Revisiting Human-in-the-Loop Object Retrieval with Pre-trained Vision Transformers

Kawtar Zaher, Olivier Buisson, and Alexis Joly

Motivation

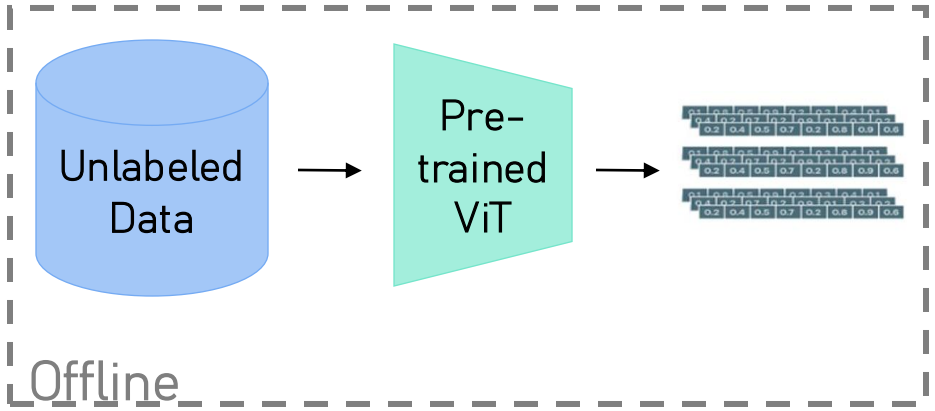
- **Context:** Audiovisual archives with 22 million hours of data, no labels.

- **Objective:** A user looking for a specific concept starting from a query, e.g. a car from a specific era.

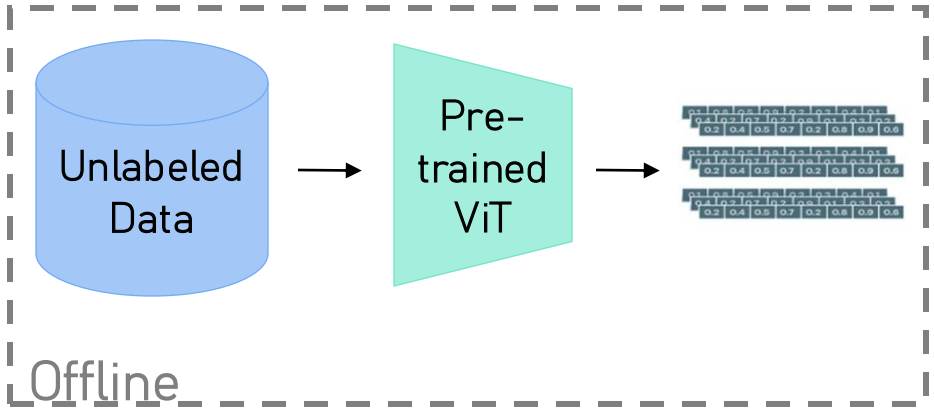
- **Limitation:** A single retrieval pass based on similarity search returns near-duplicates or doesn't recognize the concept.

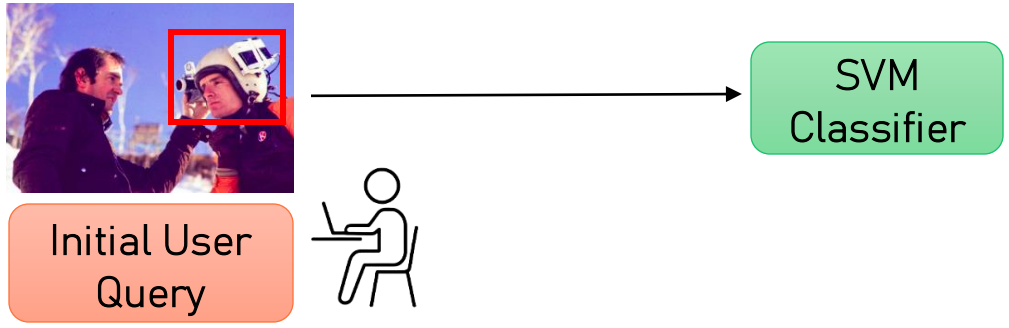
> Need for iterative refinement with user interaction.

Interactive Retrieval Process

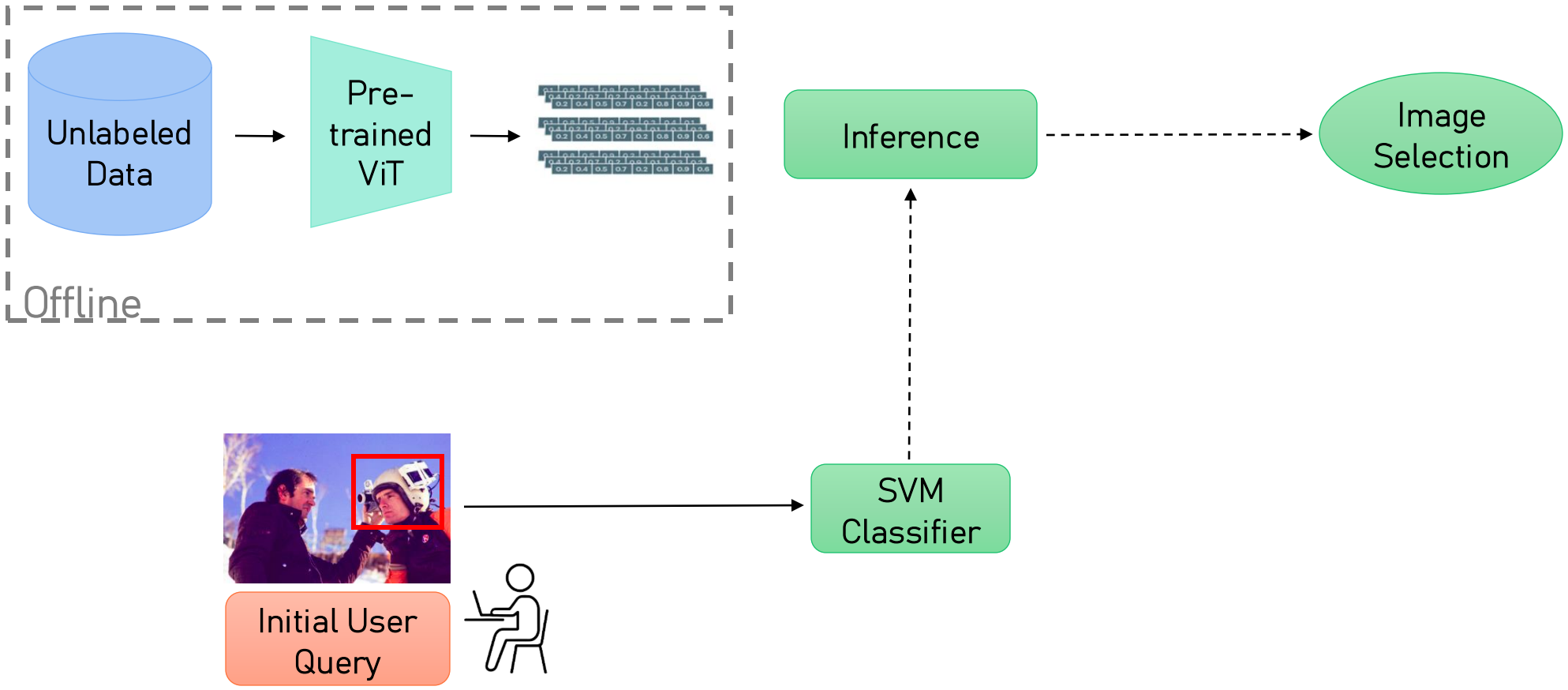


Interactive Retrieval Process

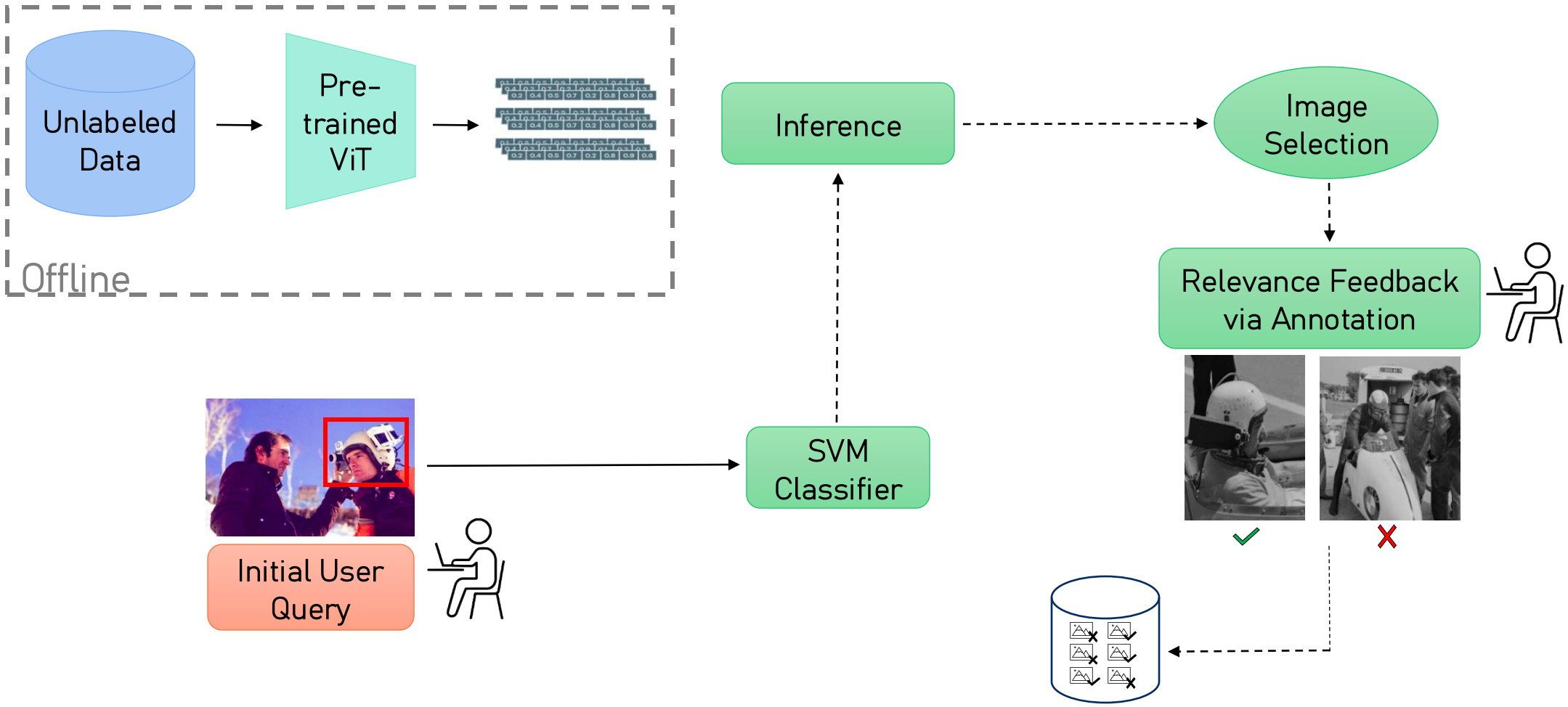




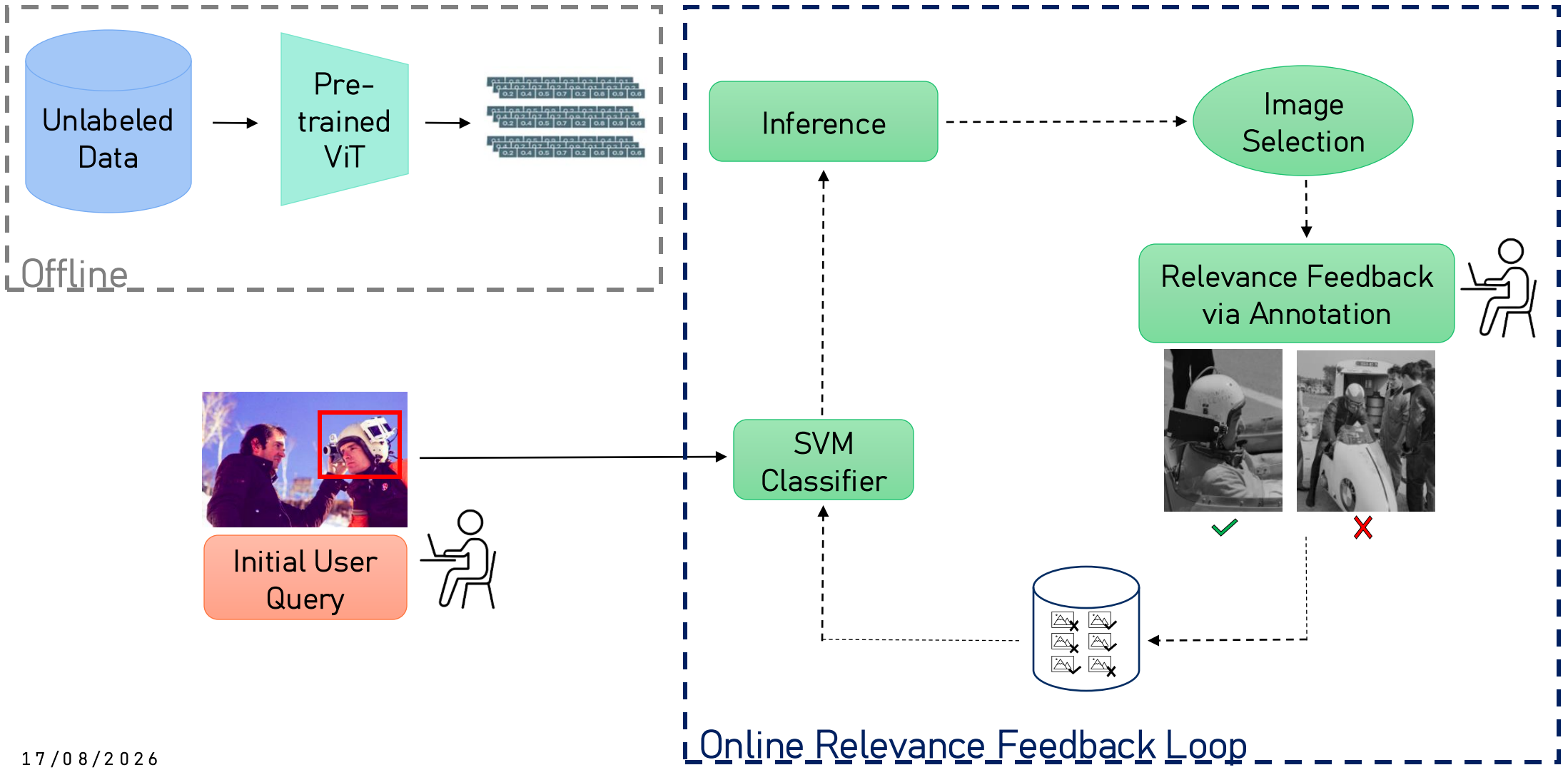
Interactive Retrieval Process



Interactive Retrieval Process



Interactive Retrieval Process

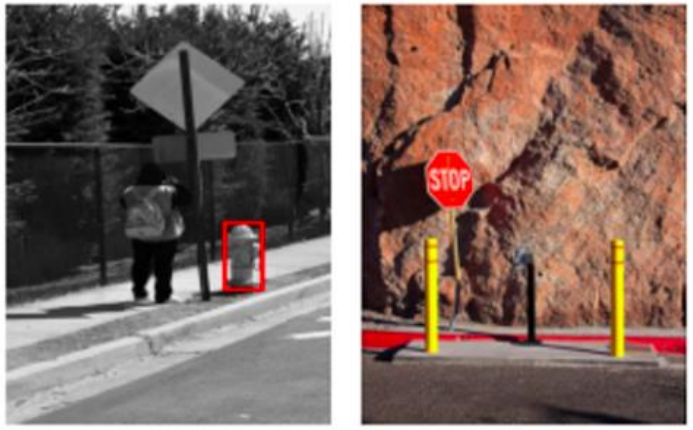


Which Image Descriptor?

User Query Global

cooking

User Query Global



road

- **Global descriptor (cls token):** Global context with loss of finer details.

> Need for adequate descriptors for small objects within cluttered scenes.

Representation Strategies – Local Only

- Split the image into M regular crops (2x2 or 4x4).

- Extract each crop's descriptor:
  - Feed forward the crop into the network (costly).
  - Use the mean of corresponding ViT patch tokens (cheap).

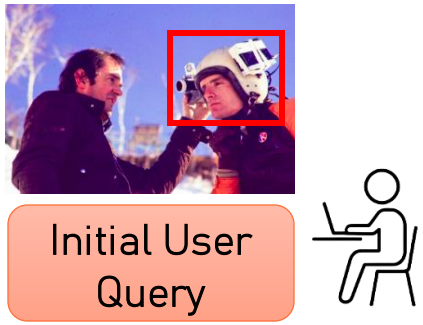
> Each crop counts as a sample?

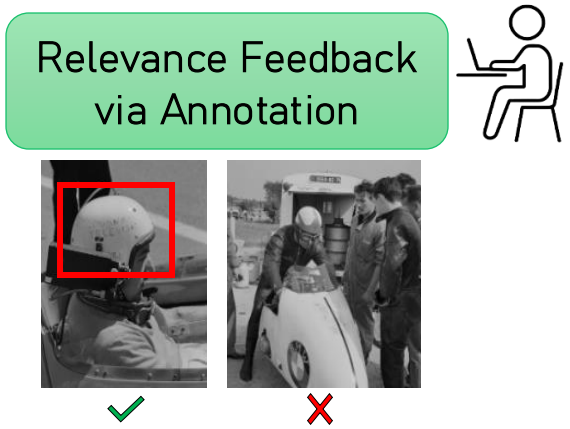
> Only choose one crop as a sample?

> How to choose this crop?

Representation Strategies – Local Only

- Choose the crop based on a new form of user annotations: bounding boxes.





Representation Strategies – Local Only

- Choose the crop based on a new form of user annotations: bounding boxes.

- One crop: max overlap with object – **LO-One**.
- All crops: if overlap (+); else (-) – **LO-All**.

- One crop: random or mean – **LO-One-Rand/Proto**
- All crops: (-) – **LO-All**.

Representation Strategies – Global Local

- Same construction for local descriptors.

- Add global information (cls token) via :
  - Concatenation – **GL-concat**.
  - Pooling – **GL-pool**.

- Fusion with:
  - One crop – **GL-*-One-Rand/Proto**.
  - All crops – **GL-*-All**.

Image Selection for User RF

- How to select the images to return to the user for annotation ?
- Consider an Active Learning criterion, return images that score the maximum.
- **Multiple crops:** score each crop, then consider the maximum crop score as the image score.

> If one crop fires, the image should fire.

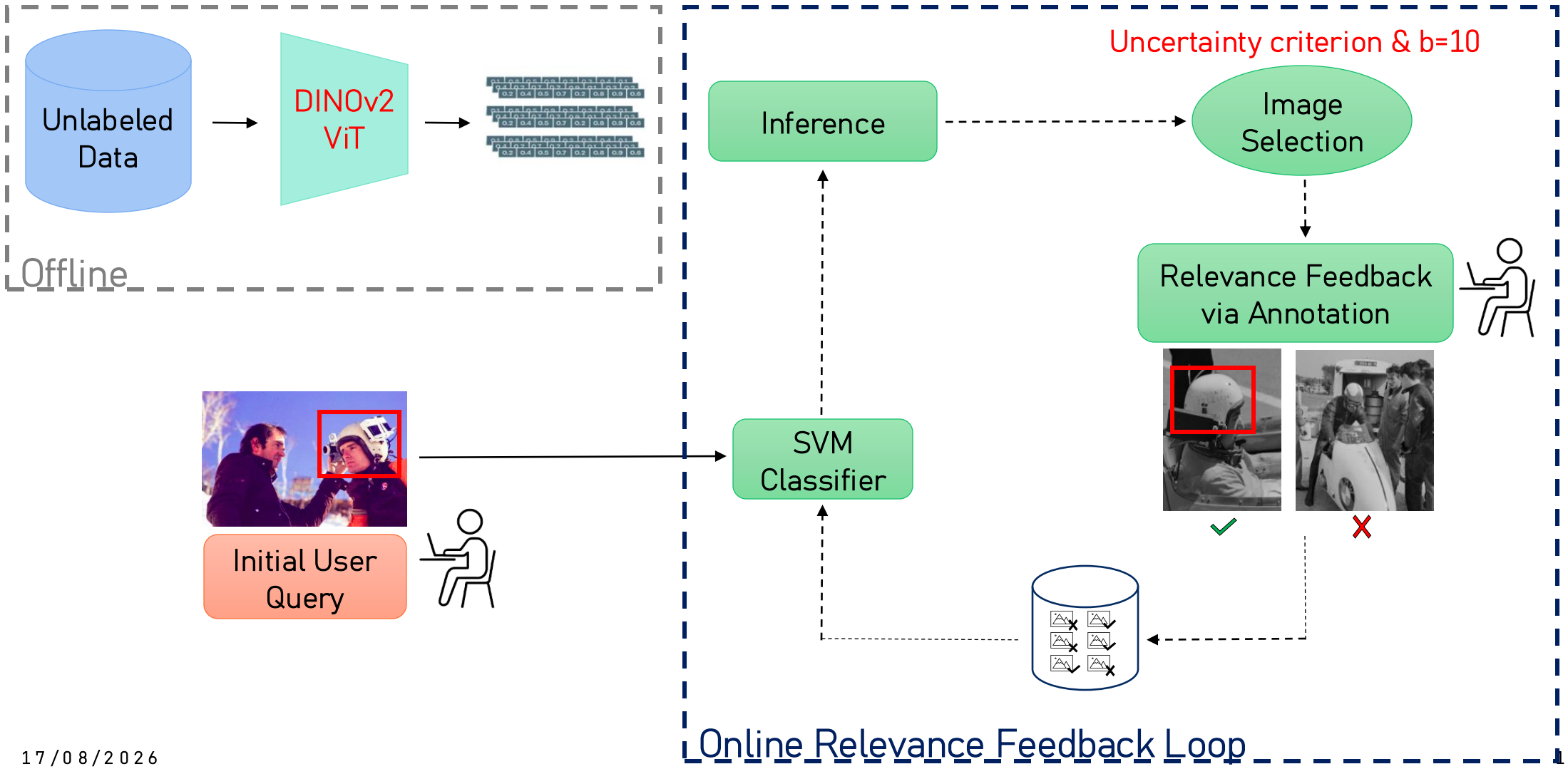
Evaluation

- Evaluate both the classifier quality and the user satisfaction.

- Classifier quality:
  - **Mean Average Precision score:** can the classifier retrieve all relevant images first?

- User satisfaction:
  - **Class Coverage score:** do the selected images cover all class modes or are they only duplicates?

Interactive Retrieval Process



Experiments – choosing the local strategy

Representation with M=4 (2x2) crops		PascalVOC2012			Coco2017		
		mAP	Cov	sum	mAP	Cov	sum
Global		0.829	0.652	-	0.628	0.522	-
Feed forward crops	***LO-One-Proto***	0.73 (-11.94%)	0.142 (-78.22%)	-90.16%	0.489 (-22.13%)	0.095 (-81.8%)	-103.9%
	LO-One-Rand	0.866 (4.46%)	0.618 (-5.21%)	-0.75%	**0.683 (8.76%)**	0.53 (1.53%)	10.29%
	LO-All	0.865 (4.34%)	0.634 (-2.76%)	1.58%	0.651 (3.66%)	0.574 (9.96%)	13.62%
Avg patch tokens	***LO-One-Proto***	0.668 (-19.42%)	0.143 (-78.07%)	-97.49%	0.403 (-35.83%)	0.093 (-82.18%)	-118%
	LO-One-Rand	0.871 (5.07%)	0.597 (-8.44%)	-3.37%	0.663 (5.57%)	0.514 (-1.53%)	4.04%
	LO-All	**0.899 (8.44%)**	**0.642 (-1.53%)**	**6.91%**	0.676 (7.64%)	**0.593 (13.6%)**	**21.24%**

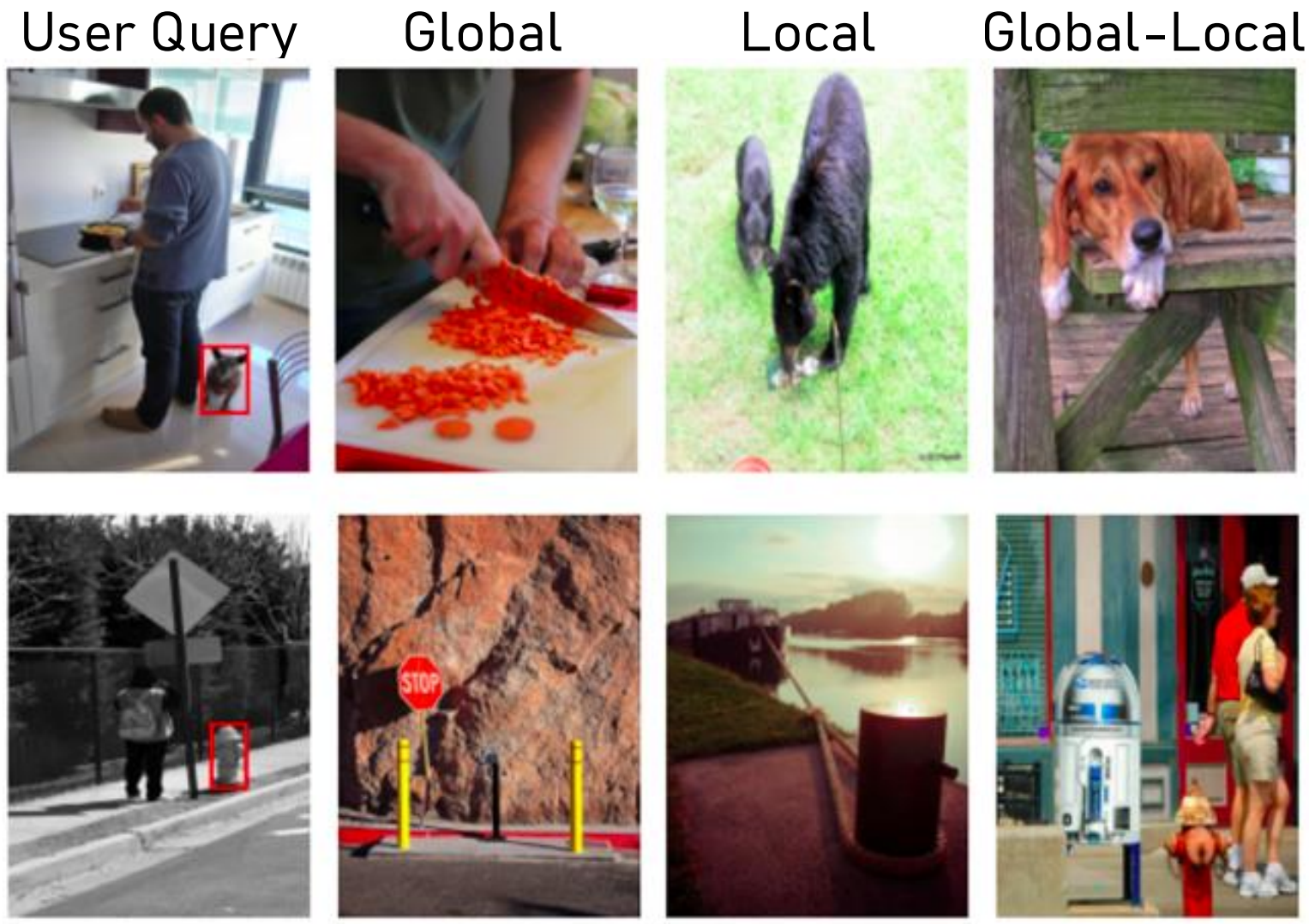
Experiments – choosing the strategy

Representation with avg patch tokens		PascalVOC2012			Coco2017		
		mAP	Cov	sum	mAP	Cov	sum
Global		0.829	0.652	-	0.628	0.522	-
4 crops (2x2)	***LO-All***	0.899 (8.44%)	0.642 (-1.53%)	6.91%	0.676 (7.64%)	0.593 (13.6%)	21.24%
	GL-concat-All	0.9 (8.56%)	0.674 (3.37%)	11.93%	0.714 (13.69%)	0.566 (8.43%)	22.12%
	GL-pool-All	0.889 (7.24%)	**0.695 (6.6%)**	**13.84%**	0.674 (7.32%)	0.596 (14.18%)	21.5%
16 crops (4x4)	***LO-All***	**0.913 (10.13%)**	0.559 (-14.26%)	-4.13%	0.711 (13.22%)	0.576 (10.34%)	23.56%
	GL-concat-All	0.907 (9.41%)	0.652 (0.0%)	9.41%	**0.732 (16.56%)**	0.564 (8.05%)	24.61%
	GL-pool-All	0.896 (8.08%)	0.651 (-0.15%)	7.93%	0.69 (9.87%)	**0.603 (15.52%)**	**25.39%**

Experiments – objects sizes

	M = 2x2 = 4	M = 4x4 = 16
Object Size	**PascalVOC2012**	**Coco2017**
Small (<1/16 image)	44%	74.7%
Medium (1/16 - 1/4 image)	27.3%	15.9%
Large (>1/4 image)	28.7%	9.4%

Visual Results



Key Takeways

- Hybrid Global-Local descriptors guarantee global context & fine-grained details.
- Concatenation favours precision, pooling favours diversity.
- Crop granularity is data-dependent.

Revisiting Human-in-the-Loop Object Retrieval with Pre-trained Vision Transformers

Thank you! Any Questions?