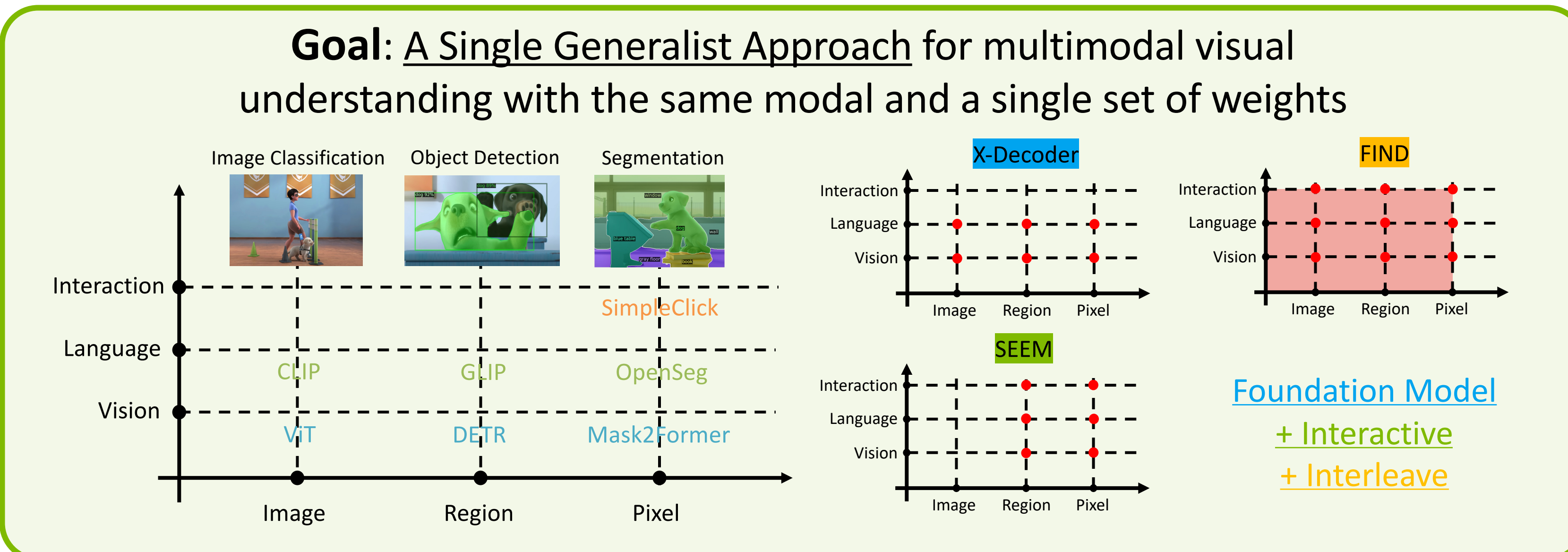


# Segment Everything Everywhere All at Once

Xueyan Zou<sup>\*^</sup>, Jianwei Yang<sup>\*^</sup>, Hao Zhang<sup>\*</sup>, Feng Li<sup>\*</sup>, Linjie Li, Jianfeng Wang,  
Lijuan Wang, Jianfeng Gao<sup>§</sup>, Yong Jae Lee<sup>§</sup>



\*Equal Contribution  
§ Equal Advising  
▲ Project Lead  
^ Main Technical Contribution



### Generic Segmentation + Language Description

**FIND** Foundation Models

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.

**Interactive Segmentation**

Interleaved Visual Retrieval & Grounding (is playing with) [a red frisbee]. (is sitting on) [a wooden bench].

Interleaved Text/Visual Retrieval & Grounding (is skiing) [on the slope] (under) [the clear sky]. (is surfing on) [a small wave in the vast ocean].

Xueyan Zou, Linjie Li, Jianfeng Wang, Jianwei Yang, Mingyu Ding, Zhengyuan Yang, Feng Li, Hao Zhang, Shilong Liu, Arul Aravinthan, Yong Jae Lee<sup>§</sup>, Lijuan Wang<sup>§</sup>. "FIND: Interfacing Foundation Models' Embeddings." ArXiv, 2023.

### Generalized Decoding for Pixel, Image, and Language

Xueyan Zou<sup>\*</sup>, Ziyi Dou<sup>\*</sup>, Jianwei Yang<sup>\*^</sup>, Zhe Gan, Linjie Li, Chunyuan Li, Xiyang Dai, Harkirat Behl, Lu Yuan, Nanyun Peng, Lijuan Wang, Yong Jae Lee<sup>§</sup>, Jianfeng Gao<sup>§</sup>.

Image Captioning, Image-Text Retrieval, Referring Segmentation, Open-Vocabulary Panoptic Segmentation, Open-Vocabulary Instance Segmentation, Referring Captioning, Referring Image Editing.

SEEM Architecture: Learnable Queries, Image Features, Text Prompt, Memory Prompt, Visual Prompt. Joint Image-Text Representation Space. Text Encoder, Image Encoder, Visual Sampler. Cross Attention, Self Attention. Machine Loop, Human Loop. SEEM-Decoder. Memory Prompt, Text Prompt, Visual Prompt.

### SEEM Architecture

Panoptic, Instance, Semantic, Point, Box, Scribble, Text/Audio, Cross Style, Text+Visual.

SEEM Architecture: Learnable Queries, Image Features, Text Prompt, Memory Prompt, Visual Prompt. Joint Image-Text Representation Space. Text Encoder, Image Encoder, Visual Sampler. Cross Attention, Self Attention. Machine Loop, Human Loop. SEEM-Decoder. Memory Prompt, Text Prompt, Visual Prompt.

(a) Model Architecture, (b) Human-Model Interaction, (c) Queries and Prompt Interaction, (d) Self-Attention Mask.

Xueyan Zou<sup>\*</sup>, Jianwei Yang<sup>\*^</sup>, Hao Zhang<sup>\*</sup>, Feng Li<sup>\*</sup>, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao<sup>§</sup>, Yong Jae Lee<sup>§</sup>. "SEEM: Segment Everything Everywhere All at Once." NeurIPS, 2023.

### Quantitative Results:

Table 1: One model for segmentation on a wide range of segmentation tasks. SEEM is the first model to simultaneously support generic segmentation, referring segmentation, and interactive segmentation, as well as prompt compositionality. (#Concurrent work. - indicates the model does not have capability for the task, \* indicates do not have reported number.)

Method	Segmentation Data	Type	Generic Segmentation			Referring Segmentation			Interactive Segmentation					
			PQ	mAP	mIoU	cloU	RetCOCO AP50	5-NoCR5	10-NoCR5	20-NoCR5	PascalVOC 5-NoC90	10-NoC90	20-NoC90	
Mask2Former (T) [6]	COCO (0.12M)	Segmentation	53.2	43.3	63.2	-	-	-	-	-	-	-	-	-
Mask2Former (B) [6]	COCO (0.12M)	Segmentation	56.4	46.3	67.1	-	-	-	-	-	-	-	-	-
Mask2Former (L) [6]	COCO (0.12M)	Segmentation	57.8	48.6	67.4	-	-	-	-	-	-	-	-	-
PanoSegFormer (B) [45]	COCO (0.12M)	Segmentation	55.4	-	-	-	-	-	-	-	-	-	-	-
LAVT (B) [53]	Ref-COCO (0.03M)	Segmentation	-	-	-	61.2	*	*	-	-	-	-	-	-
PolyFormer (B) [17]	Ref-COCO+VG+... (0.16M)	Segmentation	-	-	-	69.3	*	*	-	-	-	-	-	-
PolyFormer (L) [17]	Ref-COCO+VG+... (0.16M)	Segmentation	-	-	-	71.1	*	*	-	-	-	-	-	-
PanoSegFormer (B) [45]	COCO+LVIS (0.12M)	Segmentation	-	-	-	-	-	-	2.19	*	*	2.57	*	2.25
PseudoClick (<T) [54]	COCO (0.12M)	Segmentation	-	-	-	-	-	-	1.94	*	*	3.52	*	2.88
FocalClick (T) [21]	COCO (0.12M)	Segmentation	-	-	-	-	-	-	2.97	*	*	2.88	*	2.38
FocalClick (B) [21]	COCO (0.12M)	Segmentation	-	-	-	-	-	-	2.46	*	*	1.94	2.19	1.96
SimpleClick (B) [20]	COCO+LVIS (0.12M)	Segmentation	-	-	-	-	-	-	1.75	1.93	1.64	1.72	1.67	1.84
SimpleClick (L) [20]	COCO+LVIS (0.12M)	Segmentation	-	-	-	-	-	-	1.52	1.64	1.72	1.64	1.83	1.98
SimpleClick (H) [20]	COCO+LVIS (0.12M)	Segmentation	-	-	-	-	-	-	1.51	1.64	1.76	1.64	1.83	1.98
ViViM (L) [55]	COCO (0.12M)	Segmentation	45.8	*	-	-	-	-	-	-	-	-	-	-
Pix2Seq v2 (B) [56]	COCO (0.12M)	Segmentation	52.6	41.3	62.4	59.8	-	-	-	-	-	-	-	-
X-Decoder (T) [11]	COCO (0.12M)	Segmentation	56.2	45.8	66.0	64.5	-	-	-	-	-	-	-	-
X-Decoder (B) [11]	COCO (0.12M)	Segmentation	56.9	46.7	67.5	64.6	-	-	-	-	-	-	-	-
X-Decoder (L) [11]	COCO (0.12M)	Segmentation	-	-	-	70.0	-	-	-	-	-	-	-	-
UNINEXT (T) [48]	Images+Video (3M)	Segmentation	-	-	-	44.9	-	-	-	-	-	-	-	-
UNINEXT (L) [48]	Images+Video (3M)	Segmentation	-	-	-	49.6	-	-	-	-	-	-	-	-
UNINEXT (H) [48]	Images+Video (3M)	Segmentation	-	-	-	73.4	-	-	-	-	-	-	-	-
Painter (L) [57]	COCO+ADE+NYUV2 (0.16M)	Generalist	43.4	*	*	-	-	-	-	-	-	-	-	-
#SegGPT (L) [50]	COCO+ADE+NYUV2 (0.16M)	Generalist	34.4	*	*	-	-	-	-	-	-	-	-	-
#SAM (B) [36]	SAM (11M)	Generalist	-	-	-	-	-	-	2.47	2.65	3.28	2.23	3.13	4.12
#SAM (L) [36]	SAM (11M)	Generalist	-	-	-	-	-	-	1.85	2.15	2.60	2.01	2.46	3.12
#SAM (H) [36]	SAM (11M)	Generalist	-	-	-	-	-	-	1.82	2.13	2.55	1.98	2.43	3.11
SEEM (T)	COCO+LVIS (0.12M)	Generalist	50.8	39.7	62.2	60.9	65.7	74.8	1.72	2.30	3.37	1.97	2.83	4.41
SEEM (B)	COCO+LVIS (0.12M)	Generalist	56.1	46.4	66.3	65.0	69.6	78.2	1.56	2.04	2.92	1.77	2.47	3.79
SEEM (L)	COCO+LVIS (0.12M)	Generalist	57.5	47.7	67.6	65.6	70.3	78.9	1.51	1.95	2.77	1.71	2.36	3.61
SEEM (T)	COCO+LVIS (0.12M)	Composition	-	-	-	70.4	71.7	82.1	1.72	2.28	3.32	1.97	2.82	4.37
SEEM (B)	COCO+LVIS (0.12M)	Composition	-	-	-	76.2	77.8	87.8	1.56	2.03	2.91	1.77	2.46	3.76
SEEM (L)	COCO+LVIS (0.12M)	Composition	-	-	-	75.1	76.9	86.8	1.52	1.97	2.81	1.72	2.38	3.64

### Table 2: One model for all kinds of mask interactions. SEEM has strong generalization capability on different input mask types.

Method	COCO					Open Image					ADE				
	Point 1-IoU	Stroke 1-IoU	Scribble 1-IoU	Polygon 1-IoU	Box 1-IoU	Point 1-IoU	Stroke 1-IoU	Scribble 1-IoU	Polygon 1-IoU	BoX 1-IoU	Point 1-IoU	Stroke 1-IoU	Scribble 1-IoU	Polygon 1-IoU	BoX 1-IoU
SimpleClick (B)	49.0	33.1	65.1	48.6	42.5	48.6	29.5	54.2	49.5	42.7	47.0	19.0	52.1	48.3	37.2
SimpleClick (L)	38.9	33.9	68.8	39.2	34.7	37.5	29.1	59.8	35.2	31.2	36.8	16.4	56.4	41.7	29.5
SimpleClick (H)	59.0	37.3	71.5	45.3	52.4	54.1	32.6	64.7	39.9	49.3	52.8	18.4	58.3	46.8	41.8
SAM (B)	58.6	22.8	34.2	44.5	50.7	62.3	28.4	39.2	45.8	53.6	51.0	21.9	31.1	31.0	58.8
SAM (L)	64.7	44.4	57.1	60.7	50.9	65.3	45.9	55.7	57.8	52.4	57.4	45.8	53.1	45.8	58.7
SAM (H)	80.6	57.8	77.8	57.8	63.6	80.4	67.7	79.8	79.9	79.9	80.4	69.4	79.2	78.3	88.5
SEEM (T)	78.9	81.0	81.2	72.2	73.7	67.1	69.4	69.5	63.1	60.9	65.4	67.3	67.3	59.0	53.4
SEEM (B)	81.7	82.8	83.5	76.0	75.7	67.6	69.0	68.7	64.2	60.3	66.4	68.6	67.7	60.5	53.6
SEEM (L)	83.4	84.6	84.1	76.5	76.9	66.8	67.8	67.6	62.4	60.1	65.5	66.6	66.3	58.1	54.1

### Zero Shot In-Context Prompting

Referring Image, Referring (Frame 0), Frame 5, Frame 10, Frame 30, Frame 90.

### Interactive Segmentation

Positive (Green), Negative (Blue).

Xueyan Zou<sup>\*</sup>, Jianwei Yang<sup>\*^</sup>, Hao Zhang<sup>\*</sup>, Feng Li<sup>\*</sup>, Linjie Li, Jianfeng Wang, Lijuan Wang, Jianfeng Gao<sup>§</sup>, Yong Jae Lee<sup>§</sup>. "SEEM: Segment Everything Everywhere All at Once." NeurIPS, 2023.