Do You Keep an Eye on What I Ask? Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Yeongjae Cho, Keonwoo Kim, Taebaek Hwang, Sungzoon Cho

Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Problem Definition

- **Growing interest in Large Vision-Language Models (LVLMs) with advancements in language models**
- **Object Hallucination: Describing non-existent objects or incorrect details**
- **Reduced reliability in Visual Question Answering and Image Captioning tasks**

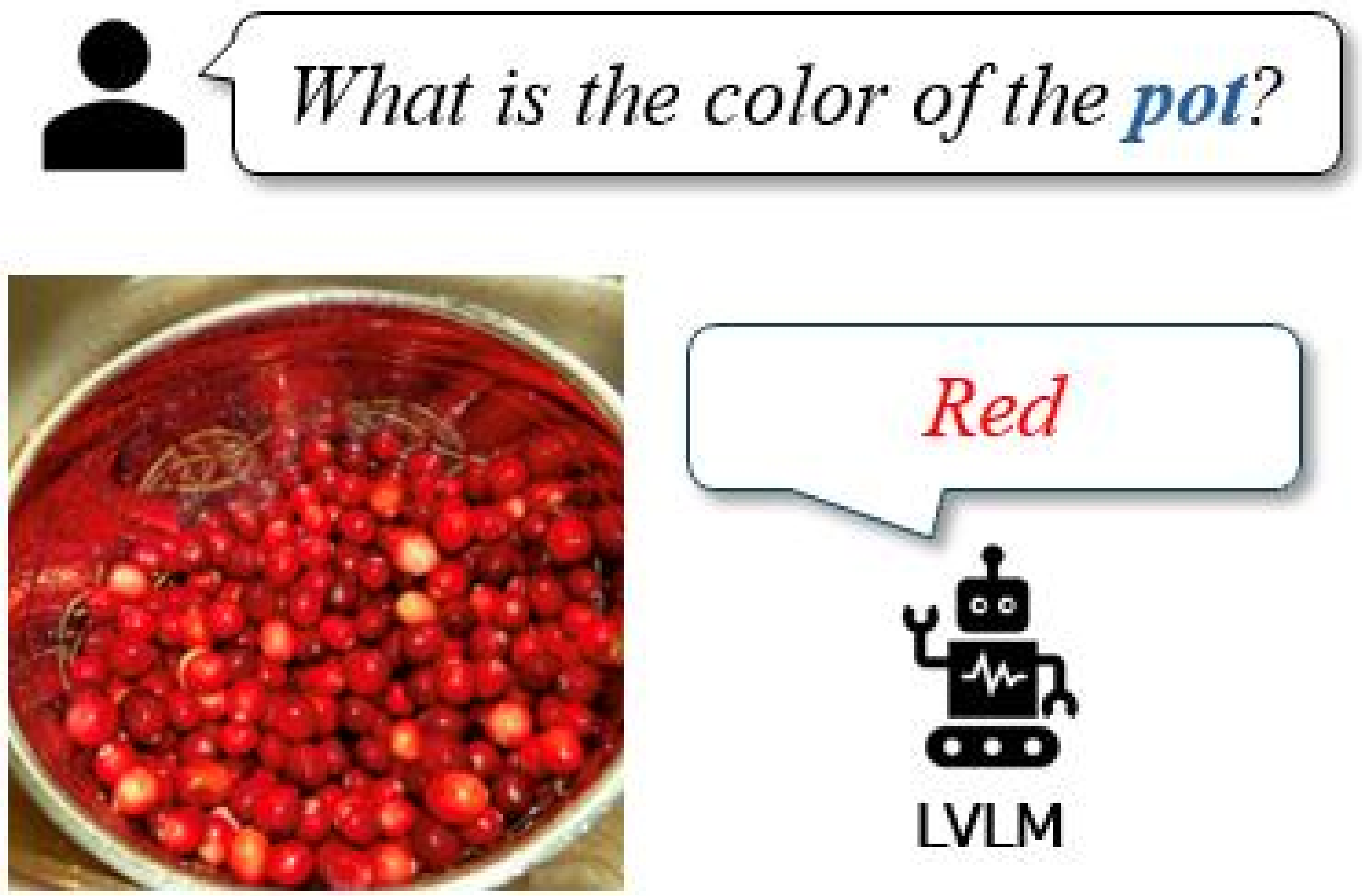

Fig. Example of Object Hallucination in Visual Question Answering Task

Fig. Example of Object Hallucination in Image Captioning Task

Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Motivation

- **Some Object Hallucination cases can be easily resolved using the Crop & Resize technique.**
- **Established two assumptions to mitigate Object Hallucination:**
    a. **Reducing the number of unnecessary objects**
    b. **Ensuring high resolution**

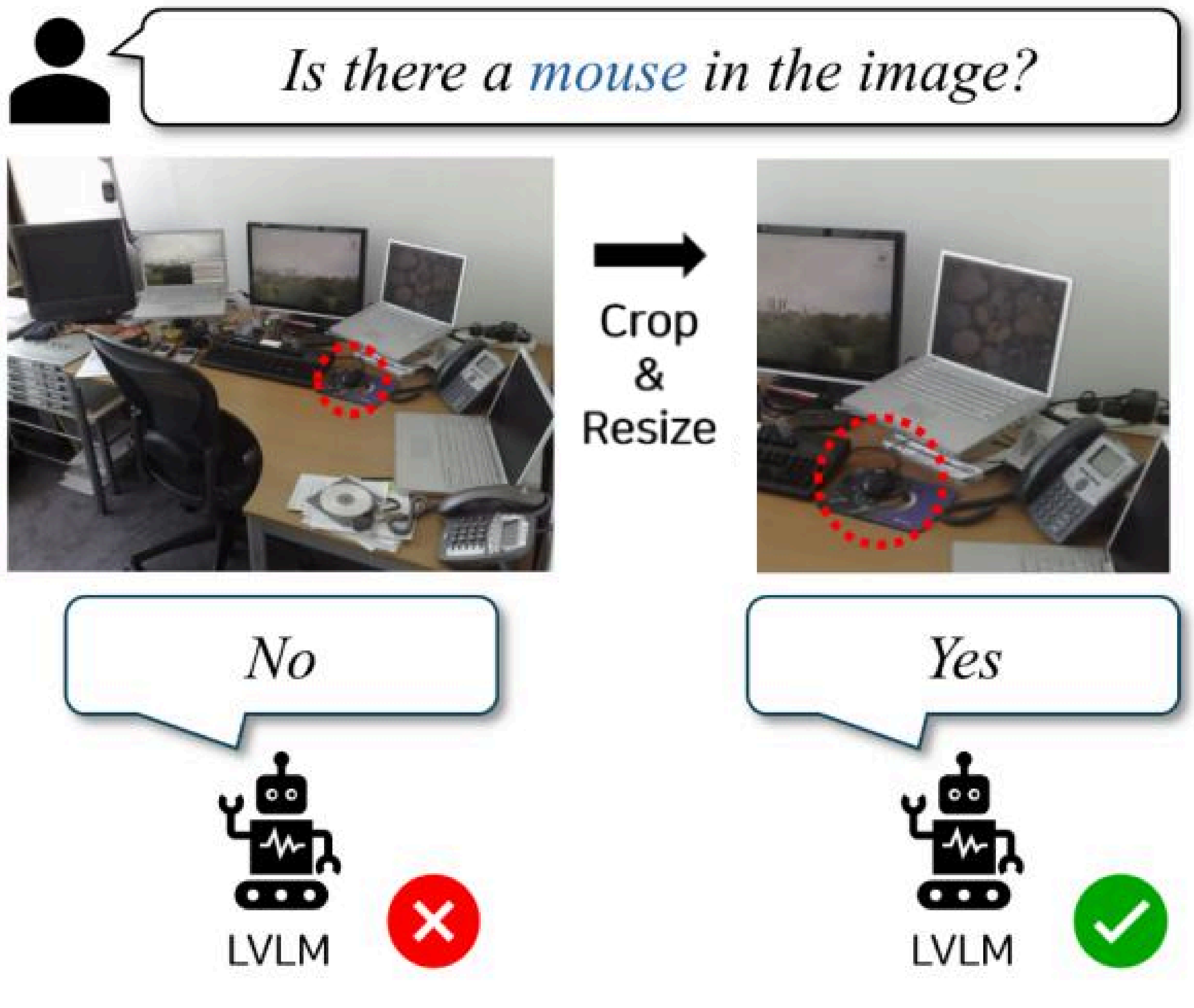

Fig. Example of Object Hallucination and its Modified Case

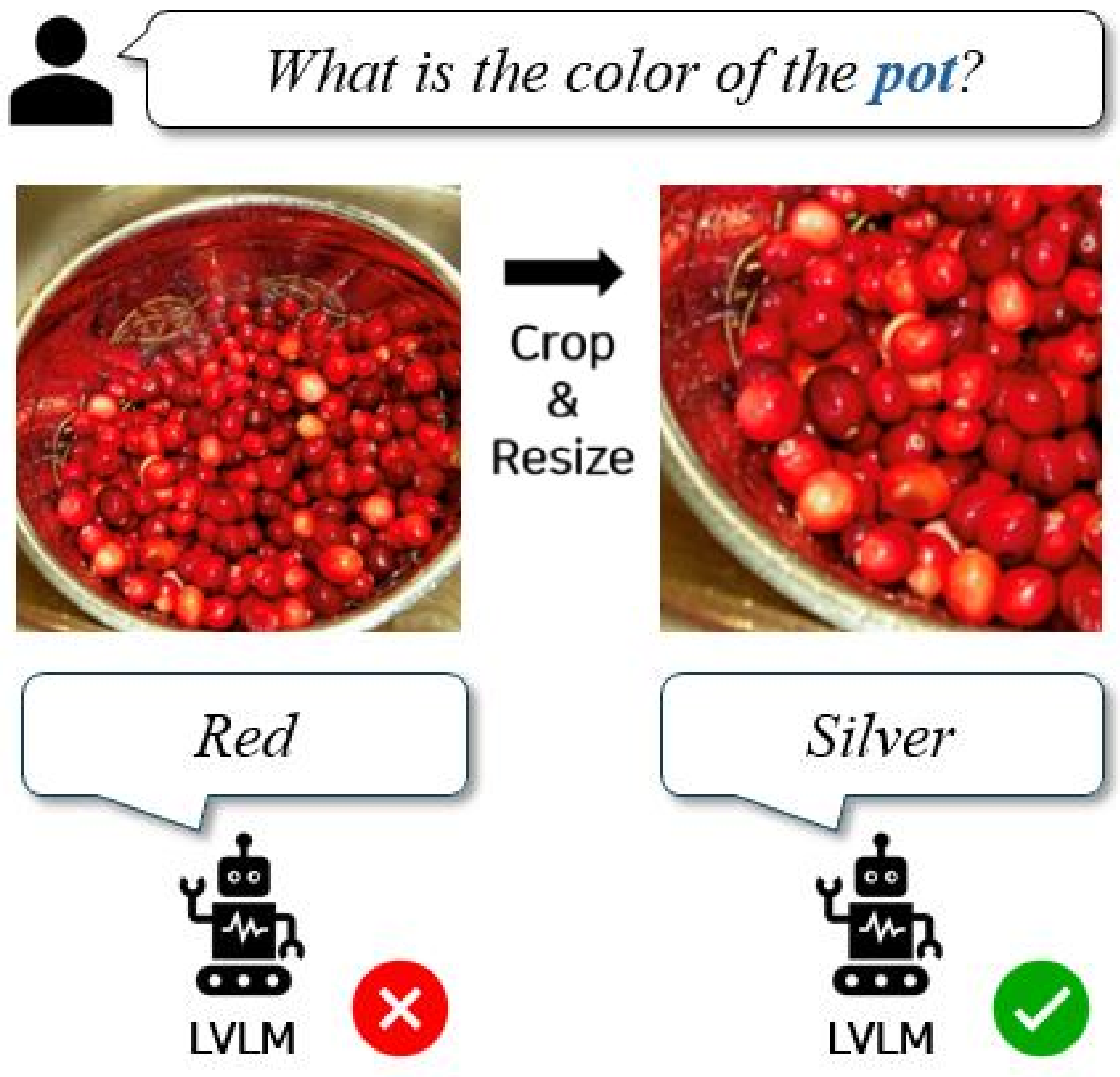

Fig. Example of Object Hallucination and its Modified Case

Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Method

- **Proposed Ensemble Decoding (ED) utilizing attention-guided weights and sub-image logit distribution.**
- **Introduced an optimized version (FastED) and ED Adaptive Plausibility Constraint.**

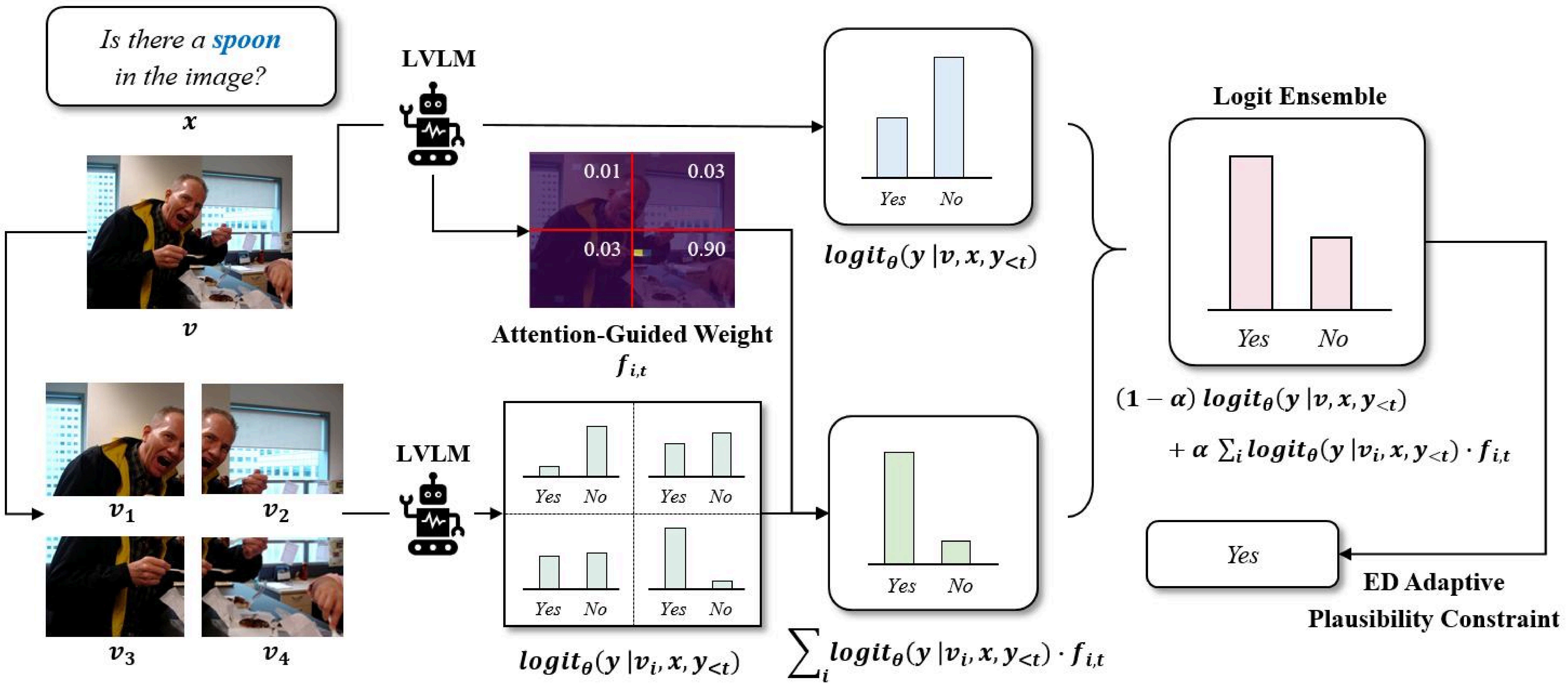

Fig. Overall Pipeline of Ensemble Decoding

Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Method

- **Proposed Ensemble Decoding (ED) utilizing attention-guided weights and sub-image logit distribution.**
- **Introduced an optimized version (FastED) and ED Adaptive Plausibility Constraint.**

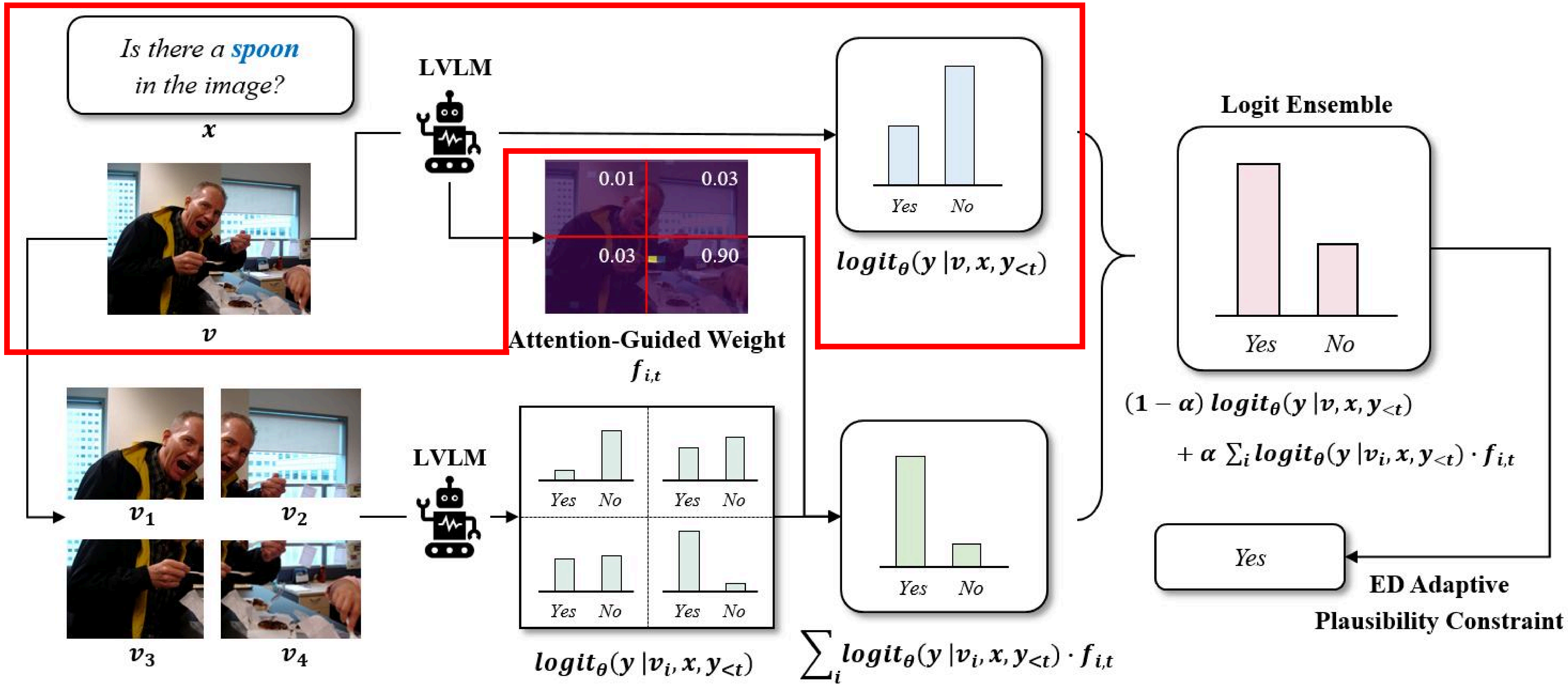

Fig. Overall Pipeline of Ensemble Decoding

Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Method

- **Proposed Ensemble Decoding (ED) utilizing attention-guided weights and sub-image logit distribution.**
- **Introduced an optimized version (FastED) and ED Adaptive Plausibility Constraint.**

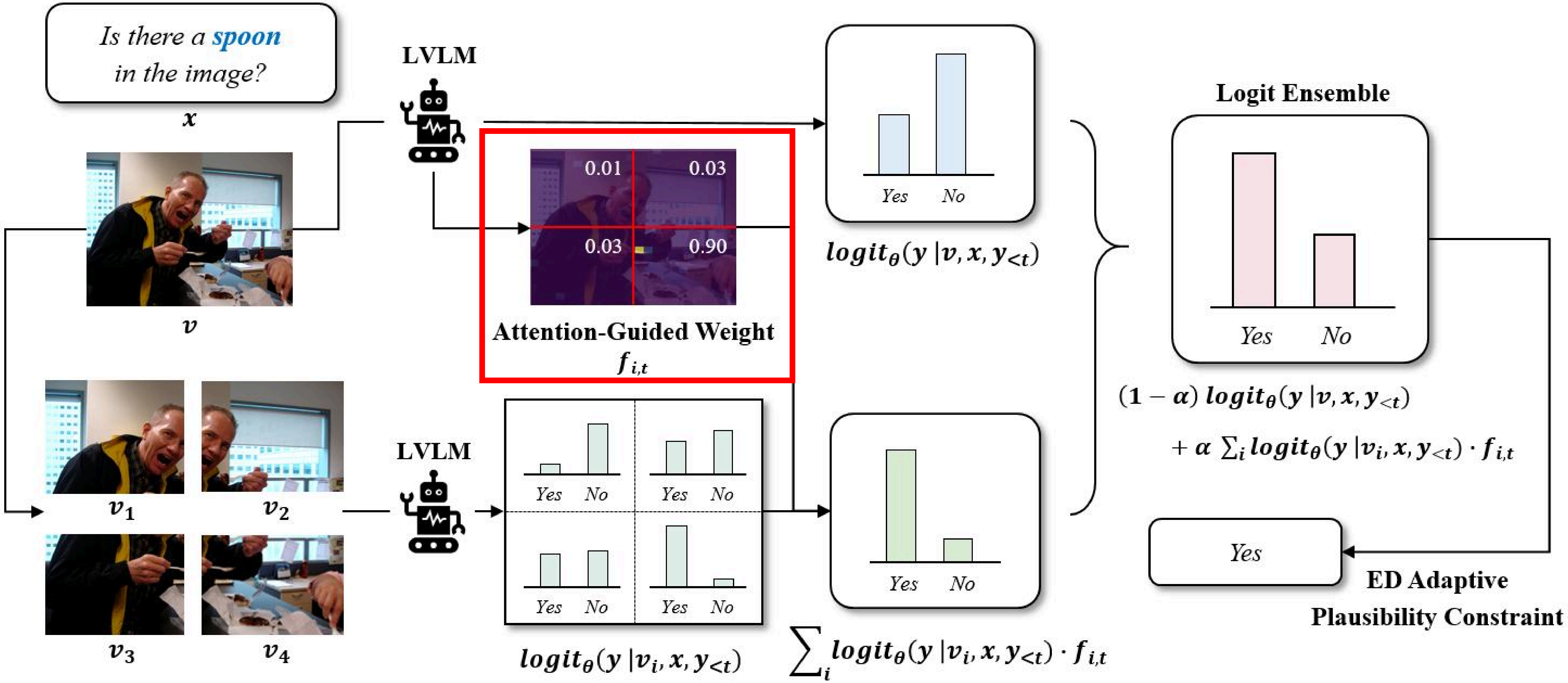

Fig. Overall Pipeline of Ensemble Decoding

Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Method

- **Proposed Ensemble Decoding (ED) utilizing attention-guided weights and sub-image logit distribution.**
- **Introduced an optimized version (FastED) and ED Adaptive Plausibility Constraint.**

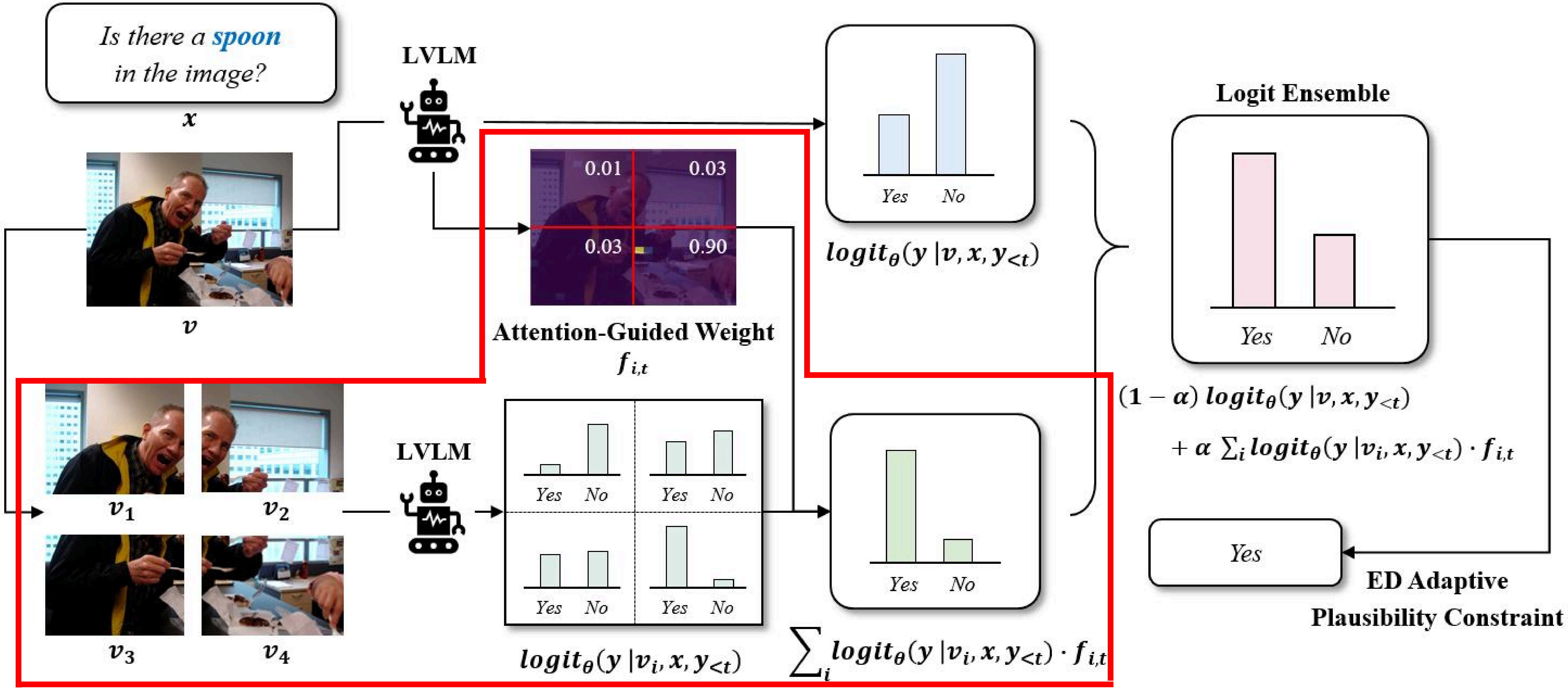

Fig. Overall Pipeline of Ensemble Decoding

Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Method

- **Proposed Ensemble Decoding (ED) utilizing attention-guided weights and sub-image logit distribution.**
- **Introduced an optimized version for image captioning (FastED) and ED Adaptive Plausibility Constraint.**

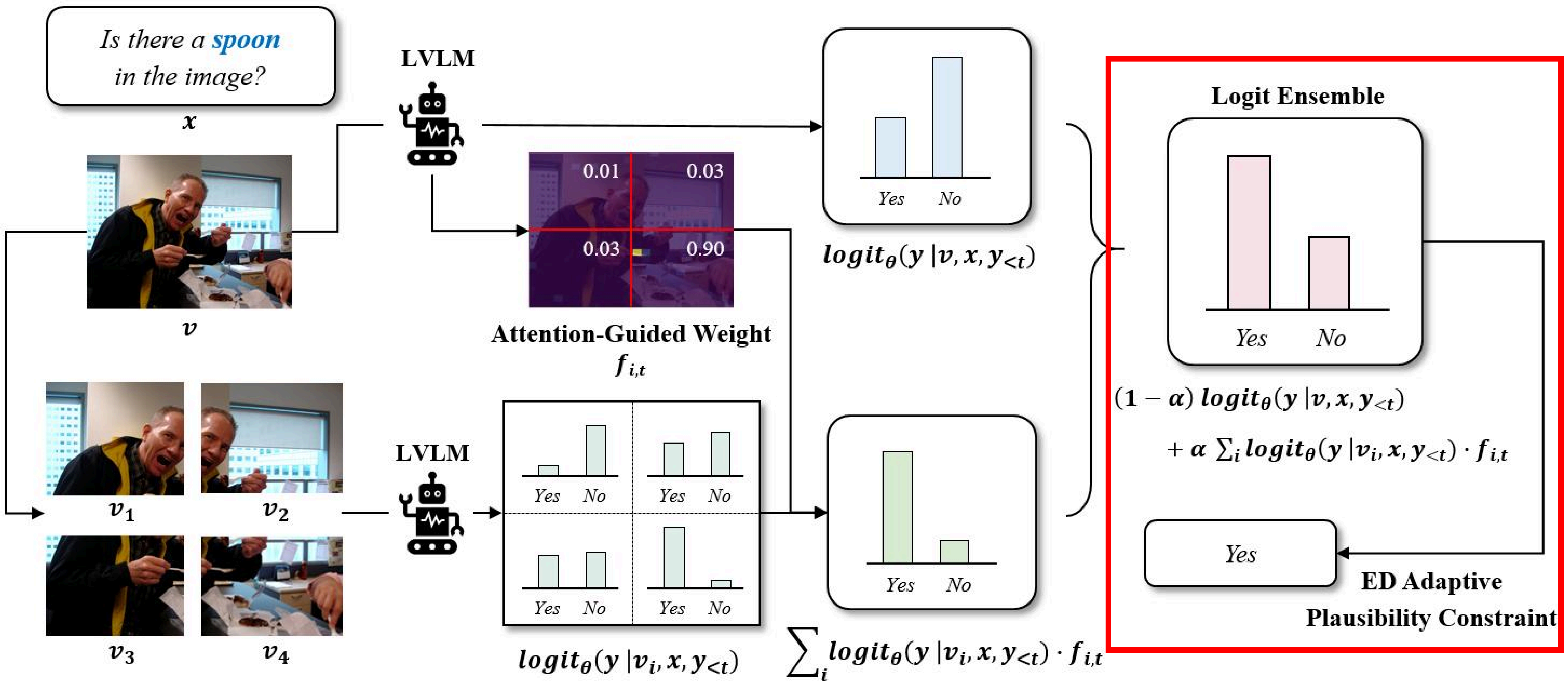

Fig. Overall Pipeline of Ensemble Decoding

Mitigating Multimodal Hallucination via Attention-Guided Ensemble Decoding

ICLR 2025

Experimental Results

- **Achieved the highest F1 Score and Accuracy on the hallucination benchmark (POPE).**
- **Outperformed other decoding strategies across all metrics in CHAIR, which requires generating long-form answers.**

Tab. Experimental Results of VQA Hallucination Dataset (POPE)

Setting	Decoding	Precision	Recall	F1 Score	Accuracy
Random	Regular	88.84	76.76	82.28	83.49
	DOLA	87.59	81.27	84.19	84.78
	OPERA	94.52	79.80	86.45	87.53
	VCD	87.15	86.68	86.83	86.84
	AGLA	94.41	82.08	87.71	88.54
	ED	93.40	86.41	**89.68**	**90.08**
Popular	Regular	82.47	76.76	79.34	79.98
	DOLA	84.11	76.22	80.61	79.75
	OPERA	88.00	79.80	83.50	84.21
	VCD	87.15	80.59	83.37	82.65
	AGLA	87.88	82.08	84.68	85.14
	ED	86.12	86.41	**86.00**	**86.09**
Adversarial	Regular	76.11	76.80	76.26	76.03
	DOLA	77.27	75.47	76.16	76.32
	OPERA	82.16	79.76	80.69	80.88
	VCD	73.43	86.47	79.28	77.31
	AGLA	81.20	82.10	81.36	81.13
	ED	79.75	86.47	**81.90**	**82.75**

Tab. Experimental Results of Image Captioning Dataset (CHAIR)

Decoding	$CHAIR_S\downarrow$	$CHAIR_I\downarrow$	Recall↑	Average Length
Regular	51.0	15.2	75.2	102.2
DOLA	57.0	15.9	78.2	97.5
OPERA	47.0	14.6	78.5	95.3
VCD	51.0	14.9	77.2	101.9
AGLA	**43.0**	14.1	78.9	98.8
ED	**43.0**	**14.0**	**82.5**	100.1

Qualitative Evaluation

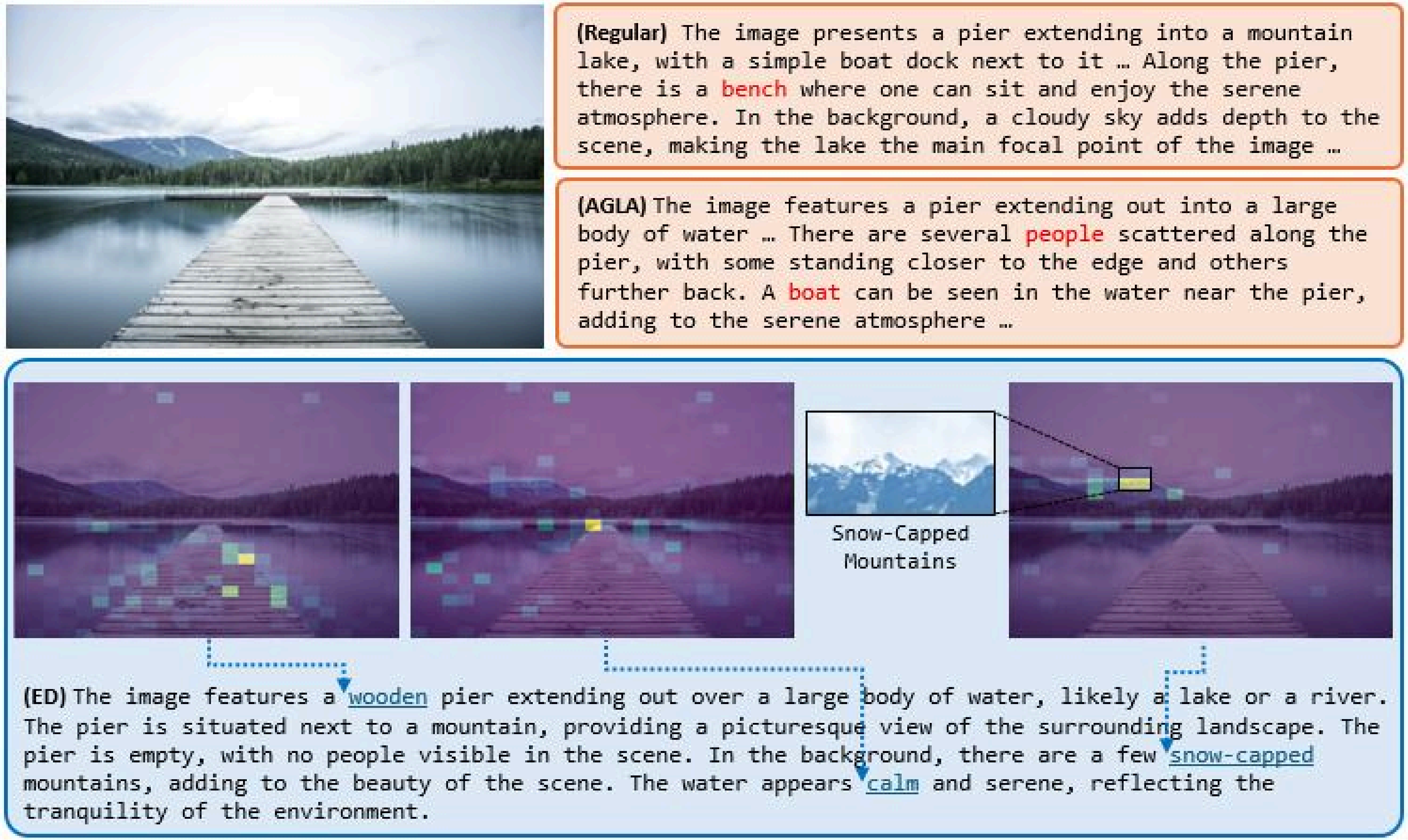

Fig. Generated captions using Regular, AGLA, and ED decoding strategies

Conclusion

- **Confirmed the superior performance of ED and FastED in hallucination benchmarks and image captioning tasks.**
- **FastED improves speed and accuracy, while ED is advantageous for detailed tasks, allowing selective use based on user needs.**
- **Proposed ED Adaptive Plausibility Constraint, better suited for ED and FastED compared to previous constraints.**
- **Minimized external dependencies by actively leveraging the model's inherent visual capabilities without external modules.**
- **Applicable to models without additional training and offers scalability in terms of model size.**

Thank You!

ICLR 2025

Yeongjae Cho, Keonwoo Kim, Taebaek Hwang, Sungzoon Cho