

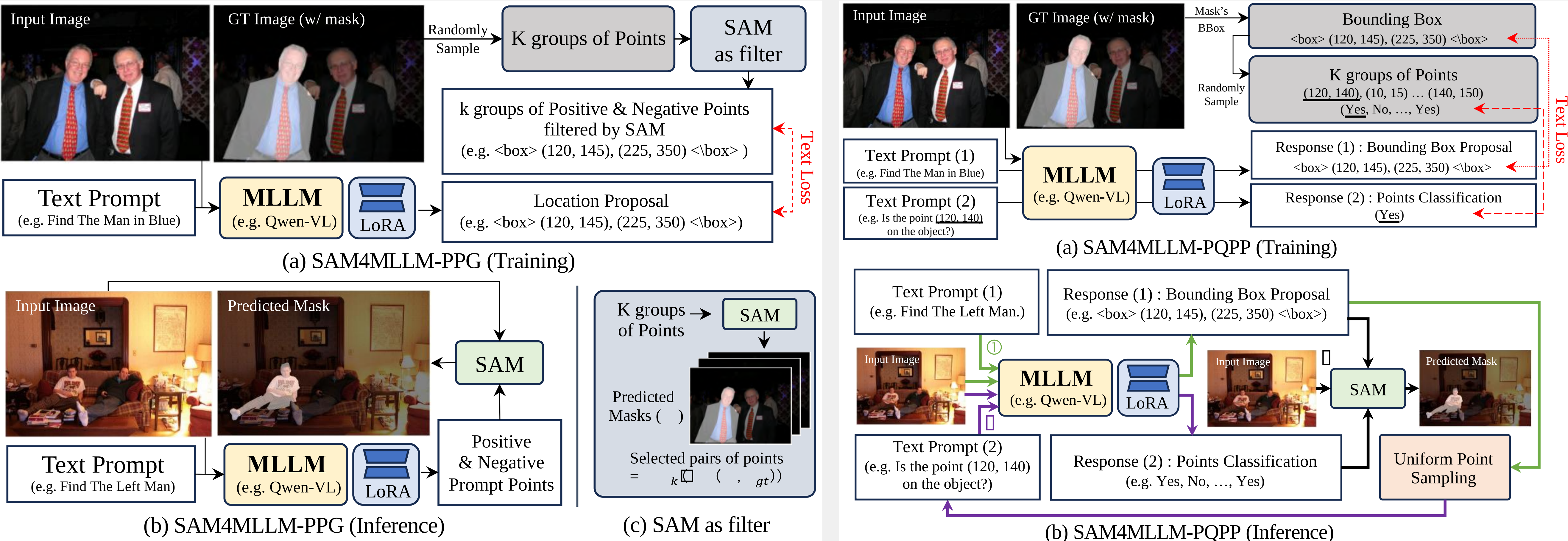
子計畫名稱：可信賴的生活場景文字識別技術
子計畫主持人：陳祝嵩
展示技術：多模態大型語言模型進行引述表達影像分割

SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. ECCV (2024).
Chen, Y.C., Li, W.H., Sun, C., Wang, Y.C.F., and Chen, C.S.

摘要

我們提出了一種創新的作法 (SAM4MLLM), 結合了分割任意模型 (Segment Anything Model, SAM) 與多模態大語言模型 (Multi-Modal Large Language Models, MLLMs) 進行表達分割 (Referring Expression Segmentation, RES)。我們的方法不需對現有模型架構進行過多的修改或加入額外的特徵，而是使用了主動查詢的策略，與 MLLM 進行多輪對話，並生成 SAM 所需要的提示詞 (Prompt) 以進行影像分割。我們提出了兩種對話流程來高效微調既有的 MLLM 分別為：提示點生成 (PPG) 以及主動查詢提示點 (PQPP)，其中 PPG 的流程簡單直接，而 PQPP 則會和 MLLM 進行兩輪對話，可以進一步提升分割表現。這兩種方法巧妙地結合了 SAM 的分割能力及 MLLMs 的強大語言理解與生成能力，在語言框架內實現了精細的影像切割，更重要的是我們的作法不需在學習過程中增加過多的計算成本，便可以在多個 RES 資料集上展現我們方法的有效性與泛用性。

方法



數據實驗結果 (於 refCOCO, refCOCO+, refCOCOg, GRES)

Method	refCOCO			refCOCO+			refCOCOg		
	val	testA	testB	val	testA	testB	val(U)	test(U)	
<i>LLM based (13B)</i>									
PerceptionGPT-13B [CVPR-24]	75.3	79.1	72.1	68.9	74.0	61.9	70.7	71.9	
GSVA-Llama2-13B [CVPR-24]	79.2	81.7	77.1	70.3	73.8	63.6	75.7	77.0	
<i>traditional methods</i>									
MAttNet [CVPR-18]	56.51	62.37	51.70	46.67	52.39	40.08	47.64	48.61	
STEP [ICCV-19]	60.04	63.46	57.97	48.19	52.33	40.41	-	-	
LSCM [ECCV-20]	61.37	64.99	59.55	49.34	53.12	43.50	-	-	
VLT [ICCV-21]	65.65	68.29	62.73	55.50	59.20	49.36	52.99	56.65	
SeqTR [ECCV-22]	67.26	69.79	64.12	54.14	58.93	48.19	55.67	55.64	
CRIS [CVPR-22]	70.5	73.2	66.1	65.3	68.1	53.7	59.9	60.4	
LAVT [CVPR-22]	72.7	75.8	68.8	62.1	68.4	55.1	61.2	62.1	
ReLA [CVPR-23]	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	
X-Decoder [CVPR-23]	-	-	-	-	-	-	64.6	-	
PolyFormer-L [CVPR-23]	76.94	78.49	74.83	72.15	75.71	66.73	71.15	71.17	
VPD [ICCV-2023]	73.25	-	-	62.69	-	-	61.96	-	
ETRIS [ICCV-2023]	71.06	74.11	66.66	62.23	68.51	52.79	60.28	60.42	
SEEM [NeurIPS-23]	-	-	-	-	-	-	65.7	-	
<i>LLM based (7B)</i>									
LISA-7B [CVPR-24]	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	
PixelLM-7B [CVPR-24]	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	
PerceptionGPT-7B [CVPR-24]	75.1	78.6	71.7	68.5	73.9	61.3	70.3	71.7	
GSVA-7B [CVPR-24]	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3	
GLaMM-7B [CVPR-24]	79.5	83.2	76.9	72.6	78.7	64.6	74.2	74.9	
SAM4MLLM-7B-PPG	76.2	80.1	72.0	71.2	75.9	64.3	74.2	74.3	
SAM4MLLM-7B-PQPP	77.1	80.9	72.5	71.5	76.8	64.7	74.5	75.2	
SAM4MLLM-7B-PQPP-LLaVA1.6	79.6	82.8	76.1	73.5	77.8	65.8	74.5	75.6	
SAM4MLLM-8B-PQPP-LLaVA1.6	79.8	82.7	74.7	74.6	80.0	67.2	75.5	76.4	

Method	Validation Set			Test Set A			Test Set B		
	gIoU	cloU	N-acc.	gIoU	cloU	N-acc.	gIoU	cloU	N-acc.
<i>LLM-based (13B) model</i>									
LISA-13B [CVPR-24]	65.24	63.96	57.49	69.99	71.00	55.43	62.11	62.29	56.34
GSVA-13B [CVPR-24]	70.04	66.38	66.02	73.29	72.79	64.72	65.45	63.20	62.47
<i>traditional methods</i>									
MAttNet [CVPR-18]	48.24	47.51	41.15	59.30	58.66	44.04	46.14	45.33	41.32
LTS [CVPR-21]	52.70	52.30	-	62.64	61.87	-	50.42	49.96	-
VLT [ICCV-21]	52.00	52.51	47.17	63.20	62.19	48.74	50.88	50.52	47.82
CRIS [CVPR-22]	56.27	55.34	-	63.42	63.82	-	51.79	51.04	-
LAVT [CVPR-22]	58.40	57.64	49.32	65.90	65.32	49.25	55.83	55.04	48.46
ReLA [CVPR-23]	63.60	62.42	56.37	70.03	69.26	59.02	61.02	59.88	58.40
<i>LLM based (7B)</i>									
LISA-7B [CVPR-24]	61.63	61.76	54.67	66.27	68.50	50.01	58.84	60.63	51.91
GSVA-7B [CVPR-24]	66.47	63.29	62.43	71.08	69.93	65.31	62.23	60.47	60.56
SAM4MLLM-7B-PPG	68.37	65.66	63.71	69.05	69.62	65.96	63.71	62.35	61.25
SAM4MLLM-7B-PQPP	68.96	66.33	62.96	70.54	70.13	65.82	63.98	63.21	61.61
SAM4MLLM-8B-PQPP-LLaVA1.6	71.86	67.83	66.08	74.15	72.22	63.92	65.29	63.42	59.99

視覺實驗結果 (於 ReasonSeg)

