CLIP and Image Generation

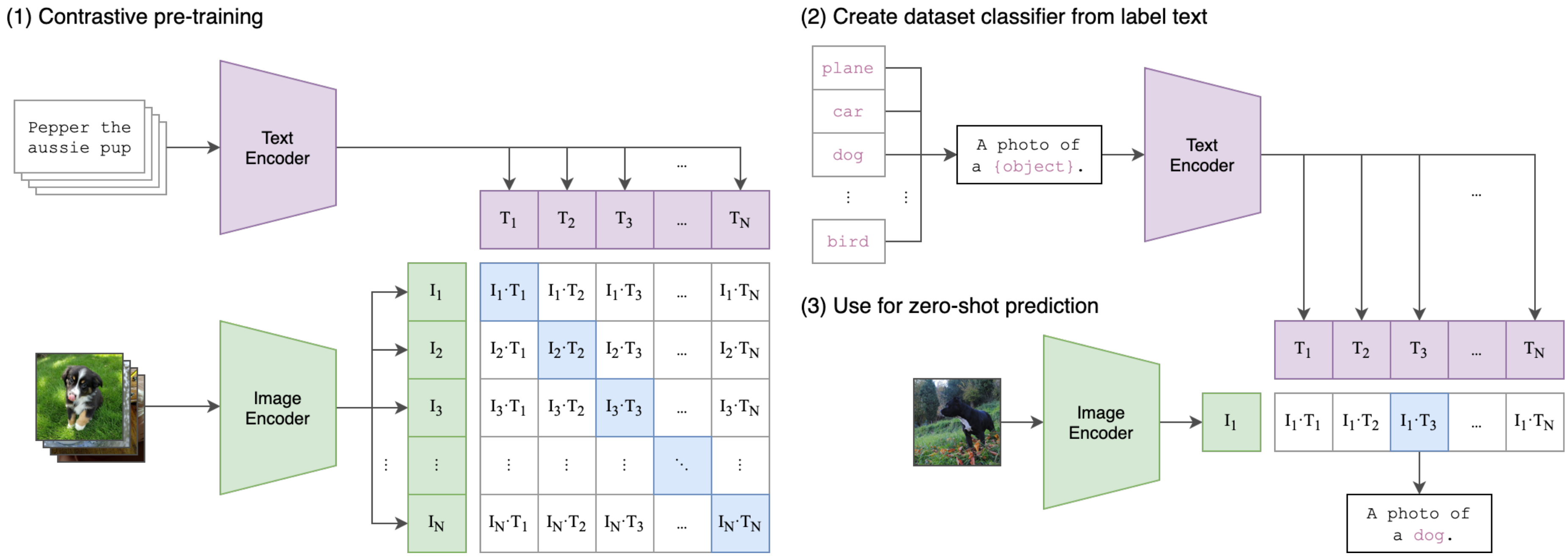
MENG Xiangqiao 22041201r

Motivation of CLIP

- What problems exist in the research of CV?
  - Labeling datasets is labor-intensive and expensive;
  - General visual network is hard to migrate to a new task;
  - Poor generalization ability.

- What did OpenAI do?
  - Bring abstract concepts in NLP to CV;
  - 400M dataset;

CLIP Contrastive Language-Image Pre-Training



(1) Contrastive pre-training

(2) Create dataset classifier from label text

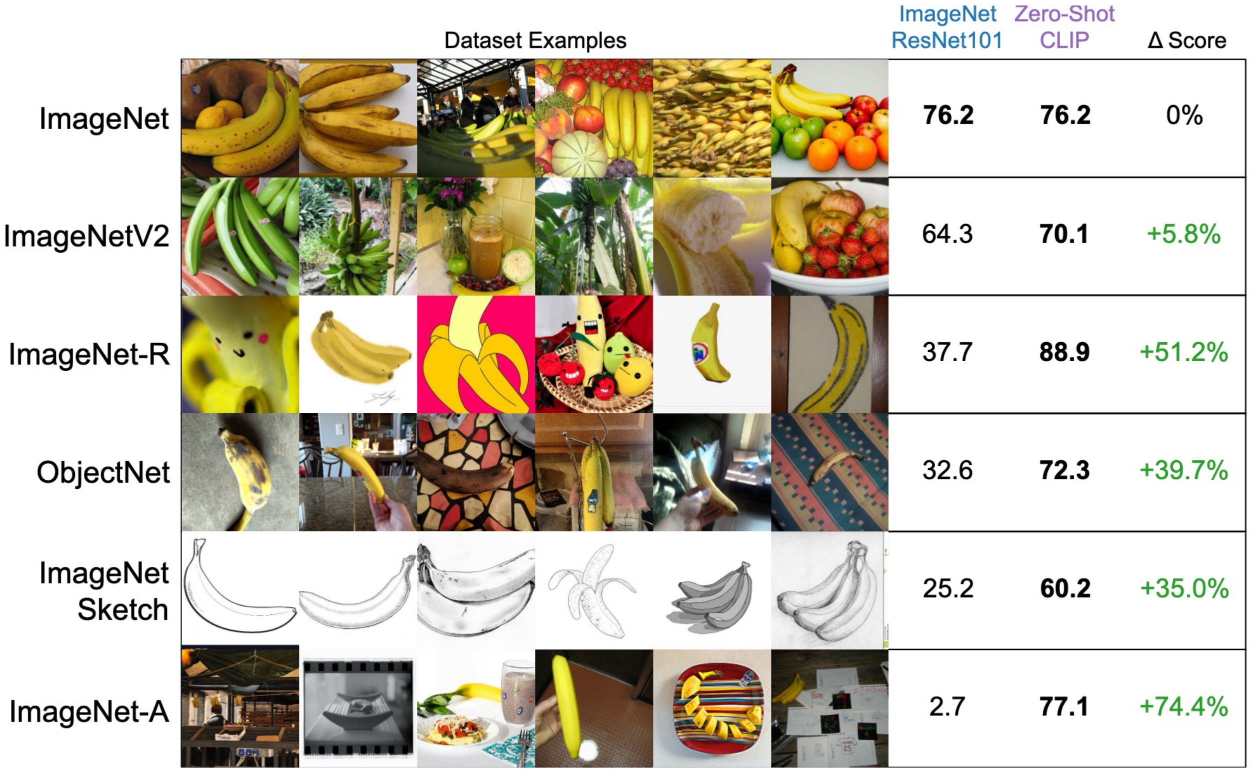
(3) Use for zero-shot prediction

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$
I_2	$I_2 \cdot T_1$	$I_2 \cdot T_2$	$I_2 \cdot T_3$	...	$I_2 \cdot T_N$
I_3	$I_3 \cdot T_1$	$I_3 \cdot T_2$	$I_3 \cdot T_3$	...	$I_3 \cdot T_N$
⋮	⋮	⋮	⋮	⋱	⋮
I_N	$I_N \cdot T_1$	$I_N \cdot T_2$	$I_N \cdot T_3$	...	$I_N \cdot T_N$

	T_1	T_2	T_3	...	T_N
I_1	$I_1 \cdot T_1$	$I_1 \cdot T_2$	$I_1 \cdot T_3$	...	$I_1 \cdot T_N$

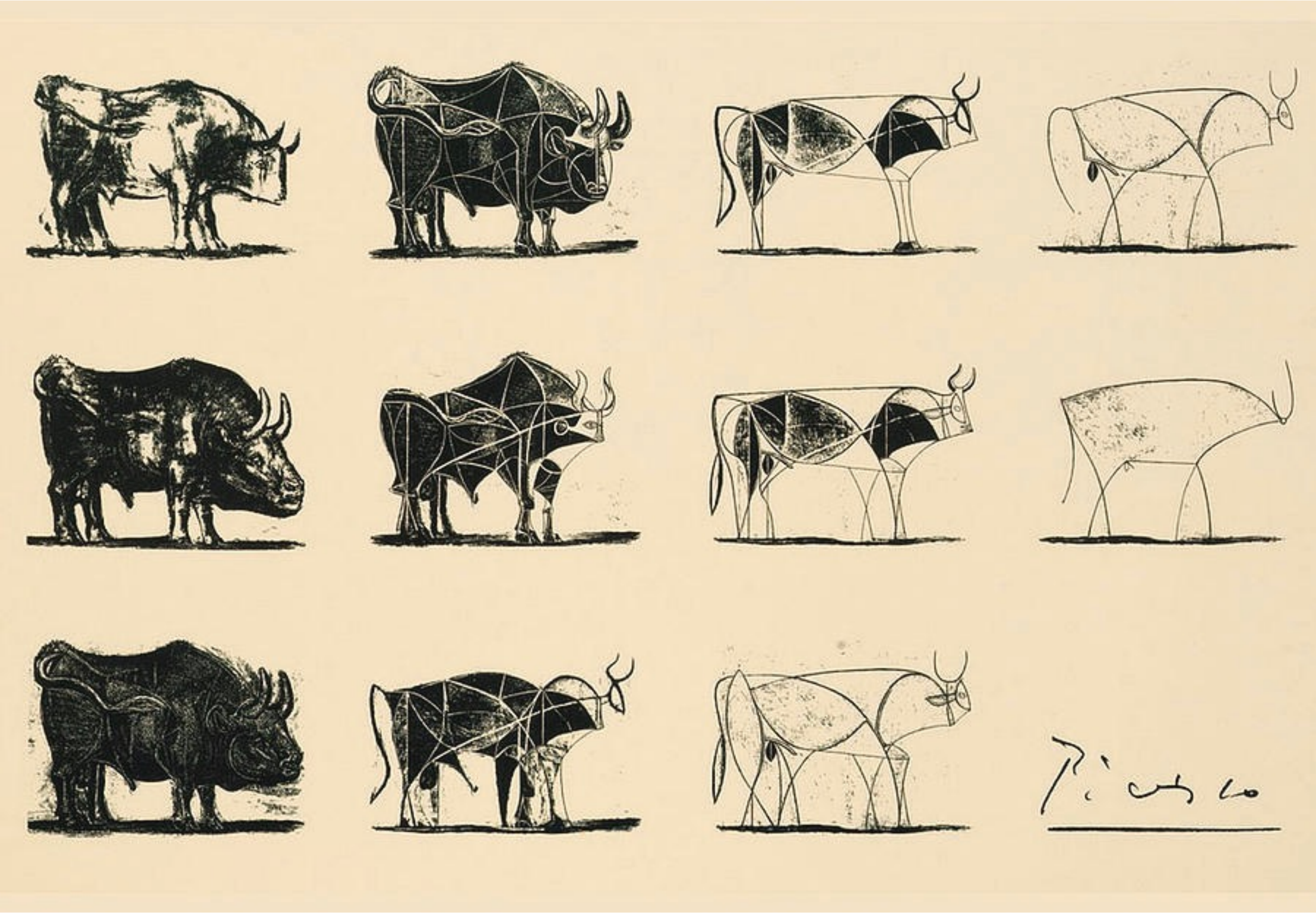
Advantages of CLIP

- Why can CLIP do image generation?
  - Excellent generalization ability

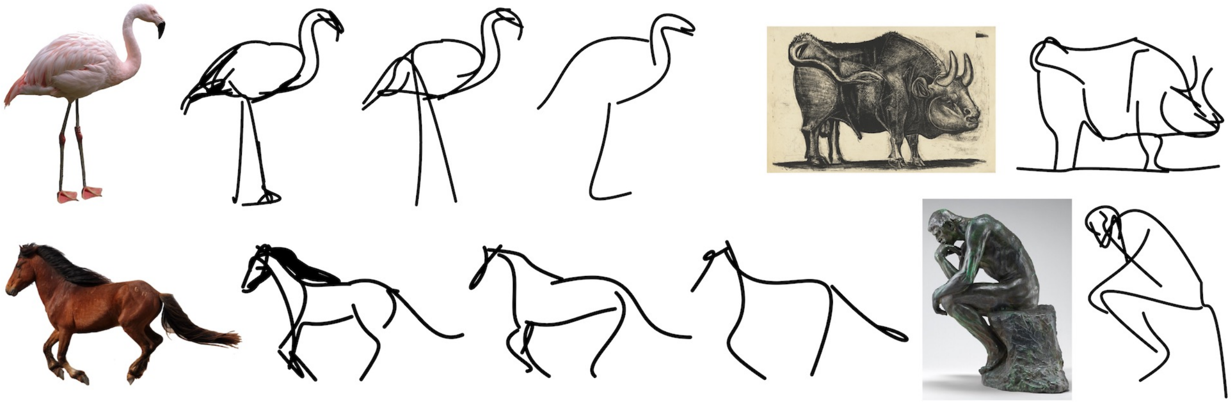


	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		**76.2**	**76.2**	0%
ImageNetV2		64.3	**70.1**	+5.8%
ImageNet-R		37.7	**88.9**	+51.2%
ObjectNet		32.6	**72.3**	+39.7%
ImageNet Sketch		25.2	**60.2**	+35.0%
ImageNet-A		2.7	**77.1**	+74.4%

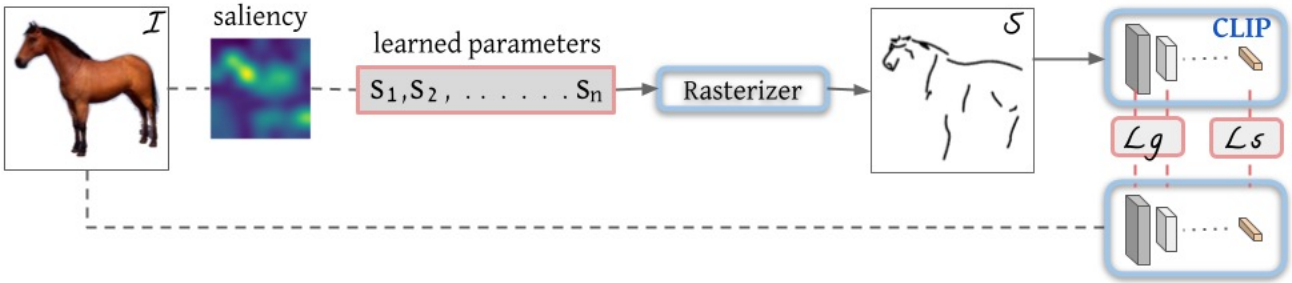
Picasso “Le Taureau”



CLIPasso: Semantically-Aware Object Sketching

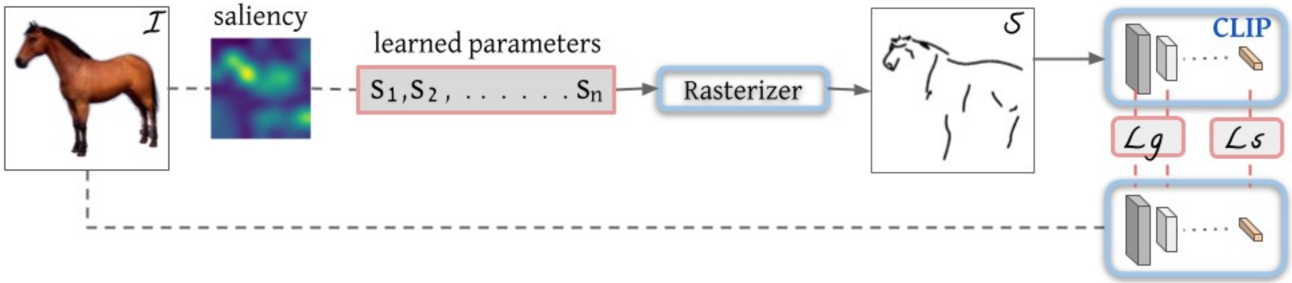


- Pipeline



CLIPasso: Semantically-Aware Object Sketching

- Pipeline



- Initial input: A series of Bezier curves' control points.

$$S(t) = P_0(1 - t)^3 + 3P_1 t(1 - t)^2 + 3P_2 t^2(1 - t) + P_3 t^3$$

CLIPasso: Semantically-Aware Object Sketching

- Geometric Loss

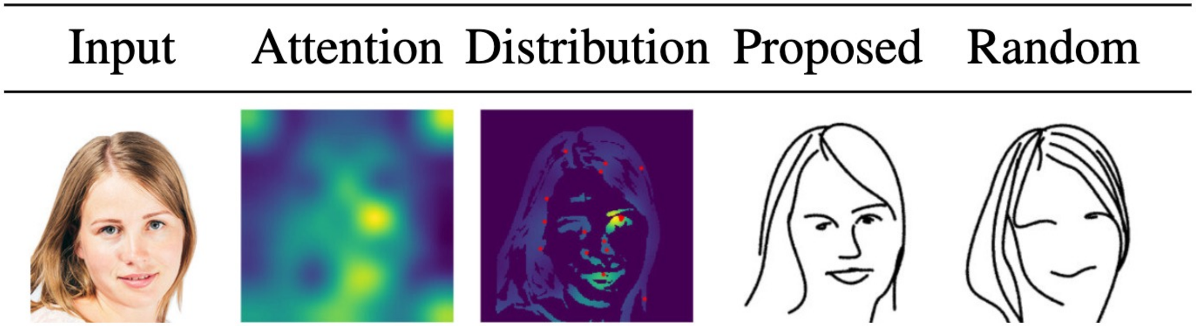
$$L_{geometric} = \sum_{l} \left\| CLIP_l(I) - CLIP_l\big(R(\{S_i\})\big) \right\|_2^2$$

- Semantics Loss

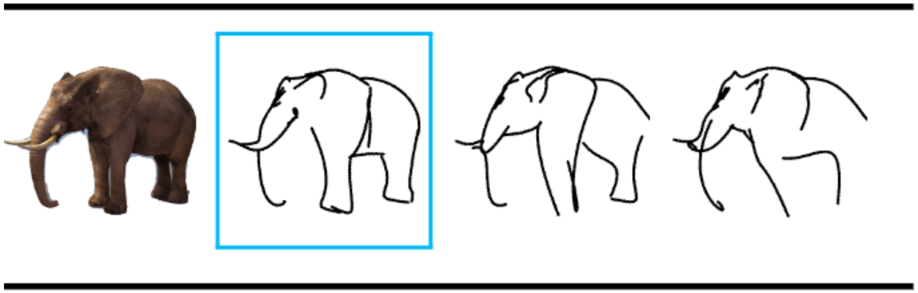
$$L_{semantic} = dist(CLIP(I), CLIP(R(\{S_i\}))$$

CLIPasso: Semantically-Aware Object Sketching

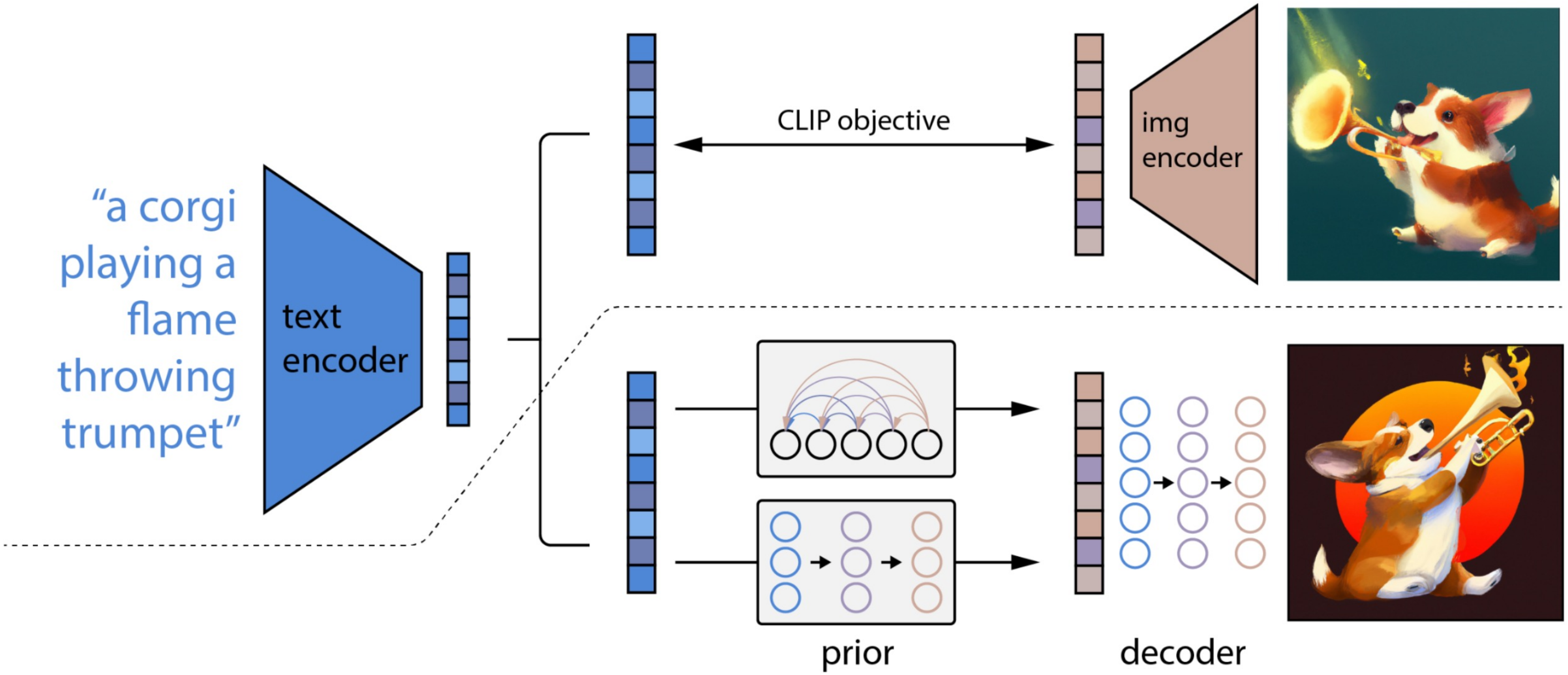
- Initialization
  - Using ViT to generation attention map.



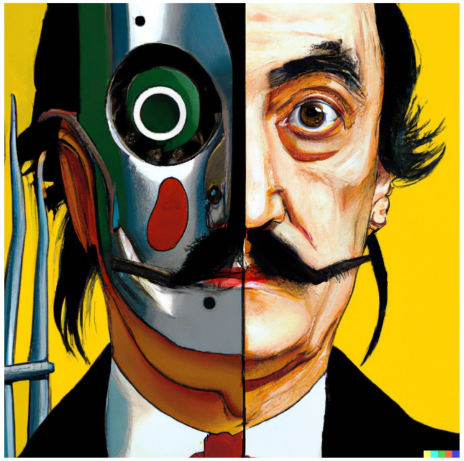
- Result Selection



CLIP in Image Generation – DALL-E.2



CLIP in Image Generation – DALL-E.2



vibrant portrait painting of Salvador Dalí with a robotic half face

a shiba inu wearing a beret and black turtleneck

a close up of a handpalm with leaves growing from it

a propaganda poster depicting a cat dressed as french emperor napoleon holding a piece of cheese

an espresso machine that makes coffee from human souls, artstation

panda mad scientist mixing sparkling chemicals, artstation

a corgi's head depicted as an explosion of a nebula

a teddy bear on a skateboard in times square

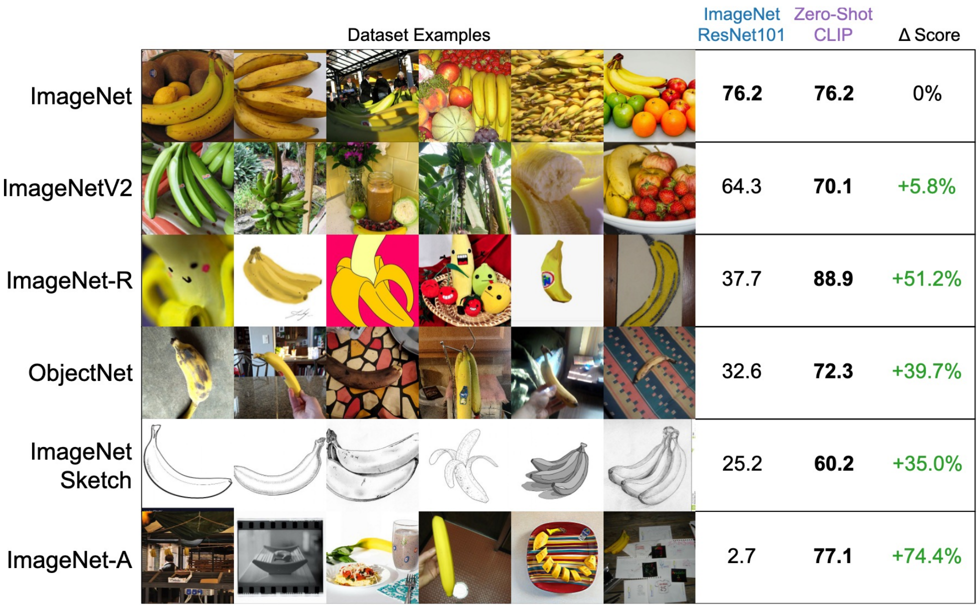
Discussion

- NLP Supervision
  - Compared with single label, a sentence consists of multi concepts;
  - Multi-concepts help minimize the ambiguity.

"Remote" vs "A photo of remote"

Discussion

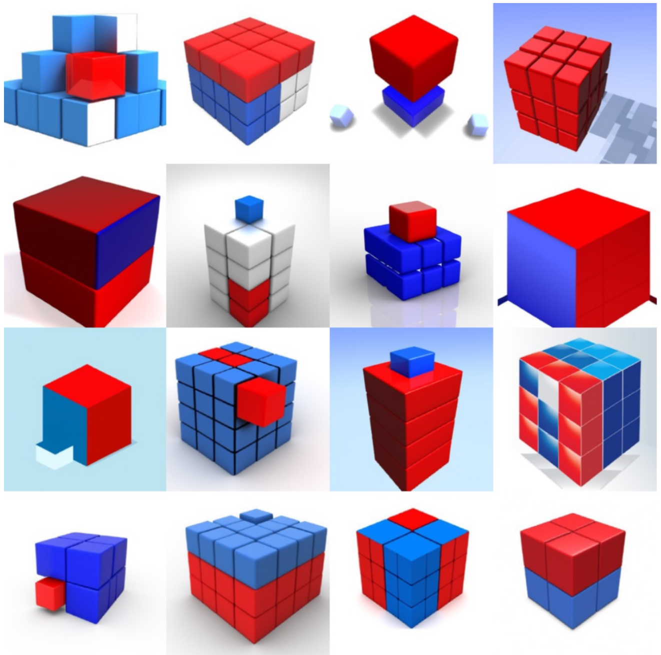
- NLP Supervision
  - Compared with single label, a sentence consists of multi concepts;
  - Multi-concepts help minimize the ambiguity.
  - More robust to distribution shift.



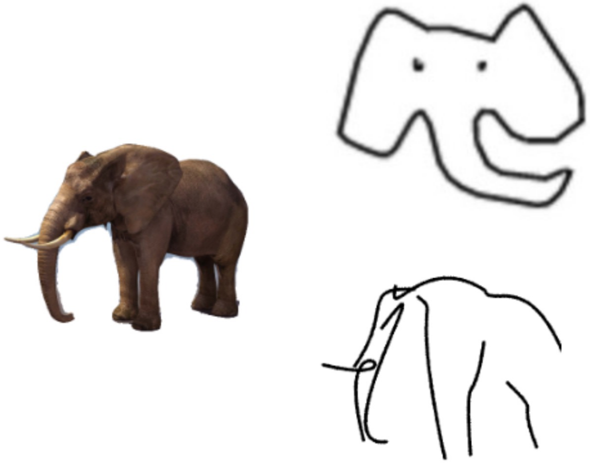
	Dataset Examples	ImageNet ResNet101	Zero-Shot CLIP	Δ Score
ImageNet		**76.2**	**76.2**	0%
ImageNetV2		64.3	**70.1**	+5.8%
ImageNet-R		37.7	**88.9**	+51.2%
ObjectNet		32.6	**72.3**	+39.7%
ImageNet Sketch		25.2	**60.2**	+35.0%
ImageNet-A		2.7	**77.1**	+74.4%

Discussion

- Limitation
  - Lack of understanding of attributes.



DALL-E2's result *"A red cube* ***on top of*** *a blue cube."*



Geometric Loss is necessary in CLIPasso

Discussion

- Conclusion
  - Data Preparation:
    - The labels can be non-fixed;
    - Web-scale pre-training is used in multi-modal tasks;
  - Data Processing:
    - The result of CLIP's image encoder is treated as ground truth in DALL-E 2.
    - The result of CLIP's similarity score is used in Loss of CLIPasso.

Reference

[1] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[2] Ramesh A, Dhariwal P, Nichol A, et al. Hierarchical text-conditional image generation with clip latents[J]. arXiv preprint arXiv:2204.06125, 2022.

[3] Vinker Y, Pajouheshgar E, Bo J Y, et al. Clipasso: Semantically-aware object sketching[J]. arXiv preprint arXiv:2202.05822, 2022.

Thank you!

Discussion

- Limitation
  - CLIPasso‘s performance reduced for images with background.
  - The number of strokes is determined, and the model cannot be adjusted adaptively. In order to draw more like a human, the strokes should be generated sequentially.