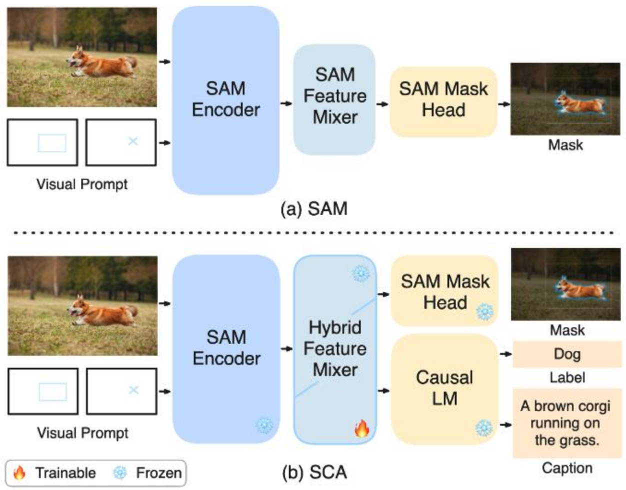
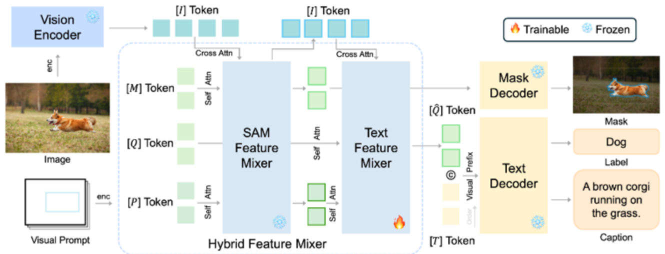
Segment and Caption Anything

Xiaoke Huang[1], Jianfeng Wang[2], Yansong Tang[1], Zheng Zhang[2], Han Hu[2], Jiwen Lu[1], Lijuan Wang[2], Zicheng Liu[3]
[1]Tsinghua University, [2]Microsoft, [3]AMD

Introduction



Method



We found that the regional features of SAM (Segment Anything Model) can be used for regional captioning.

Thus we proposed a lightweight query-based feature mixer to connect SAM with Causal Language Model.

Project Page & Code

Comparison

Method	M	C
ASM [20] (Zero-shot)†	12.6	44.2
ASM (Finetuned)†	18.0	145.1
GPT4RoI [24] (7B)†	17.4	145.2
GPT4RoI (13B)†	17.6	146.8
GPT4RoI (7B)‡	16.4	122.3
SCA (GPT2-large, VG)	17.4	148.8
SCA (LLAMA-3B, VG)	17.4	149.8
SCA (GPT2-large, Pretrain+VG)	17.5	149.8

Pre-train or not

Pretrain	C	M	S
*No Pretrain**	127.9	15.8	27.7
COCO [54] (img. 117K, cls. 80)†	130.2	16.0	28.0
V3Det [94] (img. 183K, cls. 13K)†	130.4	16.0	28.0
O365 [81] (img. 1M, cls. 365)†	134.5	16.3	28.7

Anything Mode

Training Recipe

M. LR	T.D.	T.D. LR	C	M	S
1e-4	GPT2 -large	5e-6	135.6	16.3	28.5
		1e-6	134.8	16.2	28.5
		5e-7	134.5	16.2	28.5
		1e-7	135.6	16.4	28.8
		0.0	136.0	16.5	28.9
5e-5	GPT2 -large	5e-6	129.1	15.7	27.5
		1e-6	131.4	15.9	28.0
		5e-7	131.2	16.0	28.0
		1e-7	132.5	16.1	28.2
		0.0	131.7	16.1	28.2
1e-4	GPT2	5e-6	134.1	16.2	28.4
		1e-6	134.7	16.3	28.7
		5e-7	134.5	16.2	28.7
		1e-7	133.2	16.1	28.6
		0.0	132.3	15.9	28.9
5e-5	GPT2	5e-6	131.3	16.0	28.0
		1e-6	131.1	16.0	28.1
		5e-7	130.6	15.9	28.1
		1e-7	130.4	15.9	28.2
		0.0	126.3	15.4	27.9

Segment and Caption Anything

Xiaoke Huang[1†] Jianfeng Wang[2] Yansong Tang[1*] Zheng Zhang[2]
Han Hu[2] Jiwen Lu[3] Lijuan Wang[2] Zicheng Liu[4]

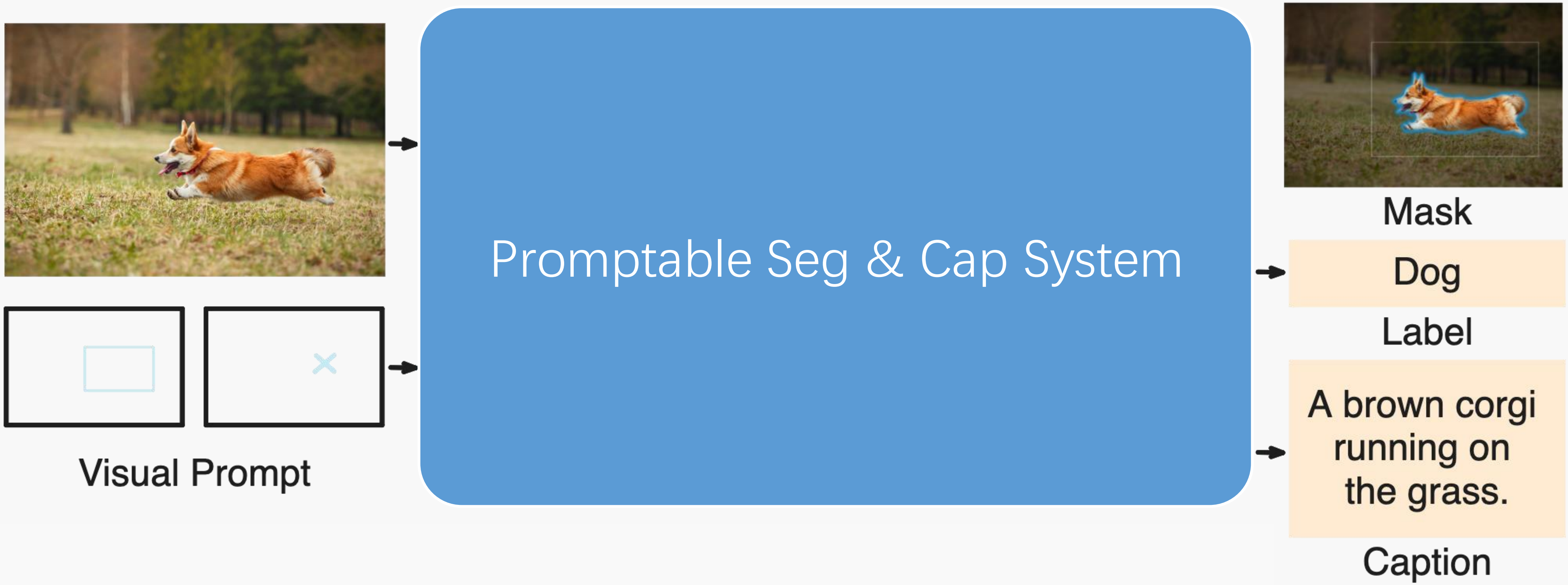
[1]Shenzhen International Graduate School, Tsinghua University [2]Microsoft
[3]Department of Automation, Tsinghua University [4]Advanced Micro Devices

{hxk21@mails.,tang.yansong@sz.,lujiwen@}tsinghua.edu.cn
{jianfw,zhez,lijuanw}@microsoft.com ancientmooner@gmail.com zicheliu@amd.com

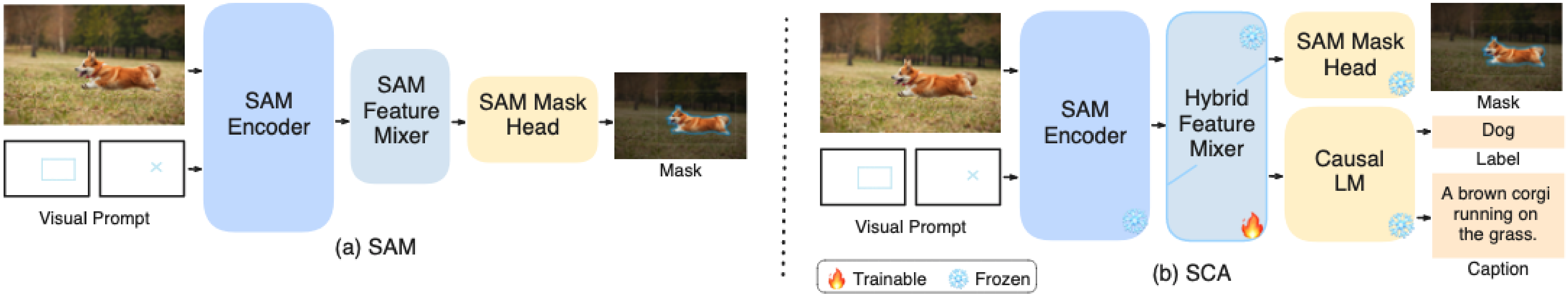
Outline

- Intro
- Preliminary
  - Task: Image/Dense Caption
  - Model: SAM / Language Modeling
- Method
- Results
- Conclusion, and What's Next?

Task


Visual Prompt
Promptable Seg & Cap System
Mask
Dog
Label
A brown corgi running on the grass.
Caption

Segment and Caption Anything



tl;dr

1. Despite the absence of semantic labels in the training data, SAM implies high-level semantics sufficient for captioning.
2. SCA (b) is a lightweight augmentation of SAM (a) with the ability to generate regional captions.
3. On top of SAM architecture, we add a fixed pre-trained language mode, and a optimizable lightweight hybrid feature mixture whose training is cheap and scalable.

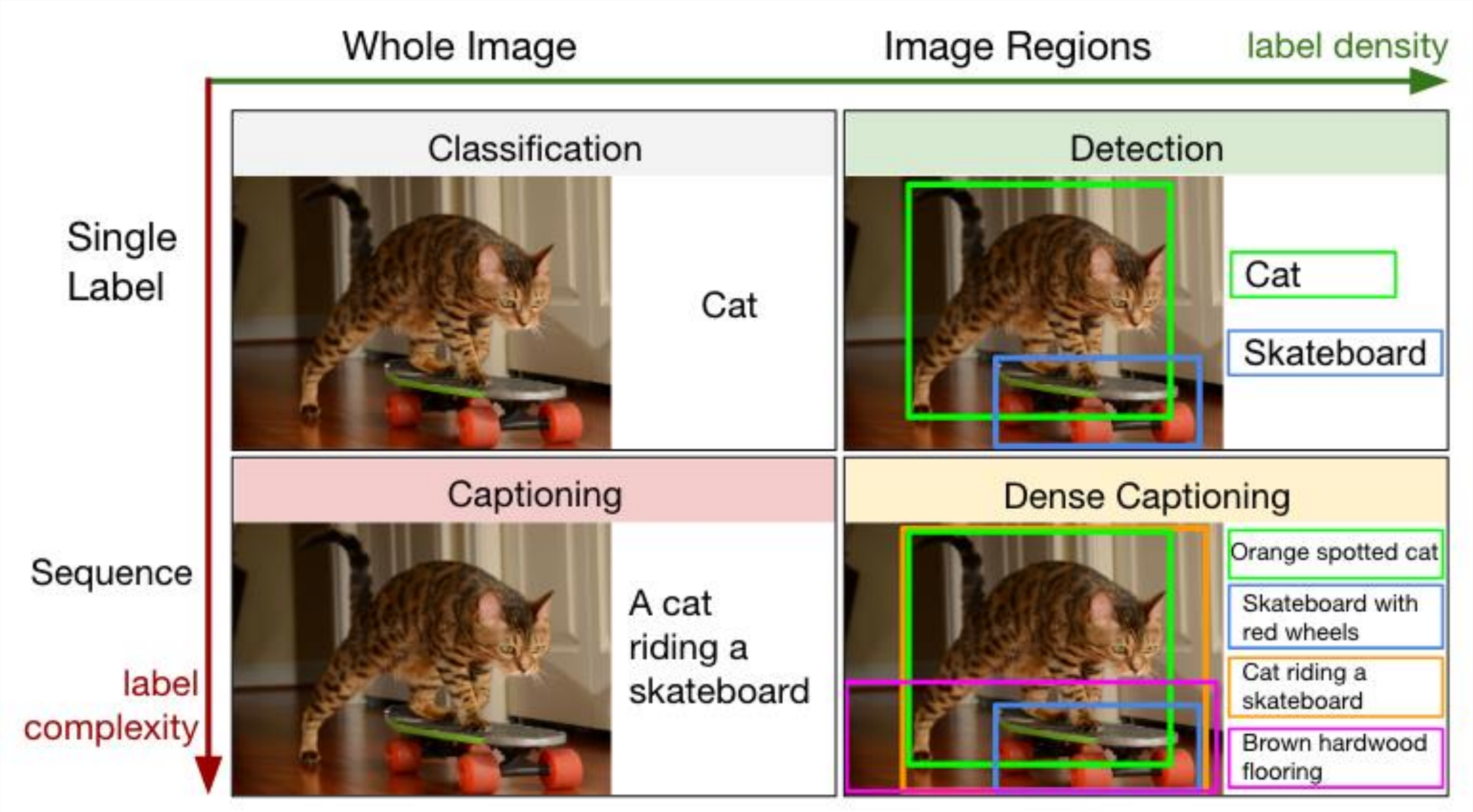
Preliminary

- Task: Image/Dense Caption
- Model: SAM / Language Modeling

i-VisionGroup

Preliminary

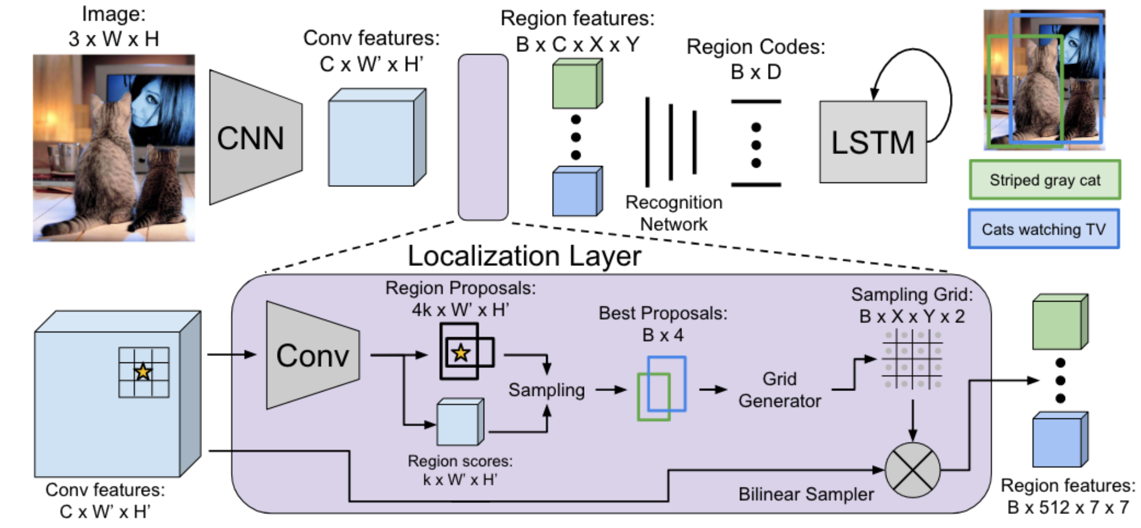
Task: Image/Dense Caption

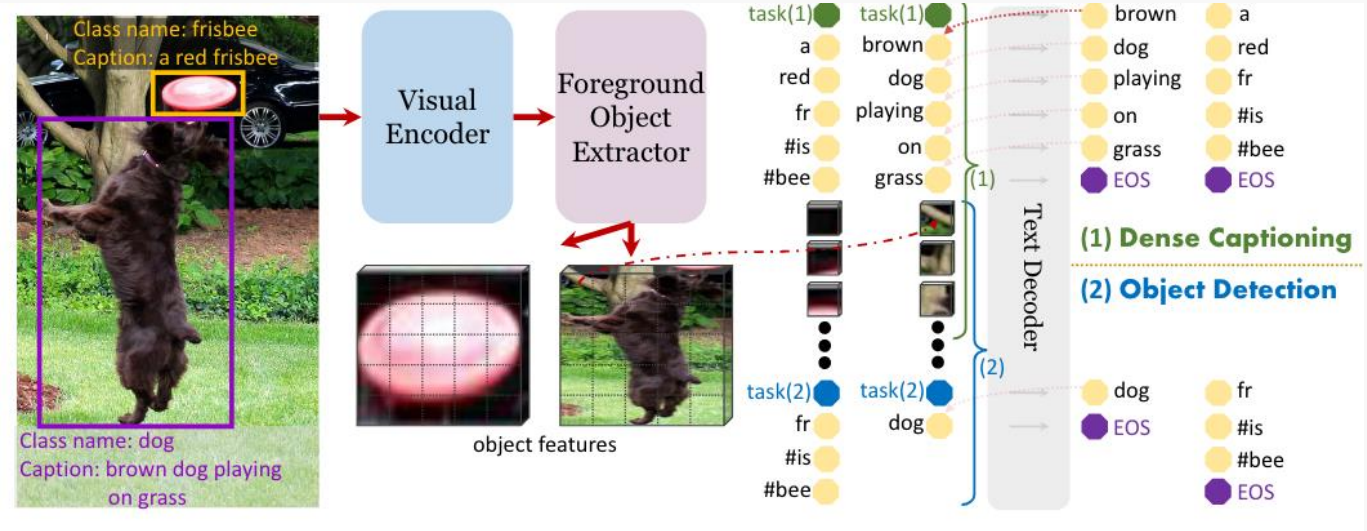


DenseCap: Fully Convolutional Localization Networks for Dense Captioning

Preliminary

- Task: Image/Dense Caption
- Model: Image encoder + Causal Language Model





- The Evaluation metrics are hard.
  - Localization + description

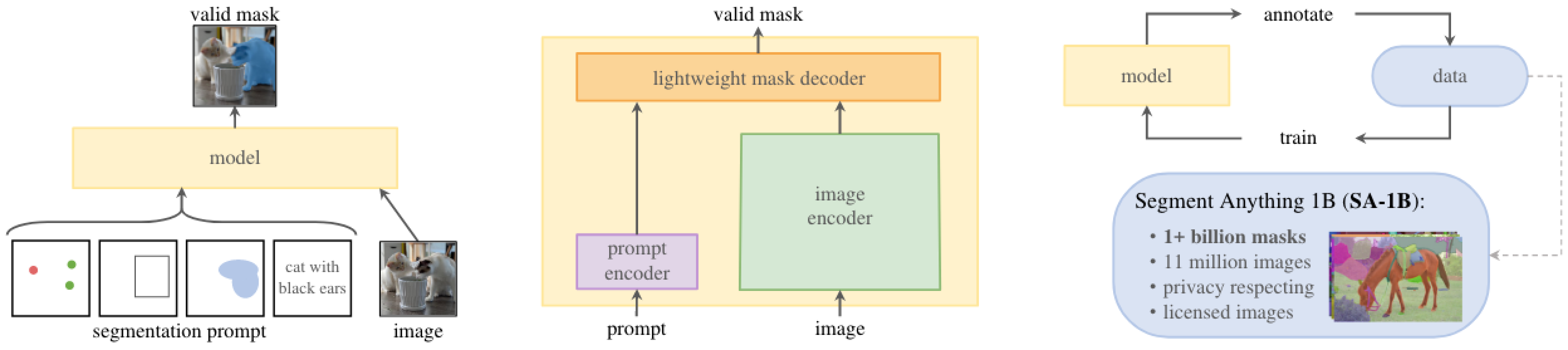
DenseCap: Fully Convolutional Localization Networks for Dense Captioning
GRiT: A Generative Region-to-text Transformer for Object Understanding

Preliminary

- Model: SAM / Language Model

Preliminary

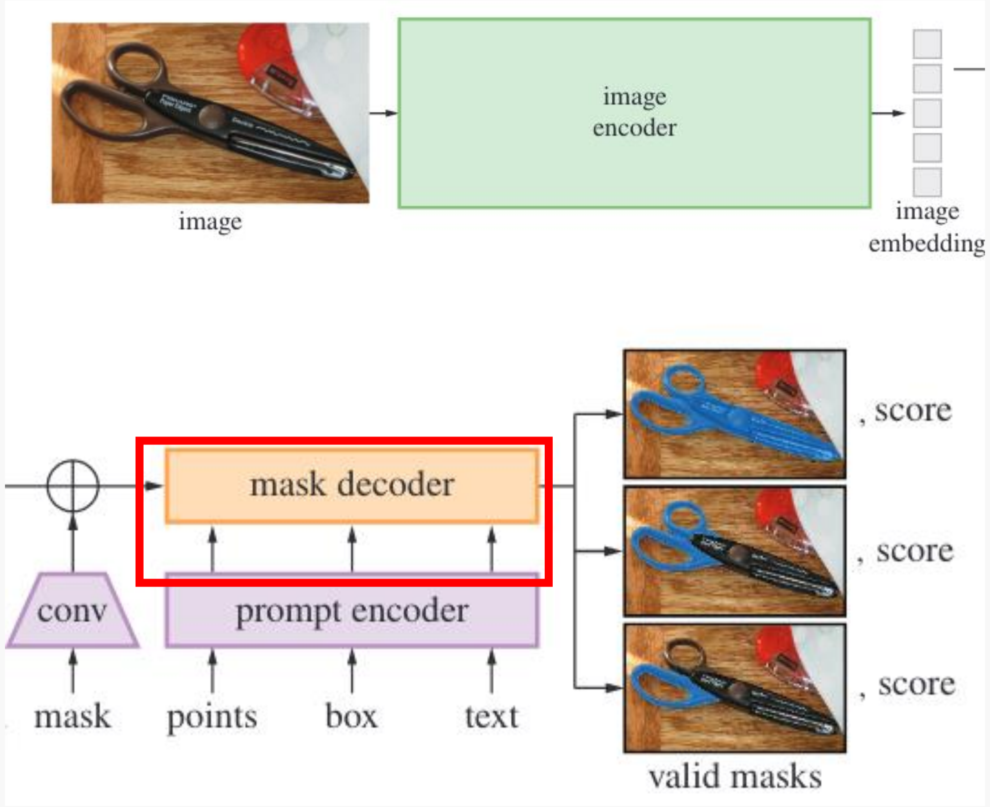
Model: SAM

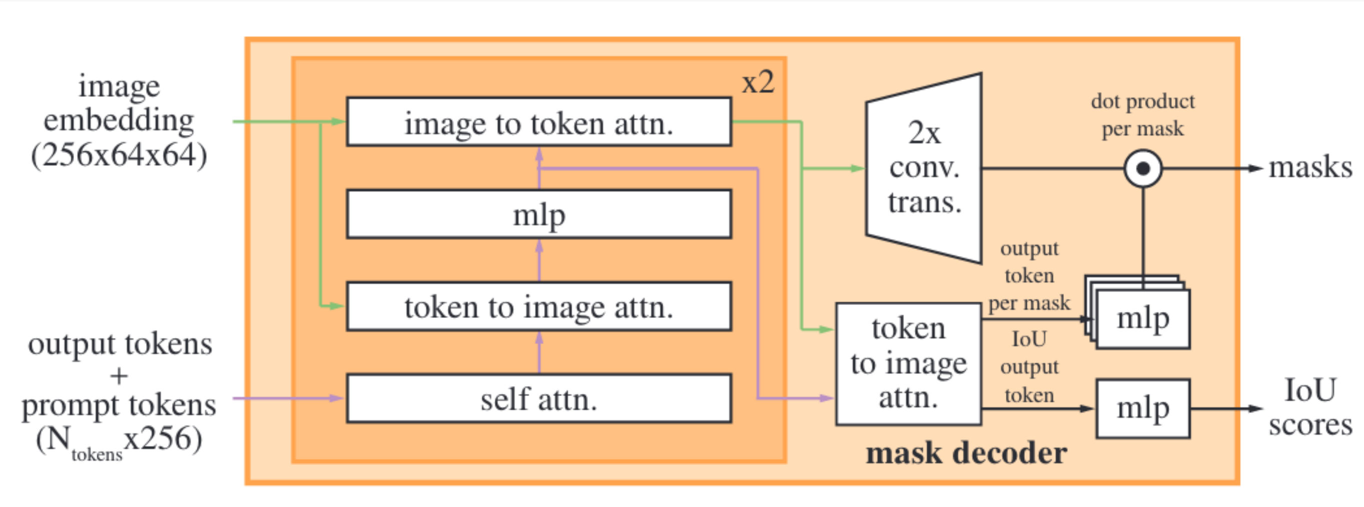


(a) **Task**: promptable segmentation (b) **Model**: Segment Anything Model (**SAM**) (c) **Data**: data engine (top) & dataset (bottom)

Preliminary

SAM's Model

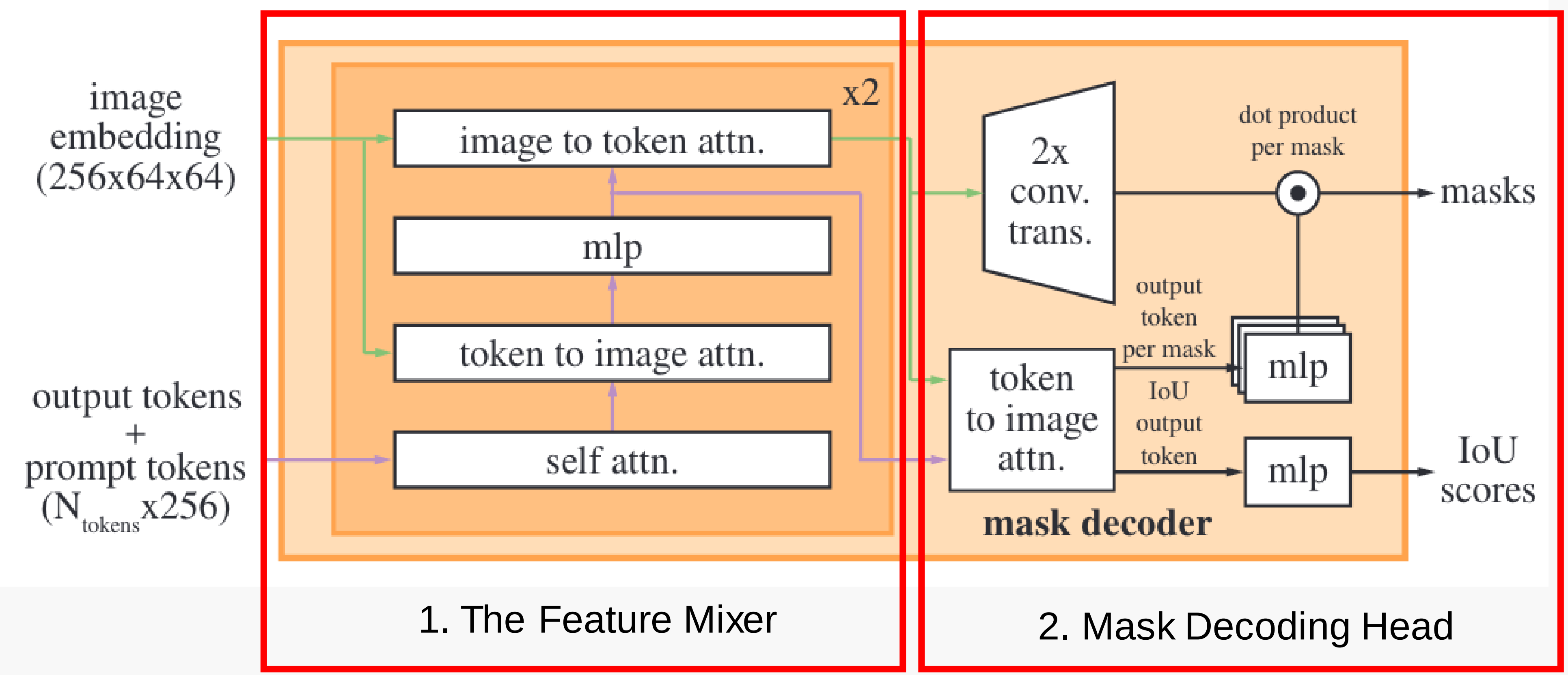




Learn about Model and Training Recipe

Preliminary

SAM's Mask Decoder



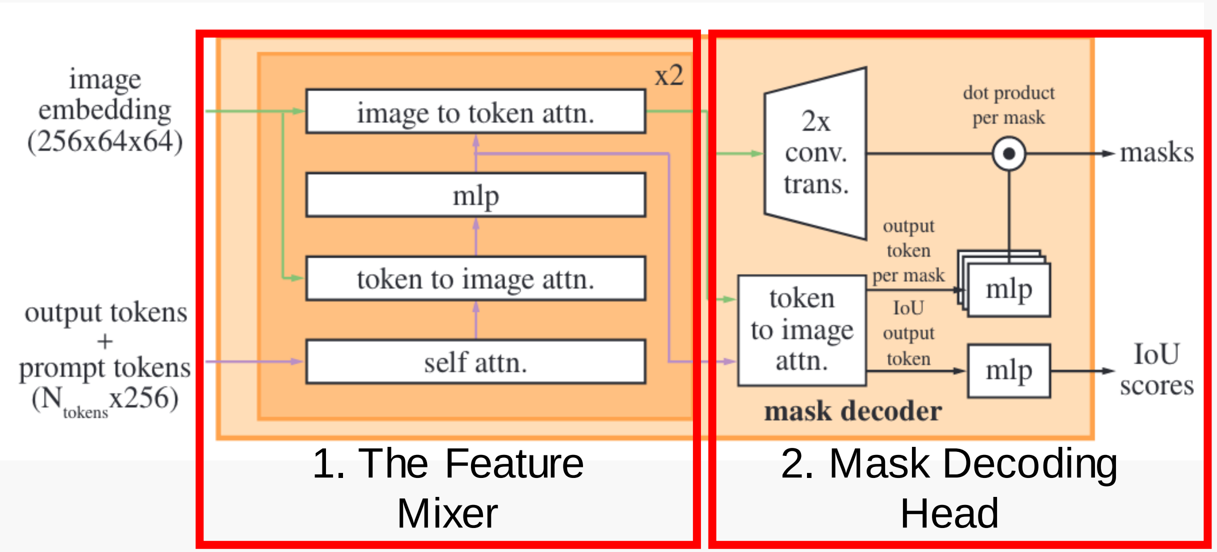
Feature Mixer: DETR

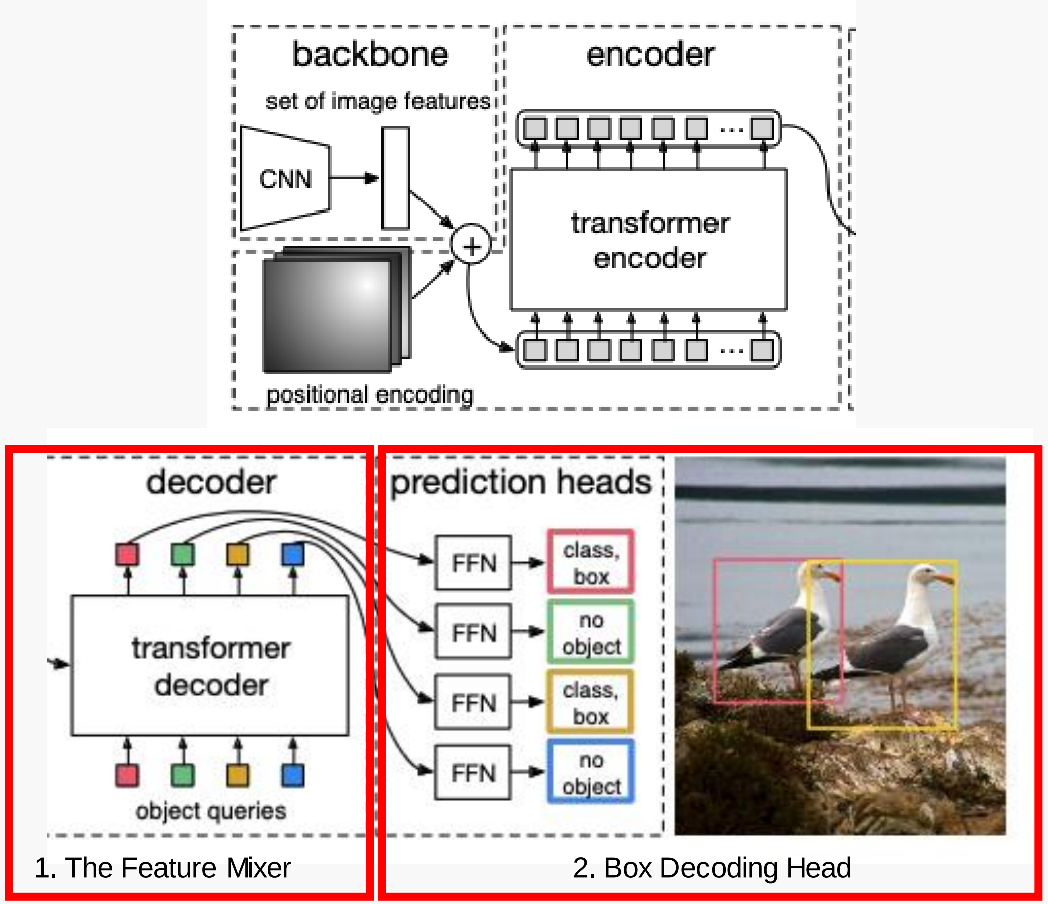
Segment Anything
End-to-End Object Detection with Transformers

i-VisionGroup

Preliminary

SAM’s Mask Decoder





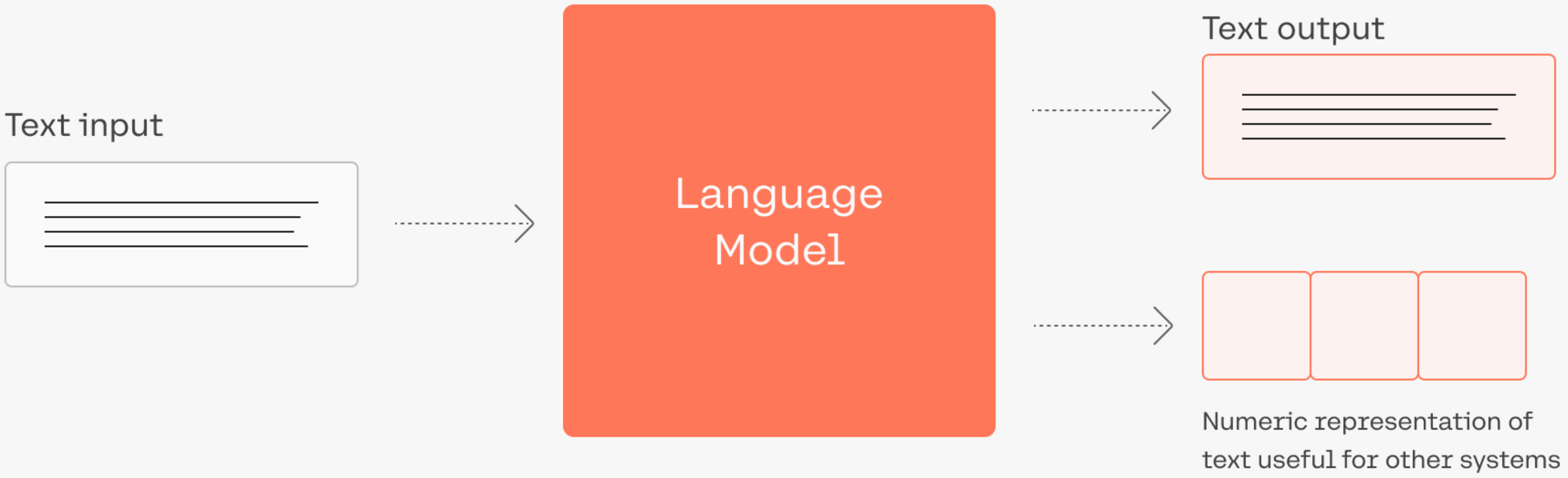
Feature Mixer: DETR

Segment Anything
End-to-End Object Detection with Transformers

Preliminary

Language Modeling:

- Modeling the dist. of language (Markov Chain)

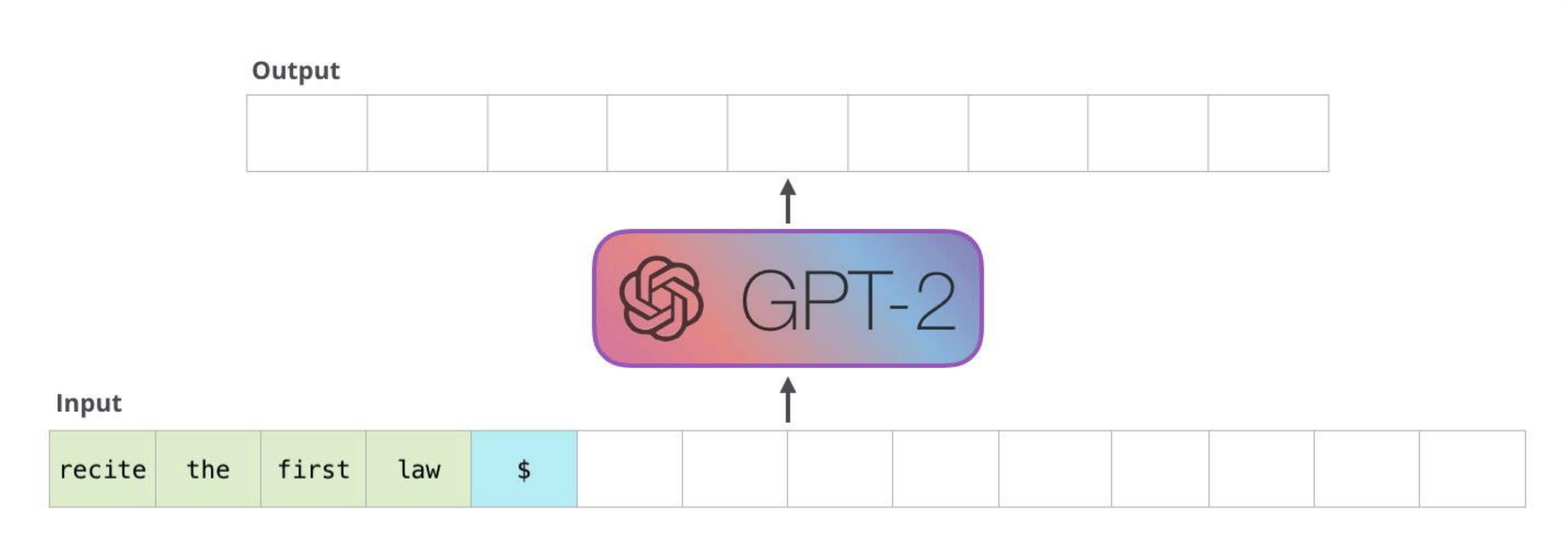


https://docs.cohere.com/docs/introduction-to-large-language-models

Preliminary

Language Modeling:

- Autoregressive + Transformer Decoder

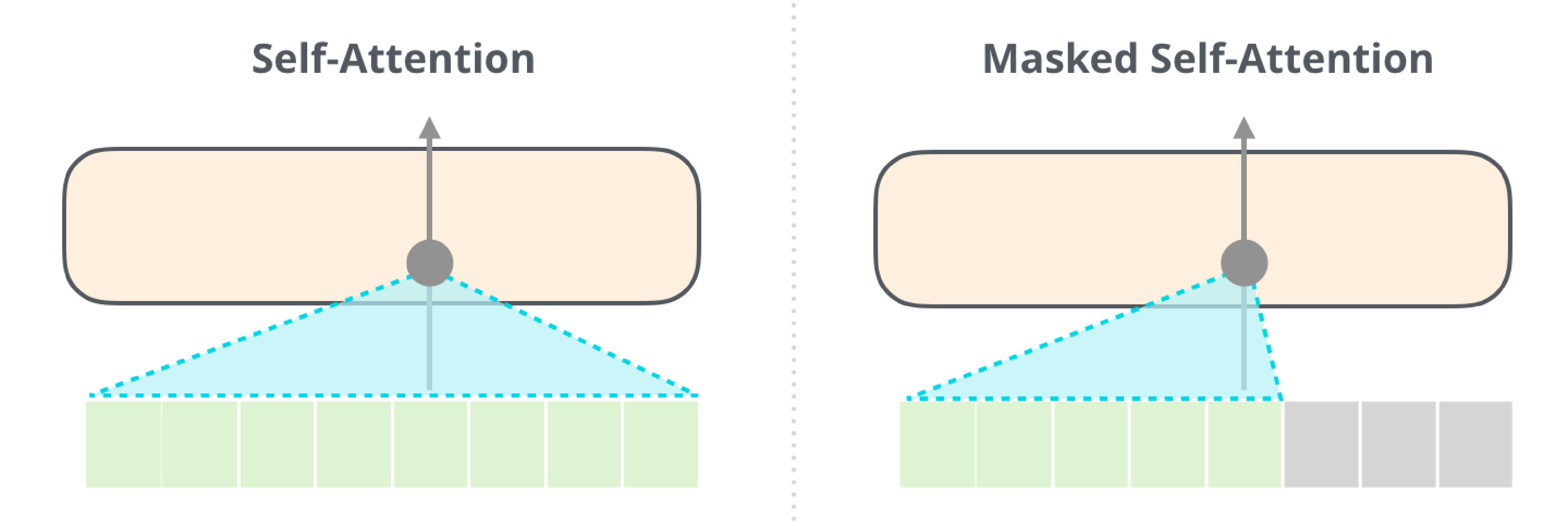


http://jalammar.github.io/illustrated-gpt2/

Preliminary

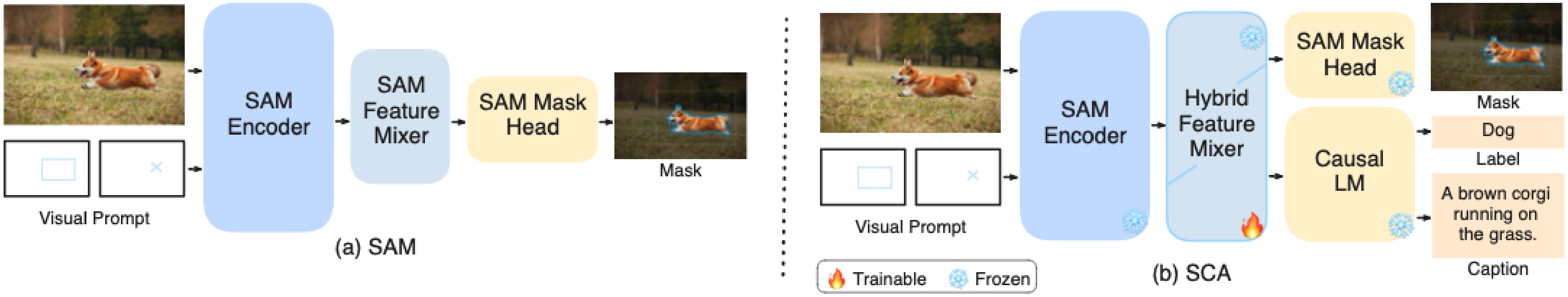
Language Modeling:

- Causal Language Model: a special form of masked modeling (Bert)



http://jalammar.github.io/illustrated-gpt2/

Segment and Caption Anything

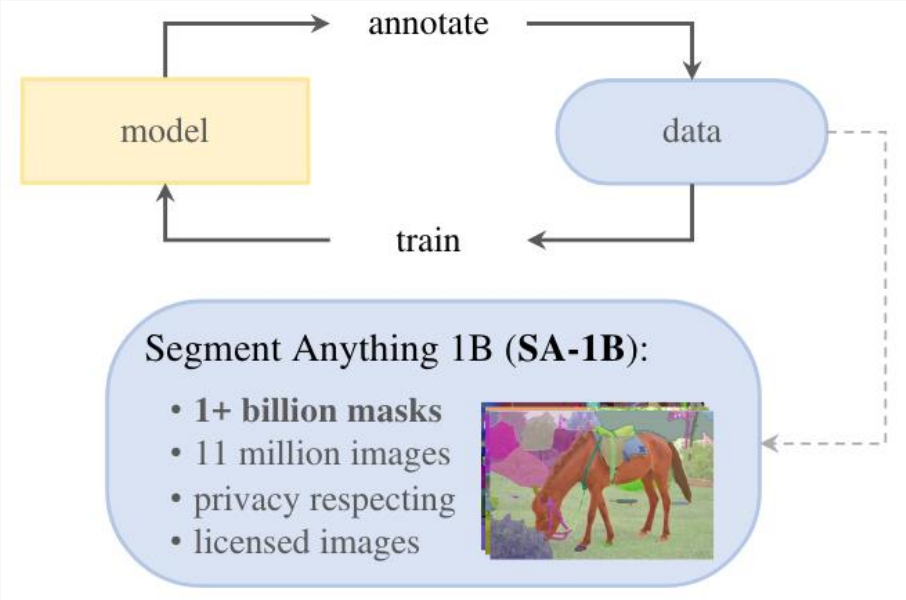


tl;dr

1. Despite the absence of semantic labels in the training data, SAM implies high-level semantics sufficient for captioning.
2. SCA (b) is a lightweight augmentation of SAM (a) with the ability to generate regional captions.
3. On top of SAM architecture, we add a fixed pre-trained language mode, and a optimizable lightweight hybrid feature mixture whose training is cheap and scalable.

Motivation

SAM as a data engine:



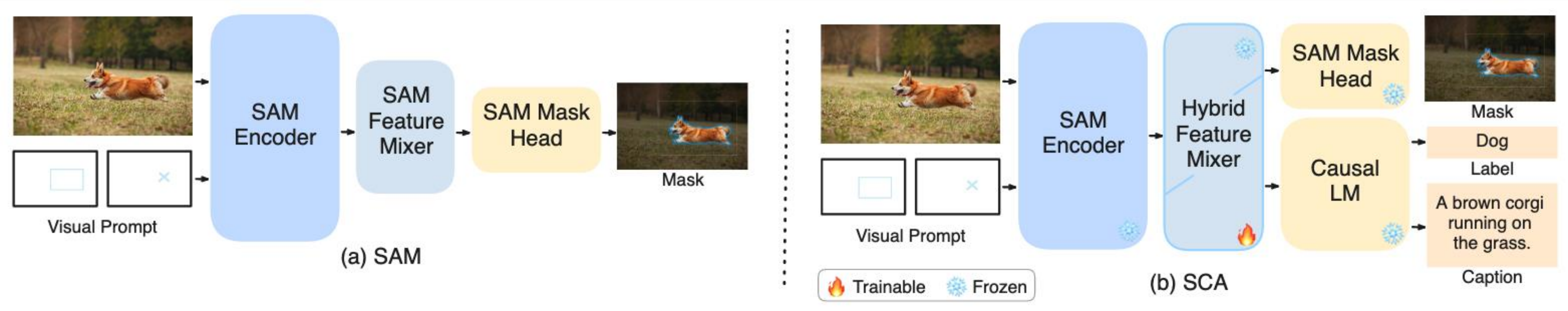
Assisted-manual stage. In the first stage, resembling classic interactive segmentation, a team of professional annotators labeled masks by clicking foreground / background object points using a browser-based interactive segmentation tool powered by SAM. Masks could be refined using pixel-precise "brush" and "eraser" tools. Our model-assisted annotation runs in real-time directly inside a browser (using precomputed image embeddings) enabling a truly interactive experience. We did not impose semantic constraints for labeling objects, and annotators freely labeled both "stuff" and "things" [1]. We suggested annotators label objects they could name or describe, but did not collect these names or descriptions. Annotators were asked to label objects in order of prominence and were encouraged to proceed to the next image once a mask took over 30 seconds to annotate.

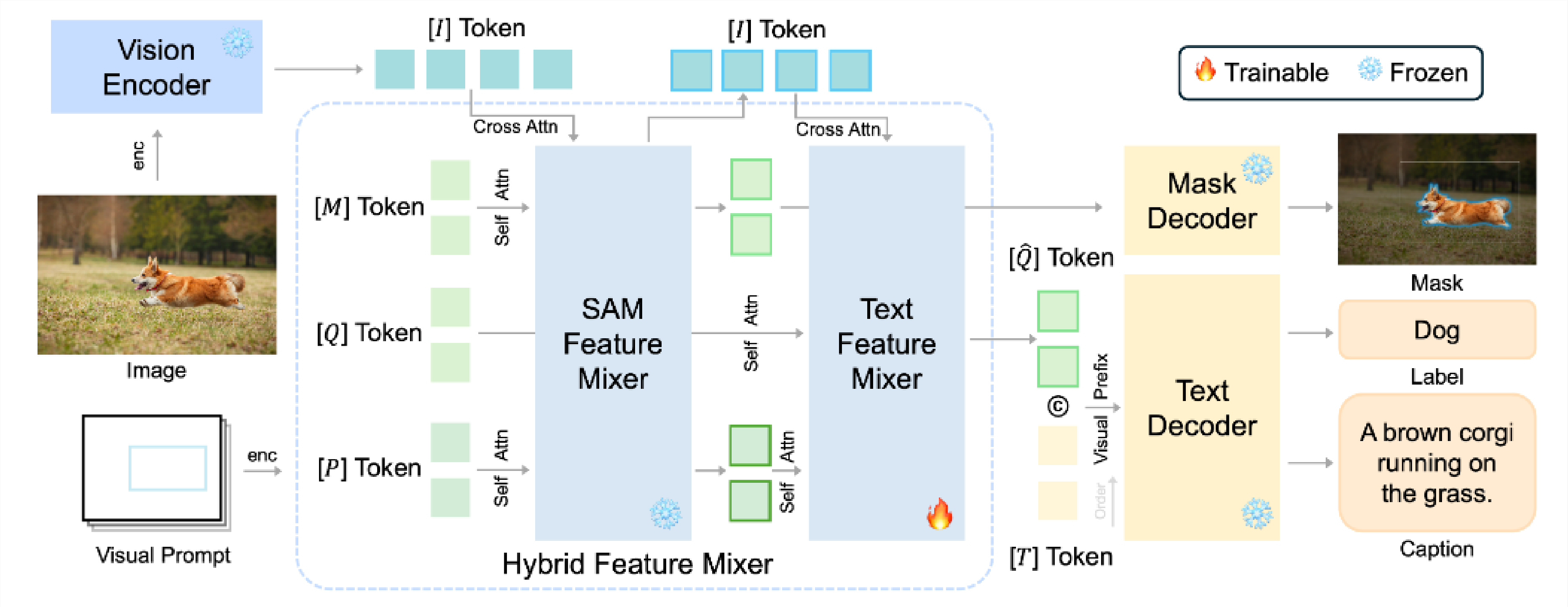
hypothesis:

- Though there is an **absence** of semantic labels, can SAM still implies **high-level semantics** sufficient for captioning?

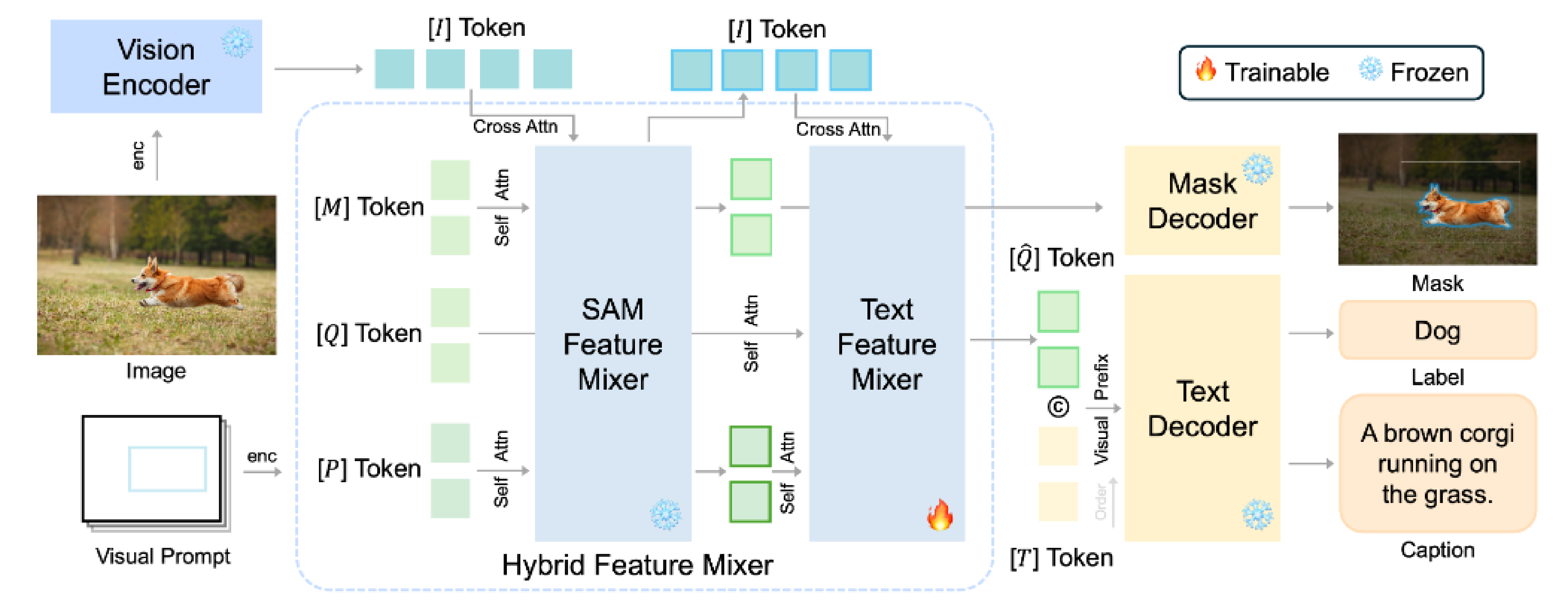
Segment Anything

Method


SAM Encoder
SAM Feature Mixer
SAM Mask Head
Mask
Visual Prompt
(a) SAM
SAM Encoder
Hybrid Feature Mixer
SAM Mask Head
Causal LM
Mask
Dog
Label
A brown corgi running on the grass.
Caption
Visual Prompt
Trainable
Frozen
(b) SCA


Vision Encoder
enc
Image
Visual Prompt
enc
[I] Token
[I] Token
Cross Attn
Cross Attn
[M] Token
[Q] Token
[P] Token
Self Attn
SAM Feature Mixer
Text Feature Mixer
Hybrid Feature Mixer
Trainable
Frozen
Mask Decoder
Text Decoder
[Q̂] Token
[T] Token
Visual
Prefix
Order
Mask
Dog
Label
A brown corgi running on the grass.
Caption

Method



Vision Encoder
enc
Image
Visual Prompt
enc
[I] Token
[I] Token
Cross Attn
Cross Attn
[M] Token
Self Attn
[Q] Token
[P] Token
Self Attn
SAM Feature Mixer
Self Attn
Self Attn
Text Feature Mixer
Hybrid Feature Mixer
Trainable
Frozen
Mask Decoder
[Q̂] Token
Visual Prefix
Order
[T] Token
Text Decoder
Mask
Dog
Label
A brown corgi running on the grass.
Caption

Method

Implementation details

- Direct Train on VG for 200K steps
- 100K Pretrain O365/COCO + 100K Finetune on VG
- 64 V100 GPUs to pre-train 32 V100 GPUs to finetune

Data Epoch*	
Batch Size*	8
# Reg / Img	16
Steps	200000
# Img	77398
# Reg	3684063
Img Epoch*	20.67
Reg Epoch*	6.95
GPU Type	V100-16GB
# GPUs	8

Model Details	
Input	a) 1024x1024 Long side: 1024 Short side: padding b) Large Scale Jitter c) Horizontal Flip
Loss	a) Cross Entropy Loss b) Label Smooth (0.1)
Text Decoder	a) GPT2-large b) Open LLAMA 3B v2
# Query Tokens	8
# Mixer Layers	12
# Task Tokens	6
Opt. Module	Text Feat. Mixer
# Opt. Params	19.4 M

Results

Dataset

dataset	type	total samples	total regions	total sents	total tokens	total words
COCO[11]	Region recognition	117,266	860,001	860,001	1,275,513	942,822
V3Det [18]	Region recognition	183,348	1,357,351	1,357,351	3,984,388	2,126,318
Objects365 [15]	Region recognition	1,742,289	25,407,598	25,407,598	49,264,696	32,341,116
Visual Genome [10]	Region captioning	77,398	3,684,063	3,684,063	21,392,494	19,740,221
RefCOCOg [21]	Referring Expression	24,698	48,599	92,671	834,305	785,259

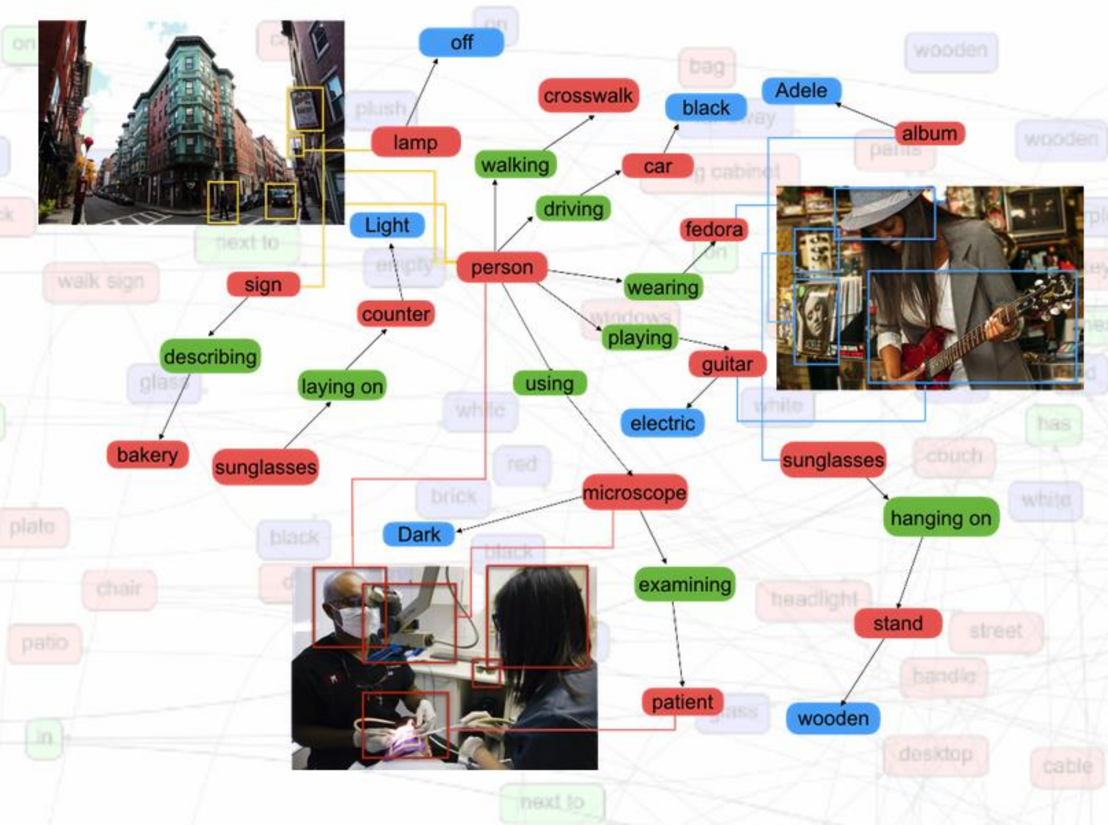
Metrics

- Reference-based metrics: Cider-D, METEOR, ROUGE, BLEU, …
- Noun / Verb IOU

清華大學 i-VisionGroup

Results

Dataset: Visual Genome



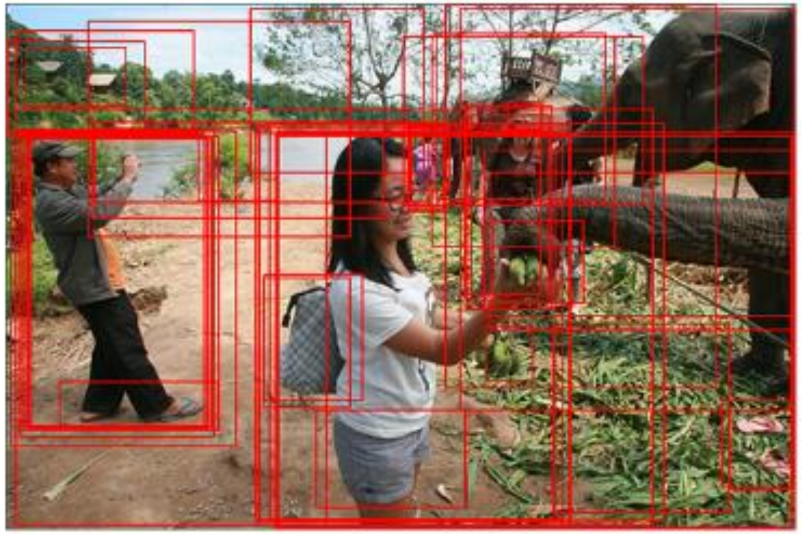
Visual Genome is a dataset, a knowledge base, an ongoing effort to connect structured image concepts to language.

108,077 Images
5.4 Million Region Descriptions
1.7 Million Visual Question Answers
3.8 Million Object Instances
2.8 Million Attributes
2.3 Million Relationships
Everything Mapped to Wordnet Synsets

Read our paper.

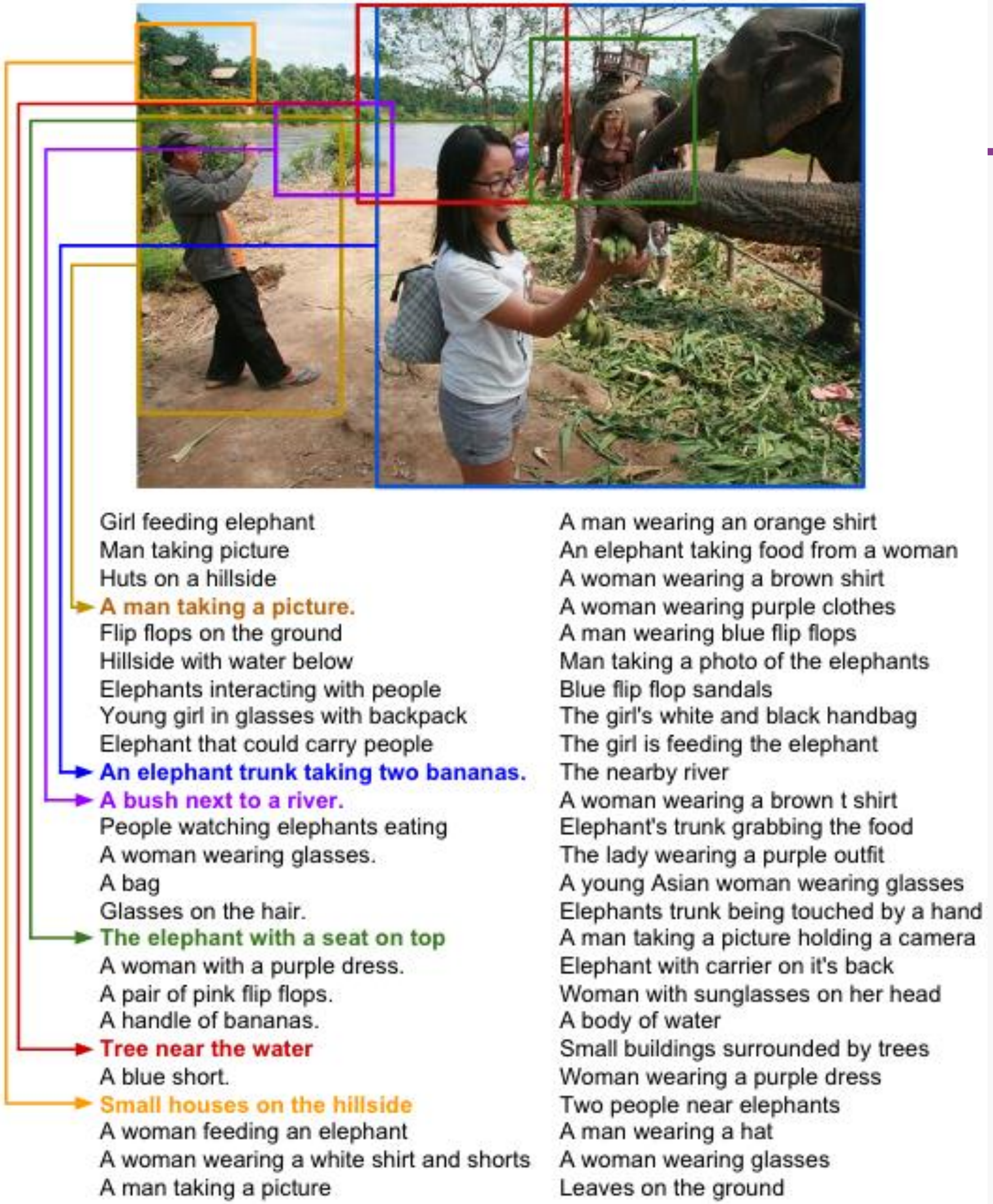
Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations

Dataset: Visual Genome



(b)

Fig. 14: (a) An example image from the dataset with its region descriptions. We only display localizations for 6 of the 42 descriptions to avoid clutter; all 50 descriptions do have corresponding bounding boxes. (b) All 42 region bounding boxes visualized on the image.



(a)

清華大學 i-VisionGroup

Results

Comparison w/ Baselines

Table 1. Comparison with baselines. "C": CIDEr-D [83], "M": METEOR [5], "S": SPICE [2], "B": BLEU [61], "R": ROUGE [49], "(F)": Fuzzy. For all metrics, the higher the better. The best, the second best, the third best scores are marked as red, orange, yellow, respectively. ∗: The captioners used in [86]. †: We pre-train the model for 100K steps, then finetune it on VG for 100K steps. ‡: When no pertaining is applied, we train the model on VG for 200K steps. Thus they have similar training costs.

Method	C	M	S	B@1	B@2	B@3	B@4	R	Noun	Verb	Noun (F)	Verb (F)
SAM+BLIP-base	43.8	9.6	12.6	16.8	7.8	3.9	2.1	19.8	21.4	3.0	49.6	8.2
SAM+BLIP-large*	25.3	11.0	12.7	14.1	6.5	3.2	1.6	18.5	27.3	4.3	56.2	12.4
SAM+GIT-base	65.5	10.1	17.1	23.6	11.7	7.1	4.8	21.8	22.7	1.4	49.8	3.0
SAM+GIT-base-coco	67.4	11.2	17.5	24.4	12.6	7.5	4.9	23.1	25.6	2.5	52.7	5.2
SAM+GIT-base-textcaps	45.6	11.6	15.0	18.4	8.9	4.7	2.7	21.8	26.1	3.5	54.2	7.4
SAM+GIT-large*	68.8	10.5	17.8	24.2	12.3	7.4	5.0	22.4	24.5	1.8	51.6	3.7
SAM+GIT-large-coco	71.8	12.2	18.8	24.6	12.9	7.7	4.9	24.4	28.9	3.4	55.8	6.7
SAM+GIT-large-textcaps	59.2	12.6	17.5	20.9	10.5	6.0	3.6	23.6	29.4	3.7	56.5	7.2
SAM+BLIP2-OPT-2.7B-coco	30.4	11.3	12.0	14.4	7.1	3.6	1.9	19.3	26.7	4.7	55.0	12.1
SAM+BLIP2-OPT-2.7B*	59.7	11.7	16.7	19.6	9.8	5.3	3.0	22.7	26.6	4.5	53.7	9.7
SAM+BLIP2-OPT-6.7B-coco	30.4	12.2	13.1	14.7	7.3	3.8	2.0	19.9	29.7	4.7	57.8	11.7
SAM+BLIP2-OPT-6.7B	56.6	11.7	16.2	19.0	9.5	5.0	2.8	22.3	26.7	4.4	53.9	10.1
GRiT	142.2	17.2	30.5	36.0	22.1	15.2	11.2	34.5	39.5	4.3	63.3	7.2
SCA (GPT2-large, VG)‡	148.8	17.4	31.2	38.0	23.9	16.6	12.1	35.5	41.5	4.8	65.0	7.6
SCA (LLAMA-3B, VG)‡	149.8	17.4	31.3	38.0	23.9	16.7	12.2	35.5	41.2	4.5	64.6	7.1
SCA (GPT2-large, Pretrain+VG)†	149.8	17.5	31.4	38.2	24.1	16.8	12.2	35.7	41.7	4.8	65.1	7.5

Results

Comparison w/ Baselines

white letter on black sticker
there is a sign that says this thing there is on the floor
the word the is written on the sign
sign in a subway station
a sign in the mirror
a sign on the wall
a sign in the mirror
a sign on the mirror says the thirst

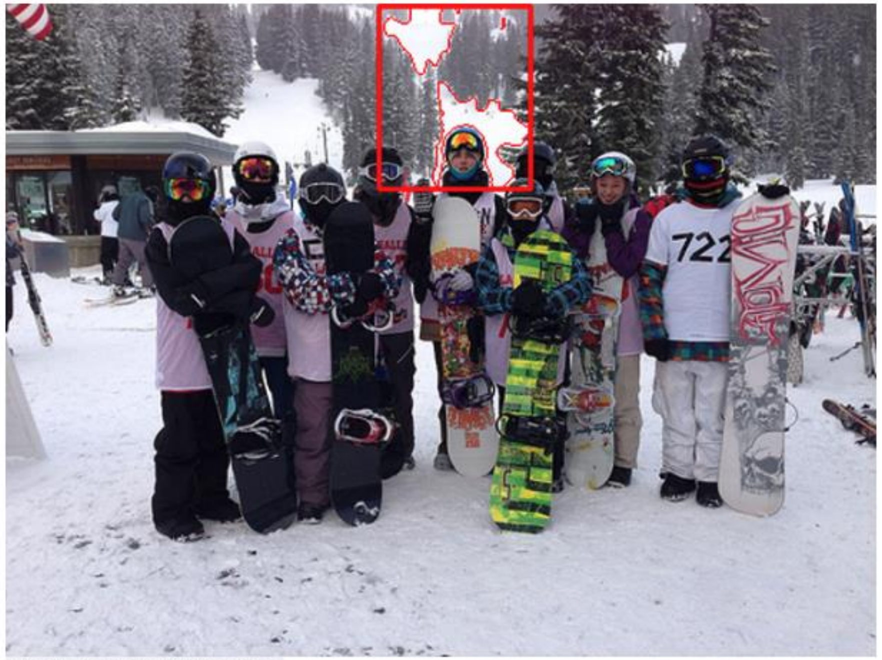
the letter a on a red shirt
there is a woman sitting at a table with a laptop computer
a woman sitting on a bed with a laptop
woman wearing a orange shirt
a red shirt on a woman
the shirt is red in color
a red shirt on a woman
red short sleeve shirt

(1) SAM+Captioner {GIT-large, BLIP-large, BLIP2-OPT-2.7B}, (2) GRIT [89], (3) SCA {GPT2-large+VG, LLAMA-3B+VG, GPT2- large+Pretrain+VG}, and (4) the ground truth.

Results

Comparison w/ Baselines

a woman holding a tennis racket
several people are walking on a tennis court with rackets
a group of people on a tennis court
a green soccer field
tennis court is green
a green tennis court
tennis court is green
a green and white tennis court



green leaves on the tree
skiers standing on a ski slope with trees in the background
a group of people standing in front of a mountain
snow covered pine trees
snow on the ground
Snow on the ground.
snow on the ground
skii slope seen in the background

(1) SAM+Captioner {GIT-large, BLIP-large, BLIP2-OPT-2.7B}, (2) GRIT [89], (3) SCA {GPT2-large+VG, LLAMA-3B+VG, GPT2- large+Pretrain+VG}, and (4) the ground truth.

Results

Comparison w/ Baselines and Large Multimodal Model (LMM)

Table 3. Comparison with referring Vision Large Language Models (VLLMs). “M”: Meteror, “C”: CIDEr-D. †: The scores are from the papers. ‡: We reproduced the result with “GPT4RoI-7B-delta-V0” from https://github.com/jshilong/GPT4RoI. The best, the second best, the third best scores are marked as red, orange, yellow, respectively.

Method	M	C
ASM [19] (Zero-shot)†	12.6	44.2
ASM (Finetuned)†	18.0	145.1
GPT4RoI [23] (7B)†	17.4	145.2
GPT4RoI (13B)†	17.6	146.8
GPT4RoI (7B)‡	16.4	122.3
SCA (GPT2-large, VG)	17.4	148.8
SCA (LLAMA-3B, VG)	17.4	149.8
SCA (GPT2-large, Pretrain+VG)	17.5	149.8

Table 5. The zero-shot performance on the Referring Expression Generation (REG) task. “M”: Meteror, “C”: CIDEr-D. ∗: “k” means th number of examples in the prompt. †: The scores are from the papers.

Method	RefCOCOg		RefCOCO+				RefCOCO			
	val		testA		testB		testA		testB	
	M	C	M	C	M	C	M	C	M	C
separate train/test										
Visdif [21]†	14.5	-	14.2	-	13.5	-	18.5	-	24.7	-
SLR [22]†	15.9	66.2	21.3	52.0	21.5	73.5	29.6	77.5	34.0	132.0
zero-shot										
Kosmos-2 [13]†	12.2	60.3	-	-	-	-	-	-	-	-
Kosmos-2 (k=2)*†	13.8	62.2	-	-	-	-	-	-	-	-
Kosmos-2 (k=4)*†	14.1	62.2	-	-	-	-	-	-	-	-
ASM [19]†	13.6	41.9	-	-	-	-	-	-	-	-
GRiT [20]	15.2	71.6	-	-	-	-	-	-	-	-
SCA (GPT2-large, Pretrain+VG)	15.4	71.9	21.7	29.2	20.4	57.2	20.4	27.0	20.2	66.4
SCA (GPT2-large, VG)	15.3	70.5	21.7	30.2	20.1	56.6	20.5	27.7	20.1	66.7
SCA (LLAMA-3B, VG)	15.6	74.0	22.0	30.0	20.2	56.1	20.7	27.3	20.3	65.3

Results

Ablations:

Table 2. Comparison of using different image encoders. “C”: CIDEr-D, “M”: Meteror.

Image Encoder	C	M
vit_large_patch14_clip_336.openai	67.3	10.2
vit_large_patch14_clip_224.datacompxl	59.0	9.3
eva02_large_patch14_clip_336.merged2b	53.9	8.8
vit_large_patch14_reg4_dinov2.lvd142m	76.4	11.2
vit_large_patch16_224.mae	59.6	9.4
Add optimization of sam feature mixer		
vit_large_patch14_clip_336.openai	66.7	10.1
vit_large_patch14_clip_224.datacompxl	60.3	9.5
eva02_large_patch14_clip_336.merged2b	54.2	8.8
vit_large_patch14_reg4_dinov2.lvd142m	76.1	11.1
vit_large_patch16_224.mae	59.2	9.4
SAM		
SAM-ViT-base	130.2	16.0
SAM-ViT-large	129.6	15.9
SAM-ViT-huge	130.9	16.0

Table 2. The ablation of pretraining with weak supervision. $*$: The model is trained solely on VG [36] for *100K* steps. †: The model is first pre-trained for 100K, and then it is fine-tuned for 100K. The training setting for ablations is different from that of Tab. 1.

Pretrain	C	M	S
*No Pretrain**	127.9	15.8	27.7
COCO [50] (img. 117K, cls. 80)†	130.2	16.0	28.0
V3Det [85] (img. 183K, cls. 13K)†	130.4	16.0	28.0
O365 [72] (img. 1M, cls. 365)†	134.5	16.3	28.7

Table 4. The effect of different number of layers in the feature mixer. Note that this is the *only* trainable module in our models.

# of Layers	# of Params	C	M	S
2	3.3 M	108.8	13.6	24.6
4	6.5 M	109.8	14.0	25.6
8	12.8 M	127.0	15.3	27.8
12	19.1 M	127.7	15.3	27.9
24	38.0 M	124.5	15.0	27.3

All exp. were conducted with 8 V100 GPUs.

清華大学 i-VisionGroup

Results

Ablations:

Table 5. The ablation of feature mixer design.

Method	C	M	S
ROI Align [26]	45.2	9.4	11.6
ROI Align + MLP [52]	82.5	12.1	19.3
SAM Query [35]	130.6	15.9	28.4
Text Query w/o SAM Tokens	136.6	16.4	29.2
Text Query w/ SAM Tokens	137.4	16.5	29.3

Table 7. The ablation of using data augmentation. "LM": Language model, "Aug.": Augmentation.

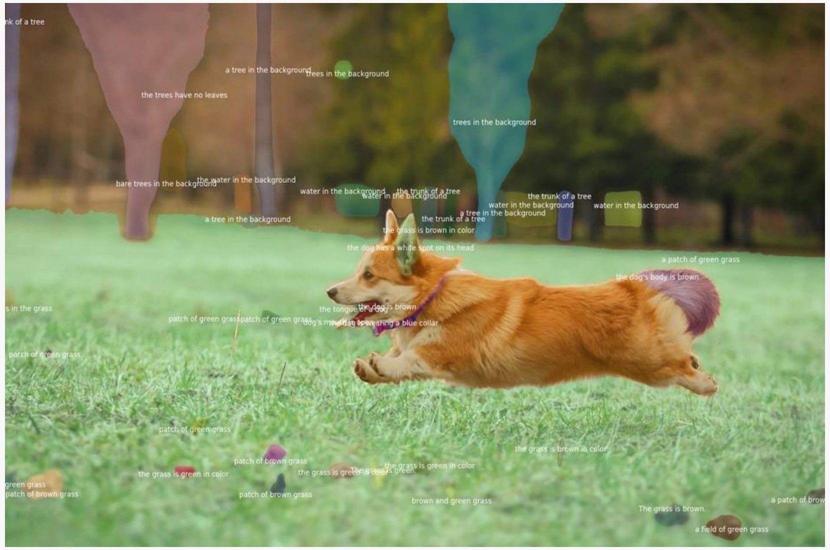
LM	Aug.	C	M	S
GPT2-large	No LSJ	137.6	16.5	29.3
	LSJ (1.0, 2.0)	140.2	16.7	29.9
	LSJ (0.1, 2.0)	140.8	16.7	29.9
LLAMA-3B	No LSJ	137.7	16.4	29.2
	LSJ (1.0, 2.0)	142.1	16.7	30.0
	LSJ (0.1, 2.0)	142.6	16.8	30.1

All exp. were conducted with 8 V100 GPUs.

Table 3. The ablation of training settings of the feature mixer ar the text decoder. "M.": Feature mixer, "T.D.": Text decoder.

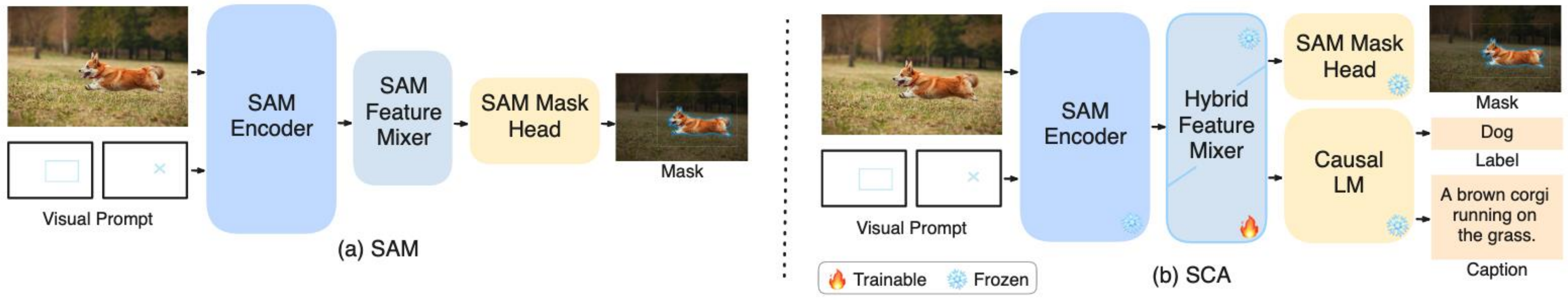
M. LR	T.D.	T.D. LR	C	M	S
1e-4	GPT2 -large	5e-6	135.6	16.3	28.5
		1e-6	134.8	16.2	28.5
		5e-7	134.5	16.2	28.5
		1e-7	135.6	16.4	28.8
		0.0	136.0	16.5	28.9
5e-5	GPT2 -large	5e-6	129.1	15.7	27.5
		1e-6	131.4	15.9	28.0
		5e-7	131.2	16.0	28.0
		1e-7	132.5	16.1	28.2
		0.0	131.7	16.1	28.2
1e-4	GPT2	5e-6	134.1	16.2	28.4
		1e-6	134.7	16.3	28.7
		5e-7	134.5	16.2	28.7
		1e-7	133.2	16.1	28.6
		0.0	132.3	15.9	28.9
5e-5	GPT2	5e-6	131.3	16.0	28.0
		1e-6	131.1	16.0	28.1
		5e-7	130.6	15.9	28.1
		1e-7	130.4	15.9	28.2
		0.0	126.3	15.4	27.9

Anything Mode



Conclusion, and What's Next?

Conclusions



tl;dr

1. Despite the absence of semantic labels in the training data, SAM implies high-level semantics sufficient for captioning.
2. SCA (b) is a lightweight augmentation of SAM (a) with the ability to generate regional captions.
3. On top of SAM architecture, we add a fixed pre-trained language mode, and a optimizable lightweight hybrid feature mixture whose training is cheap and scalable.

Conclusion, and What's Next?

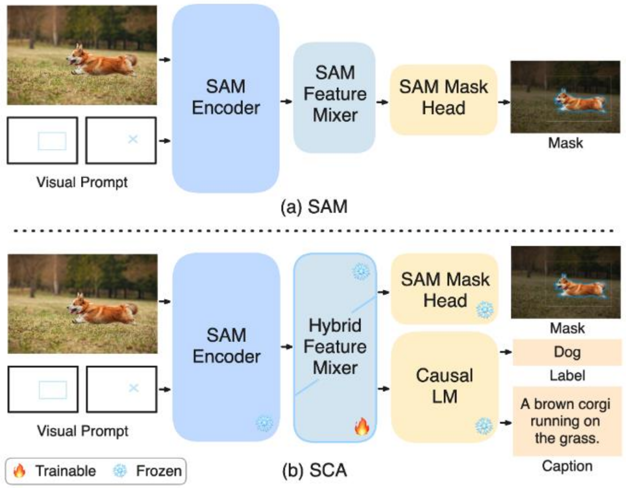
What's Next?

- **Solve Limitations**:
  - wrong attribute prediction,
  - distinguishing similar visual concepts,
  - and alignment with mask predictions.
- We hope this work serves as a stepping stone towards **scaling regional captioning data**
- and exploring **emerging abilities** in vision from low-level data or pre-trains.

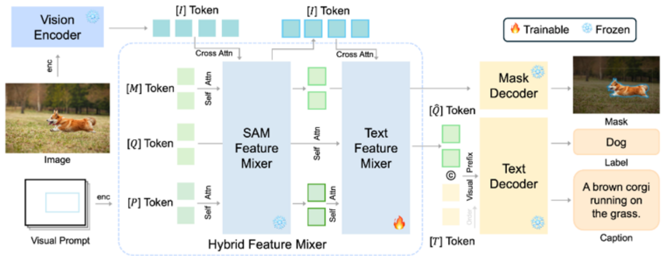
Segment and Caption Anything

Xiaoke Huang[1], Jianfeng Wang[2], Yansong Tang[1], Zheng Zhang[2], Han Hu[2], Jiwen Lu[1], Lijuan Wang[2], Zicheng Liu[3]
[1]Tsinghua University, [2]Microsoft, [3]AMD

Introduction



Method



We found that the regional features of SAM (Segment Anything Model) can be used for regional captioning.

Thus we proposed a lightweight query-based feature mixer to connect SAM with Causal Language Model.

Project Page & Code

Comparison

Method	M	C
ASM [20] (Zero-shot)†	12.6	44.2
ASM (Finetuned)†	18.0	145.1
GPT4RoI [24] (7B)†	17.4	145.2
GPT4RoI (13B)†	17.6	146.8
GPT4RoI (7B)‡	16.4	122.3
SCA (GPT2-large, VG)	17.4	148.8
SCA (LLAMA-3B, VG)	17.4	149.8
SCA (GPT2-large, Pretrain+VG)	17.5	149.8

Pre-train or not

Pretrain	C	M	S
*No Pretrain**	127.9	15.8	27.7
COCO [54] (img. 117K, cls. 80)†	130.2	16.0	28.0
V3Det [94] (img. 183K, cls. 13K)†	130.4	16.0	28.0
O365 [81] (img. 1M, cls. 365)†	134.5	16.3	28.7

Anything Mode

Training Recipe

M. LR	T.D.	T.D. LR	C	M	S
1e-4	GPT2 -large	5e-6	135.6	16.3	28.5
		1e-6	134.8	16.2	28.5
		5e-7	134.5	16.2	28.5
		1e-7	135.6	16.4	28.8
		0.0	136.0	16.5	28.9
5e-5	GPT2 -large	5e-6	129.1	15.7	27.5
		1e-6	131.4	15.9	28.0
		5e-7	131.2	16.0	28.0
		1e-7	132.5	16.1	28.2
		0.0	131.7	16.1	28.2
1e-4	GPT2	5e-6	134.1	16.2	28.4
		1e-6	134.7	16.3	28.7
		5e-7	134.5	16.2	28.7
		1e-7	133.2	16.1	28.6
		0.0	132.3	15.9	28.9
5e-5	GPT2	5e-6	131.3	16.0	28.0
		1e-6	131.1	16.0	28.1
		5e-7	130.6	15.9	28.1
		1e-7	130.4	15.9	28.2
		0.0	126.3	15.4	27.9