



Alpha-CLIP: A CLIP Model Focusing on Wherever You Want

Zeyi Sun^{*1,4}, Ye Fang^{*2,4}, Tong Wu³, Pan Zhang⁴, Yuhang Zang⁴,
Shu Kong⁵, Yuanjun Xiong⁶, Dahua Lin^{3,4}, Jiaqi Wang^{†4}

¹Shanghai Jiao Tong University ²Fudan University ³The Chinese University of Hong Kong

⁴Shanghai AI Laboratory ⁵University of Macau ⁶MThreads, Inc.

szy2023@sjtu.edu.cn, {fangye, zhangpan, zangyuhang, wangjiaqi}@pjlab.org.cn

4. 4. 2024

EffL Lab @ POSTECH

Yuji Byun

Introduction : multi-modal learning

- Multi-modal : The integration and processing of information from multiple types such as

Unimodal



Multi-modal

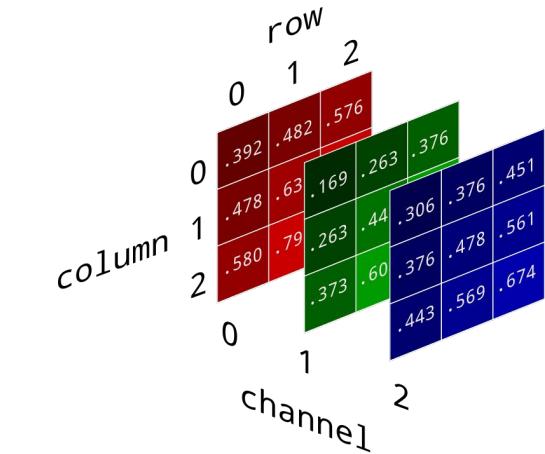
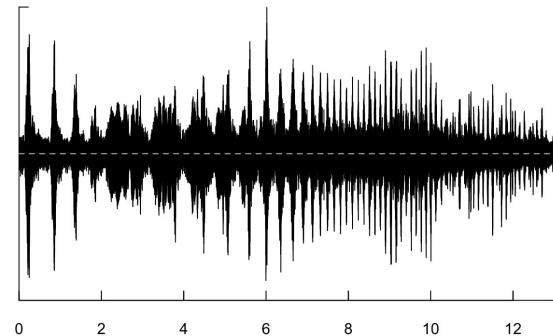


Text



Introduction : multi-modal learning

- Multi-modal learning is difficult because of..
 - Data Alignment
 - : Differences in semantics...
 - Dimensionality
 - : Different dimension...
 - Data imbalance
 - : Variances in the availability and quality...
 - And so on...



Introduction : multi-modal learning

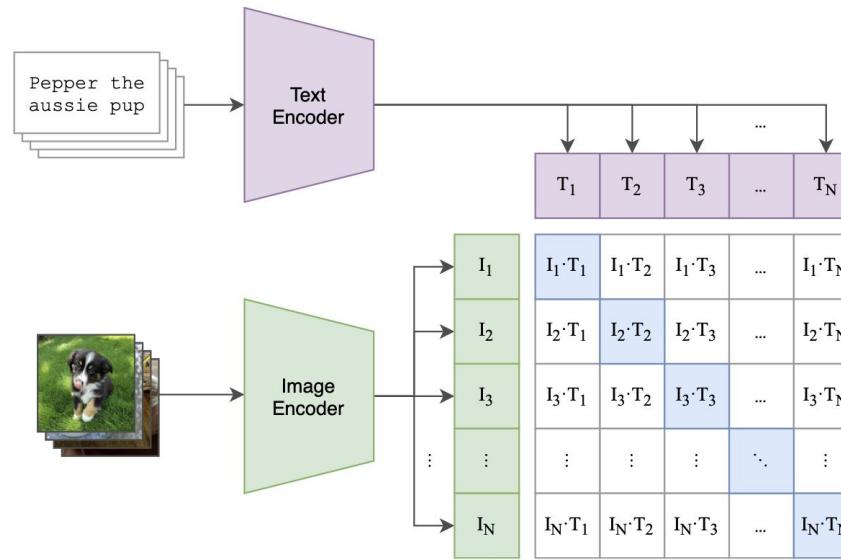
- Multi-modal learning is difficult because of..
 - Data Alignment
: Differences in semantics...
 - Dimensionality
: Different dimension...
 - Data imbalance
: Variances in the availability and quality...
 - And so on...

Dancing
Christmas Tree



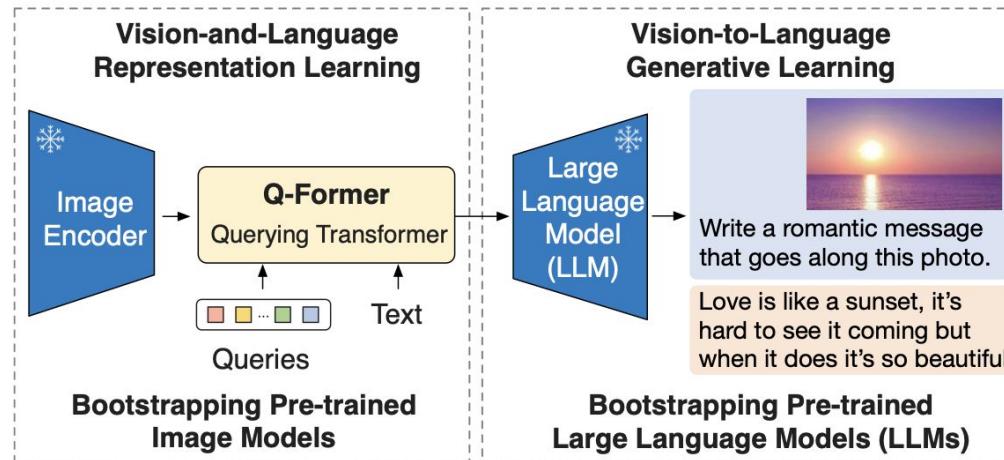
Introduction : CLIP

- CLIP^[1] : Contrastive Language-Image Pre-training



Introduction : Application of CLIP

- Multimodal Large Language model : BLIP-2[1]



[1] Li et al., BLIP-2 : Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Introduction : Application of CLIP

- Multimodal Large Language model : BLIP-2[1]



8

8

Write a conversation between the two animals.

cat: hey dog, can i ride on your back?
dog: sure, why not?
cat: i'm tired of walking in the snow.



8

8

What are shown in the photo?

A man and a chicken.

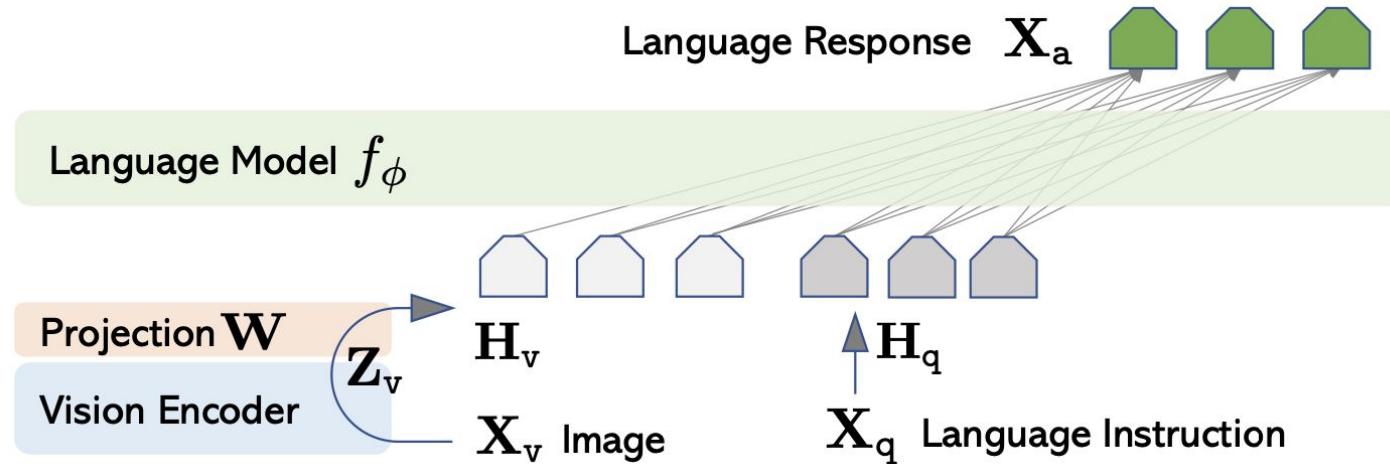
What does the man feel and why?

He is scared of the chicken because it is flying at him.

[1] Li et al., BLIP-2 : Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models

Introduction : Application of CLIP

- Multimodal Large Language model : LLAVA[1]





User

Introducti



What are the meals that I can cook with these?

• Multi



With the variety of food items stored in the refrigerator, you can prepare several meals. For example, you can create a fruit salad using the strawberries, blueberries, and carrots. Additionally, you can make a delicious smoothie using the fruits and vegetables, which would be a healthy and nutritious snack or breakfast option. The milk and eggs can be used to create various dishes, such as a milk-based smoothie, a baked casserole, or scrambled eggs. The juice can also be used to make a fresh fruit and juice mix or as a refreshing homemade popsicle. The possibilities are vast, and the meals you can create will depend on your preferences and dietary needs.

I would like to do the fruit salad. Show me the recipe.



User



Certainly! Here's a simple fruit salad recipe using the items in your refrigerator:

Ingredients:

- 2 cups fresh strawberries, sliced
- 1 cup fresh blueberries
- 1 large carrot, grated
- Juice of half a lemon
- A handful of chopped fresh parsley or mint
- A few ice cubes
- Optional: honey or maple syrup (adjust to taste)

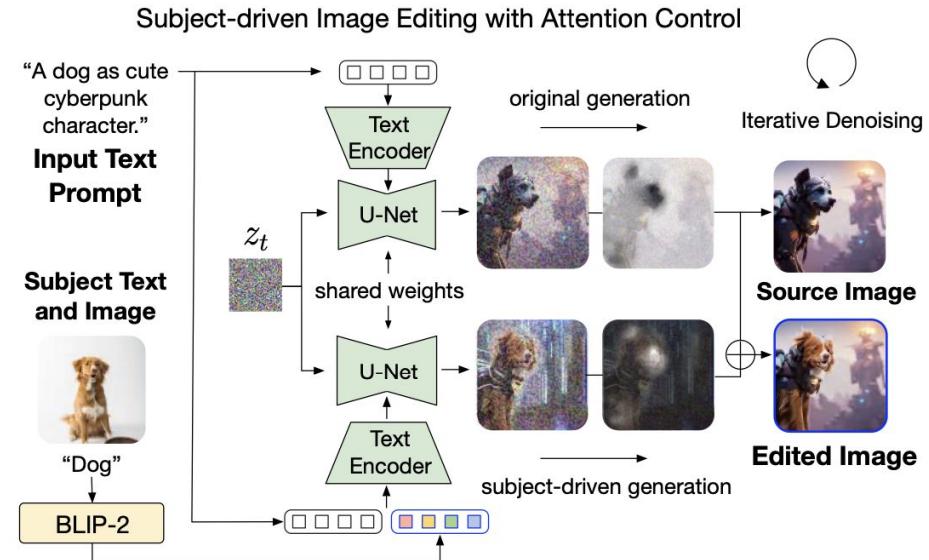
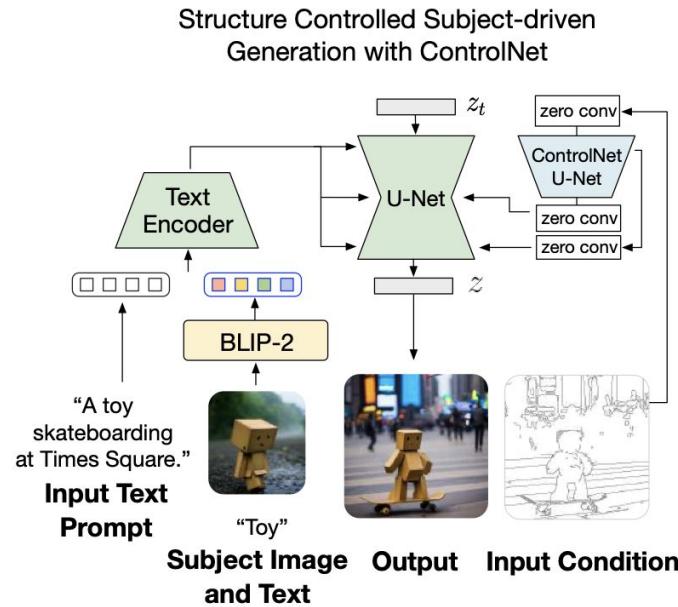
Instructions:

1. In a large bowl, combine the sliced strawberries, blueberries, and carrot. Gently toss them together to mix.
2. Drizzle the juice of half a lemon over the fruit and vegetables to prevent them from oxidizing and add a refreshing citrus flavor.
3. Sprinkle the chopped fresh parsley or mint over the fruit and gently toss once more.
4. If desired, add a drizzle of honey or maple syrup for a touch of sweetness. Adjust the amount according to your taste preferences.
5. Cover the bowl and refrigerate the fruit salad for at least 30 minutes or up to a few hours. This allows the flavors to meld and creates a more refreshing salad.
6. Serve the fruit salad in cups or bowls, garnished with a sprinkle of fresh parsley or mint, if desired.

This fruit salad recipe is perfect for a healthy snack or side dish, and with the variety of ingredients available in your refrigerator, it should be easy to put together quickly.

Introduction : Application of CLIP

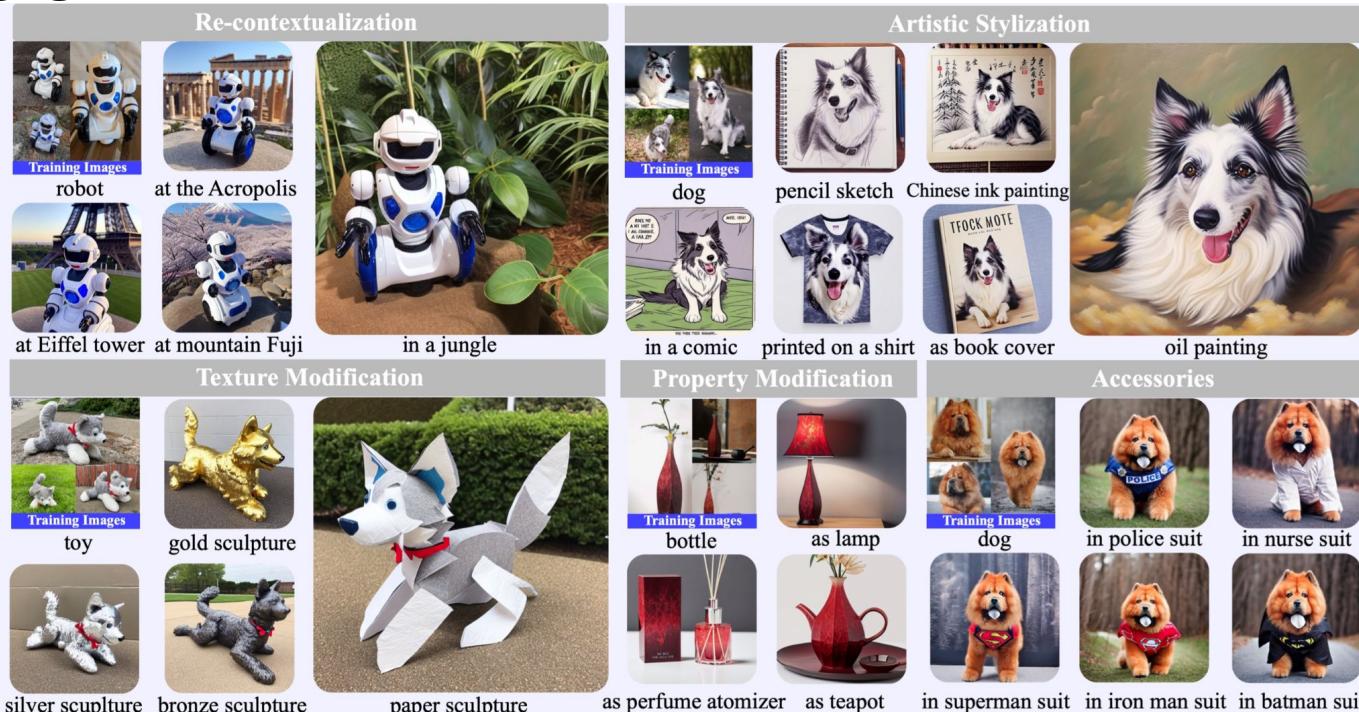
- Image generation model : BLIP-Diffusion[1]



[1] Li et al., BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing

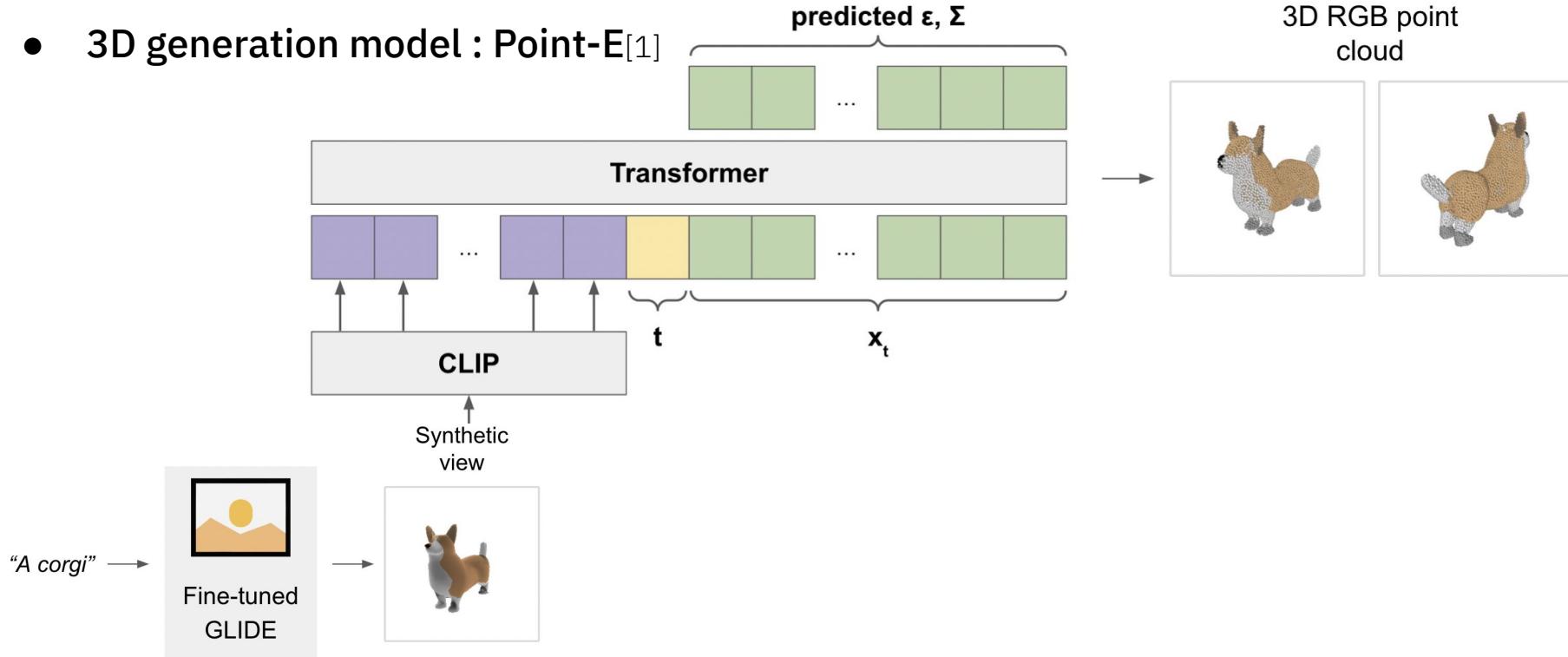
Introduction : Application of CLIP

- Image generation model : BLIP-Diffusion[1]



Introduction : Application of CLIP

- 3D generation model : Point-E^[1]



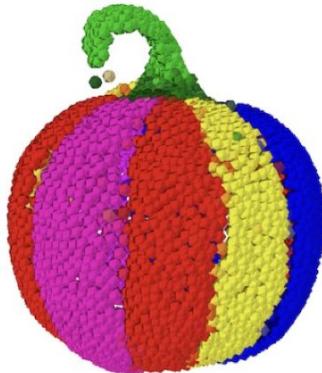
[1] Nichol et al., Point-E: A System for Generating 3D Point Clouds from Complex Prompts

Introduction : Application of CLIP

- 3D generation model : Point-E^[1]



“a corgi wearing a
red santa hat”



“a multicolored rainbow
pumpkin”



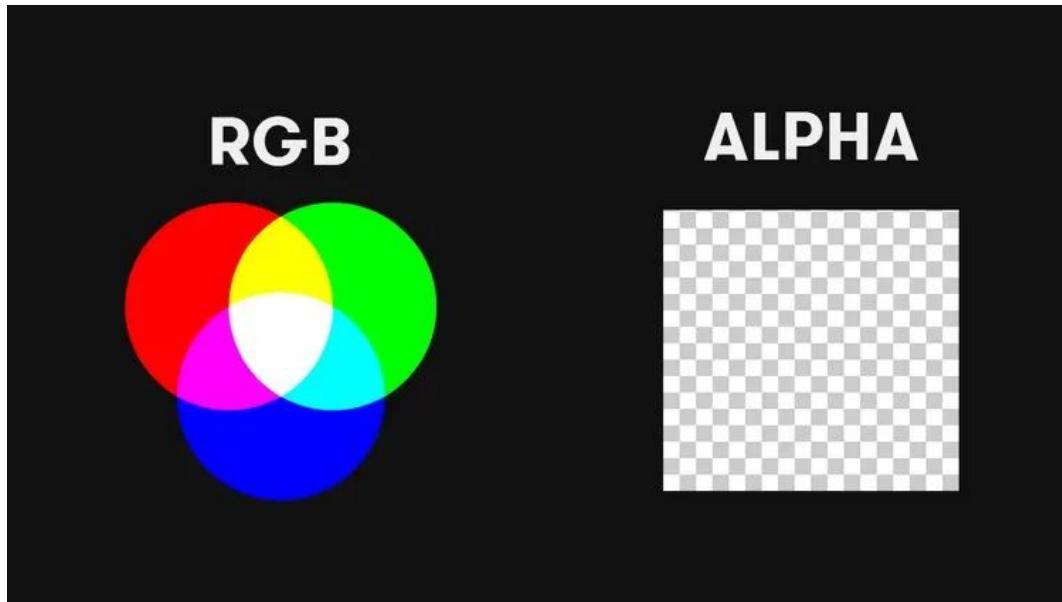
“an elaborate fountain”



“a traffic cone”

Introduction : RGBA

- RGBA : RGB + Alpha channel

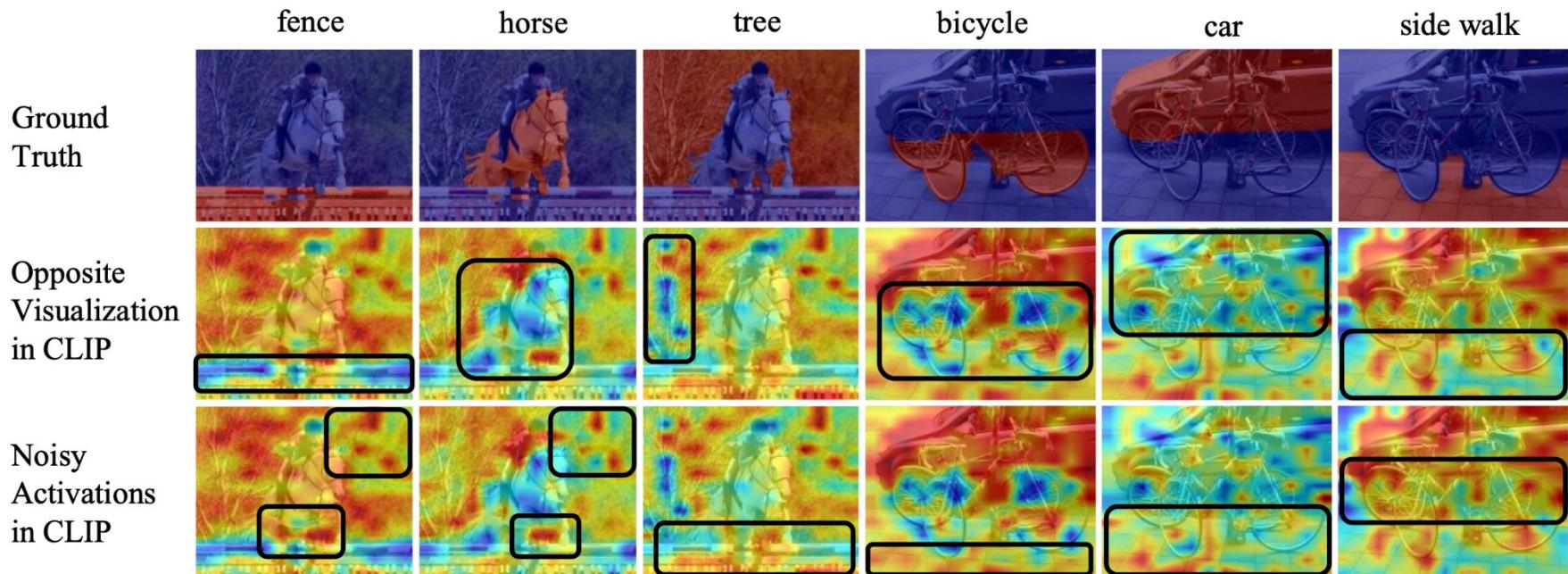


Introduction : RGBA

- RGBA : RGB + Alpha channel



Introduction : Limitation of CLIP



Introduction : Limitation of CLIP



Introduction : Limitation of CLIP



Alpha-CLIP



cropping

😢 disrupt context
rectangular area only

masking

😢 omit context

red circle

😢 change image content

feature masking

😢 omit context
patch-level granularity

Alpha-CLIP

😊 keep context
pixel-level granularity

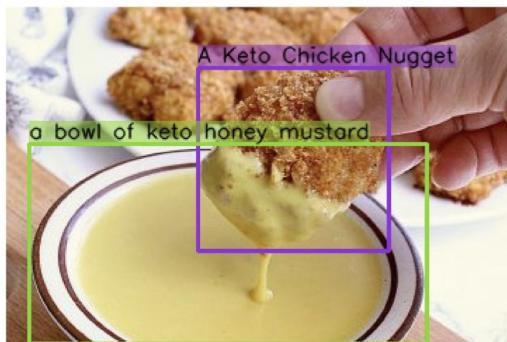
Alpha-CLIP : Data pipeline

Grounding Data



Box-text pairs

SAM



RS



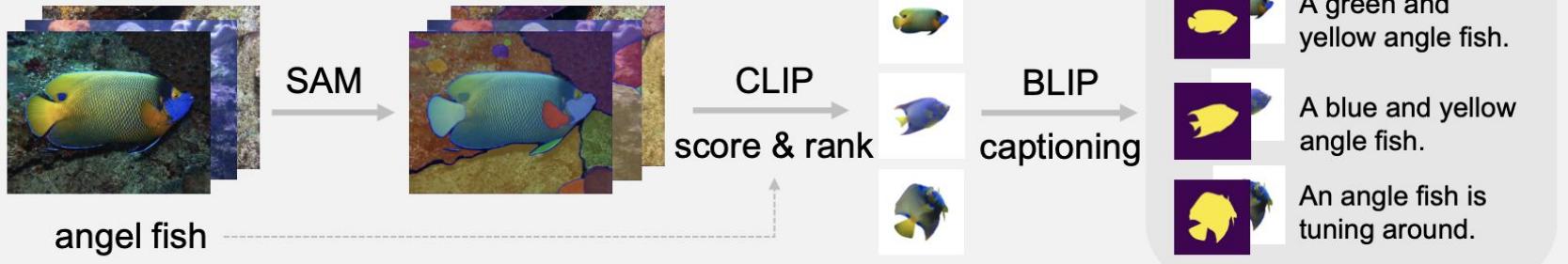
A female athlete
Volleyball
A sandy court

- Grounding Data from GRIT(Grounded Image-Text pairs)^[1] dataset

[1] Peng et al., KOSMOS-2: Grounding Multimodal Large Language Models to the World

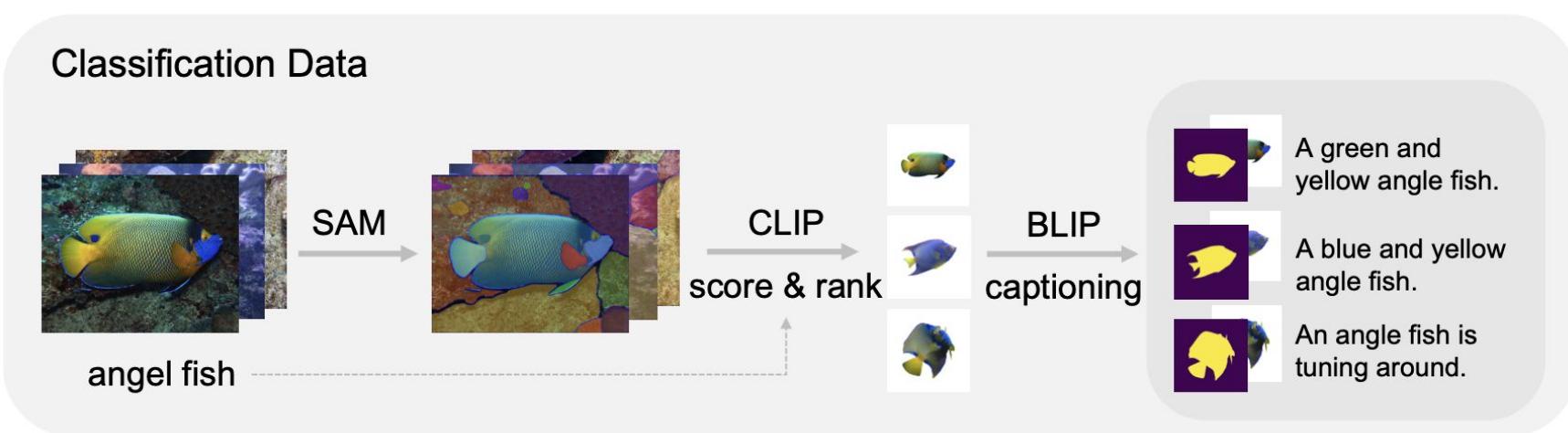
Alpha-CLIP : Data pipeline

Classification Data



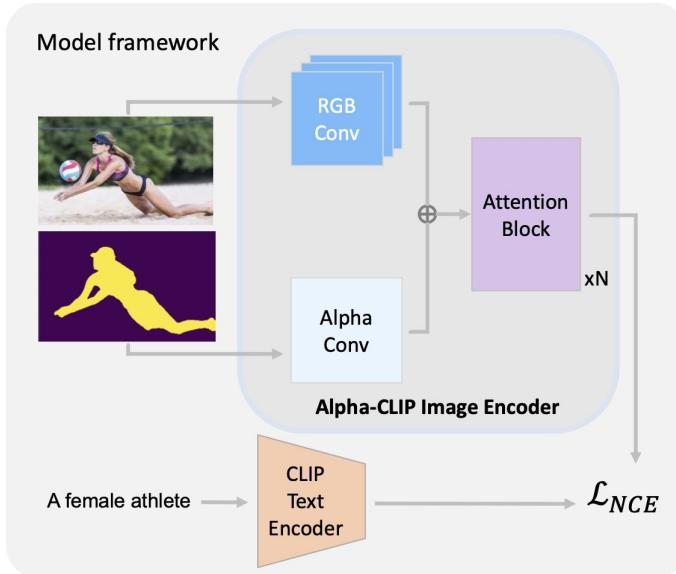
- Classification data from ImageNet dataset

Alpha-CLIP : Data pipeline



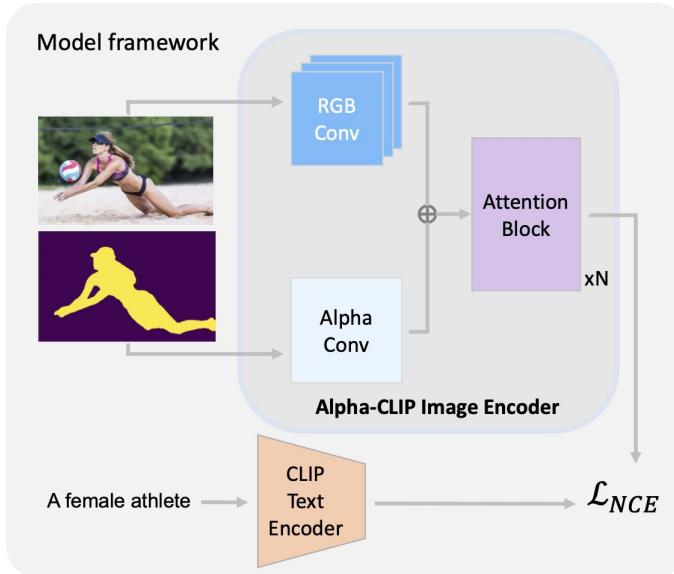
1. Automatically generating mask for each image
2. Cropping the foreground object -> Centering -> Enlarging
3. Calculating class label scores with CLIP and sorting the mask by class based
4. Selecting top-ranked mask, and captioning with BLIP-2

Alpha-CLIP : Fine-tuning pipeline



- Adding Alpha Conv layer parallel to the RGB Conv layer in first ViT layer
- Alpha channel input is set to range from [0,1]
 - 1 = foreground / 0 = background

Alpha-CLIP : Fine-tuning pipeline



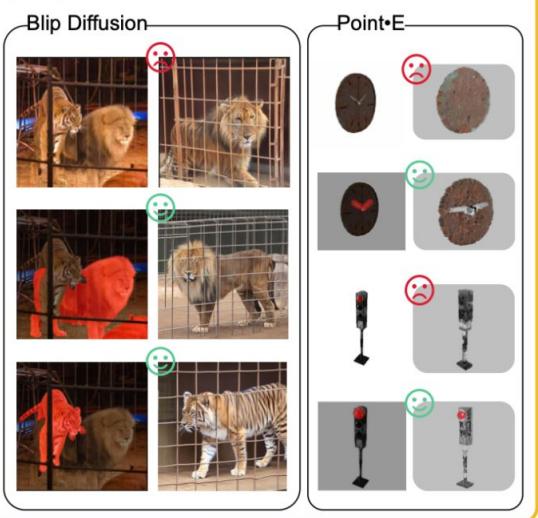
- Training only CLIP image encoder (freezing text encoder)
- To preserve CLIP's global recognition, adopting a data sampling generation
 - Occasionally setting the alpha channel to full 1 (0.1)

Alpha-CLIP!

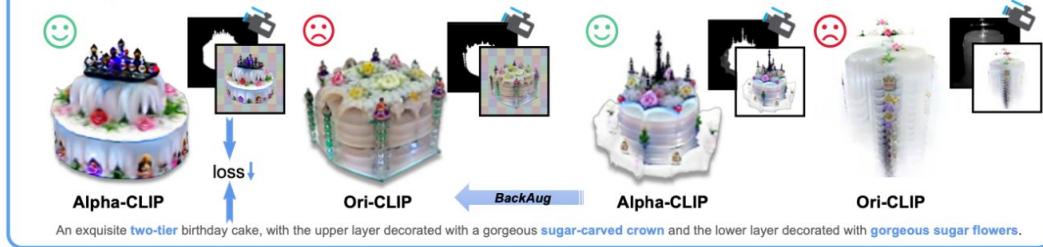
Alpha-CLIP+LLM



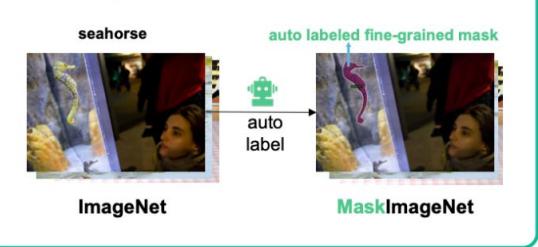
Alpha-CLIP+Diffusion



Alpha-CLIP+NeRF



Alpha-CLIP+SAM



Experiments

- Zero-shot classification
 - MaskCLIP isn't directly compatible with BLIP-2, BLIP-Diffusion, LLaVA...

Methods	ViT-B/16		ViT-L/14	
	Top-1	Top-5	Top-1	Top-5
Original CLIP [1]	66.48	88.90	73.48	91.60
MaskAdaptedCLIP [2]	57.86	79.12	63.50	86.34
Red Circle [3]	65.37	88.68	73.37	92.09
MaskCLIP* [4]	67.86	89.40	77.04	93.39
Alpha-CLIP(ours)	68.89	90.51	77.41	94.45

[1] Radford et al., Learning Transferable Visual Models From Natural Language Supervision(ICML'21)

[2] Liang et al., Open-vocabulary semantic segmentation with mask-adapted clip

[3] Shtedritski et al., What does clip know about a red circle?

[4] Zhou et al., Extract free dense labels from clip

Experiments

- Zero-shot classification
 - With different alpha map levels

Model	Alpha Map	Top-1	Top-5
CLIP [1]	-	73.48	91.60
	whole image	73.37	91.75
	rectangular box	75.62	93.34
Alpha-CLIP	mask	77.41	94.45

Experiments

- Region level captioning

Model	RefCOCOg		Visual Genome	
	METEOR	CIDEr	METEOR	CIDEr
GRIT [1]	15.2	71.6	17.1	142.0
Kosmos-2 [2]	14.1	62.3	-	-
GPT4RoI [3]	-	-	17.4	145.2
GLaMM [4]	16.2	105.0	18.6	157.8
Alpha-CLIP+LLaVA	16.7	109.2	18.9	160.3

[1] Wu et al., Grit : A generative region-to-text transformer for object understanding

[2] Peng et al., Kosmos-2 : Grounding multimodal large language models to the world

[3] Zhang et al., Ipt4roi: Instruction tuning large language model on region-of-interest

[4] Rasheed et al., Glamm: Pixel grounding large multimodal model

Experiments

- Region level captioning

Alpha-CLIP in LLaVA

 What is in the bowl?  The bowl contains apples and eggs.	 The image features a unique and creative sculpture of a fire-breathing dragon or a creature with flames coming out of its head. The sculpture is made of a combination of red and black colors, giving it a striking appearance. The dragon is standing on a table, and it appears to be a part of a larger artistic display. The sculpture's intricate design and the use of vibrant colors make it an eye-catching piece of art.
 What is in the bowl?  The bowl contains flour.	 The image features a red sculpture of a creature, possibly a dragon or a monster, with a pair of red glasses on top of its head. The glasses are placed in such a way that they appear to be part of the creature's head, giving it a unique and eye-catching appearance. The sculpture is positioned on a table, drawing attention to its vibrant colors and interesting design.

Experiments

- Region level captioning(BLIP-2)



A red telephone on a table



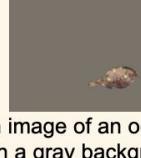
A red mushroom on a table



A green telephone on a table

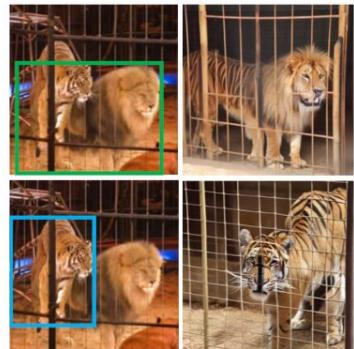
Experiments

- Region level captioning(BLIP-2)

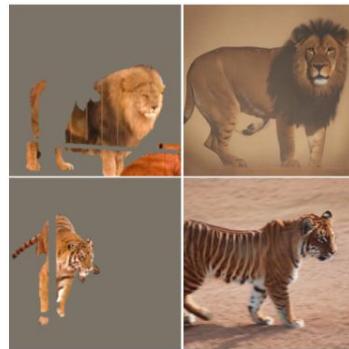
image level masking result	feature level masking result	Alpha-CLIP result
	 Image of a shell on a gray background	An image of a shell with the word person on it
	 A picture of a cat in a cage	A small shell sitting on the sand
	 A black and red file cabinet on a table	A small sea creature in a shell on the sand
	 A picture of a book on a gray background	A man is working on a computer
		A black and red microwave
		A man is using a computer
		image

Experiments

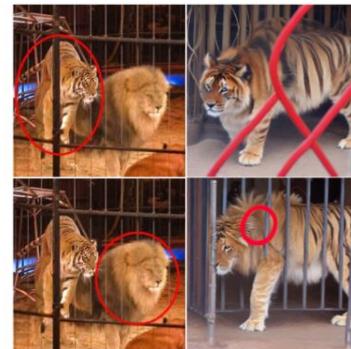
- Image variation(BLIP-diffusion)



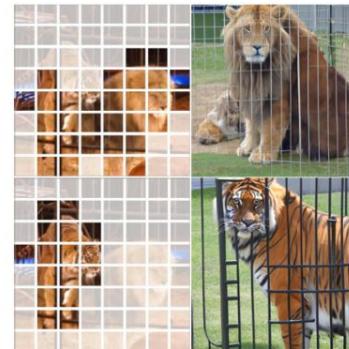
cropping
disrupt context
rectangular area only



masking
omit context



red circle
change image content



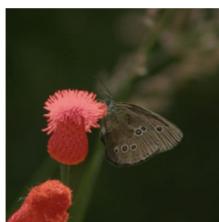
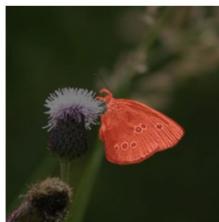
feature masking
omit context
patch-level granularity



Alpha-CLIP
keep context
pixel-level granularity

Experiments

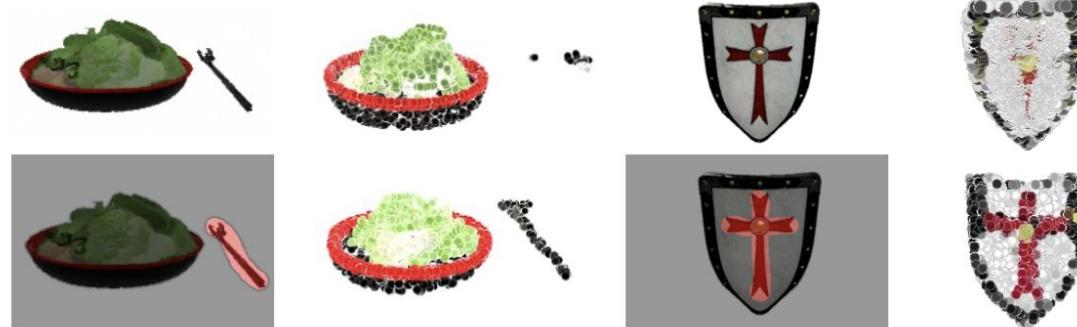
- Image variation(BLIP-diffusion)



Experiments

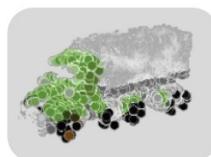
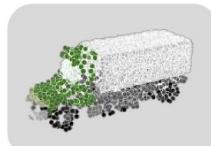
- 3D object generation

Alpha-CLIP in Point-E



Experiments

- 3D object generation



Rectifying missing part

Emphasizing specific part

Experiments

- 3D object generation
 - BackAug, derived Dream Fields[1], improves consistency and makes output less diffuse

Alpha-CLIP in PureCLIPNeRF



A Baroque church with ornate reliefs on the walls, soaring vaults and domes, and rich gold decorations throughout.



A porcelain plate displays juicy meat, broccoli and brown toast.

BackAug ✗
Alpha CLIP ✗

BackAug ✗
Alpha CLIP ✓

BackAug ✓
Alpha CLIP ✗

BackAug ✓
Alpha CLIP ✓

Limitations

- Limited capability to focus on multiple objects or model relationships between different object
- Restricted the alpha channel (the binary values of 0 and 1)

Thank you