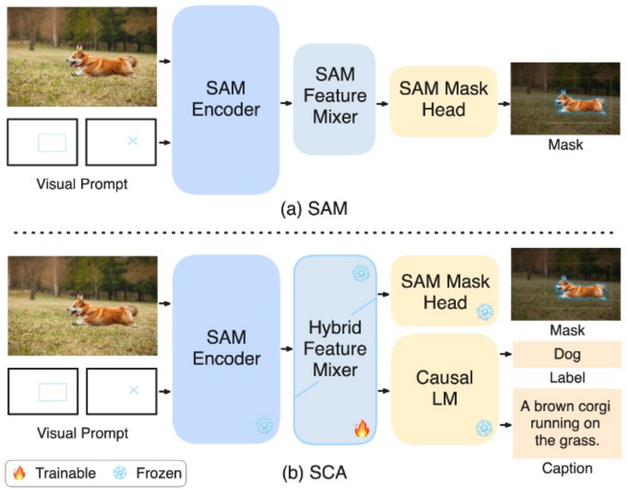
Segment and Caption Anything

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

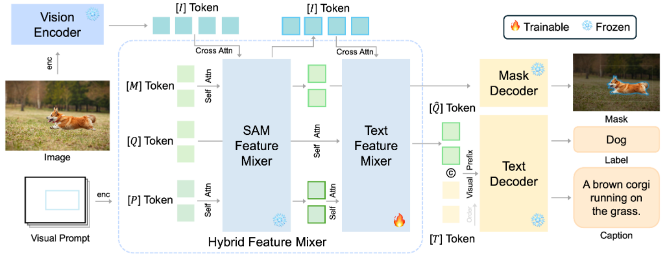
Introduction



(a) SAM

(b) SCA

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