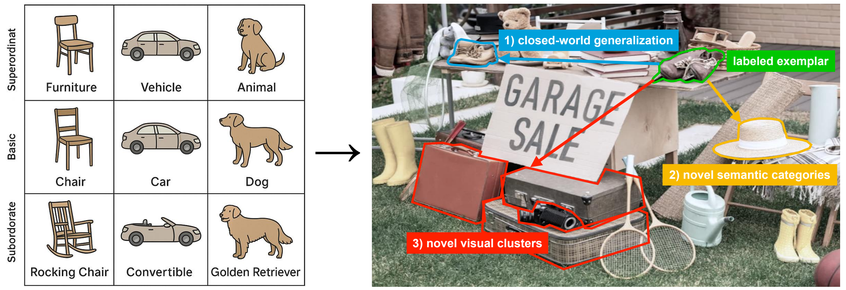
Open Ad-hoc Categorization with Contextualized Feature Learning

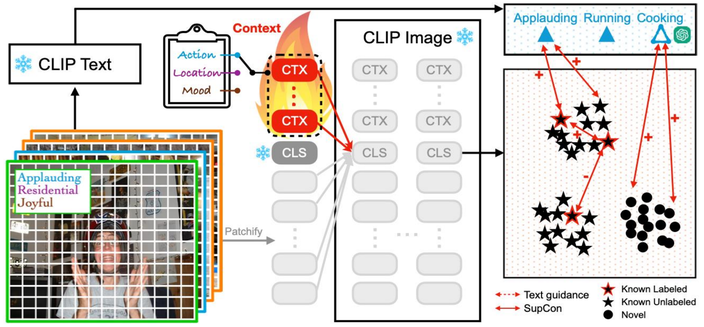
Zilin Wang* Sangwoo Mo* Stella X. Yu Sima Behpour Liu Ren

From Basic-Level to Open Ad-hoc Categorization



From Fixed to Open Concepts with Contextual Framing

OAK: CLIP w/ Context Tokens and Clustering



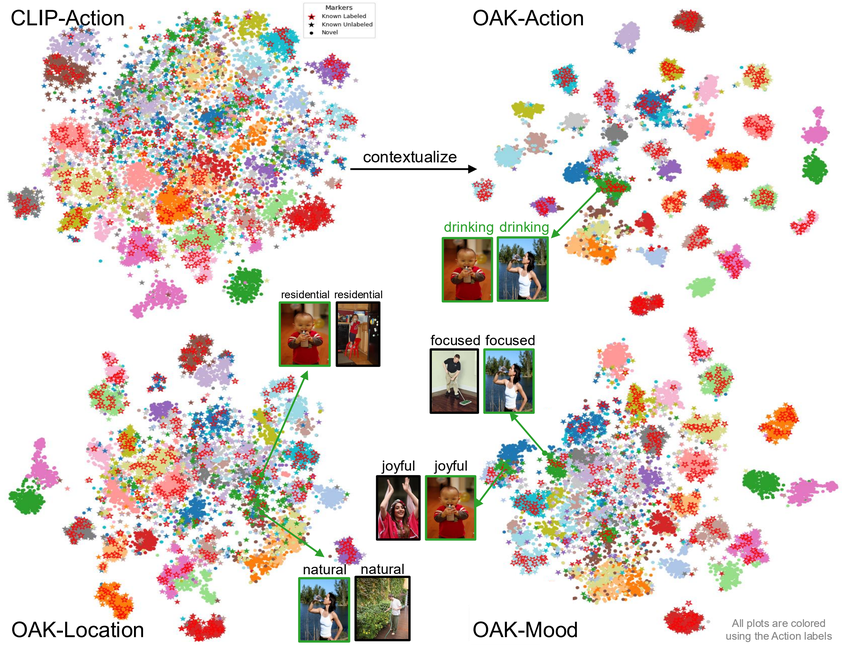
1. **Architecture**: Build on CLIP to leverage its general knowledge
2. **Context tokens:** Modulate ViT to contextualize visual features
3. **Top-Down Guidance**

   CLIP Baseline: *A photo of a { cat, dog, ..., elephant, bird, ... }* → LLM/GT
4. **Bottom-Up Guidance**

   GCD Baseline:

$$\text{OAK} = \underset{\text{Loss}}{\text{CLIP}}\left(\begin{matrix}\text{labeled,}\\ \text{pseudo-labeling,}\\ \text{by cluster}\end{matrix}\right) + \underset{\text{Loss}}{\text{Contrastive}}\left(\begin{matrix}\text{supervised,}\\ \text{self-supervised}\end{matrix}\right)$$

Contextual Features Group Images to Reveal Concepts



Significant Gains on Novel Concepts across Contexts

Method	Known			Novel			Overall			
	Action	Location	Mood	Action	Location	Mood	Action	Location	Mood	Omni
CLIP-ZS	**96.2**	60.4	**87.4**	-	-	-	-	-	-	-
+ LLM vocab	93.5	58.3	75.9	38.6	34.2	35.4	65.2	47.5	55.0	43.0
+ GT vocab	93.6	59.6	78.8	80.3	59.7	65.8	86.7	59.7	72.1	38.3
SS-KMeans	67.0	56.3	43.2	53.9	71.1	78.2	60.2	62.9	61.3	47.7
GCD	89.4	75.4	64.3	67.8	80.8	40.6	78.3	77.8	52.1	52.3
OAK (ours)	88.9	**83.9**	68.8	**85.1**	**88.4**	**87.4**	**86.9**	**85.9**	**78.4**	**70.3**

Discovered Clusters Align w/ Human Naming

GT	Naming
running	*jogging*
gardening	*weeding*
dining area	*kitchen*
joyful	*exhilarated*
⋮	⋮

Contextual Framing Shifts Attention for Interpretation