Draft-and-Refine with Visual Experts

✨ Highlight

CVPR JUNE 3-7, 2026 DENVER COLORADO

Quantifiable Metrics for Visual Agent Reasoning

SungHeon Jeong · Ryozo Masukawa · Jihong Park · Sanggeon Yun · Wenjun Huang · Hanning Chen
Mahdi Imani · Mohsen Imani

Presenter

Who is speaking today?

Sungheon Jeong

Ph.D. Student

Research Focus

AI Agent | Multimodal Language Model | Hallucination

Affiliation

BIASLab @ UC Irvine | advisor Mohsen Imani

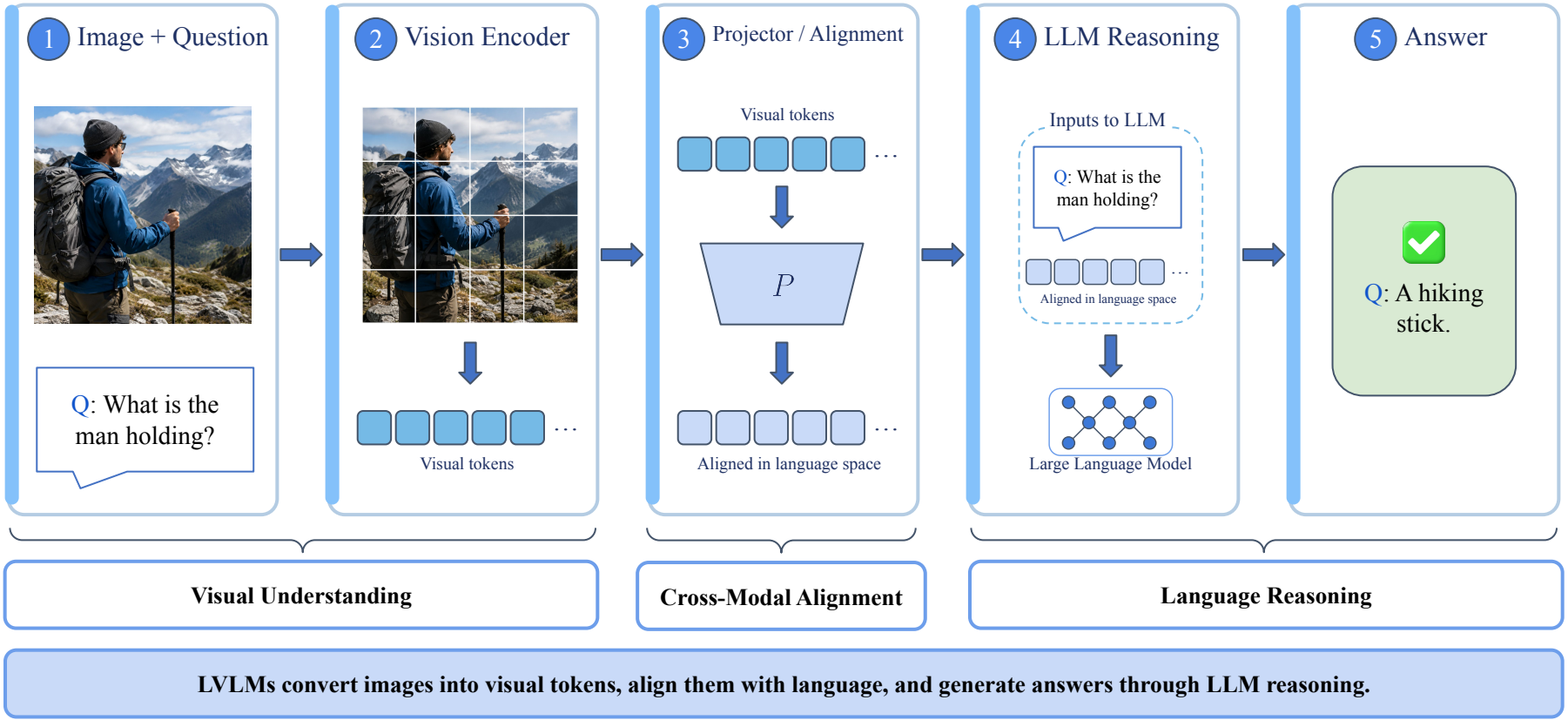
Contact

sungheoj@uci.edu | jshoney15@gmail.com

This presentation introduces DnR, a framework for more evidence-grounded multimodal reasoning agent.

Part 1: How LVLMs Work

From visual input to language answer



LVLMs convert images into visual tokens, align them with language, and generate answers through LLM reasoning.

Part 1: LVLM Challenges

The shadow of language bias

Language Bias

Model over-rely on learned language priors and ignore image content.

Q: What animal is in the picture?

Model thinking

Zebras are usually black and white.
So this must be a zebra

Model answer

A zebra.

Relies on prior knowledge (zebras are black & white) instead of the image.

Visual Ignorance

Even with images provided, models generate answers based only on text context.

Q: How many bicycle are there? Answer with a number

Model thinking

In typical street scenes, there are usually some bicycles.

Model answer

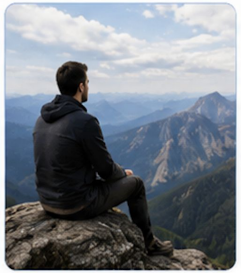
2

Ignores the visual evidence and guesses from common patterns.

Hallucination

Models confidently describe content that doesn't exist in the image

Q: What is the man holding?



Model thinking

People often hold items in outdoor photos. It looks like a cup.

Model answer

A coffee cup.

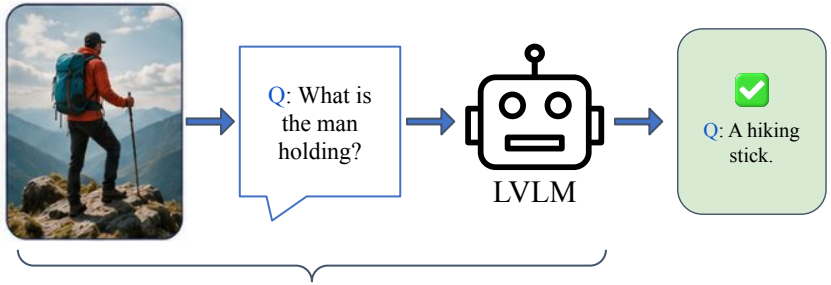
Invents details that are not present, high confidence.

These challenges stem from over-reliance on language priors rather than grounded visual understanding. They lead to unreliable and unsafe model behavior in real-world applications

Part 1: Motivation

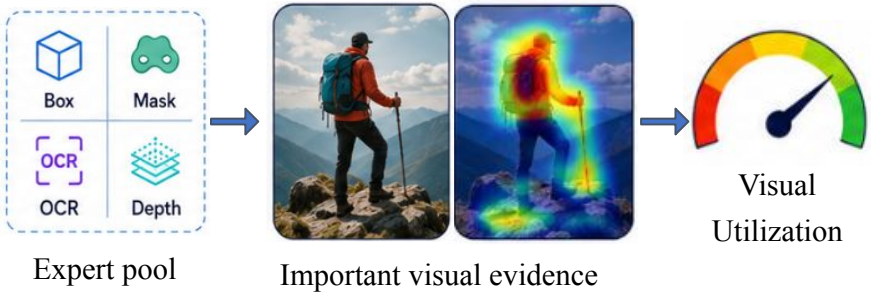
Can visual grounding guide expert selection?

1 A correct answer is not enough



A plausible answer does not guarantee grounded visual reasoning.

2 Measure grounding to choose experts



Expert pool

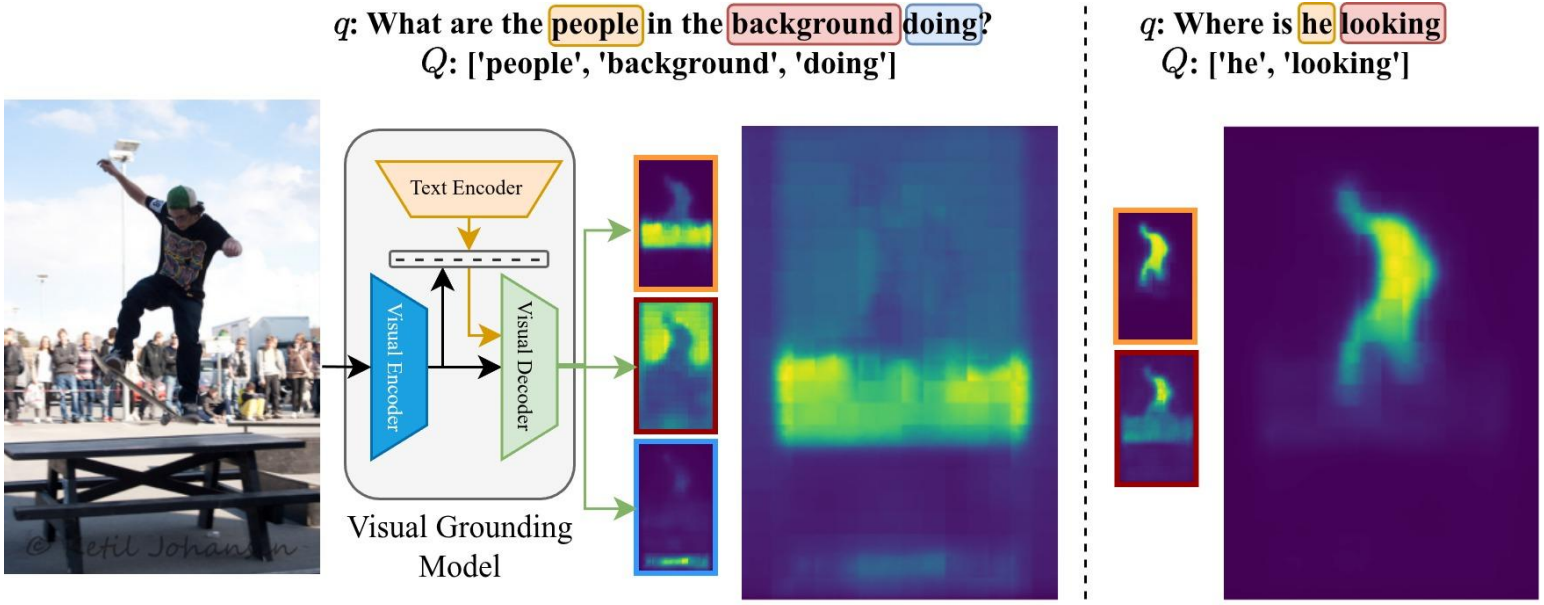
Important visual evidence

If we can quantify how much the model uses important visual evidence, we can choose visual experts in a grounded way.

Core Idea: Measure how much an LVLM uses important visual evidence, then use that signal to guide visual expert selection.

Part 2: Quantify important visual region

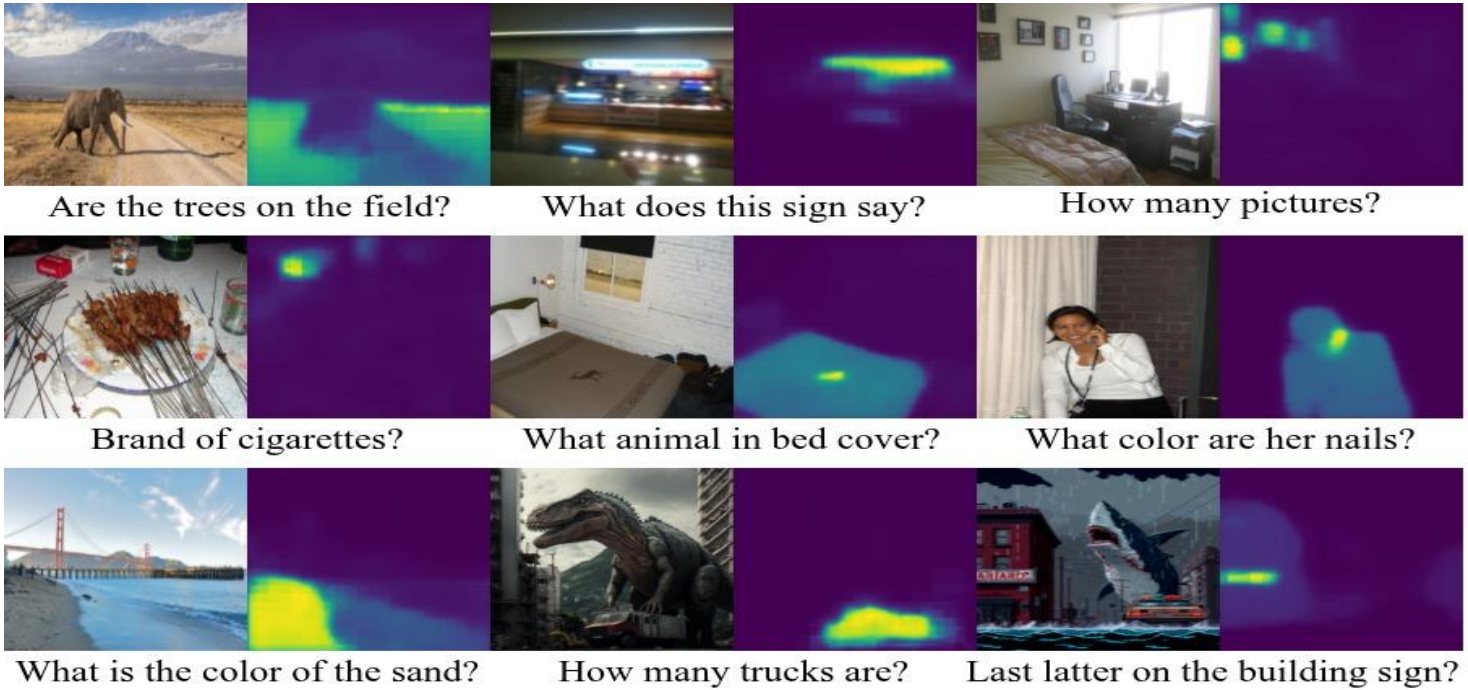
Query Decomposition: What is important part of image given question?



Key idea: The important visual region should depend on the question, not only on the image.

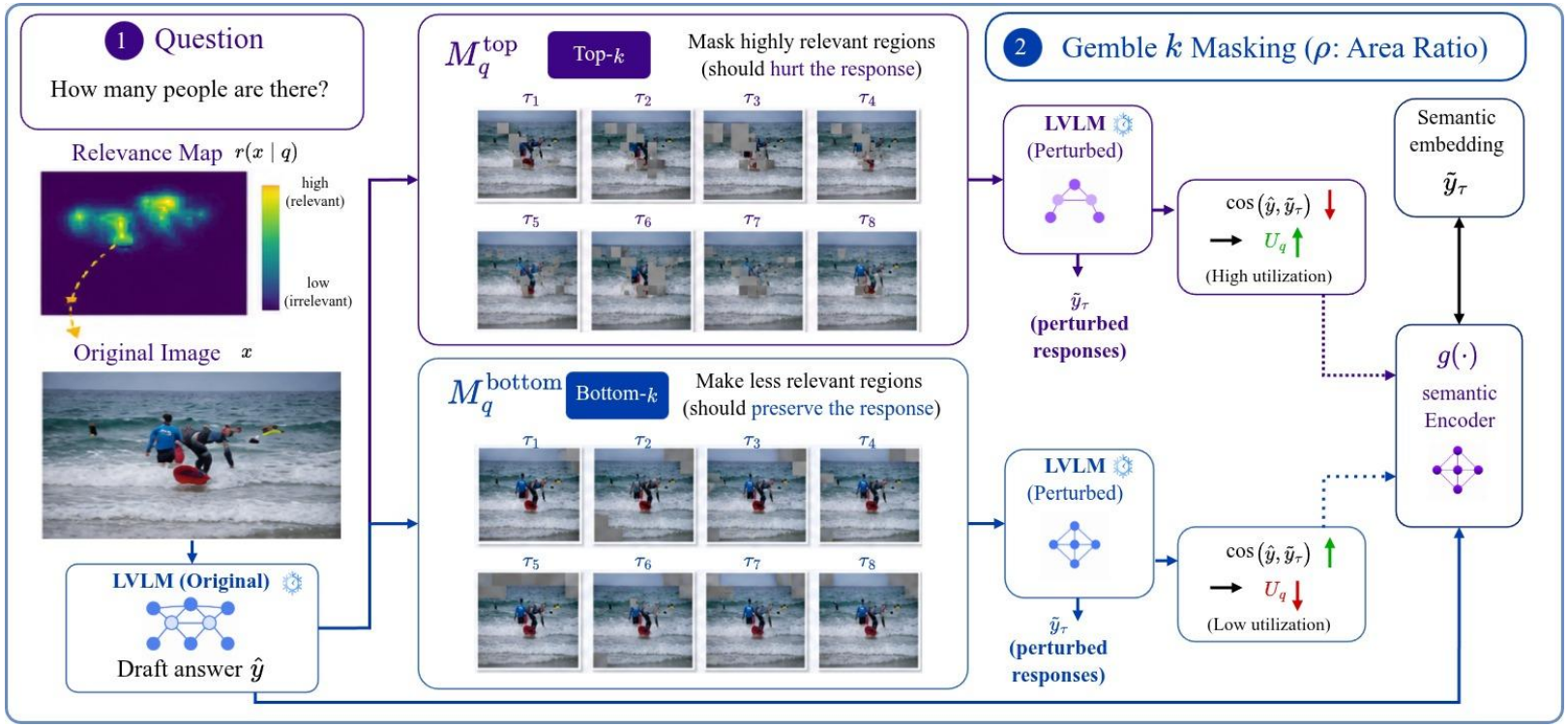
Part 2: Quantify important visual region

Relevance Map ($r(x|q)$): A question-conditioned map that localizes the image regions most relevant to answering the question.



Part 2: Quantify important visual region

Utilization Scoring via Top-k / Bottom-k Perturbation

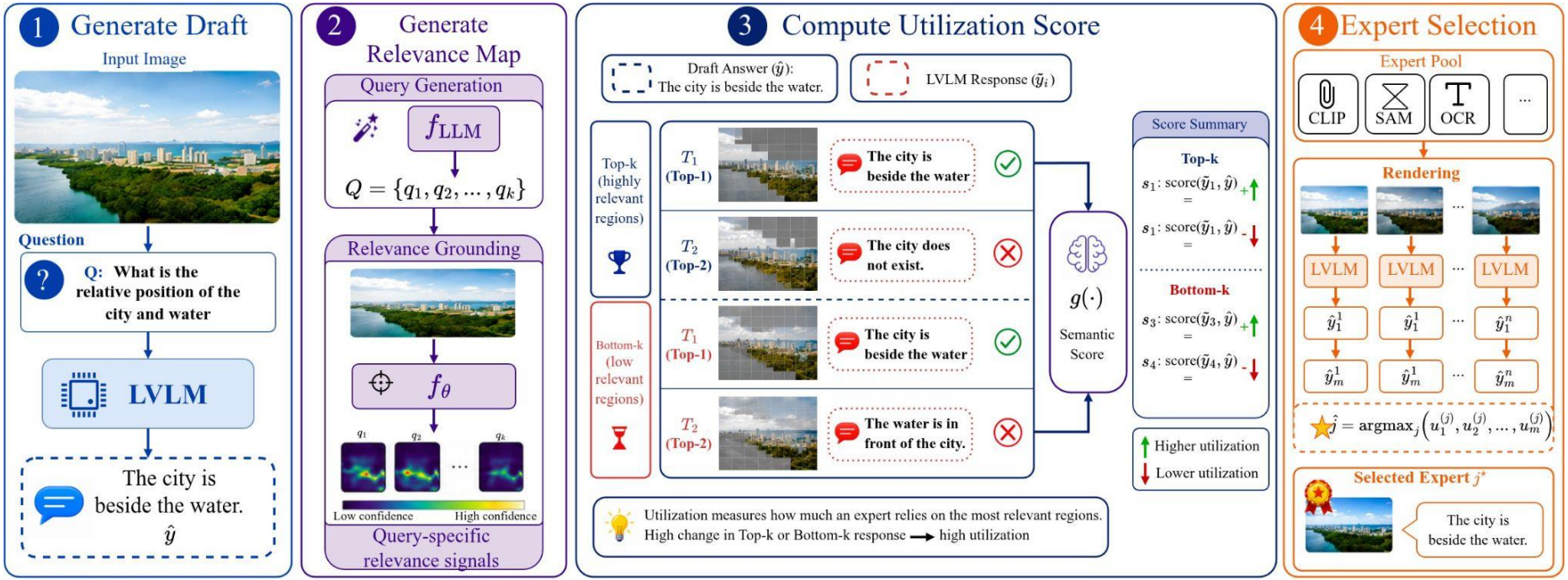


Part 2: Quantify important visual region

Rendering: How can we use the knowledge of Visual Experts?

Part 3: Draft and Refine

Overview of the entire Framework.



Part 4: Comprehensive evaluation

Draft/DnR results across 7 base LVML with Revision Rate, Degradation, Pearson metric.

VLM Backbone	VQA					Knowledge VQA			Comprehensive Benchmarks		
	VQAv2	GQA	VizWiz	TextVQA	OCR-VQA	OK-VQA	A-OKVQA	ScienceQA	MME	MMBench	SEED-Bench
IDEFICS (Draft / DnR)	37.8 / **47.85**	24.1 / **25.5**	**43.33** / 36.67	30.15 / **30.32**	47.24 / **47.74**	43.5 / **44.4**	69 / **69.5**	43.95 / **44.02**	1392.11 / **1431.58**	50.01 / **50.26**	32.16 / **32.75**
Revision Rate	29.8	1.5	19.8	2	3.2	2.9	0.2	0.2	0.6	0.4	0.3
Correction / Degradation	46.2 / 14.3	51.3 / 1.1	6.3 / 24.1	2.3 / 0.3	12.5 / 2.1	33.3 / 1.4	8.6 / 1.9	94.3 / 0.4	80.3 / 16.6	43.8 / 12.2	33.3 / 19.6
Pearson/Spearman	0.143 / 0.064	0.449 / 0.364	0.248 / 0.259	0.026 / 0.073	0.129 / 0.222	0.351 / 0.210	0.1 / 0.166	0.351 / 0.277	0.12 / 0.148	0.12 / 0.06	0.354 / 0.375
InstructBLIP	76.4 / **77.75**	38.24 / **39.77**	37.83 / **38.67**	52.43 / **54.1**	80.4 / **81.91**	49.2 / **50.2**	79.11 / **81.09**	50.5 / **52.5**	1294.74 / **1295.31**	51.84 / **52.89**	51.46 / **53.8**
Revision Rate	7.2	13	24.2	2.1	1.2	4.5	1.8	1.2	0.1	1.6	3.2
Correction / Degradation	9.1 / 5.4	35.3 / 1.2	16.1 / 2	52.3 / 25.25	24.5 / 3.4	39.5 / 18.4	42.9 / 28.6	77.6 / 4.3	92.1 / 4.3	55.3 / 4.4	36.4 / 9.1
Pearson/Spearman	0.243 / -0.024	0.290 / 0.286	0.290 / 0.286	0.066 / 0.076	0.426 / 0.415	0.168 / 0.09	0.136 / 0.104	0.608 / 0.421	0.152 / 0.128	0.01 / 0.02	0.062 / 0.145
MiniGPTv2	32.6 / **34.1**	25.3 / **27.6**	59.17 / **60.67**	36.67 / **36.68**	56.78 / **58.79**	18.1 / **19.8**	38.58 / **41.01**	28.77 / **29.41**	878.95 / **910.53**	37.63 / **39.21**	29.82 / **31.58**
Revision Rate	17.5	3.1	3.8	15.5	1.8	4.5	6.8	0.2	0.2	0.4	7.6
Correction / Degradation	12.9 / 2.9	8.3 / 0.5	41.2 / 6.7	1.6 / 0.1	11.1 / 0.1	11.1 / 0.1	14.8 / 7.2	94.3 / 0.3	50.32 / 12.1	66.7 / 1.2	30.8 / 15.7
Pearson/Spearman	0.026 / 0.051	0.194 / 0.378	0.304 / 0.234	0.492 / 0.129	0.065 / 0.124	0.206 / 0.168	0.108 / 0.122	0.719 / 0.807	0.347 / 0.35	0.422 / 0.360	0.155 / 0.131
LLaVA 1.6	80.9 / **82.81**	61.5 / **64.2**	76.83 / **76.99**	64.49 / **64.59**	73.37 / **74.87**	55.1 / **56.1**	73.5 / 73.5	72.9 / **73.8**	1694.74 / **1721.05**	76.32 / **77.89**	66.08 / **66.67**
Revision Rate	5.1	4.5	1.8	1	3.9	2	0.8	1.4	0.4	0.5	0.3
Correction / Degradation	66.7 / 9.2	33.3 / 3.1	14.3 / 2.1	25 / 0.7	66.7 / 16.7	37.5 / 24.9	33.3 / 7.8	49.8 / 0.4	98.3 / 0.6	23.3 / 7.3	95.5 / 0.4
Pearson/Spearman	0.509 / 0.639	0.288 / 0.328	0.156 / 0.154	0.155 / 0.231	0.626 / 0.849	0.297 / 0.161	0.110 / 0.128	0.454 / 0.459	0.270 / 0.307	0.109 / 0.116	0.112 / 0.01
PaliGemma	73.2 / **75.2**	58.3 / **59.9**	79.17 / **80.17**	65.83 / **66.5**	68.34 / **70.35**	57.2 / **58.5**	85.2 / **85.5**	89.1 / **89.7**	1434.21 / **1444.74**	70.09 / **71.05**	57.89 / **59.65**
Revision Rate	9.2	1.5	2.8	1.2	3.5	3.2	1	0.5	0.2	0.5	4.1
Correction / Degradation	36.4 / 27.3	16.7 / 0.8	82.1 / 2.3	42.3 / 21.3	28.6 / 0.5	60.15 / 23.1	12.3 / 0.3	92.1 / 1.3	51.1 / 12.5	23.5 / 2.1	64.3 / 31.7
Pearson/Spearman	0.292 / 0.291	0.276 / 0.399	0.307 / 0.251	0.136 / 0.03	0.321 / 0.329	0.165 / 0.128	0.441 / 0.454	0.803 / 0.853	0.08 / 0.07	0.09 / 0.1	0.129 / 0.191
CogVLM	82.05 / **82.85**	56.13 / **57.74**	48.5 / **50.33**	68.51 / **69.68**	82.91 / 82.91	58.6 / 58.6	84.5 / 84.5	61.5 / **62.22**	1384.21 / **1423.68**	76.84 / **77.89**	58.48 / **59.06**
Revision Rate	5.1	2.5	4.8	2.30	0.20	0.8	0.7	1	0.6	0.3	0.3
Correction / Degradation	66.7 / 2.3	32.2 / 1.3	5.7 / 0.4	42.9 / 28.6	1.3 / 0.1	33.1 / 4.6	3.4 / 2.1	75.43 / 21.33	66.6 / 31.1	70.1 / 2.9	98.1 / 0.3
Pearson / Spearman	0.684 / 0.735	0.223 / 0.357	0.220 / 0.136	0.318 / 0.11	0.263 / 0.28	0.500 / 0.623	0.163 / 0.14	0.499 / 0.412	0.420 / 0.470	0.01 / 0.03	0.153 / 0.227
Qwen2.5-VL	83.95 / **85.45**	57.02 / **58.31**	73.01 / **73.83**	83.92 / **84.25**	72.36 / **74.86**	58.2 / **58.3**	74.4 / **76.4**	86.5 / **87.1**	2268.42 / **2276.32**	86.05 / **86.84**	80.12 / **81.29**
Revision Rate	10.6	8.8	1.2	0.20	2.5	3.8	1.5	0.4	0.3	1.2	0.5
Correction / Degradation	40 / 4.3	16.7 / 8.3	41.3 / 1.0	33.3 / 5.7	89.3 / 3.4	13.3 / 6.7	33.2 / 16.8	82.4 / 5.3	65.6 / 32.1	55.5 / 42.3	29.4 / 17.6
Pearson/Spearman	0.389 / 0.509	0.136 / 0.222	0.19 / 0.224	0.230 / 0.112	0.499 / 0.425	0.02 / 0.023	0.43 / 0.467	0.644 / 0.703	0.5 / 0.5	0.188 / 0.083	0.170 / 0.07

Part 4: Hallucination results

Draft/DnR results categorized by Hallucination, Misperception, Grounded, and Correct responses.

Model	HaloQuest			MMHal-Bench				VizWiz				COCO Caption		
	↓H	M	C↑	↓H	M	G	C↑	↓H	M	G	C↑	↓H	M	G↑
IDEFICS	43.34/**40.87**	22.05/23.76	34.62/**35.37**	13.54/**11.50**	44.79/44.75	12.50/11.46	29.17/**32.29**	**12.06**/17.09	32.16/30.65	19.10/22.11	**36.68**/30.15	28.14/**22.11**	32.66/37.69	39.20/**40.20**
InstructBLIP	33.73/**33.02**	20.48/21.54	45.79/**45.44**	35.42/**28.12**	38.54/45.83	9.38/8.33	16.67/**17.71**	23.62/**18.59**	23.62/26.13	20.60/22.11	32.16/**33.17**	37.69/**23.62**	29.15/34.17	33.16/**42.22**
MiniGPT-v2	20.22/**19.42**	29.75/29.52	50.03/**51.06**	35.42/**27.08**	48.96/54.17	6.25/8.33	9.38/**10.42**	11.56/**10.55**	26.13/24.85	10.05/12.06	52.26/**52.54**	–	–	–
LLaVA 1.6	26.33/**25.65**	13.96/14.57	59.71/**59.79**	8.33/**4.17**	26.04/26.04	21.88/23.96	43.75/**45.83**	1.01/**0.53**	13.07/17.06	20.10/16.58	65.83/65.83	24.62/**20.60**	29.15/30.65	46.24/**48.75**
PaliGemma	20.67/**16.41**	14.96/16.63	64.37/**66.96**	12.50/**8.33**	31.25/33.33	14.58/9.38	41.67/**48.96**	2.01/**1.11**	12.56/10.45	14.57/17.09	70.85/**71.36**	30.65/**28.14**	26.63/26.13	42.72/**45.73**
CogVLM	19.24/**17.82**	15.60/16.72	65.16/**65.46**	6.25/**4.17**	52.08/48.96	9.38/17.71	**32.29**/29.17	20.60/**16.08**	15.08/17.59	22.61/24.12	41.71/**42.21**	41.21/**36.18**	21.61/24.62	37.19/**39.19**
Qwen2.5-VL	3.48/**3.07**	11.90/12.30	84.63/84.63	4.17/**3.12**	20.83/14.58	31.25/35.42	43.75/**46.88**	0.49/**0.21**	10.55/12.86	25.14/22.10	63.82/**64.83**	48.24/**26.13**	24.62/29.15	27.14/**44.72**

Part 4: Query-Driven Expert Activation

Different question queries activate different visual experts, suggesting that expert selection should be grounded in the visual demands of the question.

Part 5: Future Work

Towards More Faithful, General, and Efficient Visual Reasoning

DnR demonstrates that query-conditioned relevance and utilization can effectively guide selective refinement.
We suggest to extend this framework in several direction.

1. **Better Relevance Modeling**
   Use richer query decomposition and multi-scale grounding (e.g., attributes, relations, temporal cues) to produce more precise and complete relevance maps.

2. **Stronger Utilization Signals**
   Explore alternative perturbation strategies and semantic metrics to measure faithfulness more robustly and reduce noise.

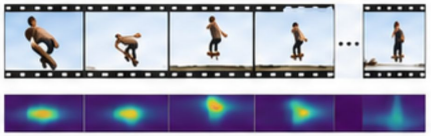
3. **Smarter Expert Selection**
   Move beyond single-step selection: learn a policy to route to the most helpful expert(s) or combine multiple experts in a principled way.

4. **Multi-step Refinement**
   Iteratively refine with updated relevance and utilization until sufficient improvement or a confidence criterion is met.

Possible Extensions

Temporal / Video Understanding

Extend relevance grounding and utilization to video with motion-aware queries.



Finer-grained and Structured Experts

Incorporate more specialized experts (e.g., counting, spatial reasoning, chart) and structured outputs.

Learning to Predict Utilization

Train a lightweight predictor to estimate utilization without test-time perturbation for efficiency.

Broader Domains & Modalities

Apply DnR to document, chart, medical, and 3D date with domain-specific experts.

Long-term Goal Build a general, efficient, and interpretable refinement system that reliably grounds reasoning in the right visual evidence across tasks and domains.

Open Challenges

Faithfulness vs. Performance
Further study the trade-off and develop metrics that better reflect faithful reasoning.

Perturbation Design
Design perturbations that are more human-aligned and less disruptive to semantics and reasoning

Expert Integration
How to best fuse expert evidence with the LVLM in a consistent and controllable way.

Efficiency & Scalability
Reduce computation for relevance, utilization, and expert calls to scale to longer contexts.

CVPR JUNE 3-7, 2026 DENVER COLORADO

Thank You

We appreciate your attention to our research on the **Draft-and-Refine** framework for enhancing LVLM reliability and efficiency through quantifiable visual utilization.

For more information, please contact us at sungheoj@uci.edu or visit our project page.