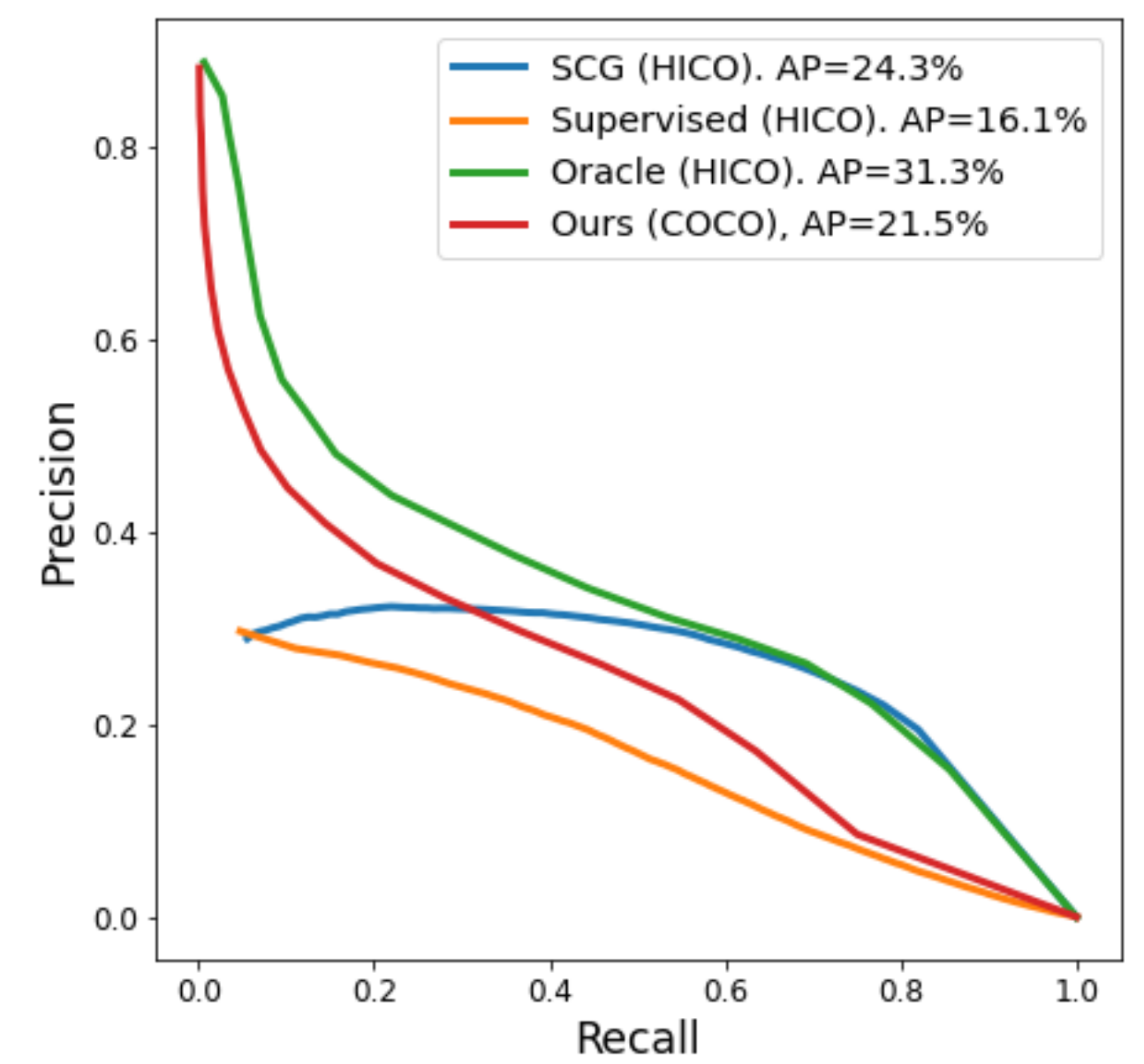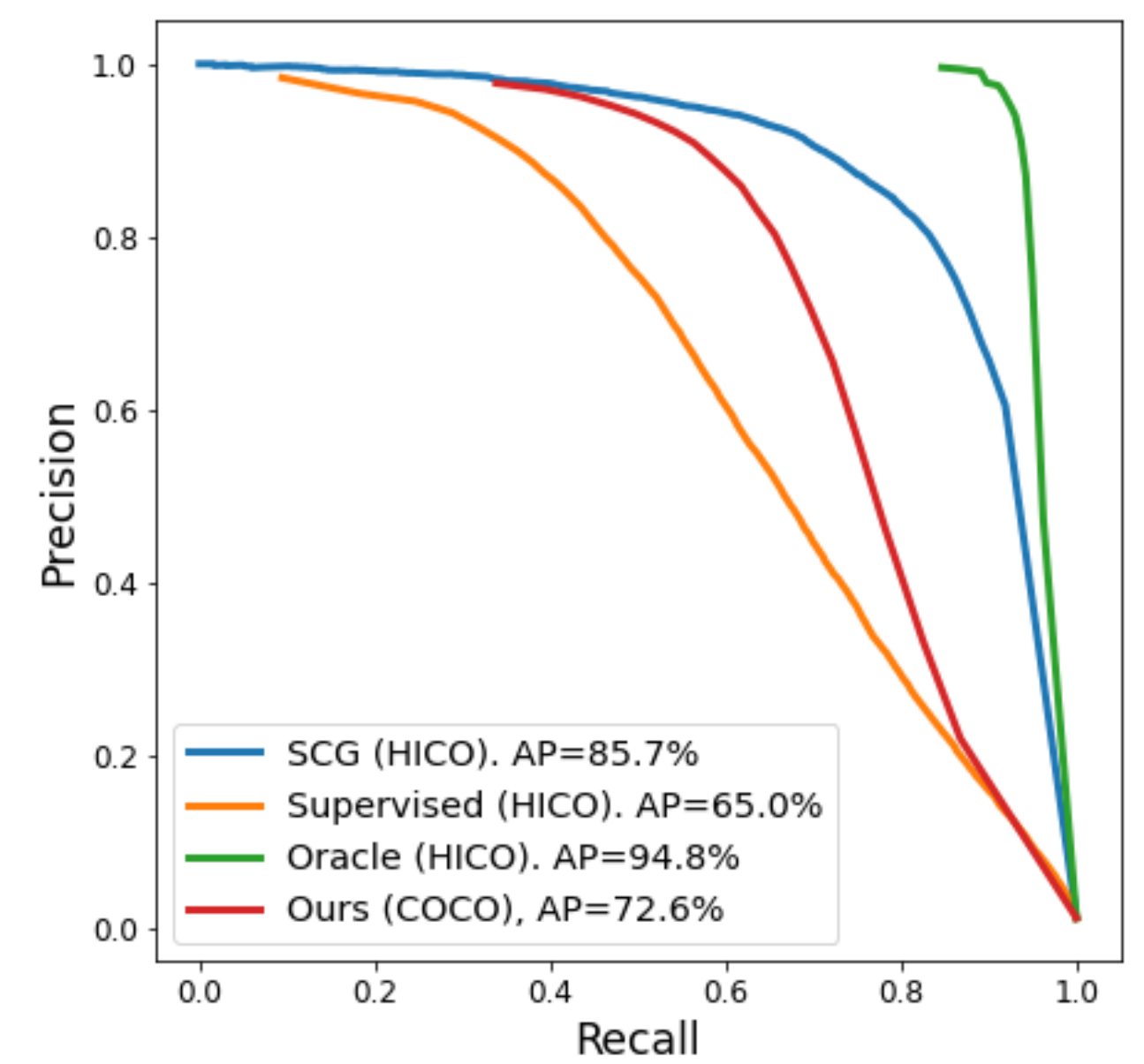# Contextual Visual Feature Learning for Zero-Shot Recognition of Human-Object Interactions
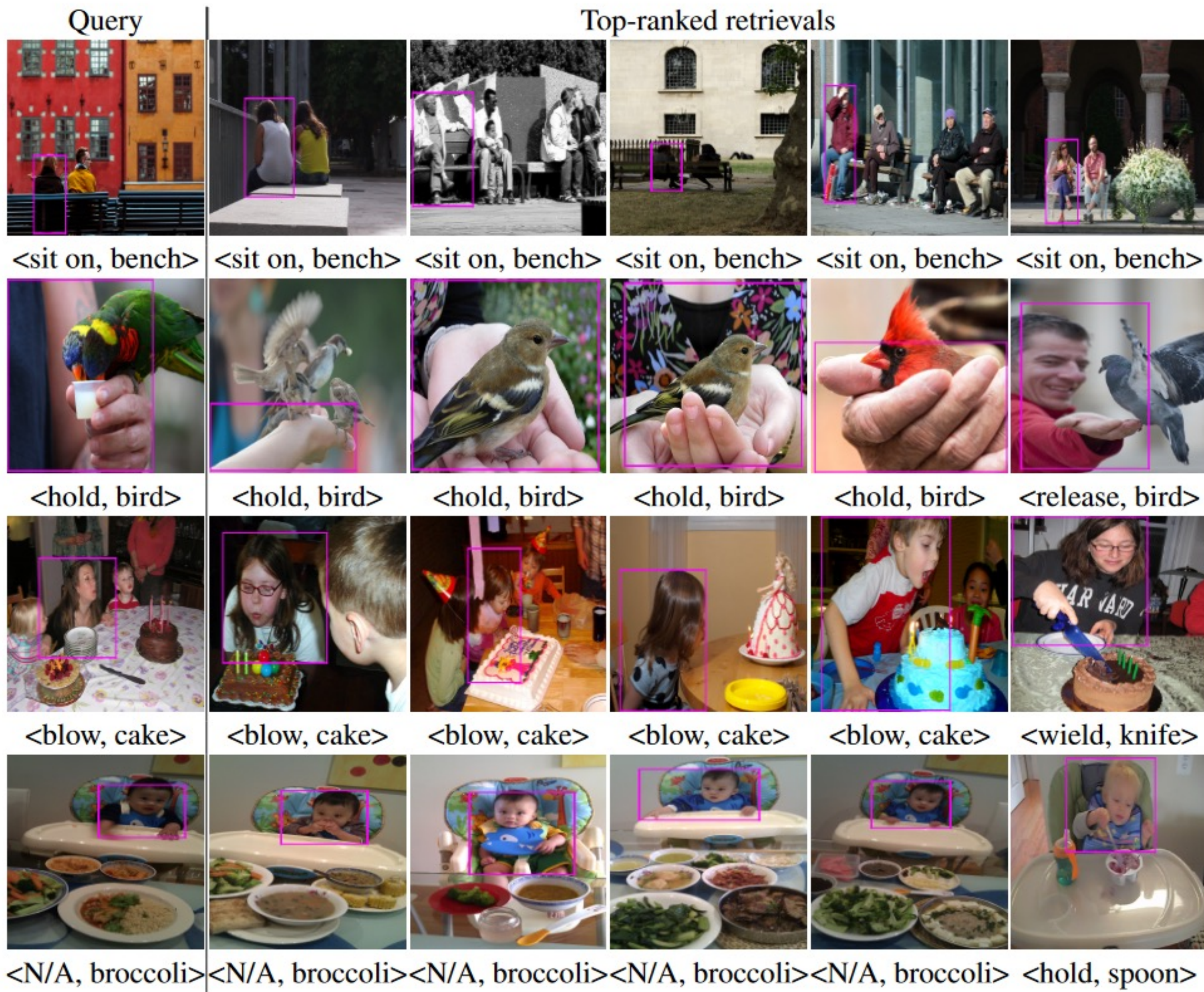
Tsung-Wei Ke   Dong-Jin Kim   Stella X. Yu   Liang Gou   Liu Ren
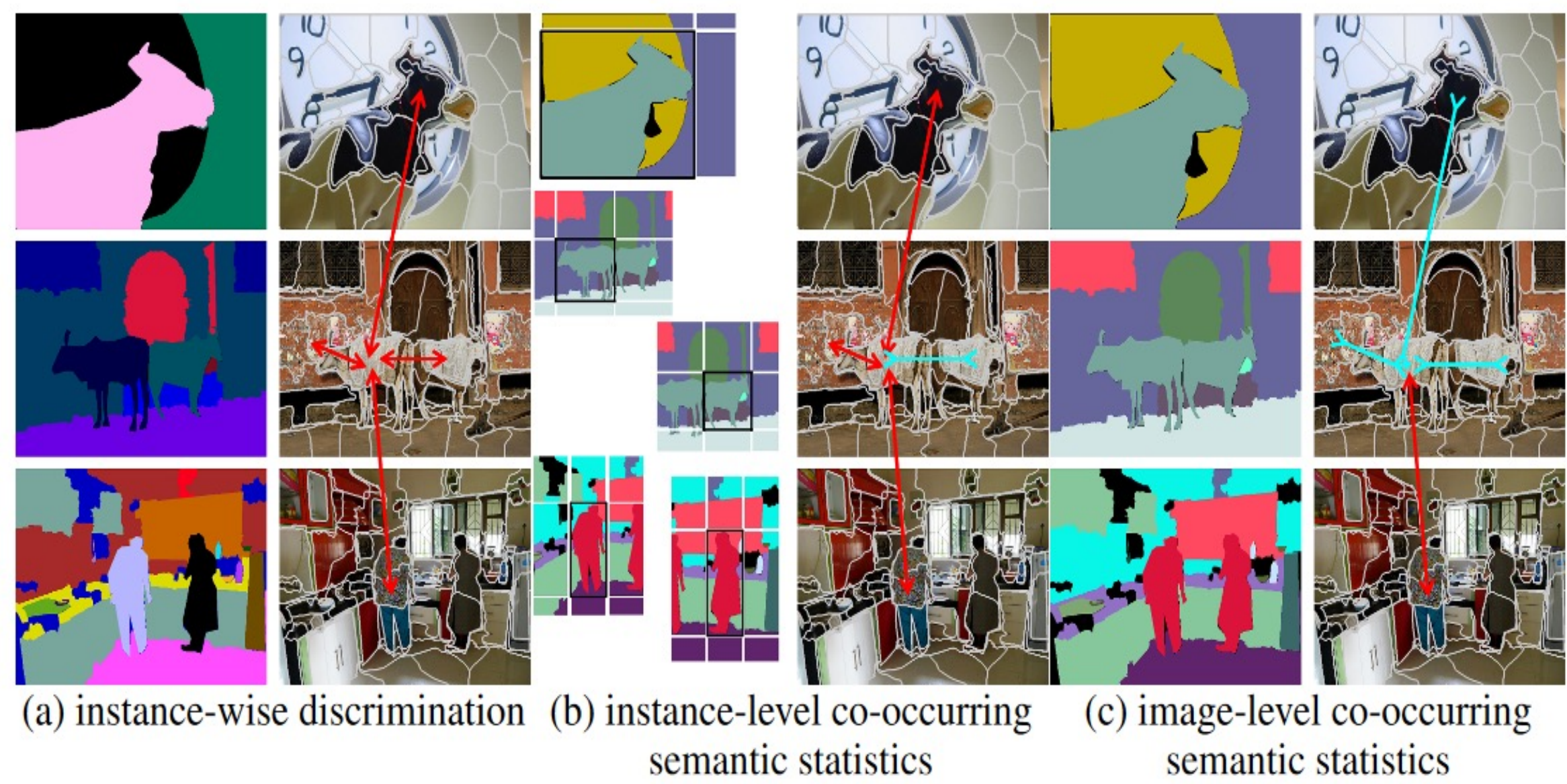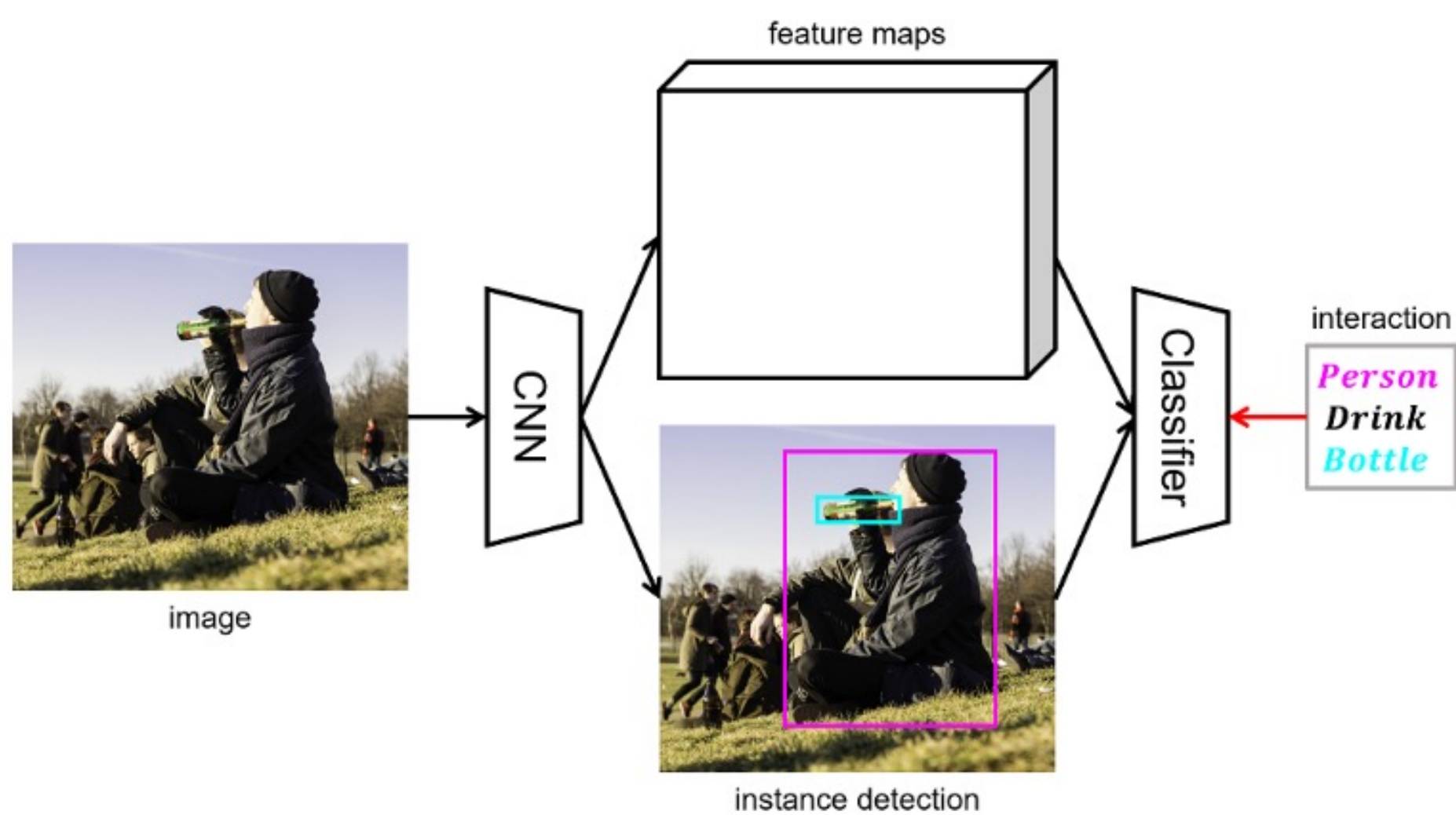
High-level Contextual Semantics Emerge from Our Learned Features

| Query | Top-ranked retrievals |

<sit on, bench> | <sit on, bench> | <sit on, bench> | <sit on, bench> | <sit on, bench> | <sit on, bench>

<hold, bird> | <hold, bird> | <hold, bird> | <hold, bird> | <hold, bird> | <release, bird>

<blow, cake> | <blow, cake> | <blow, cake> | <blow, cake> | <blow, cake> | <wield, knife>

<N/A, broccoli> | <N/A, broccoli> | <N/A, broccoli> | <N/A, broccoli> | <N/A, broccoli> | <hold, spoon>

SCG (HICO). AP=85.7%
Supervised (HICO). AP=65.0%
Oracle (HICO). AP=94.8%
Ours (COCO). AP=72.6%

SCG (HICO). AP=24.3%
Supervised (HICO). AP=16.1%
Oracle (HICO). AP=31.3%
Ours (COCO), AP=21.5%

Baselines: Classify Human-Object Pairs. Training Needs Interaction Labels.

We Contrast Pixels on 3 Types of Relationships

(a) instance-wise discrimination (b) instance-level co-occurring semantic statistics (c) image-level co-occurring semantic statistics

Our Work: Maps Pixels to Latent Features. Training Doesn't Need Interaction Labels.

Query | Top Retrievals

Hold Bottle | Hold Bottle | Hold Bottle | Hold Bottle | Hold Bottle | Drink Bottle

We Model Contexts with Spatially Co-occurred Semantics

Query | Top Retrievals   Query | Top Retrievals