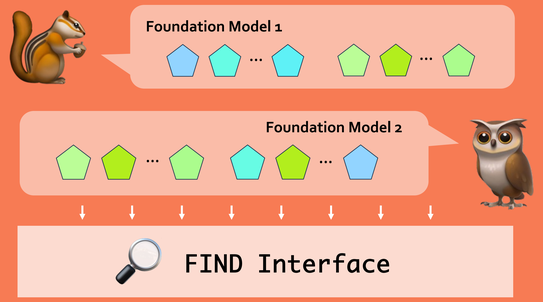
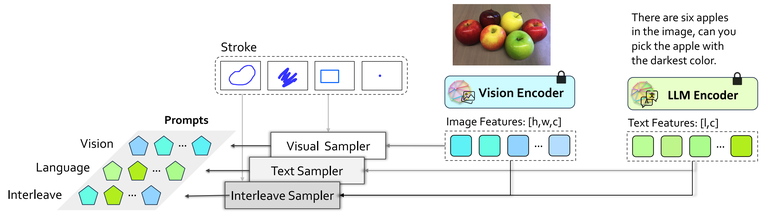
Interfacing Foundation Models' Embeddings



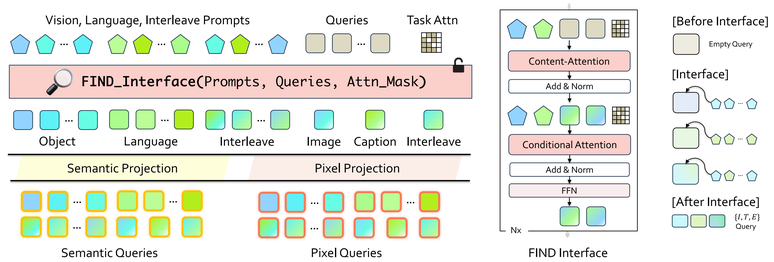
GitHub

Xueyan Zou, Linjie Li, Jianfeng Wang, Jianwei Yang, Mingyu Ding, Junyi Wei, Zhengyuan Yang, Feng Li, Hao Zhang, Shilong Liu, Arul Aravinthan, Yong Jae Lee*, Lijuan Wang*
* Equal Advising

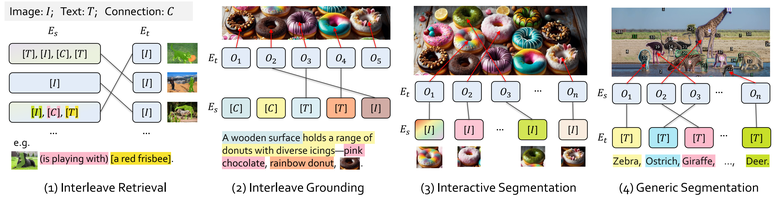
Tokenization



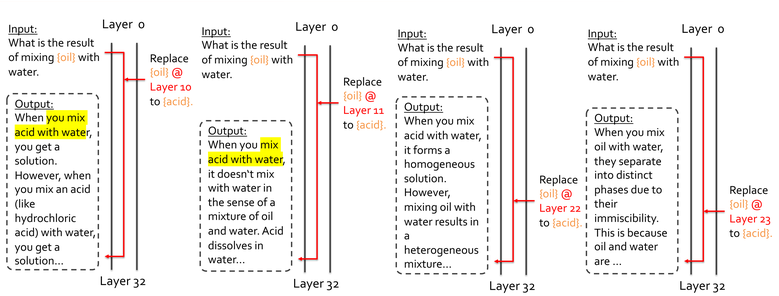
Reasoning



Detokenization



Motivation (LLaMA)



Tokenization | Reasoning | Detokenization

Qualitative Results

Interleave Segmentation


Generic Segmentation

This creature is the tallest land animal, known for its exceptionally long neck and legs, and distinctive coat of brown patches separated by lighter lines. It is native to Africa, where it browses on the higher branches of trees, primarily feeding on leaves and shoots.


Long-Context Segmentation


Interactive Segmentation

Interleave Grounding

Interleave Retrieval

Quantitative Results

		Generic Segmentation			Grounded Segmentation					Interactive Segmentation			Image-Text Retrieval					
		COCO			RefCOCO-g	COCO-Entity		COCO-Paragraph		Pascal VOC			COCO-Karpathy		COCO-Entity		COCO-Paragraph	
	Data	Joint	PQ	mAP	mIoU	cIoU	mIoU	cIoU	mIoU	cIoU	mIoU	Point	Circle	Box	IR@1	TR@1	IR@1	TR@1	IR@1	TR@1
*Mask2Former (T) [8]	COCO (0.12M)	-	53.2	43.3	63.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
*Mask2Former (B) [8]	COCO (0.12M)	-	56.4	46.3	67.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
*Mask2Former (L) [8]	COCO (0.12M)	-	**57.8**	48.6	67.4	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
Grounding-SAM (H) [27]	Grounding (5M)	✓	-	-	-	-	-	58.9	57.7	56.1	56.6	-	-	-	-	-	-	-	-	-
SAM (B) [19]	SAM (11M)	-	-	-	-	-	-	-	-	-	-	58.2	-	61.8	-	-	-	-	-	-
SAM (L) [19]	SAM (11M)	-	-	-	-	-	-	-	-	-	-	58.1	-	63.5	-	-	-	-	-	-
*SEEM (T) [53]	COCO+LVIS (0.12M)	✗	50.8	39.7	62.2	60.9	65.7	54.3	56.1	52.6	54.6	83.5	86.0	71.8	-	-	-	-	-	-
*SEEM (B) [53]	COCO+LVIS (0.12M)	✗	56.1	46.4	66.3	65.0	69.6	57.2	58.7	56.1	57.4	87.3	88.8	75.5	-	-	-	-	-	-
*SEEM (L) [53]	COCO+LVIS (0.12M)	✗	57.5	47.2	**67.6**	65.6	70.3	54.8	57.8	53.8	56.7	88.5	**89.6**	76.5	-	-	-	-	-	-
X-Decoder (T) [52]	COCO+ITP (4.12M)	✗	52.6	41.3	62.4	59.8	*	-	-	-	-	-	-	-	40.7 / 49.3	55.0 / 66.7	46.5 / 52.6	48.0 / 55.6	54.8 / 62.3	58.5 / 66.1
X-Decoder (B) [52]	COCO+ITP (4.12M)	✗	56.2	45.8	66.0	64.5	*	-	-	-	-	-	-	-	50.2 / 54.5	66.8 / 71.2	49.2 / 56.9	51.3 / 58.1	58.1 / 67.5	62.5 / 70.1
X-Decoder (L) [52]	COCO+ITP (4.12M)	✗	56.9	46.7	67.5	64.6	*	-	-	-	-	-	-	-	56.4 / 58.6	73.1 / 76.1	58.1 / 60.0	**59.9** / 62.7	58.7 / 71.6	**72.0** / 74.1
CLIP/ImageBind (H) [13, 9]	ITP (400M)	✓	-	-	-	-	-	-	-	-	-	-	-	-	49.4	65.9	53.4	57.6	59.6	64.8
FROMAGe (L) [20]	CC (12M)	✗	-	-	-	-	-	-	-	-	-	-	-	-	27.5	37.8	27.4	33.1	32.8	41.3
BLIP-2 (L) [23]	COCO+IPT (130.1M)	✗	-	-	-	-	-	-	-	-	-	-	-	-	**63.4** / 58.1	**74.4** / 65.2	**59.1** / 53.8	59.8 / 55.4	**66.3** / 64.6	65.8 / 60.1
FIND (T)	COCO (0.12M)	✓	51.0	42.3	62.0	61.1	65.3	68.5	62.5	65.0	59.4	84.3	85.8	74.5	40.4	53.0	51.0	51.5	61.2	62.9
FIND (B)	COCO (0.12M)	✓	55.5	49.0	65.7	65.3	69.3	69.5	63.0	**67.2**	60.1	86.3	88.0	75.0	45.8	60.6	56.3	56.7	65.5	69.1
FIND (L)	COCO (0.12M)	✓	56.7	**50.8**	67.4	**65.9**	**70.5**	**69.7**	**64.2**	66.6	**61.2**	**88.5**	89.5	**77.4**	46.3	61.9	57.2	58.2	**67.2**	68.6

	Interleave Grounding			Interleave Retrieval				Generic Segmentation								
	COCO-Entity		COCO-Paragraph			COCO-Entity	COCO-Paragraph	Class		Visual Context			Description			
	cIoU	mIoU	AP50	cIoU	mIoU	AP50	IR@5	IR@10	IR@5	TR@5	PQ	mAP	mIoU	PQ	mAP	mIoU	PQ	mAP	mIoU
Mask2Former (L) [8]	-	-	-	-	-	-	-	-	-	-	57.8	48.6	67.4	-	-	-	-	-	-
Grounding-SAM (H) [27]	58.9	57.7	63.2	56.1	56.6	62.5	-	-	-	-	-	-	-	-	-	-	-	-	-
CLIP/ImageBind (H) [13, 9]	-	-	-	-	-	-	51.4	61.3	58.7	68.9	-	-	-	-	-	-	-	-	-
FROMAGe (L) [20]	-	-	-	-	-	-	24.1	34.2	26.0	36.6	-	-	-	-	-	-	-	-	-
BLIP-2 (L) [23]	-	-	-	-	-	-	20.8 / 34.3	25.8 / 47.7	22.1 / 39.3	27.1 / 54.7	-	-	-	-	-	-	-	-	-
X-Decoder (T) [52]	-	-	-	-	-	-	23.6	32.2	25.6	35.5	52.6	41.3	62.4	-	-	-	18.5	15.9	22.5
X-Decoder (B) [52]	-	-	-	-	-	-	26.7	35.8	32.1	42.0	56.2	46.3	67.1	-	-	-	20.8	15.0	24.7
X-Decoder (L) [52]	-	-	-	-	-	-	26.8	36.2	32.2	43.4	57.8	48.6	67.4	-	-	-	23.5	21.1	21.7
SEEM (T) [53]	67.6	67.2	75.8	65.9	65.7	74.4	-	-	-	-	50.8	39.7	62.2	-	-	-	18.6	15.7	16.0
SEEM (B) [53]	69.4	69.2	77.8	69.2	68.6	77.3	-	-	-	-	56.1	46.4	66.3	-	-	-	22.9	21.6	20.0
SEEM (L) [53]	68.3	69.0	77.5	67.7	68.4	77.0	-	-	-	-	**56.9**	46.7	**67.5**	-	-	-	24.0	26.4	18.7
FIND (T)	74.9	68.1	79.5	73.2	66.4	77.7	43.5	57.1	49.4	63.9	51.0	42.3	62.0	41.8	32.3	51.6	19.5	30.2	**35.5**
FIND (B)	76.3	69.7	**81.8**	**75.1**	68.0	79.7	51.4	64.6	60.5	73.4	55.5	49.0	65.7	47.1	36.7	53.6	16.5	26.7	26.7
FIND (L)	**76.3**	**69.7**	81.7	74.7	**68.6**	**79.7**	**53.4**	**66.7**	**62.7**	**75.0**	56.7	**50.8**	67.4	**49.5**	**38.9**	**57.1**	**27.0**	**31.2**	26.8

Ablation Results

		COCO			g-Ref	Entity	VOC	Karpathy		Entity	
		PQ	mAP	mIoU	cIoU	cIoU	Point	IR@1	TR@1	IR@1	TR@1
Language Level	[-1]	48.3	39.1	61.2	61.3	73.0	82.6	38.9	52.2	50.3	50.8
	[-6]	47.8	38.8	60.4	60.3	72.9	81.3	38.1	49.9	48.1	47.5
	[-12]	**48.5**	**39.0**	**61.4**	61.3	**73.0**	**82.6**	**40.4**	**54.0**	**50.8**	**51.9**
	[-18]	48.2	39.0	61.1	62.2	72.6	82.2	40.1	52.7	50.6	50.5
	[-24]	48.5	38.8	61.5	**61.6**	72.9	82.6	40.2	52.2	50.5	51.3
	[-30]	48.1	39.2	61.1	60.1	73.3	82.4	37.9	49.3	49.4	50.0

		Generic Segmentation						Grounding	Interactive	Retrieval	
		Class		Description				g-Ref	VOC	COCO-Karpathy	
Vision	Language	PQ	mAP	mIoU	PQ	mAP	mIoU	cIoU	1-IoU	IR@1	TR@1
X-Decoder (T) (52)	UniCL (43)	48.5	39.0	61.4	12.4	20.7	18.9	61.3	82.6	40.4	54.0
X-Decoder (T) (52)	LLaMa (38)	48.5	38.9	61.2	19.5	30.2	35.5	61.6	82.5	40.2	52.2
SAM (B) (19)	UniCL (43)	42.5	37.6	53.6	4.5	17.7	17.9	64.9	81.6	29.1	39.5
SAM (B) (19)	LLaMa (38)	42.5	36.9	53.0	6.1	15.6	16.6	58.9	81.5	27.0	35.5