[NeurIPS 2026]

CoMet: Context and Multiplicity Decomposition for Multimodal Uncertainty Estimation

Sanghyuk Chun
sanghyukc@princeton.edu

Joint work with William Yang, Amaya Dharmasiri and Olga Russakovsky

Outline

- Introduction: Uncertainty estimation in multimodal large language models (MLLMs)
- Existing approaches and challenges
- CoMet: Context and multiplicity decomposition
- Experiments
- Conclusion

Outline

- Introduction: Uncertainty estimation in multimodal large language models (MLLMs)
- Existing approaches and challenges
- CoMet: Context and multiplicity decomposition
- Experiments
- Conclusion

Why does “uncertainty” matter for MLLMs?

- Multimodal Large Language Models (MLLMs) are increasingly used in high-stakes applications, such as visual assistance for people who are blind.

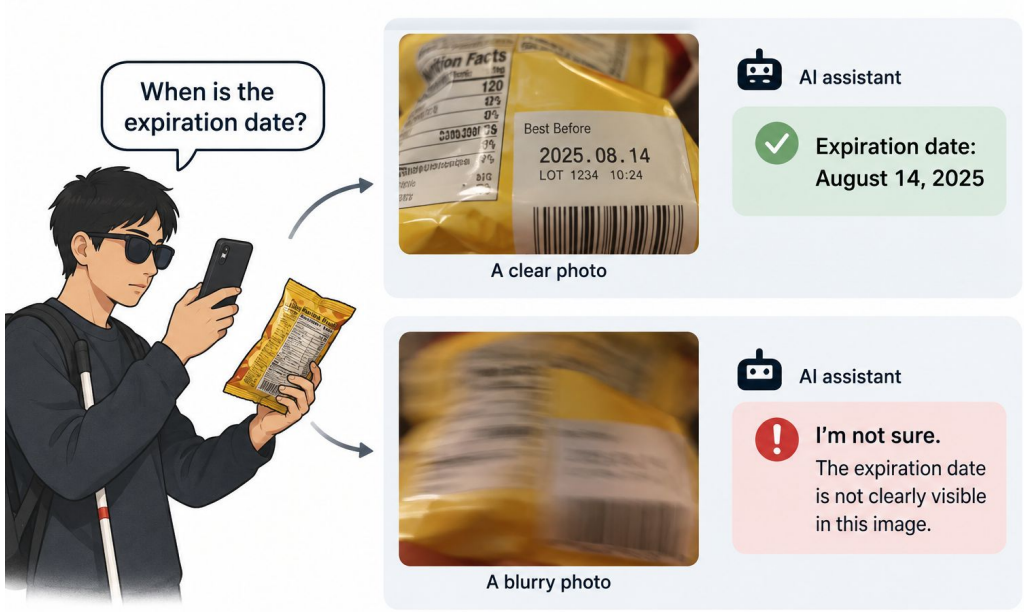
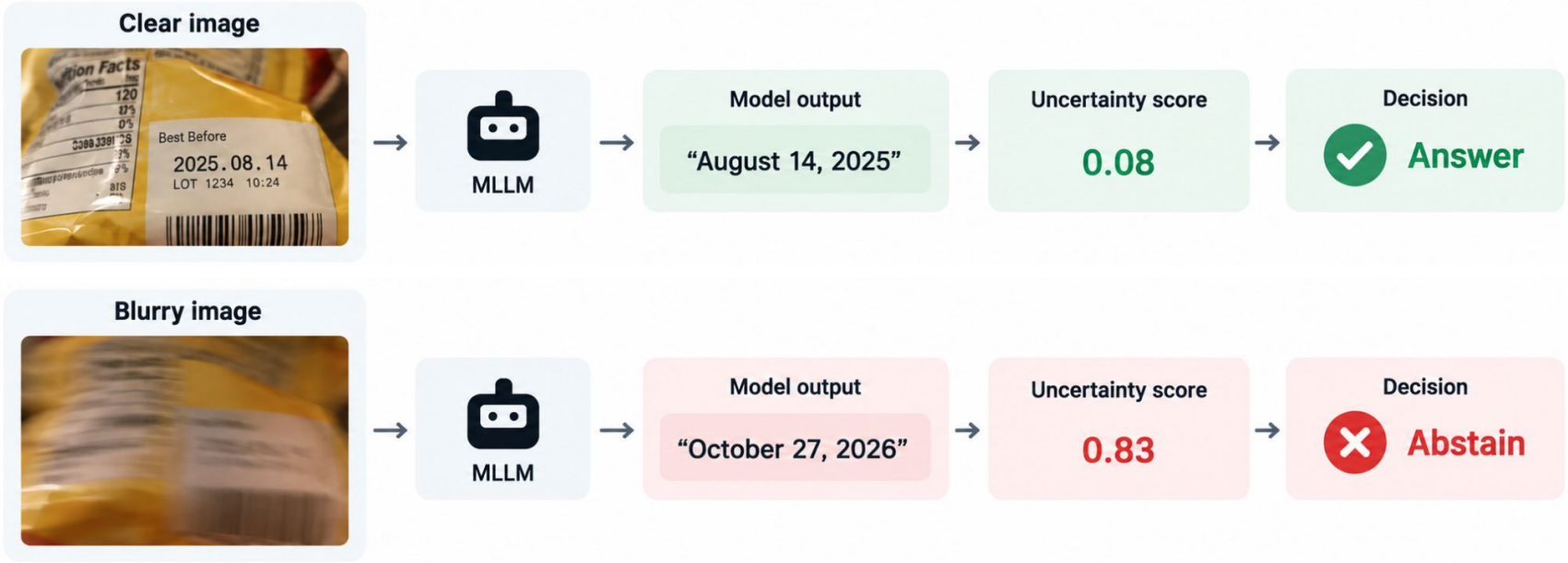


Image generated by ChatGPT

Princeton Visual AI Lab

PRINCETON UNIVERSITY

How can we know when to trust the model?



- Uncertainty estimation enables us to decide when to trust (or not to trust) the answer
- Example applications: failure prediction, dataset diagnosis, hallucination detection, or active learning

Image generated by ChatGPT

Princeton Visual AI Lab
PRINCETON UNIVERSITY

How can we estimate uncertainty?

- Uncertainty quantification (UQ) for **closed-set tasks** (when the valid outputs are fixed and enumerable, e.g., **classification**) is relatively straightforward

Input



Q: What animal is shown?

Answer candidates (pre-defined, fixed)	Prediction score (normalized)
Duck	0.03
Rabbit	0.02
Cat	0.07
Zebra	0.88

Uncertainty (Entropy)

$$-\sum_i p_i \log p_i$$

0.695

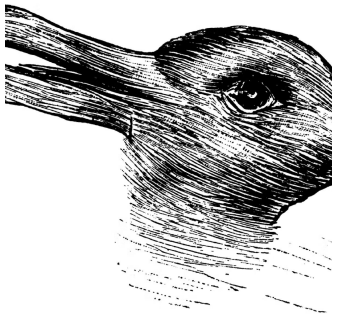
Entropy is a natural uncertainty score:
lower entropy → more confident

Princeton Visual AI Lab
PRINCETON UNIVERSITY

How can we estimate uncertainty?

- Uncertainty quantification (UQ) for **closed-set tasks** (when the valid outputs are fixed and enumerable, e.g., **classification**) is relatively straightforward

Input



Q: What animal is shown?

Answer candidates (pre-defined, fixed)	Prediction score (normalized)
Duck	0.49
Rabbit	0.49
Cat	0.01
Zebra	0.01

Uncertainty (Entropy)

$$-\sum_i p_i \log p_i$$

1.141

Entropy is a natural uncertainty score: higher entropy → less confident

Princeton Visual AI Lab | PRINCETON UNIVERSITY

But what about open-ended tasks?

Input	Answer candidates (unknown, no explicit candidates)	Prediction score (unnormalized)
 Q: What is happening in this image?	There are two zebras	-2.3
	Two zebras are in front of a castle	-3.1
	It looks cloudy	1.5
	⋮	⋮

$$-\sum_i p_i \log p_i$$

We cannot compute the entropy score anymore!

Goal 1. Estimating uncertainty for open-ended tasks

Outline

- Introduction: Uncertainty estimation in multimodal large language models (MLLMs)
- **Existing approaches and challenges**
- CoMet: Context and multiplicity decomposition
- Experiments
- Conclusion

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Challenges in MLLM uncertainty estimation

MLLM outputs are **open-ended**

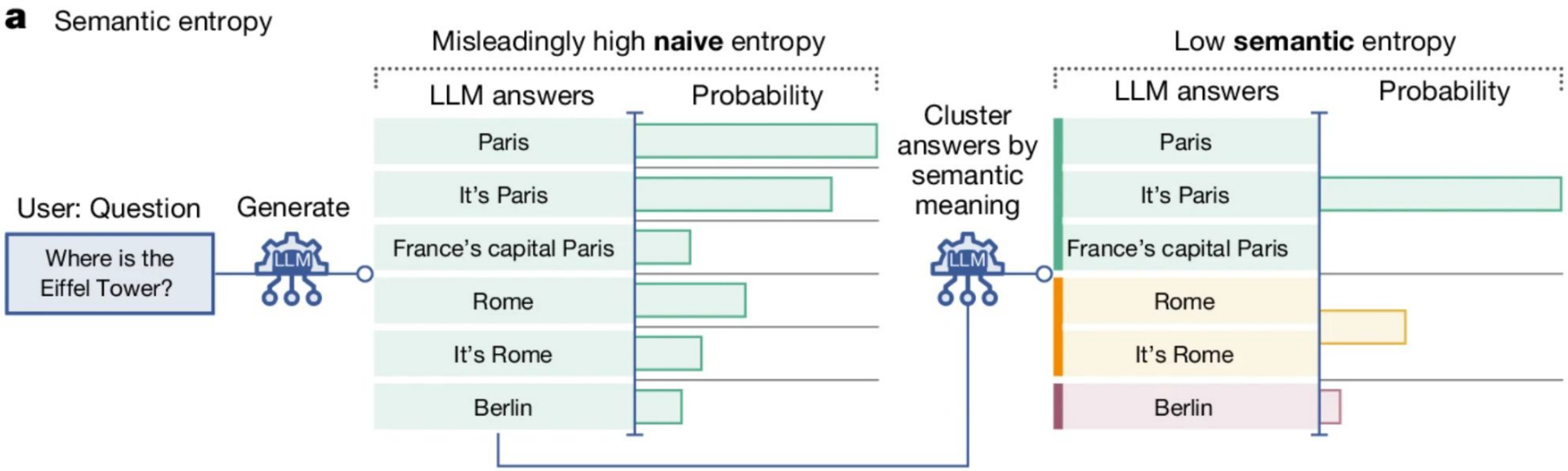
We often require **generating answers multiple times**

It is often unclear how uncertainty scores should be **interpreted**

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Previous solution (1): UQ by repeated generation



"Consistency" (or semantic) Entropy

Previous solution (1): UQ by repeated generation

✅ Enables entropy-based uncertainty estimation for MLLMs

❌ Repeated generation is computationally expensive

❌Sensitive to generation hyperparameters
e.g., prompt, temperature, decoding algorithm, …

Goal 2. We aim to remove the dependency on autoregressive generation to reduce the generation-related hyperparameter sensitivity and its cost

Previous solution (2): Self-assessing the generated answer



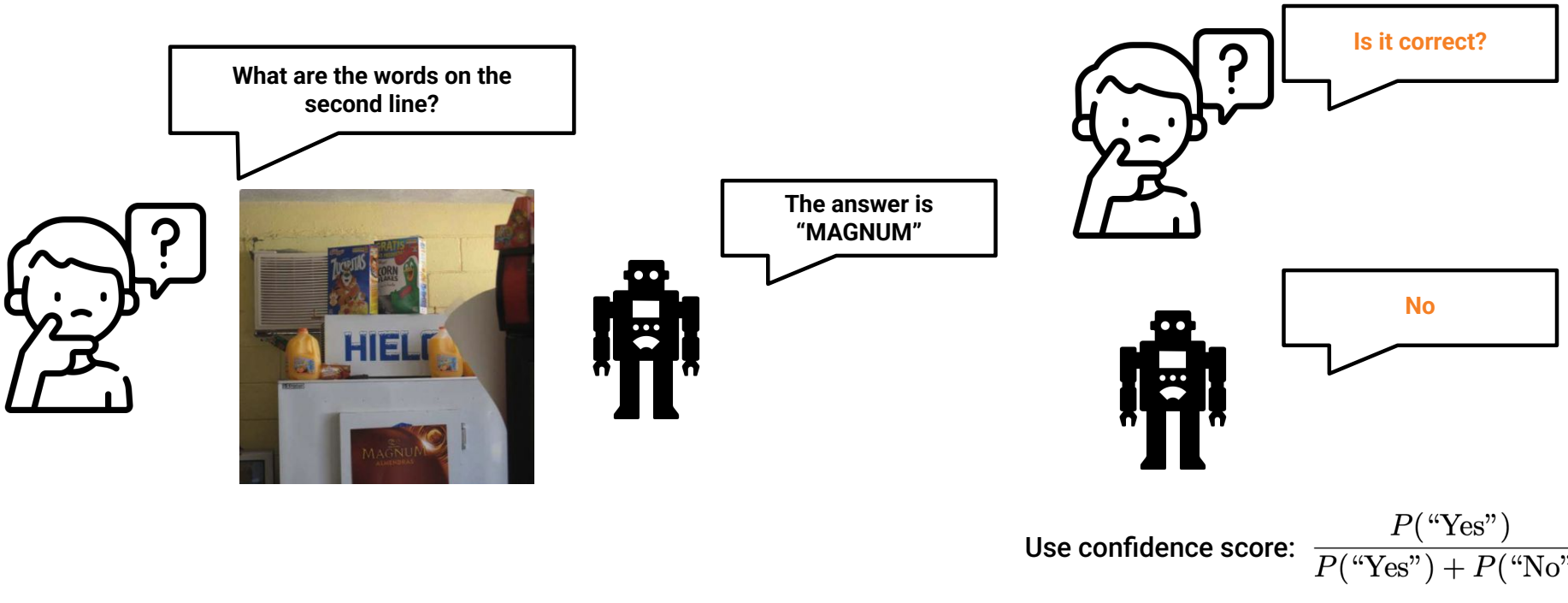

```
Image: {x}
Question: {t}
Proposed Answer: {y}
How likely is the possible answer to be correct? Briefly explain, then write
only one score from 0 to 10 inside <confidence></confidence> tags.
```


Tian, et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. EMNLP 2023

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Previous solution (2): Self-assessing the generated answer



Use confidence score: $\frac{P(\text{“Yes”})}{P(\text{“Yes”}) + P(\text{“No”})}$

Kadavath, et al. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022).

Previous solution (2): Self-assessing the generated answer

✅ Leverages the model's ability to assess an answer for UQ

❌ Still requires generating the answer to assess

❌It is unclear what the score actually measures, so sometimes we need additional fine-tuning of the (M)LLM **[a, b]**

Goal 3. We aim to achieve a theoretically sound uncertainty estimation method, which interprets the results well and works without fine-tuning the MLLM backbone

[a] Kapoor, et al. Large language models must be taught to know what they don't know. NeurIPS 2024
[b] Xiao, et al. VL-Calibration: Decoupled Confidence Calibration for Large Vision-Language Models Reasoning. ACL 2026

Outline

- Introduction: Uncertainty estimation in multimodal large language models (MLLMs)
- Existing approaches and challenges
- **CoMet: Context and multiplicity decomposition**
- Experiments
- Conclusion

Recap: Challenges in MLLM uncertainty estimation

MLLM outputs are **open-ended**

We often require **generating answers multiple times**

It is often unclear how uncertainty scores should be **interpreted**

Can we address all three challenges?

Problem setting (1): Answer candidate “y”



“Image” x



What is this animal?

“Question” t



Cat

“An answer candidate” y

$$p(y \mid t)$$

p(y | x, t) = p(y | t)

We need more information to connect x and y

Problem setting (2): Binary matching variable “m”



“Image” x

What is this animal?

“Question” t

Cat

“An answer candidate” y

Is “Cat” (y) a plausible answer for the image “x” and the question “t”?
No (m=0)

“Matching” m

Problem setting (2): Binary matching variable “m”



“Image” x

What is this animal?

“Question” t

Zebra

“An answer candidate” y

Is “Zebra” (y) a plausible answer for the image “x” and the question “t”?
Yes (m=1)

“Matching” m

Formulate uncertainty estimation using x, t, y and m

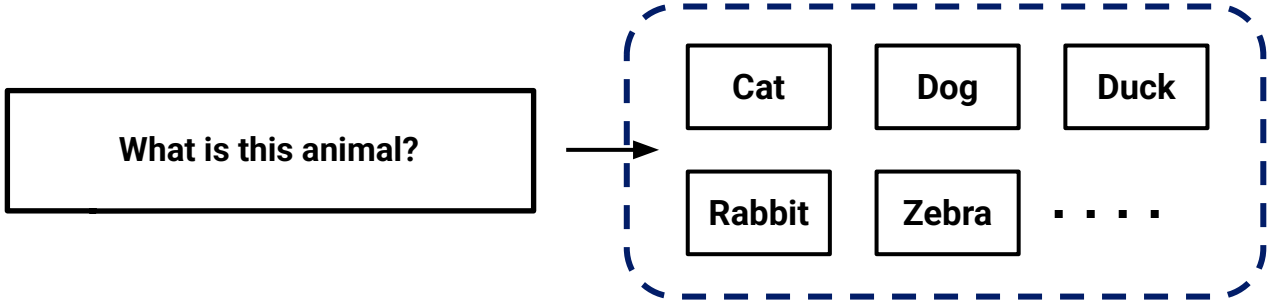
- Our goal is to quantify uncertainty over plausible answers given image x and question t, whose distribution can be represented by p(y | x, t, m=1)

$$p(y \mid x, t, m=1) = \frac{p(m=1 \mid x, t, y)p(y \mid x, t)}{\int p(m=1 \mid x, t, y')p(y' \mid x, t)dy'} = \frac{p(m=1 \mid x, t, y)p(y \mid t)}{\int p(m=1 \mid x, t, y')p(y' \mid t)dy'}.$$

$$\text{uncertainty}(x, t) = \text{Entropy}\big(p(y \mid x, t, m=1)\big)$$

Candidate distribution p(y | t)

$$p(y \mid x, t, m=1) = \frac{p(m=1 \mid x, t, y)p(y \mid x, t)}{\int p(m=1 \mid x, t, y')p(y' \mid x, t)dy'}$$
$$= \frac{p(m=1 \mid x, t, y)p(y \mid t)}{\int p(m=1 \mid x, t, y')p(y' \mid t)dy'}.$$



"Question" t

"All answer candidates for given t"

$$p(y \mid t) = \frac{1}{\int_\Omega dy} := \frac{1}{K_t}.$$

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Matching probability p(m=1 | x, t, y)

$$p(y \mid x, t, m=1) = \frac{p(m=1 \mid x, t, y)p(y \mid x, t)}{\int p(m=1 \mid x, t, y')p(y' \mid x, t)dy'}$$
$$= \frac{p(m=1 \mid x, t, y)p(y \mid t)}{\int p(m=1 \mid x, t, y')p(y' \mid t)dy'}.$$



What is this animal?

Cat

→ "Is this answer correct?"

"Image" x

"Question" t

"An answer candidate" y

$$p(m=1 \mid x, t, y) = \frac{p(\text{“Yes”})}{p(\text{“Yes”}) + p(\text{“No”})}$$

Princeton Visual AI Lab

PRINCETON UNIVERSITY

CoMet uncertainty decomposition

Context and Multiplicity Decomposition (CoMet) for Multimodal Uncertainty Estimation

$$\pi_y := p(y \mid x, t, m = 1)$$

$$\text{uncertainty}(x, t) = \text{Entropy}\big(p(y \mid x, t, m = 1)\big)$$

$$= \underbrace{\log K_t}_{:= u_t(t)} + \underbrace{\Big[-\log\Big(\big\{\text{CV}_{Y \sim p(\cdot \mid t)}[p(m = 1 \mid x, t, Y)]\big\}^2 + 1\Big)\Big]}_{:= u_{x \mid t}(x \mid t)}$$

$$= \text{context-specific uncertainty} + \text{multiplicity-specific uncertainty}$$

$$p(y \mid t) = \frac{1}{\int_\Omega dy} := \frac{1}{K_t}.$$

$$\text{CV}(s) = \frac{\sqrt{\text{Var}(s)}}{\mathbb{E}(s)}$$

Princeton Visual AI Lab

PRINCETON UNIVERSITY

What do the two uncertainty terms measure?

How broad is the answer candidate space induced by the question *t*?

How many of those answer candidates remain plausible with *x*?

$$u(x,t) = \underbrace{u_t(t)}_{\text{context-specific}} + \underbrace{u_{x|t}(x \mid t)}_{\text{multiplicity-specific}}$$

Context-specific uncertainty

How broad is the answer candidate space induced by the question *t*?

How many of those answer candidates remain plausible with *x*?

$$u(x,t) = \underbrace{u_t(t)}_{\text{context-specific}} + \underbrace{u_{x|t}(x \mid t)}_{\text{multiplicity-specific}}$$

Context-specific uncertainty

- **Before seeing an image *x*,** the question can induce the uncertainty
- **"How broad is the answer candidate space induced by the context *t*?"**
- In terms of this uncertainty, **t1 < t2 < t3**

Is it a zebra?	vs.	How many zebras in the photo?	vs.	What is happening in this image?
t1		t2		t3
"Yes" / "No"		Nonnegative Integers		A very broad and general answer space

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Multiplicity-specific uncertainty

How broad is the answer candidate space induced by the question *t*?

How many of those answer candidates remain plausible with *x*?

$$u(x,t) = \underbrace{u_t(t)}_{\text{context-specific}} + \underbrace{u_{x|t}(x \mid t)}_{\text{multiplicity-specific}}$$

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Multiplicity-specific uncertainty

- **After seeing an image *x*,** the image reduces the answer candidate space induced by ***t***
- **"How many answer candidates remain plausible with the observed image?"**

Context

How many zebras in the photo?

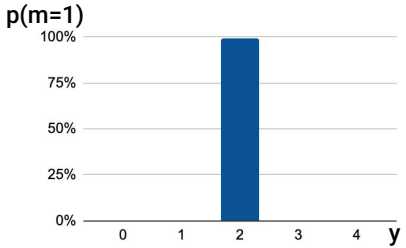
How many stripes on the zebra on the right?

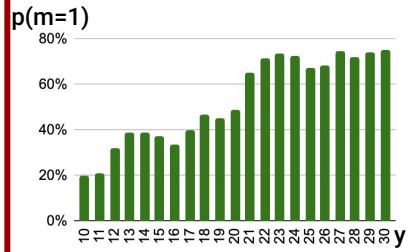
Context-plausible answers

Nonnegative Int: [0, 1, 2, … N]

Nonnegative Int: [0, 1, 2, … N]

p(m=1 | x, t, y)





More uncertain



Image input: x

Multiplicity-specific uncertainty

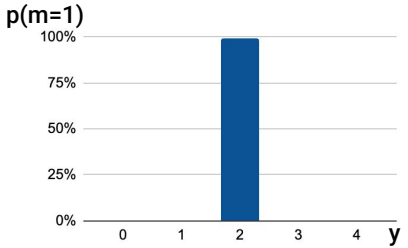
- After seeing an image *x*, the image reduces the answer candidate space induced by *t*
- "How many answer candidates remain plausible with the observed image?"
- $$u_{x|t}(x \mid t) := -\log\left(\left\{\mathrm{CV}_{Y\sim p(\cdot|t)}[p(m=1 \mid x,t,Y)]\right\}^2 + 1\right)$$

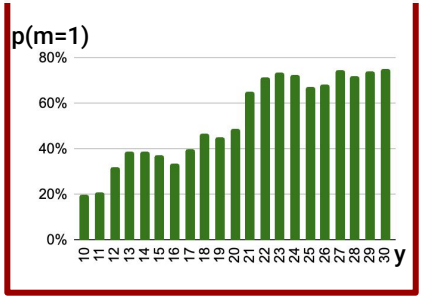
$$\text{where, } \mathrm{CV}(s) = \frac{\sqrt{\mathrm{Var}(s)}}{\mathbb{E}(s)}$$

Context: How many zebras in the photo? / How many zebras on the zebra on the right?

Context-plausible answers: Nonnegative Int: [0, 1, 2, … N] / Nonnegative Int: [0, 1, 2, … N]

p(m=1 | x, t, y)





More uncertain

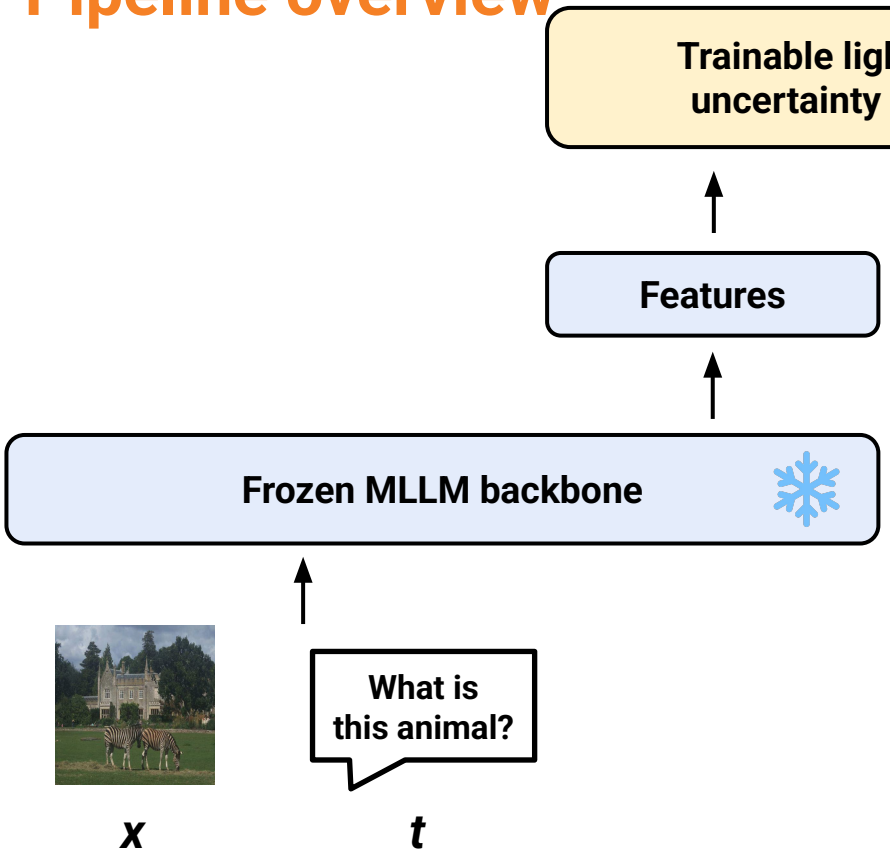


Image input: x

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Pipeline overview



Output 1: $\log K_t$

(Context-specific uncertainty)

Output 2: $-\log\left(\left\{\mathrm{CV}_{Y\sim p(\cdot|t)}[p(m=1\mid x,t,Y)]\right\}^2+1\right)$

(Multiplicity-specific uncertainty)

Goal: Train uncertainty modules with a regression loss with the target uncertainty values

No generation / sampling / or candidate sets at inference

What do we need to construct the training targets?

$$u(x,t) = \underbrace{\log K_t}_{:=u_t(t)} + \underbrace{\left[-\log\left(\left\{\mathrm{CV}_{Y\sim p(\cdot|t)}[p(m=1\mid x,t,Y)]\right\}^2+1\right)\right]}_{:=u_{x|t}(x|t)}$$

- We need to estimate the following two terms:

$$p(y\mid t) = \frac{1}{\int_\Omega dy} := \frac{1}{K_t}.$$

$$p(m=1\mid x,t,y)$$

What do we need to construct the training targets?

$$u(x,t) = \underbrace{\log K_t}_{:=u_t(t)} + \underbrace{\left[-\log\left(\left\{\mathrm{CV}_{Y\sim p(\cdot|t)}[p(m=1\mid x,t,Y)]\right\}^2+1\right)\right]}_{:=u_{x|t}(x|t)}$$

- We need to estimate the following two terms:

$$p(y\mid t) = \frac{1}{\int_\Omega dy} := \frac{1}{K_t}.$$

$$p(m=1\mid x,t,y)$$

What do we need to construct the training targets?

$$u(x,t) = \underbrace{\log K_t}_{:=u_t(t)} + \underbrace{\left[-\log\left(\left\{\mathrm{CV}_{Y\sim p(\cdot|t)}\left[p(m=1 \mid x,t,Y)\right]\right\}^2 + 1\right)\right]}_{:=u_{x|t}(x|t)}$$

- We need to estimate the following two terms:

$$p(y \mid t) = \frac{1}{\int_\Omega dy} := \frac{1}{K_t}.$$

$$p(m=1 \mid x,t,y)$$

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Estimating matching probability using MLLMs

Prompts for matching probability verification

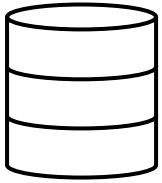
```
Image: {x}
Question: {t}
Proposed Answer: {y}
Only answer Yes if the image clearly supports the answer.
If the image is missing, unclear, or the answer is not visually confirmed,
answer No.
Answer Yes or No.
```

$$p(m = 1 \mid x, t, y) = \frac{p(\text{“Yes”})}{p(\text{“Yes”}) + p(\text{“No”})}$$

(Simplified)

Princeton Visual AI Lab
PRINCETON UNIVERSITY

Approximating p(y | t) using text clustering

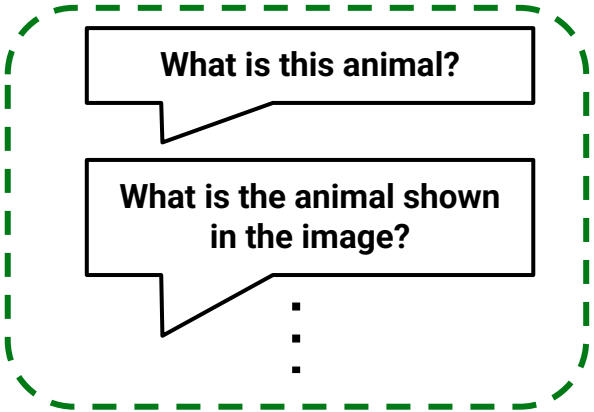


{ (x, t, y) }
(Images, Questions, Answers)

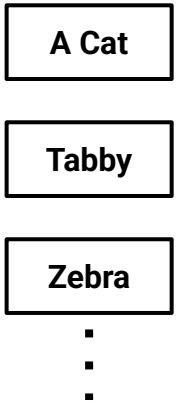


Cluster similar questions

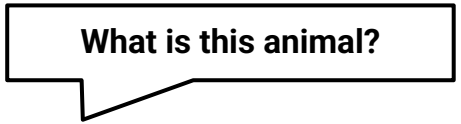
Clustered questions *t*



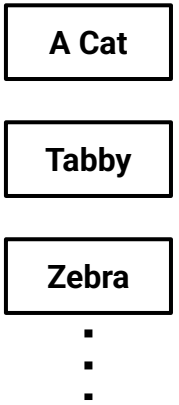
Answer candidates Y_t



Question *t*

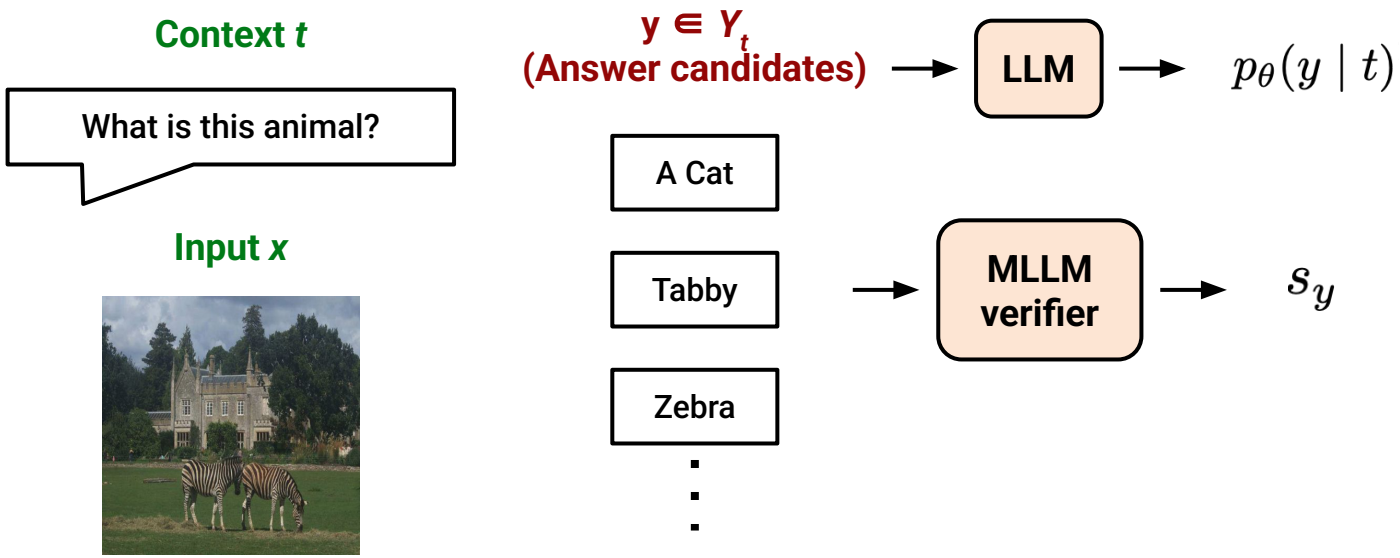


The answer candidate set Y_t approximates p(y | t)



Constructing training targets

- Step 1. Collecting matching probability scores for all answer candidates of (x, t)



Constructing training targets

- Step 2. Aggregating collected scores over the answer candidates to get $u_{x|t}$ and u_t

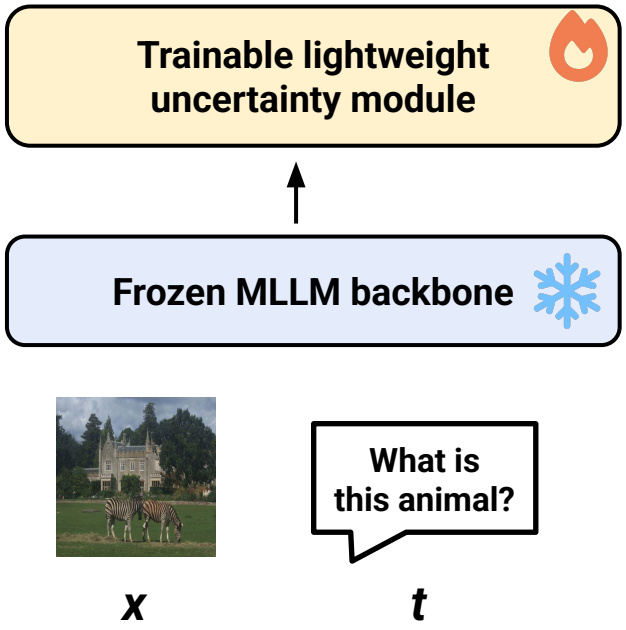
$$u_t = -\log \sum_{y \in \mathcal{Y}_t} p_\theta(y \mid t)^2$$

$$u_{x|t} = 2 \log Z_1 - \log Z_2$$

$$Z_1 = \frac{1}{|\mathcal{Y}_t|} \sum_{y \in \mathcal{Y}_t} s_y \quad Z_2 = \frac{1}{|\mathcal{Y}_t|} \sum_{y \in \mathcal{Y}_t} s_y^2$$

Training

- Step 3. Train the uncertainty module with the constructed training targets



(Regression)

$$u_t = -\log \sum_{y \in \mathcal{Y}_t} p_\theta(y \mid t)^2$$

$$u_{x|t} = 2 \log Z_1 - \log Z_2$$

$$Z_1 = \frac{1}{|\mathcal{Y}_t|} \sum_{y \in \mathcal{Y}_t} s_y \quad Z_2 = \frac{1}{|\mathcal{Y}_t|} \sum_{y \in \mathcal{Y}_t} s_y^2$$

Princeton Visual AI Lab

PRINCETON UNIVERSITY

CoMet addresses all three challenges

MLLM outputs are **open-ended**

CoMet enables uncertainty estimation for **open-ended tasks**

We often require **generating answers multiple times**

CoMet can estimate uncertainty **directly from inputs without any generation**

It is often unclear how uncertainty scores should be **interpreted**

CoMet decomposes uncertainty into two components: **context and multiplicity**

Outline

- Introduction: Uncertainty estimation in multimodal large language models (MLLMs)
- Existing approaches and challenges
- CoMet: Context and multiplicity decomposition
- **Experiments**
- Conclusion

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Benchmarks

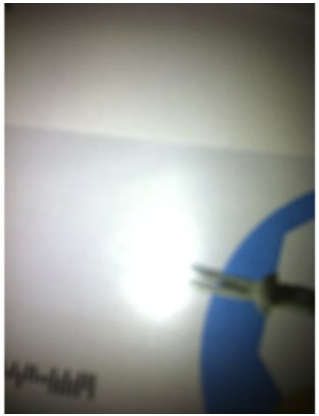
Openset multimodal VQA



Q: What are the words on the second line?

A: Almendras

Unreliability detection



Q: Who is this mail for? Is it answerable?

A: unanswerable

Multiple choice VQA



Question: The patient is a post-mature newborn. What is the most likely diagnosis? <image 1>
Option:
(A) Transient tachypnea of the newborn
(B) Congenital lobar emphysema
(C) Hyaline membrane disease
(D) Meconium aspiration syndrome

A: (D)

Antol, et al. "Vqa: Visual question answering." ICCV 2015
Gurari, et al., Vizwiz grand challenge: Answering visual questions from blind people, CVPR 2018
Yue et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI, CVPR 2024

Benchmarks

Openset multimodal VQA

3 datasets

VQA v2
VizWiz
OK-VQA

Unreliability detection

2 datasets

VizWiz (unanswerable)
HallucinationBench

Multiple choice VQA

3 datasets

MMMU
MMMU Pro
MMStar

4 backbones

Qwen3-VL family (2B, 4B, 8B)
InternVL3.5 family (1B)

Antol, et al. "Vqa: Visual question answering." ICCV 2015
Gurari, et al., Vizwiz grand challenge: Answering visual questions from blind people, CVPR 2018
Yue et al., MMMU: A Massive Multi-discipline Multimodal Understanding and Reasoning Benchmark for Expert AGI, CVPR 2024

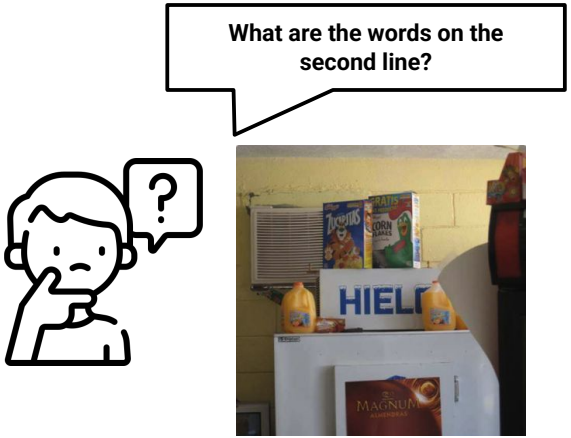
Comparison methods

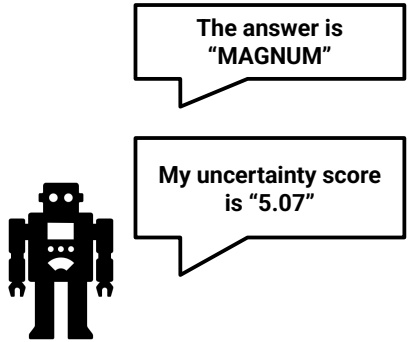
- **“Baseline”**
  - Negative Log Likelihood (or Perplexity)
- **“Uncertainty estimation with repeated generation”**
  - “Consistency” (or semantic) Entropy [a, b]
- **“Self-assessment methods”**
  - P(Correct) [c]
  - Verbalized confidence [d]
- **“NLL + Input perturbation”**
  - VL Uncertainty [e]

[a] Farquhar, et al., Detecting hallucinations in large language models using semantic entropy, Nature 2024
[b] Wang, et al. Self-consistency improves chain of thought reasoning in language models, ICLR 2023
[c] Kadavath, et al. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022).
[d] Tian, et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. EMNLP 2023
[e] Zhang, et al. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation, arXiv 2024

How can we evaluate different uncertainty estimators?

- We can treat this as **a binary classification problem**:
  - Input: uncertainty score by **the uncertainty estimator** (real-valued scalar)
  - Output: correct or incorrect by **the MLLM** (binary classification)



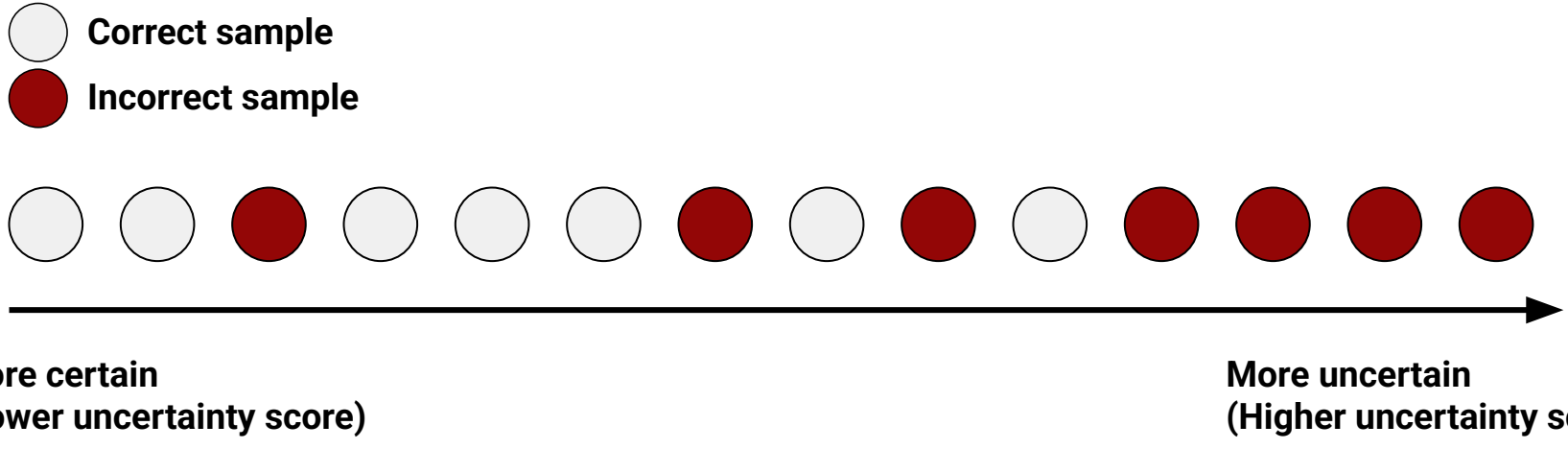


Incorrect answer **(target binary class)**

5.07 **(input value of a binary classifier)**

How can we evaluate different uncertainty estimators?

- We can treat this as **a binary classification problem**:
  - Input: uncertainty score by **the uncertainty estimator** (real-valued scalar)
  - Output: correct or incorrect by **the MLLM** (binary classification)
- We will use AUROC, a threshold-free metric



Princeton Visual AI Lab

PRINCETON UNIVERSITY

CoMet shows better results across three settings

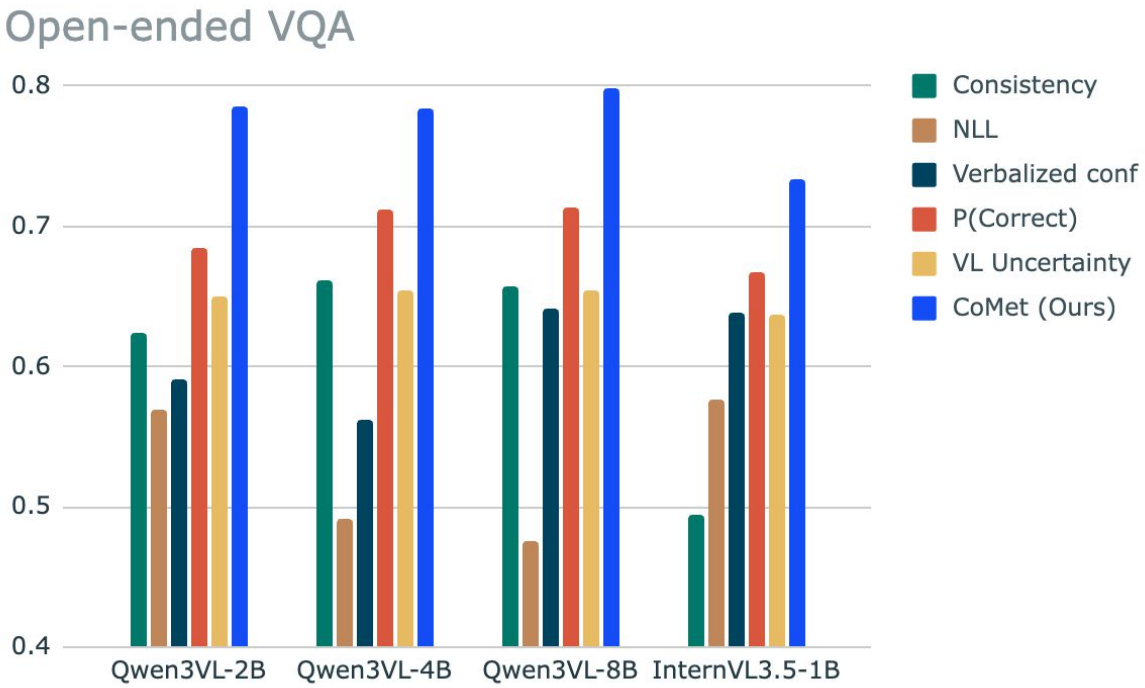
Setting	Best comparison	CoMet	Gain
Openset multimodal VQA	71.4	**79.9**	**+8.5**
Unreliability detection	63.5	**70.1**	**+6.6**
Multiple choice VQA	73.3	**78.3**	**+5.0**

Backbone: Qwen3VL-8B-Instruct
Metric: Average AUROC

Princeton Visual AI Lab

PRINCETON UNIVERSITY

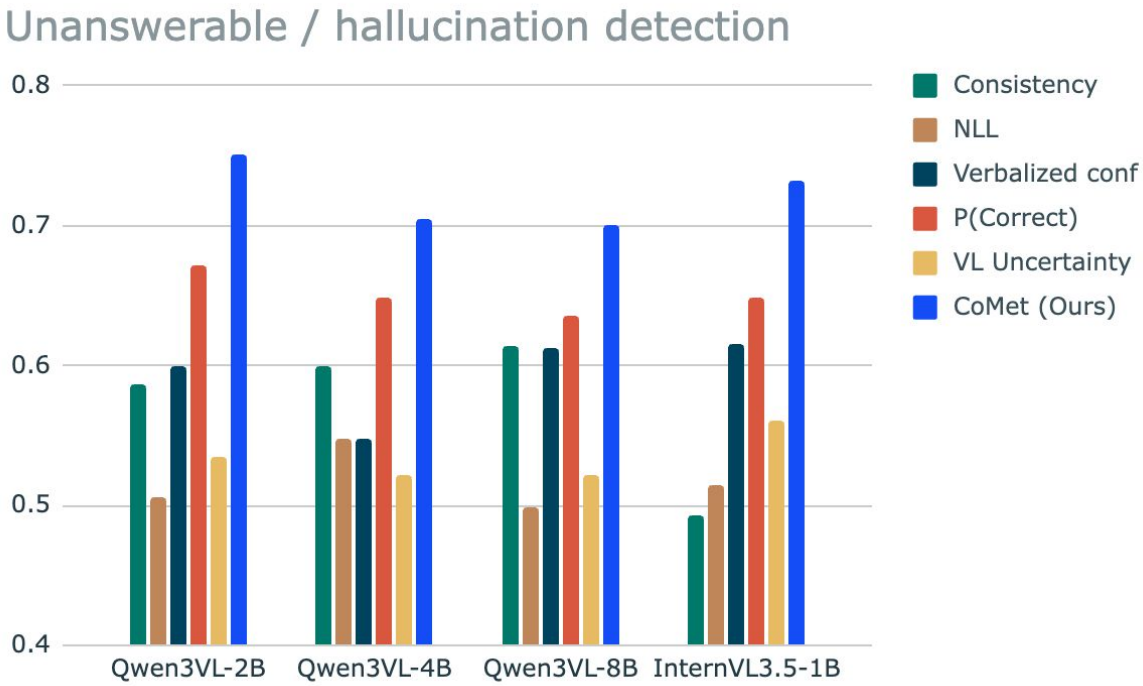
Results



Princeton Visual AI Lab

PRINCETON UNIVERSITY

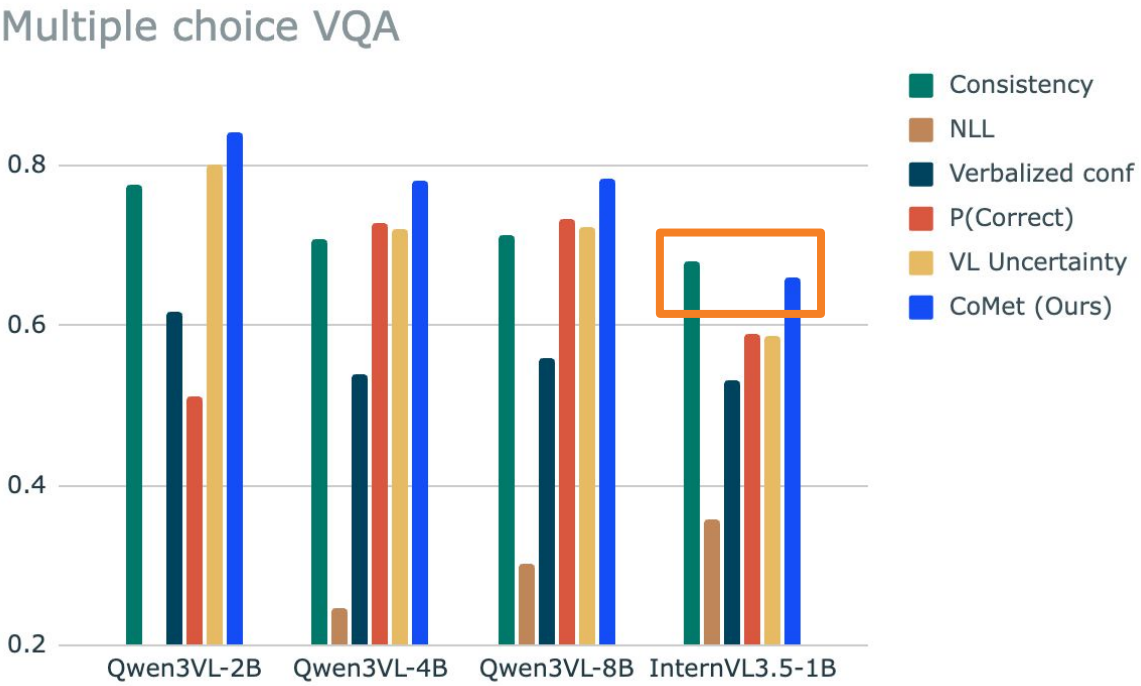
Results



Princeton Visual AI Lab

PRINCETON UNIVERSITY

Results



Princeton Visual AI Lab

PRINCETON UNIVERSITY

Ablation: Proposed decomposition helps better UQ

Objective	Avg AUROC	
u_t only	68.6	(-11.2)
$u_{x\|t}$ only	78.0	(-1.8)
$u_t + u_{x\|t}$ (CoMet)	**79.8**	

Backbone: Qwen3VL-8B-Instruct
Metric: Average AUROC

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Ablation: Sensitivity to the verifier model

Verifier model	Avg AUROC
Qwen3VL-2B (original)	**79.8**
Qwen3VL-8B	78.9 (-0.9)
InternVL3.5-1B	78.0 (-1.8)

Candidate-level disagreement:
Qwen-2B vs. Qwen-8B: 19.3%
Qwen-2B vs. InternVL: 35.0%

Aggregated Z1, Z2-level disagreement:
Qwen-2B vs. Qwen-8B: 22.6%
Qwen-2B vs. InternVL: 28.0%

Backbone: Qwen3VL-2B-Instruct
Metric: Average AUROC

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Ablation: Sensitivity to the candidate construction

Candidate source	# Questions	Avg AUROC
Cambrian clusters (proposed)	3k	78.6
Qwen3-8B generation	3k	79.2
Cambrian clusters (proposed)	200k	79.8

Backbone: Qwen3VL-2B-Instruct
Metric: Average AUROC

Generating 28 candidates for 3k questions required about 250 L40 GPU-hours; clustering 200k questions took under 2.5 hours on one L40

Ablation: Sensitivity to the candidate construction

Prompts for candidate construction study

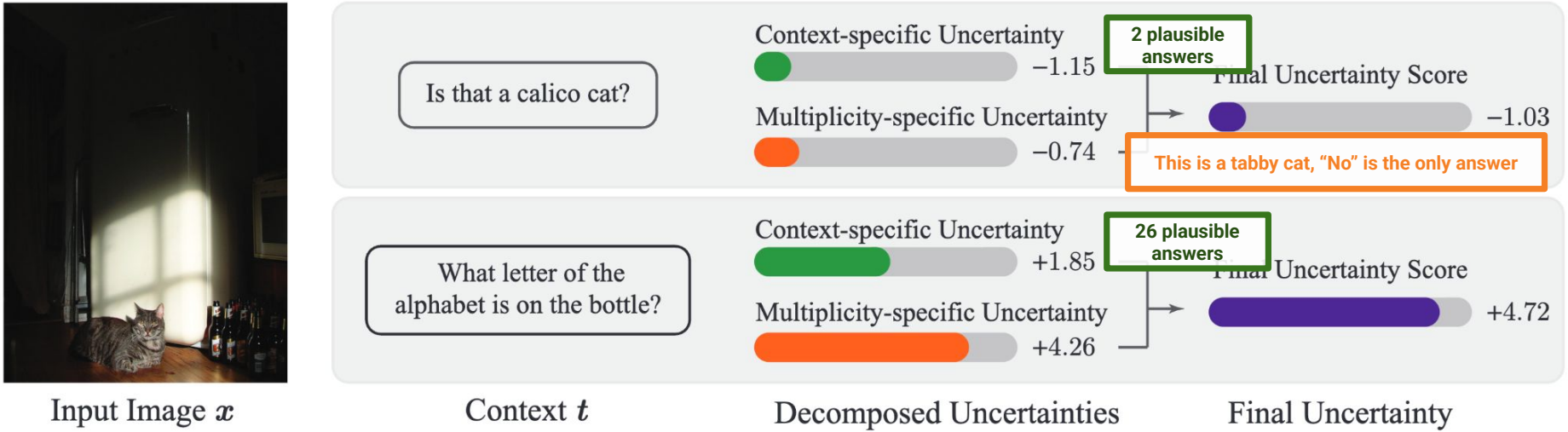
```
You are constructing plausible candidate answers for a visual question.
Generate exactly one short answer phrase. Do not explain. Do not output
multiple answers. Question: {t}, Known valid answers (examples, not
exhaustive): {in-context examples}. Generate one additional plausible
answer, preferably different from the examples above. Answer:
```

Generating 28 candidates for 3k questions required about 250 L40 GPU-hours; clustering 200k questions took under 2.5 hours on one L40

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Qualitative examples



Backbone: Qwen3VL-8B-Instruct

Outline

- Introduction: Uncertainty estimation in multimodal large language models (MLLMs)
- Existing approaches and challenges
- CoMet: Context and multiplicity decomposition
- Experiments
- **Conclusion**

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Conclusion

- **Takeaway:** Multimodal uncertainty can be decomposed into two components:
  - **Context-specific uncertainty:** “How broad is the answer candidate space induced by the context t?”
  - **Multiplicity-specific uncertainty:** “How many of those answer candidates remain plausible with x?”
- CoMet shows effective uncertainty estimation without generating answers at inference time, and providing interpretable and understandable uncertainty measures.

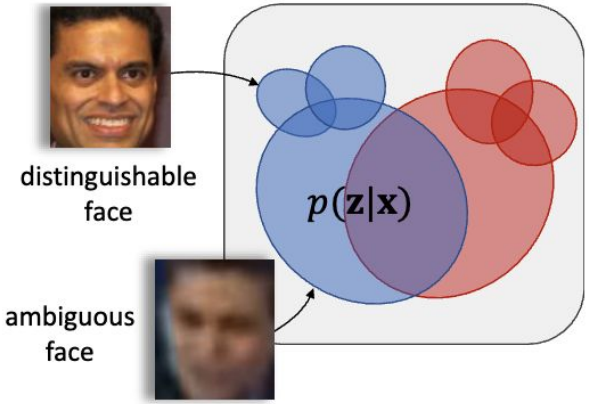
Thank you!
Any questions?

Appendix

Appendix
— Introduction

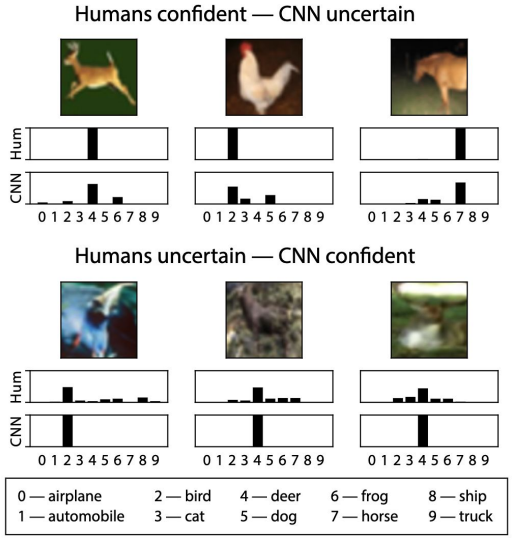
What is the source of unimodal uncertainty?

- In classification/retrieval, the source of uncertainty (what makes a model mispredict) frequently comes from the ambiguity or quality of the input
- Note that human disagreement and model uncertainty are not necessarily aligned



What color is the **mouse**?

"Mouse" may refer to either an animal or to a computer device



Luke Vilnis and Andrew McCallum. Word representations via gaussian embedding, ICLR 2015
Yichun Shi and Anil K. Jain, Probabilistic Face Embeddings, ICCV 2019
Joshua C. Peterson, Ruairidh M. Battleday, Thomas L. Griffiths, and **Olga Russakovsky**, Human uncertainty makes classification more robust, ICCV 2019

Back to the original slide

What should an uncertainty score do?

Correct



Is this taken during day or night?

Score: -5.88

Correct



How many zebras in the photo?

Score: -3.87

Correct



Should a school bus driver be taking pictures?

Score: 1.60

Incorrect



What are the words on the second line?

Score: 5.07

More certain

"Threshold"

More uncertain

Back to the original slide

Example applications of uncertainty scores



Question: The patient is a post-mature newborn. What is the most likely diagnosis?

- **Self-evaluation:** is the produced answer likely correct?
  (e.g., MLLM answer: “Meconium aspiration syndrome”, Is it correct?)
- **Abstention:** should the system defer or ask for more information?
  (e.g., Can we answer the question correctly?, or can say “unanswerable”?)
- **Hallucination detection:** is the answer visually supported?
  (e.g., “Meconium aspiration syndrome and *leg fracture*”)
- **Self-improvement:** which examples should be revisited?

Appendix
— CoMet derivation

CoMet uncertainty decomposition

Context and Multiplicity Decomposition (CoMet) for Multimodal Uncertainty Estimation

$$\pi_y := p(y \mid x, t, m = 1)$$

$$\text{uncertainty}(x, t) = \text{Entropy}\big(p(y \mid x, t, m = 1)\big)$$

$$:= \boxed{-\log \int \pi_y^2 dy} = -\log \int \frac{p(m = 1 \mid x, t, y)^2 p(y \mid t)^2}{\left(\mathbb{E}_{y' \sim p(\cdot \mid t)}\left[p(m = 1 \mid x, t, y')\right]\right)^2} dy$$

Rényi-2 entropy

$$= \underbrace{\log K_t}_{:= u_t(t)} + \underbrace{\left[-\log\left(\left\{\text{CV}_{Y \sim p(\cdot \mid t)}\left[p(m = 1 \mid x, t, Y)\right]\right\}^2 + 1\right)\right]}_{:= u_{x \mid t}(x \mid t)}$$

$$= \text{context-specific uncertainty} + \text{multiplicity-specific uncertainty}$$

$$p(y \mid t) = \frac{1}{\int_\Omega dy} := \frac{1}{K_t}.$$

$$\text{CV}(s) = \frac{\sqrt{\text{Var}(s)}}{\mathbb{E}(s)}$$

Back to the original slide

Rényi-2 entropy

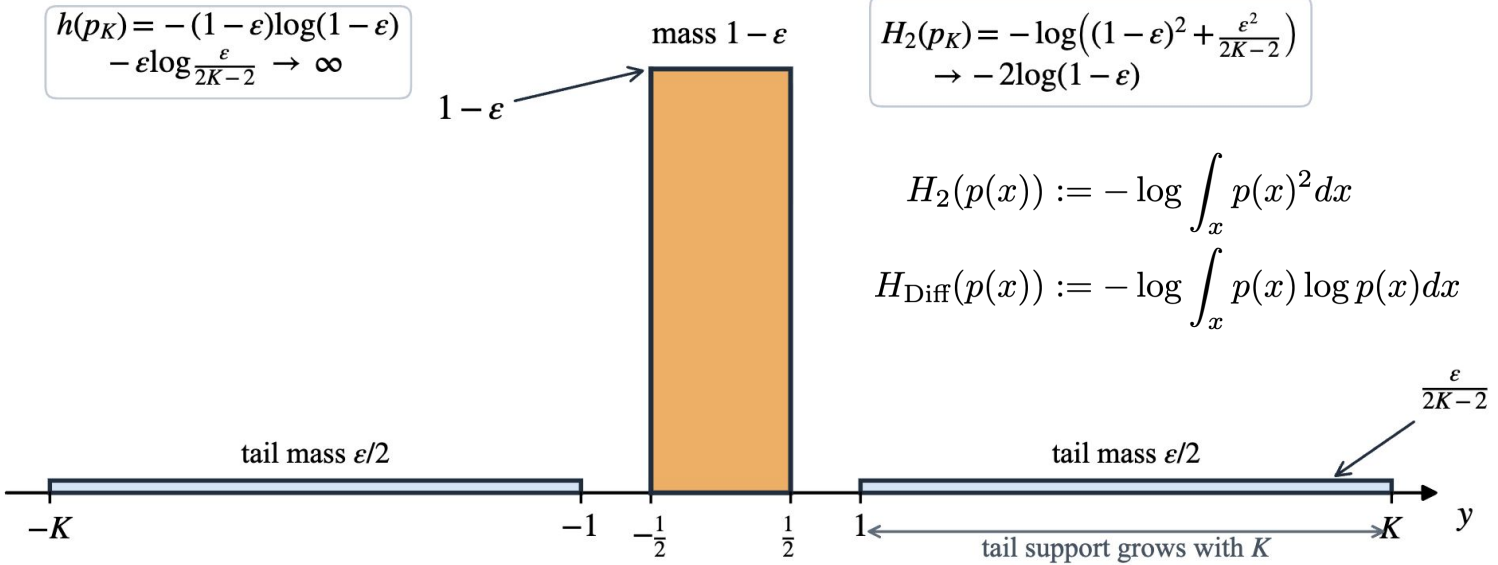
- It has merits over information entropy, specifically when we have continuous variables

$$h(p_K) = -(1-\varepsilon)\log(1-\varepsilon) - \varepsilon\log\frac{\varepsilon}{2K-2} \to \infty$$

$$H_2(p_K) = -\log\left((1-\varepsilon)^2 + \frac{\varepsilon^2}{2K-2}\right) \to -2\log(1-\varepsilon)$$

$$H_2(p(x)) := -\log\int_x p(x)^2 dx$$

$$H_{\text{Diff}}(p(x)) := -\log\int_x p(x)\log p(x) dx$$



Appendix
— CoMet training

Back to the original slide

Estimating matching probability using MLLMs

Prompts for matching probability verification

```
Image: {x}
Question: {t}
Proposed Answer: {y}
Only answer Yes if the image clearly supports the answer.
If the image is missing, unclear, or the answer is not visually confirmed,
answer No.
Answer Yes or No.
```

$$p(m = 1 \mid x, t, y) = \text{sigmoid}\left(\beta \log \frac{p(\text{“Yes”} \mid x, t, y)}{p(\text{“No”} \mid x, t, y)} + \gamma\right)$$

When beta = 1, gamma = 0

$$p(m = 1 \mid x, t, y) = \frac{p(\text{“Yes”})}{p(\text{“Yes”}) + p(\text{“No”})}$$

Beta and gamma are hyperparameters
(we use beta=0.2, gamma=−0.1)

Back to the original slide

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Why do we use the MLLM-as-verifier?

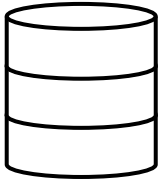
- We can train a matching probability estimator, which works well in retrieval/classification tasks [a, b, c]
- However, we found that training a reliable matching probability module is difficult
  - A matching prob module trained on a Instruction-tuning dataset is not generalizable to the other domains (e.g., open-ended VQA)
  - It easily suffers from the context shortcut (e.g., 84% of questions in the Cambrian dataset are unique, therefore, it is too easy to determine matching based on t, while ignoring x)

# Unique answers	1	2	3–5	6–10	11–20	21–50	>50
# samples	811,806	120,364	22,470	7,038	2,615	945	250

Chun, et al. Probabilistic embeddings for cross-modal retrieval, CVPR 2021
Chun, Improved probabilistic image-text representations, ICLR 2024
Chun, et al. Probabilistic language-image pre-training, ICLR 2025

Back to the original slide

An example clustering result



{ (x, t, y) }
(Images, Questions, Answers)



What is the main object in this image?

The main object in the image is a laptop computer.
The main object in the image is a red stop sign.
Boat
Counter
The setting of the image is a beach.
The image is set in a living room.
…

Back to the original slide

Approximating K_t

- We need to estimate the support of the p(y | t), which is a function of “t”

$$u_t(t) = \log K_t = -\log p(y \mid t) \underset{\text{(If } p(y \mid t) \text{ is uniform)}}{=} -\log \mathbb{E}_{Y \sim p(\cdot|t)}[p(y \mid t)] = H_2(p(y \mid t)).$$

$$\approx \widehat{u}_t(t) = -\log \sum_{y' \in \mathcal{Y}_t} p_\theta(y' \mid t)^2,$$

This can be computed by autoregressive log probability of LLM

Back to the original slide

Estimating uncertainty terms

- Context-specific term

$$u_t(t) = H_2(p(y \mid t)) \approx \widehat{u}_t(t) = -\log \sum_{y' \in \mathcal{Y}_t} p_\theta(y' \mid t)^2,$$

- Multiplicity-specific term

$$\widehat{Z}_1(x,t) = \frac{1}{|\mathcal{Y}_t|} \sum_{y' \in \mathcal{Y}_t} p(m=1 \mid x,t,y') \quad \widehat{Z}_2(x,t) = \frac{1}{|\mathcal{Y}_t|} \sum_{y' \in \mathcal{Y}_t} p(m=1 \mid x,t,y')^2.$$

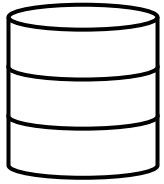
- Final objective function:

$$\widehat{u}(x,t) = \lambda \widehat{u}_t(t) + 2\log \widehat{Z}_1(x,t) - \log \widehat{Z}_2(x,t).$$

  - We use “lambda” term here (=0.25), to handle approximation error of u_t

Back to the original slide

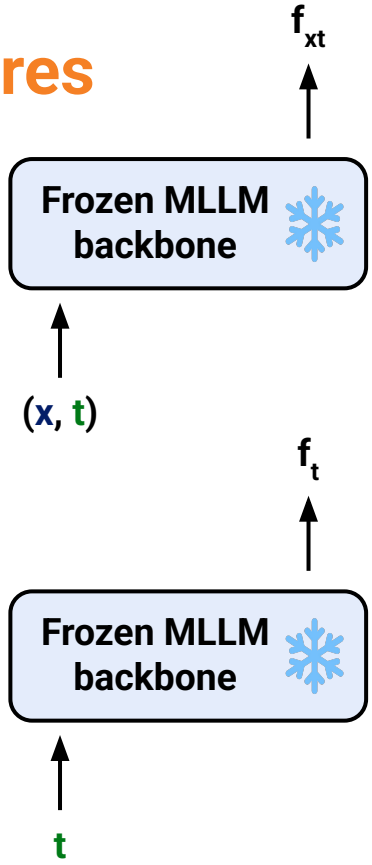
Pre-extracting features

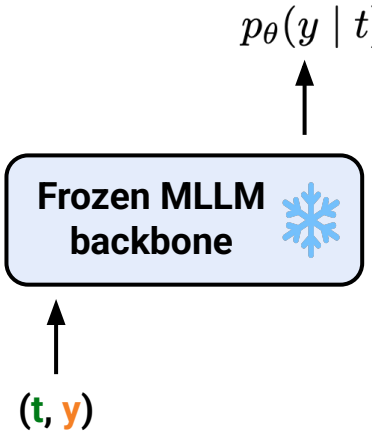


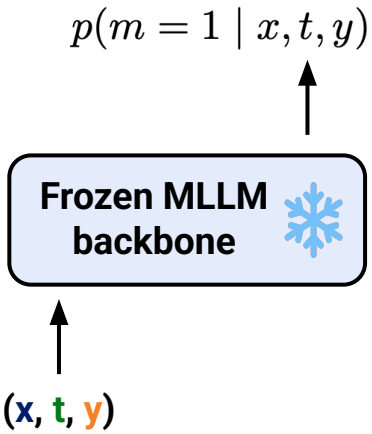
{ (x, t, y) }

Cambrian dataset [a]

200,000 (x, t, Y_t) rows
Each row has at most 32 candidates







Back to the original slide

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Lightweight adaptation

- Cambrian and the downstream benchmarks have very different distributions
- Also, CoMet does not use the actual correct/incorrect information from an MLLM
- We adapt the last layer of CoMet uncertainty module using the ranking loss

How many zebras in the photo?



(Correct answer)

The answer is “2”

My uncertainty score is “-3.87”

$$\mathbb{E}_{i\in\mathcal{I}^+, j\in\mathcal{I}^-}\left[-\text{softplus}(u'_j - u'_i)\right]$$

Apply a kind of contrastive loss to make incorrect samples have a higher uncertainty score

What are the words on the second line?



(Incorrect answer)

The answer is “MAGNUM”

My uncertainty score is “5.07”

Back to the original slide

Training setting

- Pre-training on the pseudo-candidates collected from the Cambrian dataset
  - 200,000 (image, question, candidate set) samples
  - Each sample has at most 32 candidates (randomly selected)
  - Using a regression loss with the proposed target (L1 loss)
  - Takes approximately 5~6 hours
- Adaptation on the downstream datasets
  - 900 MMMU validation samples
  - 10,000 VQA v2 held-out samples
  - Using a ranking loss with correct/incorrect information of the given MLLM (Contrastive loss using a softplus function)
  - Only training the last attention pooling layer and linear projection layer
  - Takes less than 10 minutes

Appendix
— Experiment settings

Back to the original slide

Training dataset statistics

- Pre-training uncertainty heads
  - 200,000 (x, t, Y_t) rows built from the Cambrian dataset
- Adapting uncertainty heads
  - 900 MMMU validation samples
  - 10k held-out VQA v2 samples

Back to the original slide

Benchmark dataset statistics

Table D.1: **Overview of evaluation benchmarks.** The detailed statics of the datasets are shown. We report the number of samples for each dataset and its answer type. We also report the performance of Qwen3VL-2B, -4B, -8B-Instruct models and InternVL3.5-1B model on each dataset. For open-ended VQAs, we report the original VQA score and separately report the threshold-based accuracy used in our experiments in the brackets. †: The original VQA v2 has 214,354 images, but we randomly select 10k samples for our evaluation. ‡: These samples are used for the adaptation. *: These samples are used for model selection and hyperparameter tuning.

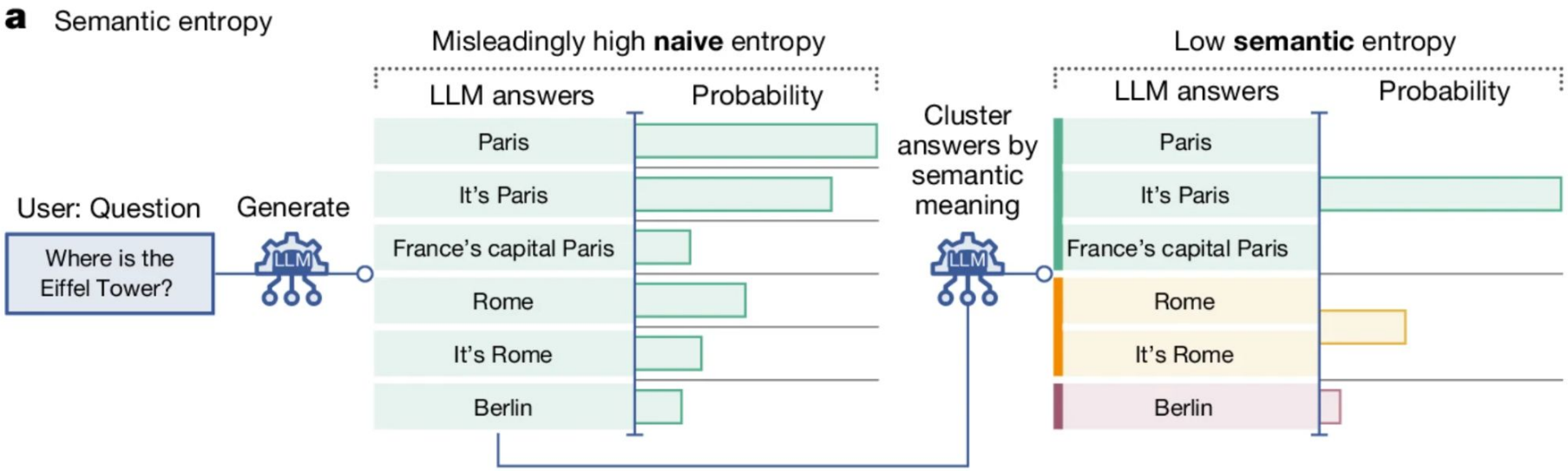
	Dataset stats		Performance			
	# Samples	Answer type	Qwen 2B	Qwen 4B	Qwen 8B	Intern 1B
VQA v2 (Validation) [22]	10,000†	Open-ended	62.2 (61.6)	82.8 (82.5)	71.6 (71.3)	76.5 (76.0)
VQA v2 (Adaptation)	10,000‡	Open-ended	-	-	-	-
Vizwiz (Validation) [23]	4319	Open-ended	38.9 (37.2)	46.4 (45.0)	40.8 (39.3)	41.2 (39.0)
OK-VQA (Validation) [24]	5046	Open-ended	46.1 (49.3)	61.2 (64.8)	55.4 (58.5)	54.9 (59.0)
HallusionBench [25]	1129	Yes / No	40.4	61.2	61.5	45.6
MMMU (Test) [26]	10500	MCQ	26	43.4	47.5	37.4
MMMU (Validation)	900‡*	MCQ	-	-	-	-
MMMU (Dev)	150*	MCQ	-	-	-	-
MMMU Pro [27]	1730	MCQ	20.6	32.2	35.9	28
MMStar [28]	1500	MCQ	41.5	48.3	49.3	39.1

Negative Log Likelihood, or Perplexity

$$\ell(x, t, y) := -\sum_{i=1}^{|y|} \log p_\theta(y_i \mid x, t, y_{<i})$$

We use the length-normalized version

“Consistency” (or semantic) Entropy



“Consistency” (or semantic) Entropy

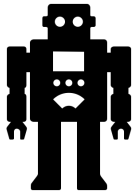
Farquhar, et al., Detecting hallucinations in large language models using semantic entropy, Nature 2024

P(Correct)



What are the words on the second line?

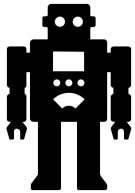




The answer is "MAGNUM"



Is it correct?

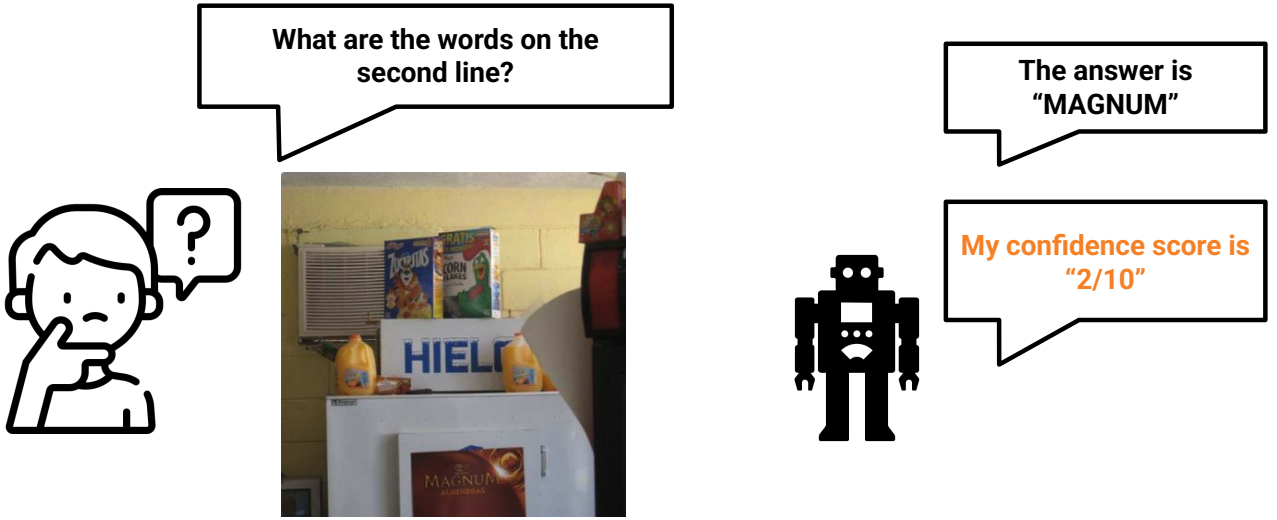


No

Use confidence score: $\frac{P(\text{“Yes”})}{P(\text{“Yes”}) + P(\text{“No”})}$

Kadavath, et al. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022).

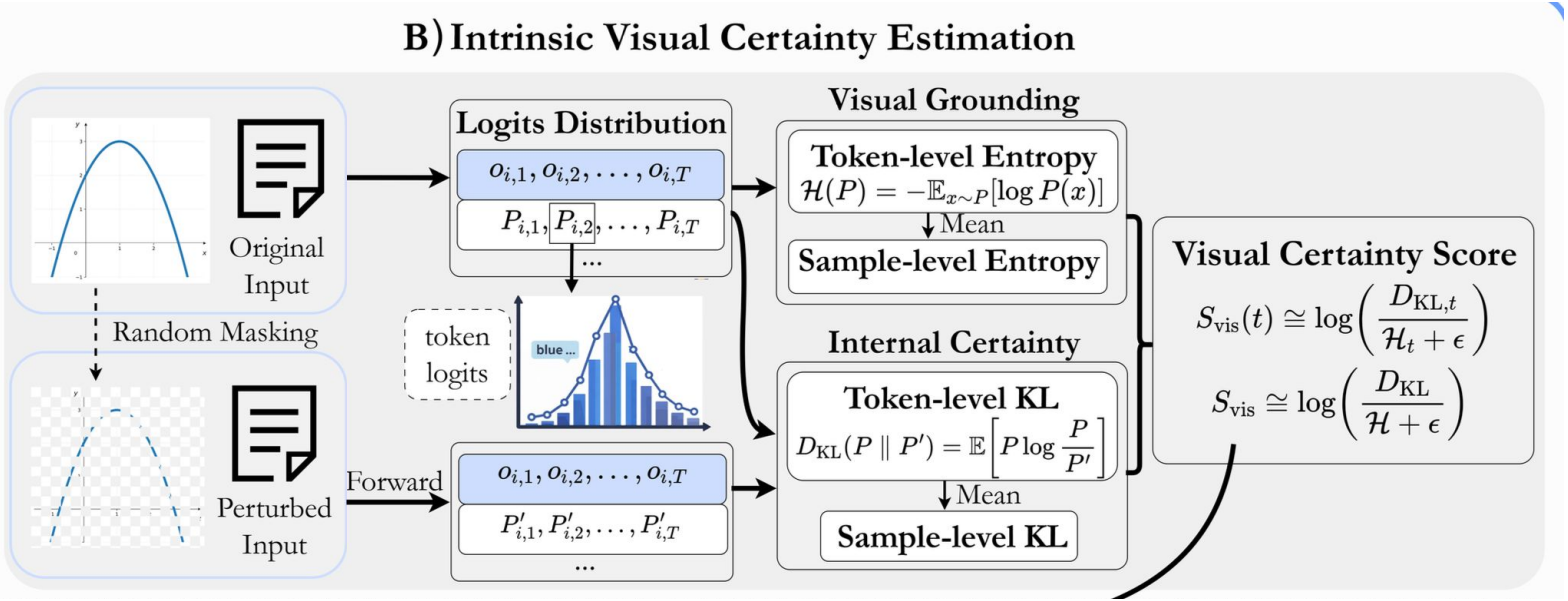
Verbalized confidence




```
Image: {x}
Question: {t}
Proposed Answer: {y}
How likely is the possible answer to be correct? Briefly explain, then write
only one score from 0 to 10 inside <confidence></confidence> tags.
```


Tian, et al. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. EMNLP 2023

VL Uncertainty



VL Uncertainty

Zhang, et al. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation, arXiv 2024

Back to the original slide

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Prompts for comparison methods

[Loglikelihood]

```
Image: {x}
Question: {t}
Proposed Answer: {y}
```

[Verbalized Confidence]

```
Image: {x}
Question: {t}
Proposed Answer: {y}
How likely is the possible answer to be correct? Briefly explain, then write
only one score from 0 to {confidence_scale} inside <confidence></confidence>
tags.
```

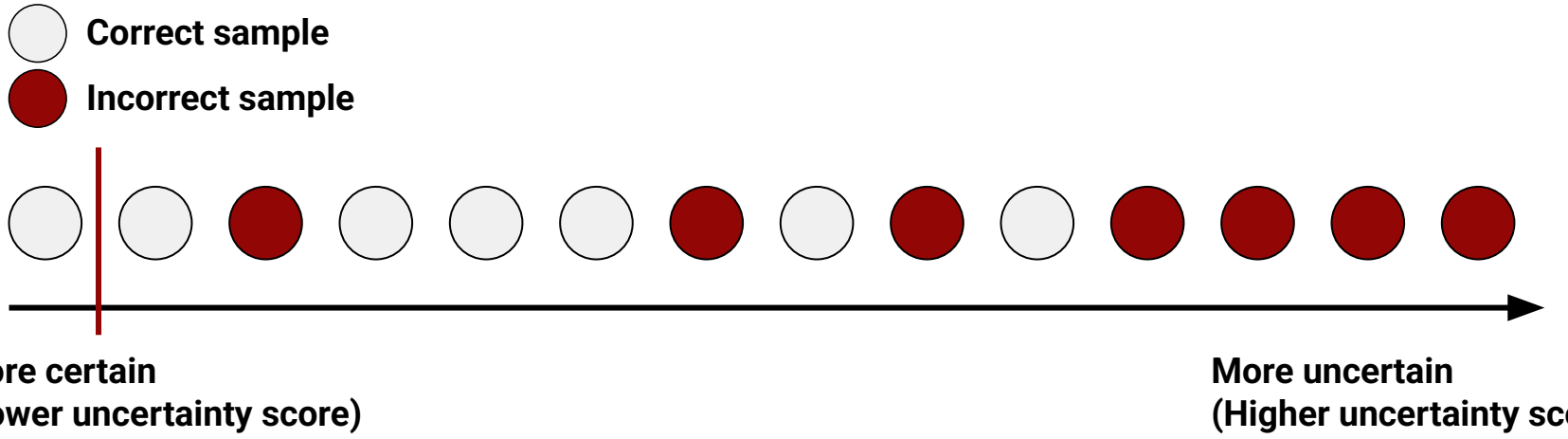
[P(Correct)]

```
Image: {x}
Question: {t}
Proposed Answer: {y}
Is the proposed answer correct?
(A) True
(B) False
The proposed answer is:
```

Back to the original slide

How can we evaluate uncertainty estimators?

- One score threshold leads to one true positive rate (TPR) / false positive rate (FPR)
- We change the threshold from the smallest to the largest values, and see the change of the TPR-FPR values, which is called as “area under the ROC curve”, or **AUROC**



Appendix
— Experiment results

Back to the original slide

CoMet gives better UQ than direct post-hoc training

Objective	Avg AUROC
BCE	77.2
Ranking	77.5
CoMet	**79.8**

$$\mathbb{E}_{i\in\mathcal{I}^+, j\in\mathcal{I}^-}\left[-\text{softplus}(u'_j - u'_i)\right]$$

Backbone: Qwen3VL-2B-Instruct
Metric: Average AUROC
Using the same features, architecture and adaptation labels
BCE, Ranking are trained from the scratch

Back to the original slide

Unified target vs. proposed decomposition

Objective	Avg AUROC
Shannon entropy	72.9
Renyi-2 entropy	73.5
CoMet	**79.8**

$$p_y = \mathrm{softmax}_y(s_y),$$

$$\text{where} \quad s_y := \log \frac{p(\text{“Yes”} \mid x, t, y)}{p(\text{“No”} \mid x, t, y)}$$

Backbone: Qwen3VL-2B-Instruct
Metric: Average AUROC
Using the same features and verifier labels

Back to the original slide

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Adaptation dataset

Adaptation	Avg AUROC
None	59.2
Cambrian source-only	68.5
VQA+MMMU (proposed)	79.8
VL-Uncertainty	67.8

Backbone: Qwen3VL-2B-Instruct
Metric: Average AUROC

Back to the original slide

Parameter study

- Relatively small impact on the choice of lambda and beta, gamma

$$u(x,t) = \lambda u_t(t) + u_{x|t}(x,t) \qquad p(m=1 \mid x,t,y) = \text{sigmoid}\left(\beta \log \frac{p(\text{“Yes”} \mid x,t,y)}{p(\text{“No”} \mid x,t,y)} + \gamma\right)$$

Table E.4: **Impact of λ.** We use Qwen3VL-8B-Instruct as the backbone. Hyperparameters are selected on the MMMU validation split (the first column).

λ	MMMU val	VQA v2	VizWiz (val)	OK-VQA	MMMU Test	MMMU Pro	MMStar	Average
0	.837	.834	.734	.735	.753	.707	**.857**	.780
0.25	**.842**	.836	.815	.746	.770	**.726**	.853	**.798**
0.5	.822	**.841**	.796	.741	**.772**	.715	.835	.789
1	.828	.839	**.821**	**.751**	.767	.707	.852	.795

Table E.5: **Impact of β, γ.** Details are the same as Tab. E.4.

β/γ	MMMU val	VQA v2	VizWiz (val)	OK-VQA	MMMU Test	MMMU Pro	MMStar	Average
0.2/-0.1	**.842**	.836	**.815**	**.746**	**.770**	.726	**.853**	**.798**
1/0	.835	**.837**	.796	**.746**	.760	.736	.832	.792

Back to the original slide

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Runtime comparison

Methods	Throuput (s)
Consistency	34.87
NLL	5.95
Verbalized Conf	6.10
P(Correct)	6.01
VL Uncertainty	16.33
CoMET (Ours)	**5.88**

Backbone: Qwen3VL-2B-Instruct
The number of maximum generation tokens is set to 4096

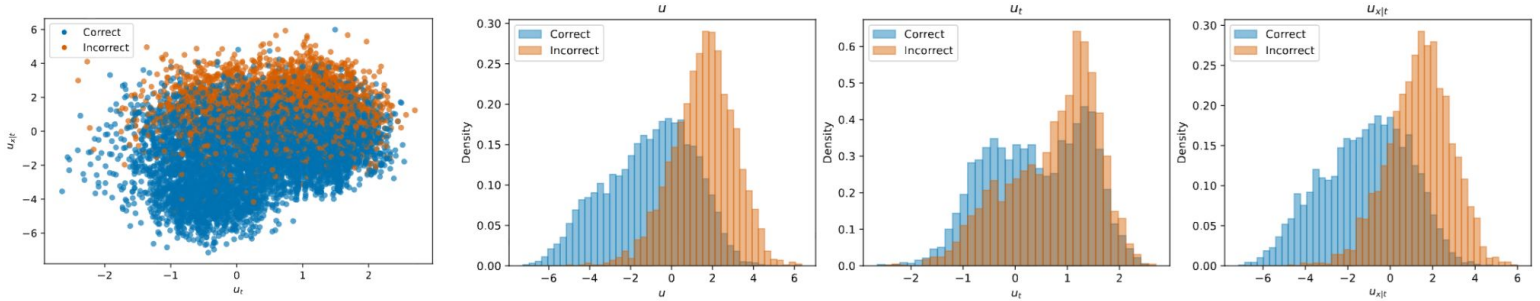
Back to the original slide

Additional parameters

Model	Backbone param (B)	Additional CoMet param (M)	Increase (%)
Qwen3VL-2B	2.13	50.4	+2.4%
Qwen3VL-4B	4.44	52.4	+1.2%
Qwen3VL-8B	8.77	58.3	+0.7%
InternVL3.5-1B	1.06	46.5	+4.4%

Back to the original slide

Uncertainty histogram



(a) u_t vs. $u_{x|t}$.

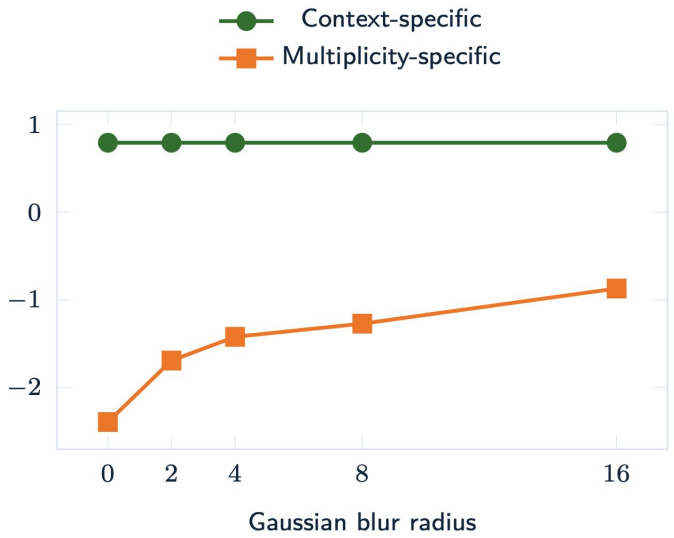
(b) Histogram of each uncertainty

Backbone: Qwen3VL-2B-Instruct
Dataset: VQA v2 (test samples)

Back to the original slide

Intervention example

- Fixing t, and blurring the image x leads to
  - Fixed u(t)
  - Increasing u(x | t)
- Tested on the VQA v2 dataset





Back to the original slide

Intervention example

- Fixing x, and expanding t by
  - Selecting “how many” questions from VQA v2 (525 samples)
  - Change the question as
    “{original question}. Answer with a single integer from 0 to {K-1}”
  - Larger K means the answer space becomes broader, i.e., it will have a larger u_t
  - This is unrealistic because the number of the plausible answers to (x, t) remains “1” regardless of K; in general, it is increased, therefore u_{xt} is getting increased by K

	$K = 20$	$K = 30$	$K = 40$
u_t	1.99	2.04	2.09
$u_{x\|t}$	-1.81	-1.90	-2.25

$$u_{x|t}(x \mid t) := -\log\left(\left\{\mathrm{CV}_{Y\sim p(\cdot|t)}[p(m=1 \mid x,t,Y)]\right\}^2 + 1\right)$$

(Increasing)
(decreasing)

$$\text{where, } \mathrm{CV}(s) = \frac{\sqrt{\mathrm{Var}(s)}}{\mathbb{E}(s)}$$

Back to the original slide

Intervention example

- Fixing x, and expanding t by
  - Selecting “how many” questions and “what color” from VQA v2 (1,809 samples)
  - Change the question as MCQ, such as
    “{original question}. Choices: {K number of candidates}”
  - E.g., What color is the floor? Choices: A. yellow B. brown C. red D. pink
  - u_{xt} is generally increased (because there could be “similar” colors in the choices)

	$K=2$	$K=4$	$K=8$	Original	
u_t	-.546	-.390	.018	1.786	(Increasing)
$u_{x\|t}$	-2.990	-2.404	-1.825	-2.469	(Increasing)

Back to the original slide

Most certain samples

Prediction: `Correct`
u: `-6.4965`
u_t: `0.1047`
$u_{x|t}$: `-6.5227`
Context t: Is this indoors or outdoors?
Input x:



Prediction: `Correct`
u: `-5.8762`
u_t: `0.1136`
$u_{x|t}$: `-5.9046`
Context t: Is this taken during day or night?
Input x:



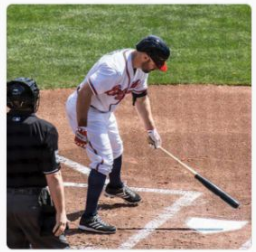
Prediction: `Correct`
u: `-5.0620`
u_t: `-0.1937`
$u_{x|t}$: `-5.0136`
Context t: Is this water or snow?
Input x:



Prediction: `Correct`
u: `-4.8878`
u_t: `1.4556`
$u_{x|t}$: `-5.2517`
Context t: How many people?
Input x:



Prediction: `Correct`
u: `-4.4962`
u_t: `2.1177`
$u_{x|t}$: `-5.0256`
Context t: What sport is this?
Input x:



Prediction: `Correct`
u: `-4.4546`
u_t: `0.3690`
$u_{x|t}$: `-4.5469`
Context t: How many zebras are in the photo?
Input x:



Sorted by its uncertainty (right is more uncertain)

Back to the original slide

Most uncertain samples

Prediction: Incorrect
u: 7.4382
u_t: 1.5318
$u_{x|t}$: 7.0553
Context t: How many calories?
Input x:



Prediction: Incorrect
u: 7.1284
u_t: 0.8759
$u_{x|t}$: 6.9094
Context t: How many train windows?
Input x:



Prediction: Incorrect
u: 7.0486
u_t: −0.6632
$u_{x|t}$: 7.2144
Context t: Do these people exercise a lot?
Input x:



Prediction: Correct
u: 6.9888
u_t: 1.4068
$u_{x|t}$: 6.6371
Context t: How many windows are there?
Input x:



Prediction: Correct
u: 6.7517
u_t: −0.6885
$u_{x|t}$: 6.9238
Context t: Are these children now grown up?
Input x:



Prediction: Correct
u: 6.7162
u_t: 1.6327
$u_{x|t}$: 6.3081
Context t: How many clay jars are there?
Input x:



Sorted by its uncertainty (left is more uncertain)

Back to the original slide

Princeton Visual AI Lab

PRINCETON UNIVERSITY

Same context but different images

Context t : **Was this photo taken recently?**

Context Uncertainty (u_t) : **−0.84**

Uncertainty (u) : **−1.28** Multiplicity Unc. ($u_{x\|t}$) : **−1.14**	Uncertainty (u) : **−1.20** Multiplicity Unc. ($u_{x\|t}$) : **−0.99**	Uncertainty (u) : **+4.34** Multiplicity Unc. ($u_{x\|t}$) : **+4.47**	Uncertainty (u) : **+4.46** Multiplicity Unc. ($u_{x\|t}$) : **+4.67**
			
Model Prediction: **Correct**	Model Prediction: **Correct**	Model Prediction: **Incorrect**	Model Prediction: **Incorrect**

Sorted by its uncertainty (right is more uncertain)

Back to the original slide

Same image but different contexts



Prediction	u	u_t	$u_{x\|t}$	Context t
Correct	-2.5025	-1.3541	-2.1640	Is the man skateboarding?
Incorrect	5.8100	1.2954	5.4861	How tall is the wave in this picture?



Prediction	u	u_t	$u_{x\|t}$	Context t
Correct	-2.7855	1.8233	-3.2413	How many animals?
Correct	-0.7946	-1.6813	-0.3743	Is the giraffe behind a fence?
Incorrect	0.5428	-0.9541	0.7813	Is this giraffe at the zoo?
Incorrect	5.0513	-0.5124	5.1794	How tall is the giraffe?

Appendix
— Limitation and future work

Back to the original slide

Limitations

- First, we need the uniform assumption for p(y | t), however,
  - p(y | t) itself could be non-uniform (e.g., “how many people in this image?”, the answers of this question might not be uniform)
  - When we approximate K_t, we actually break the uniform assumption; that is why we introduce lambda, but it is theoretically not rigorous
  - We can derive a non-uniform version of CoMet using p(y | t), but we found that this might introduce another model-specific bias to estimate p(y | t) during the training
- Second, CoMet needs the additional adaptation stage to achieve a high performance

Back to the original slide

Potential future directions

- Extending the problem to the other settings, such as
  - (Reasoning model)
    X: multimodal input
    T: reasoning trace
    Y: answer
  - (Agentic AI)
    X: environment or harness
    T: user input
    Y: answer
- The training-based CoMet might not work properly in these settings; we might need a training-free CoMet from the derivation we have