南京大学
NANJING UNIVERSITY

SAM介绍

蔡文朴
caiwenpu@smail.nju.edu.cn

大数据智能研究组
南京大学计算机科学与技术系
软件新技术国家重点实验室

2023年7月11日

Segment Anything (SAM)

NANJING UNIVERSITY

Segment Anything

Alexander Kirillov[1,2,4] Eric Mintun[2] Nikhila Ravi[1,2] Hanzi Mao[2] Chloe Rolland[3] Laura Gustafson[3]
Tete Xiao[3] Spencer Whitehead Alexander C. Berg Wan-Yen Lo Piotr Dollár[4] Ross Girshick[4]
[1]project lead [2]joint first author [3]equal contribution [4]directional lead
Meta AI Research, FAIR

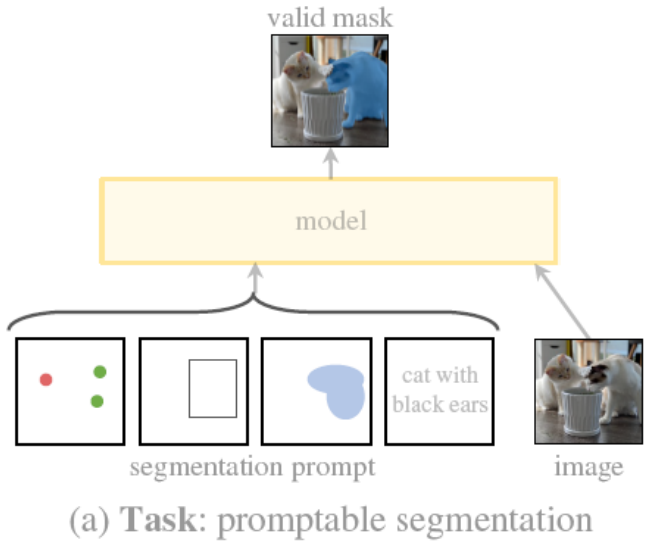
Motivation

- A foundation model for image segmentation.
- A promptable model and pre-train it on a broad dataset with powerful generalization.
- Transfer to downstream segmentation on new data distributions using prompt engineering.

Segment Anything (SAM)

南京大学 NANJING UNIVERSITY

SAM任务



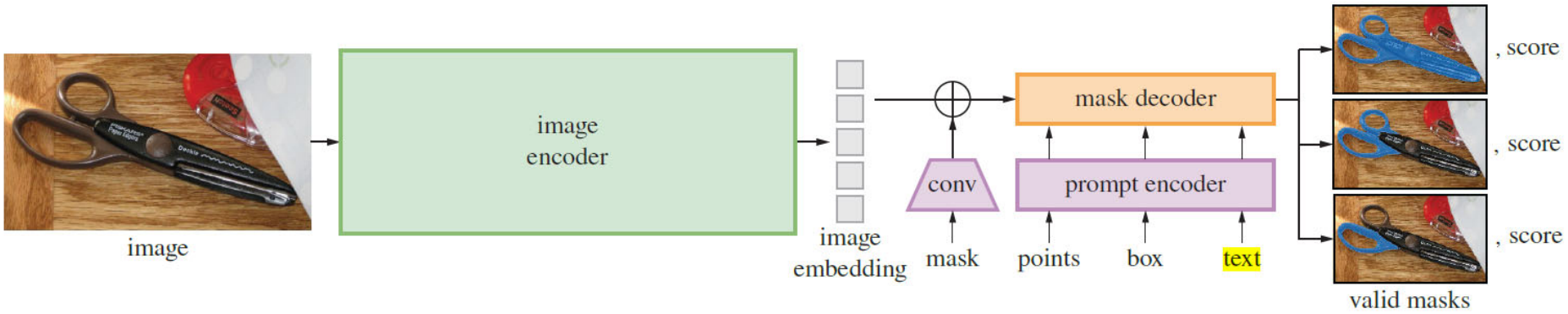
(a) **Task**: promptable segmentation

点 框 文本

Segment Anything (SAM)

NANJING UNIVERSITY

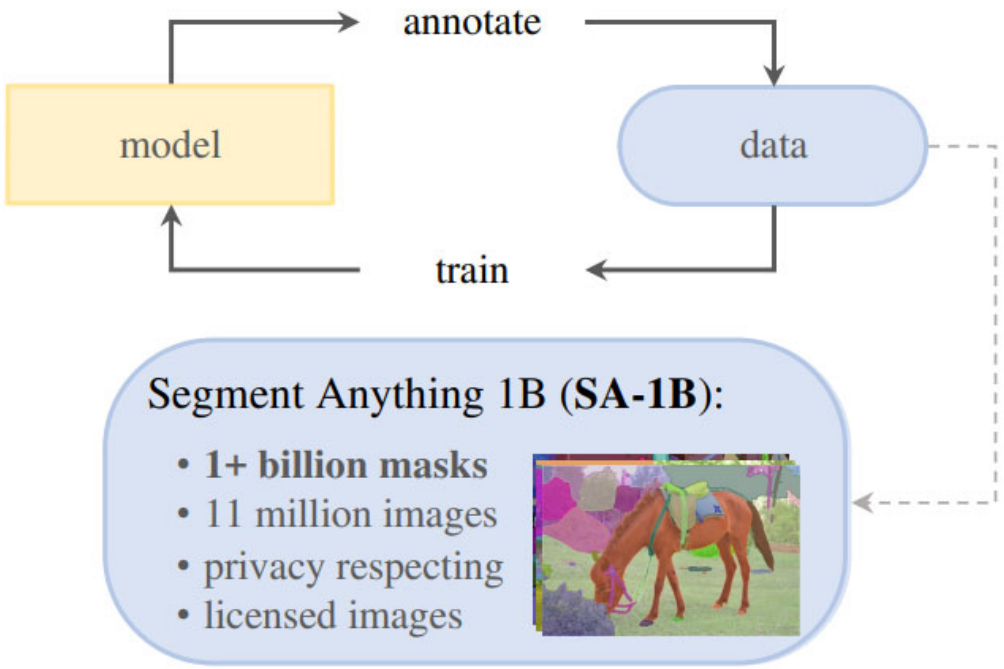
SAM模型



- Image encoder: MAE pretrained ViT.
- prompt encoder: positional encodings summed with learned embeddings
- mask decoder: lightweight Transformer, reuse image embedding for various prompts.
- Resolving ambiguity: 3 mask outputs, (whole, part, and subpart). rank by iou score.
- Losses and training: mask pixel loss; randomly sampling prompts in 11 rounds per mask
- Inference
  - w/ prompt: inference one-time with the prompt
  - w/o prompt:

    1. sample grid prompt from the image (e.g 32x32), inference many times with all the prompts.
    2. Postprocess: NMS, IOU-threshold, stability-filter, remove small component, fill holes

SAM数据



(1) a model-assisted manual annotation stage
- Firstly, SAM was trained using common public segmentation dataset
- Label masks by clicking and refining.

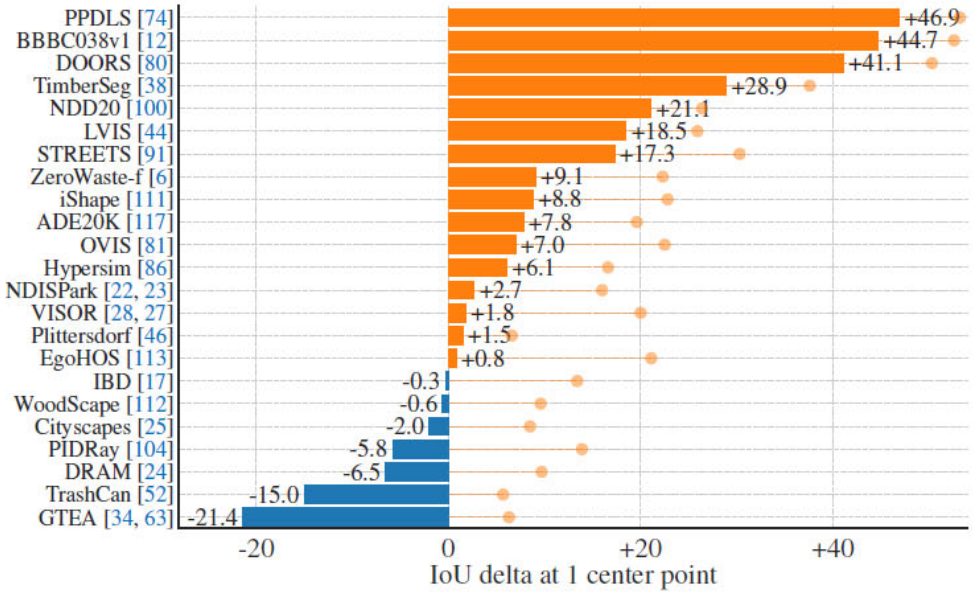
(2) a semi-automatic stage
- Present annotators with images prefilled with confident masks.
- Ask them to annotate any additional unannotated objects.

(3) a fully automatic stage
- model generates masks without annotator input.

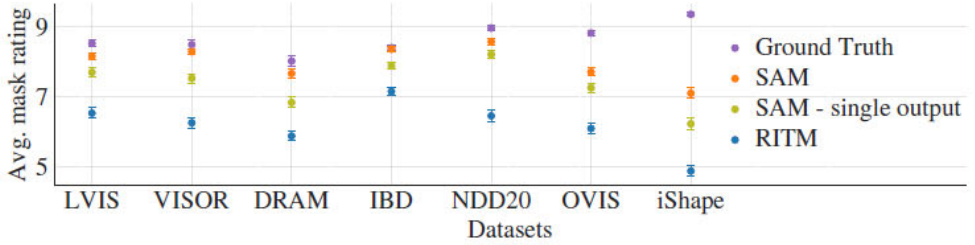
Segment Anything (SAM)

南京大学 NANJING UNIVERSITY

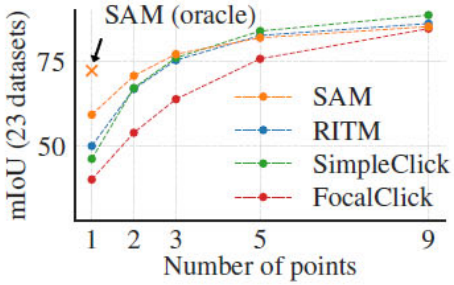
Compare with interactive segmenters.



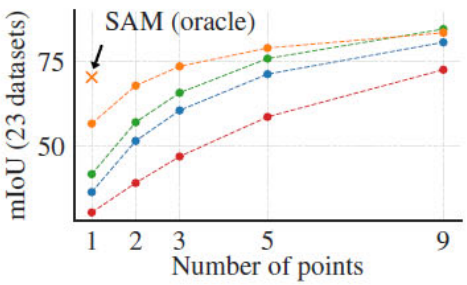
(a) SAM *vs.* RITM [92] on 23 datasets



(b) Mask quality ratings by human annotators



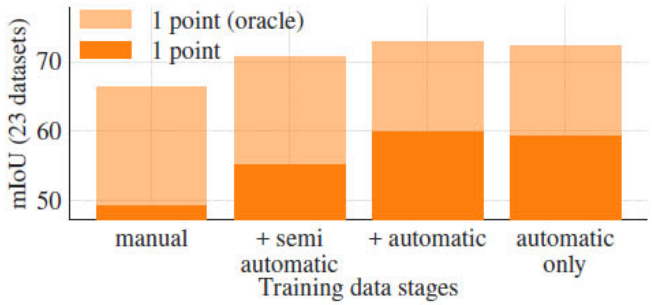
(c) Center points (default)

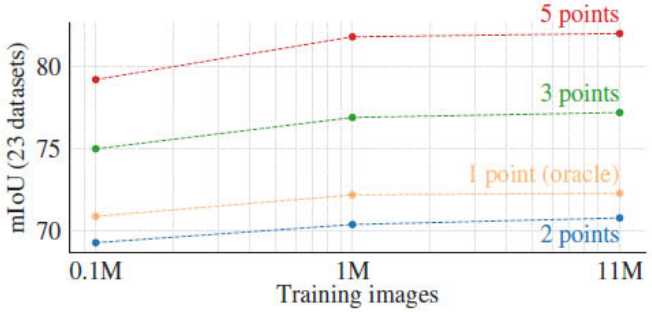


(d) Random points

circles show “oracle” results of the most relevant of SAM’s 3 predictions.

As the points increase, gap between methods decreases.





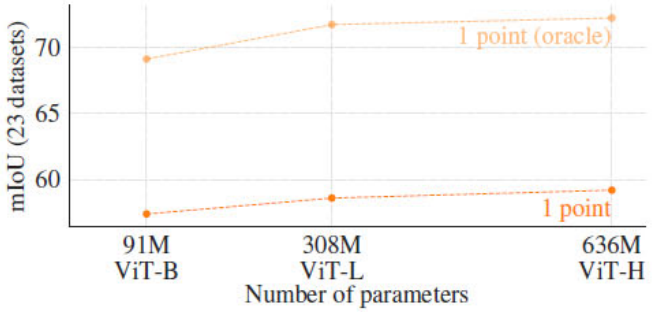


Figure 13: Ablation studies of our data engine stages, image encoder scaling, and training data scaling. (Left) Each data

Segment Anything (SAM)

NANJING UNIVERSITY

Zero-Shot Edge Detection

method	year	ODS	OIS	AP	R50
HED [108]	2015	.788	.808	.840	.923
EDETR [79]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [13]	1986	.600	.640	.580	-
Felz-Hutt [35]	2004	.610	.640	.560	-
SAM	2023	.768	.786	.794	.928

Table 3: Zero-shot transfer to edge detection on BSDS500.

Zero-Shot Object Proposals

underperforms ViTDet-H on small objects and frequent objects.

	mask AR@1000						
method	all	small	med.	large	freq.	com.	rare
ViTDet-H [62]	63.0	51.7	80.8	87.0	63.1	63.3	58.3
zero-shot transfer methods:							
SAM – single out.	54.9	42.8	76.7	74.4	54.7	59.8	62.0
SAM	59.3	45.5	81.6	86.9	59.1	63.9	65.8

Table 4: Object proposal generation on LVIS v1. SAM is applied zero-shot, *i.e.* it was not trained for object proposal generation nor did it access LVIS images or annotations.

Zero-Shot Instance Segmentation

	COCO [66]				LVIS v1 [44]			
method	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L
ViTDet-H [62]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5

Table 5: Instance segmentation results. SAM is prompted with ViTDet boxes to do zero-shot segmentation. The fully-supervised ViTDet outperforms SAM, but the gap shrinks on the higher-quality LVIS masks. Interestingly, SAM outperforms ViTDet according to human ratings (see Fig. 11).

COCO ground truth quality is relatively low, ViTDet learns the specific biases of COCO masks.

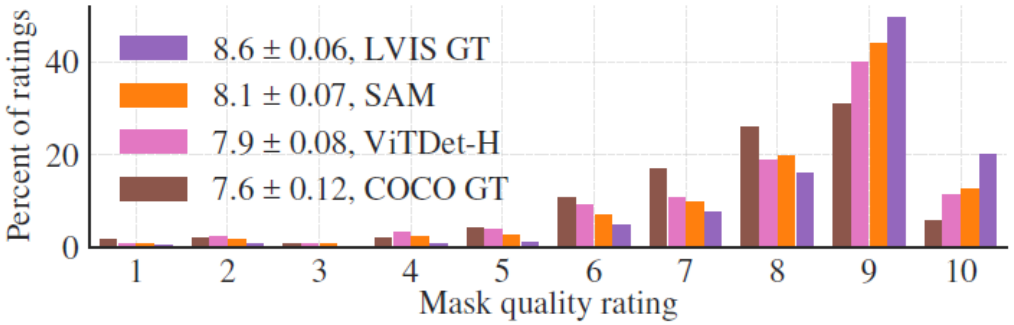


Figure 11: Mask quality rating distribution from our human study for ViTDet and SAM, both applied to LVIS ground truth boxes. We also report LVIS and COCO ground truth quality. The legend shows rating means and 95% confi-

Segment Anything (SAM)

NANJING UNIVERSITY

- Conclusion
  - Strengths:
    - Excellet zero-shot segmentation performance across a wide range of tasks.
  - Weaknesses：
    - SAM can miss fine structures and can hallucinate small disconnected components.
    - No mask semantic label.
    - Slow inference w/o prompt.

Example error for ‘Incorrect holes in the mask’: This mask

Example error for ‘Poor edge quality’: The mask has poor

SAM improvement

- SAM prompt
  - SegGPT, Seem, PerSAM
- SAM fine-tuning
  - HQ-SAM
- SAM acceleration
  - FastSAM
- SAM application
  - Inpainting，Matte Anything

NANJING UNIVERSITY

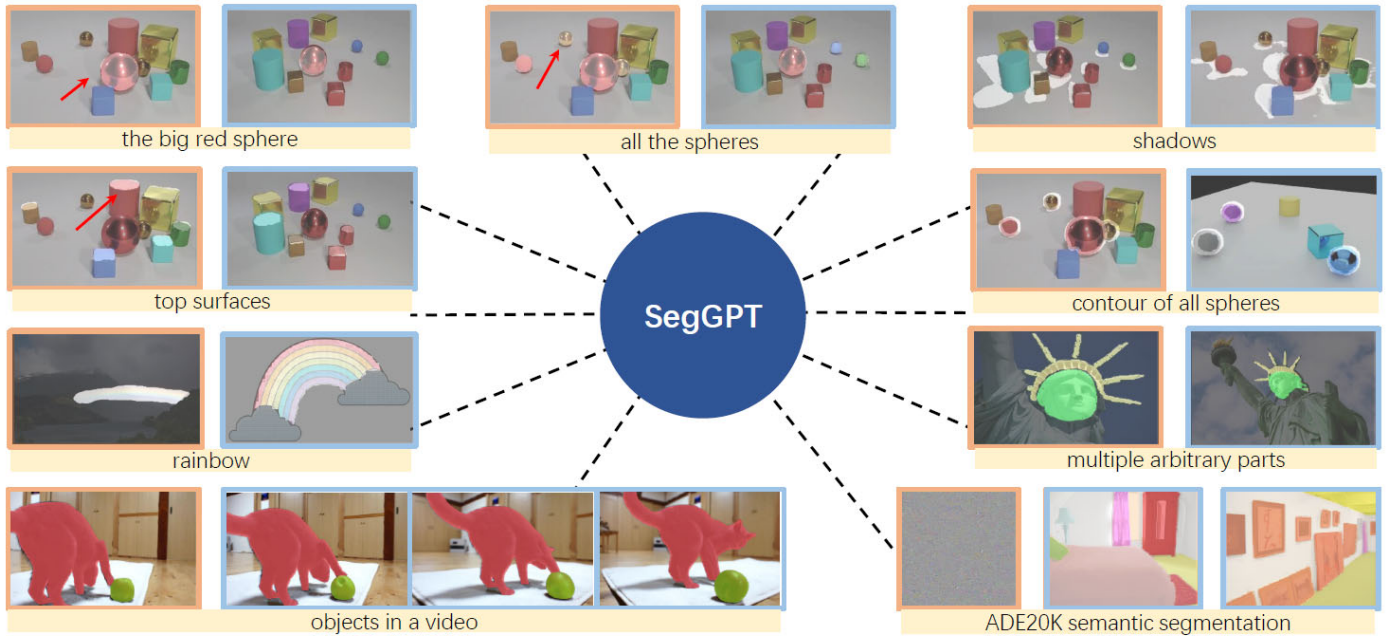
SegGPT: Segmenting Everything In Context

Xinlong Wang[1*] Xiaosong Zhang[1*] Yue Cao[1*] Wen Wang[2] Chunhua Shen[2] Tiejun Huang[1,3]

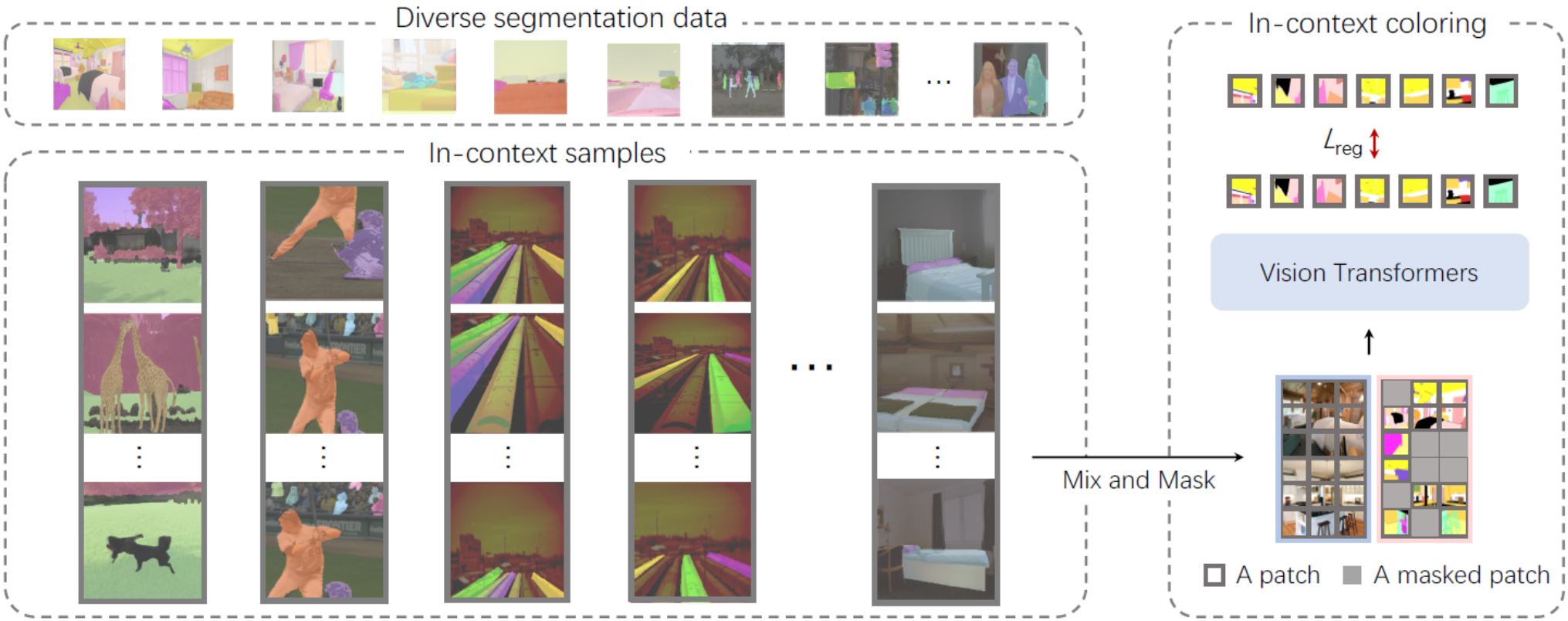
[1] Beijing Academy of Artificial Intelligence [2] Zhejiang University [3] Peking University

Code & Demo: https://github.com/baaivision/Painter

- Motivation: train a single model that is capable of solving diverse and unlimited segmentation tasks.

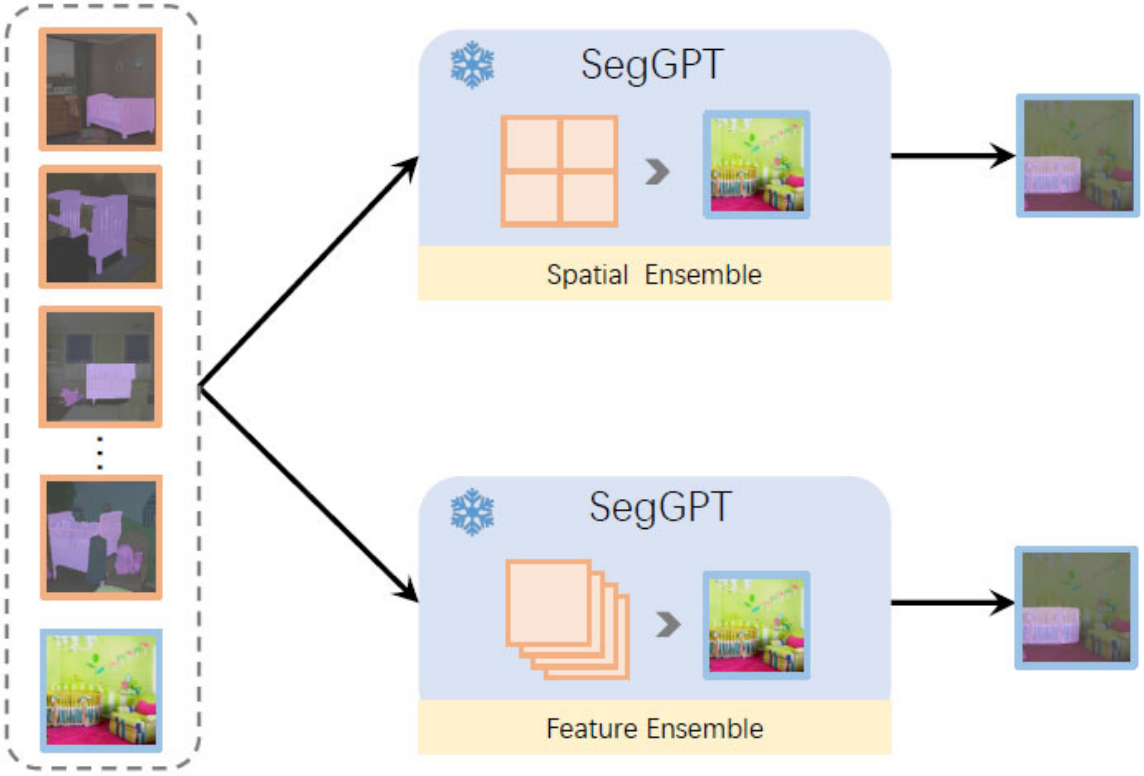


SegGPT: Segmenting Everything In Context. arXiv'2023



- In-Context Coloring: re-coloring the target images GT masks.
  - two pairs of images as input with randomly masked gt
  - train the model to reconstruct the missing pixels.
  - focus on the contextual information of the image rather than specific color information to determine the task. flexible on task definition and is capable of handling out-of-domain tasks.
- Inference: conditional image/gt + input image/masked-gt -> input gt

context ensemble for multi-example inference



multiple examples concatenated in n × n grid

features of the query image are averaged after each attention layer.

Figure 3: Illustration of our proposed context ensemble strategies for multi-example inference: the spatial ensemble (top) and the feature ensemble (bottom). The spatial en-

In-Context Tuning

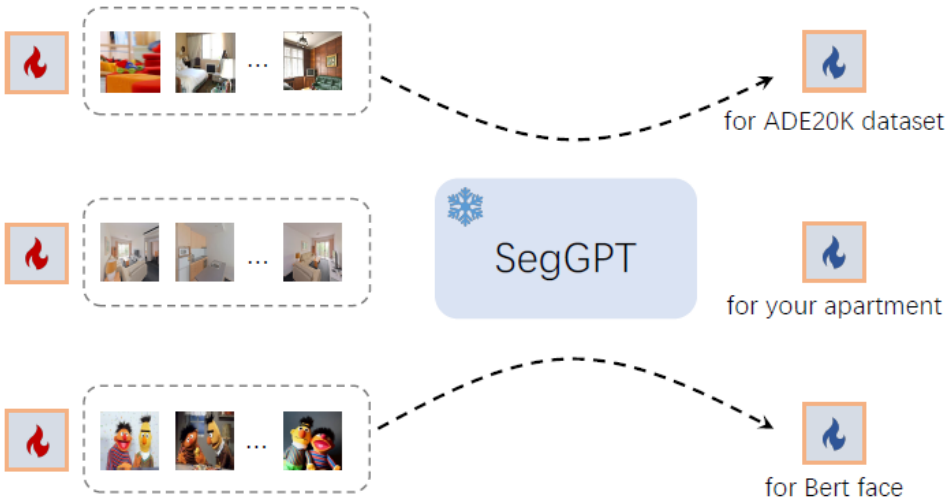


Figure 4: Illustration of in-context tuning on different task specifications. For in-context tuning, we freeze the whole pre-trained model and only optimize the learnable image tensor which serves as the input context. We can perform the

freeze the whole model.
train a learnable image tensor (conditional image/gt) as the input context.

Experiment: only train one generalist model with a mixture of public datasets （ADE20K, Cityscape, COCO, PASCAL VOC）

method	venue	COCO-20^i one-shot	COCO-20^i few-shot	PASCAL-5^i one-shot	PASCAL-5^i few-shot
specialist model					
HSNet [35]	ICCV'21	41.2	49.5	66.2	70.4
HSNet*		41.7	50.7	68.7	73.8
VAT [19]	ECCV'22	41.3	47.9	67.9	72.0
VAT*		42.9	49.4	72.4	76.3
FPTrans [53]	NeurIPS'22	47.0	58.9	68.8	78.0
FPTrans*		56.5	65.5	77.7	83.2
generalist model					
Painter	CVPR'23	32.8	32.6	64.5	64.6
SegGPT	this work	56.1	67.9	83.2	89.8

method	venue	mIoU one-shot	mIoU few-shot
trained on FSS-1000			
DAN [43]	ECCV'20	85.2	88.1
HSNet [35]	ICCV'21	86.5	88.5
SSP [15]	ECCV'22	87.3	88.6
VAT [19]	ECCV'22	90.3	90.8
DACM [50]	ECCV'22	90.8	91.7
not trained on FSS-1000			
Painter	CVPR'23	61.7	62.3
SegGPT	this work	85.6	89.3

Table 2: Quantitative results on few-shot semantic segmentation on FSS-1000. SegGPT achieves remarkable results although not trained on FSS-1000

method	venue	YouTube-VOS 2018 [52] $\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$	DAVIS 2017 [37] $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	MOSE [12] $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
with video data												
AGAME [21]	CVPR'19	66.0	66.9	-	61.2	-	70.0	67.2	72.7	-	-	-
AGSS [29]	ICCV'19	71.3	71.3	65.5	75.2	73.1	67.4	64.9	69.9	-	-	-
STM [36]	ICCV'19	79.4	79.7	84.2	72.8	80.9	81.8	79.2	84.3	-	-	-
AFB-URR [27]	NeurIPS'20	79.6	78.8	83.1	74.1	82.6	74.6	73.0	76.1	-	-	-
RDE [25]	CVPR'22	83.3	81.9	86.3	78.0	86.9	86.1	82.1	90.0	48.8	44.6	52.9
SWEM [31]	CVPR'22	82.8	82.4	86.9	77.1	85.0	84.3	81.2	87.4	50.9	46.8	54.9
XMem [9]	ECCV'22	86.1	85.1	89.8	80.3	89.2	87.7	84.0	91.4	57.6	53.3	62.0
without video data												
Painter	CVPR'23	24.1	27.6	35.8	14.3	18.7	34.6	28.5	40.8	14.5	10.4	18.5
SegGPT	this work	74.7	75.1	80.2	67.4	75.9	75.6	72.5	78.6	45.1	42.2	48.0

Table 3: Quantitative results of video object segmentation on YouTube-VOS 2018, DAVIS 2017, and MOSE. Notably, Painter

in-domain on COCO-20/PASCAL-5.

out-of-domain on FSS-1000，Youtube, DAVIS, MOSE.

method	venue	mIoU
specialist model		
FCN [32]	CVPR'15	29.4
RefineNet [28]	CVPR'17	40.7
DPT [39]	ICCV'21	49.2
Mask2Former [8]	CVPR'22	57.7
generalist model		
Painter	CVPR'23	49.9
SegGPT	this work	39.6

Table 5: Results on ADE20K semantic segmentation.

method	venue	PQ
specialist model		
PanopticFPN [23]	CVPR'19	40.3
SOLOv2 [47]	NeurIPS'20	42.1
Mask2Former [8]	CVPR'22	57.8
UViM [24]	NeurIPS'22	45.8
generalist model		
Painter	CVPR'23	43.4
SegGPT	this work	34.4

Table 6: Results on COCO panoptic segmentation.

underperforming specialist methods in specific in-domain task

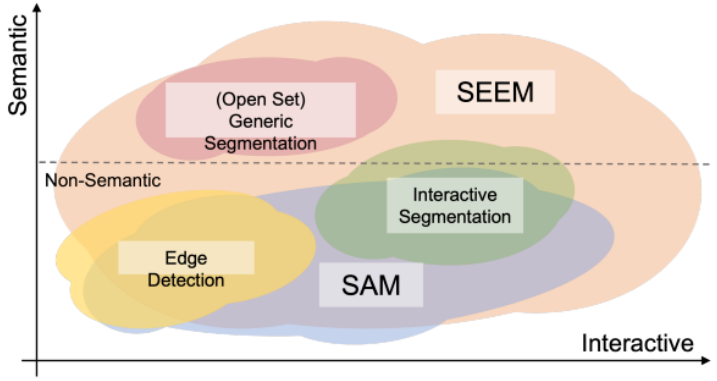
- The model needs to rely on context examples (rather than specific color) to determine the task, making optimization much more difficult.

NANJING UNIVERSITY

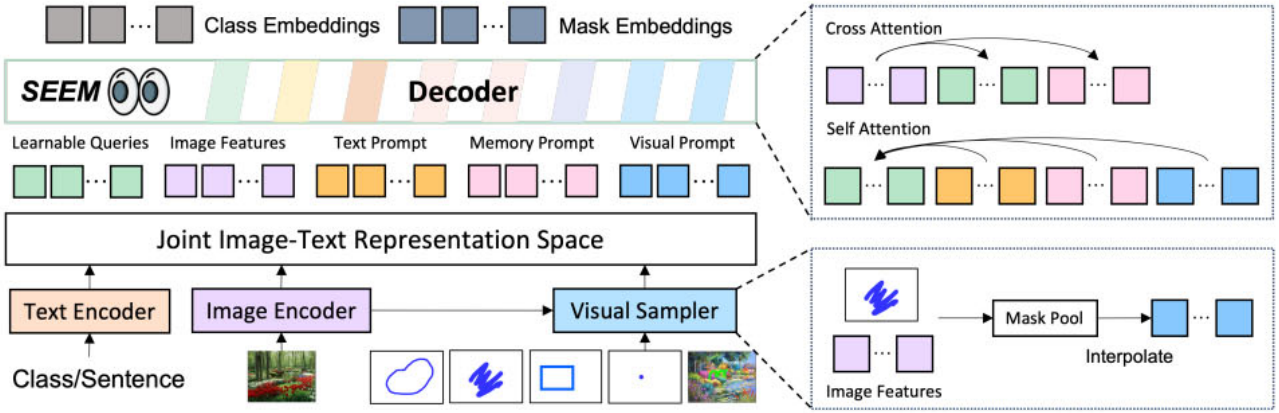
Segment Everything Everywhere All at Once

Xueyan Zou*§1, Jianwei Yang*‡2, Hao Zhang*♯, Feng Li*♯, Linjie Li†, Jianfeng Gao¶‡, Yong Jae Lee¶§
§ University of Wisconsin-Madison ‡ Microsoft Research at Redmond ♯ HKUST † Microsoft Cloud & AI
*Equal Contribution ¶ Equal Advisory Contribution 1. Main Technical Contribution 2. Project Lead
{xueyan,yongjaelee}@cs.wisc.edu {jianwyan,jfgao,linjli}@microsoft.com {hzhangcx,fliay}@connect.ust.hk

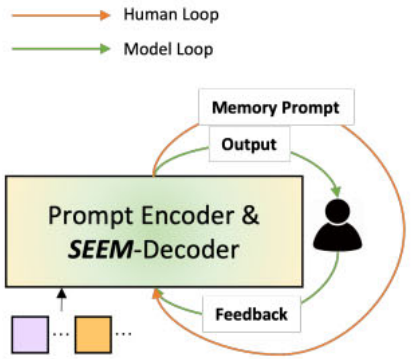
- Movtivation: segment every object with semantics (everything), cover every pixel in the image (everywhere), and support all compositions of prompts (all at once).
  - Versatility：points, masks, text, boxes, and a referred region of another similar image.
  - Compositionality ： joint visual-semantic space for visual and textual prompts to compose queries.
  - Interactivity：memory prompts for multiple interaction rounds of refinement.
  - Semantic awareness：output an open-set semantic to any output segmentation.



Segment Everything Everywhere All at Once. arXiv'2023



(a) Model Architecture



(b) Human-Model Interaction

inspired by X-Decoder(CVPR' 2023)，add Interactive component

$$\langle \mathbf{O}_h^m, \mathbf{O}_h^c \rangle = \mathbf{Decoder}(\mathbf{Q}_h; \langle \mathbf{P}_t, \mathbf{P}_v, \mathbf{P}_m \rangle | \mathbf{Z}) \quad (1)$$

$$\mathbf{M} = \mathbf{MaskPredictor}(\mathbf{O}_h^m) \quad (2)$$

$$\mathbf{C} = \mathbf{ConceptClassifier}(\mathbf{O}_h^c) \quad (3)$$

Versatility $\mathbf{P}_v = \mathbf{VisualSampler}(\mathbf{s}, \hat{\mathbf{Z}})$

Compositionality $ID_v \leftarrow \mathbf{Match}(\mathbf{O}_h^m \cdot \mathbf{P}_v + \mathbf{IoU}_{mask})$
$ID_t \leftarrow \mathbf{Match}(\mathbf{O}_h^c \cdot \mathbf{P}_t + \mathbf{IoU}_{mask})$

Textual and visual embedding remain inherently different. the visual and textual prompts can be simply concatenated in inference

Interactivity $\mathbf{P}_m^l = \mathbf{MaskedCrossAtt}(\mathbf{P}_m^{l-1}; \mathbf{M}_p | \mathbf{Z})$

Semantic-aware: labels computed $\mathbf{O}_h^c$ and the vocabularies text embedding.

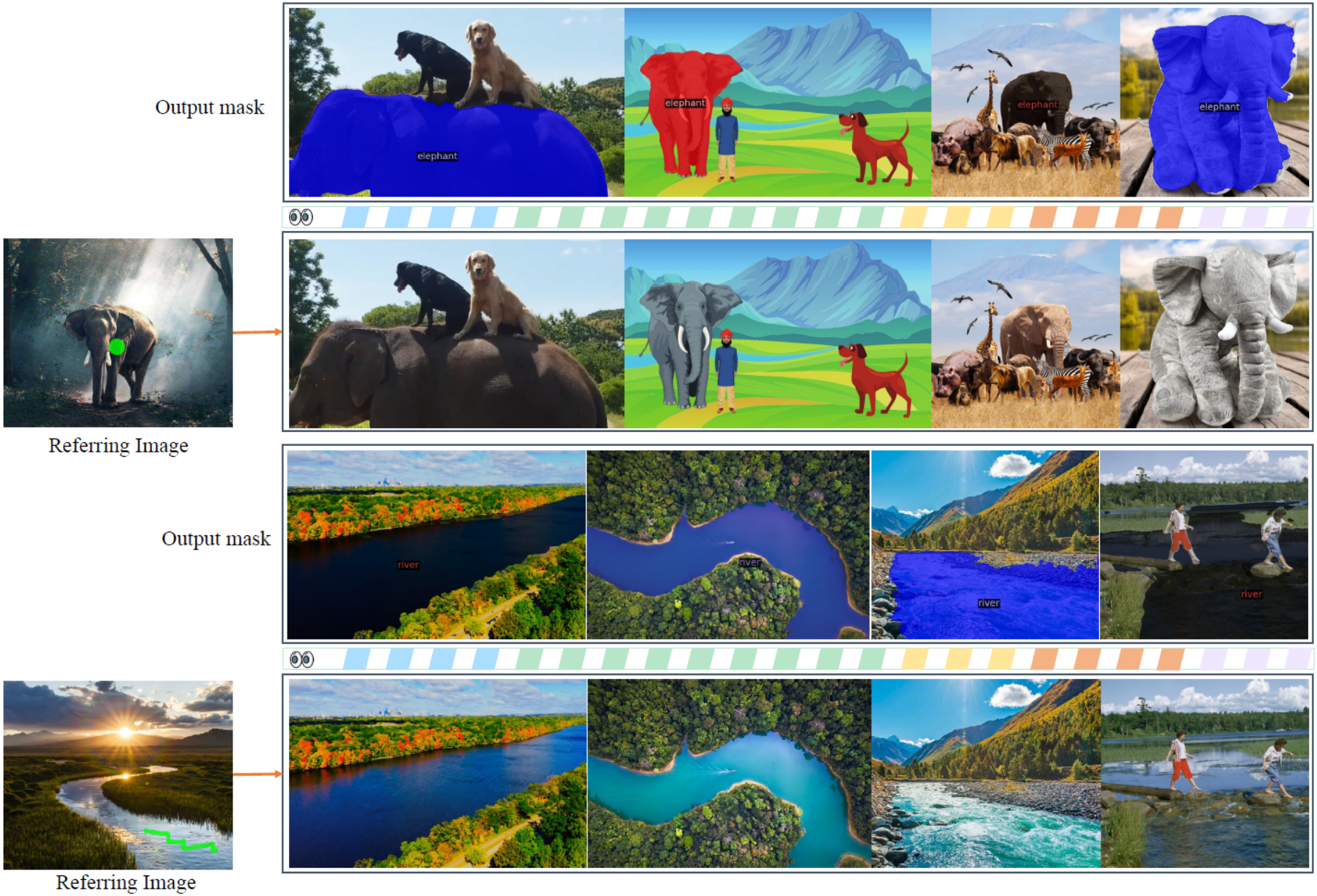
Segment Everything Everywhere All at Once. arXiv'2023

南京大学 NANJING UNIVERSITY

Method	Segmentation Data	Type	Generic Segmentation COCO			Referring Segmentation RefCOCOg			Interactive Segmentation Pascal VOC		Interactive Segmentation SBD	
			PQ	mAP	mIoU	cIoU	mIoU	AP50	NoC85	NoC90	NoC85	NoC90
Mask2Former (T) [8]	COCO (0.2M)	Segmentation	53.2	43.3	63.2	-	-	-	-	-	-	-
Mask2Former (B) [8]	COCO (0.2M)		56.4	46.3	67.1	-	-	-	-	-	-	-
Pano/SegFormer (B) [50]	COCO (0.2M)		55.4	*	*	-	-	-	-	-	-	-
LAVT (B) [56]	Ref-COCO (0.2M)		-	-	-	61.2	*	*	-	-	-	-
RITM (<T) [46]	COCO (0.2M)	Interactive	-	-	-	-	-	-	2.19	2.57	3.59	5.71
PseudoClick (<T) [33]	COCO (0.2M)		-	-	-	-	-	-	1.94	2.25	3.79	5.11
FocalClick (T) [7]	COCO (0.2M)		-	-	-	-	-	-	2.97	3.52	4.56	6.86
FocalClick (B) [7]	COCO (0.2M)		-	-	-	-	-	-	2.46	2.88	3.53	5.59
SimpleClick (B) [32]	COCO+LVIS (0.2M)		-	-	-	-	-	-	2.06	2.38	3.43	5.62
X-Decoder (T) [65]	COCO (0.2M)	Generalist	52.6	41.3	62.4	59.8	*	*	-	-	-	-
X-Decoder (B) [65]	COCO (0.2M)		56.2	45.8	66.0	64.5	*	*	-	-	-	-
SAM (B) [22]	SAM (11M)		-	-	-	-	-	-	3.30	4.20	6.50	9.76
SEEM (T)	COCO+LVIS (0.2M)		50.6	39.5	61.2	56.6	62.7	70.9	4.12	5.23	6.81	10.1
SEEM (B)	COCO+LVIS (0.2M)		56.2	46.8	65.3	63.2	68.3	76.6	3.41	4.33	6.67	9.99
SEEM (T)	COCO+LVIS (0.2M)	Composition	*	*	*	62.3	66.3	75.1	*	*	*	*
SEEM (B)	COCO+LVIS (0.2M)		*	*	*	65.7	69.8	77.0	*	*	*	*

The model is trained on a small scale of segmentation data

Segment Everything Everywhere All at Once. arXiv'2023



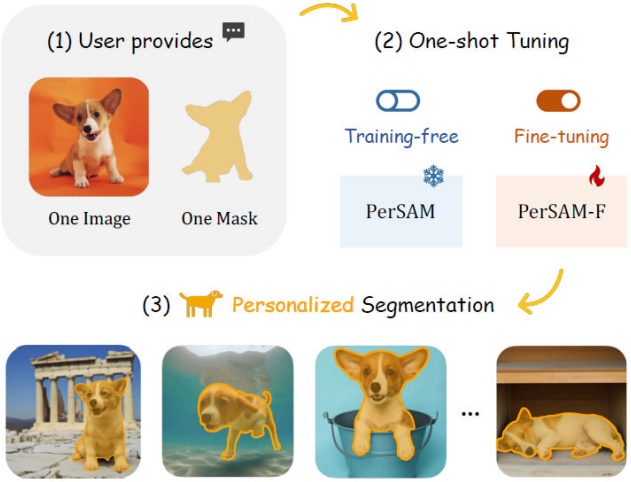
NANJING UNIVERSITY

Personalize Segment Anything Model with One Shot

Renrui Zhang[1,2], Zhengkai Jiang[3], Ziyu Guo[1], Shilin Yan[1], Junting Pan[2]
Hao Dong[4], Peng Gao[1], Hongsheng Li[2]

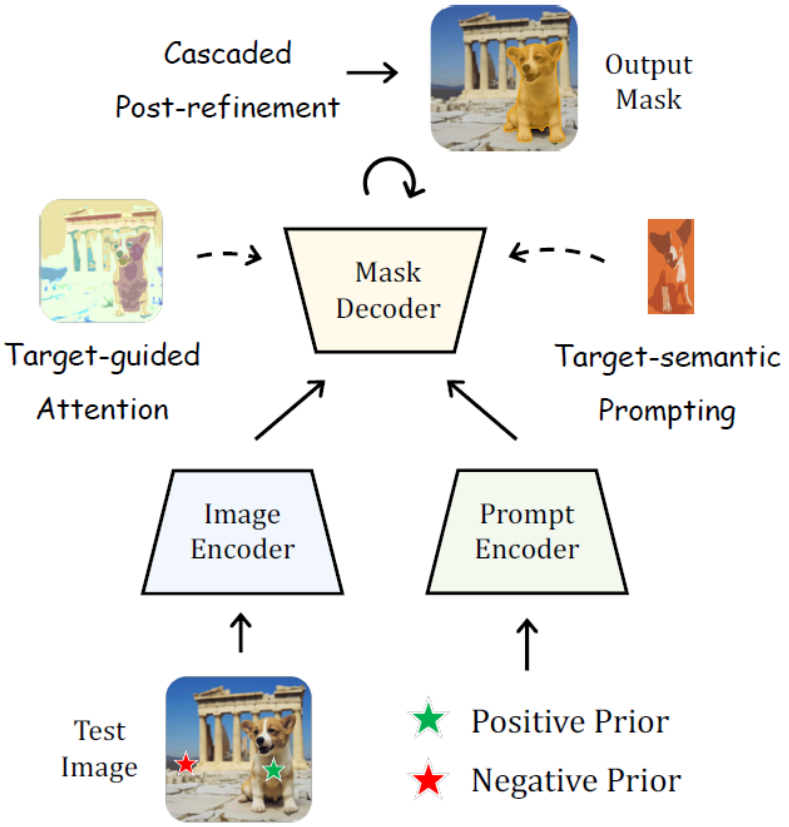
[1]Shanghai Artificial Intelligence Laboratory [2]CUHK MMLab
[3]Tencent Youtu Lab [4]CFCS, School of CS, Peking University
{zhangrenrui, gaopeng}@pjlab.org.cn, zhengkjiang@tencent.com

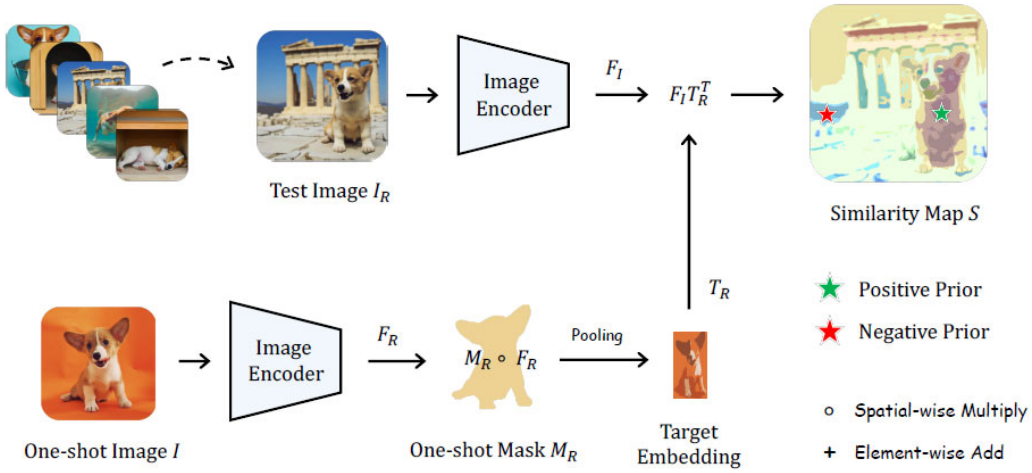
- Motivation:
  - customizing SAM for specific visual concepts without man-powered prompting.
  - To further alleviate the mask ambiguity, we present an efficient one-shot fine-tuning variant, PerSAM-F.



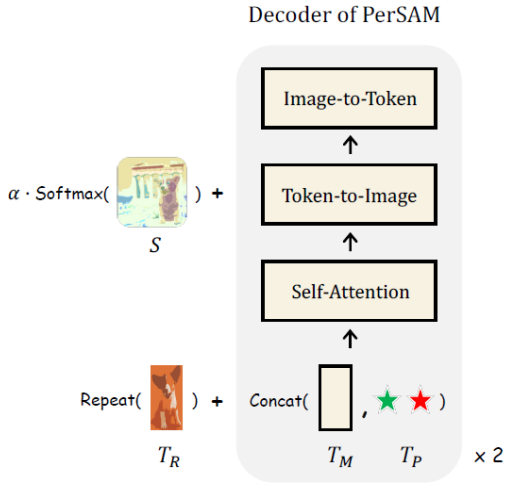
NANJING UNIVERSITY

Method: Training free PerSAM



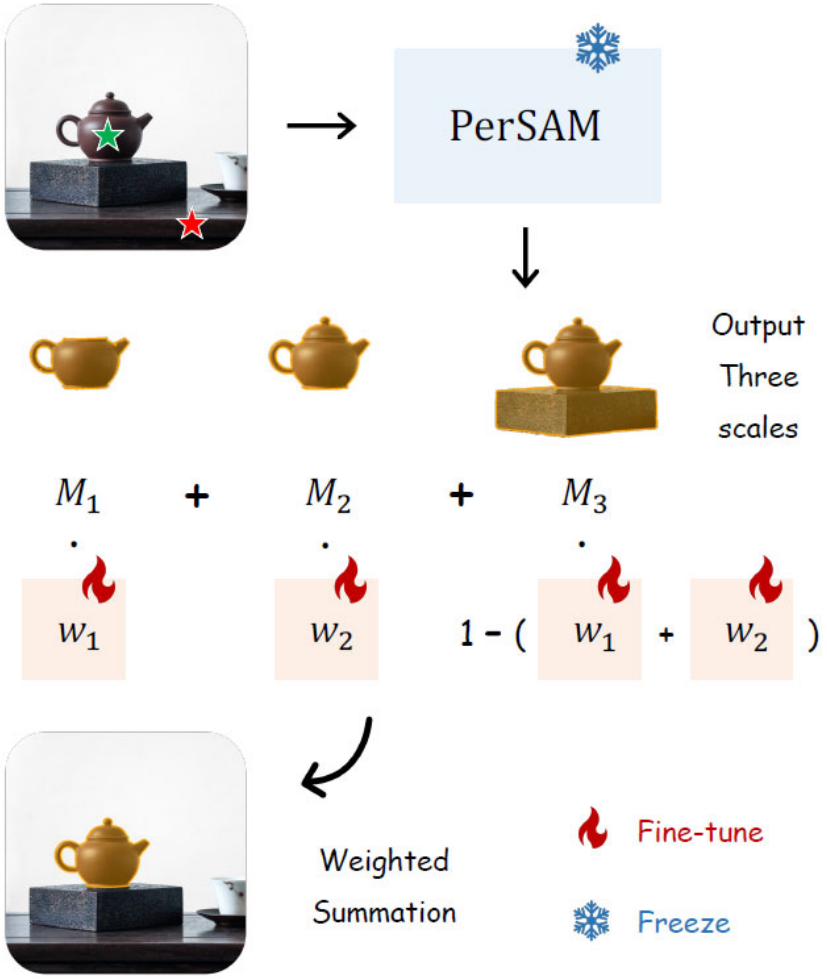


Positive-negative Location Prior



Target-guided Attention & Target semantic Prompting.

- Method: Finetuning of PerSAM-F, alleviate the ambiguity of segmentation scales without manpower



only fine-tunes the 2 parameters within 10 seconds.

Method	mIoU	Param.	Can	Barn	Clock	Cat	Back-pack	Teddy Bear	Duck Toy	Thin Bird	Red Cartoon	Robot Toy
Existing Methods												
Painter [51]	56.35	354M	19.06	3.21	42.89	94.06	88.05	93.04	33.27	20.92	98.19	64.99
Visual Prompting [2]	65.88	383M	61.23	58.55	59.23	76.60	66.67	79.75	89.93	67.35	81.03	72.37
SEEM [63]	80.50	157M	88.80	74.34	53.93	94.53	90.92	96.72	98.30	70.64	97.16	89.58
SegGPT [53]	94.26	354M	96.62	63.79	92.56	94.13	94.40	93.67	97.15	92.60	97.33	96.19
Our Approach												
PerSAM	89.32	**0**	96.17	38.91	96.19	90.70	95.39	94.64	97.31	93.73	96.96	60.56
PerSAM-F	**95.33**	2	96.72	97.50	96.10	92.27	95.52	95.19	97.31	93.96	97.11	96.67
Improvement	+6.01		+0.55	+58.59	-0.09	+1.57	+0.13	+0.55	+0.0	+0.23	+0.15	+36.11

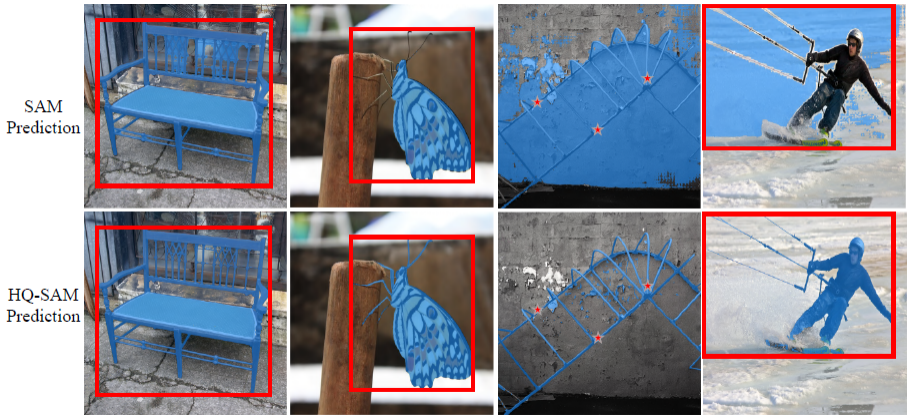
Variant	mIoU	Gain
Only Positive Prior	69.11	-
+ Negative Prior	72.47	+3.63
+ Post-refinement	83.91	+11.44
+ Target. Attention	85.82	+1.91
+ Target. Prompting	89.32	+3.50
+ Fine-tuning	95.33	+6.01

How to extend to multiple targets?

Segment Anything in High Quality

Lei Ke*[1,2] Mingqiao Ye*[1] Martin Danelljan[1] Yifan Liu[1] Yu-Wing Tai[2]
Chi-Keung Tang[2] Fisher Yu[1]
[1]ETH Zürich [2]HKUST

- Motivation: SAM suffers from two key problems:
  - 1) Coarse mask boundaries, often even neglecting the segmentation of thin object structures.
  - 2) Incorrect predictions, broken masks, or large errors in challenging cases.



Method: reuses and preserves the pre-trained model weights of SAM, while only introducing minimal additional parameters and computation.

- learnable HQ-Output Token input to SAM's mask decoder
- a refined feature set to achieve accurate mask details.
- Freeze SAM model avoid catastrophic knowledge forgetting of SAM.

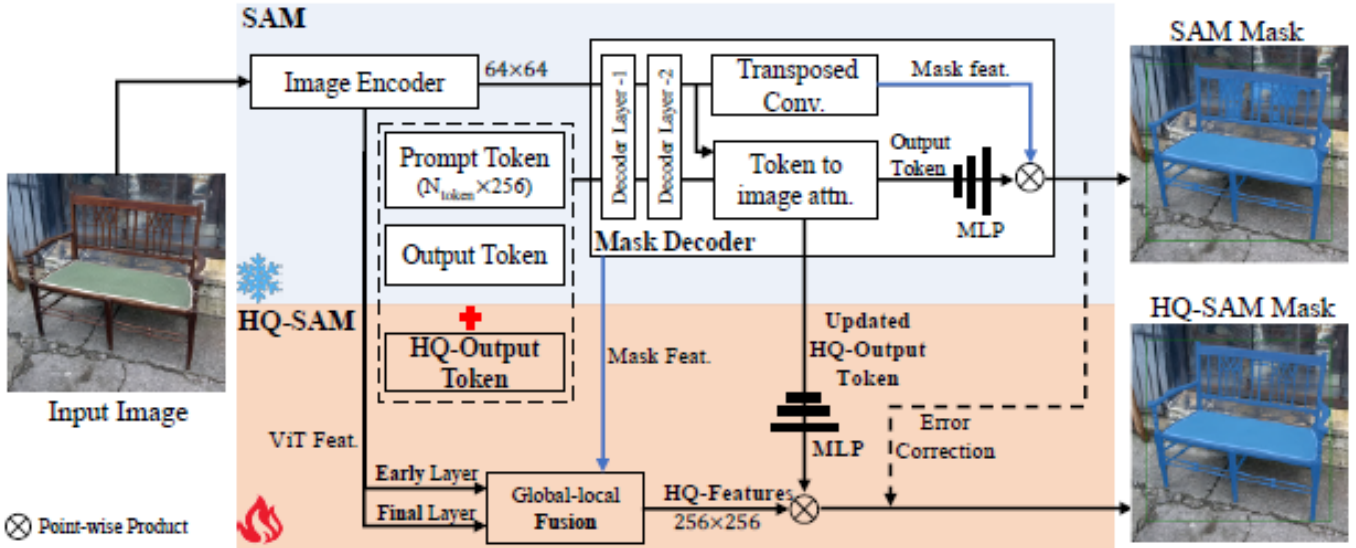
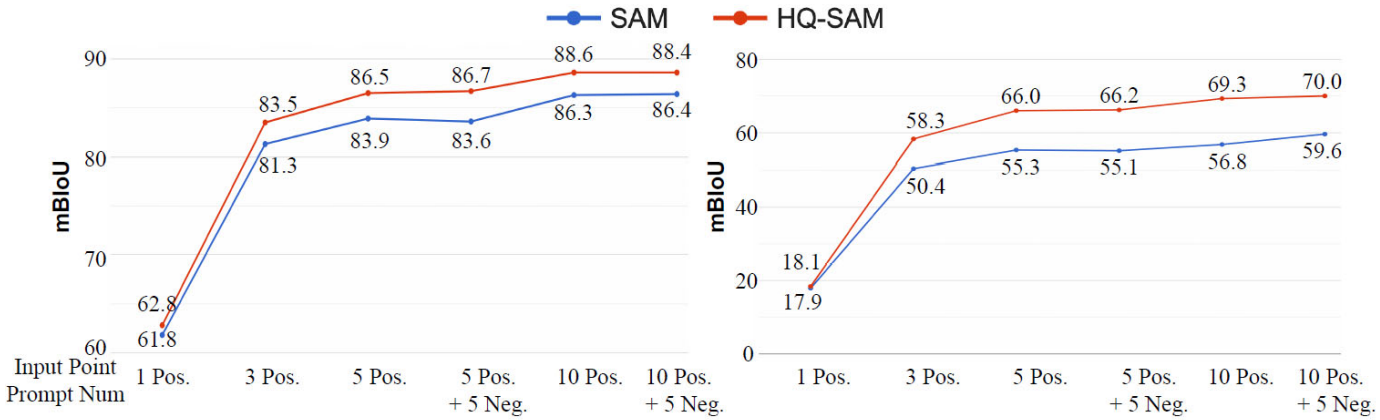


Table 4: Comparison with model finetuning or extra post-refinement [6]. For the COCO dataset, we use a SOTA detector FocalNet-DINO [51] trained on the COCO dataset as our box prompt generator.

Model	Four HQ datasets		COCO				
	mIoU	mBIoU	AP_B	AP	AP_L	AP_M	AP_S
SAM (baseline)	79.5	71.1	33.3	48.5	63.9	53.1	**34.1**
Add Context Token [54]	85.2	77.0	31.9	47.2	65.1	51.2	31.9
Extra Post-refinement [6]	80.9	74.6	2.8	13.4	43.4	9.4	0.0
Finetune SAM's decoder	87.6	79.5	9.0	19.5	45.2	15.8	4.7
Finetune SAM's output token	87.6	79.7	33.7	48.7	66.0	52.3	33.6
HQ-SAM (Ours)	**89.1**	**81.8**	**34.4**	**49.5**	**66.2**	**53.8**	33.9



(a) Performance comparison on **COIFT**

(b) Performance comparison on **DIS**

Encoder and mask-decoder are Still freeze, limiting the model Transfer ablility.

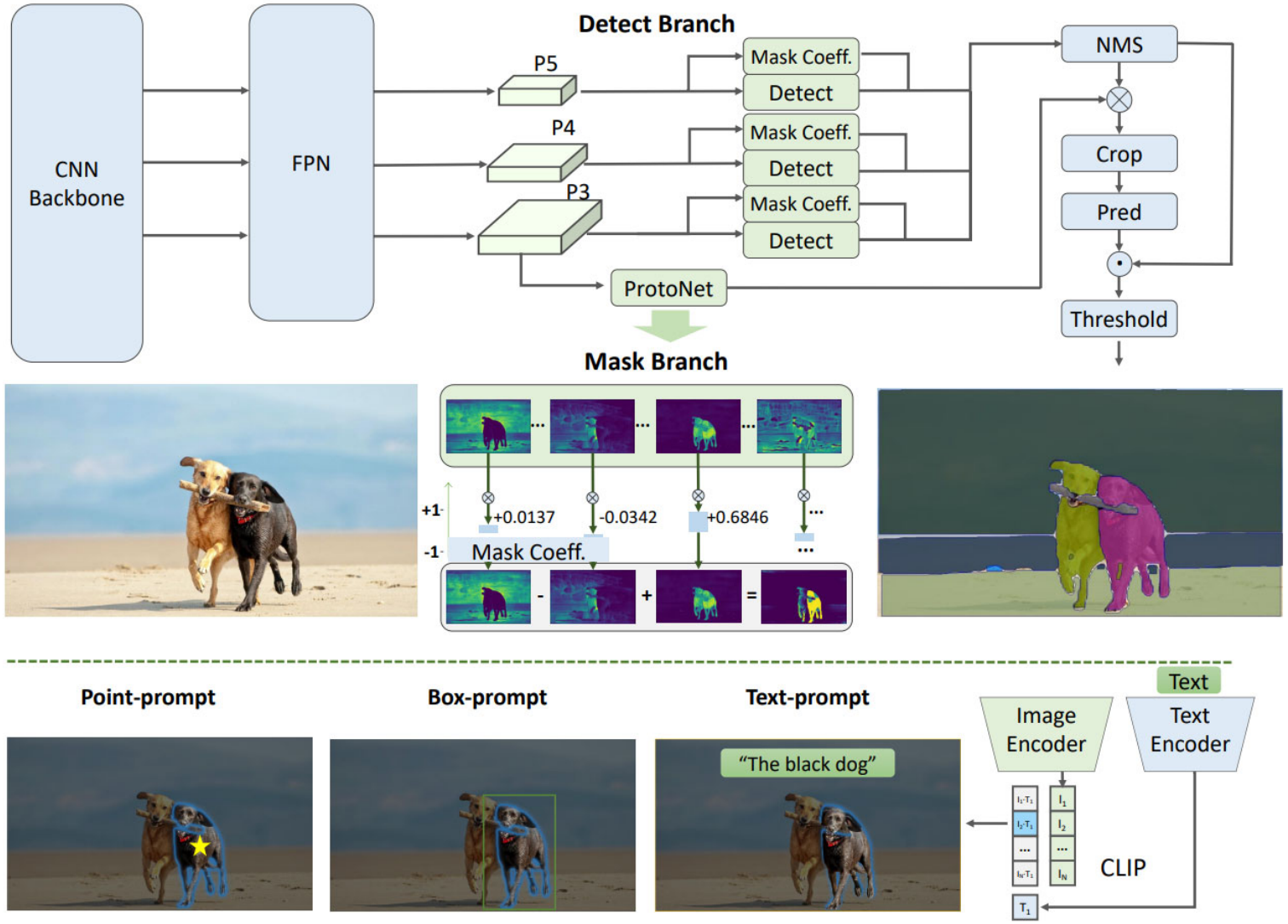
南京大学 NANJING UNIVERSITY

Fast Segment Anything

Xu Zhao [1,3] Wenchao Ding [1,2] Yongqi An [1,2] Yinglong Du [1,2]
Tao Yu [1,2] Min Li [1,2] Ming Tang [1,2] Jinqiao Wang [1,2,3,4]
Institute of Automation, Chinese Academy of Sciences, Beijing, China[1]
School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China[2]
Objecteye Inc., Beijing, China[3]
Wuhan AI Research, Wuhan, China[4]

- 动机：
  - Accelerate SAM inference.
- 方法：
  - Decouple the task of SAM into two parts: 1. Segment all objects 2. Select the object corresponding to the prompt. Inference speed is independent of the number of prompts.
  - Replace Transformer with CNN, reducing inference overhead (including w/ prompt and w/o prompt)
  - using only 1/50 of the SA-1B dataset published by SAM authors.

FastSAM



YOLO-V8-seg

FastSAM

南京大学 NANJING UNIVERSITY

method	params	Running Speed under Different Point Prompt Numbers (ms) 1	10	100	E(16×16)	E(32×32*)	E(64×64)
SAM-H [20]	0.6G	446	464	627	852	2099	6972
SAM-B [20]	136M	110	125	230	432	1383	5417
FastSAM (Ours)	68M				40		

method	year	ODS	OIS	AP	R50
HED [37]	2015	.788	.808	.840	.923
EDETR [30]	2022	.840	.858	.896	.930
zero-shot transfer methods:					
Sobel filter	1968	.539	-	-	-
Canny [6]	1986	.600	.640	.580	-
Felz-Hutt [9]	2004	.610	.640	.560	-
SAM [19]	2023	.768	.786	.794	.928
FastSAM	2023	.750	.790	.793	.903

Table 2. Zero-shot transfer to edge detection on BSDS500. Evaluation Data of other methods is from [20].

method	bbox AR@1000 all	small	med.	large
ViTDet-H [23]	65.0	53.2	83.3	91.2
zero-shot transfer methods:				
SAM-H E64	52.1	36.6	75.1	88.2
SAM-H E32	50.3	33.1	76.2	89.8
SAM-B E32	45.0	29.3	68.7	80.6
FastSAM (Ours)	57.1	44.3	77.1	85.3

Table 4. Object proposal generation on LVIS v1. FastSAM and SAM is applied zero-shot, *i.e.* it was not trained for object proposal generation nor did it access LVIS images or annotations.

	COCO [26]				LVIS v1 [13]			
method	AP	AP^S	AP^M	AP^L	AP	AP^S	AP^M	AP^L
ViTDet-H [23]	51.0	32.0	54.3	68.9	46.6	35.0	58.0	66.3
zero-shot transfer methods (segmentation module only):								
SAM	46.5	30.8	51.0	61.7	44.7	32.5	57.6	65.5
FastSAM	37.9	23.9	43.4	50.0	34.5	24.6	46.2	50.8

Table 6. Instance segmentation results. Fastsam is prompted with ViTDet boxes to do zero-shot segmentation. The fully-supervised

对比SAM仍然有明显差距

NANJING UNIVERSITY

Inpaint Anything: Segment Anything Meets Image Inpainting

Tao Yu[1] Runseng Feng[1] Ruoyu Feng[1] Jinming Liu[2] Xin Jin[2] Wenjun Zeng[2] Zhibo Chen[1]
[1]University of Science and Technology of China [2]Eastern Institute for Advanced Study
{yutao666, fengruns, ustcfry}@mail.ustc.edu.cn
{jmliu, jinxin, wenjunzeng}@eias.ac.cn, chenzhibo@ustc.edu.cn

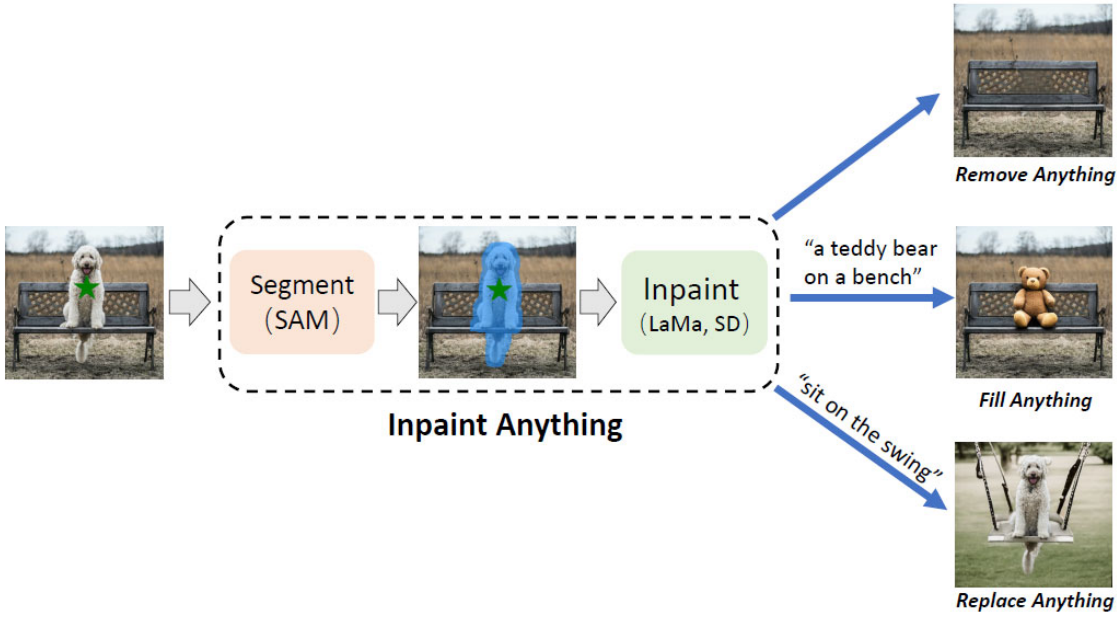


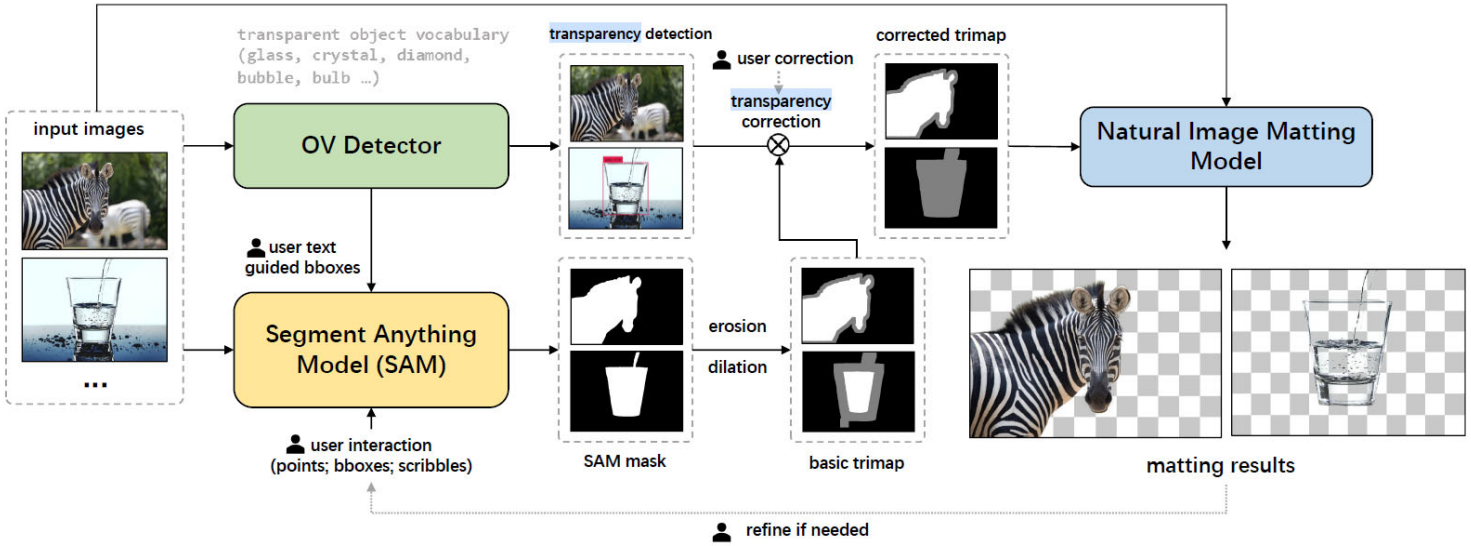
Figure 1: Illustration of our Inpaint Anything. Users can select any object in an image by clicking on it. With powerful vision models, *e.g.*, SAM[7], LaMa [13] and Stable Diffusion (SD) [11], Inpaint Anything is able to remove the object smoothly (*i.e.*, Remove Anything). Further, by inputting text prompts, users can fill the object with any desired content (*i.e.*, Fill Anything) or replace the background of it arbitrarily (*i.e.*, Replace Anything).

NANJING UNIVERSITY

Matte Anything: Interactive Natural Image Matting with Segment Anything Models

Jingfeng Yao, Xinggang Wang*, Lang Ye, and Wenyu Liu

School of EIC, HUST



$$\alpha(x,y) = \begin{cases} 1 & if\ (x,y) \in F \\ M(i,t) & if\ (x,y) \in U \\ 0 & if\ (x,y) \in B \end{cases}$$

Prior: a transparent region will not be the foreground region in trimap.
Detect transparent object for correctiong of SAM output.

南京大学
NANJING UNIVERSITY

LLM介绍

蔡文朴

caiwenpu@smail.nju.edu.cn

大数据智能研究组

南京大学计算机科学与技术系

软件新技术国家重点实验室

2023年7月11日

NANJING UNIVERSITY

GLM: General Language Model Pretraining with Autoregressive Blank Infilling

Zhengxiao Du[*1,2] **Yujie Qian**[*3] **Xiao Liu**[1,2] **Ming Ding**[1,2] **Jiezhong Qiu**[1,2]
Zhilin Yang[†1,4] **Jie Tang**[†1,2]
[1]Tsinghua University [2]Beijing Academy of Artificial Intelligence (BAAI)
[3]MIT CSAIL [4]Shanghai Qi Zhi Institute
zx-du20@mails.tsinghua.edu.cn yujieq@csail.mit.edu
{zhiliny,jietang}@tsinghua.edu.cn

Motivation

- Bidirection LM (encoder) e.g. BERT:
  - No interdependencies of mask tokens, poor in NLG (Natural Langeuage Generation) tasks
- Unidirection LM (decoder) e.g. GPT
  - unidirectional context, poor in NLU (Natural Langeuage Understanding) tasks.
- Bidirection LM + Unidirection LM (encoder-decoder) e.g. T5:
  - unifies NLU and NLG, but requires more parameters due to two modules i.e. encoder and decoder.
- none of above performs the best for all tasks including NLU and NLG.

GLM

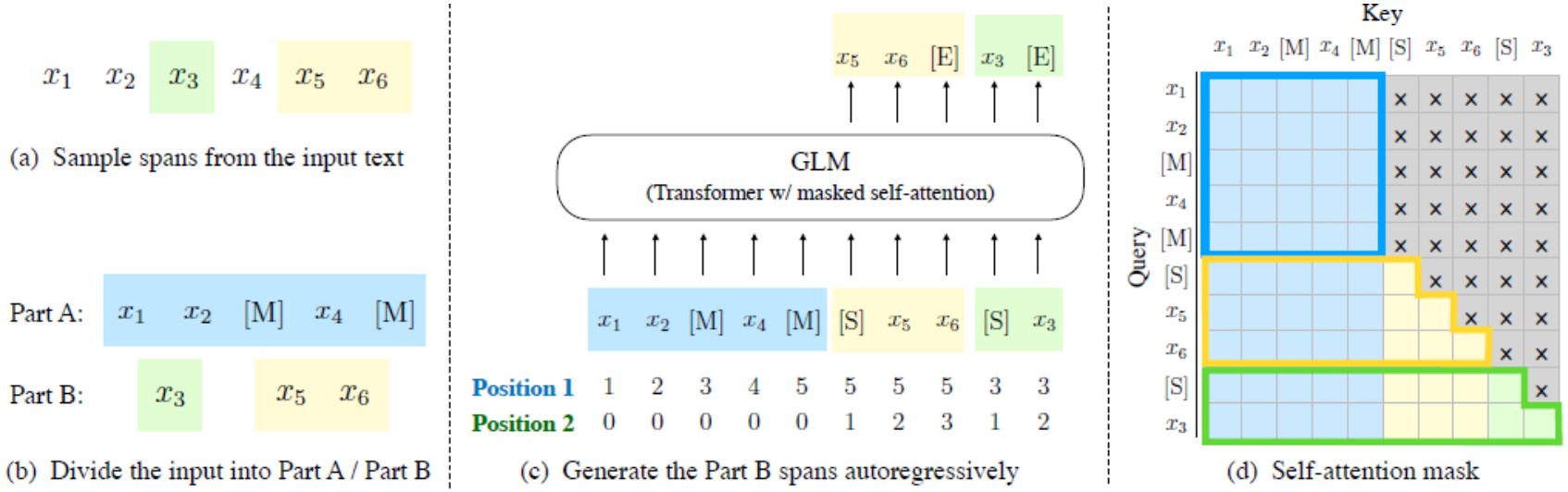
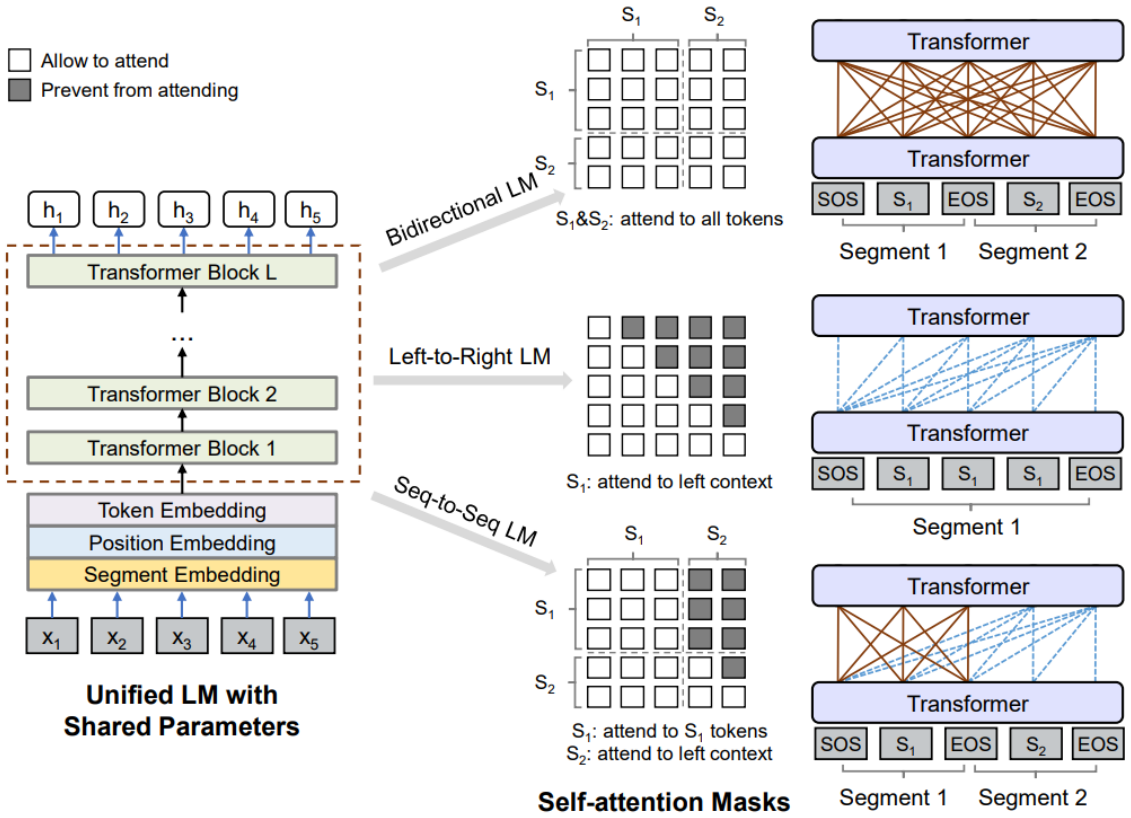


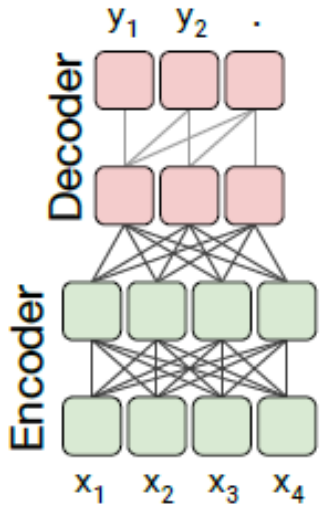
Figure 2: GLM pretraining. (a) The original text is $[x_1, x_2, x_3, x_4, x_5, x_6]$. Two spans $[x_3]$ and $[x_5, x_6]$ are sampled.

Feature: Seq-to-Seq LM; span; permute.

南京大学 NANJING UNIVERSITY

- Seq-to-Seq LM
  - Compare to Bidirection LM: No interdependencies of mask tokens/spans(Part B).
  - Compare to Unidirection LM: unidirectional context in prefix (Part A).
  - Compare to Bidirection LM + Unidirection LM: more parameters.





Dong et.al. Unified Language Model Pre-training for Natural Language Understanding and Generation. NeurIPS'2019

span:

- 2D Positional Encoding to fit span token.
- Span based [MASK] ensures that the model is not aware of the length of the [MASK]. To fit downstream tasks as usually the length of the generated text is unknown beforehand.
- Previous token based [MASK] need to know [MASK] number to predict [MASK] tokens.

permute

permute the order of the spans capture the interdependencies between different spans, from XL-Net [1]

$$\max_{\theta} \mathbb{E}_{\boldsymbol{z}\sim Z_m}\left[\sum_{i=1}^{m}\log p_\theta(\boldsymbol{s}_{z_i}|\boldsymbol{x}_{\text{corrupt}}, \boldsymbol{s}_{\boldsymbol{z}_{<i}})\right]$$

$$p_\theta(\boldsymbol{s}_i|\boldsymbol{x}_{\text{corrupt}}, \boldsymbol{s}_{\boldsymbol{z}_{<i}}) = \prod_{j=1}^{l_i} p(s_{i,j}|\boldsymbol{x}_{\text{corrupt}}, \boldsymbol{s}_{\boldsymbol{z}_{<i}}, \boldsymbol{s}_{i,<j})$$

- XLNet (Unidirection LM w/ permute) vs GPT (Unidirection LM w/o permute)
  - Context: "Thom Yorke is the singer of Radiohead, Who is the singer of Radiohead".
  - Target: "Thom Yorke ".
  - The representations of "Thom Yorke" are not dependent on "Radiohead" with GPT language modeling and thus they will not be chosen as the answer.
  - GPT cover only unidirectional dependency and XL-Net is able to cover all dependencies expectation. over all factorization orders.
  - In seq-to-seq LM, permute can cover all dependencies between [MASK] spans.

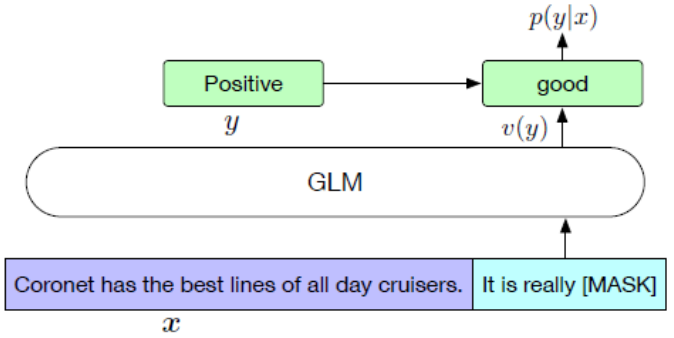
[1]Zhilin Yang et.al. XLNet: Generalized Autoregressive Pretraining for Language Understanding. NuerIPS'2019

Multitask Pretraining

- Short spans and is suited for NLU tasks.
- For NLG with longer text tasks: Document-level and Sentence-level span.

Finetuning GLM

- NLU tasks:
  - Typical Linear classifier: inconsistency between pretraining and finetuning.
  - Reformulate NLU tasks as generation t
- Generation tasks
  - Mask token at the end



GLM ACL'2022

NANJING UNIVERSITY

- Comparison with BERT (NAACL'2019):
  - Bidirection LM
  - No permutation:
  - No Span:
- Comparison with XLNet (NeurIPS'2019):
  - Unidirection LM
  - With permutation:
  - No Span:
- Comparison with T5 (JMLR'2020); Bart (ACL'2020):
  - Bidirection LM + Unidirection LM .
  - With Span:
  - No permutation:
- Comparison with UniLM (NeurIPS'2019) :
  - Unified Bidirection LM, Unidirection LM and Seq-to-Seq LM
  - No Span
  - No permutation

- Experiments:
  - Pretrain Dataset: BooksCorpus and English Wikipedia, same as BERT.
  - Model: We train GLM_Base and GLM_Large with the same architectures as BERT_Base and BERT_Large, containing 110M and 340M parameters respectively
  - Training: 250K steps, half of RoBERTa and BART

Table 1: Results on the SuperGLUE dev set.

Model	ReCoRD F1/Acc.	COPA Acc.	WSC Acc.	RTE Acc.	BoolQ Acc.	WiC Acc.	CB F1/Acc.	MultiRC F1a/EM	Avg
Pretrained on BookCorpus and Wikipedia									
$\text{BERT}_{\text{Base}}$	65.4 / 64.9	66.0	65.4	70.0	74.9	**68.8**	70.9 / 76.8	68.4 / 21.5	66.1
GLM_{Base}	**73.5 / 72.8**	**71.0**	**72.1**	**71.2**	**77.0**	64.7	**89.5 / 85.7**	**72.1 / 26.1**	**70.7**
$\text{BERT}_{\text{Large}}$	76.3 / 75.6	69.0	64.4	73.6	80.1	**71.0**	94.8 / 92.9	71.9 / 24.1	72.0
$\text{UniLM}_{\text{Large}}$	80.0 / 79.1	72.0	65.4	76.5	80.5	69.7	91.0 / 91.1	**77.2 / 38.2**	74.1
$\text{GLM}_{\text{Large}}$	81.7 / 81.1	76.0	**81.7**	74.0	**82.1**	68.5	96.1 / 94.6	77.1 / 36.3	77.0
GLM_{Doc}	80.2 / 79.6	77.0	78.8	76.2	79.8	63.6	**97.3 / 96.4**	74.6 / 32.1	75.7
GLM_{Sent}	80.7 / 80.2	77.0	79.8	79.1	80.8	70.4	94.6 / 93.7	76.9 / 36.1	76.8
GLM_{410M}	81.5 / 80.9	80.0	**81.7**	**79.4**	81.9	69.0	93.2 / **96.4**	76.2 / 35.5	78.0
GLM_{515M}	**82.3 / 81.7**	**85.0**	**81.7**	79.1	81.3	69.4	95.0 / **96.4**	**77.2** / 35.0	**78.8**
Pretrained on larger corpora									
T5_{Base}	76.2 / 75.4	73.0	79.8	78.3	80.8	67.9	94.8 / 92.9	76.4 / 40.0	76.0
T5_{Large}	85.7 / 85.0	78.0	**84.6**	84.8	84.3	71.6	96.4 / 98.2	80.9 / 46.6	81.2
$\text{BART}_{\text{Large}}$	88.3 / 87.8	60.0	65.4	84.5	84.3	69.0	90.5 / 92.9	81.8 / 48.0	76.0
$\text{RoBERTa}_{\text{Large}}$	89.0 / 88.4	**90.0**	63.5	87.0	**86.1**	**72.6**	96.1 / 94.6	**84.4 / 52.9**	81.5
$\text{GLM}_{\text{RoBERTa}}$	**89.6 / 89.0**	82.0	83.7	**87.7**	84.7	71.2	**98.7 / 98.2**	82.4 / 50.1	**82.9**

Table 2: Results of abstractive summarization on the CNN/DailyMail and XSum test sets.

Model	CNN/DailyMail			XSum		
	RG-1	RG-2	RG-L	RG-1	RG-2	RG-L
BERTSumAbs (Liu and Lapata, 2019)	41.7	19.4	38.8	38.8	16.3	31.2
$\text{UniLMv2}_{\text{Base}}$ (Bao et al., 2020)	43.2	20.4	40.1	44.0	21.1	36.1
T5_{Large} (Raffel et al., 2020)	42.5	20.7	39.8	40.9	17.3	33.0
$\text{BART}_{\text{Large}}$ (Lewis et al., 2019)	**44.2**	**21.3**	**40.9**	45.1	22.3	**37.3**
$\text{GLM}_{\text{RoBERTa}}$	43.8	21.0	40.5	**45.5**	**23.5**	**37.3**

Table 6: Ablation study on the SuperGLUE dev set. (T5 ≈ GLM – shuffle spans + sentinel tokens.)

Model	ReCoRD F1/Acc.	COPA Acc.	WSC Acc.	RTE Acc.	BoolQ Acc.	WiC Acc.	CB F1/Acc.	MultiRC F1a/EM	Avg
$BERT_{Large}$	76.3 / 75.6	69.0	64.4	73.6	80.1	**71.0**	94.8 / 92.9	71.9 / 24.1	72.0
$BERT_{Large}$ (reproduced)	**82.1 / 81.5**	63.0	63.5	72.2	80.8	68.7	80.9 / 85.7	77.0 / 35.2	71.2
$BERT_{Large}$ (cloze)	70.0 / 69.4	**80.0**	76.0	72.6	78.1	70.5	93.5 / 91.1	70.0 / 23.1	73.2
GLM_{Large}	81.7 / 81.1	76.0	**81.7**	74.0	**82.1**	68.5	**96.1 / 94.6**	77.1 / 36.3	**77.0**
– cloze finetune	81.3 / 80.6	62.0	63.5	66.8	80.5	65.0	89.2 / 91.1	72.3 / 27.9	70.0
– shuffle spans	82.0 / 81.4	61.0	79.8	54.5	65.8	56.3	90.5 / 92.9	76.7 / 37.6	68.5
+ sentinel tokens	81.8 / 81.3	69.0	78.8	**77.3**	81.2	68.0	93.7 / 94.6	**77.5 / 37.7**	76.0

WSC [MASK] include Multiple token

Sentinel: uses different sentinel tokens instead of a single [MASK] token to represent different masked spans.

THANKS

南京大學

NANJING UNIVERSITY