

# EgoBlind-RA

## Risk-Adaptive Egocentric Visual Assistance for Blind Users

Julia Kim   Xander Backus

MAS.S60

May 12, 2026



# Why Urgency Matters for Blind Users

- **253M** people worldwide live with visual impairment

# Why Urgency Matters for Blind Users

- **253M** people worldwide live with visual impairment
- AI assistants interpret their surroundings via **egocentric video**



# Why Urgency Matters for Blind Users

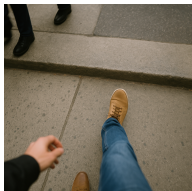
- **253M** people worldwide live with visual impairment
- AI assistants interpret their surroundings via **egocentric video**



- Benchmarks (EgoBlind, VizWiz) + VLMs (LLaVA, Kimi-VL) apply a **uniform policy** to all queries

## Why Urgency Matters for Blind Users

**fast, brief answer**

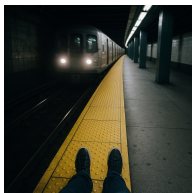


*"Is a step in front of me?"*

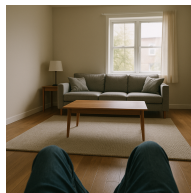
**detailed answer**



*"What cereal brand is this?"*



*"Is anything in my way?"*



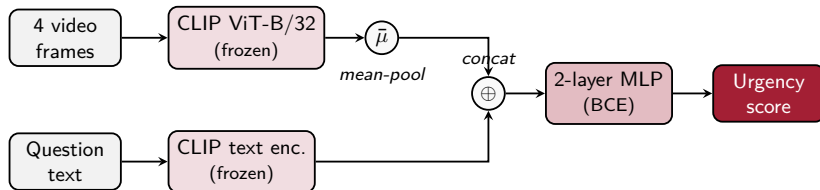
*"Is anything in my way?"*

# Our Goal

**Our goal:** classify query urgency from video + text,  
*then* route to a response calibrated to the safety stakes.

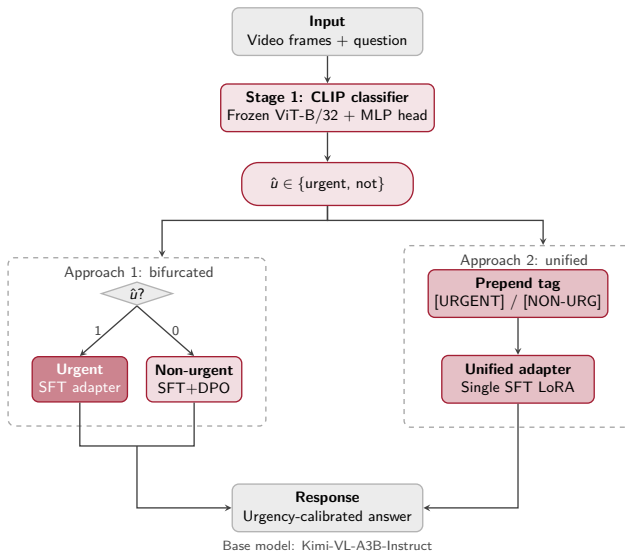
# Stage 1: CLIP Urgency Classifier

$$\text{urgency} = \text{MLP}(\text{mean-pool}(\text{CLIP}_v(\text{frames})) \oplus \text{CLIP}_t(\text{question}))$$



- Test ROC-AUC: **0.905** → strong signal from frozen features

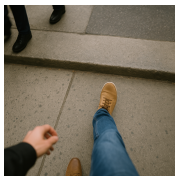
## Stage 2: Two Architectural Approaches





# SFT Dramatically Improves Over Baseline

For each  $(v, q)$ , the model learns to mimic a length-calibrated reference answer.



*"Is a step in front of me?"*

**shortest** ref: "Yes, step."



*"What cereal brand is this?"*

**longest** ref: "Honey Bunches of Oats..."

|                           | Loss ↓       | Similarity ↑ | Mean words |
|---------------------------|--------------|--------------|------------|
| Baseline (uniform policy) | 0.607        | 0.350        | 84.1       |
| Approach 1                | 0.556        | 0.393        | 4.5        |
| Approach 2                | <b>0.375</b> | <b>0.589</b> | 2.2        |

# Key SFT Discoveries

- **Vision conditioning is essential:** text-only training collapses to “Yes”/“No” for 71% of urgent queries

# Key SFT Discoveries

- **Vision conditioning is essential:** text-only training collapses to “Yes”/“No” for 71% of urgent queries
- **Light LoRA prevents overfitting:** rank 8,  $q\_proj + v\_proj$ , 1 epoch (vs. rank 16, all projections, 5+ epochs)

# DPO: Architecture Determines Whether It Helps

For each  $(q, v)$ , DPO sees two candidate responses and learns to prefer the better one.



*"What cereal brand is this?"*

## Chosen

*"Honey Bunches of Oats, on the counter next to a bowl."*

## Rejected

*"Cereal."*

## Approach 2 (unified)

DPO regresses SFT

| Method    | Loss ↓ |
|-----------|--------|
| SFT       | 0.375  |
| SFT → DPO | 0.433  |

# DPO: Architecture Determines Whether It Helps

For each  $(q, v)$ , DPO sees two candidate responses and learns to prefer the better one.



*“What cereal brand is this?”*

## Chosen

*“Honey Bunches of Oats, on the counter next to a bowl.”*

## Rejected

*“Cereal.”*

## Approach 2 (unified)

DPO regresses SFT

| Method    | Loss ↓       |
|-----------|--------------|
| SFT       | <b>0.375</b> |
| SFT → DPO | <b>0.433</b> |

## Approach 1 (bifurcated)

hurts urgent, helps non-urgent

|         | Urgent       | Non-urg.     |
|---------|--------------|--------------|
| SFT     | <b>0.505</b> | 0.268        |
| SFT+DPO | <b>0.459</b> | <b>0.298</b> |

# End-to-End Pipeline

How does each architecture handle *real* classifier routing?

| Routing strategy              | Approach 1   | Approach 2   |
|-------------------------------|--------------|--------------|
|                               | Loss ↓       | Loss ↓       |
| Baseline (no routing)         | 0.607        | 0.607        |
| <b>CLIP classifier (ours)</b> | <b>0.556</b> | <b>0.375</b> |
| Oracle (ground-truth)         | 0.549        | 0.375        |
| Gap to oracle                 | 0.007        | <b>0.000</b> |

For blind users, **Approach 2 is safer**: misrouted urgent queries still get a competent answer.

*Thank You*

**Code:** [github.com/juliavekim/EgoBlind-RA](https://github.com/juliavekim/EgoBlind-RA)  
[github.com/xabackus/mmai](https://github.com/xabackus/mmai)

**Classifier:** [huggingface.co/julia225/egoblind-ra-clip-urgency](https://huggingface.co/julia225/egoblind-ra-clip-urgency)