ZOOM BEHAVIOR INSIGHT

Presented by:
Ofir Duek , Aviv Meir , Rotem Aloni

PROJECT REVIEW

Problem

In online classes, instructors can't reliably infer students' behaviors from a single webcam feed (e.g., gaze, privacy, headphones, or object usage).

Our Goal

Given a single Zoom-like webcam frame, predict a discrete behavior vector:
Gaze, Headphones, Environment, Privacy, Object-in-hand (incl. Unknown).

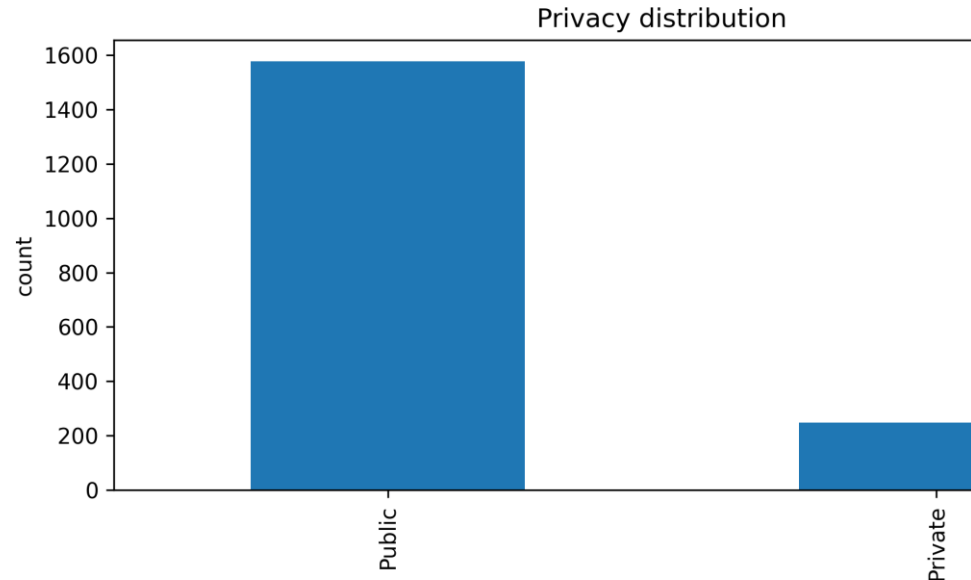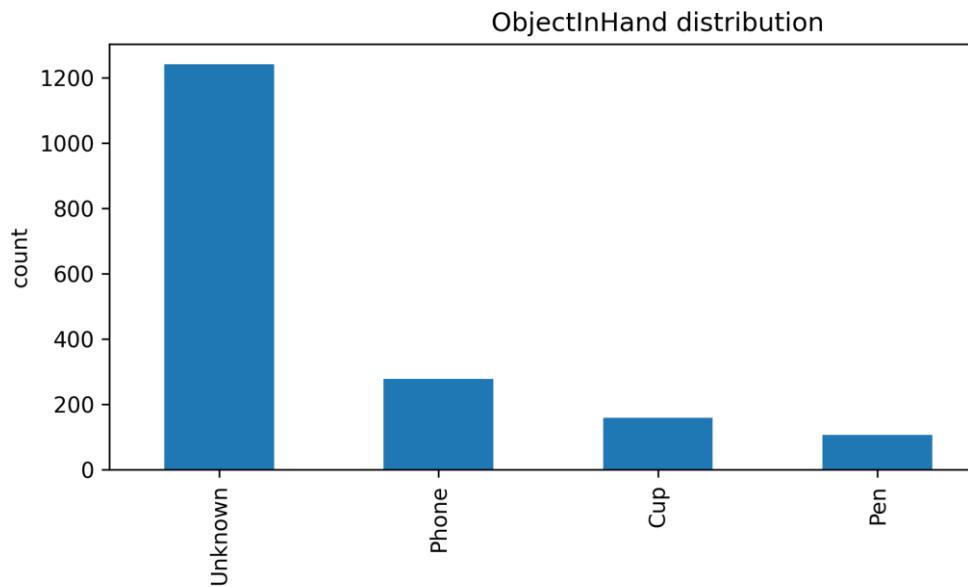## What's new since the first presentation:

- Switched from 0 - 100% statistics to categorical behavior labels.
- Set strict labeling rules with Unknown when evidence is missing.
- Started the data pipeline: real Zoom-like samples + synthetic augmentation.

Novelty / Contribution

We output multiple interpretable behaviors (not a single engaged score) and use controlled synthetic augmentation to cover rare cases.

RELATED WORK

Paper (Year)	Task	Signals / Cues	Method / Model	Data	How it relates to us
Stungage (2022)[1]	Engagement monitoring	Webcam video + meeting context	Real-time system + ML pipeline	Online meeting sessions	We predict a multi-attribute behavior vector from a single Zoom-like frame (not see engagement score).
Facial behavior engagement (2022/2023)[2]	Engagement estimation	Face cues (gaze, head pose)	Sequence model (LSTM on face features)	Online learning webcam videos	Similar gaze setting, but we output multiple behavior labels (incl. Unknown) and add non-face cues.
Emotion + eye/head (2019)[3]	Engagement classification	Emotion + eye/head movement	ML classifier on multi-cue features	Webcam e-learning sessions	Multi-cue idea fits; we extend cues to privacy / environment / object-in-hand and handle missing evidence with Unknown.

ObjectInHand distribution

Privacy distribution

DATASET OVERVIEW + EDA

- Our dataset combines real Zoom-like photos and synthetic images.

- Each image is labeled with a 5-attribute discrete behavior vector, including Unknown when evidence is missing.

- EDA: class distributions per attribute to identify imbalance and guide augmentation.

REAL DATA

- Real Zoom-like images from team-captured photos and public online datasets.

- Used as the most realistic reference for training and evaluation.

Labels:
Gaze = Not_Camera,
Headphones =
With_Headphones,
Environment = Outdoor,
Privacy = Private ,
ObjectInHand = Unknown

Labels:
Gaze = Not_Camera,
Headphones =
Without_Headphones,
Environment = Outdoor,
Privacy = Public,
ObjectInHand = pen

STABLE DIFFUSION (TEXT-TO-IMAGE)

- Generated images from scratch using prompts.

- Used to control behaviors and increase rare cases.

- Sample prompts:
  1. Zoom-style webcam photo, gaze at camera, with headphones, outdoor public background with blurred people, holding a pen in hand. Realistic, natural daylight, shallow depth of field, sharp subject.
  2. Zoom-style webcam photo, eyes closed, with headphones, indoor private setting, holding a cup in hand. Realistic, soft light, shallow depth of field, centered framing.

INPAINTING BACKGROUND REPLACEMENT

- Replace the background only to control Environment and Privacy labels.

- Keep the same subject to reduce confounding factors.

- Produces paired samples: Before → After with consistent appearance.

Same subject, different context

Before

After

BASELINE MODEL (RESNET18 MULTI-TASK) - RESULTS

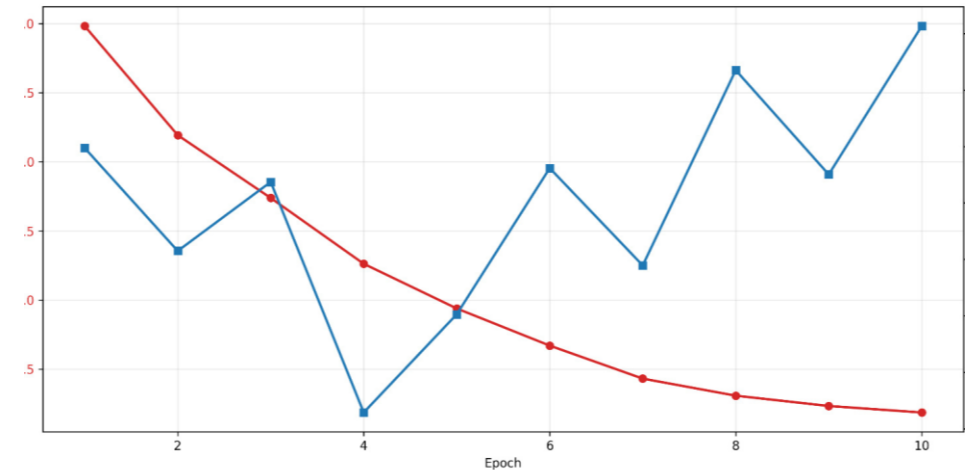Baseline Training: Train Loss vs Validation Avg Accuracy

Baseline setup:

- **Model:** ResNet18 (ImageNet pretrained), with one classification head per behavior.

- **Training:** 10 epochs, batch=32, Adam lr=1e-3, 80/20 train–val split.

Key takeaway:

- Train loss decreases strongly while validation avg accuracy changes only mildly → baseline overfitting.

- Validation performance is uneven across tasks (best: Privacy, hardest: ObjectInHand).

Final Val Acc	Task
0.959	Privacy
0.870	Environment
0.759	Headphones
0.736	Gaze
0.683	ObjectInHand

NEXT STEPS

Train an improved model on the finalized dataset (stronger baseline + regularization).

EDA-driven data balancing: add targeted real/synthetic samples for rare classes.

Evaluate on a held-out real test set and report per-behavior metrics.

Error analysis + prototype demo: input one Zoom-like frame → output the behavior vector.