Mitigating Hallucinations in Vision-Language Models using Depth and Spatial-aware Key-Value Cache Refinement

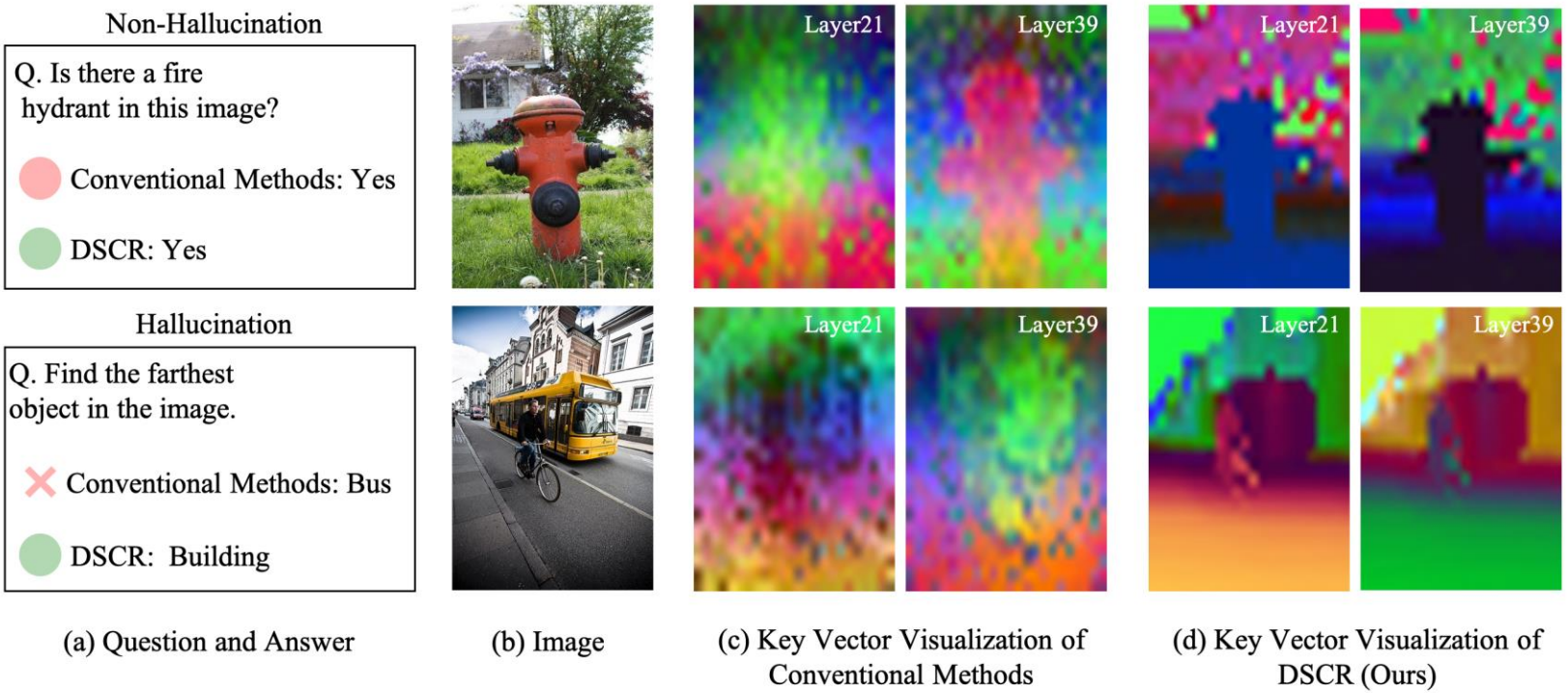
Gusang Lee

Outline

- Problem Statement
- Proposed Method
- Experiment Results
- Discussion & Conclusion

Problem Statement

Problems of VLM

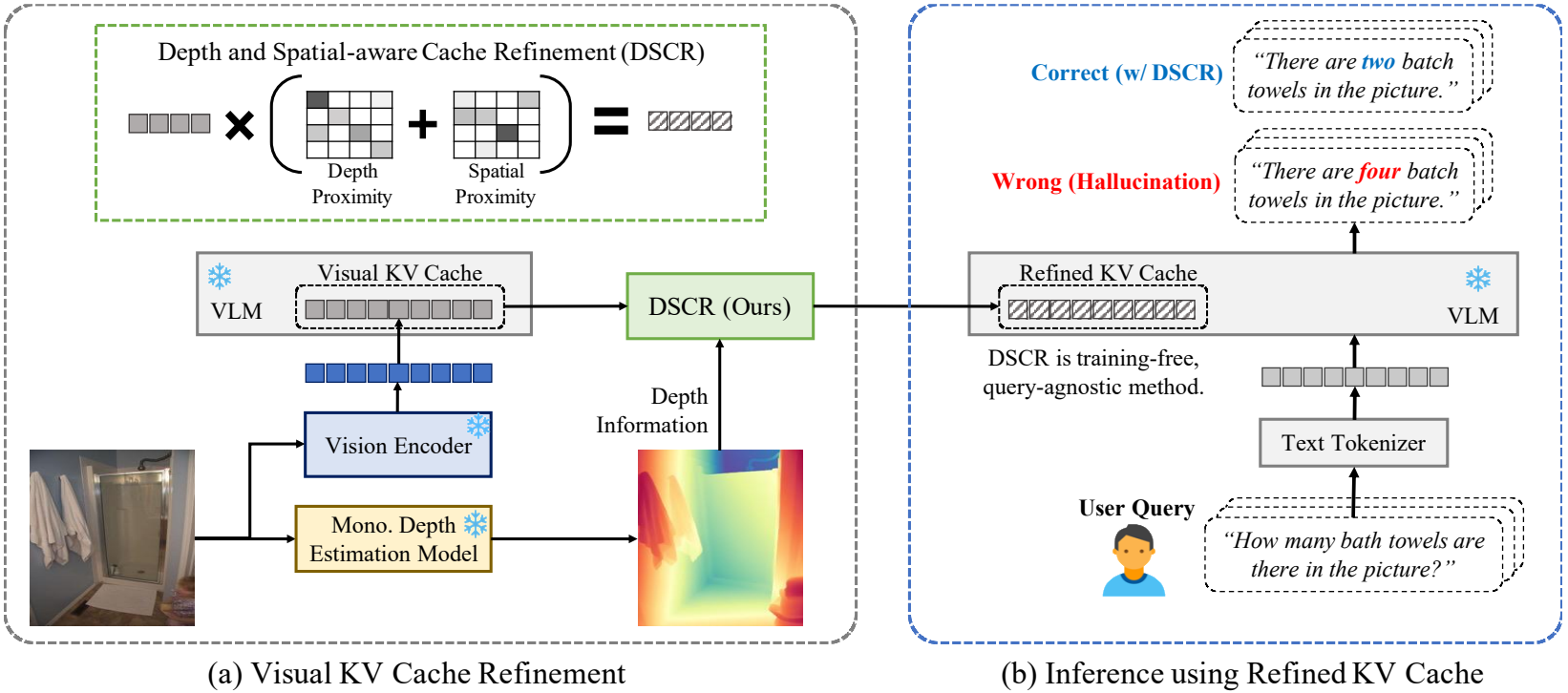


(a) Question and Answer (b) Image (c) Key Vector Visualization of Conventional Methods (d) Key Vector Visualization of DSCR (Ours)

- The figure below shows that hallucination is not merely an output error, but stems from the collapse of internal visual representations
- In non hallucinated cases, the key vectors of adjacent patches remain consistently similar and become clearer in deeper layers, whereas under hallucination this structure becomes nearly isotropic, blurs object boundaries, and weakens correct visual grounding.

Proposed Method

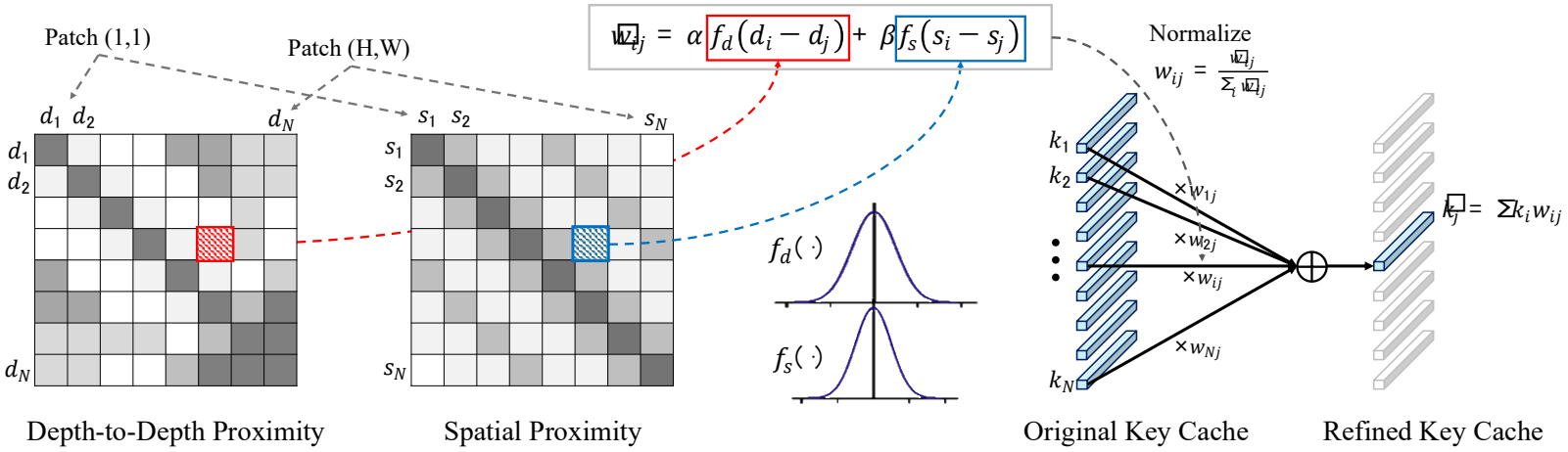
System Overview



(a) Visual KV Cache Refinement (b) Inference using Refined KV Cache

- DSCR establishes a strong association between relevant visual tokens in the attention blocks, mitigating hallucinations in the VLMs.
- The VLMs with DSCR produce the accurate answer (blue) unlike the original VLMS (red). Note that DSCR is training-free, model-invariant, and query-agnostic.

Proposed Method DSCR



- Depth-to-depth and spatial proximity maps show the difference in depth values and distances between image patch pairs, with darker shades indicating smaller differences.
- Using the importance weights derived from the proximity scores, refined cache entries are computed as a weighted sum of the original ones.
- The same process is applied to Value cache refinement.

Experimental Results

Result of MME Dataset

Model	Metric	Baseline	VCD	OPERA	HALC	DAMO	AGLA	DSCR (Ours)	VCD +DSCR	OPERA +DSCR	HALC +DSCR	DAMO +DSCR	AGLA +DSCR
LLaVA-1.5	Existence	190.00	190.00	**195.00**	190.00	190.00	190.00	**195.00**	**195.00**	**195.00**	**195.00**	190.00	190.00
	Count	143.33	143.33	**160.00**	153.33	148.33	135.00	**160.00**	143.33	**160.00**	**160.00**	155.00	155.00
	Position	**120.00**	**120.00**	**120.00**	**120.00**	105.00	**120.00**	**120.00**	115.00	**120.00**	**120.00**	110.00	**120.00**
	Color	165.00	165.00	165.00	170.00	160.00	165.00	**175.00**	170.00	**175.00**	**175.00**	165.00	**175.00**
	OCR	117.50	117.50	117.50	125.00	117.50	117.50	**140.00**	110.00	**140.00**	**140.00**	117.50	**140.00**
	Posters	156.85	156.85	157.19	141.10	151.37	**162.67**	135.96	155.14	135.37	135.96	148.97	151.37
	Total	892.68	892.68	914.69	899.43	872.20	890.17	925.96	888.47	925.27	925.96	886.47	**931.37**
LLaVA-1.6	Existence	**180.00**	**180.00**	**180.00**	175.00	**180.00**	175.00	**180.00**	**180.00**	170.00	**180.00**	**180.00**	170.00
	Count	153.33	138.33	**158.33**	141.67	150.00	146.67	156.67	136.67	135.00	156.67	155.00	143.33
	Position	93.33	93.33	98.33	101.67	96.67	91.67	105.00	110.00	105.00	105.00	101.67	**113.33**
	Color	**180.00**	**180.00**	**180.00**	155.00	175.00	175.00	175.00	175.00	145.00	175.00	155.00	150.00
	OCR	132.50	**147.50**	132.50	132.50	117.50	140.00	132.50	**147.50**	**147.50**	132.50	132.50	**147.50**
	Posters	150.34	147.60	150.34	142.47	151.37	**157.53**	152.40	148.63	132.88	152.38	142.47	144.52
	Total	889.51	886.77	899.51	848.30	870.54	885.87	**901.56**	897.80	835.38	**901.56**	866.63	868.69
Qwen-VL	Existence	170.00	165.00	170.00	170.00	160.00	170.00	175.00	165.00	165.00	**185.00**	153.33	175.00
	Count	145.00	145.00	150.00	145.00	150.00	145.00	**155.00**	145.00	150.00	145.00	**155.00**	145.00
	Position	98.33	98.33	98.33	98.33	108.33	93.33	103.33	98.33	101.67	108.33	**121.67**	96.67
	Color	180.00	180.00	180.00	180.00	**185.00**	180.00	**185.00**	180.00	180.00	**185.00**	180.00	175.00
	OCR	87.50	110.00	87.50	87.50	**110.00**	87.50	87.50	102.50	87.50	80.00	**110.00**	80.00
	Posters	144.18	146.23	144.52	144.18	152.05	146.58	166.78	148.92	160.62	**178.42**	145.55	154.79
	Total	825.01	844.57	830.35	825.01	865.39	822.41	872.61	839.81	844.78	**881.76**	865.55	826.46
Qwen2.5-VL	Existence	**195.00**	**195.00**	190.00	185.00	190.00	190.00	190.00	**195.00**	185.00	190.00	190.00	190.00
	Count	160.00	165.00	165.00	135.00	160.00	165.00	165.00	**170.00**	165.00	165.00	165.00	165.00
	Position	160.00	160.00	160.00	153.33	160.00	**165.00**	155.00	160.00	155.00	155.00	155.00	**165.00**
	Color	185.00	190.00	185.00	**195.00**	185.00	190.00	185.00	190.00	190.00	185.00	185.00	190.00
	OCR	**177.50**	155.00	170.00	155.00	170.00	162.50	**177.50**	162.50	**177.50**	**177.50**	**177.50**	162.50
	Posters	165.41	167.12	165.41	164.73	167.47	167.81	167.47	168.49	166.10	**169.52**	168.49	167.81
	Total	1042.91	1032.12	1035.41	988.06	1032.47	1040.31	1039.97	**1045.99**	1043.60	1042.02	1040.99	1040.31

- Overall, DSCR achieves the most consistent gains in total score across all models, outperforming the baseline and existing hallucination mitigation methods, with especially strong improvements on object focused categories such as existence, count, and position.

Result of Pope Dataset

Setting	Model	w/DSCR	POPE (GQA) Acc.	POPE (GQA) Prec.	POPE (GQA) Rec.	POPE (GQA) F1	RePOPE (MSCOCO) Acc.	RePOPE (MSCOCO) Prec.	RePOPE (MSCOCO) Rec.	RePOPE (MSCOCO) F1
Random	LLaVA-1.5	×	0.90	0.92	**0.87**	0.89	0.92	0.93	0.88	0.90
		✓	0.90	**0.93**	0.85	0.89	0.92	0.95	0.87	0.90
	Qwen-VL	×	0.84	0.84	**0.83**	0.84	0.92	0.92	0.88	0.90
		✓	**0.85**	**0.91**	0.77	0.84	91	**0.93**	0.87	0.90
	mPLUG-Owl2	×	**0.85**	0.91	**0.77**	**0.84**	0.63	0.62	0.87	0.70
		✓	0.84	**0.93**	0.75	0.83	**0.70**	**0.68**	**0.89**	**0.78**
Popular	LLaVA-1.5	×	0.86	0.84	**0.87**	0.86	0.91	0.92	0.86	0.89
		✓	**0.87**	**0.87**	0.85	0.86	**0.90**	**0.93**	**0.85**	**0.89**
	Qwen-VL	×	0.72	0.68	**0.83**	0.75	0.89	0.87	0.88	0.88
		✓	**0.80**	**0.81**	0.77	**0.79**	**0.88**	**0.88**	**0.87**	**0.88**
	mPLUG-Owl2	×	0.79	0.80	**0.77**	**0.79**	0.60	0.55	0.87	0.68
		✓	0.79	**0.82**	0.75	0.78	**0.66**	**0.60**	**0.90**	**0.74**
Adversarial	LLaVA-1.5	×	0.81	0.78	**0.87**	0.82	0.89	0.87	0.87	0.87
		✓	**0.82**	**0.80**	0.85	0.82	**0.89**	**0.89**	**0.85**	**0.87**
	Qwen-VL	×	0.75	0.72	0.77	0.77	0.89	0.87	0.88	0.88
		✓	**0.78**	**0.79**	0.77	**0.78**	**0.88**	**0.88**	**0.87**	**0.88**
	mPLUG-Owl2	×	0.77	0.76	**0.77**	0.77	0.58	0.53	0.85	0.66
		✓	0.77	**0.79**	0.75	0.77	**0.63**	**0.58**	**0.88**	**0.72**

- Across POPE and RePOPE, DSCR consistently improves F1 scores across different settings and models, demonstrating robust gains in hallucination mitigation and object level visual grounding.

Result of Another Dataset

- Evaluation results on the CHAIR dataset using LLaVA-1.5 model.

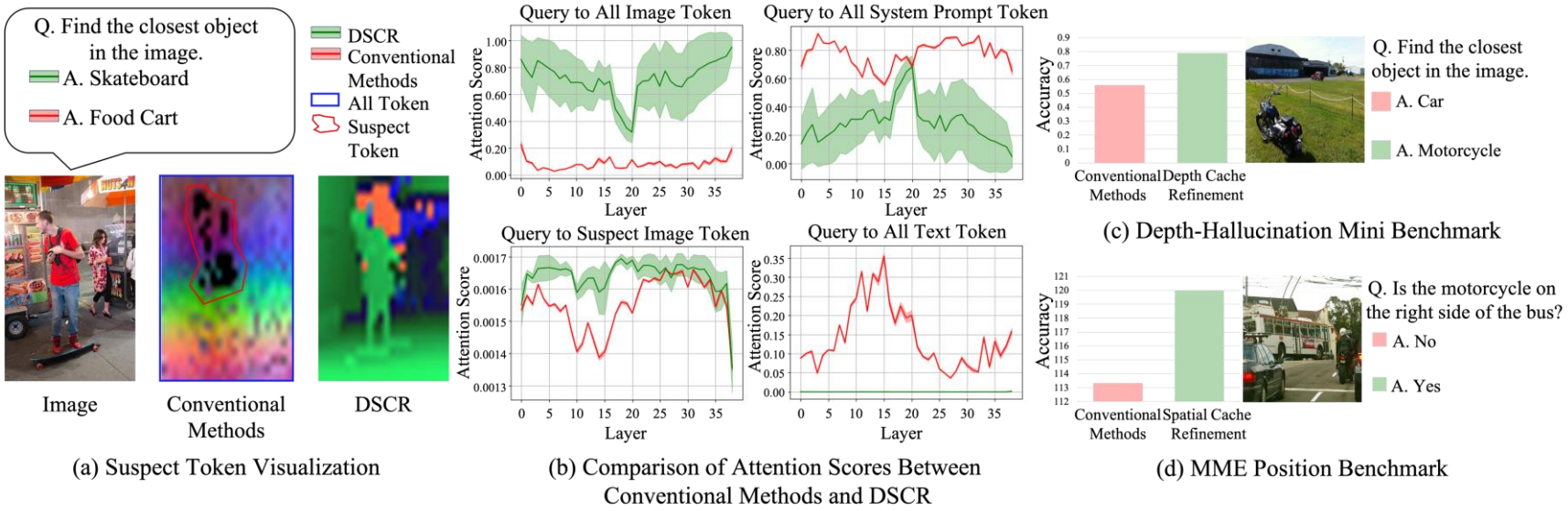
Method	$CHAIR_S \downarrow$	$CHAIR_I \downarrow$	Recall ↑	Avg. Len.
Baseline	48.2	12.5	78.2	99.4
VCD	54.4	14.4	**79.0**	101.6
OPERA	**39.2**	11.6	73.5	83.9
DSCR	**39.2**	**11.2**	73.4	96.1

- Evaluation results on the VQAv2 dataset using LLaVA-1.5 and Qwen2.5-VL.

Metric	LLaVa-1.5		Qwen2.5-VL	
	Baseline	DSCR	Baseline	DSCR
Overall Accuracy	79.87	**80.13**	84.27	**84.73**

- On CHAIR, DSCR achieves the lowest hallucination scores, matching or outperforming prior methods on $\mathrm{CHAIR_S}$ and $\mathrm{CHAIR_I}$ while maintaining competitive recall.
- On AMBER, DSCR also improves overall accuracy over the baseline for both LLaVA 1.5 and Qwen2.5 VL, showing that its benefits generalize across generative and discriminative hallucination benchmarks.

Analysis



- (a) Visualization of a hallucination case where DSCR reduces incorrect predictions by suppressing suspect tokens and enhancing the spatial structure of key vectors.
- (b) Layer-wise attention scores comparing Conventional Methods and DSCR across image, suspect, prompt, and text tokens.
- (c) Accuracy comparisons from the Hallucination-Depth Mini Benchmark, comparing the Conventional Methods with depth-refined KV cache.
- (d) Results on MME where spatial refinement is applied to KV caches.

Analysis

Absolute Depth Perception

Q. Find the closest object in the image.
Prompt : Please select the correct answer (only give the key such as 'a', 'b', 'c', 'd').
a. Glass
b. Salt shaker
c. Dessert sign
d. Phone

Q. Find the farthest object in the image.
Prompt : Please select the correct answer (only give the key such as 'a', 'b', 'c', 'd').
a. Hand
b. Glasses
c. Salt shaker
d. Phone

Perspective-Aware Size Perception

Q. Which object would be the largest in real life if placed at the same distance?
Prompt : Please select the correct answer (only give the key such as 'a', 'b', 'c', 'd').
a. Cup
b. Menu board
c. Phone
d. Salt shaker

Q. Which object would be the smallest in real life if placed at the same distance?
Prompt : Please select the correct answer (only give the key such as 'a', 'b', 'c', 'd').
a. Person
b. Menu board
c. Glass
d. Salt shaker

- The left shows the input image, and the right displays four sample questions grouped by reasoning type.
- Top : Questions assessing Absolute Depth Perception (e.g., identifying the closest or farthest object in the scene).
- Bottom: Questions assessing Perspective-Aware Size Perception, which test whether the model can infer real-world object size based on perspective.
- Correct answers are highlighted in green.

Ablation Study

Table 5: Inference time and GPU memory consumption per image for various methods.

Method	Time (sec/img)	GPU Mem. (MiB)
Baseline	9.35	29950.3
DSCR (Ours)	11.06	32813.7
VCD	15.13	29979.6
OPERA	39.37	37717.3
HALC	31.47	32890.7
DAMO	11.80	29965.2
AGLA	23.97	33448.1
VCD + DSCR	16.31	32841.1
OPERA + DSCR	42.47	40569.8
HALC + DSCR	34.53	36704.4
DAMO + DSCR	13.48	32828.6
AGLA + DSCR	25.08	36306.0

Table 6: COCO Image captioning performance with and without DSCR.

Method	BLEU-4	CIDEr	SPICE
Baseline	0.122	0.529	0.162
DSCR	**0.235**	**0.909**	**0.193**

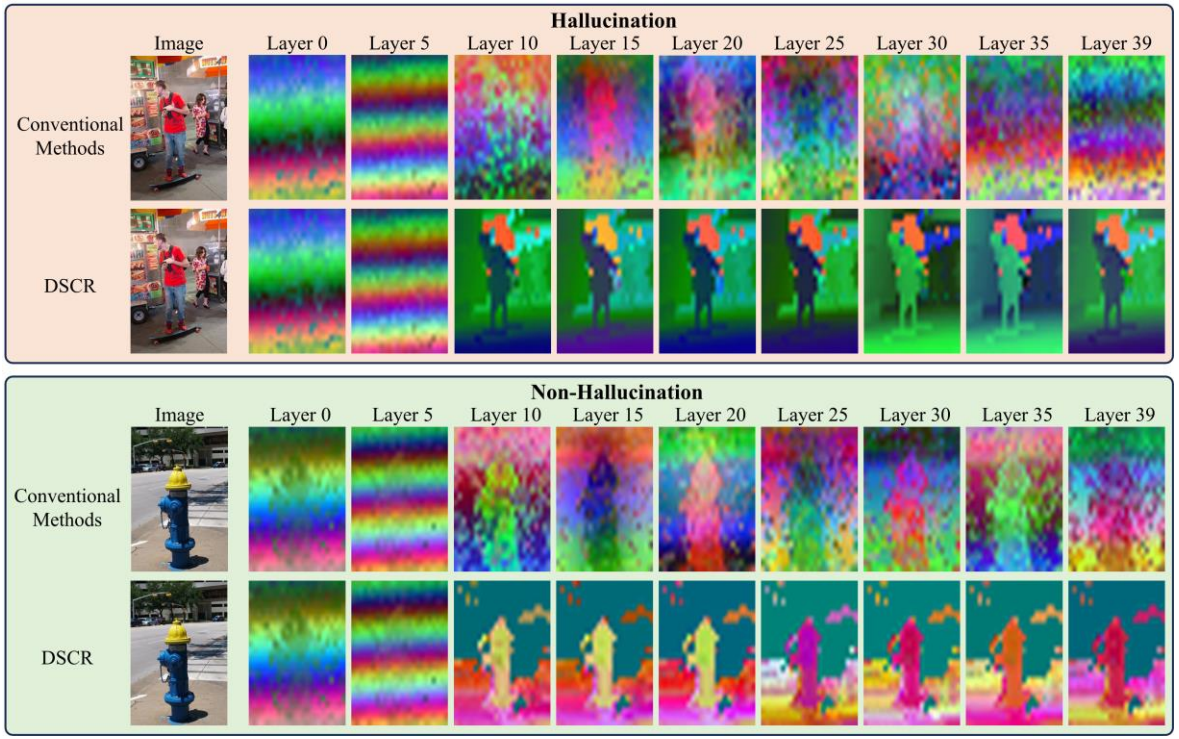
Table 7: Hyperparameter settings used across all models and datasets.

H.Params	σ_d	σ_s	α	β	Layers
Value	0.6	0.6	0.6	0.8	10–39

Table 8: GPU memory, per-image inference time, and performance comparisons of DSCR using different depth estimators.

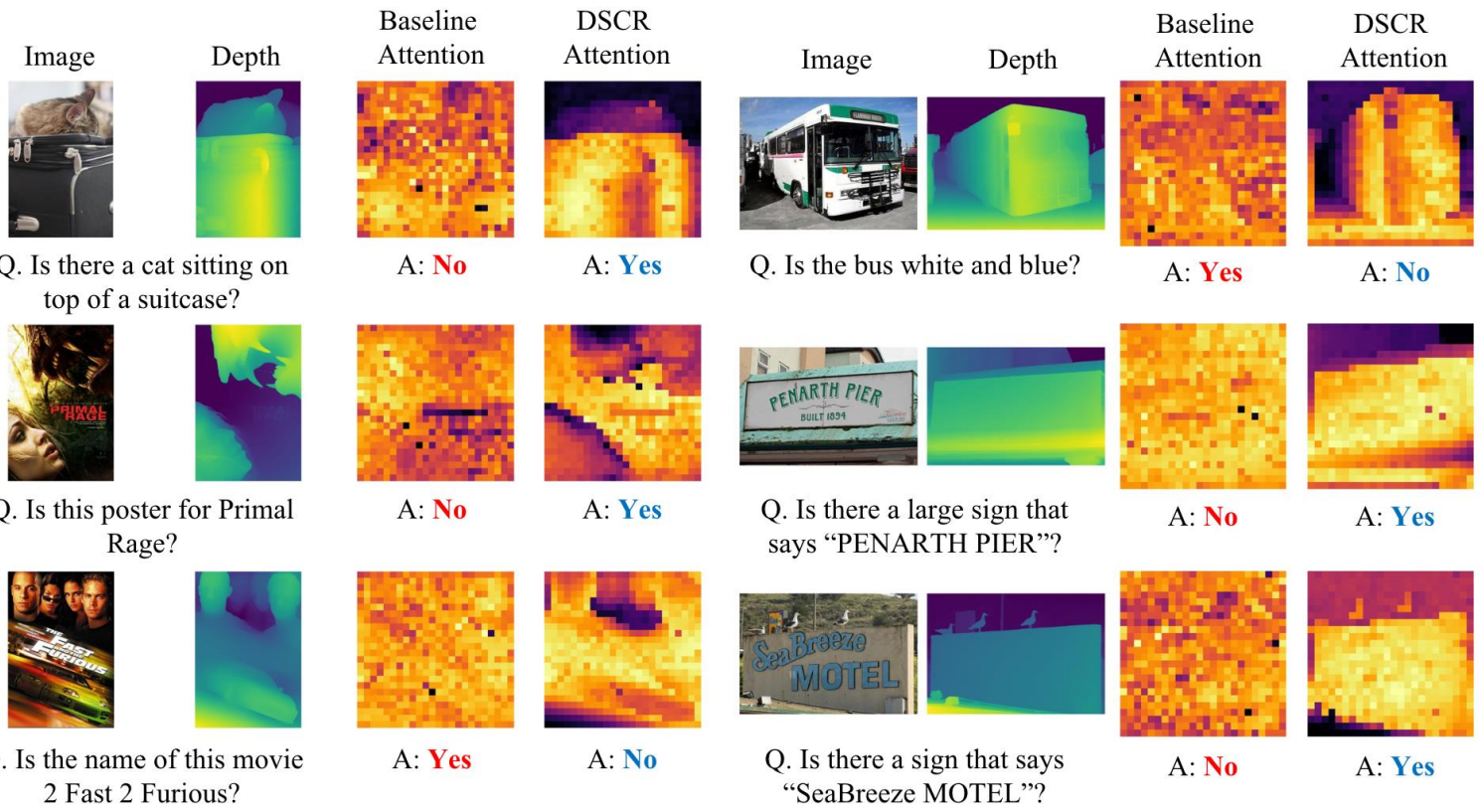
Depth Model	GPU (MiB)	Time (sec/img)	OCR	Color	Count	Existence	Position	Posters
Depth-Anything-v2 (Yang et al., 2024b)	2134.1	1.34	132.5	175.0	160.0	195.0	120.0	140.48
MiDaS-Lite (Ranftl et al., 2020)	1264.2	1.15	125.0	180.0	160.0	195.0	111.67	132.65
DPT-Lite (Ranftl et al., 2021)	526.1	1.54	125.0	180.0	160.0	195.0	111.67	132.65

PCA visualizations of key vectors across layers



- PCA visualizations of key vectors across layers for a hallucination example (top) and a non-hallucination example (bottom). In each block, the top row is the baseline model and the bottom row is DSCR.
- The visualization shows that DSCR produces more object-aligned patterns and clearer separation between foreground and background, especially in middle and upper layers.

Visual Question-Answering Example



- VQA examples, including image, depth, and query-to-image attention heatmaps before and after applying DSCR.

Conclusion

Conclusion

- We proposed DSCR, a novel and zero-shot method designed to mitigate the hallucination problem in VLMs.
- DSCR enhances the reliability of VLMs by refining the internal Key-Value (KV) caches of visual tokens based on their depth and spatial relationships, without requiring additional training, architectural modifications, or changes to the inference process.
- By leveraging a lightweight depth estimation model, DSCR effectively reduces both object-level and attribute-level hallucinations, as evidenced by substantial improvements on MME and POPE hallucination benchmarks.
- The simplicity and efficiency of DSCR make it a promising solution for enhancing the trustworthiness of VLMs, thereby facilitating their adoption in critical real-world applications where accuracy and reliability are paramount.

Thank You