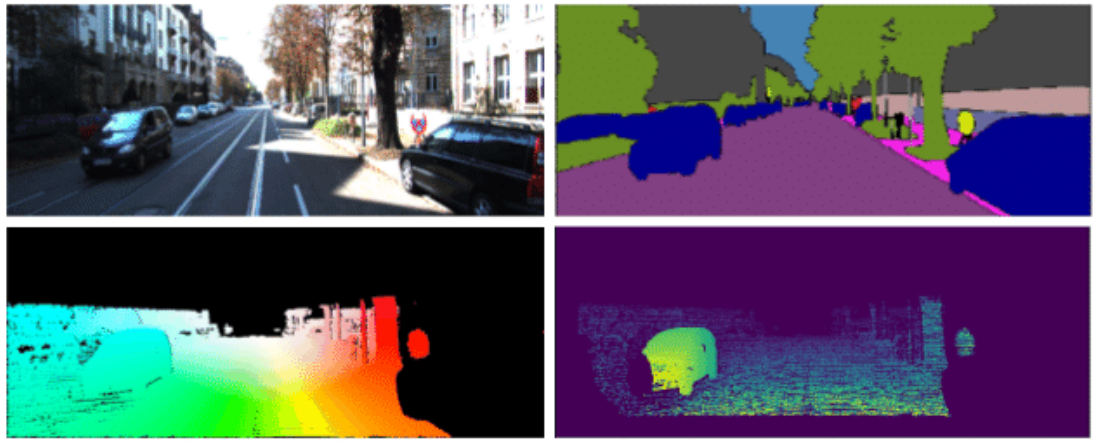
Better, Stronger, Faster: Tackling the Trilemma in MLLM-based Segmentation with Simultaneous Textual Mask Prediction

Jiazhen Liu, Mingkuan Feng, Long Chen

May, 2026

Challenge in Dense perception



Demonstration of dense perception

1. Architectural Conflict

- **MLLMs:** 1D token-by-token generation.
- **Perception:** 2D pixel-level outputs.

2. Objective Conflict

- **MLLMs:** Optimize for high-level, abstract logic.
- **Perception:** Demand low-level, local details.

Existing Approaches:

- **Paradigm 1**: Embedding Prediction

  LISA [CVPR'24], PixelLM [CVPR'24], M²SA [ICLR'25], READ [CVPR'25]

  Resolve architectural conflict

- **Paradigm 2**: Next-token Prediction

  VisionLLM [NIPS'24], Seg-Zero [arXiv'25], SegAgent [CVPR'25], Text4Seg [ICLR'25]

  Resolve objective conflict

- **Paradigm 3 (Ours)**: All-token Prediction

  STAMP [CVPR'26]

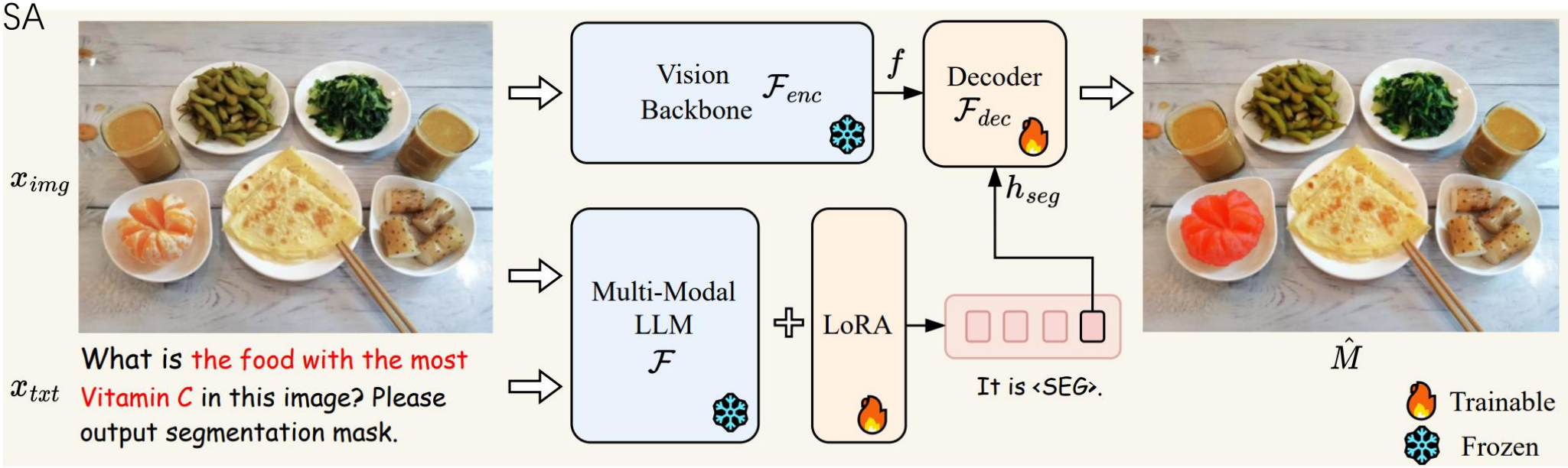
  Resolve trilemma
  Better, Faster, Stronger;

Challenge in Dense perception

Current solution:

1. Add additional pixel-level decoder and new training objective. (Embedding Prediction)

LISA



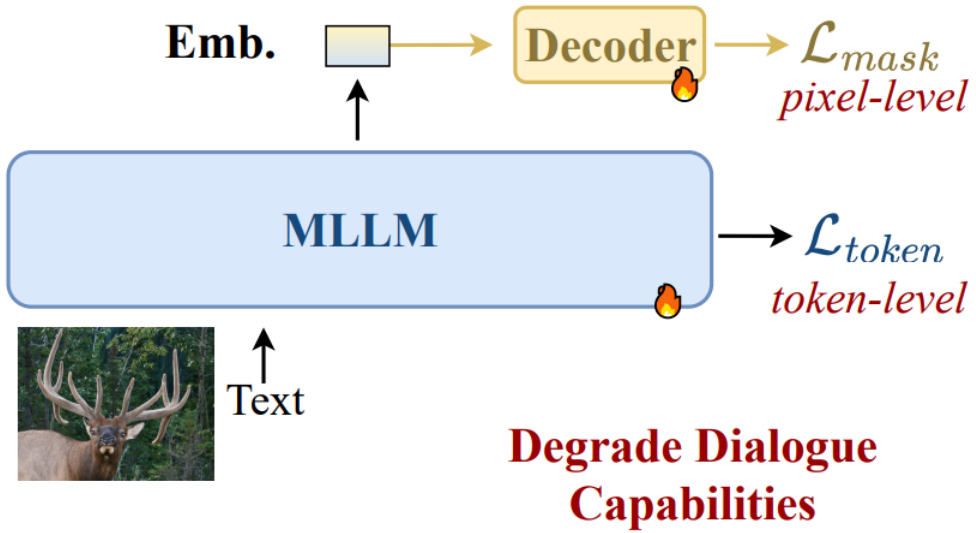
X Lai et al., LISA: Reasoning Segmentation via Large Language Model, CVPR 2024

Similar works: GSVA (CVPR'24), PixelLM (CVPR'24), READ (CVPR'25) ···

Challenge in Dense perception

Current solution:

1. Add additional pixel-level decoder and new training objective. (Embedding Prediction)



Catastrophic forgetting:

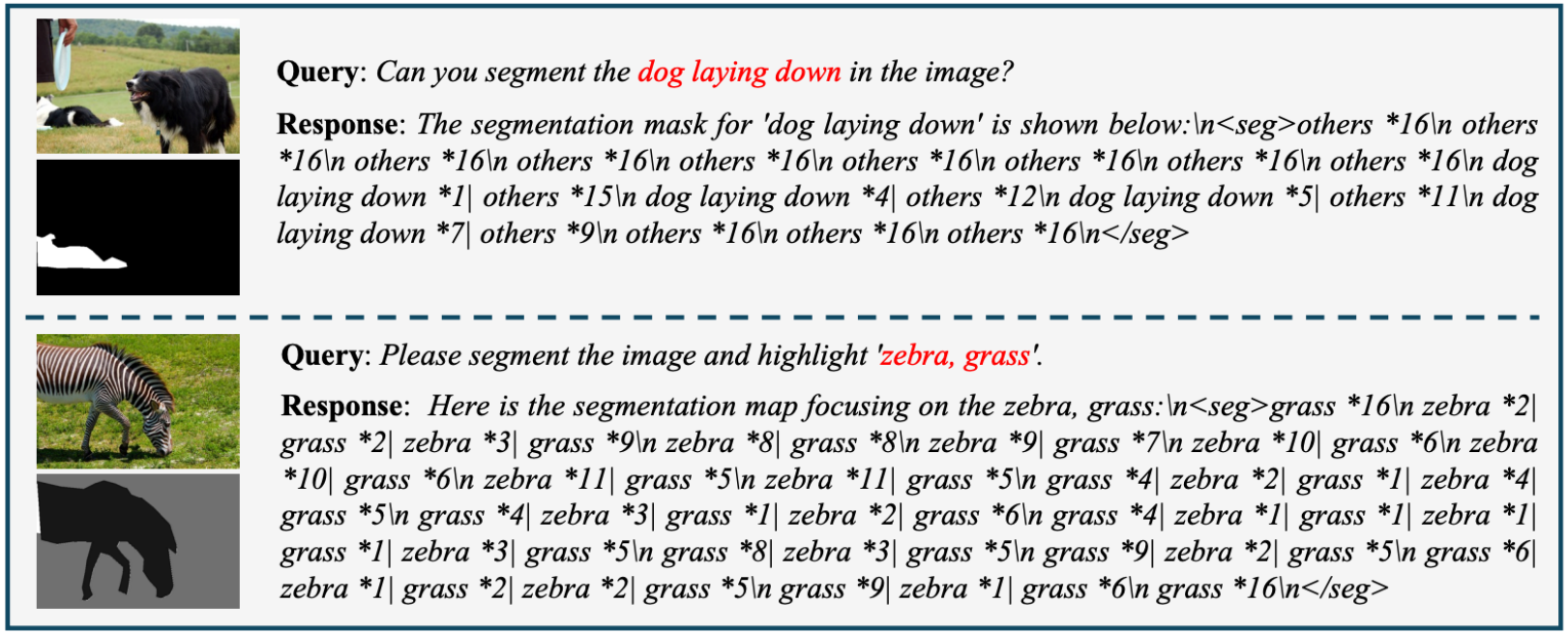
Forced adaptation of the architecture and training objectives creates internal conflicts and impairs the performance on other tasks.

Methods	Training Data	VQA			
		MMMU	MMBench	MMStar	ScienceQA
LISA-7B	Mix	0	0	0	0

Challenge in Dense perception

Current solution:

2. Represent masks as text tokens and refine details via post-processing. (Next-token Prediction)



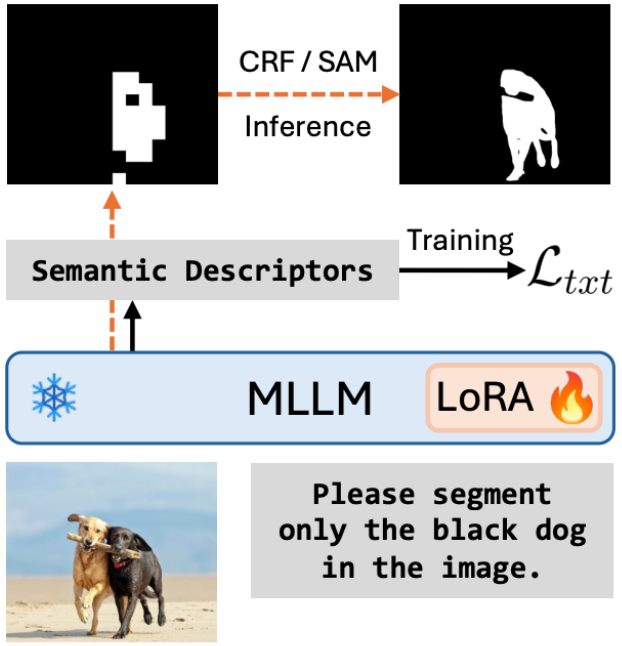
Query: *Can you segment the dog laying down in the image?*

Response: *The segmentation mask for 'dog laying down' is shown below:\n<seg>others *16\n others *16\n others *16\n others *16\n others *16\n others *16\n others *16\n others *16\n others *16\n dog laying down *1| others *15\n dog laying down *4| others *12\n dog laying down *5| others *11\n dog laying down *7| others *9\n others *16\n others *16\n others *16\n</seg>*

Query: *Please segment the image and highlight 'zebra, grass'.*

Response: *Here is the segmentation map focusing on the zebra, grass:\n<seg>grass *16\n zebra *2| grass *2| zebra *3| grass *9\n zebra *8| grass *8\n zebra *9| grass *7\n zebra *10| grass *6\n zebra *10| grass *6\n zebra *11| grass *5\n zebra *11| grass *5\n grass *4| zebra *2| grass *1| zebra *4| grass *5\n grass *4| zebra *3| grass *1| zebra *2| grass *6\n grass *4| zebra *1| grass *1| zebra *1| grass *1| zebra *3| grass *5\n grass *8| zebra *3| grass *5\n grass *9| zebra *2| grass *5\n grass *6| zebra *1| grass *2| zebra *2| grass *5\n grass *9| zebra *1| grass *6\n grass *16\n</seg>*

Text4Seg



M Lan et al., Text4Seg: Reimagining Image Segmentation as Text Generation, ICLR 2025

Similar works: SegZero (arXiv'25), SegAgent (CVPR'25), VisionLLM (NIPS'24) ···

Challenge in Dense perception

Current solution:

2. Represent masks as text tokens and refine details via post-processing. (Next-token Prediction)

Drawbacks:

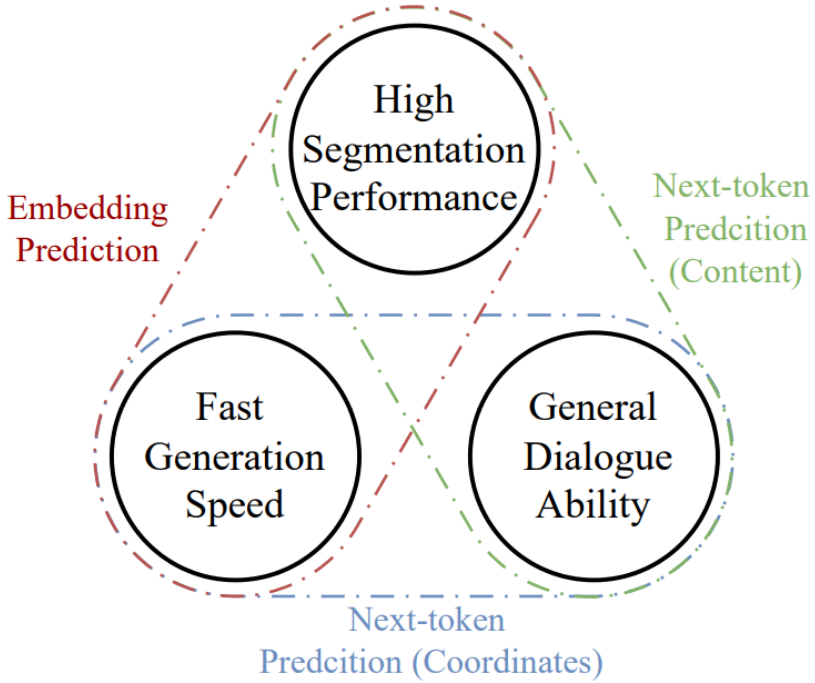
Extremely slow decoding speed due to massive mask tokens.

To segment one object, Text4Seg needs 6s.

Challenge in Dense perception

Our solution:

We want the model solve the “Trilemma”:

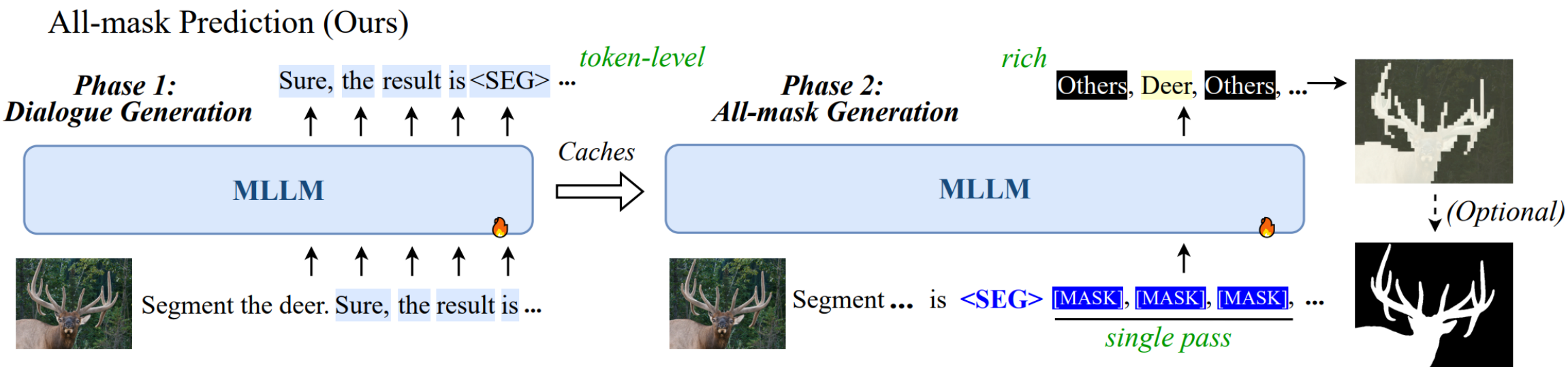


Only by solving this trilemma can dense perception efficiently assist reasoning without degrading general capabilities

Challenge in Dense perception

Our solution:

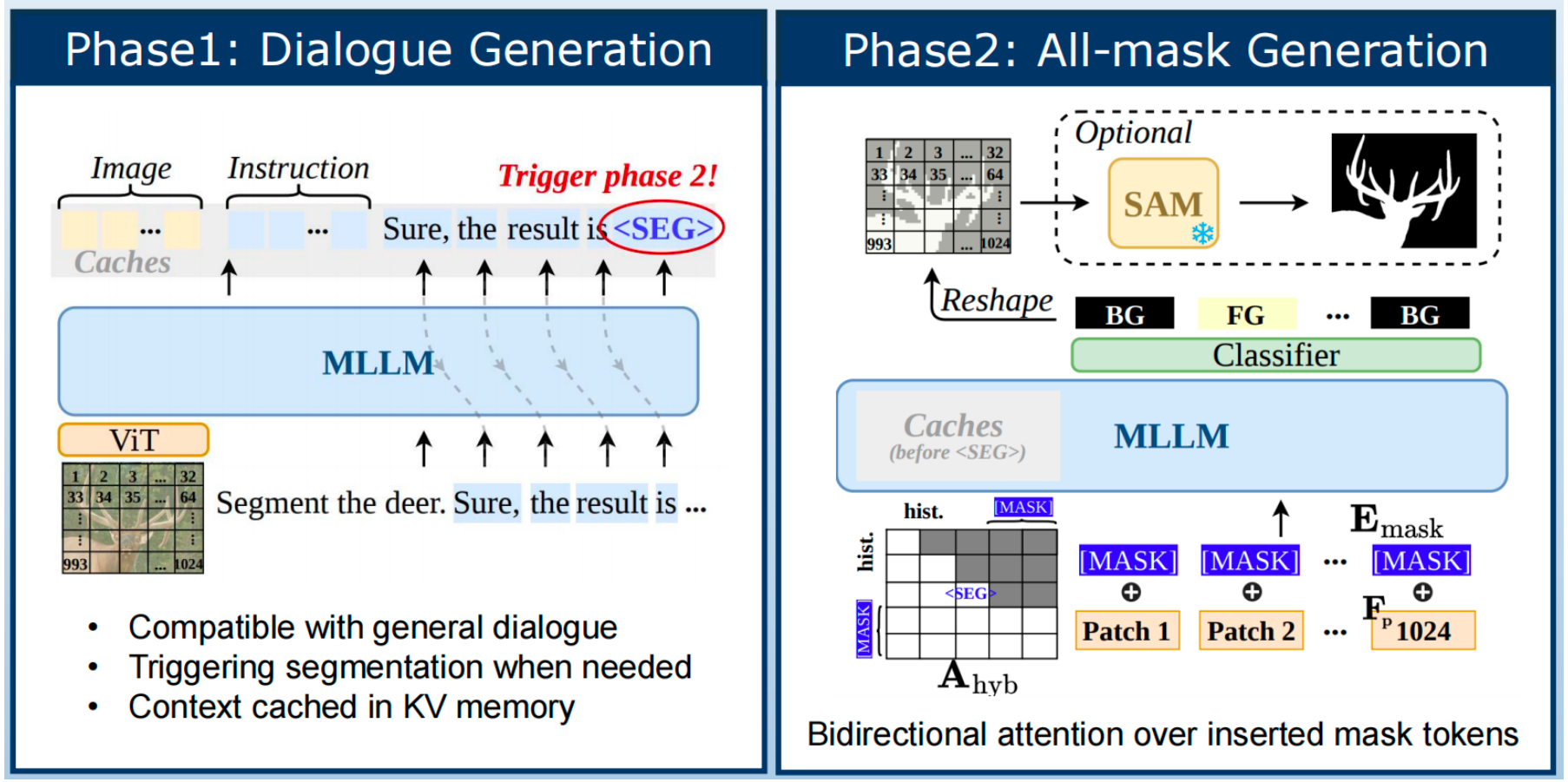
We want the model solve the “Trilemma”: we propose STAMP (CVPR, 26)



Jiazhen Liu et al., Better, Stronger, Faster: Tackling the Trilemma in MLLM-based Segmentation with Simultaneous Textual Mask Prediction, CVPR 2026

STAMP achieves a **Better** integration of dense perception into MLLMs, enabling **Faster** inference of perception maps to ground **Stronger** subsequent reasoning.

STAMP:


Phase1: Dialogue Generation
Image
Instruction
Trigger phase 2!
Sure, the result is <SEG>
Caches
MLLM
ViT
Segment the deer. Sure, the result is ...
Compatible with general dialogue
Triggering segmentation when needed
Context cached in KV memory
Phase2: All-mask Generation
Optional
SAM
Reshape
BG
FG
BG
Classifier
Caches
(before <SEG>)
MLLM
hist.
[MASK]
<SEG>
Patch 1
Patch 2
1024
Bidirectional attention over inserted mask tokens

STAMP:

Good segmentation performance

RefCOCO

Method	LLM	RefCOCO			RefCOCO+			RefCOCOg		Avg.
		val	testA	testB	val	testA	testB	val(U)	test(U)	
Specialised Baselines										
HIPIE (NIPS'24) [31]	BERT	78.3	-	-	66.2	-	-	69.8	-	-
ReLA (CVPR'23) [15]	BERT	73.8	76.5	70.2	66.0	71.0	57.7	65.0	66.0	68.3
PolyFormer-L (CVPR'23) [19]	BERT	76.0	78.3	73.3	69.3	74.6	61.9	69.2	70.2	71.6
UNINEXT-L(CVPR'24) [36]	BERT	80.3	82.6	77.8	70.0	74.9	62.6	73.4	73.7	74.4
Paradigm: Embedding Prediction										
PixelLM (CVPR'24) [28]	Vicuna-7B	73.0	76.5	68.2	66.3	71.7	58.3	69.3	70.5	69.2
LISA (CVPR'24) [12]	Vicuna-7B	74.9	79.1	72.3	65.1	70.8	58.1	67.9	70.6	69.9
LISA (CVPR'24) [12]	Vicuna-13B	76.0	78.8	72.9	65.0	70.2	58.1	69.5	70.5	70.1
GSVA (CVPR'24) [35]	Vicuna-7B	77.2	78.9	73.5	65.9	69.6	59.8	72.7	73.3	71.4
READ (CVPR'25) [25]	Vicuna-7B	78.1	80.2	73.2	68.4	73.7	60.4	70.1	71.4	71.9
GSVA (CVPR'24) [35]	Vicuna-13B	78.2	80.4	74.2	67.4	71.5	60.9	74.2	75.6	72.8
Paradigm: Token Prediction										
Mask-Decoder-Free										
Text4Seg (ICLR'25) [13]	Vicuna-13B	74.1	76.4	72.4	68.5	72.8	63.6	69.1	70.1	70.9
Text4Seg (ICLR'25) [13]	InternLM2.5-7B	74.7	77.4	71.6	68.5	73.6	62.9	70.7	71.6	71.4
STAMP	Qwen2-2B	77.7	79.4	76.1	73.4	76.4	69.7	74.9	75.1	75.3
STAMP	Qwen2-7B	**78.1**	**79.2**	**76.8**	**74.7**	**77.6**	**70.9**	**75.7**	**76.2**	**76.2**
With Mask Decoder										
Seg-Zero (arXiv'25) [21]	Qwen2.5-3B	-	79.3	-	-	73.7	-	-	71.5	-
Seg-Zero (arXiv'25) [21]	Qwen2.5-7B	-	80.3	-	-	76.2	-	-	72.6	-
SegLLM (ICLR'25) [32]	Vicuna-7B	80.2	81.5	75.4	70.3	73.0	62.5	72.6	73.6	73.6
Text4Seg (ICLR'25) [13]	Vicuna-7B	79.3	81.9	76.2	72.1	77.6	66.1	72.1	73.9	74.9
Text4Seg (ICLR'25) [13]	InternLM2.5-7B	79.2	81.7	75.6	72.8	77.9	66.5	74.0	75.3	75.4
SegAgent (CVPR'25) [42]	Qwen-7B	79.7	81.4	76.6	72.5	75.8	66.9	75.1	75.2	75.4
Text4Seg (ICLR'25) [13]	Vicuna-13B	80.2	82.7	77.3	73.7	78.6	67.6	74.0	75.1	76.2
STAMP	Qwen2-2B	81.9	83.7	79.5	77.1	80.5	72.7	78.5	78.8	79.1
STAMP	Qwen2-7B	**83.1**	**84.5**	**80.8**	**79.4**	**82.8**	**74.6**	**79.9**	**80.4**	**80.7**

gRefCOCO

Method	LLM	Val Set		Test Set A		Test Set B		Avg.
		gIoU	cIoU	gIoU	cIoU	gIoU	cIoU	
Mask-Decoder-Free								
Text4Seg	InternLM2.5-7B	70.0	66.1	69.4	70.9	63.1	64.1	67.3
Text4Seg	Vicuna-13B	70.3	66.9	69.8	71.4	63.8	64.4	67.8
STAMP	Qwen2-2B	73.9	70.3	73.6	73.7	67.5	68.1	71.2
STAMP	Qwen2-7B	**74.4**	**70.9**	**73.8**	**74.7**	**68.1**	**69.1**	**71.8**
With Mask Decoder								
LAVT	BERT	58.4	57.6	65.9	65.3	55.8	55.0	59.7
LISA	Vicuna-7B	61.6	61.8	66.3	68.5	58.8	60.6	62.9
ReLA	BERT	63.6	62.4	70.0	69.3	61.0	59.9	64.4
LISA	Vicuna-13B	63.5	63.0	68.2	69.7	61.8	62.2	64.7
GSVA	Vicuna-7B	66.5	63.3	71.1	69.9	62.2	60.5	65.6
Text4Seg	InternLM2.5-7B	74.4	69.1	75.1	73.8	67.3	66.6	71.1
Text4Seg	Vicuna-13B	74.8	69.8	75.1	74.3	68.0	67.1	71.5
STAMP	Qwen2-2B	76.4	72.2	76.5	75.7	70.0	69.8	73.4
STAMP	Qwen2-7B	**77.6**	**73.6**	**77.6**	**77.2**	**71.4**	**71.6**	**74.8**

ReasonSeg

Method	LLM	Val		Test		Avg.
		gIoU	cIoU	gIoU	cIoU	
OVSeg	Vicuna-7B	28.5	18.6	26.1	20.8	23.5
LISA	Vicuna-7B	53.6	52.3	48.7	48.8	50.9
SegLLM	Vicuna-7B	57.2	54.3	52.4	48.4	53.1
Text4Seg	Qwen2-7B	59.1	49.5	57.1	52.1	54.5
Seg-Zero	Qwen2.5-7B	62.6	62.0	57.5	52.0	58.5
READ	Vicuna-7B	59.8	**67.6**	58.5	58.6	61.1
STAMP	Qwen2-2B	**65.1**	63.9	**62.7**	**60.9**	**63.2**

STAMP:

Dialogue capability is maintained, and even promotes segmentation capability.

Methods	Training Data	VQA					RES (val)		
		MMMU	MMBench	MMStar	ScienceQA	TextVQA	RefC	RefC+	RefCg
LLaVA-1.5-7B	VQA	35.7	66.5	33.1	68.4	55.0	n.a.	n.a.	n.a.
Qwen2-VL-2B	VQA	**38.3**	66.5	42.1	70.2	**70.7**	n.a.	n.a.	n.a.
LISA-7B	Mix	0	0	0	0	0	74.9	65.1	67.9
READ-7B	Mix	1.1	0	14.4	23.2	22.6	78.1	68.4	70.1
Text4Seg-7B	Mix	34.0	54.8	33.4	68.1	55.0	77.5	70.7	73.4
STAMP-2B	Seg.	n.a.	n.a.	n.a.	n.a.	n.a.	81.9	77.1	78.5
STAMP-2B	Mix	37.8	**68.7**	**42.4**	**72.6**	69.7	**82.2**	**77.3**	**79.0**

STAMP:

Fast inference speed.