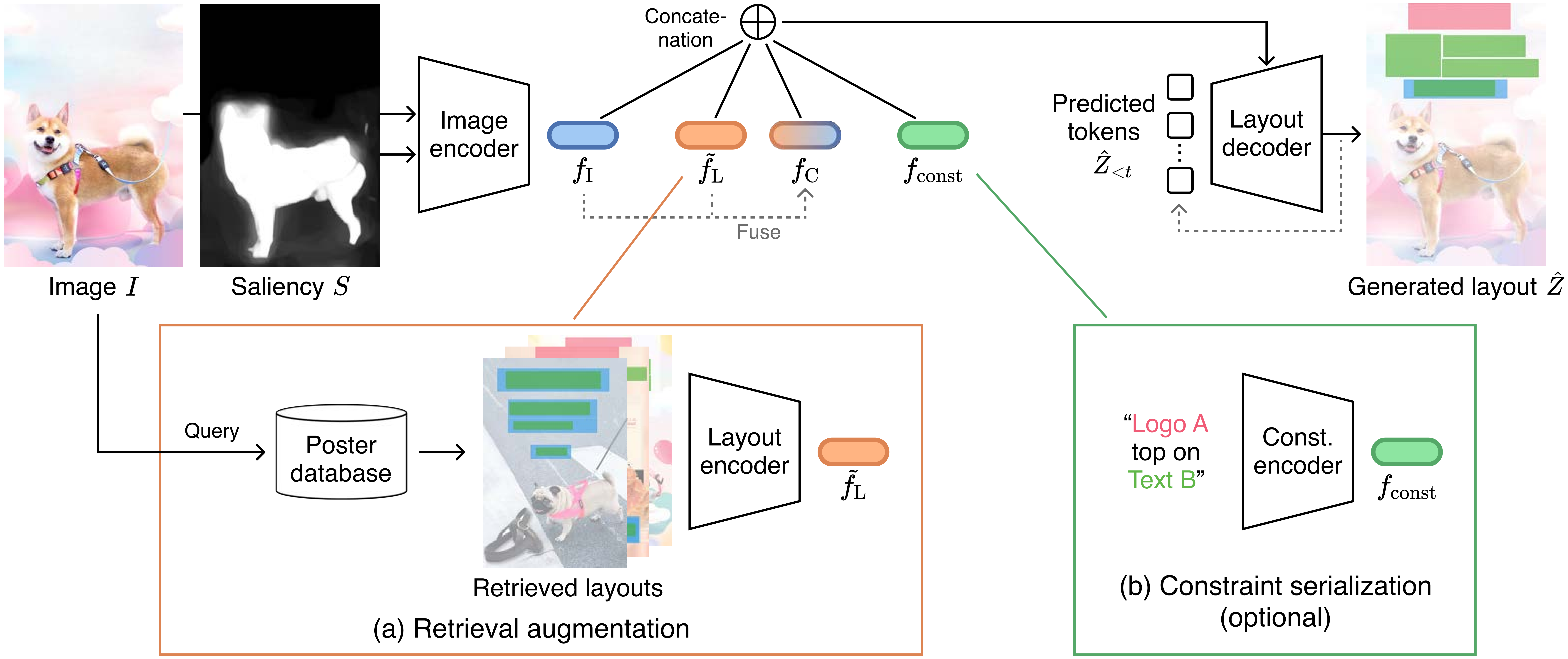
I: image, *S*: saliency map



In exp, $B = 128$

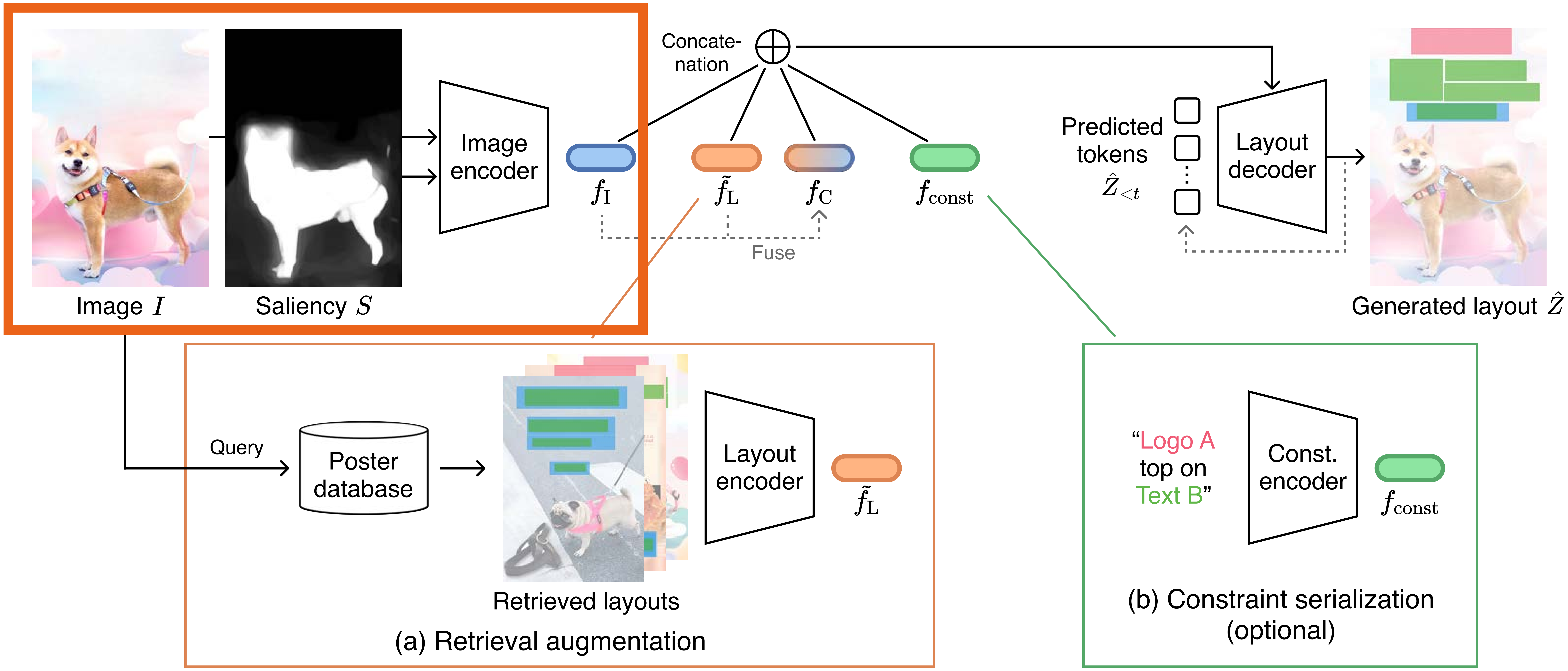
Overview of RALF

Propose a **Retrieval-Augmented Layout Transformer (RALF)**



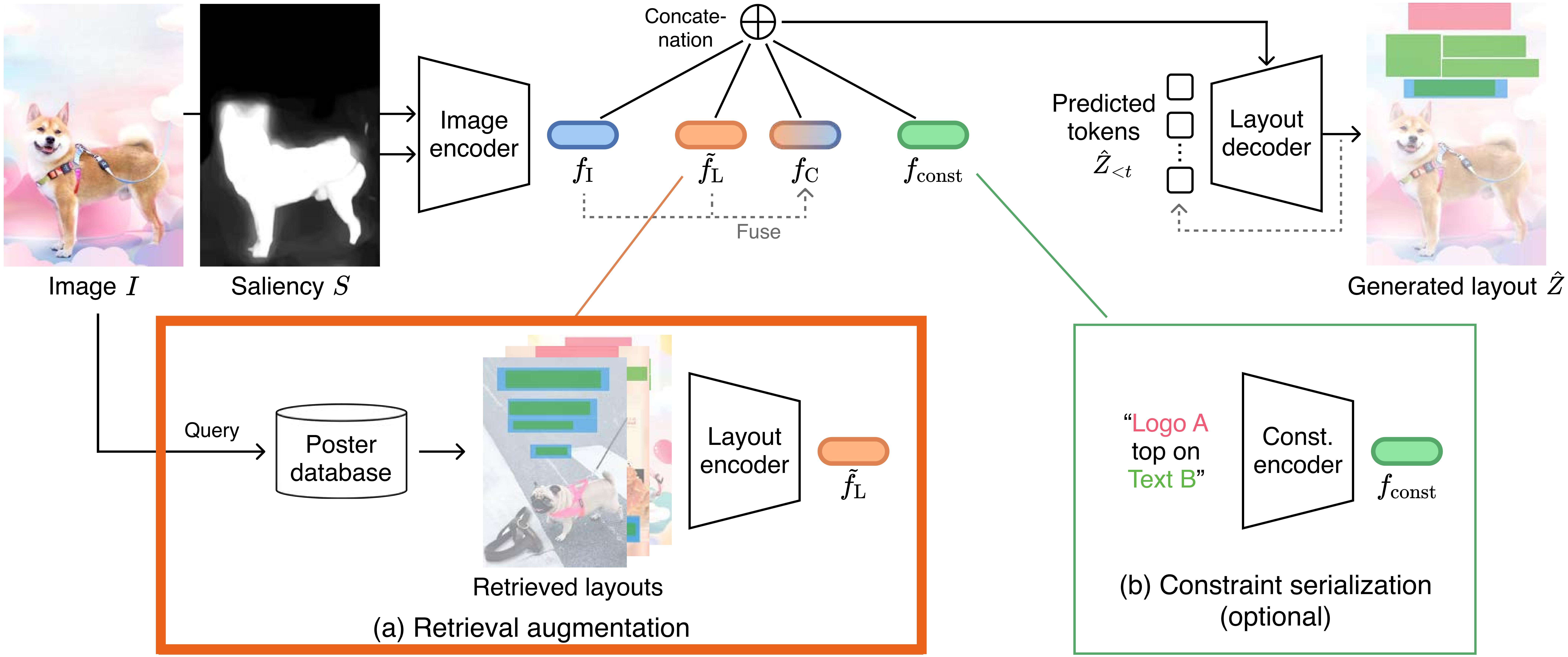
Overview of RALF

Encodes an input canvas image and a saliency map



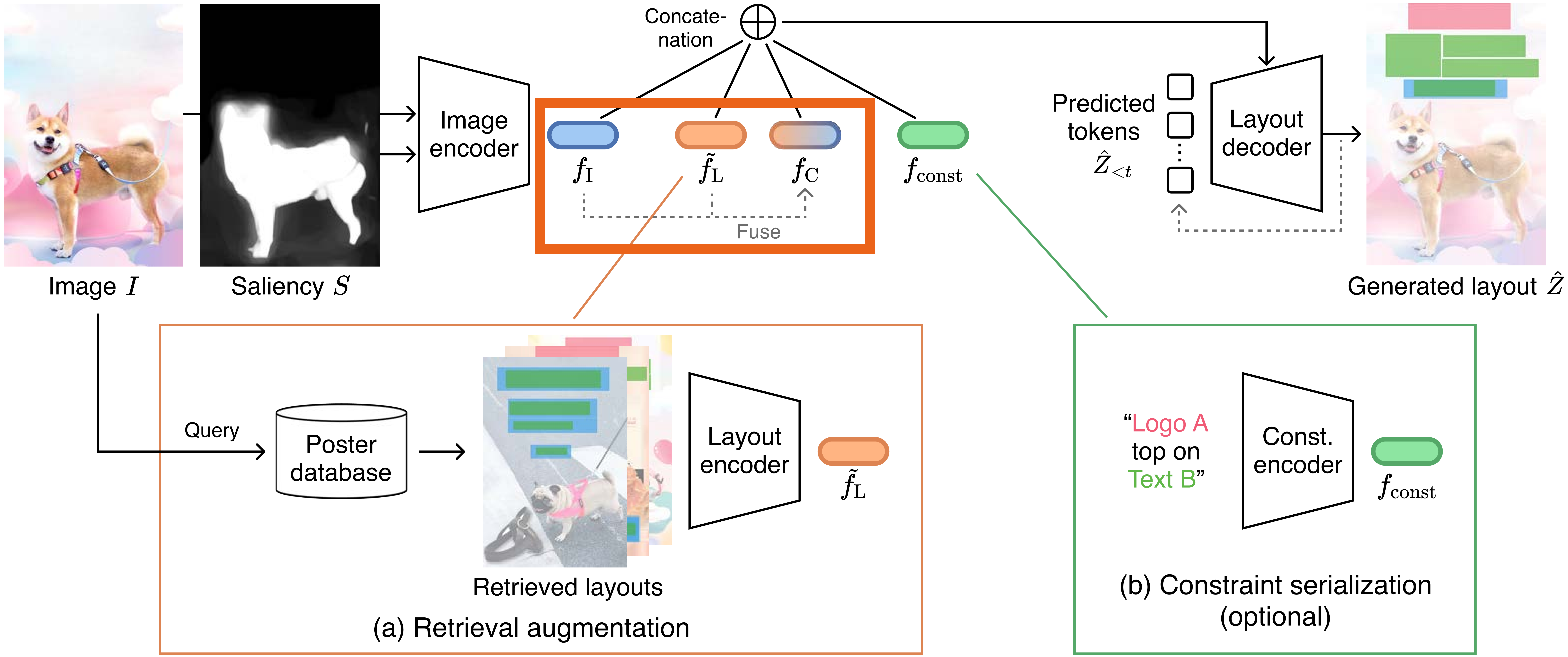
Overview of RALF

Retrieves nearest neighbor layout examples based on the similarity of an input image.



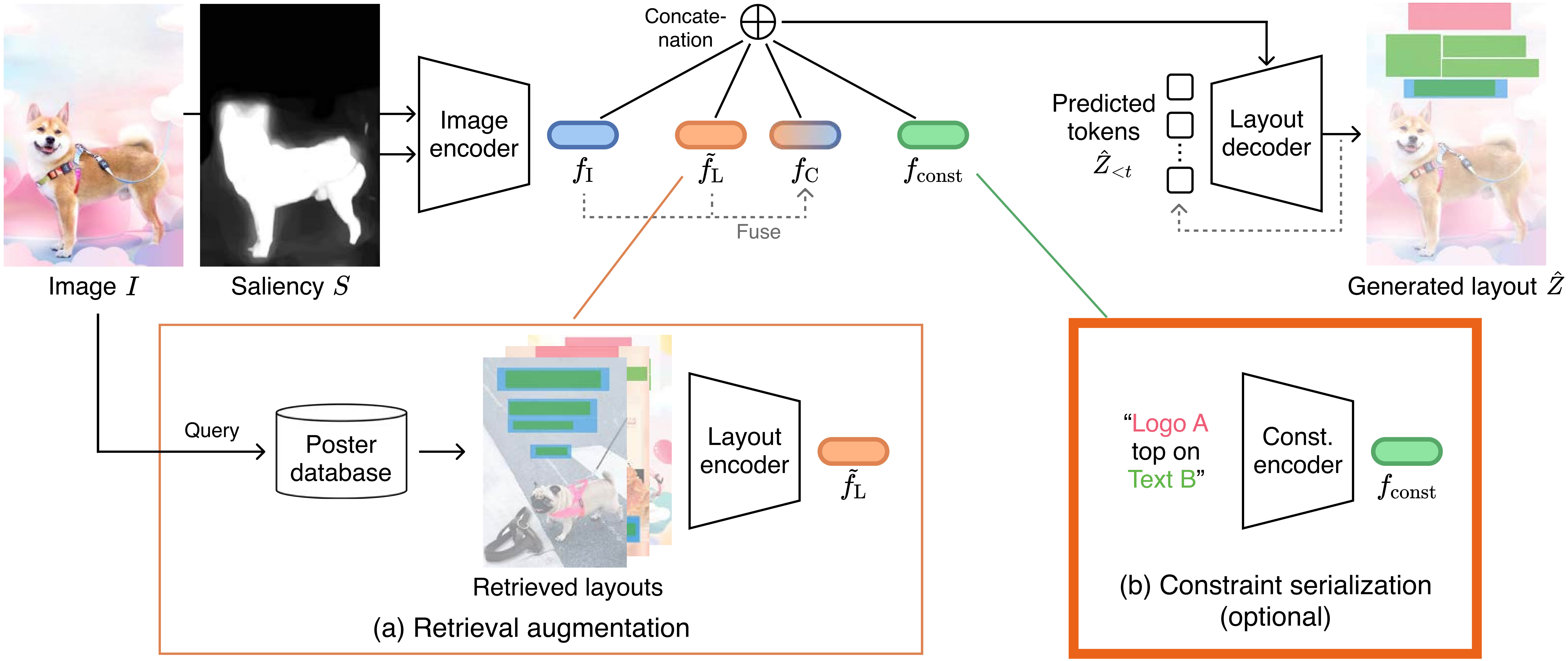
Overview of RALF

Fuses the features of retrieved layouts with the image feature using cross-attention.



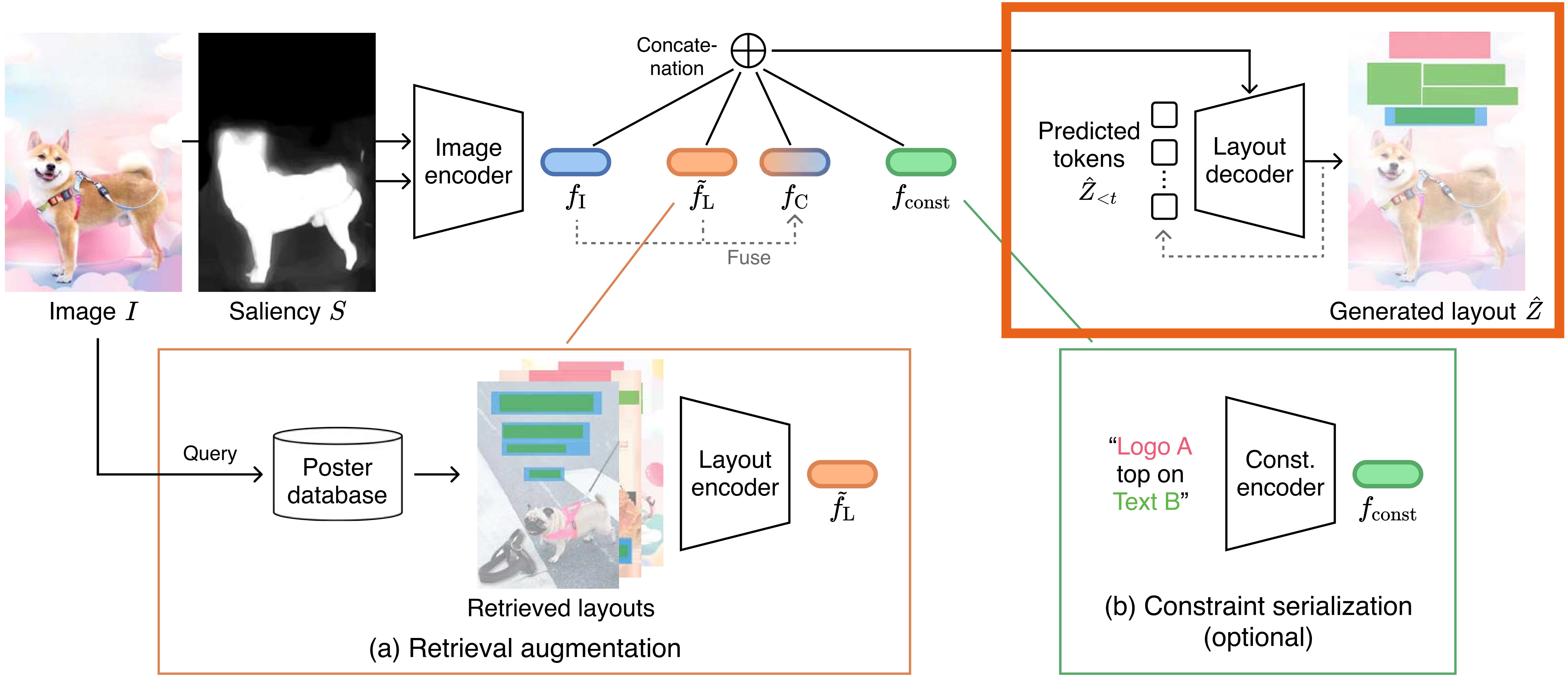
Overview of RALF

Incorporates user-specified constraints following LayoutFormer++ [Jiang+ CVPR23], which tokenizes constraints.



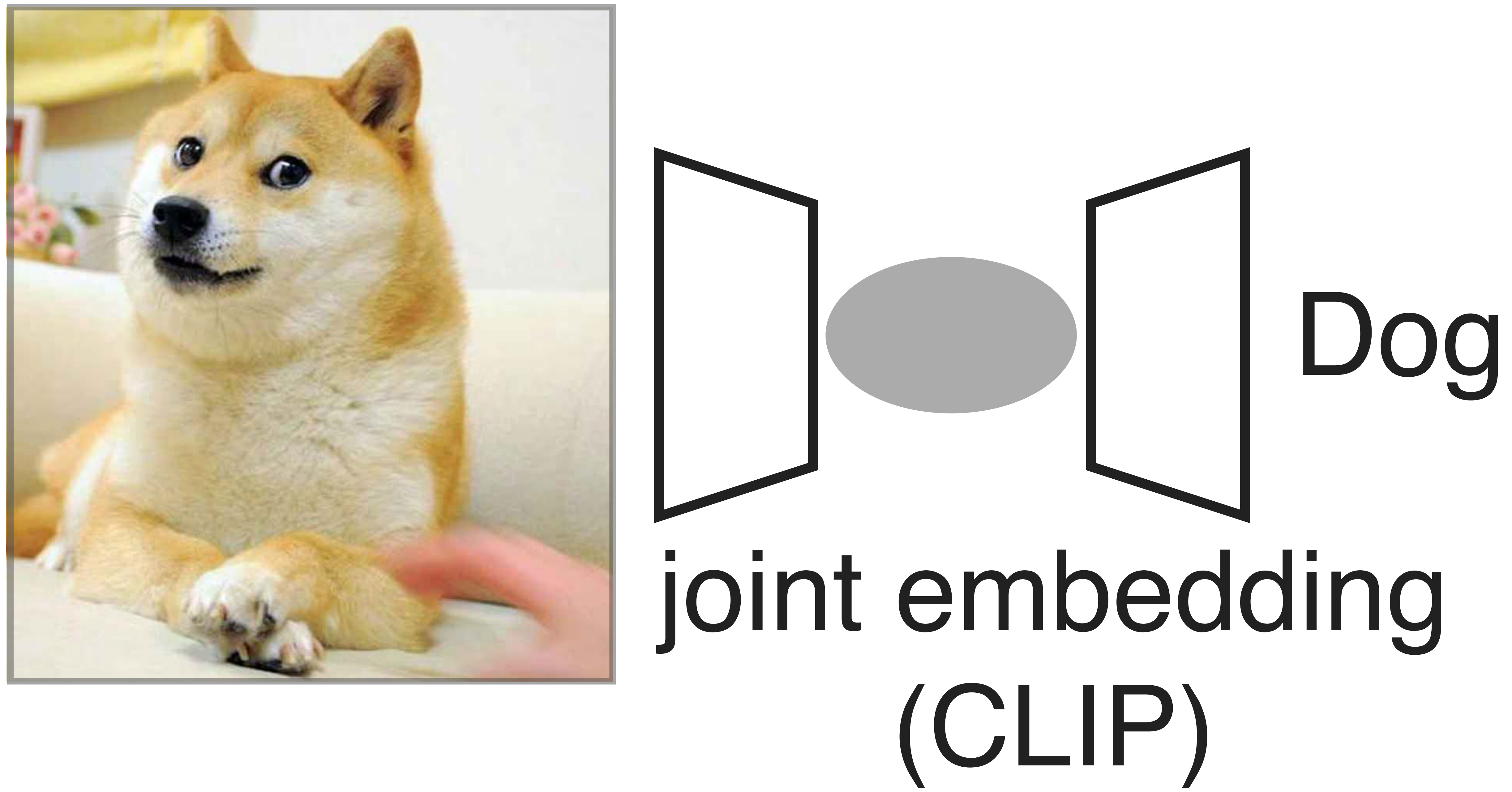
Overview of RALF

Autoregressively generates a layout.



Layout Retrieval

- A challenge lies in ***the absence of joint embedding for image–layout retrieval***, unlike CLIP for image—text retrieval.



No joint embedding!

image—text retrieval **image—layout retrieval**

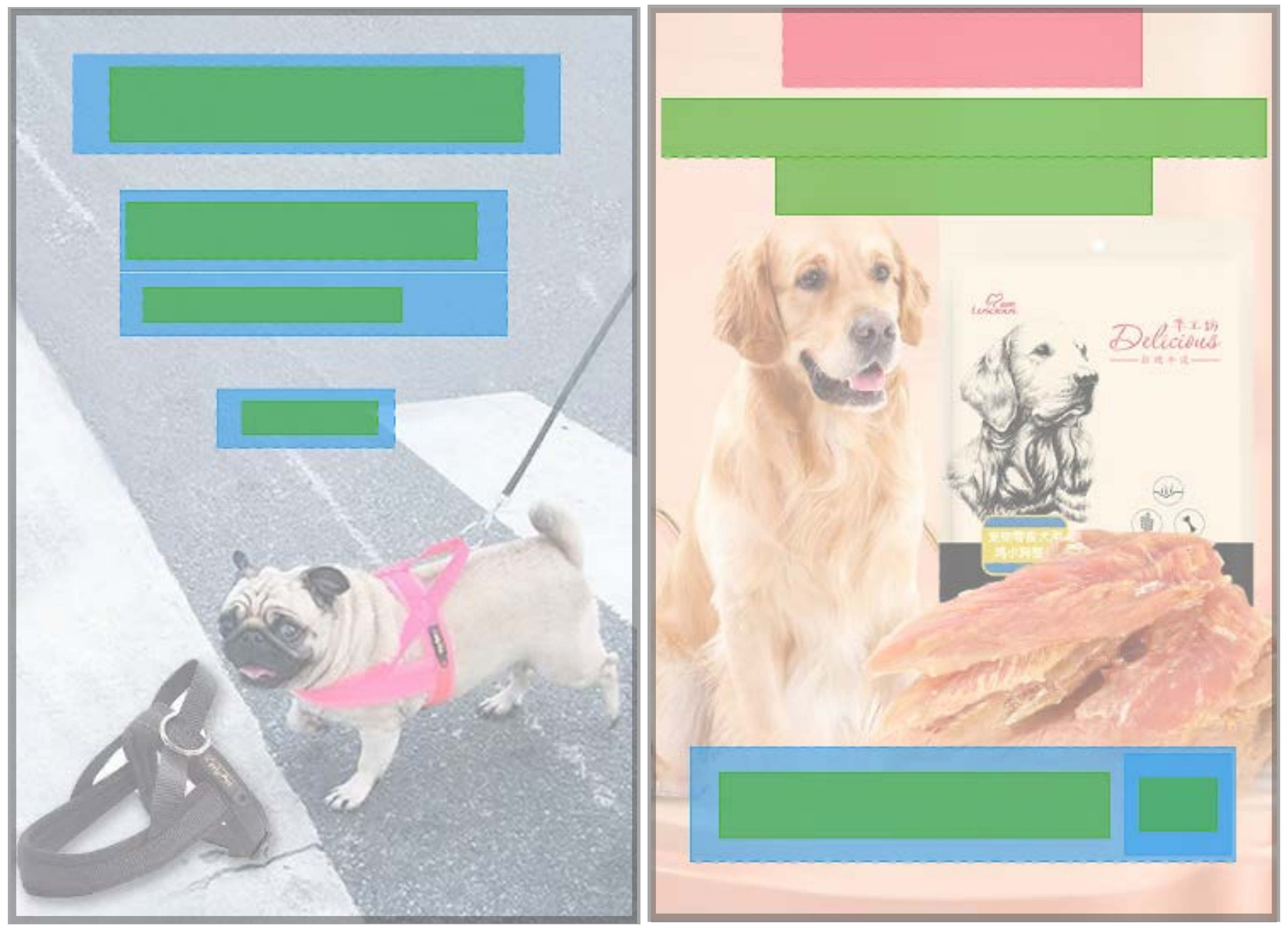
Layout Retrieval

- We hypothesize that ***given an image–layout pair $(\tilde{I}, \tilde{L})$, $\tilde{L}$ is more likely to be useful when $\tilde{I}$ is similar to I***.

GT layout

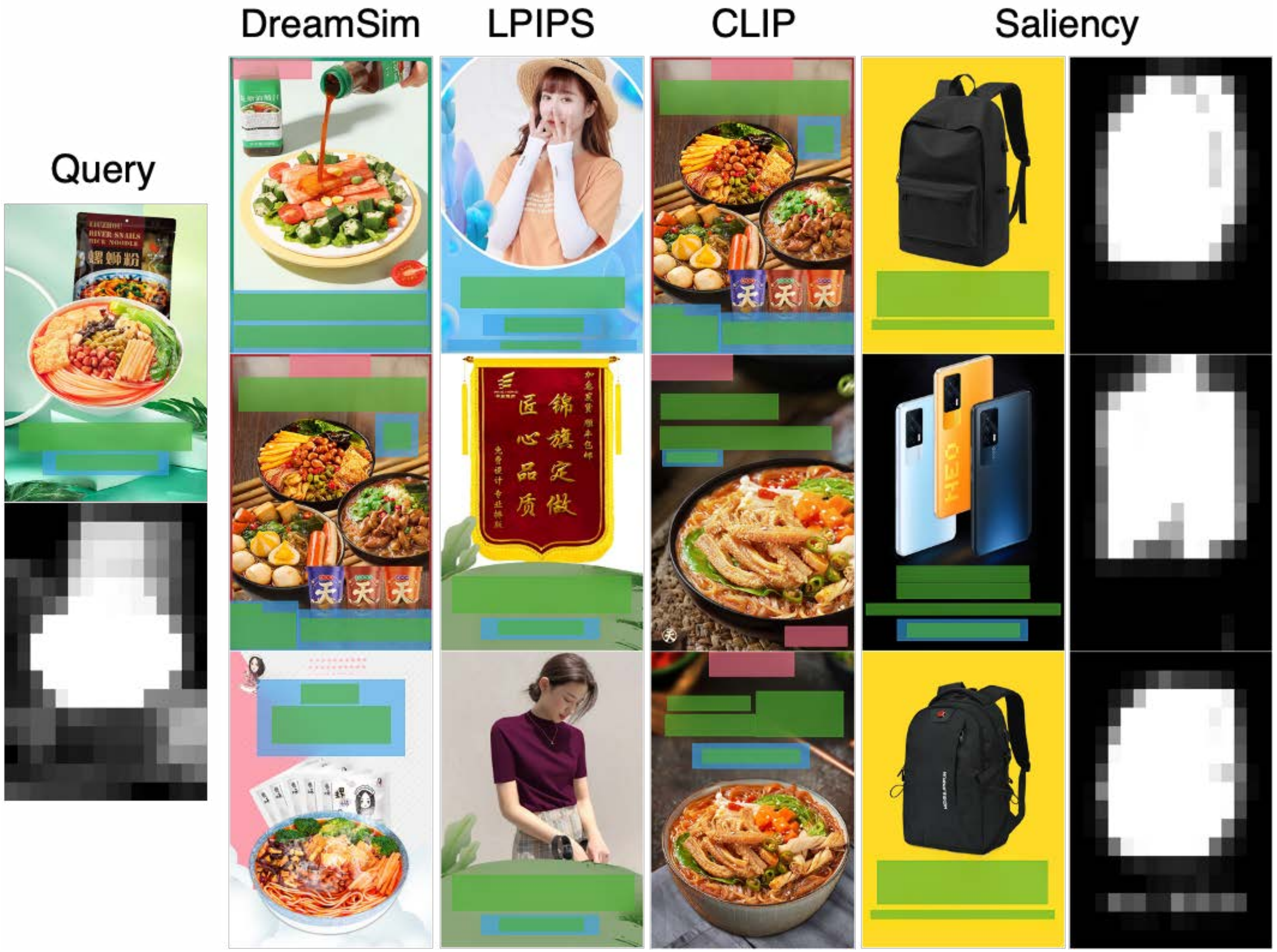
Image I (query)

Retrieve images $\tilde{I}$ using similarity, then use paired layout $\tilde{L}$

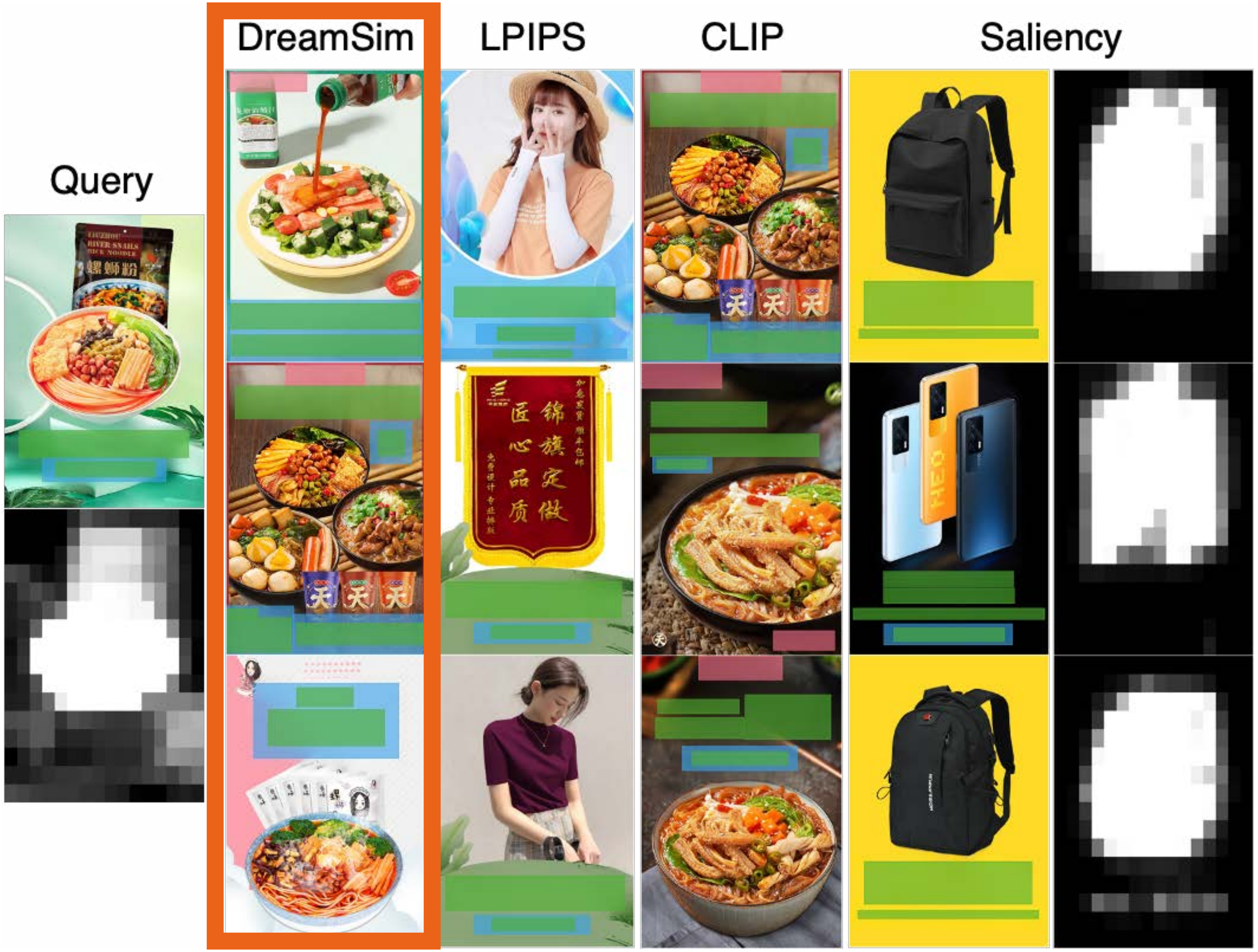


$(\tilde{I}_1, \tilde{L}_1)$ $(\tilde{I}_2, \tilde{L}_2)$

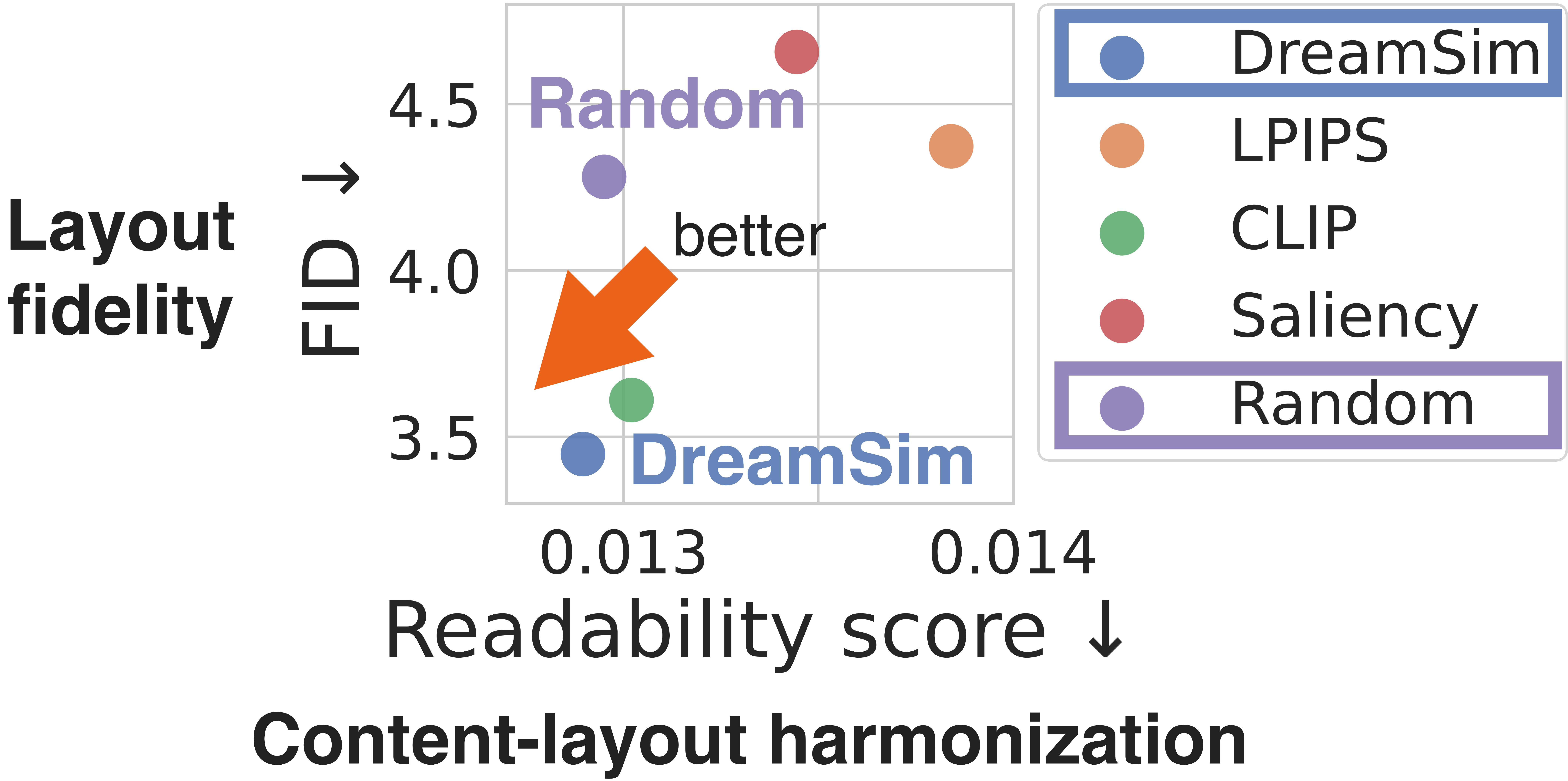
Layout Retrieval



Layout Retrieval

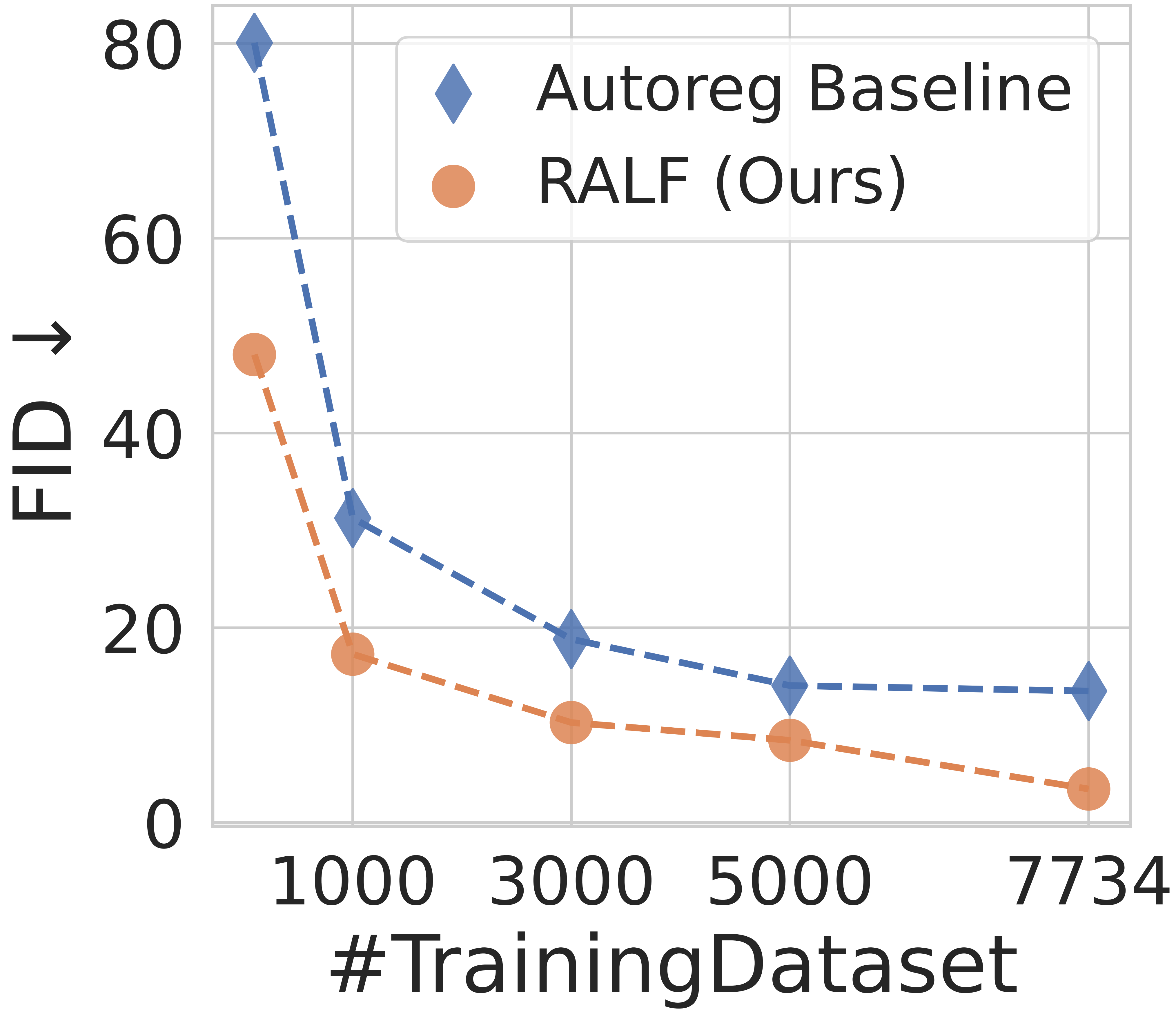


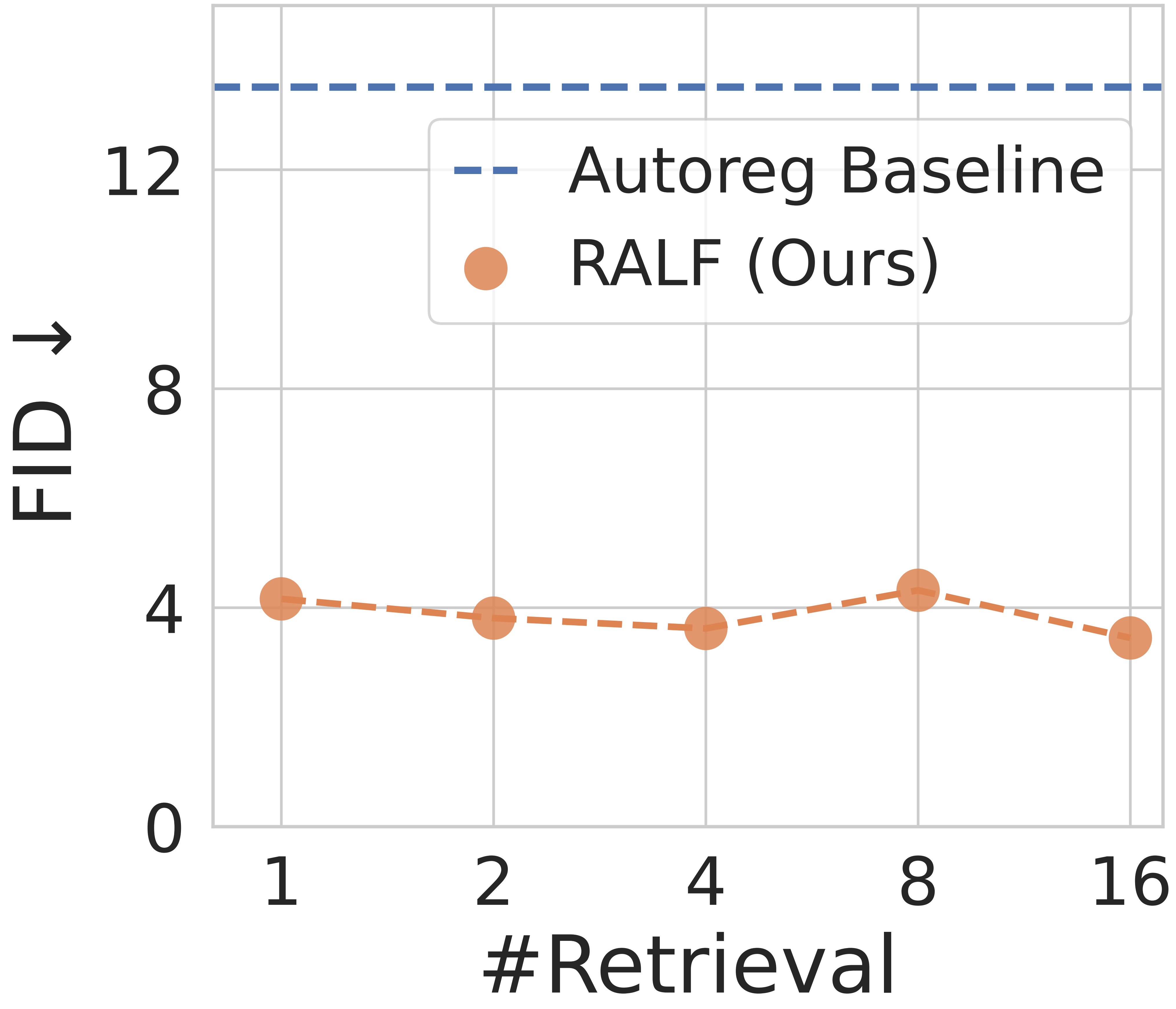
Layout Retrieval



Analysis

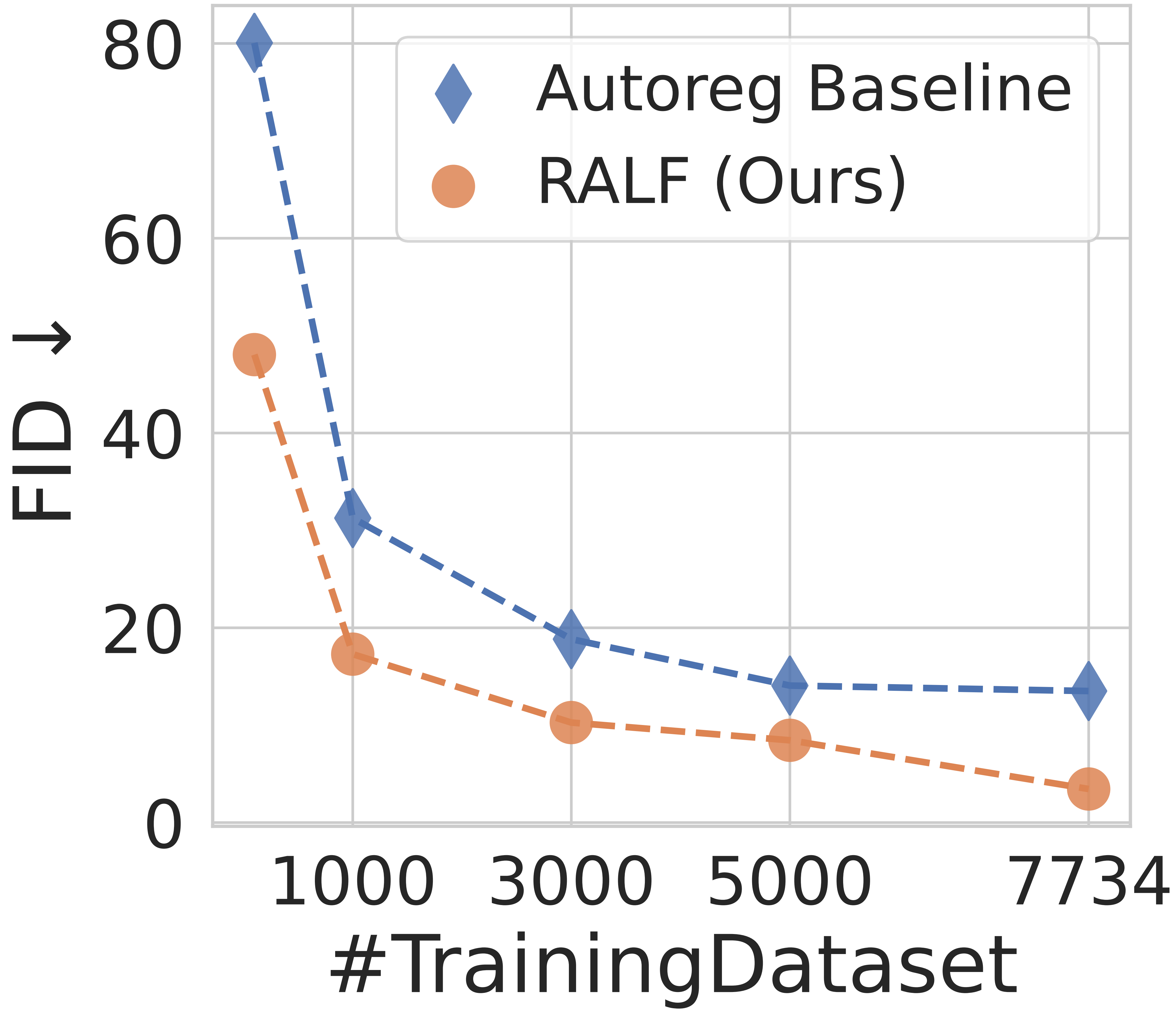
How effective RALF?

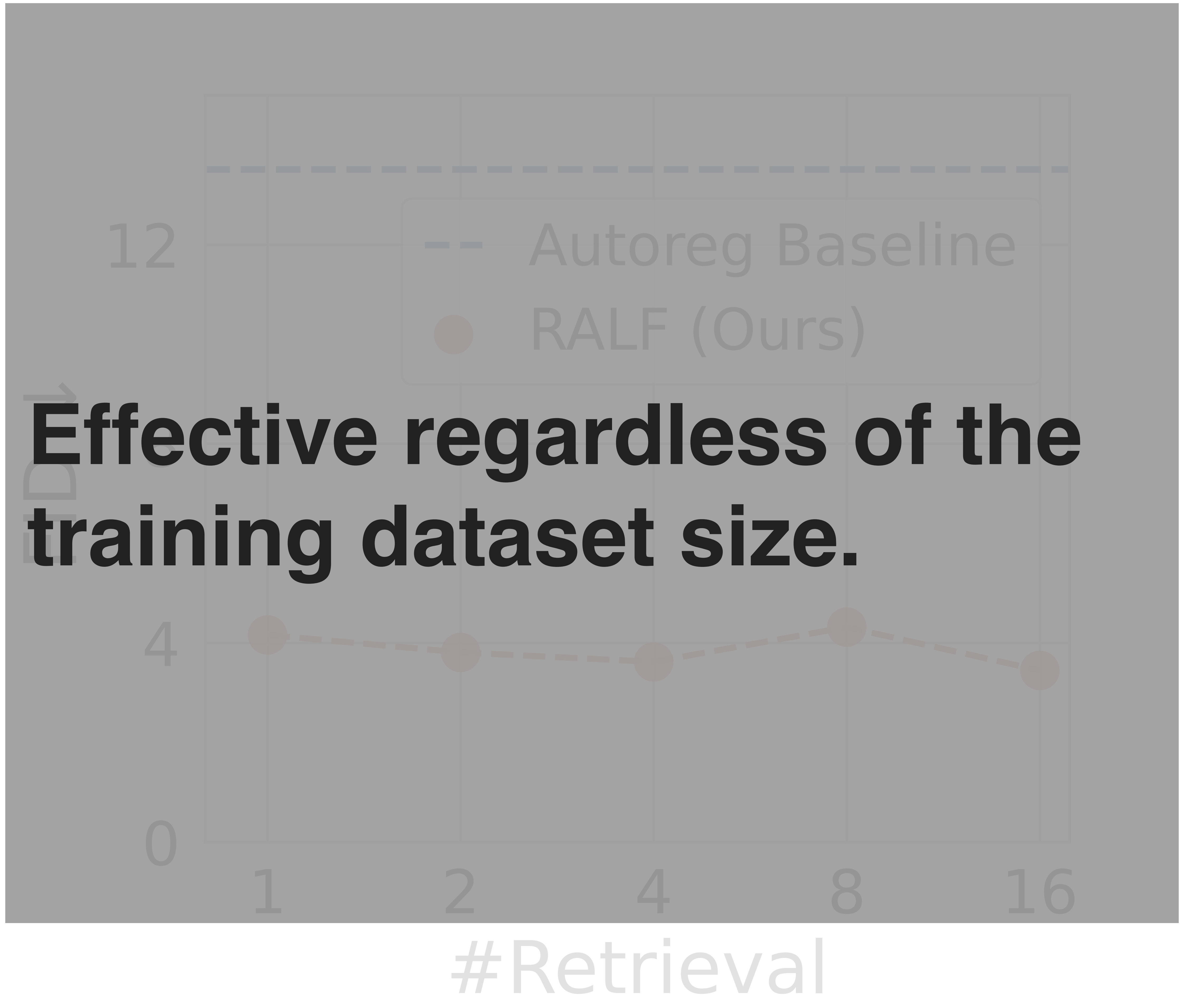




Analysis

How effective RALF?

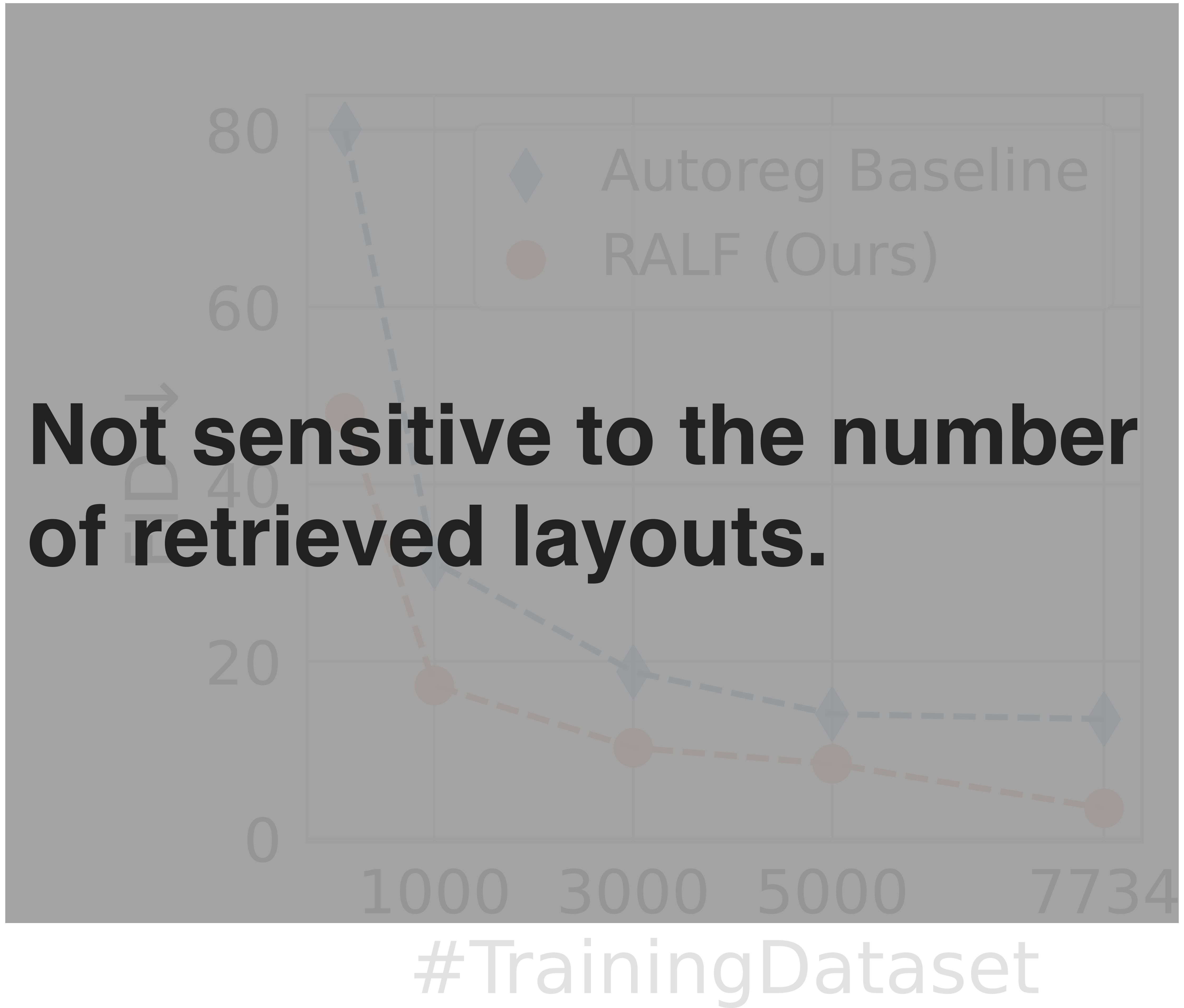




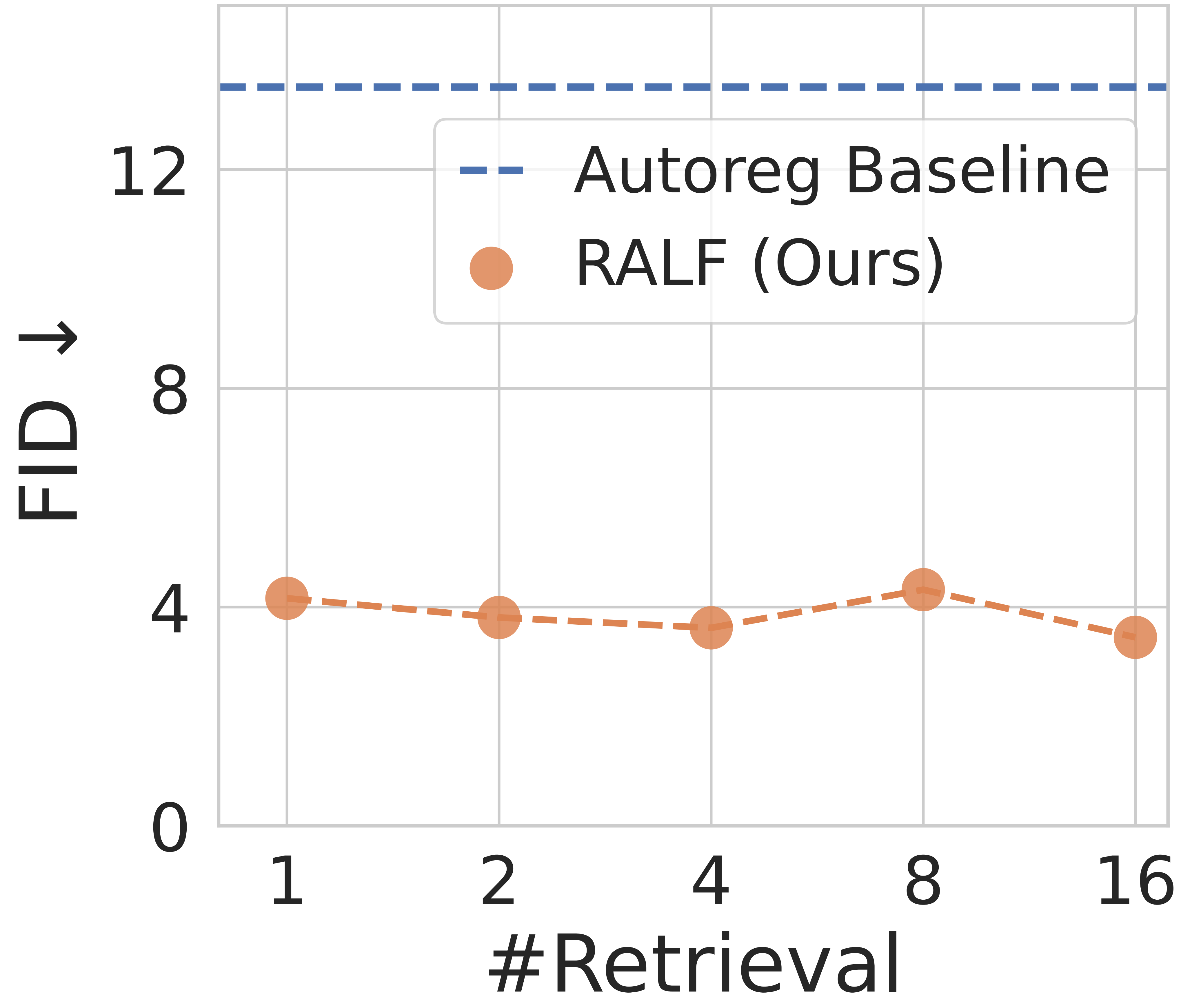
Effective regardless of the training dataset size.

Analysis

How effective RALF?



Not sensitive to the number of retrieved layouts.



Analysis

How different *K* affects the output? Compare *K*=1 with *K*=16

Similar results

***K*=1**

Reference

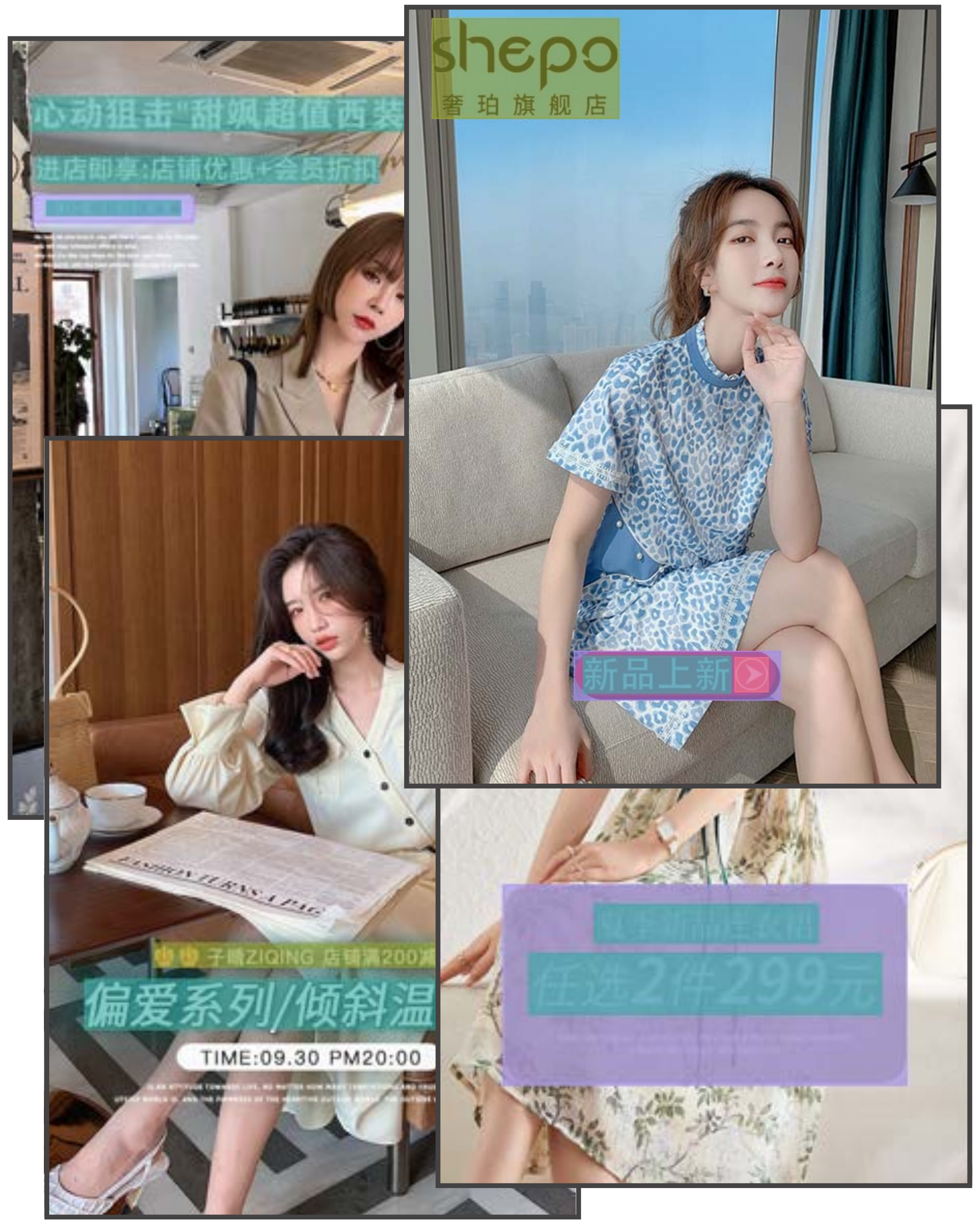
Output

Analysis

How different *K* affects the output? Compare *K*=1 with *K*=16

Diverse and plausible results.

***K*=16**

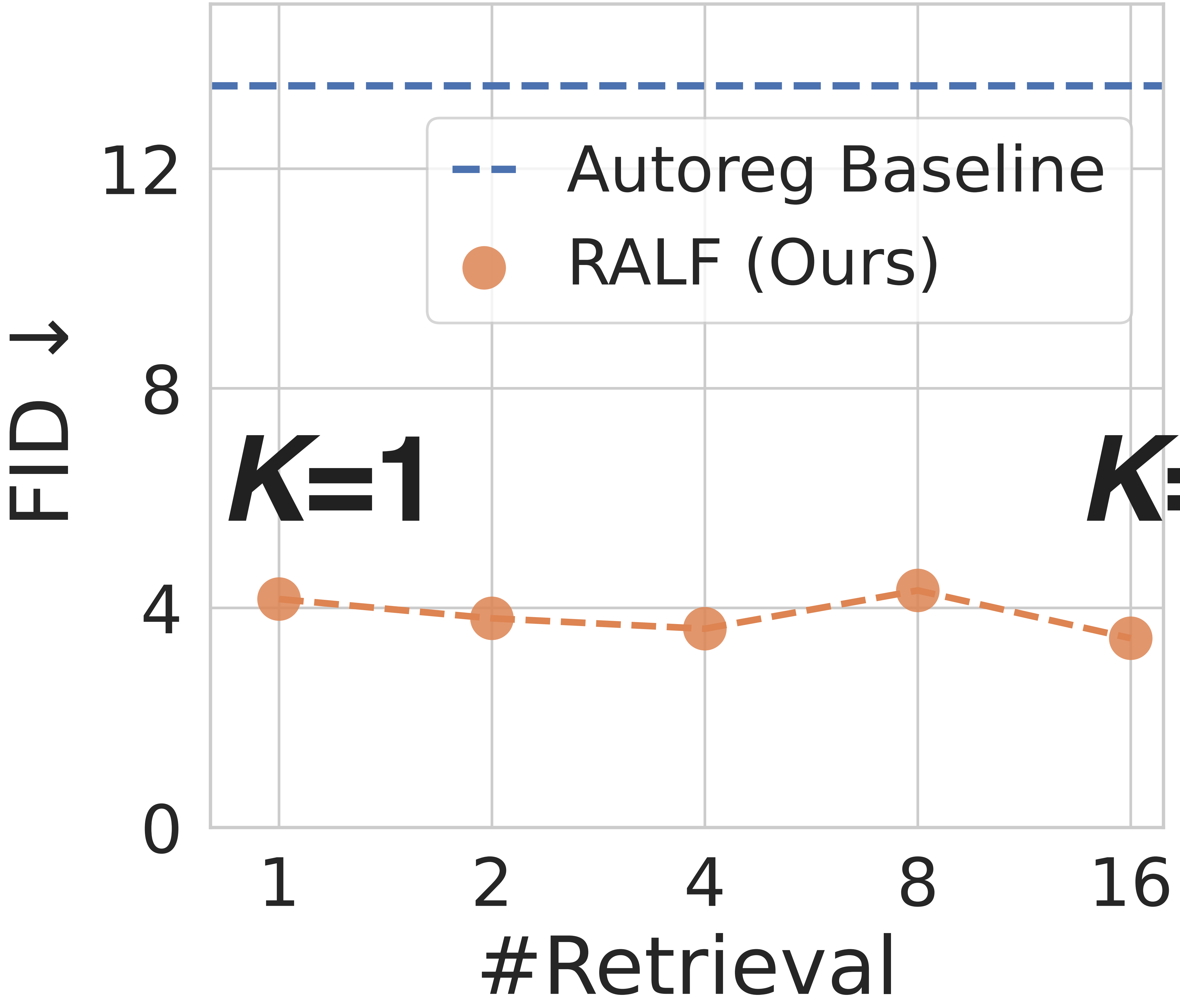


Reference

Output

Analysis

Limitation of FID.

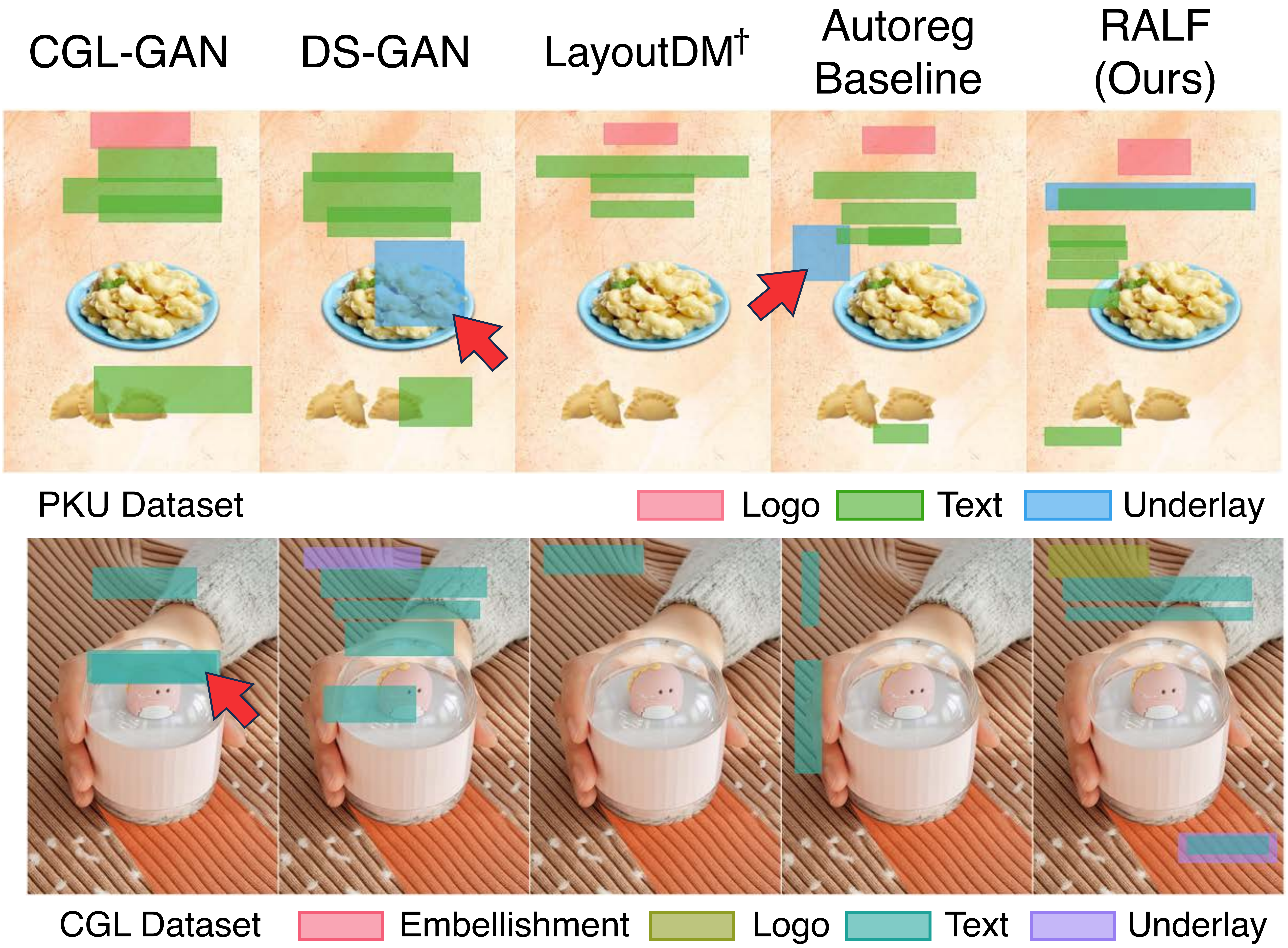


***K*=1**

***K*=16**

Same FID 🤔

Qualitative Results



Quantitative Results

Unconstrained generation results on PKU dataset **[Hsu+ CVPR23]**

Train:Test:Val = 7,734:1,000:1,000

Content: an overlap of saliency object and layout

Method	#Params	PKU				
		Content		Graphic		
		Occ ↓	Rea ↓	Und ↑	Ove ↓	FID↓
Real Data	-	0.112	0.0102	0.99	0.0009	1.58
Top-1 Retrieval	-	0.212	0.0218	0.99	0.002	1.43
CGL-GAN [53]	41M	0.138	0.0164	0.41	0.074	34.51
DS-GAN [18]	30M	0.142	0.0169	0.63	0.027	11.80
ICVT [7]	50M	0.146	0.0185	0.49	0.318	39.13
LayoutDM† [19]	43M	0.150	0.0192	0.41	0.190	27.09
Autoreg Baseline	41M	0.134	0.0164	0.43	0.019	13.59
RALF (Ours)	43M	**0.119**	**0.0128**	**0.92**	**0.008**	**3.45**

Quantitative Results

Real data is supposed to be upper bound

Validation data

	arams	PKU				
		Content		Graphic		
		Occ ↓	Rea ↓	Und ↑	Ove ↓	FID↓
Real Data	-	0.112	0.0102	0.99	0.0009	1.58
Top-1 Retrieval	-	0.212	0.0218	0.99	0.002	1.43
CGL-GAN [53]	41M	0.138	0.0164	0.41	0.074	34.51
DS-GAN [18]	30M	0.142	0.0169	0.63	0.027	11.80
ICVT [7]	50M	0.146	0.0185	0.49	0.318	39.13
LayoutDM† [19]	43M	0.150	0.0192	0.41	0.190	27.09
Autoreg Baseline	41M	0.134	0.0164	0.43	0.019	13.59
RALF (Ours)	43M	**0.119**	**0.0128**	**0.92**	**0.008**	**3.45**

Quantitative Results

Just top-1 retrieval is the worst in content metrics

“Retrieval-augmented” generation is important

Top-1 retrieved layout

		PKU				
		Content		Graphic		
		Occ ↓	Rea ↓	Und ↑	Ove ↓	FID↓
Real Data	-	0.112	0.0102	0.99	0.0009	1.58
Top-1 Retrieval	-	0.212	0.0218	0.99	0.002	1.43
CGL-GAN [53]	41M	0.138	0.0164	0.41	0.074	34.51
DS-GAN [18]	30M	0.142	0.0169	0.63	0.027	11.80
ICVT [7]	50M	0.146	0.0185	0.49	0.318	39.13
LayoutDM† [19]	43M	0.150	0.0192	0.41	0.190	27.09
Autoreg Baseline	41M	0.134	0.0164	0.43	0.019	13.59
RALF (Ours)	43M	**0.119**	**0.0128**	**0.92**	**0.008**	**3.45**

Quantitative Results

RALF significantly outperforms the baselines

Baseline methods

		PKU				
		Content		Graphic		
		Occ ↓	Rea ↓	Und ↑	Ove ↓	FID↓
Real Data	-	0.112	0.0102	0.99	0.0009	1.58
Top-1 Retrieval	-	0.212	0.0218	0.99	0.002	1.43
CGL-GAN [53]	41M	0.138	0.0164	0.41	0.074	34.51
DS-GAN [18]	30M	0.142	0.0169	0.63	0.027	11.80
ICVT [7]	50M	0.146	0.0185	0.49	0.318	39.13
LayoutDM† [19]	43M	0.150	0.0192	0.41	0.190	27.09
Autoreg Baseline	41M	0.134	0.0164	0.43	0.019	13.59
RALF (Ours)	43M	**0.119**	**0.0128**	**0.92**	**0.008**	**3.45**

Quantitative Results

Baseline methods + Retrieval augmentation

Method	Retrieval	Occ ↓	Rea ↓	Und ↑	Ove ↓	FID↓
CGL-GAN		**0.138**	**0.0164**	0.41	0.074	34.51
CGL-GAN	✓	0.144	**0.0164**	**0.63**	**0.039**	**13.28**
LayoutDM†		0.150	0.0192	0.41	0.190	27.09
LayoutDM†	✓	**0.123**	**0.0144**	**0.51**	**0.091**	**10.03**

Quantitative Results

Out-of-domain generalization

e.g. **Train / DB**: CGL dataset, **Test**: PKU dataset

Train	Test	Method	Occ ↓	Rea ↓	Und ↑	Ove ↓
CGL	PKU	Autoreg Baseline	0.176	0.0276	0.84	0.037
		RALF (Ours)	**0.144**	**0.0249**	**0.96**	**0.023**
PKU	CGL	Autoreg Baseline	0.341	0.0464	0.29	0.037
		RALF (Ours)	**0.286**	**0.0355**	**0.79**	**0.036**

Quantitative Results

Constrained generation

Category → Size + Position

Relationship

Method	PKU				
	Content		Graphic		
	Occ ↓	Rea ↓	Und ↑	Ove ↓	FID↓
$\mathbf{C \rightarrow S + P}$					
CGL-GAN	0.132	0.0158	0.48	0.038	11.47
LayoutDM†	0.152	0.0201	0.46	0.172	20.56
Autoreg Baseline	0.135	0.0167	0.43	0.028	10.48
RALF (Ours)	**0.124**	**0.0138**	**0.90**	**0.010**	**2.21**

Relationship					
Autoreg Baseline	0.140	0.0177	0.44	0.028	10.61
RALF (Ours)	**0.122**	**0.0141**	**0.85**	**0.009**	**2.23**

Conclusion

Thank you!

- **Retrieval augmentation effectively addresses the data scarcity problem.**
- **Propose RALF**: Retrieval-augmented Layout Transformer
  - Retrieval augmentation + Autoregressive Transformer.
- **Show that RALF successfully generates high-quality layouts**, significantly outperforming baselines.

Contact:
horita@hal.t.u-tokyo.ac.jp

@udoooom