



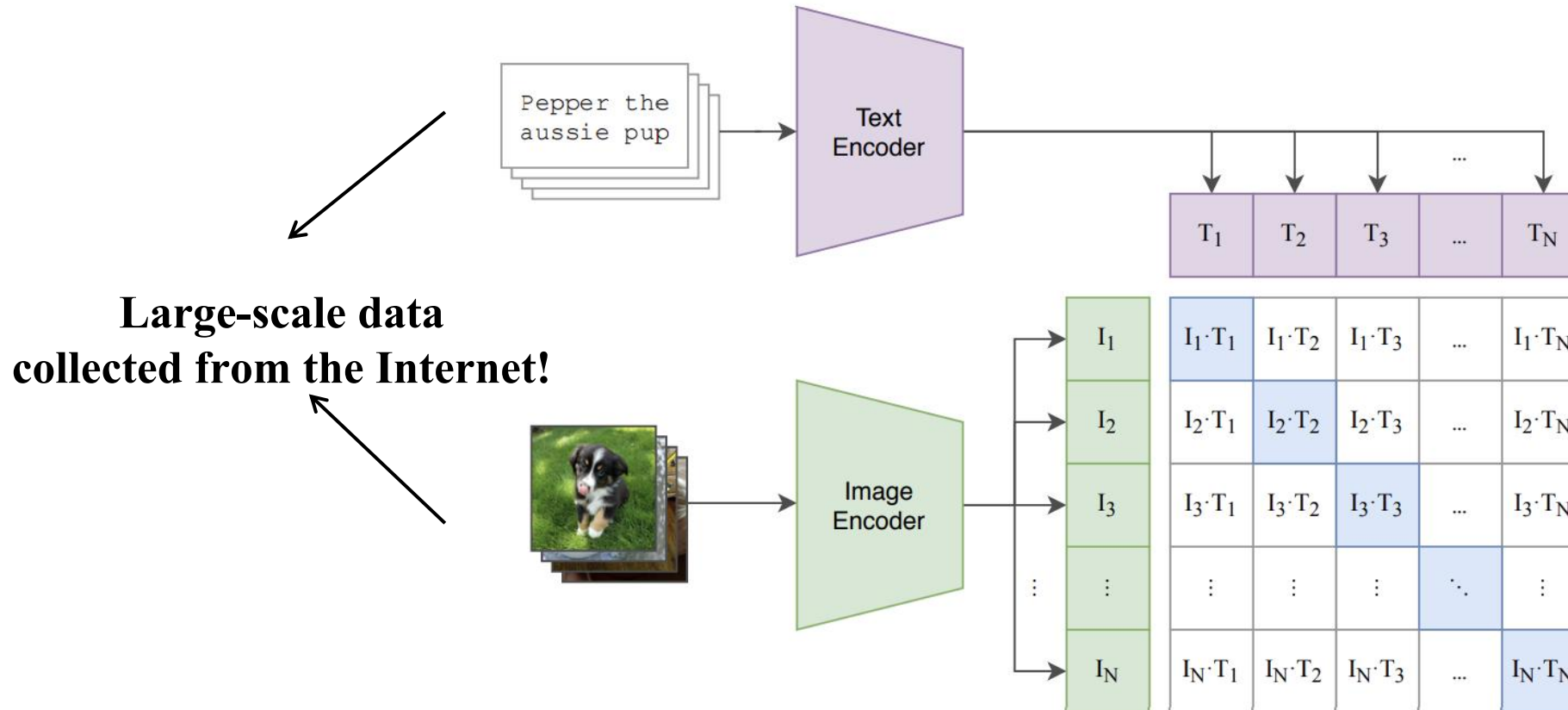
EUROPEAN
CONFERENCE
ON COMPUTER
VISION

Parrot Captions Teach CLIP to Spot Text

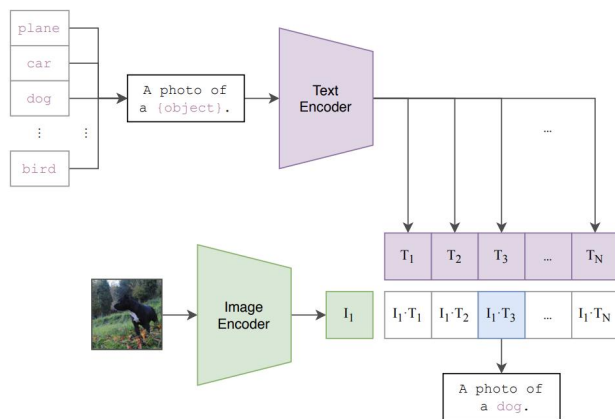
Yiqi Lin^{1,2*}, Conghui He^{1*†}, Alex Jinpeng Wang^{2*}, Bin Wang^{1*},
Weijia Li³, Mike Zheng Shou²



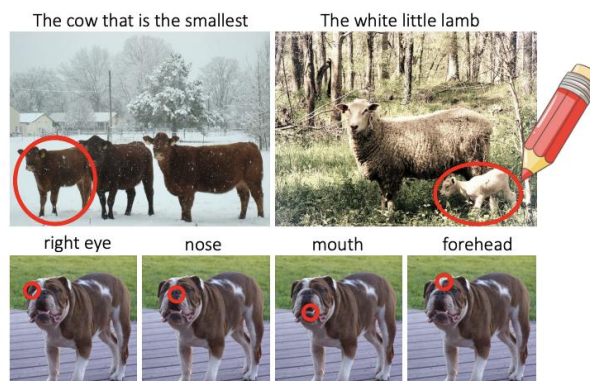
- **Contrastive Language-Image Pre-Training (CLIP)**



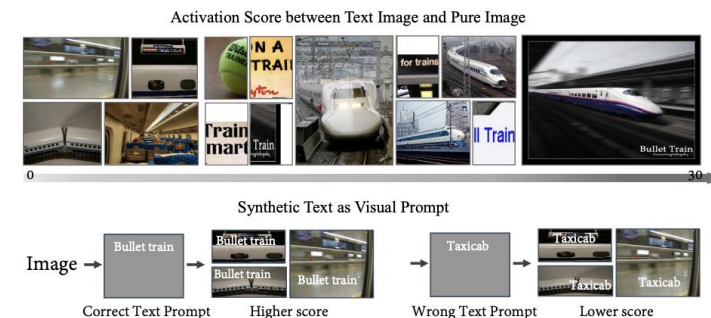
- Emergent Abilities in CLIP



Zero-Shot Classification
[Radford et al. ICML'21]



Red Circle
[Shtedritski et al. ICCV'23]



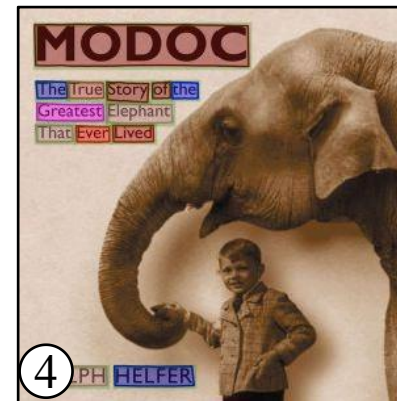
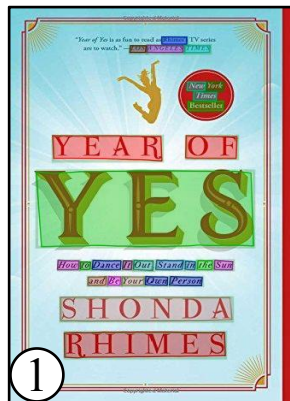
Text Spotting (OCR)
[Shi et al. ICCV'23]



What enables this ability?

- **Let's Look at Some Data!**

Top5% CLIP Score in LAION-2B



- 1). Year of Yes: How to Dance It Out, Stand In the Sun and Be Your Own Person by Shonda Rhimes.
- 2). National Association of Student Financial Aid Administrators Presents 2015 NASFAA What You Need to Know About Financial Aid.
- 3). Kids Again (feat. Amy Allen) by Artist Vs Poet.
- 4). Modoc: The True Story of the Greatest Elephant That Ever Lived, Ralph Helfer.
- 5). Everton Mints 150g Jar.

Parrot captions

➤ **Does CLIP Simply Parroting Text in Images?**

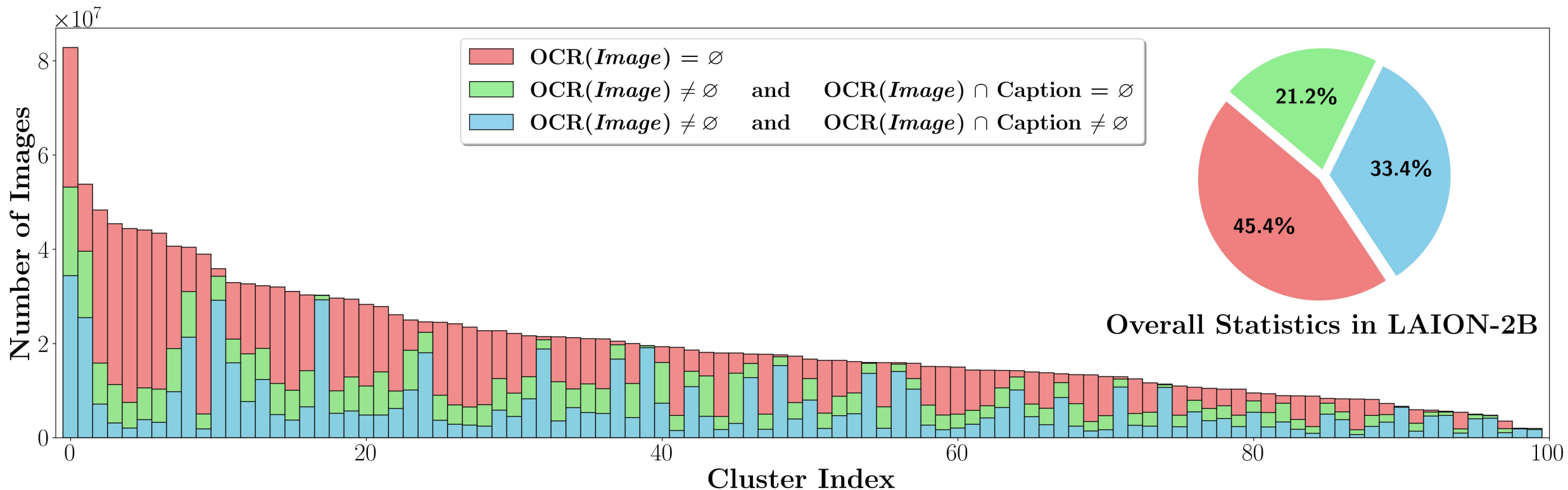
- **Profiling LAION-2B Data**

- Divide LAION 2B into 4000 clusters.
- Top CLIP score samples from 50 clusters.



Posters, Book Covers, Advertisements, and even Slides!

• Profiling LAION-2B Data

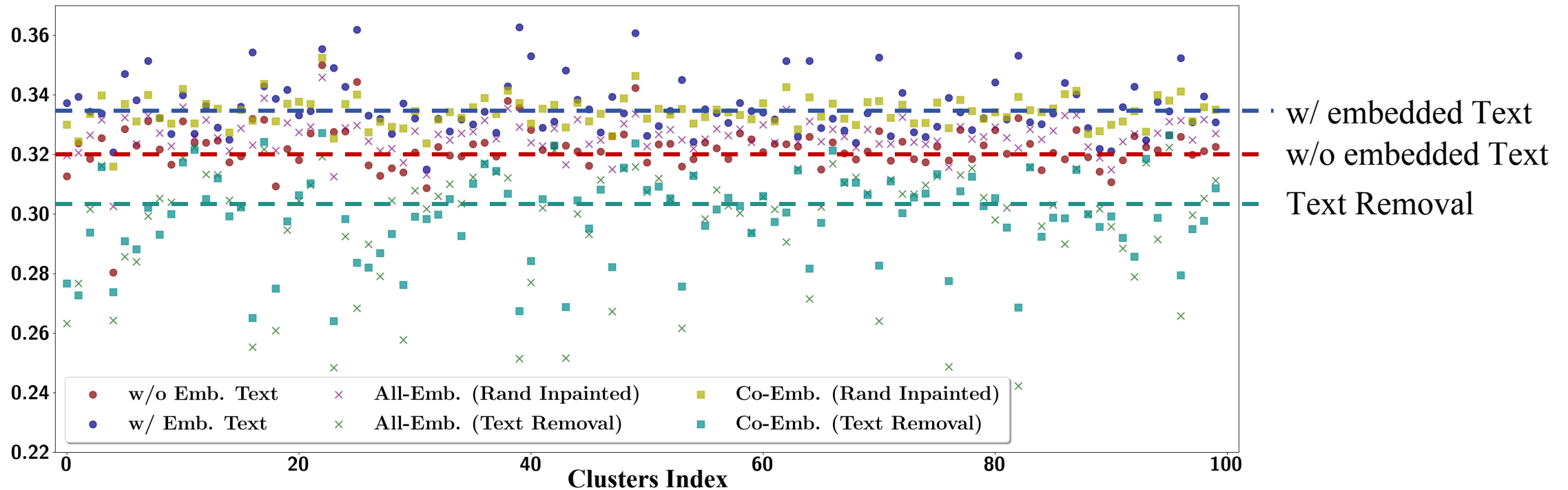
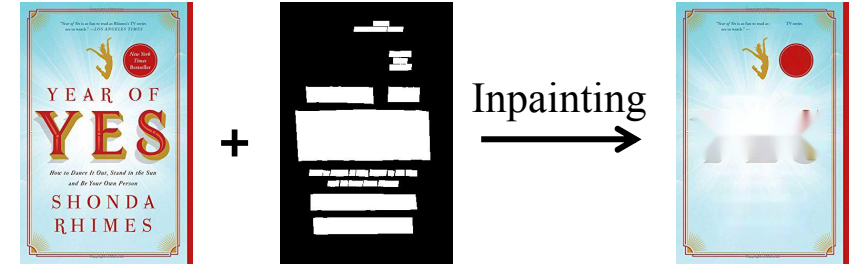


- **54.6%** of images contain embedded text.
- **33.4%** of captions have words that overlap with embedded text.

NOTE: LAION-2B dataset collection uses OpenAI's CLIP score > 0.28 as filtering!

Inspecting Pre-Trained CLIP Models

- Text Removal by inpainting.
- CLIP score difference before and after Text Removal.



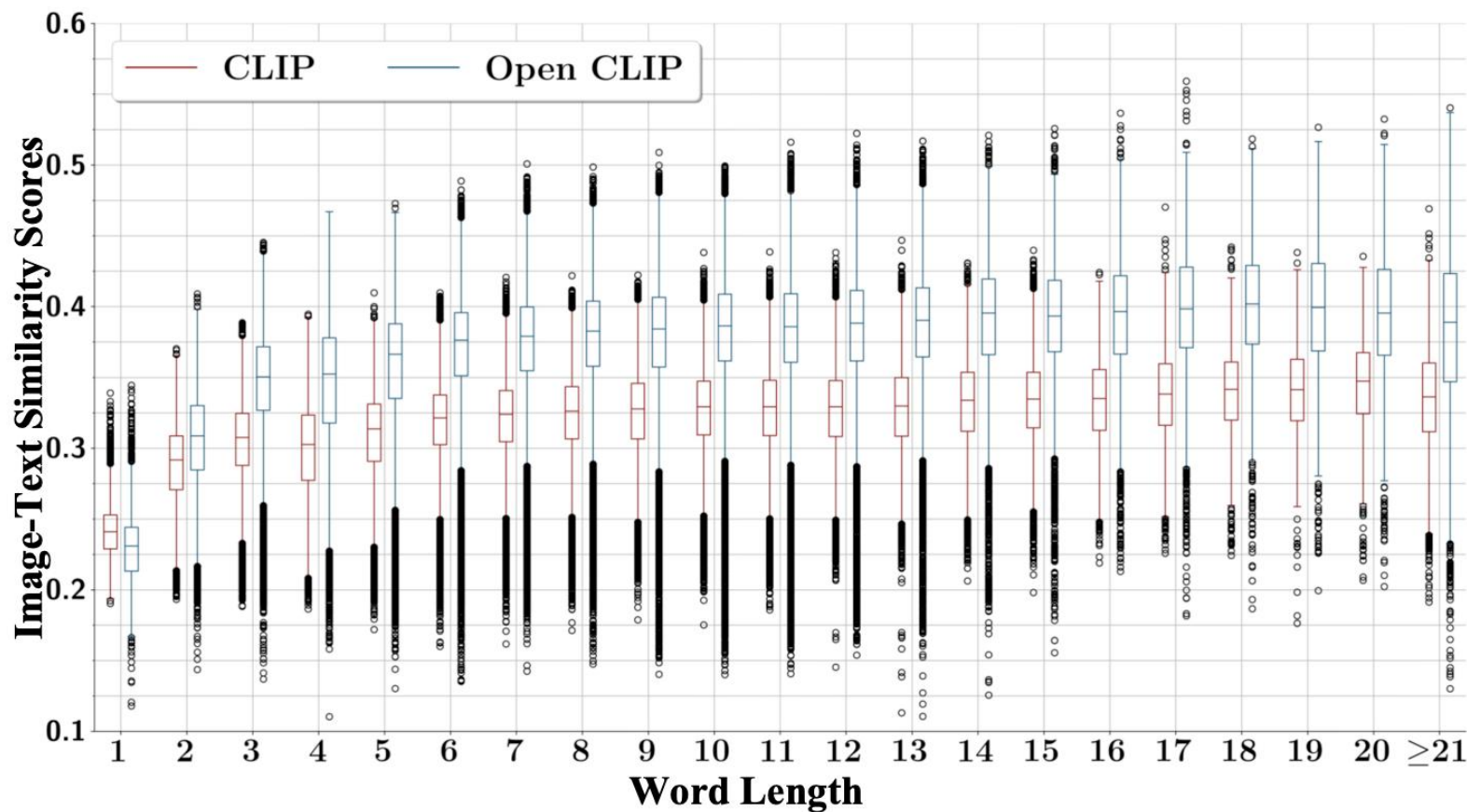
- Images with embedded text usually get higher CLIP scores.
- Embedded text can dominate the CLIP score measurement.

- **Inspecting Pre-Trained CLIP Models**



CLIP Score
↔ “Architectural”

- CLIP scores of synthetic single-word images.

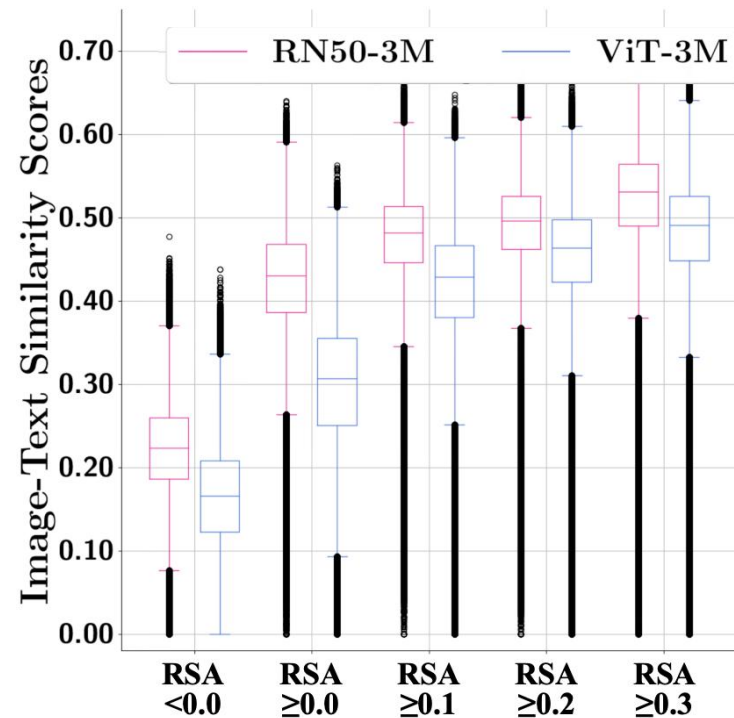


- OpenCLIP (LAION-2B) model has a stronger text spotting ability than OpenAI’s CLIP (WIT 400M).

- **Training on Embedded Text Curated Data**

- RSA: CLIP Score(Raw Image) - CLIP Score(Text Removal Image).
- Evaluate the Zero-shot DataComp benchmark and Synthetic Image CLIP score.

Data (3M)	Model	Avg.S(●)	IN	Ret.	Avg.
RSA < 0.0	RN50	0.319	0.181	0.220	0.239
RSA ≥ 0.0	RN50	0.339	0.126	0.180	0.215
RSA ≥ 0.1	RN50	0.351	0.041	0.123	0.148
RSA ≥ 0.2	RN50	0.360	0.017	0.094	0.109
RSA ≥ 0.3	RN50	0.376	0.009	0.075	0.097
RSA < 0.0	ViT-B	0.319	0.123	0.159	0.198
RSA ≥ 0.0	ViT-B	0.339	0.079	0.129	0.185
RSA ≥ 0.1	ViT-B	0.351	0.031	0.103	0.134
RSA ≥ 0.2	ViT-B	0.360	0.012	0.080	0.103
RSA ≥ 0.3	ViT-B	0.376	0.006	0.070	0.096



- Images with embedded text (parrot captions) generally **reduce** dataset quality.
- The model learns **stronger text spotting ability** with **more biased data** without learning vision-language semantics.

- **Training on Embedded Text Curated Data**

- Finetuning pre-trained BLIP on downstream general and text-orient tasks(**gray color**).

BLIP Data (3M)	Visual Question Answering (Acc)			Image Captioning (CIDEr)		Text-to-Image Retrieval (R@1)		Image-to-Text Retrieval (R@1)	
	VQAv2	TextVQA	ST-VQA	COCO	TextCaps	COCO	TextCaps	COCO	TextCaps
RSA < 0.0	70.79	14.16	9.64	115.7	44.9	48.25	36.85	64.72	54.7
RSA ≥ 0.0	70.03	18.76	11.81	111.9	84.5	46.25	68.61	62.92	81.23
RSA ≥ 0.1	68.14	19.48	13.33	105.6	96.1	39.96	68.13	54.64	79.37
RSA ≥ 0.2	66.01	21.06	11.85	98.7	94.4	33.03	64.17	47.12	75.33
RSA ≥ 0.3	64.20	18.44	12.04	95.26	91.1	26.64	60.11	37.3	70.24

- Parrot captions can benefit the text-orient downstream tasks while requiring careful data mixing trade-off.

- **Takeaways**

- LAION-2B dataset has a significant bias towards text spotting.
- Released CLIP models exhibit a strong text spotting bias, which makes CLIP-filtered datasets inherently biased.
- CLIP models can easily learn text spotting ability while failing to connect the vision-language semantics.



Thank you!

Paper: <https://arxiv.org/pdf/2312.14232>

Project: <https://linyq17.github.io/CLIP-Parrot-Bias/>