Mitigating Manifold Departure: Uncertainty-Aware Subspace Rectification for Trustworthy MLLM Decoding

Yingxuan Zhuang, Jingxiao Yang, Miao Pan, Cheng Tan, Yuxiang Cai, Siwei Tan, Chen Zhi, Xuhong Zhang, Jianwei Yin, Jintao Chen

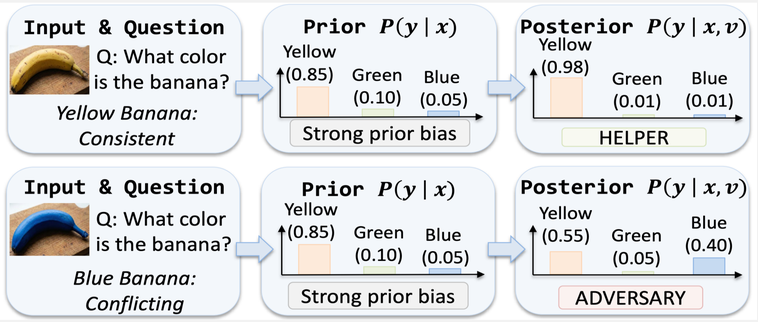
chenjintao@zju.edu.cn

ICML International Conference On Machine Learning

Introduction

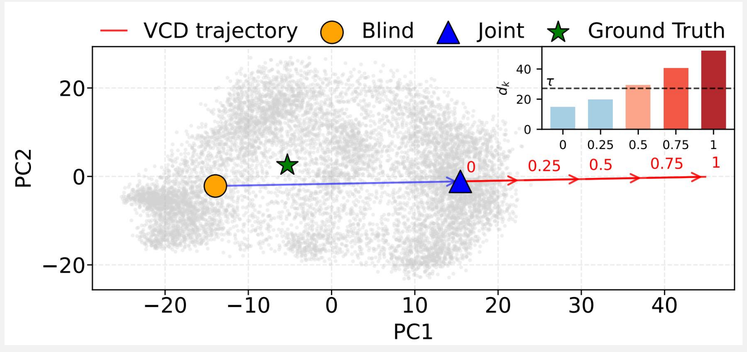
Background and Objective. MLLMs often create object hallucinations due to over-reliance on language prior overriding weak or ambiguous visual evidence.

The Dual Role of Priors.



Priors serve as anchors (yellow banana) and induce hallucinations (blue banana).

Core Diagnosis: Manifold Departure

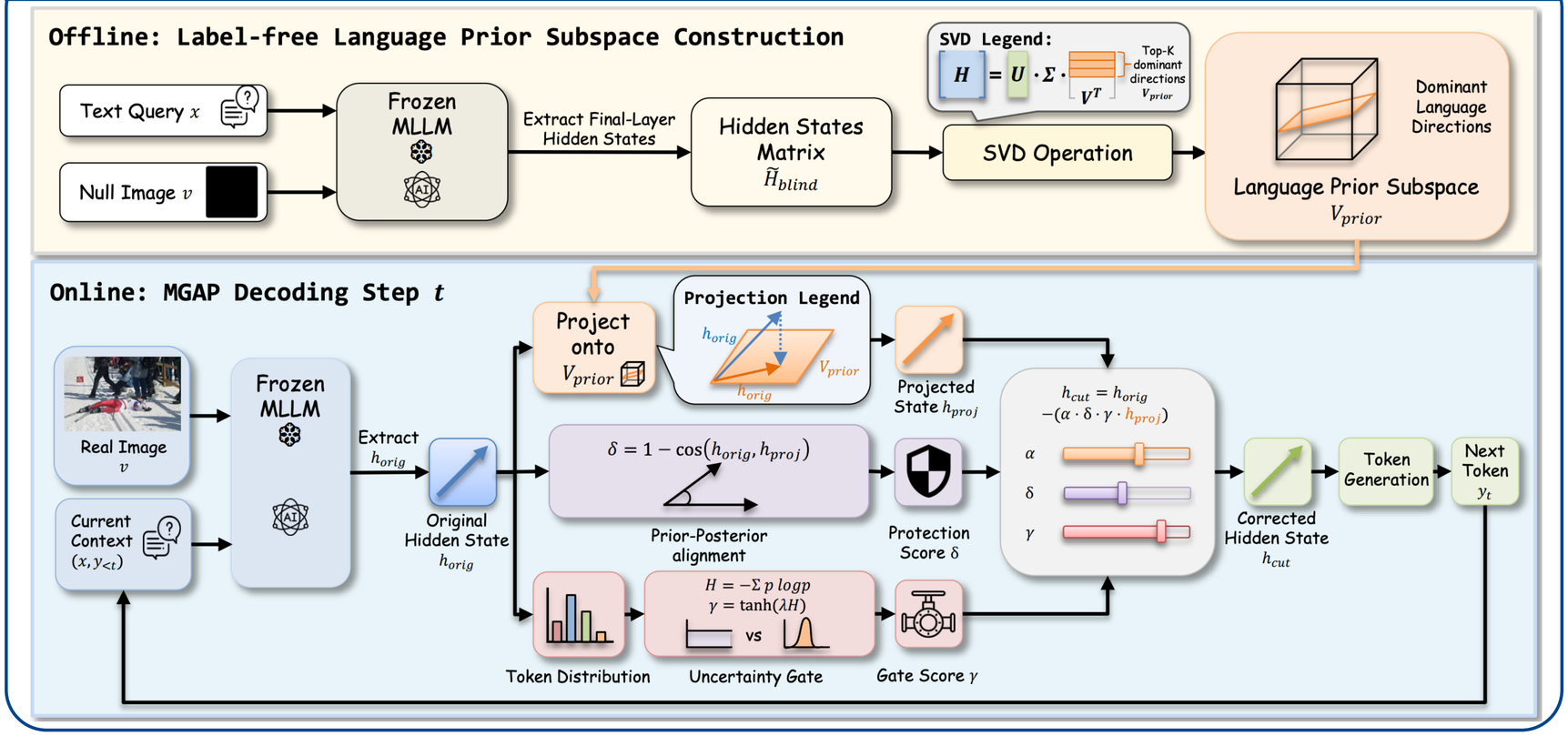


Structure-agnostic linear prior suppression push hidden state towards low-density regions, disrupting valid semantic configurations.

Contribution

- We identify Manifold Departure as a geometric failure mode of linear, training-free prior suppression, explaining why such methods can degrade performance even when priors align with vision.
- We propose MGAP, which learns a language-prior subspace from blind states and applies adaptive, subspace selective projection to suppress hallucinations without disturbing orthogonal semantics.
- Experiments on POPE and CHAIR show that MGAP achieves stronger trade-offs between hallucination mitigation and descriptive fidelity than prior training-free decoding baselines.

Method



Experiments

POPE Evaluation. Main results on the POPE benchmark. MGAP consistently outperforms training-free decoding methods.

Backbone	Method	Random			Popular			Adversarial		
		Acc.	Prec.	F1	Acc.	Prec.	F1	Acc.	Prec.	F1
LLaVA v1.5 7B	Vanilla	88.88	89.70	89.76	86.23	83.28	86.80	80.16	74.93	82.05
	VCD	87.57	86.23	87.79	84.23	80.76	85.08	78.56	73.47	80.66
	ICD	89.47	92.84	89.04	87.50	89.04	87.24	82.70	81.10	83.13
	HalTrapper	88.67	88.77	88.65	85.40	83.31	85.84	80.30	76.02	81.80
	DeCo	89.86	92.41	89.58	87.72	89.36	87.31	83.18	82.47	83.71
	MoD	89.24	91.37	89.17	87.03	87.96	86.91	82.51	80.74	83.02
	Ours	**90.63**	**93.69**	**90.29**	**88.10**	**91.50**	**87.59**	**84.59**	**85.06**	**84.46**
Qwen3-VL 8B	Vanilla	91.53	89.08	91.79	84.20	78.28	85.70	80.00	73.20	82.56
	VCD	90.67	88.80	90.83	83.27	79.41	84.53	81.36	74.92	83.60
	ICD	91.74	91.21	91.75	84.80	80.63	85.90	82.52	76.53	83.89
	HalTrapper	91.60	92.51	91.93	85.94	82.20	86.06	82.14	78.45	82.57
	DeCo	**91.96**	92.36	**91.98**	86.01	83.02	83.81	82.74	79.18	83.81
	MoD	91.87	91.94	91.86	85.76	82.71	83.74	82.61	78.92	83.74
	Ours	91.83	**93.67**	91.94	**86.40**	**84.13**	**86.84**	**83.13**	**79.30**	**84.17**

CHAIR Evaluation. MGAP significantly reduces hallucination while maintaining competitive precision and F1.

Method	LLaVA v1.5 7B				Qwen3-VL 8B			
	CHAIRs↓	CHAIRi↓	Prec.	F1	CHAIRs↓	CHAIRi↓	Prec.	F1
Vanilla	47.4	23.5	70.8	65.0	33.6	8.7	81.5	67.7
HalTrapper	47.2	23.5	70.6	64.9	30.8	8.8	81.6	62.3
VCD	52.8	15.8	72.6	76.5	38.4	11.7	79.5	**68.8**
ICD	51.8	14.7	73.7	77.4	34.2	8.4	81.4	68.5
CODE	49.8	13.8	76.0	75.7	31.4	8.7	81.6	64.0
Ours	**26.2**	**7.6**	**85.9**	**77.4**	**26.8**	**8.1**	**83.9**	66.2

The results demonstrate that MGAP generalizes beyond binary question answering to open-ended captioning tasks.

Inference Efficiency Comparison. MGAP is significantly more efficient than CD baselines.

Method	Total Time	Sec / Sample
Vanilla	23:57	0.48
VCD	1:46:54	2.14
ICD	1:45:18	2.11
HalTrapper	1:50:03	2.20
Ours	**52:24**	**1.05**

AMBER Evaluation. MGAP consistently outperforms well.

Model	CHAIR↓	Cover↑	Hal↓	Cog↓
LLaVA v1.5 7B	11.2	50.2	47.9	4.6
+ VCD	8.9	51.2	38.1	4.4
+ ICD	8.6	51.1	37.3	3.9
+ CODE	9.0	51.1	39.5	4.3
+ HalTrapper	8.0	51.5	36.3	**3.8**
+ MGAP (Ours)	**7.6**	**51.7**	**35.1**	**3.8**

SCAN FOR MORE !

arXiv

OmniAI Our Team