西北工业大学
NORTHWESTERN POLYTECHNICAL UNIVERSITY

空天地海一体化大数据应用技术国家工程实验室
National Engineering Laboratory for
Integrated Aero-Space-Ground-Ocean Big Data Application Technology

Agency for Science, Technology and Research
SINGAPORE

Vision-Aware Self-Verification for Hallucination Detection in Medical VLMs

廖泽慧（博士生）

空天地海一体化大数据应用技术国家工程实验室

西北工业大学，计算机学院

Email: merrical@mail.nwpu.edu.cn



Hallucinations in Medical VLMs

➢ Hallucination vs. Error

Hallucinations in Neural Machine Translation [1]

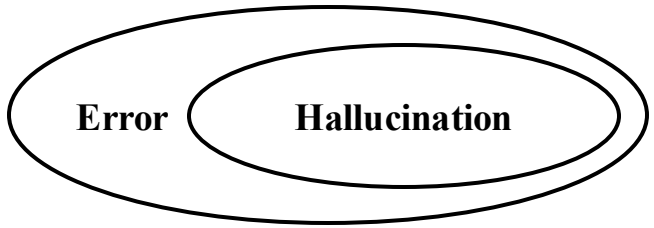
These mistranslations are completely semantically incorrect and also grammatically viable. They are untethered from the input so we name them 'hallucinations'.

Survey of Hallucination in LLMs [2]:

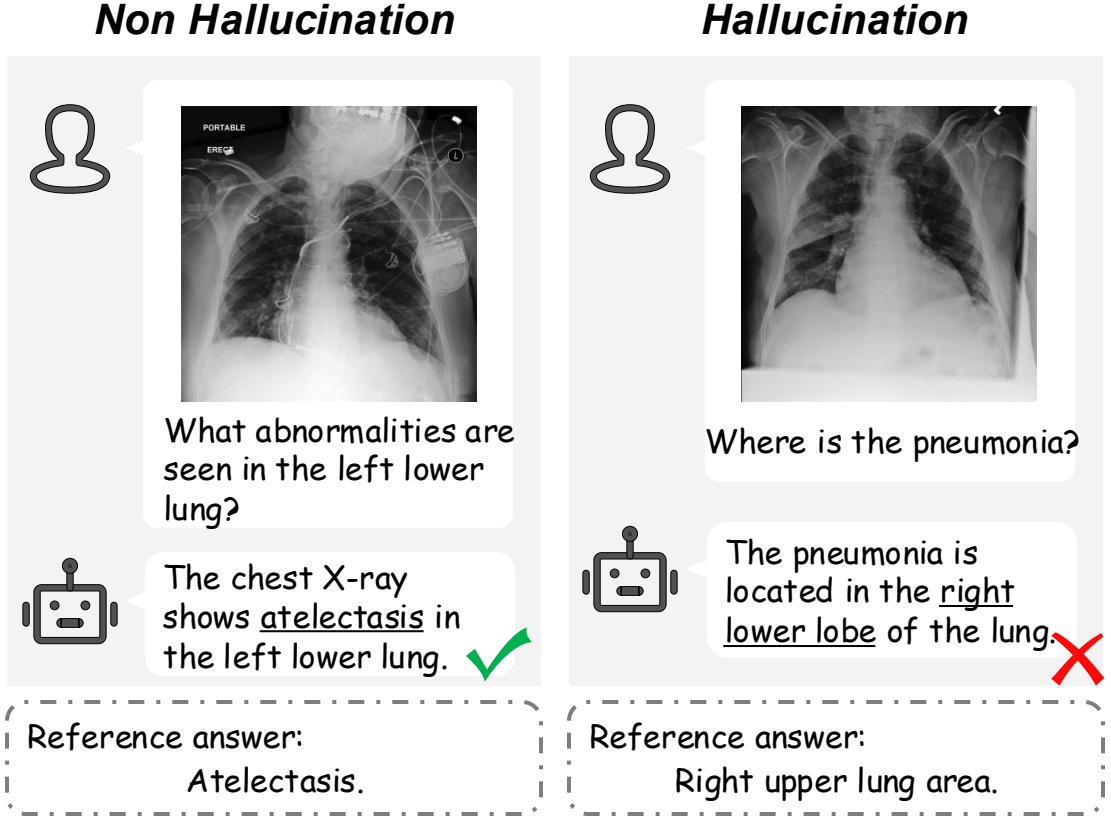
LLMs exhibit a critical tendency to produce hallucinations, resulting in content that is inconsistent with real-world facts or user inputs.

Survey of Hallucination in VLMs [3]:

…, "hallucination", or more specifically, the misalignment between factual visual content and corresponding textual generation, poses a significant challenge of utilizing LVLMs.



➢ Medical VLMs are prone to 'hallucinations'



- Hallucinations in clinical decision-making present significant risks.
- Establishing trust in MLLMs among clinicians and patients is crucial for their real-world adoption.

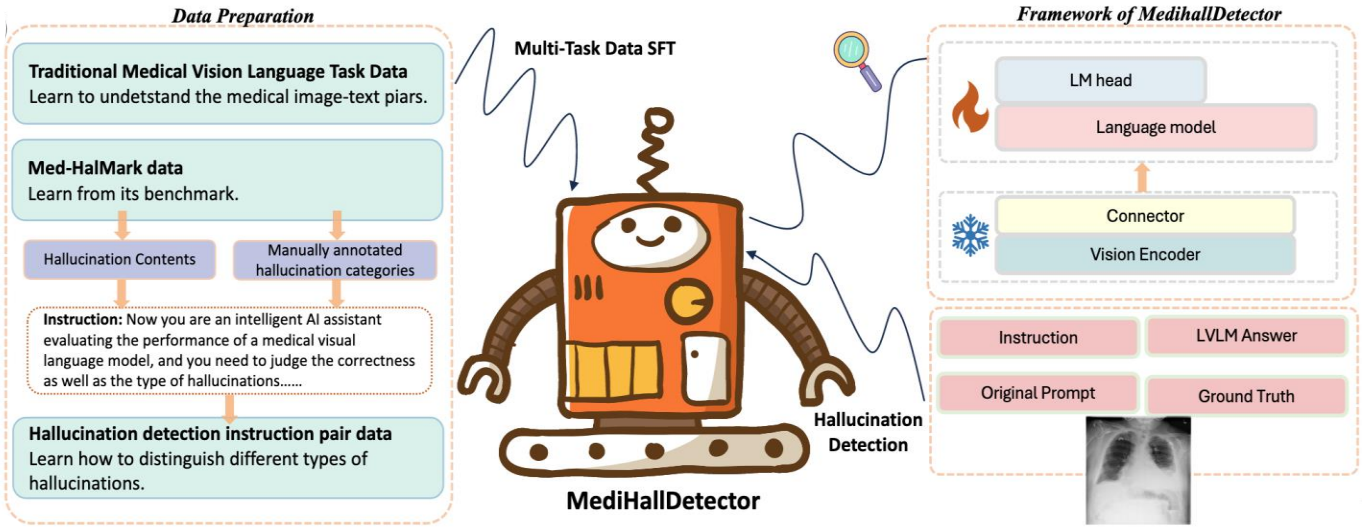
[1] Agarwal, Ashish, et al. "Hallucinations in neural machine translation." *ICLR*. 2018.
[2] Huang, Lei, et al. "A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions." *ACM TIS,* 43.2 (2025): 1-55.
[3] Liu, Hanchao, et al. "A survey on hallucination in large vision-language models." *arXiv preprint arXiv:2402.00253* (2024).

Existing Hallucination Detection Methods in VLMs/LLMs

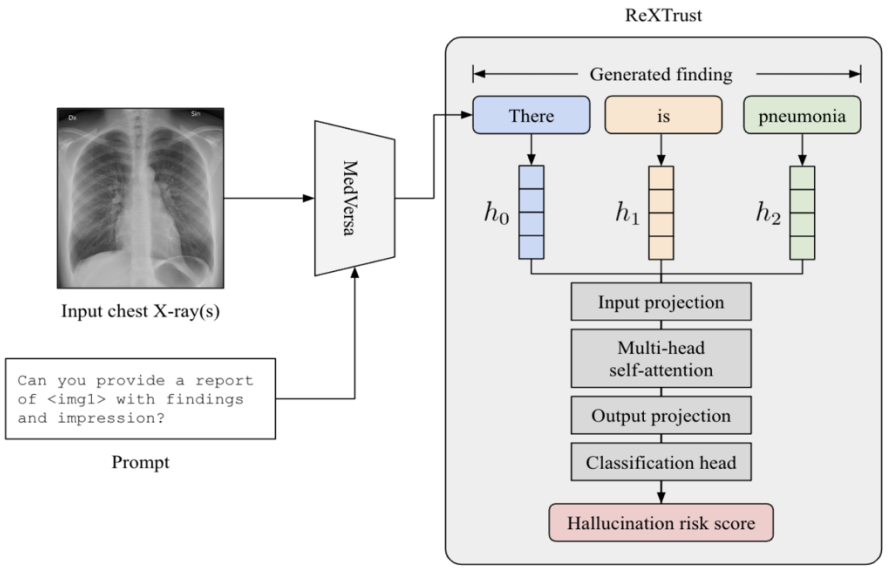
- **Recent efforts broadly fall into three categories.**

(1) Supervised Detectors
(2) External-Verification Methods
(3) Uncertainty-based Methods

These methods require costly annotated hallucination data and often generalize poorly to unseen scenarios.



MediHallDetector [1]



ReXTrust [2]

[1] Chen, Jiawei, et al. "Detecting and evaluating medical hallucinations in large vision language models." *arXiv preprint arXiv:2406.10185* (**2024**).
[2] Hardy, Romain, et al. "ReXTrust: A Model for Fine-Grained Hallucination Detection in AI-Generated Radiology Reports." ***AIMedHealth*** (**2024**).

Existing Hallucination Detection Methods in VLMs/LLMs

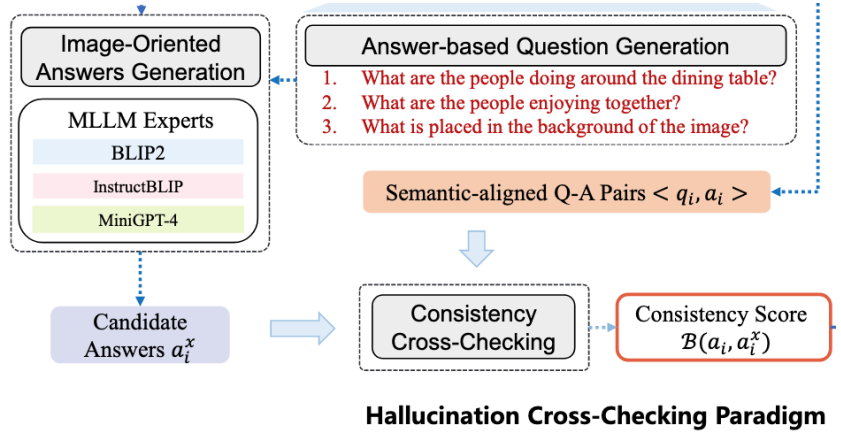
➢ **Recent efforts broadly fall into three categories.**

(1) Supervised Detectors

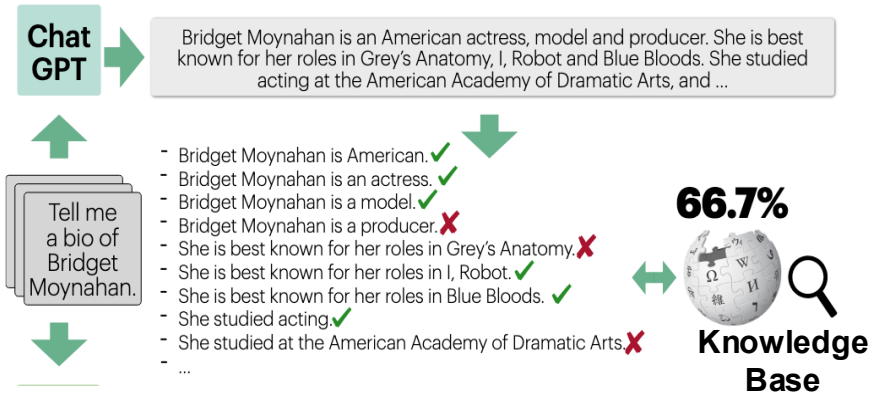
(2) External-Verification Methods

(3) Uncertainty-based Methods

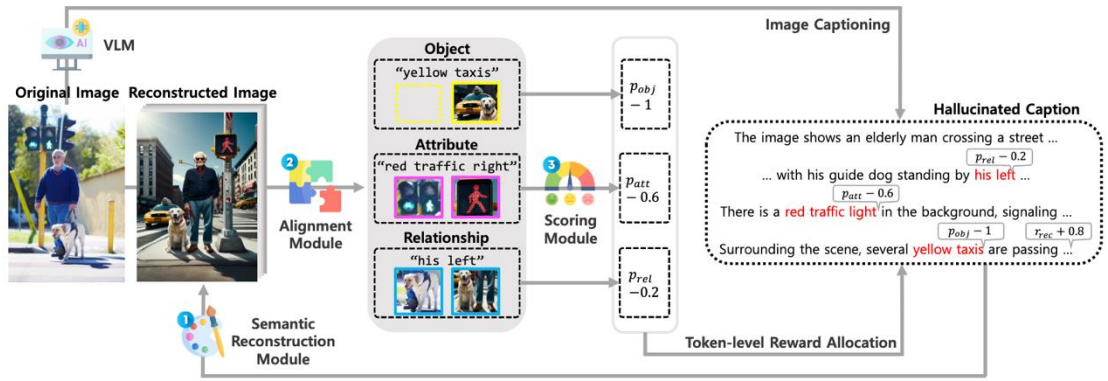
These methods rely on additional sources of information, such as other LLMs or VLMs, vision expert models, or external knowledge bases.



Hallucidoctor [1]



FActScore [2]



ESREAL [3]

[1] Yu, Qifan, et al. "Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data." ***CVPR*** **(2024)**.
[2] Min, Sewon, et al. "FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation." ***EMNLP***. **2023**.
[3] Kim, Minchan, et al. "ESREAL: Exploiting Semantic Reconstruction to Mitigate Hallucinations in Vision-Language Models." ***ECCV*** **(2024)**.

Existing Hallucination Detection Methods in VLMs/LLMs

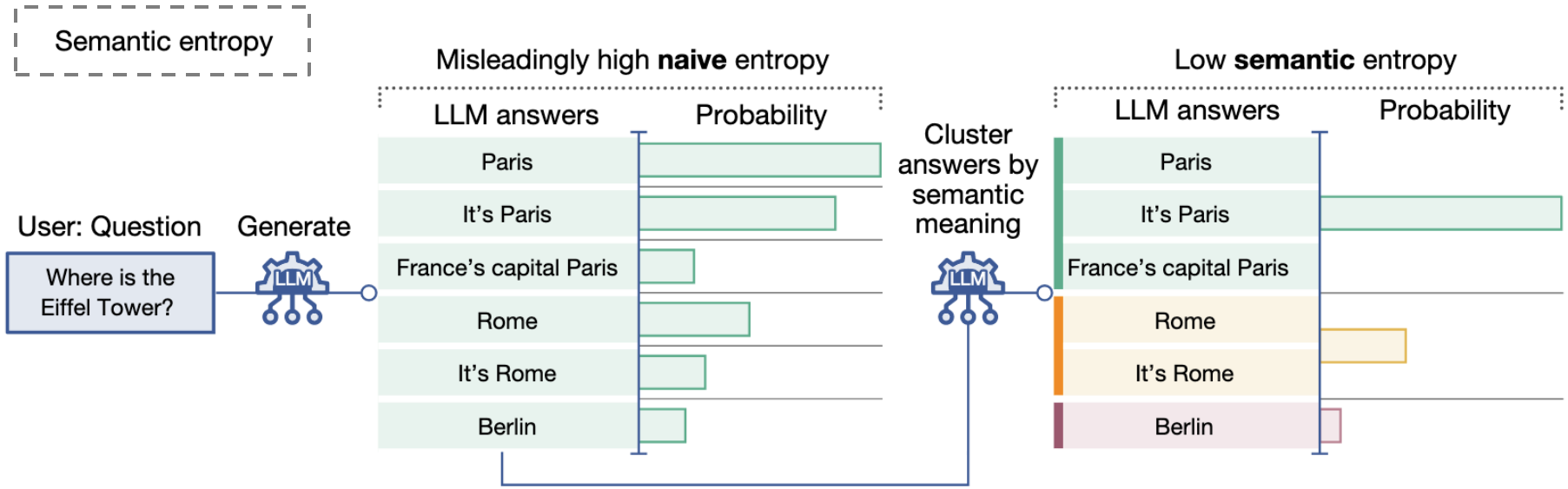
➢ **Recent efforts broadly fall into three categories.**

(1) Supervised Detectors
(2) External-Verification Methods
(3) Uncertainty-based Methods
- Token-level
- Embedding-level
- Sentence-level

Predictive uncertainty ↑
↓
Likelihood of hallucination ↑

These methods requires no auxiliary models, external knowledge bases, or task-specific fine-tuning.

A Notable Advancement in Uncertainty Estimation for LLMs: Semantic Entropy [1]



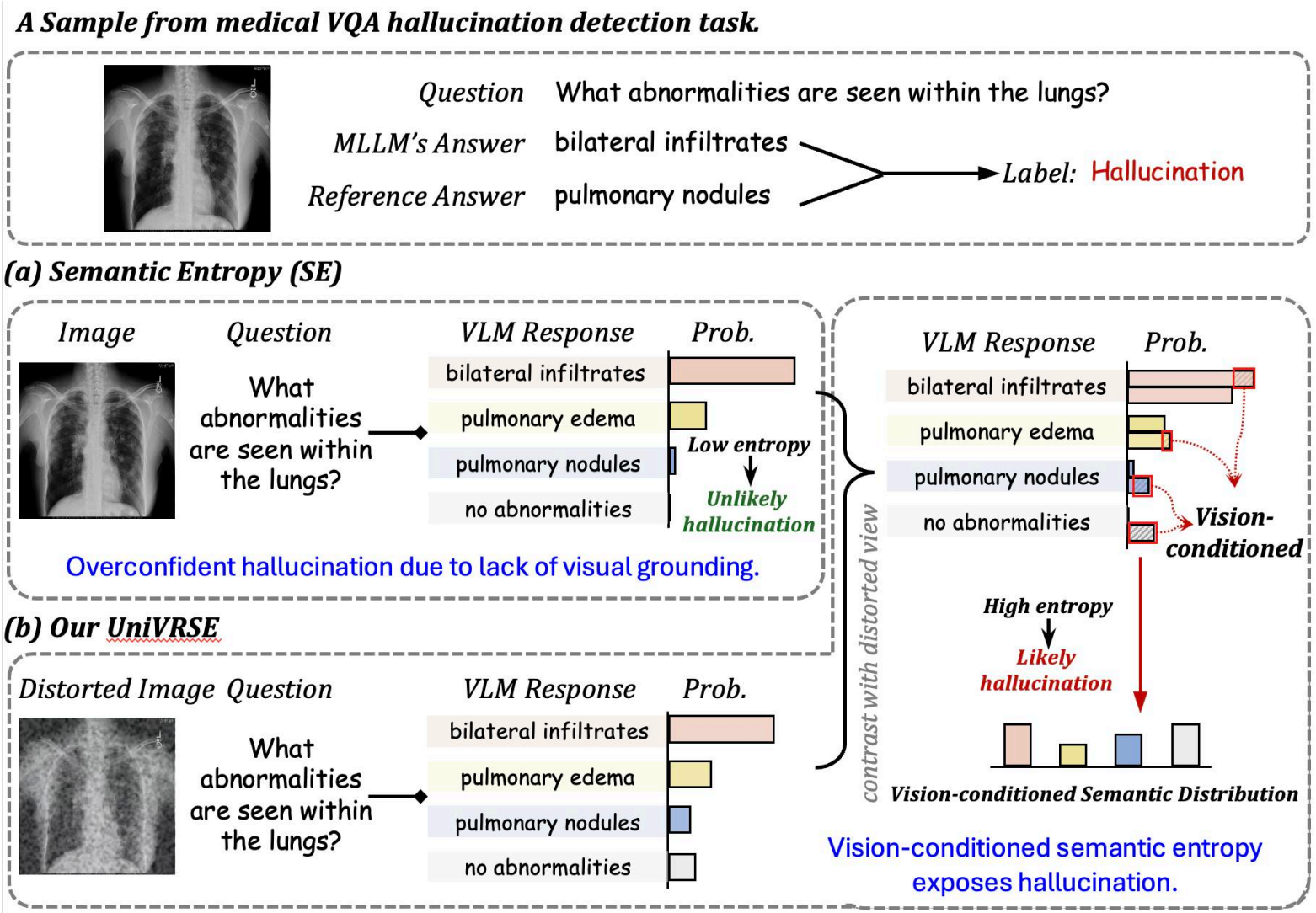
[1] Farquhar S, Kossen J, Kuhn L, et al. Detecting hallucinations in large language models using semantic entropy[J]. ***Nature***, 2024, 630(8017): 625-630.

Applying Semantic Entropy to Medical VLMs

➢ Limitation of semantic entropy in VLMs

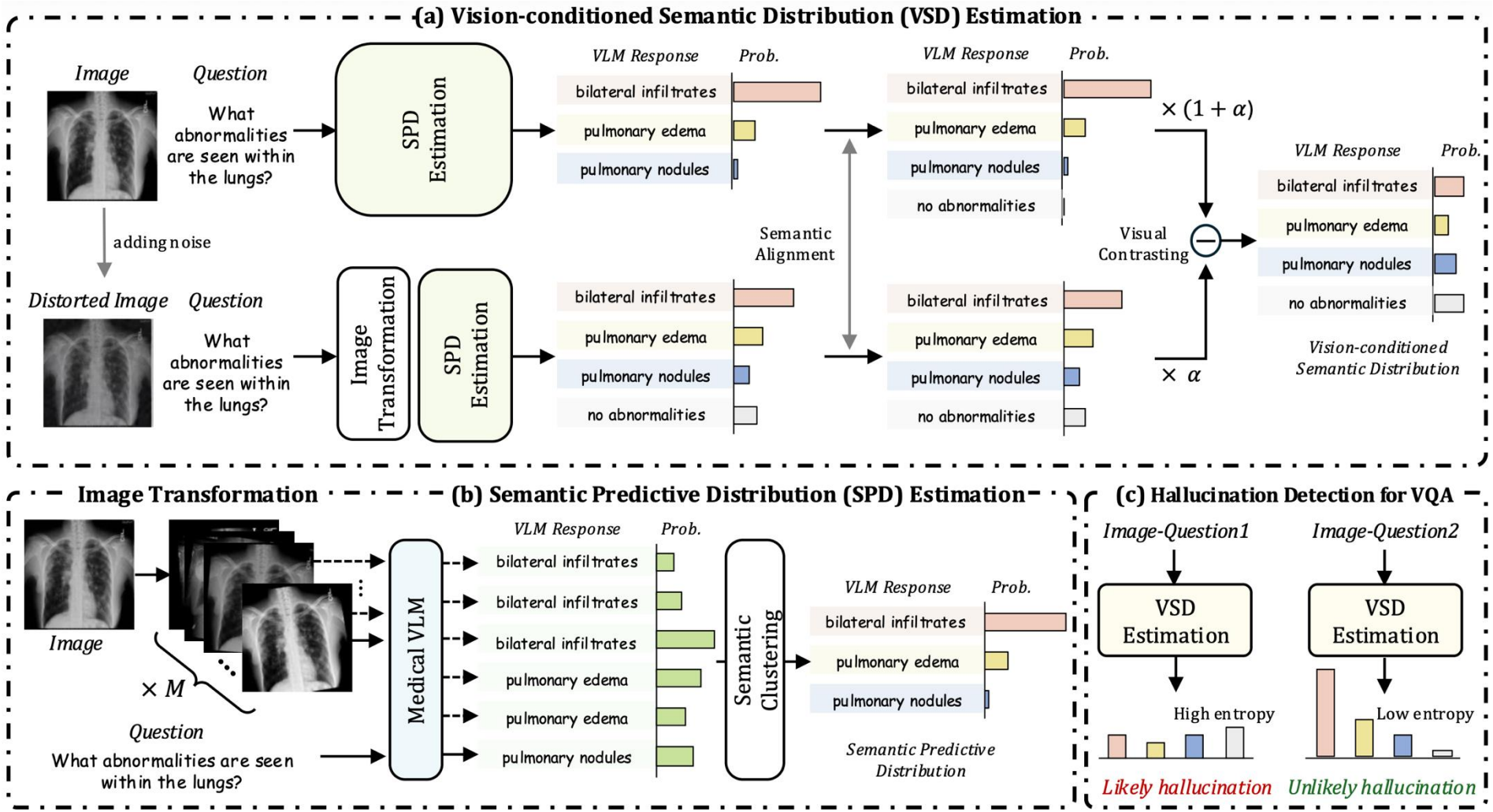
A natural adaptation of SE to VLMs involves incorporating perturbations in visual input during entropy estimation.

- However, medical VLMs often overlook visual inputs and rely predominantly on textual information when generating responses.
- This modality preference may cause overconfidence in incorrect answers, resulting in inaccurate entropy estimation.



Our Proposal

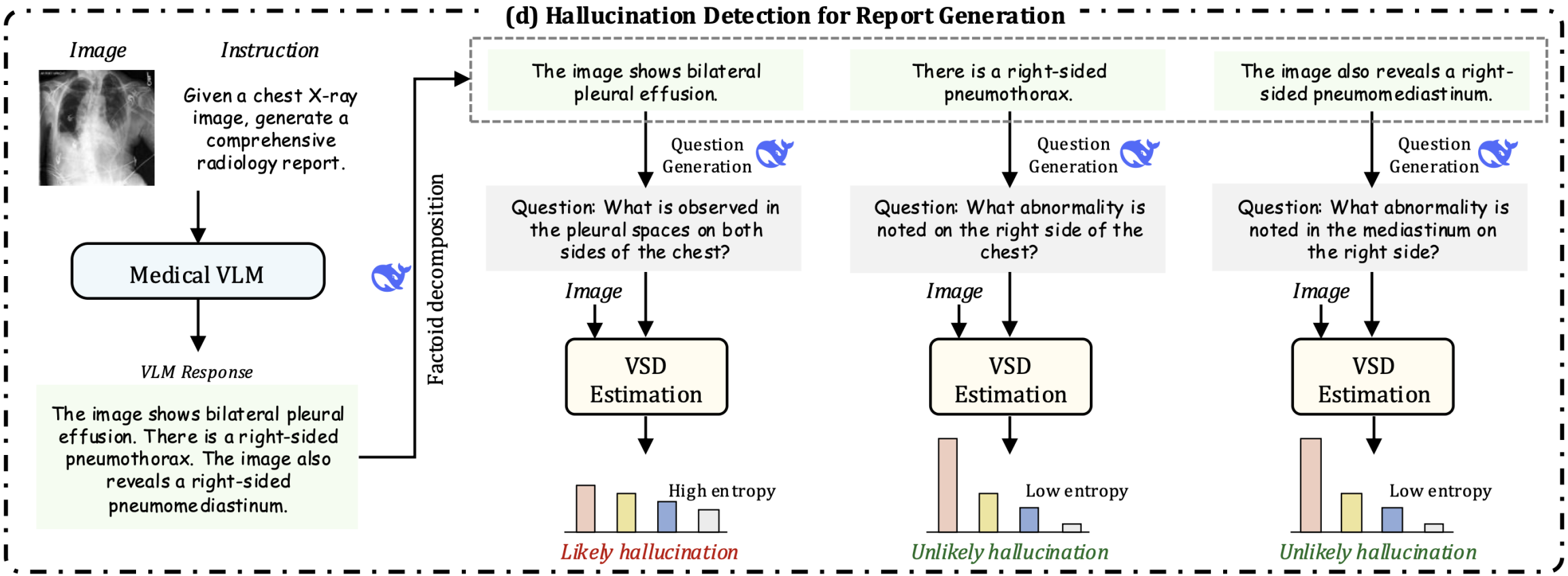
➢ Unified Vision-conditioned Response Semantic Entropy (UniVRSE)



Zehui Liao, et al., *TPAMI, Under Review*; *MICCAI, 2025*.

Our Proposal

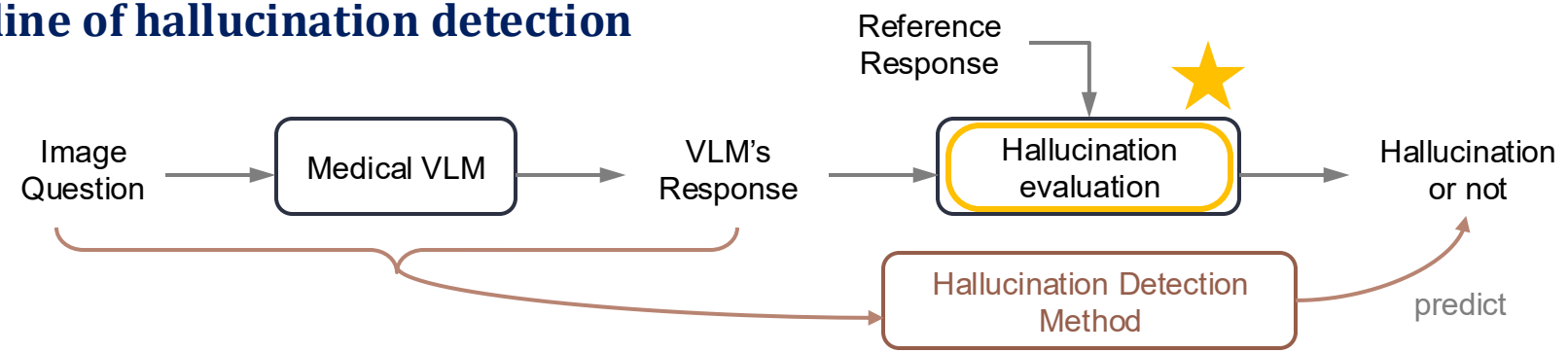
➢ Unified Vision-conditioned Response Semantic Entropy (UniVRSE)



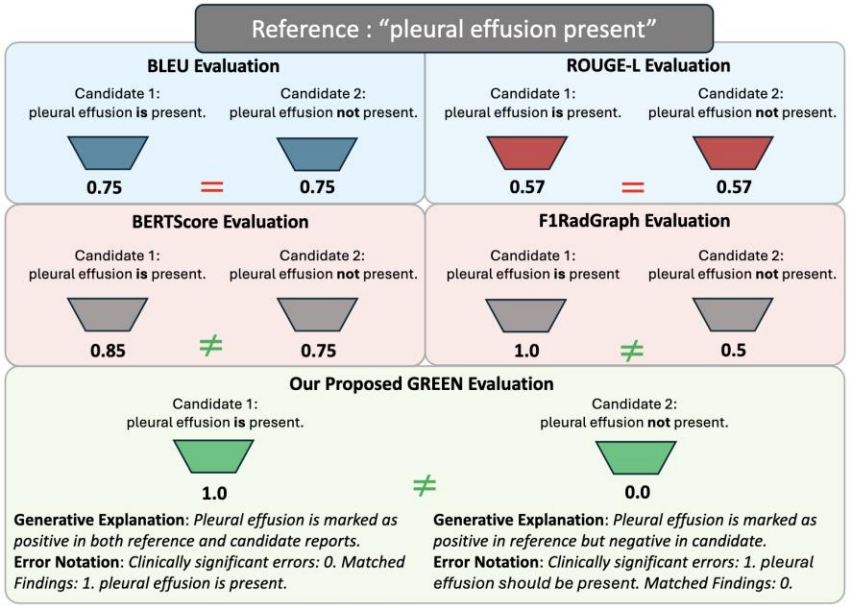
Zehui Liao, et al., *TPAMI, Under Review*; *MICCAI, 2025*.

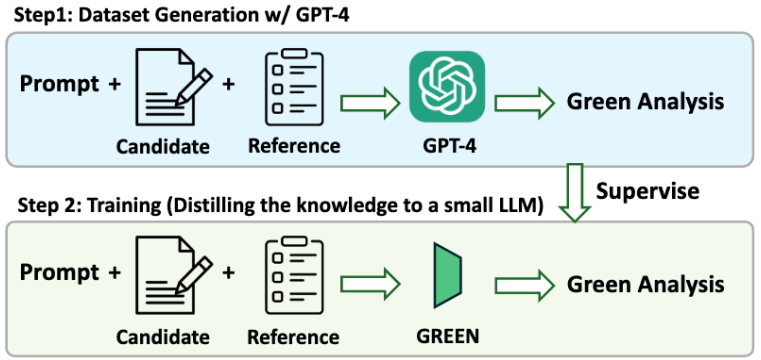
Experimental Setup

➢ Evaluation pipeline of hallucination detection



➢ Existing hallucination evaluation methods and their limitations





$$\text{GREEN} = \frac{\#\text{ matched findings}}{\#\text{ matched findings} + \sum_{i=(a)}^{(f)} \#\text{ error}_{\text{sig.},i}}$$

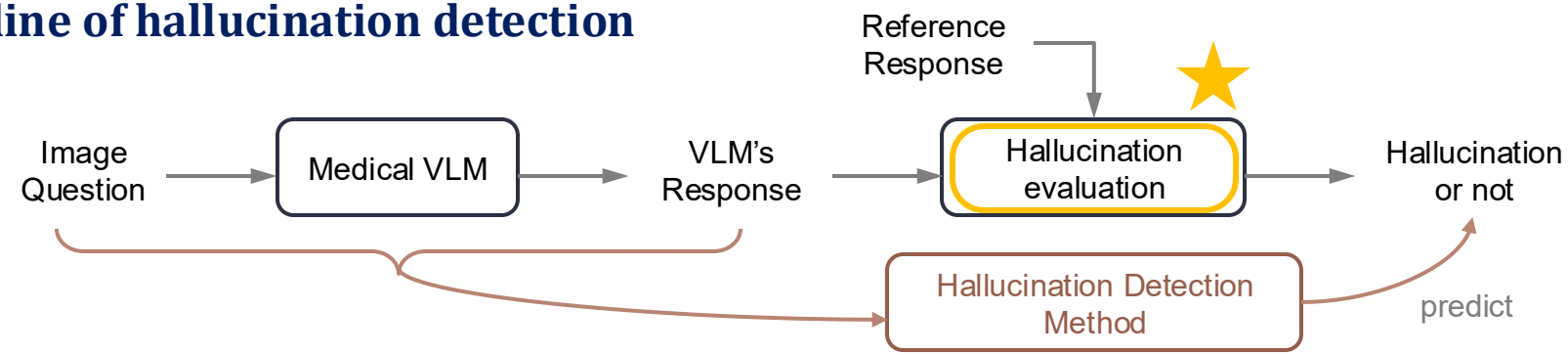
GREEN [1]

poor generalization across imaging modalities and clinical domains

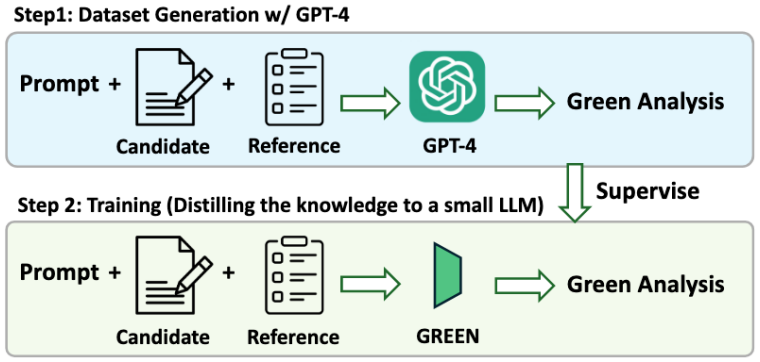
[1] Ostmeier, Sophie, et al. "GREEN: Generative Radiology Report Evaluation and Error Notation." ***EMNLP*** (**2024**).

Experimental Setup

➢ Evaluation pipeline of hallucination detection



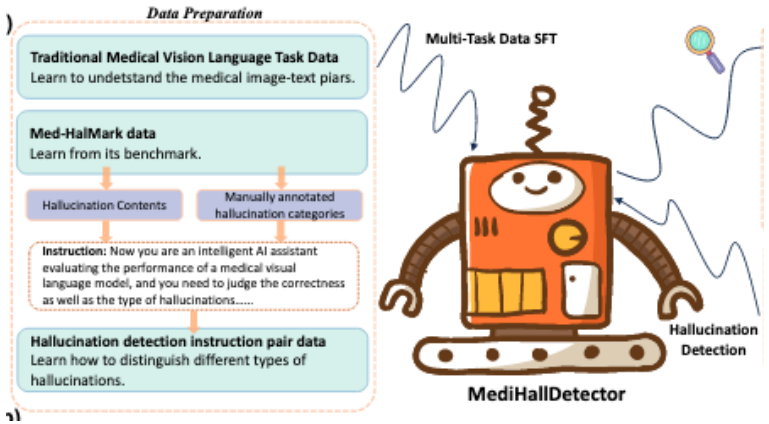
➢ Existing hallucination evaluation methods and their limitations



$$\text{GREEN} = \frac{\#\text{ matched findings}}{\#\text{ matched findings} + \sum_{i=(a)}^{(f)} \#\text{ error}_{\text{sig.},i}}$$

GREEN [1]

poor generalization across imaging modalities and clinical domains



- Catastrophic Hallucinations (Hc = 0.0),
- Critical Hallucinations (Hcr = 0.2),
- Attribute Hallucinations (Ha = 0.4),
- Prompt-induced Hallucinations (Hp = 0.6),
- Minor Hallucinations (Hm = 0.8), and
- Correct Statements (Hs = 1.0)

MediHallDoctor [2]

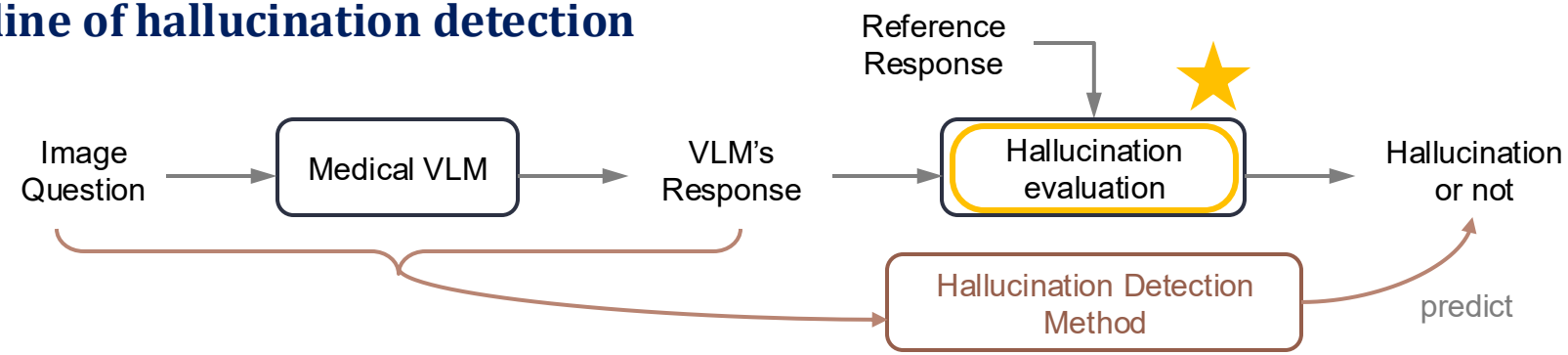
ambiguous or subjective evaluation principles

[1] Ostmeier, Sophie, et al. "GREEN: Generative Radiology Report Evaluation and Error Notation." ***EMNLP*** (**2024**).
[2] Chen, Jiawei, et al. "Detecting and evaluating medical hallucinations in large vision language models." *arXiv preprint arXiv:2406.10185* (**2024**).

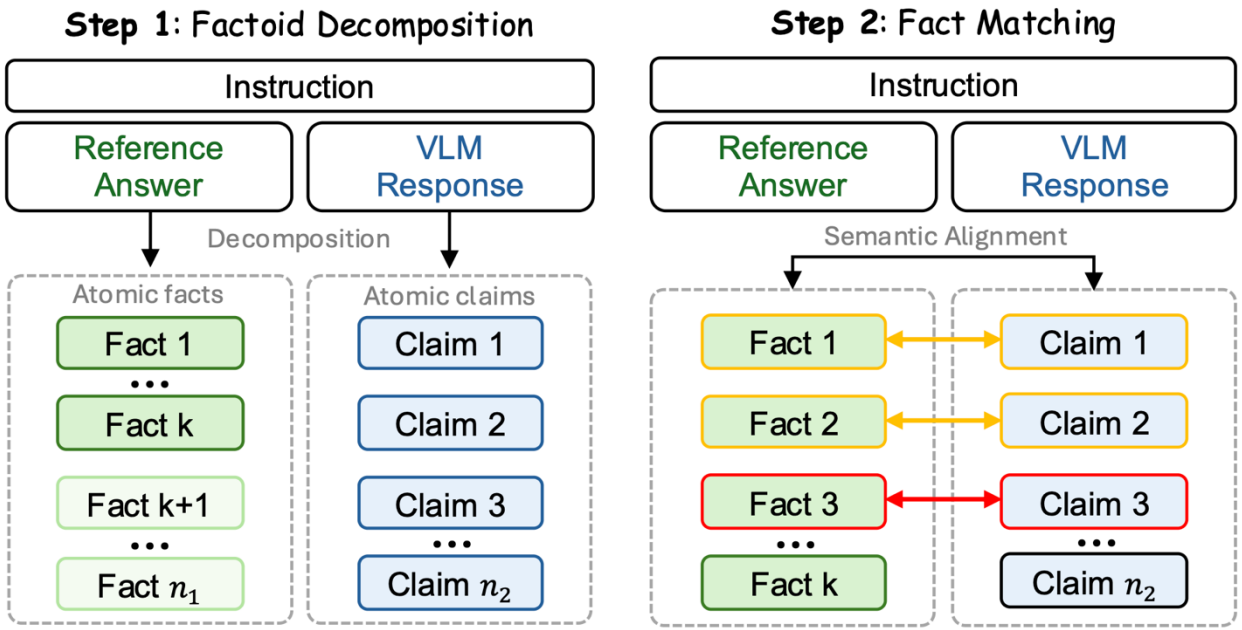
Experimental Setup

Limitation of existing methods

➢ Evaluation pipeline of hallucination detection



➢ Our ALFA, ALignment ratio of atomic Facts, a fine-grained hallucination evaluation method



Instruction-answering fact. Contextual fact.

Matched claim. Number of matched claims=m.

Hallucinated claim. Number of hallucinated claims=h.

Extraneous claim. Number of extraneous claims=e.

Ratio of matched claims $\alpha_m = \frac{m}{n_2}$

Ratio of hallucinated claims $\alpha_h = \frac{h}{n_2}$

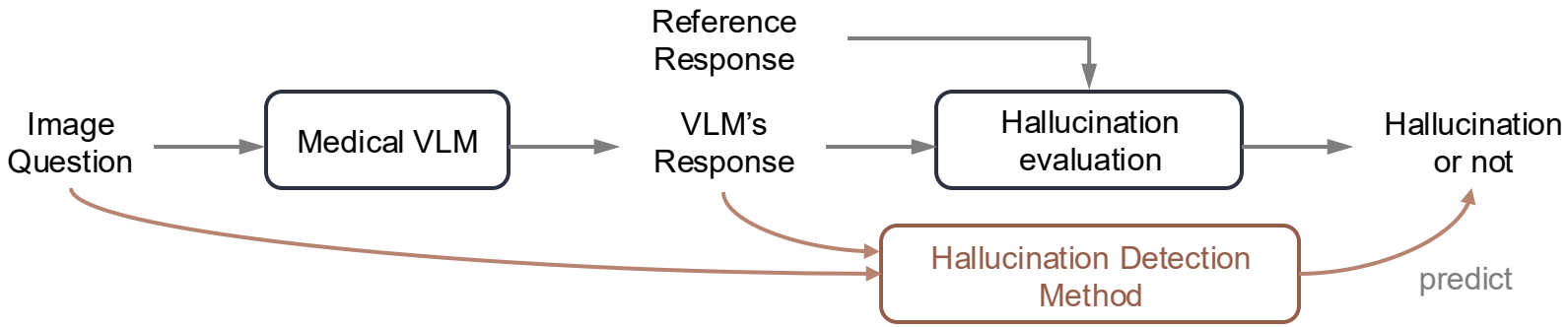
Ratio of extraneous claims $\alpha_e = \frac{e}{n_2}$

α_h serves as the hallucination label.

Zehui Liao, et al., *TPAMI, Under Review*; *MICCAI, 2025*.

Experimental Setup

➢ Evaluation pipeline of hallucination detection



➢ Evaluation metrics of hallucination detection

- **AUC**: the probability that a randomly chosen correct answer has been assigned a higher confidence / lower uncertainty score than a randomly chosen hallucinated answer.
- **Area Under ALFA Curve (AUA)**:
  - ① 'Mean α_h score at X%': mean α_h score of the model on the most-confident X% of samples identified by the respective uncertainty method.
  - ② To summarize the 'Mean α_h score at X%' from 1%-100% (interval=1%), we compute the AUA -- the total area enclosed by the mean α_h score at all cut-off percentage X%.

Experimental Results

➢ Comparison results and ablation studies

Performance of ours UniVRSE and other baselines on Medical VQA datasets.

Method	RAD-VQA		SLAKE		Path-VQA		MIMIC-Diff-VQA	
	AUC ↑	AUA ↓	AUC ↑	AUA ↓	AUC ↑	AUA ↓	AUC ↑	AUA ↓
MedGemma-4B-it [26]								
AvgProb	40.87	40.98	43.29	24.04	48.21	40.23	49.04	37.31
AvgEnt	59.21	26.94	56.58	25.86	52.40	39.65	50.89	40.35
MaxProb	41.06	38.16	42.35	26.36	48.44	39.87	49.27	**37.16**
MaxEnt	58.83	27.05	57.30	22.76	51.66	40.59	50.39	40.72
Cross-Checking	64.64	24.75	65.50	18.17	54.01	36.86	50.44	41.89
RadFlag	70.15	23.00	67.43	**16.85**	56.48	35.11	**52.51**	40.60
SE	71.87	24.79	67.93	17.44	57.19	34.88	52.42	40.76
UniVRSE	**76.25**	**18.22**	**69.56**	17.67	**59.17**	**32.98**	52.04	40.42
LlavaMed-7B [27]								
AvgProb	43.57	55.94	49.97	52.42	45.48	64.45	41.10	70.67
AvgEnt	54.76	50.08	50.78	51.07	55.59	55.94	48.68	63.78
MaxProb	41.71	55.18	49.75	51.91	48.94	59.97	40.71	69.76
MaxEnt	56.02	51.08	52.18	53.41	53.01	60.02	49.20	64.51
Cross-Checking	59.32	50.00	54.59	49.57	61.89	52.16	54.82	64.41
RadFlag	69.46	40.72	61.00	45.64	61.56	52.77	56.10	62.42
SE	72.90	40.54	66.75	43.79	63.73	50.60	56.31	62.34
UniVRSE	**74.31**	**39.02**	**68.16**	**42.07**	**66.18**	**48.72**	**58.13**	**61.05**
HuatuoGPT-Vision-7B [28]								
AvgProb	35.28	48.00	49.58	42.66	45.99	62.08	46.89	65.04
AvgEnt	66.73	29.91	47.64	44.96	54.26	56.28	54.90	66.66
MaxProb	34.20	47.13	47.35	44.28	46.14	61.45	46.61	64.17
MaxEnt	69.01	29.28	52.07	43.92	55.42	56.47	56.05	66.32
Cross-Checking	66.44	33.70	63.64	33.59	56.98	53.71	58.54	60.94
RadFlag	78.44	23.19	69.28	30.29	57.81	54.04	58.11	60.23
SE	79.02	23.52	69.60	30.42	60.78	52.96	58.34	60.51
UniVRSE	**81.17**	**22.05**	**70.89**	**29.06**	**62.89**	**51.18**	**60.71**	**59.19**

Performance of ours UniVRSE and other baselines on Medical VRG datasets.

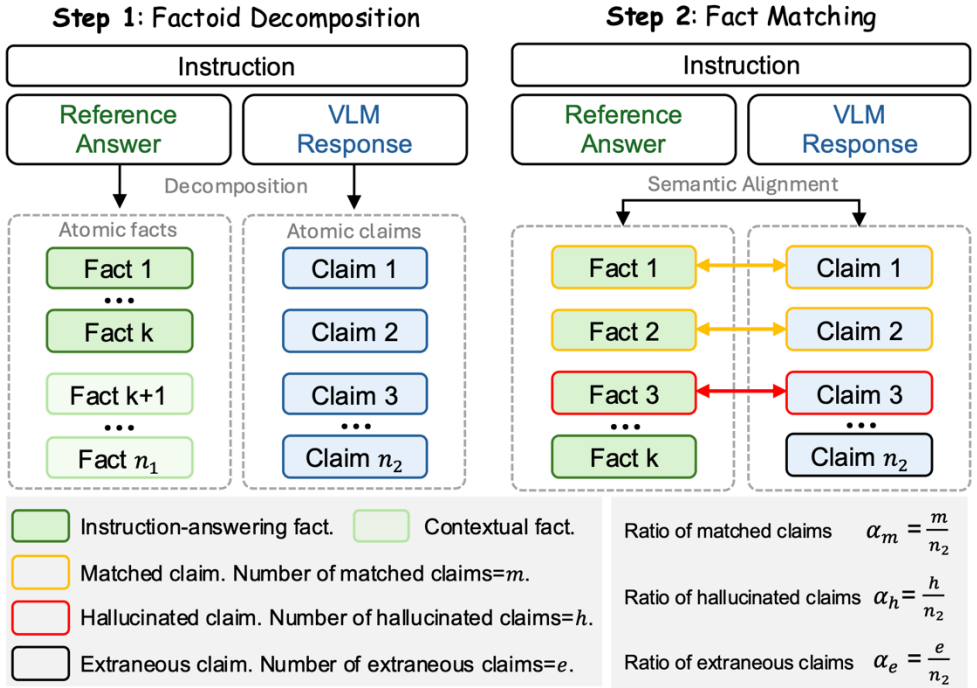
Method	CheXpertPlus		IU-Xray	
	AUC ↑	AUA ↓	AUC ↑	AUA ↓
MedGemma-4B-it [26]				
AvgProb	45.73	53.40	45.87	19.39
AvgEnt	54.16	47.10	53.97	17.11
MaxProb	45.30	55.10	44.96	21.16
MaxEnt	54.49	46.86	54.80	17.22
Cross-Checking	56.16	46.54	56.47	17.77
RadFlag	56.53	**44.53**	56.12	16.09
SE	56.79	44.61	57.78	15.07
UniVRSE	**58.96**	44.71	**63.06**	**12.38**
LlavaMed-7B [27]				
AvgProb	47.29	95.38	50.06	94.52
AvgEnt	55.32	94.27	45.33	94.44
MaxProb	48.31	95.41	43.81	95.30
MaxEnt	56.26	94.13	58.11	93.83
Cross-Checking	53.91	94.07	72.40	90.32
RadFlag	55.70	94.13	69.65	90.44
SE	60.19	93.97	75.70	90.14
UniVRSE	**62.52**	**93.08**	**78.06**	**89.03**

Performance of ours UniVRSE and its variants.

UniVRSE		AUC ↑	AUA ↓
Image Transformation	Visual Contrasting		
×	×	71.87	24.79
√	×	71.94	24.16
×	√	74.37	19.70
√	√	**76.25**	**18.22**

Experimental Results

➢ Analysis of ALFA eval.



Accuracy of GREEN and ALFA evaluation on subsets of Rad-VQA and Path-VQA datasets.

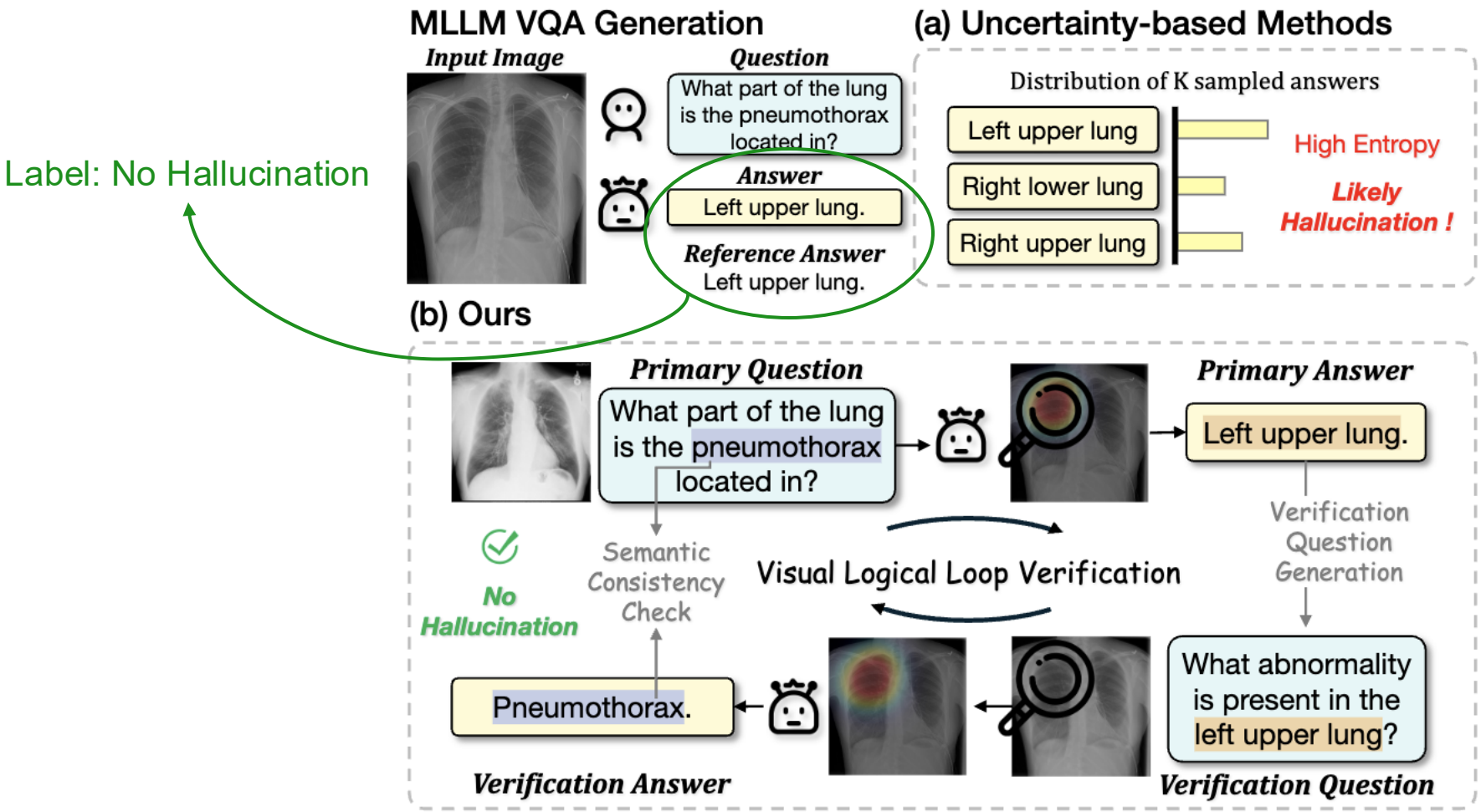
Dataset	RAD-VQA	Path-VQA
Modality	Radiology	Pathology
Acc. of Green Eval	93	76
Acc. of ALFA Eval	99	93

ALFA scores of three medical VLMs across four VQA and two VRG datasets.

VLMs	$\alpha_m \uparrow$	$\alpha_h \downarrow$	$\alpha_e \downarrow$
RAD-VQA			
MedGemma-4b-it	45.56	35.26	17.68
LLavamed-7B	27.81	52.64	18.05
HuatuoGPT-Vision-7B	**52.50**	**32.98**	**14.52**
SLAKE			
MedGemma-4b-it	**54.34**	**24.52**	21.14
LLavamed-7B	25.09	52.90	21.74
HuatuoGPT-Vision-7B	40.11	45.01	**14.76**
Path-VQA			
MedGemma-4b-it	**9.19**	**40.08**	49.95
LLavamed-7B	5.63	60.72	31.94
HuatuoGPT-Vision-7B	7.65	60.69	**31.62**
MIMIC-Diff-VQA			
MedGemma-4b-it	**18.69**	**38.27**	40.32
LLavamed-7B	8.81	67.88	**15.73**
HuatuoGPT-Vision-7B	12.92	65.46	17.38
ChexpertPlus			
MedGemma-4b-it	**25.23**	**25.39**	49.63
LLavamed-7B	3.97	75.16	**20.87**
HuatuoGPT-Vision-7B	23.94	30.13	51.41
IU-Xray			
MedGemma-4b-it	**46.26**	**9.55**	44.23
LLavamed-7B	2.64	71.57	**25.76**
HuatuoGPT-Vision-7B	34.48	21.28	50.43

Motivation

➢ Limitation of uncertainty-based methods



1. Uncertainty-based methods assess the predictive uncertainty of a VLM for a given image-question pair, rather than evaluating the correctness of the generated answer.

2. These methods necessitate additional inference procedures, typically K or 2K.

Conclusion

- We propose **UniVRSE**, a unified and model-agnostic framework for hallucination detection in medical VLMs that explicitly enhances visual guidance in semantic predictive uncertainty estimation.
- We propose **V-Loop**, a training-free, plug-and-play hallucination detection framework that verifies the factual correctness of medical VQA responses via visual logical loop verification.
- We introduce **ALFA**, a fine-grained and objective metric that evaluates factual consistency and enables automatic hallucination labeling across VQA and VRG tasks.

Paper

Code

Email: merrical@mail.nwpu.edu.cn

Thanks for your attention!