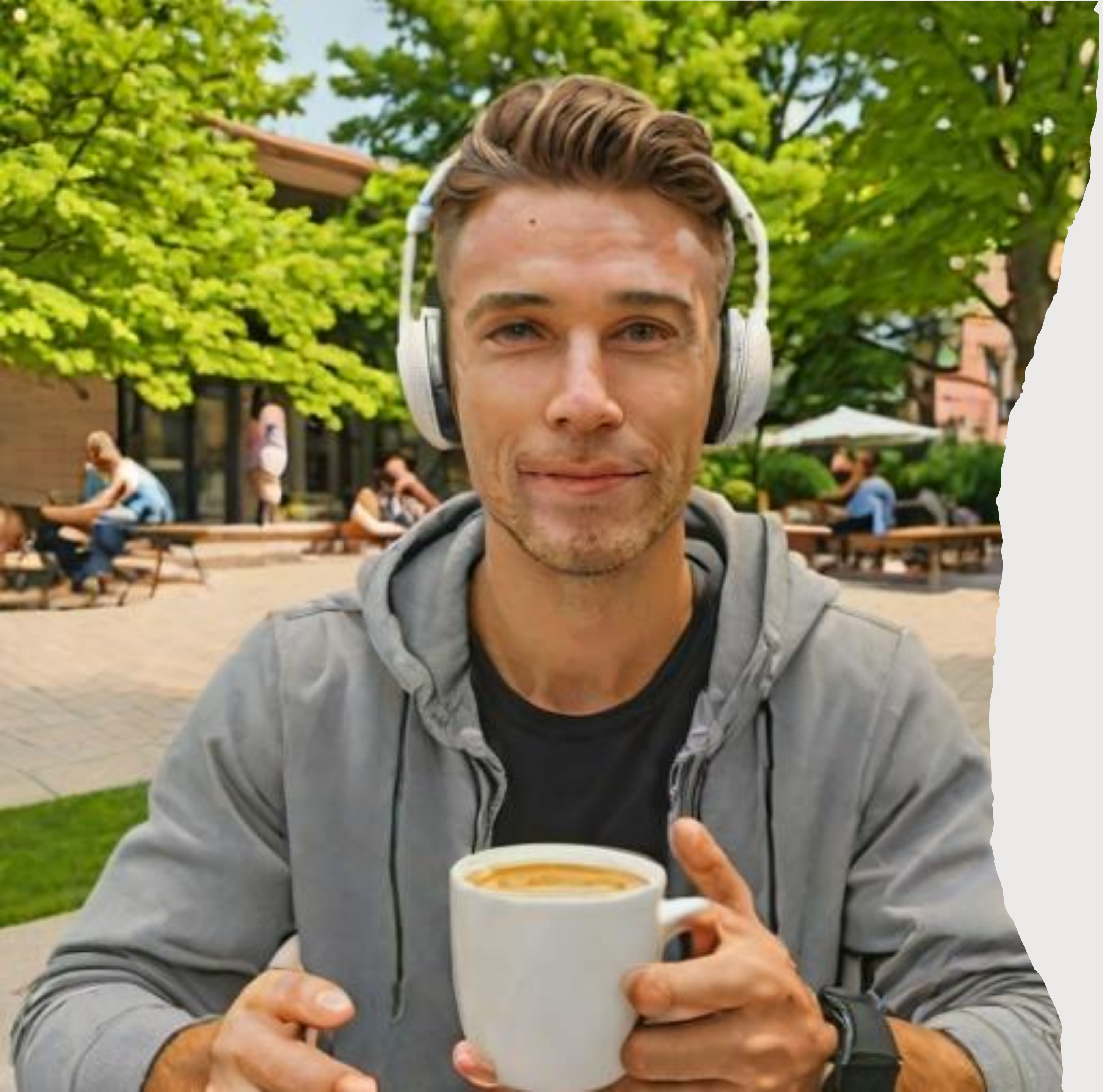
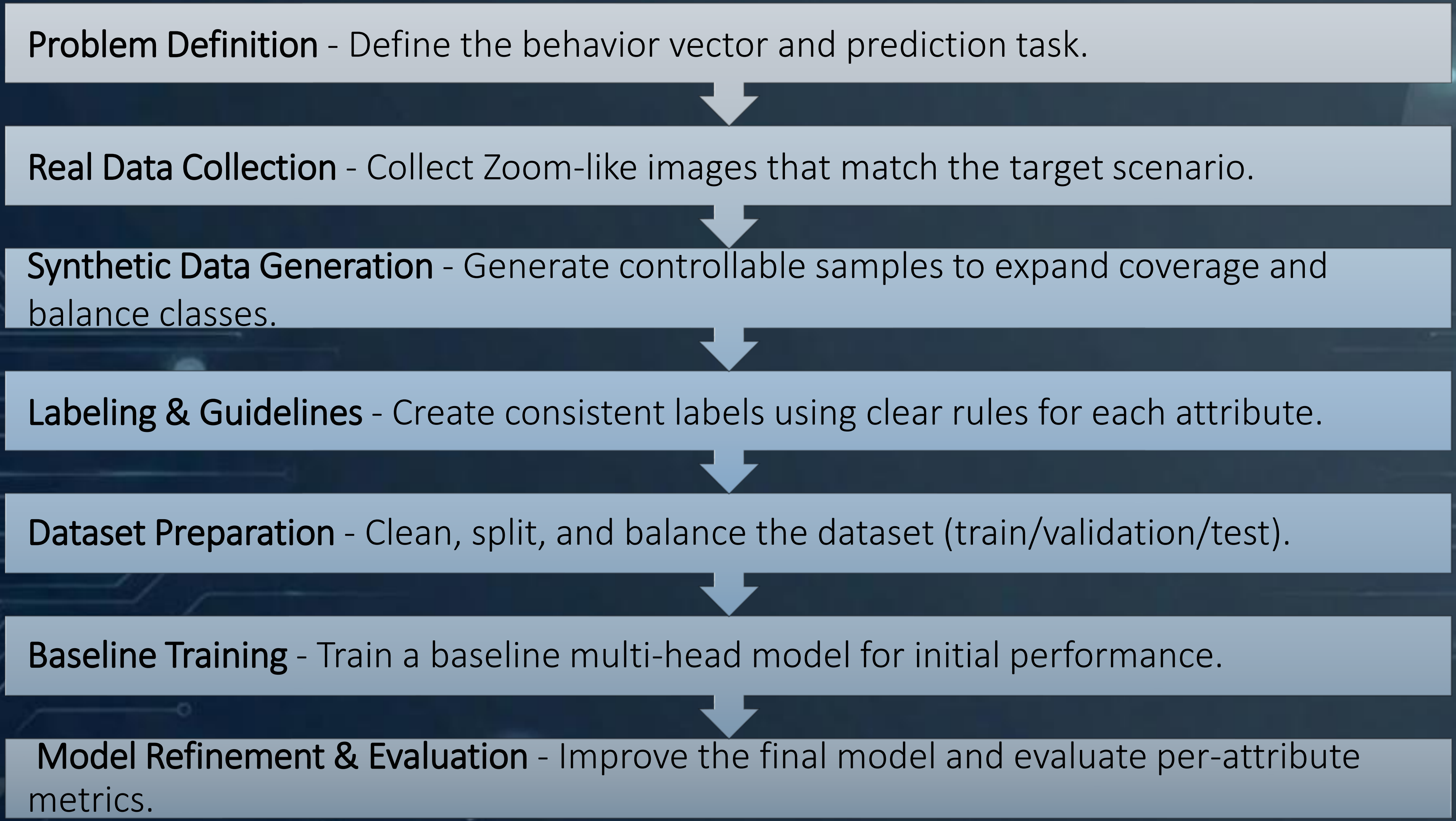
Presented by:
Ofir Duek , Aviv Meir , Rotem Aloni



From a Zoom-like Frame to a Behavior Vector

- **Goal:** Predict a multi-attribute behavior vector from a single Zoom-like image.
- **Output:** y = [Gaze, Headphones, Environment, Privacy, Object-in-hand]
- **Example:** [Camera, With_Headphones, Outdoor, Public, Cup]
- **Why it matters:** Enables scalable analysis of learning conditions in remote settings.
- **Novelty:** A dedicated dataset + controllable synthetic data generation.

Project Pipeline: From Idea to a Trained Model



Problem Definition - Define the behavior vector and prediction task.

Real Data Collection - Collect Zoom-like images that match the target scenario.

Synthetic Data Generation - Generate controllable samples to expand coverage and balance classes.

Labeling & Guidelines - Create consistent labels using clear rules for each attribute.

Dataset Preparation - Clean, split, and balance the dataset (train/validation/test).

Baseline Training - Train a baseline multi-head model for initial performance.

Model Refinement & Evaluation - Improve the final model and evaluate per-attribute metrics.

Data Creation & Labeling

Real Data (~900):

- Self-captured Zoom-like images + a small curated online set.
- ~200 fully labeled + ~700 labeled for Gaze only.

Synthetic Data (~2000):

- Sample behavior vector → auto-labels (Gaze/Headphones/Env/Privacy/Object).
- SDXL Turbo (Text-to-Image) - generate Zoom-like image from a prompt template.
- Inpainting (Background replacement): keep the person fixed (mask) + change background to control Environment/Privacy (labels updated)

Prompt Examples:

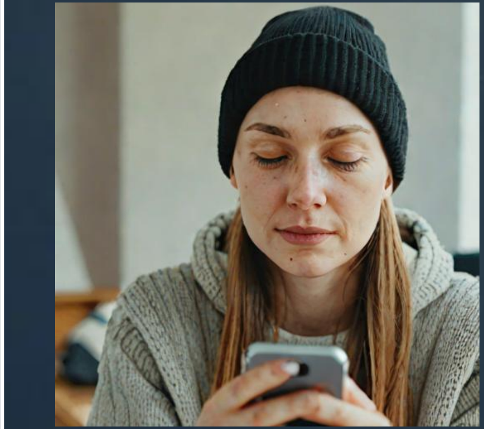
- T2I: webcam zoom call, student, looking at camera, holding phone, indoor, realistic.
- Inpaint BG: busy public coffee shop background / quiet private bedroom background.

Examples

Gaze

Environment

Headphones



Privacy

Object in hand

Models, Training & Evaluation

Baseline:

- ResNet-18 (pretrained), multi-head (one head per attribute).

Final Model:

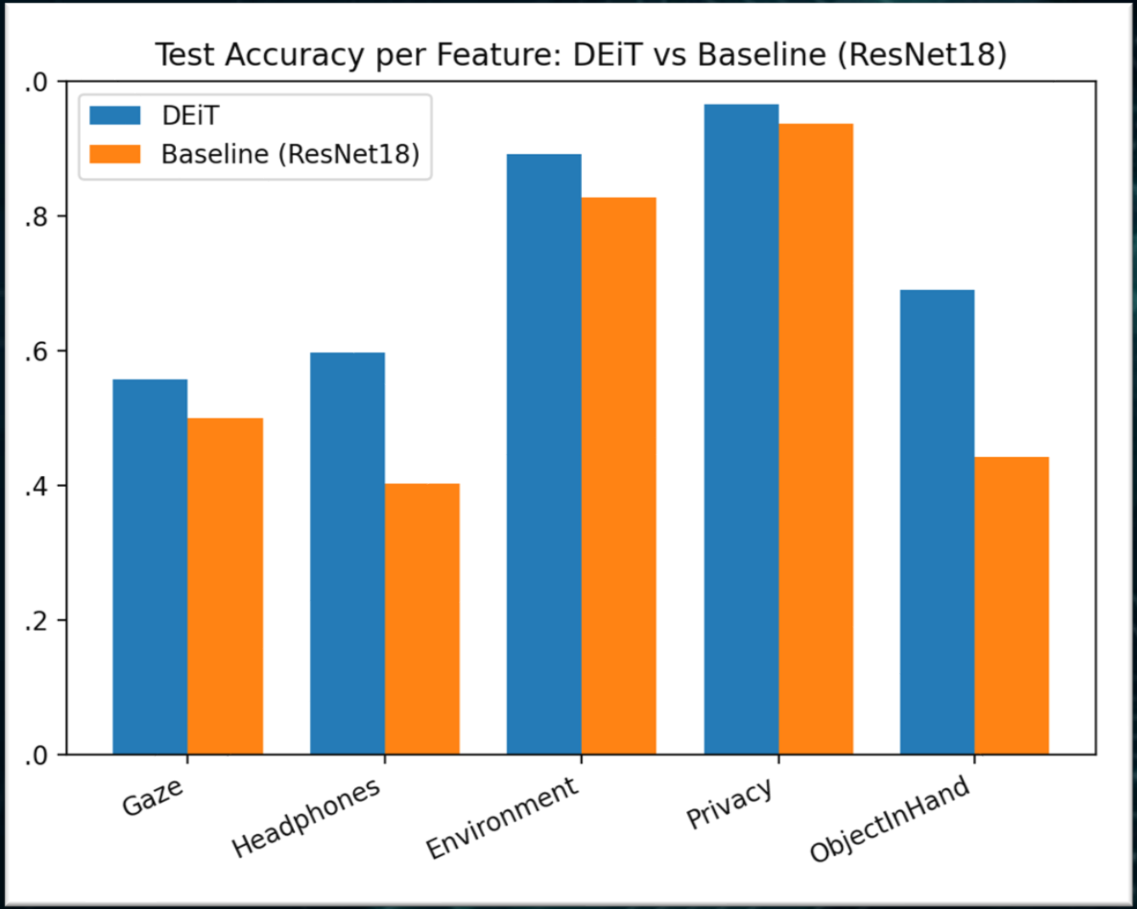
- DeiT-Base (pretrained), multi-head.
- Fine-tuned end-to-end (224×224).
- Switched to a Transformer-based model after baseline overfitting.

Evaluation:

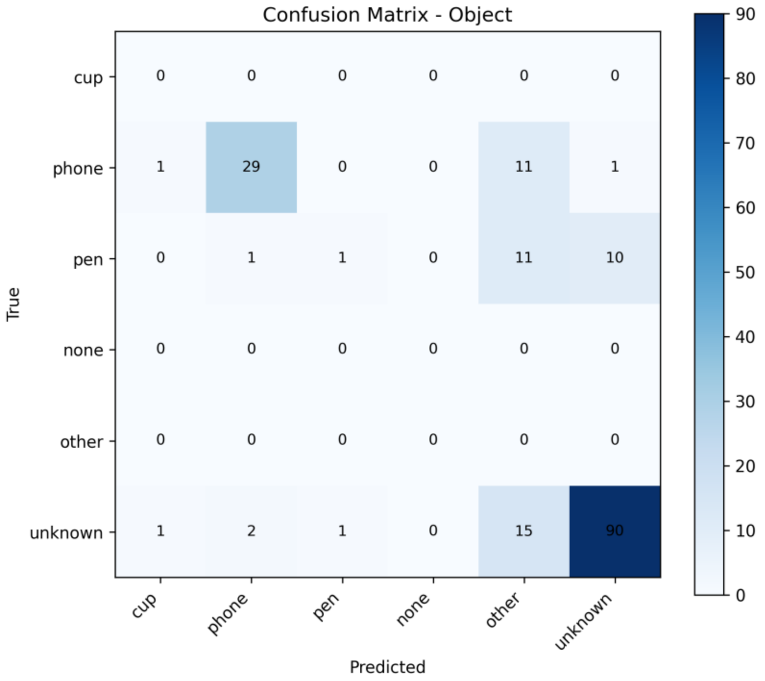
- Per-feature Accuracy + Macro-F1
- Overall: Mean Accuracy + Joint Accuracy (all 5 correct)
- 80/10/10 split, real-only test set.
- Confusion matrix (Object) for error analysis.

Results: Accuracy per Feature

- Multi-task, evaluated on a real-only test set (80/10/10 split).
- DeiT consistently outperforms the ResNet-18 baseline.
- Strongest: Privacy & Environment | Most challenging: Gaze & Object-in-Hand.

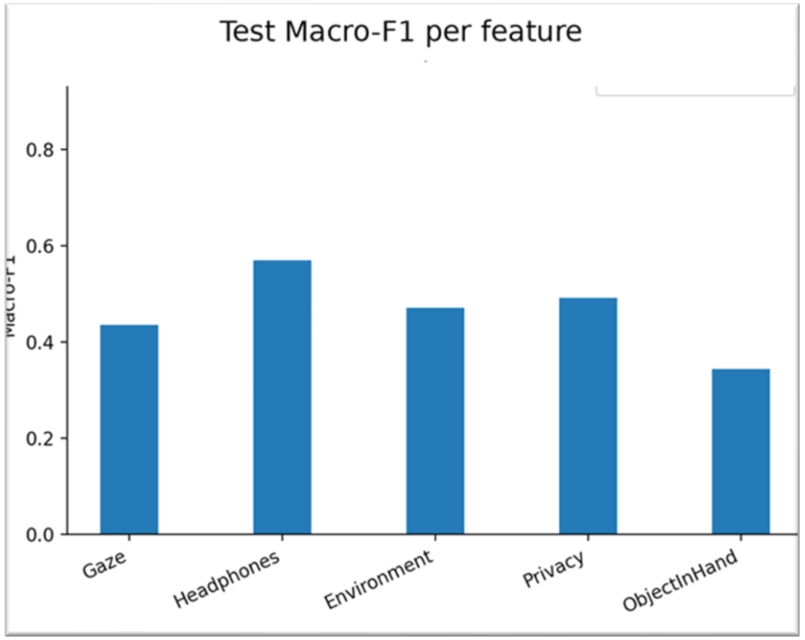


Results: Macro-F1 & Error Analysis



Confusion Matrix - Object

True \ Predicted	cup	phone	pen	none	other	unknown
cup	0	0	0	0	0	0
phone	1	29	0	0	11	1
pen	0	1	1	0	11	10
none	0	0	0	0	0	0
other	0	0	0	0	0	0
unknown	1	2	1	0	15	90



Test Macro-F1 per feature

- **Macro-F1 (class-balanced):** strongest on Headphones, weakest on Object-in-Hand (harder task).
- **Object confusion matrix:** most errors are between phone / other / unknown (similar shapes + occlusions).

Summary & Future Work

Key Achievements:

- Built a Zoom-like dataset (~2,900 images): ~900 real + ~2,000 synthetic.
- Trained a multi-task, multi-head DeiT model and improved the baseline (ResNet-18 → DeiT).

Key Takeaways:

- Real-only test set is critical for measuring real-world performance
- Strongest: Privacy & Environment | Most challenging: Gaze & Object-in-Hand.

Future Work:

- Extend from single-frame prediction to video-level focus estimation by aggregating frame-wise vectors.
- Improve Object-in-Hand with more real data and better class balance (reduce confusion: phone / other / unknown).