UOUO: Uncontextualized Uncommon Objects for Measuring Knowledge Horizons of Vision Language Models

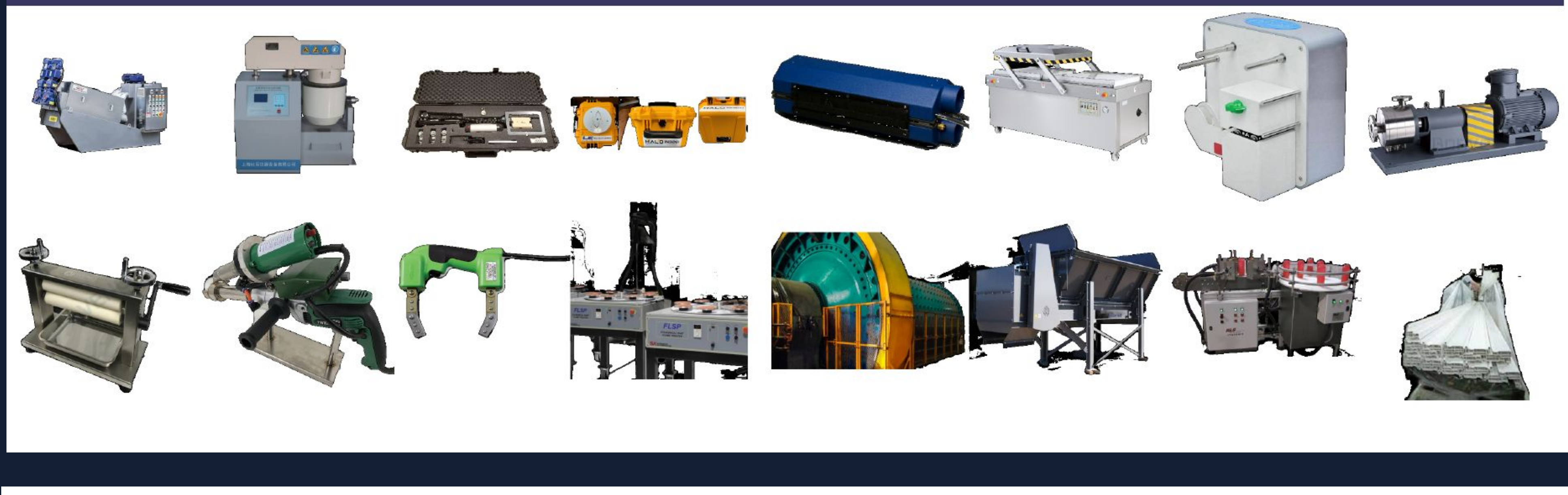
Xinyu Pi*, Mingyuan Wu*, Jize Jiang*, Haozhen Zheng*, Beitong Tian,
Chengxiang Zhai, Klara Nahrstedt, Zhiting Hu (* Indicates Equal Contribution)

Primary Contact: {mw34, jizej2, haozhen3}@illinois.edu, xpi@ucsd.edu

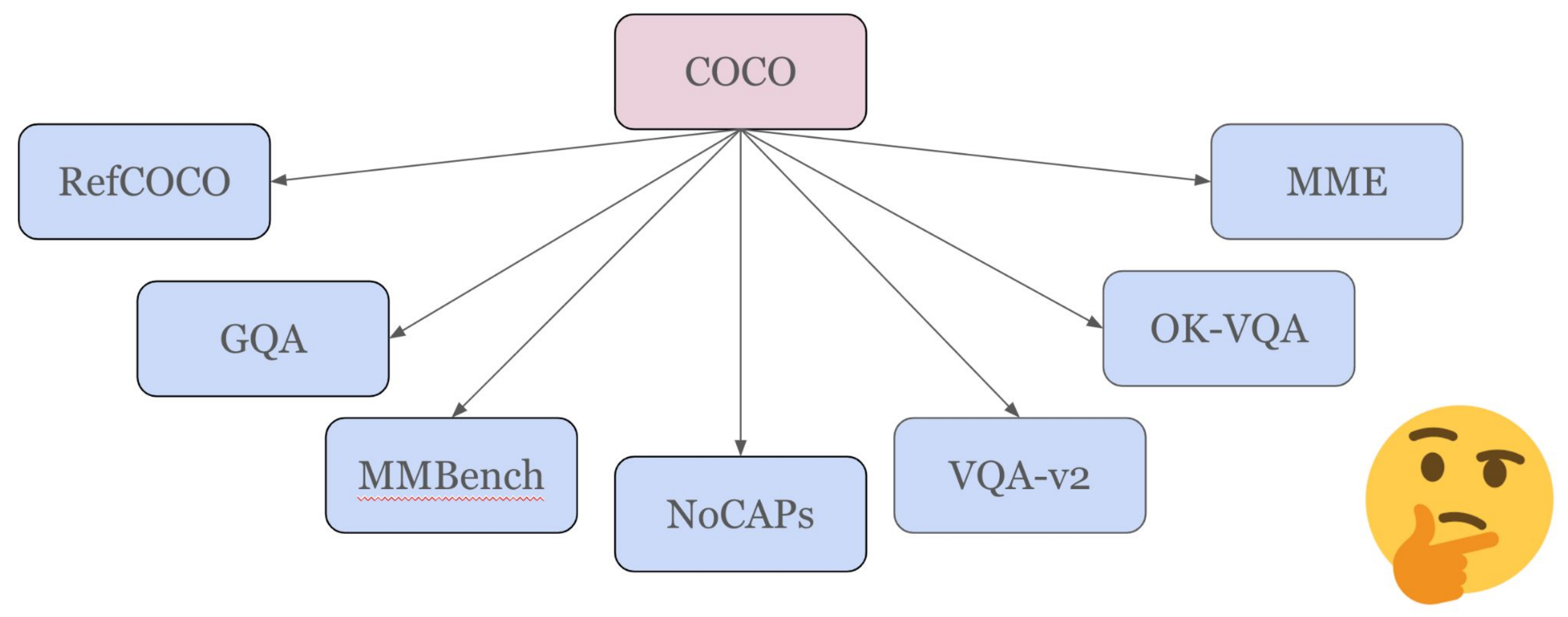
COCO vs. UOUO (Long tail, Uncommon)

COCO

UOUO



No VLM evaluation in *uncommon*



COCO

RefCOCO

MME

GQA

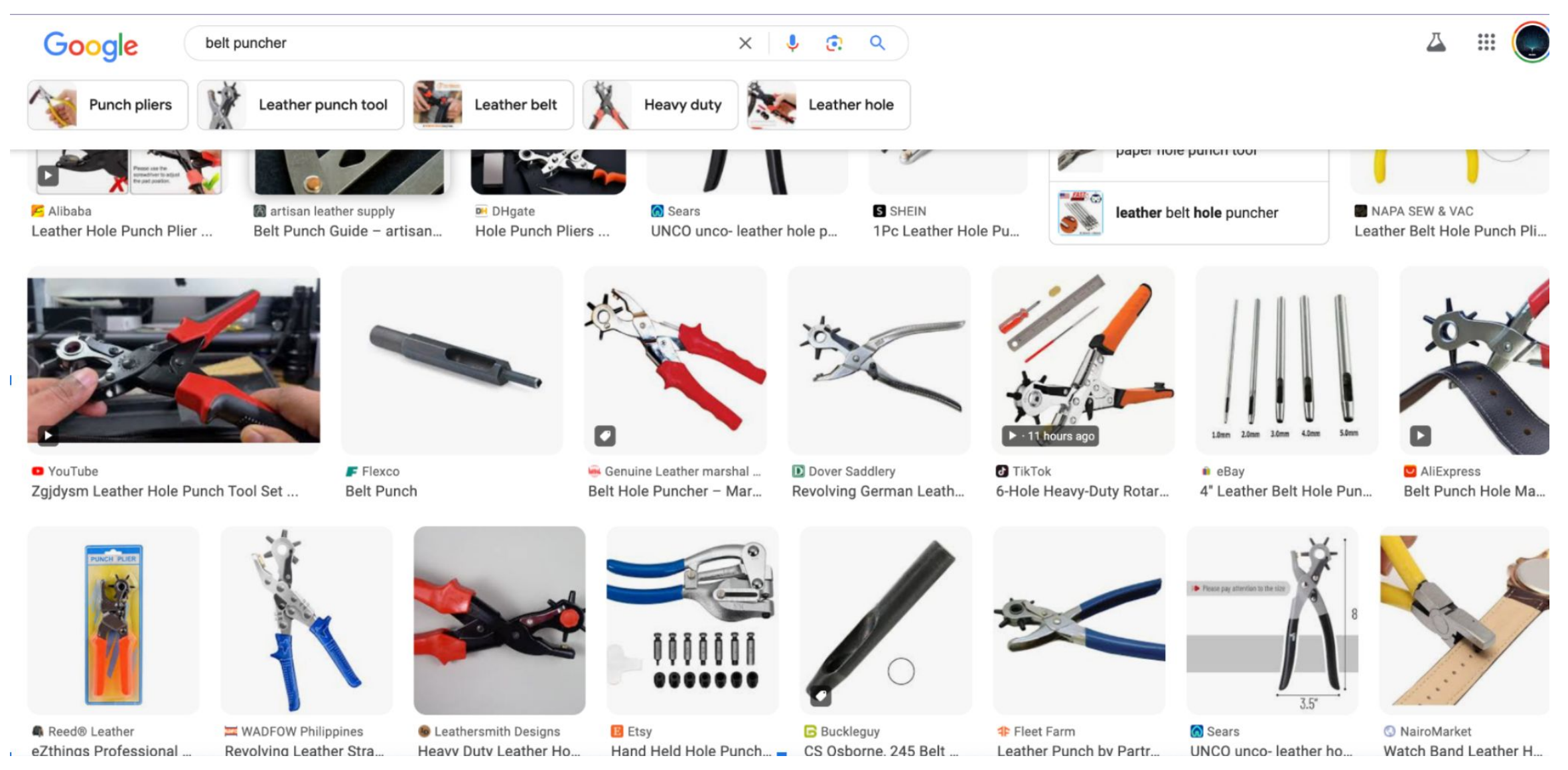
OK-VQA

MMBench

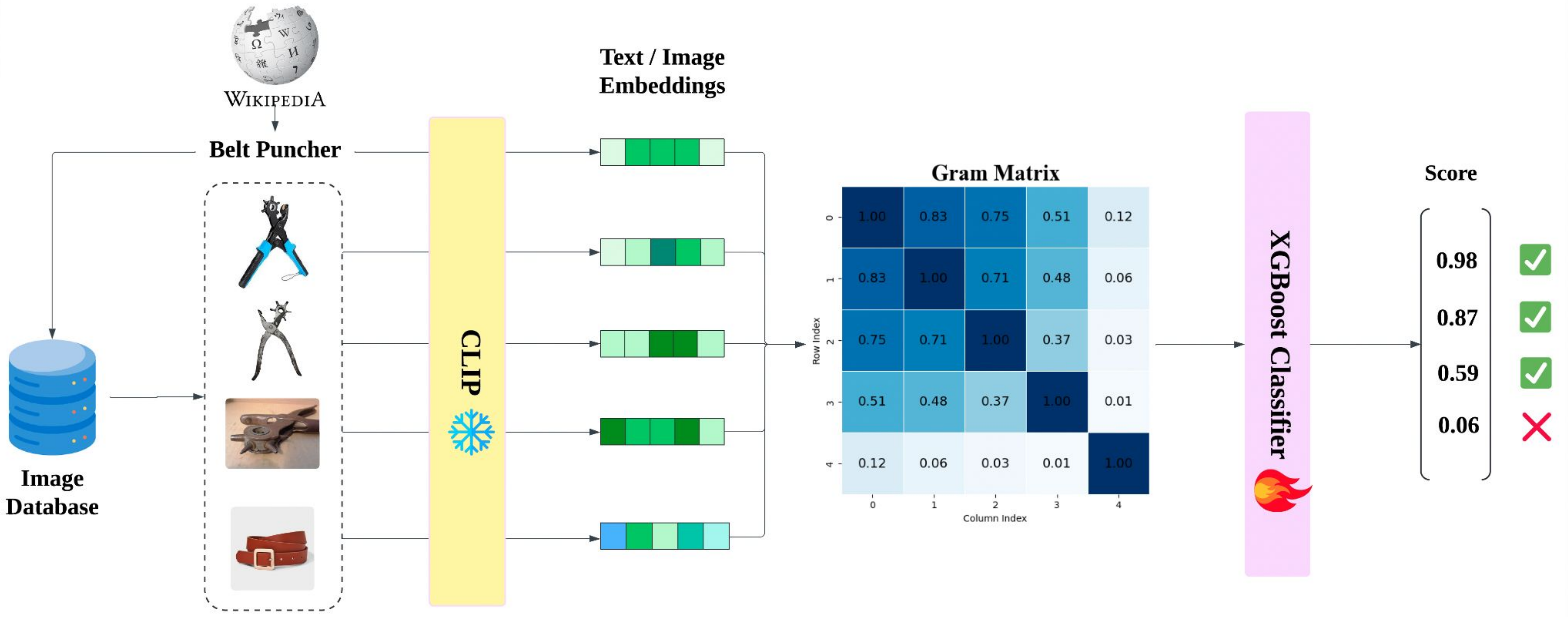
NoCAPs

VQA-v2

Web scraping -> uncommon object



Automatic Data Curation



Wikipedia

Belt Puncher

Image Database

CLIP

Text / Image Embeddings

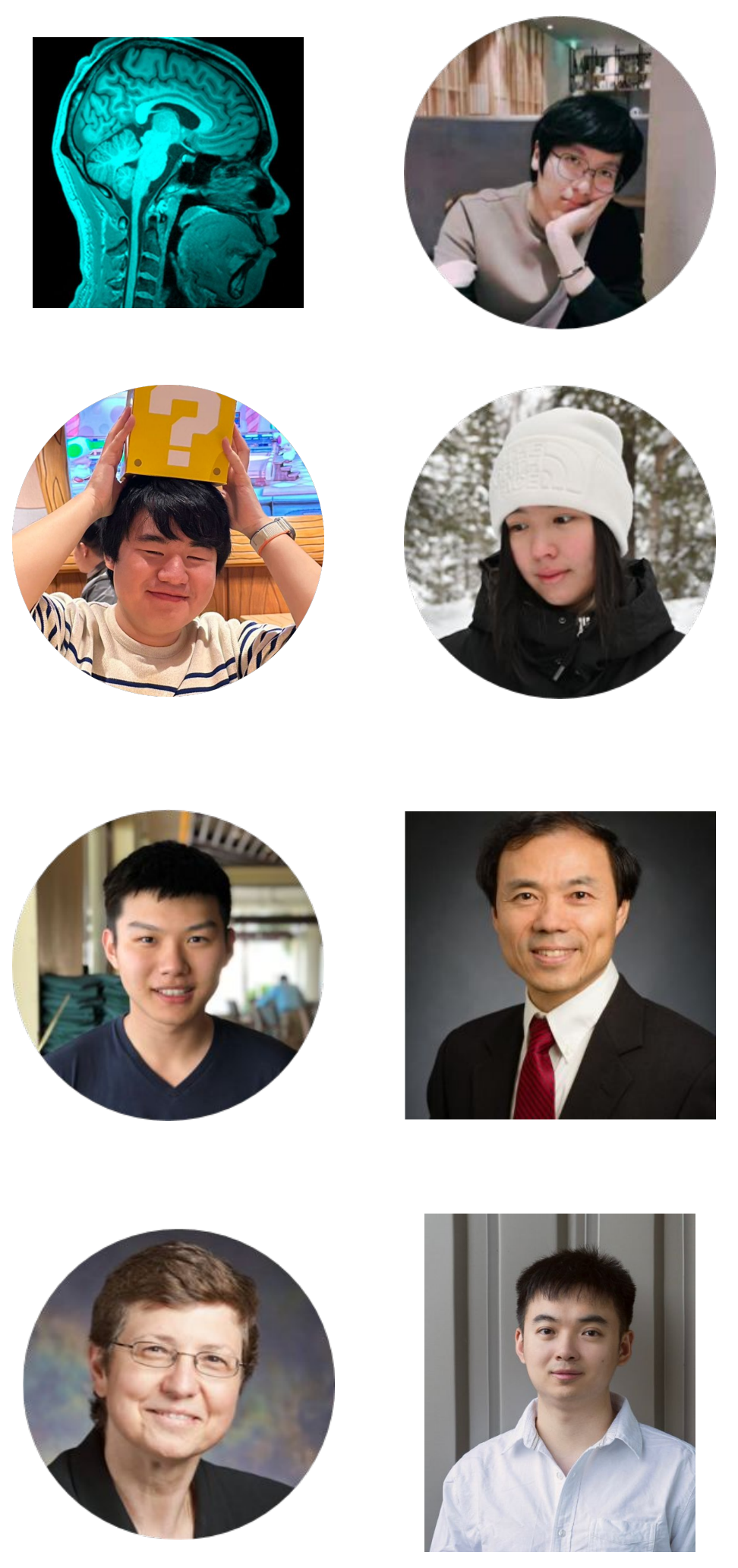
Gram Matrix

	0	1	2	3	4
0	1.00	0.83	0.75	0.51	0.12
1	0.83	1.00	0.71	0.48	0.06
2	0.75	0.71	1.00	0.37	0.03
3	0.51	0.48	0.37	1.00	0.01
4	0.12	0.06	0.03	0.01	1.00

XGBoost Classifier

Score	
0.98	✓
0.87	✓
0.59	✓
0.06	✗

Profile Photo

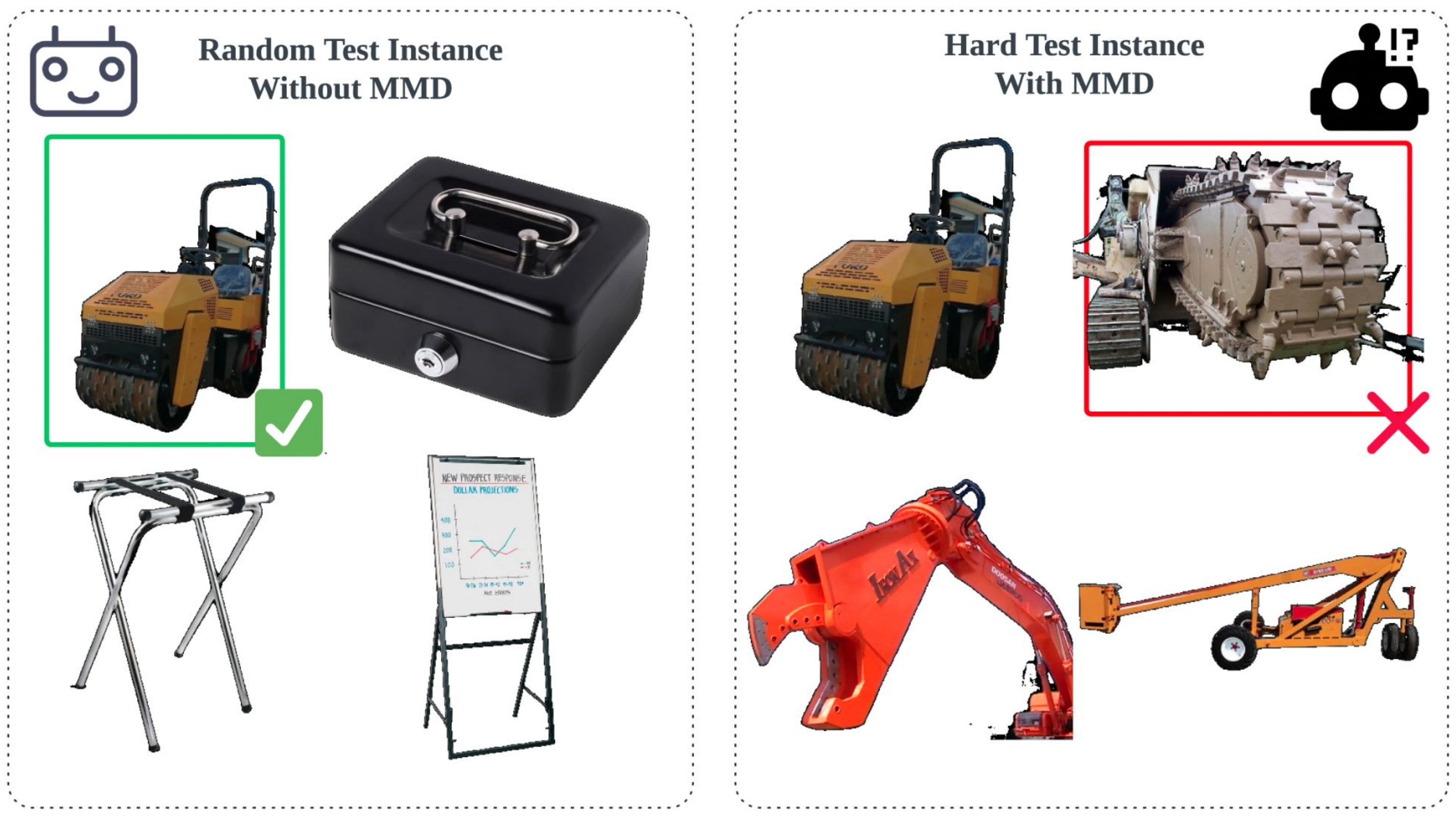


Question Generation -> Grounding

User

Please provide the bounding box coordinate(x1,y1,x2,y2) of the **road roller** in the image.

Identify the location of the **road roller** from 'top left', 'top right', 'bottom left', 'bottom right', or none



$$\mathrm{MMD}(\mathbf{x},\mathbf{y}) = k(\mathbf{x},\mathbf{x}) + k(\mathbf{y},\mathbf{y}) - 2 \cdot k(\mathbf{x},\mathbf{y})$$

Where x and y be the sets of CLIP embeddings for two different object categories, where k is a kernel function (We adopt Gaussian RBF)

Experiment Results

Model Name	mIoU-mmd	mIoU-rand	acc-mmd	acc-rand
llava-v1.5-7b	0.18	0.41	0.42	0.70
llava-v1.5-13b	0.23	0.47	0.44	0.73
llava-v1.6-vicuna-7b	0.28	0.48	0.49	0.75
llava-v1.6-vicuna-13b	0.28	0.49	0.52	0.78
llava-v1.6-34b	0.38	**0.55**	0.57	0.83
cogvlm-llama3-chat-19b	**0.49**	0.69	0.43	0.60
gemini-1.5-pro	0.27	0.27	0.63	0.80
gpt-4-turbo	0.34	0.38	0.67	**0.90**
gpt-4o	0.33	0.35	**0.68**	0.88

Take away

- **Smaller VLMs struggle with uncommon objects, especially in low MMD mosaics.**
- **A specialized dataset for benchmarking VLMs on uncommon objects.**
- **An automatic pipeline for web data scraping, curation, filtering, and generating challenging test instances for domain-specific objects**

NSF ILLINOIS