EgoBlind-RA: Towards Safer Egocentric Assistive AI for Blind Users via Risk-Adaptive Routing

Julia Kim*
Massachusetts Institute of Technology
julia225@mit.edu

Xander Backus*
Massachusetts Institute of Technology
xabackus@mit.edu

Abstract

Blind and low-vision (BLV) users depend on visual assistance systems to navigate an unpredictable physical world, yet existing multimodal systems apply a uniform inference policy irrespective of query urgency. We present **EgoBlind-RA**, a risk-adaptive framework that classifies egocentric query-video pairs as *urgent* or *non-urgent* via a lightweight CLIP-based classifier and routes each query to a response calibrated to its safety stakes. Our classifier achieves 0.905 ROC-AUC on the held-out test set. We compare two LoRA-fine-tuned architectures of Kimi-VL-A3B-Instruct on the EgoBlind benchmark: a bifurcated design with separate urgent and non-urgent adapters, and a unified prompt-tag-conditioned design. Urgency-conditioned SFT against carefully selected reference targets substantially improves over a uniform-policy baseline in both. Surprisingly, DPO regresses SFT in the unified setting whilst proving productive in the bifurcated setting, where the best result mixes an SFT urgent adapter with a DPO non-urgent adapter. The unified adapter is also robust to noisy CLIP routing at deployment, matching oracle composite loss despite 21.5% test-time tag flips. Together, these results suggest the value of preference-pair-based alignment and the cost of classifier noise both depend sharply on architectural regime.

1 Introduction

Vision-language models (VLMs) have advanced rapidly, yet their development has largely centered on users with full sensory access. For the $\sim$253M people worldwide living with visual impairment [1], AI systems that interpret egocentric video and answer natural-language questions about the immediate environment offer a transformative opportunity to support independence and personal safety. Benchmarks for this setting — EgoBlind [2] and VizWiz [3] — have begun to expose challenges specific to blind and low-vision (BLV) users: off-centre regions of interest, ambiguous spatial context, motion blur, and information needs ranging from casual to life-critical. Performance here matters beyond academic interest because the cost of a misleading answer is asymmetric, and in the safety-critical case, severe.

Yet state-of-the-art multimodal systems apply a uniform inference policy to every query, regardless of stakes. Consider two questions a blind user might ask: *"Is there a step in front of me?"* signals an immediate physical hazard and demands a fast, conservative response; *"What brand of cereal is this?"* tolerates a slower, more deliberative one. Models such as LLaVA [4] and InstructBLIP [5] treat the two identically, and the cost runs in both directions: excessive caution degrades usability on routine queries, whilst insufficient urgency on safety-critical ones risks physical harm. Urgency, moreover, is not determined by text alone — *"Is there anything in my way?"* carries very different stakes at the edge of a subway platform than in one's kitchen — so any effective response policy must condition jointly on the question and the visual scene. To our knowledge, no prior assistive system jointly classifies visual urgency and routes queries to risk-stratified response policies.

*Equal contribution.

In this paper, we propose **EgoBlind-RA**, a risk-adaptive egocentric visual assistance framework for BLV users. EgoBlind-RA is a two-stage pipeline: a lightweight CLIP-based classifier predicts whether each query–video pair is *urgent* or *non-urgent*, and the predicted urgency conditions a fine-tuned VLM. We investigate two architectures: a bifurcated design with separate urgent and non-urgent Low-Rank Adaptation (LoRA) adapters (**Approach 1**), and a unified single-adapter design conditioned by an urgency prompt-tag (**Approach 2**). Both are optimized under a regime-specific composite loss balancing accuracy, assistive utility, and generation latency.

We make three contributions. First, a frozen-backbone CLIP classifier predicts query urgency from joint video–text representations, reaching 0.905 ROC-AUC on the test set. Second, urgency-conditioned SFT with length-asymmetric reference targets — shortest-correct for urgent queries, longest-correct for non-urgent — substantially outperforms the uniform-policy baseline in both architectures. Third, two central design choices depend sharply on adapter architecture: DPO regresses SFT in the unified design but is productive in the bifurcated one, and the unified design absorbs CLIP routing errors at deployment (matching oracle composite loss despite 21.5% tag flips) whereas the bifurcated design does not. These results suggest that alignment method, classifier robustness, and adapter architecture must be co-designed.

2 Related Work

Visual assistance for BLV users. VizWiz [3] introduced an early assistive VQA system in which BLV users photographed their surroundings and asked questions that were answered by crowd workers. This work highlighted challenges characteristic of BLV-generated images, including poor framing, motion blur, and queries spanning both low- and high-stakes situations. The setting was later formalized as a large-scale VQA benchmark [6] that demonstrated BLV-generated images are substantially more difficult for existing models than standard VQA inputs. More recently, EgoBlind [2] extended the problem to egocentric video; we adopt EgoBlind as our evaluation dataset. However, none of these benchmarks evaluates whether models adapt their behavior to the urgency or safety-criticality of user queries.

Egocentric video understanding. Ego4D [7] introduced a large corpus of egocentric video supporting tasks such as episodic memory, action anticipation, and future forecasting. Embodied Question Answering [8] introduced a setting in which an agent must navigate through an environment to obtain the information needed to answer a query. However, both settings assume sighted users and treat all queries uniformly, without accounting for differences in urgency or safety criticality.

VLMs and efficient adaptation. Recent VLMs, including LLaVA [4], InstructBLIP [5], Qwen-VL [9], and Kimi-VL [10], achieve strong out-of-the-box performance on visual question answering tasks. We employ Kimi-VL as our base model. Our urgency classifier is built on representations from CLIP [11], whilst LoRA [12] enables efficient fine-tuning of multi-billion-parameter models on a single GPU.

Alignment and preference optimization. Reinforcement Learning from Human Feedback (RLHF) [13] aligns language models to human preferences using a reward model trained on pairwise comparisons. DPO [14] simplifies this process by optimizing directly from preference pairs, and is now widely adopted. Variants include Kahneman–Tversky Optimization (KTO) [15], which uses scalar utility labels instead of pairwise comparisons; Reinforcement Learning from AI Feedback (RLAIF) [16], which replaces human labels with AI feedback; and Group Relative Policy Optimization (GRPO) [17], which scores groups of sampled responses against a learned baseline rather than pairwise. We apply DPO with urgency-conditioned preference pairs and find that its effect depends sharply on adapter architecture — a failure mode not previously reported for short, vision-grounded VQA.

Risk-aware inference. Several lines of work adapt model behavior to input risk: models can abstain when uncertain [18], estimate uncertainty using semantic entropy [19], or allocate more compute to difficult inputs [20]. Importantly, these approaches are text-only and do not incorporate a learned visual urgency signal.

How our work differs. EgoBlind-RA is, to our knowledge, the first system to classify urgency from egocentric video and route queries to risk-stratified response policies in the assistive setting.

We further show that DPO's effect on SFT, and the deployment cost of classifier errors, both depend sharply on adapter architecture, a contrast to prior alignment and BLV work that treats these decisions in isolation.

3 Problem Statement

Let $\mathbf{x} = (v, q)$ denote a video-query pair, where $v \in \mathcal{V}$ is an egocentric video clip and $q \in \mathcal{Q}$ is a natural language query from a BLV user. Each pair has a binary urgency label $u \in \{0, 1\}$ and a reference answer set $\mathcal{Y}^* = \{y_1^*, \ldots, y_K^*\}$. EgoBlind-RA solves two coupled subproblems:

(1) learn an **urgency classifier** $f_\theta : \mathcal{V} \times \mathcal{Q} \to \{0, 1\}$;

(2) learn a **conditional response generator** $g_\phi : \mathcal{V} \times \mathcal{Q} \times \{0, 1\} \to \mathcal{A}$, whose objective depends on the predicted urgency $\hat{u} = f_\theta(\mathbf{x})$. We instantiate g in two ways: Approach 1 trains separate generators g_{ϕ_u}, g_{ϕ_n}, and selects between them by $\hat{u}$; Approach 2 trains a single generator g_ϕ that takes $\hat{u}$ as an input prompt-tag.

The objective for g is regime-specific: a composite loss balancing accuracy, assistive utility, and generation latency for urgent queries, and an accuracy-and-utility-only loss for non-urgent ones. We define these objectives in Section 4.3.

4 Proposed Method

As illustrated in Figure 1, EgoBlind-RA operates as a two-stage pipeline: (i) an urgency classifier predicts the risk level of each query, and (ii) a conditional response generator produces an answer matched to the predicted urgency regime. We investigate two instantiations of the second stage, motivated by an open architectural question: should risk-adaptive behaviour be encoded *structurally*, via separate adapters, or *procedurally*, via a single adapter conditioned on a prompt tag? The two approaches differ only at this point in the pipeline; classifier, loss, base model, and inference-time routing signal are otherwise shared.

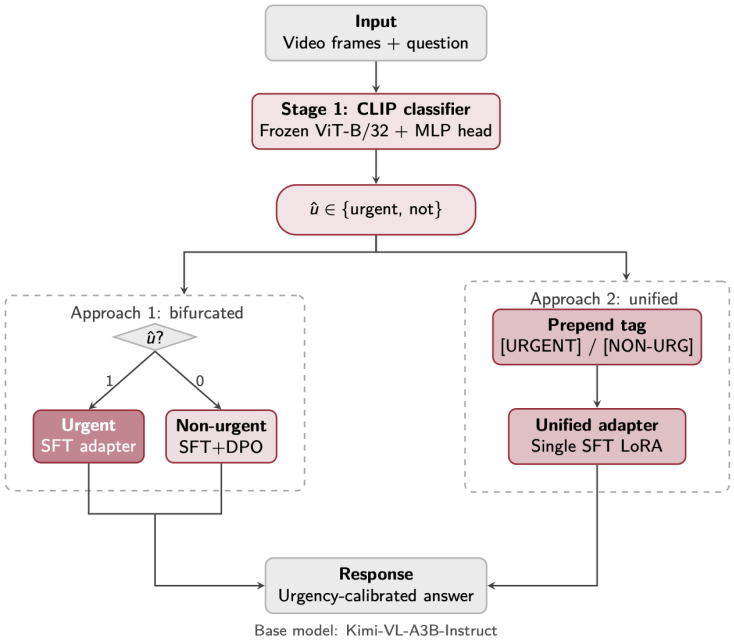

Figure 1: EgoBlind-RA pipeline. A CLIP-based classifier predicts query urgency from the joint video–text representation; the predicted urgency then routes to one of two architectural variants. Approach 1 selects between separate urgent and non-urgent LoRA adapters. Approach 2 prepends an urgency tag to a single unified adapter. Both are scored under the same regime-specific composite loss.

4.1 Stage 1: Urgency Classification

Given a video–query pair $\mathbf{x} = (v, q)$, we train a lightweight classifier f_θ to predict the binary urgency label $\hat{u} \in \{0, 1\}$. The model uses a frozen CLIP ViT-B/32 backbone: K frames sampled around the query timestamp are encoded individually, mean-pooled to produce a single visual embedding, and concatenated with the CLIP text embedding of the question. The fused representation is passed to a two-layer MLP head trained with binary cross-entropy.

Keeping the classifier lightweight lets it act as a fast preprocessing step before the response generator; freezing the backbone preserves CLIP's pretrained multimodal representations without end-to-end fine-tuning. We use a default decision threshold of 0.5 at inference; threshold tuning to reduce false negatives is discussed as future work (Section 7). Architectural and training details appear in Section 5, with results in Section 6.1.

4.2 Stage 2: Urgency-Conditioned Response Generation

We compare two architectural instantiations of the response generator (Figure 1), differing only in how urgency is incorporated. Both use the composite loss defined in Section 4.3.

Approach 1: Bifurcated adapters. We train two separate LoRA adapters on a shared base model $\mathcal{M}$. The *urgent adapter* ϕ_u is fine-tuned on $\mathcal{D}_u = \{(\mathbf{x}_i, y_i) : u_i = 1\}$ under the full loss $\mathcal{L}_{\text{urgent}}$, with the *shortest* correct reference answer as target to encourage concise, actionable responses. The *non-urgent adapter* ϕ_n is fine-tuned on $\mathcal{D}_n = \{(\mathbf{x}_i, y_i) : u_i = 0\}$ under $\mathcal{L}_{\text{acc}} + \mathcal{L}_{\text{util}}$ (no latency term), with the *longest* reference answer as target. At inference, the classifier prediction $\hat{u}$ selects the adapter:

$$y = \begin{cases} \mathcal{M}_{\phi_u}(\mathbf{x}) & \text{if } \hat{u} = 1, \\ \mathcal{M}_{\phi_n}(\mathbf{x}) & \text{if } \hat{u} = 0. \end{cases} \tag{1}$$

The hard-routing design lets each adapter specialize without compromising the other regime. It also makes the classifier a hard gate: a single misrouted query loads an adapter that has not seen its training distribution, with no graceful fallback. Section 6.4 examines the deployment-time cost of this brittleness empirically.

Approach 2: Prompt-conditioned unified adapter. A single LoRA adapter ϕ_{uni} is trained on the full dataset $\mathcal{D}_u \cup \mathcal{D}_n$, with urgency provided as a control tag prepended to the system prompt. The loss switches accordingly: $\mathcal{L}_{\text{urgent}}$ for `[URGENT]` examples and $\mathcal{L}_{\text{non-urgent}}$ for `[NON-URGENT]` examples (Section 4.3). This setup tests whether prompt conditioning alone is enough for a single model to learn risk-adaptive behaviour. Compared to Approach 1, it halves adapter storage and exposes both regimes to a single set of parameters during training, a design we expected to make the model more tolerant of routing noise but potentially less able to specialize on either regime. Whether this trade-off favours unified or bifurcated training is an empirical question we resolve in Sections 6.3–6.4.

4.3 Composite Loss

The composite loss is regime-specific. For urgent queries ($\hat{u} = 1$),

$$\mathcal{L}_{\text{urgent}} = \alpha\, \mathcal{L}_{\text{acc}} + \beta\, \mathcal{L}_{\text{util}} + \gamma\, \mathcal{L}_{\text{lat}}, \qquad \alpha + \beta + \gamma = 1, \quad \alpha, \beta, \gamma > 0. \tag{2}$$

For non-urgent queries ($\hat{u} = 0$), the latency term is dropped and the remaining weights are renormalized:

$$\mathcal{L}_{\text{non-urgent}} = \frac{\alpha}{\alpha + \beta}\, \mathcal{L}_{\text{acc}} + \frac{\beta}{\alpha + \beta}\, \mathcal{L}_{\text{util}}. \tag{3}$$

Accuracy. $\mathcal{L}_{\text{acc}}$ penalizes semantic mismatch between y and the closest reference in $\mathcal{Y}^*$:

$$\mathcal{L}_{\text{acc}}(y, \mathcal{Y}^*) = 1 - \max_{k \in [K]} S(y, y_k^*), \tag{4}$$

where $S(\cdot, \cdot) \in [0, 1]$ is token-overlap F1. The max over references rewards matching any valid answer, accommodating EgoBlind's up-to-four-answer reference sets.

Assistive utility. EgoBlind provides a $[0, 5]$ utility score capturing whether the answer is actionable for a BLV user, beyond lexical correctness. We treat its normalized complement as a loss:

$$\mathcal{L}_{\text{util}}(y, \mathcal{Y}^*) = 1 - \frac{\text{Score}(y, \mathcal{Y}^*)}{5}. \tag{5}$$

A verbose paraphrase and a terse direct answer can score similarly under S but differ sharply in utility, motivating the separate term.

Latency. For urgent queries, a correct answer arriving too late is no longer useful. We use response token count $|y|$ as a differentiable proxy for latency, with a piecewise penalty: zero below $\tau_{\min}$, polynomial growth with exponent p between $\tau_{\min}$ and $\tau_{\max}$, and a linear penalty with slope κ beyond $\tau_{\max}$ (closed form in Appendix C). The piecewise structure encodes a deliberate model of how latency interacts with usefulness: very short responses carry no latency cost, since brevity itself is a virtue; moderately long responses accumulate cost slowly, reflecting gradual usability degradation; and responses beyond $\tau_{\max}$ incur a steep penalty, modelling the regime in which the answer arrives too late to influence the user's next action. The term is dropped for non-urgent queries, where verbosity is acceptable. This asymmetry is what makes the loss *risk-adaptive* rather than merely accuracy-adaptive.

4.4 Optimization and Inference

Optimization. Both approaches are trained via SFT with LoRA [12] under the urgency-conditioned composite loss: Approach 1 trains separate adapters on $\mathcal{D}_u$ and $\mathcal{D}_n$ under their respective losses, whilst Approach 2 trains a single adapter on $\mathcal{D}_u \cup \mathcal{D}_n$ with the per-example loss switched by the prepended urgency tag.

We further apply DPO [14] to the urgent regime in both approaches: the urgent adapter in Approach 1 and the shared adapter restricted to $\mathcal{D}_u$ in Approach 2. Preference pairs are constructed by sampling k candidate responses per prompt and ranking them by a scoring function; the top-scoring response is treated as "chosen" and the bottom as "rejected." The choice of k, the scoring function, and any filtering criteria differ between the two approaches and are specified in Section 5.

Inference. Inference begins identically in both approaches: a CLIP-based classifier predicts urgency $\hat{u} = f_\theta(\mathbf{x}) \in \{0, 1\}$ from $\mathbf{x} = (v, q)$ using a 0.5 threshold. In Approach 1, $\hat{u}$ selects which adapter is loaded (Eq. 1) and the response is generated greedily from it. In Approach 2, $\hat{u}$ instead chooses the `[URGENT]` or `[NON-URGENT]` system-prompt tag for the single shared adapter.

Decoding is greedy in both approaches. We considered best-of-k sampling for the non-urgent regime, where the absence of a latency penalty admits richer search, but pilot experiments showed marginal gains relative to the added inference cost. Greedy decoding is also deterministic, a useful property in a safety-critical assistive setting.

The two approaches represent a trade-off between modularity and prompting: Approach 1 enforces regime separation structurally at the cost of doubled adapter storage, whilst Approach 2 uses a single adapter but relies on prompt conditioning for behavioural separation.

5 Experimental Methodology

Dataset. We use the EgoBlind dataset [2], which consists of egocentric video clips paired with natural-language questions from BLV users, each accompanied by ≤ 4 reference answers per question. EgoBlind organizes each question into 7 categories: information reading, safety warnings, navigation, social communication, tool use, communication & interaction, and other resources. Since the dataset does not include urgency annotations, we augment it by labeling each question as either urgent or non-urgent from the perspective of a BLV user. To do so, we prompt GPT-5.2 (`gpt-5.2-2025-12-11`, `reasoning_effort=xhigh`) with five sampled frames per video clip together with the corresponding question (full prompt in Appendix B). We validate this annotation procedure by manually reviewing a random subset of 300 questions, finding over 90% agreement with the model-generated labels.

Data splits. The training set contains 2,746 questions across 922 videos (1,474 urgent and 1,272 non-urgent), while the test set consists of 1,283 questions (668 urgent and 615 non-urgent). To train

the CLIP-based urgency classifier, we further split the training data into stratified 80/20 train and validation sets to preserve the class distribution.

Uniform-policy baseline. We evaluate a uniform-policy baseline using `kimi-vl-a3b-instruct` (Moonshot API). For each test question, we provide frames extracted from the corresponding video together with the question text, and use a single urgency-agnostic system prompt instructing the model to respond as a blind person answering a first-person query (full prompt in Appendix A). The same prompt is applied to all inputs, regardless of urgency, yielding a strict uniform-policy baseline against which urgency-aware fine-tuning is compared. We use a temperature of 0.2 and set `max_tokens` to 256.

CLIP urgency classifier. The classifier is built on a frozen `CLIP ViT-B/32` backbone (OpenAI weights). For each query timestamp t_q, we sample $K = 4$ frames uniformly from a ± 2-second window around t_q, mean-pool their CLIP image embeddings, and concatenate the result with the CLIP text embedding of the question. A two-layer MLP head takes this fused representation and predicts urgency. The CLIP backbone is frozen; only the MLP is trained, with AdamW (lr $= 10^{-4}$), binary cross-entropy loss, and 5 epochs on a single NVIDIA L40S GPU.

Fine-tuning. We fine-tune Kimi-VL-A3B-Instruct [10] via LoRA [12] on MIT's Engaging cluster (NVIDIA L40S GPUs) for both architectural variants.

Approach 1. Two separate LoRA adapters (rank 8, $\alpha = 16$, targeting `q_proj` and `v_proj`) are trained on urgency-stratified subsets using the `kimi_vl_nothink` template. The urgent adapter trains on $\mathcal{D}_u$ ($n = 1{,}474$) with AdamW (lr $= 1 \times 10^{-4}$), batch size 1, gradient accumulation 8, and 1 epoch in bf16; each example includes one video frame and the shortest reference answer as target. The non-urgent adapter trains on $\mathcal{D}_n$ ($n = 1{,}272$) for 1 epoch under the same settings, except that the learning rate is reduced by $5\times$ to 2×10^{-5} (the higher rate caused gradient divergence on longer targets), with the longest reference answer as target. DPO pairs are generated by sampling $k = 4$ responses, scoring by token-overlap F1 against the longest reference, and selecting the highest- and lowest-scoring as chosen/rejected ($\beta = 0.1$, lr $= 5 \times 10^{-6}$, 3 epochs). For the non-urgent adapter, DPO is applied via stacked adapters: the SFT adapter is merged into the base before the DPO adapter is applied as a delta.

Approach 2. A single LoRA adapter (rank 8, $\alpha = 16$, dropout 0.05, targeting `q_proj` and `v_proj`) is trained under 4-bit quantization with AdamW (cosine schedule, lr 1×10^{-4}, warmup ratio 0.1, weight decay 0.01), per-device batch size 1, gradient accumulation over 8 steps, and 1 epoch in bf16. The dataset is vision-conditioned (`cutoff_len` 2048) with urgency tags prepended to each example; for urgent questions the SFT target is the *shortest* correct reference answer, and for non-urgent questions the *longest*. We experiment with three DPO variants on top of the SFT checkpoint: (i) composite-loss-scored preference pairs ($k = 8$ candidates, three-filter selection, 975 retained pairs); (ii) token-overlap-F1-scored pairs ($k = 4$) from the SFT checkpoint; and (iii) token-overlap-F1-scored pairs from the base model. All DPO runs use $\beta = 0.1$, sigmoid loss, learning rate 5×10^{-6}, and 3 epochs.

Evaluation metrics. For the urgency classifier we report accuracy, precision, recall, F1, and ROC-AUC on both validation and test partitions. For the response generator we report the composite loss $\mathcal{L}$ along with mean output length on the EgoBlind test set. Following [2], we instantiate the similarity function $S(\cdot, \cdot)$ in Section 4.3 as sentence-transformer cosine (`all-MiniLM-L6-v2`) over the full reference set, applied uniformly across all composite-loss tables for cross-method comparison. Per-regime similarity in Approach 1's adapter-comparison tables (Table 2, Table 3) is additionally reported as token-overlap F1 against the closest reference, matching the convention in prior bifurcated-adapter ablations. We score under the canonical grid-search-optimal hyperparameters $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$, $\tau_{\min} = 8$, $\tau_{\max} = 40$, $p = 3$, $\kappa = 0.3$, selected by minimizing overall test loss across 3,888 configurations (486 satisfying $\alpha + \beta + \gamma = 1$).

6 Results and Discussion

6.1 CLIP Urgency Classifier

Results. Table 1 reports classifier performance on both splits.

Table 1: CLIP urgency classifier performance on validation and test splits.

Metric	Validation	Test
Accuracy	0.863	0.798
Precision	0.930	0.879
Recall	0.807	0.659
F1	0.864	0.759
ROC-AUC	0.938	0.905

Discussion. The strong ROC-AUC across both splits indicates that the frozen CLIP backbone captures a meaningful urgency signal from the joint video–text representation without end-to-end fine-tuning. The performance gap between validation (86.3%) and test (79.8%) is expected given that the test set contains unseen videos. The main limitation is recall on urgent cases: the model misses approximately 30% of urgent queries on the test set, misclassifying them as non-urgent, which is the most costly error in a safety-critical setting. Improving recall via threshold tuning, asymmetric loss weighting, or partial backbone unfreezing is left for future work (Section 7).

6.2 Baseline Evaluation

Table 6 reports overall performance: a composite loss of 0.607, with mean output length of 84.1 words and similarity 0.350. While the baseline is not inaccurate per se (similarity is non-trivial), it is substantially over-verbose, producing responses exceeding 80 words for questions whose reference answers are typically only a few words long. As a result, the latency component of the composite loss dominates, indicating that prompt-only urgency conditioning is insufficient to induce concise responses for safety-critical queries.

Per-category analysis. Table 7 breaks down the best-configuration composite loss by question type. Safety warnings have the highest loss (1.118), followed by tool use (0.948) and navigation (0.787). Information reading and social communication, which are predominantly non-urgent, exhibit the lowest losses. This pattern aligns with EgoBlind [2], which finds that models perform worst on categories where errors are most consequential, further motivating urgency-aware approaches.

Composite loss grid search. We evaluate the baseline across 3,888 hyperparameter configurations of the composite loss $(\alpha, \beta, \gamma, \tau_{\min}, \tau_{\max}, p, \kappa)$, of which 486 satisfy the normalization constraint $\alpha + \beta + \gamma = 1$. Among constrained settings, the best test loss is achieved with $\alpha = 0.5, \beta = 0.3, \gamma = 0.2, \tau_{\min} = 8, \tau_{\max} = 40, p = 3, \kappa = 0.3$. The top 5 configurations share $\alpha = 0.5, \beta = 0.3, \gamma = 0.2$ and $\tau_{\max} = 40$, differing only in $\tau_{\min}$ and p, suggesting relative insensitivity to latency-regime boundaries but sensitivity to the accuracy–latency trade-off weights. The optimal $\gamma = 0.2$ assigns non-trivial weight to latency, indicating that brevity is important and supporting our design choice to decouple latency penalties for non-urgent queries. We use this configuration as the canonical scoring throughout the paper.

6.3 SFT Results

6.3.1 Approach 1

Training the bifurcated adapters required resolving two issues beyond the template fix shared with Approach 2. First, text-only training (no video frames) produced degenerate outputs across all configurations: the best text-only adapter generated "Yes." or "No." for 71% of urgent queries, pattern-matching common short references without visual grounding. Including a single video frame per example resolved this. Second, aggressive LoRA configurations (rank 16, all projection targets,

$\geq$2 epochs) consistently overfitted; adopting the lighter configuration from Approach 2 stabilized training. Full ablation details appear in Appendix E.

With vision-conditioned SFT, the urgent adapter achieves 0.505 token-F1 similarity on urgent test examples and the non-urgent adapter achieves 0.268 on non-urgent examples. Under oracle routing, the combined system reaches 0.391 overall similarity (Table 2), nearly $4\times$ the uniform-policy baseline (0.103), while reducing mean output length from 84.1 to 4.0 words.

6.3.2 Approach 2

Two implementation issues initially impeded vision-conditioned fine-tuning. First, the LLaMA-Factory [21] training pipeline applied a default chat template incompatible with Kimi-VL's native format, producing mis-tokenized inputs during training and degenerate outputs at inference. We resolve this by registering a custom `kimi_vl_nothink` template that matches the model's expected dialogue structure. Second, the released `modeling_kimi_vl.py` is incompatible with `transformers` ≥ 4.50 across several API mismatches; we patch these in both the downloaded model and the runtime module cache via an idempotent setup script.[2] On applying these fixes, optimization becomes stable and training loss decreases monotonically.

Table 4 reports SFT against the uniform-policy baseline. SFT reduces composite loss from 0.607 to 0.375 (a 38.2% relative improvement), drops mean output length from 84.1 to 2.2 words, and raises reference similarity from 0.350 to 0.589. Urgency-conditioned SFT therefore enforces concise outputs without sacrificing accuracy; in fact, brevity and similarity improve together.

6.4 DPO Results

6.4.1 Approach 1

We evaluate four DPO configurations — initialized from SFT or from the base model, for each urgency regime — and a final mixed configuration. Results are summarized in Table 2.

Table 2: Approach 1 training method comparison (token-overlap F1 similarity on test set, oracle routing).

Method	**Urgent**	**Non-urg.**	**Overall**
Baseline (uniform policy)	0.124	0.080	0.103
SFT only	0.505	0.268	0.391
DPO only (from base)	0.277	0.270	0.273
SFT followed by DPO	0.459	0.298	0.381
Best mix (SFT urgent, SFT+DPO non-urgent)	0.507	0.298	0.407

For the urgent adapter, DPO on SFT degrades performance (0.505 $\rightarrow$ 0.459), consistent with the regression observed in Approach 2. DPO from base improves over the untrained model (0.124 $\rightarrow$ 0.277) but remains far below SFT. For the non-urgent adapter, DPO on SFT yields a modest gain (0.268 $\rightarrow$ 0.298). The best overall result *mixes* the SFT urgent adapter with the SFT + DPO non-urgent adapter (0.407 overall).

Cross-architecture comparison. DPO appears to degrade the urgent adapter in both approaches for the same hypothesized reason: urgent reference answers are typically 1–3 words long, so sampled preference pairs contain minimal variation, and DPO has little meaningful signal to learn from. For non-urgent queries, reference answers are substantially longer and more varied, giving DPO enough contrast between chosen and rejected responses to produce a genuine improvement. This explains why the bifurcated design can benefit from DPO — its non-urgent adapter trains exclusively on long-answer examples — while the unified design cannot, since its DPO pairs mix both regimes and the short-answer urgent examples dilute the signal.

[2] Specifics in the project repository.

End-to-end pipeline evaluation. We evaluate the full Approach 1 pipeline by replacing oracle routing with the CLIP classifier.

Table 3: End-to-end Approach 1 pipeline under different routing strategies (test set, $n = 1{,}283$).

Routing strategy	Loss	Similarity	Mean words
Baseline (uniform policy)	0.607	0.350	84.1
Zero-shot (prompted Kimi-VL)	0.597	0.346	5.1
CLIP classifier	**0.556**	**0.393**	4.5
Oracle (ground-truth labels)	0.549	0.399	4.0

A zero-shot baseline (prompting base Kimi-VL to classify urgency) achieves only 0.45% recall on urgent queries, effectively collapsing the bifurcated design. The trained CLIP classifier achieves 78.1% accuracy and 65.9% recall (F1 = 0.759), raising pipeline similarity from 0.346 (zero-shot) to 0.393 and closing 88% of the gap to oracle routing (0.399). The remaining gap is attributable to 228 false negatives: urgent queries misrouted to the non-urgent adapter. The classifier, not just the response generators, is also a primary pipeline bottleneck.

6.4.2 Approach 2

We apply DPO [14] on top of the SFT checkpoint to align the urgent-regime adapter with the urgency-aware composite reward. We train three variants to assess whether DPO improves over SFT, and whether any gains are robust to the choice of pair scoring and initialization.

Composite-loss-scored pairs. For each urgent training example, we sample $k = 8$ responses from the SFT model (temperature 0.7) and score them using $\mathcal{L}_{\text{urgent}}$. The lowest-loss response is treated as the "chosen" and the highest-loss as the "rejected." We apply three filtering criteria to remove low-quality pairs: identical responses, loss_gap < 0.2, and yes/no polarity mismatches. This yields 975 preference pairs. DPO is trained with $\beta = 0.1$, sigmoid loss, learning rate 5×10^{-6}, and 3 epochs.

Token-overlap F1 pairs. We also consider a simpler scheme where $k = 4$ sampled responses are scored using token-overlap F1 against the longest reference. The highest-scoring response is used as the "chosen" and the lowest as the "rejected." We only filter identical-candidate pairs. We evaluate this setup from both the SFT checkpoint and the base model to isolate the effect of initialization on DPO performance.

Table 4: Final composite-loss comparison across all methods on the test set ($n = 1{,}283$). All three DPO variants regress the SFT result, regardless of pair-scoring method or initialization. Mean words and similarity are reported alongside loss to characterize the regression. Percentages are computed from unrounded values.

Method	Loss	Δ vs SFT	Mean words	Similarity
Baseline (few-shot prompted Kimi-VL)	0.607	$+61.6\%$	84.1	0.350
SFT (vision LoRA)	**0.375**	—	2.2	0.589
SFT $\rightarrow$ DPO (composite-loss pairs)	0.433	$+15.5\%$	7.3	0.522
SFT $\rightarrow$ DPO (token-overlap F1 pairs)	0.526	$+40.1\%$	18.2	0.424
Baseline $\rightarrow$ DPO (token-overlap F1 pairs)	0.543	$+44.7\%$	25.2	0.407

Results. Every DPO variant regresses SFT (Table 4). Composite-loss DPO from SFT increases loss by 15.5% (0.375 $\rightarrow$ 0.433); F1 DPO from SFT lands at +40.1% (0.526); F1 DPO from the base model is worst at +44.7% (0.543). The regression is robust to pair-scoring method and to initialization.

Discussion. Two diagnostic signals explain the regression. First, verbosity drift is uniform across question types: mean output length jumps from 2.2 (SFT) to 7.3 words, with near-identical drift across yes/no ($+4.8$), what ($+5.2$), where ($+5.9$), and how ($+6.8$) questions. DPO learned a global

"be wordy" preference, not a question-conditional length policy. Second, the reward components fight each other: training pairs pointed chosen toward higher similarity *and* shorter length, yet the DPO model produces both longer outputs and lower similarity than SFT, degrading on both axes despite preferences pointing the opposite way.

For short-form egocentric VQA where references provide strong supervision, SFT against the shortest correct reference is a stronger objective than DPO. SFT enforces brevity per example, while DPO only enforces a relative ranking between two candidates — a much weaker signal when the optimum is a specific short answer rather than just a "shorter than" relation. The failure generalizes across pair-scoring schemes, with F1-scored DPO regressing further than composite-loss DPO, so it is not specific to our reward formulation.

6.4.3 Approach 2 End-to-End Pipeline

We evaluate the full Approach 2 pipeline by replacing oracle urgency labels with CLIP-predicted ones at inference time, holding the SFT model fixed.

Table 5: End-to-end Approach 2 pipeline under different routing strategies (test set, $n = 1{,}283$). The unified prompt-conditioned adapter is robust to CLIP routing errors: CLIP-predicted urgency tags at deployment match oracle GT routing under composite loss, despite 21.5% test-time tag flips.

Routing strategy	**Loss**	**Similarity**	**Mean words**
Baseline (uniform policy)	0.607	0.350	84.1
CLIP classifier	**0.375**	**0.602**	14.0
Oracle (ground-truth labels)	0.375	0.589	2.2

CLIP routing achieves the same composite loss as oracle routing (0.375 vs. 0.375) despite disagreeing with the ground-truth urgency on 21.5% of test examples. The unified adapter has been trained on both urgent and non-urgent examples and so handles either tag competently; a wrong tag at inference shifts response style (terse vs. detailed) but rarely changes whether the answer is correct. This contrasts with Approach 1, where a misrouted urgent query loads an adapter that has never seen urgent training data and produces a substantively worse response. The implication is that further improvements to the urgency classifier would yield little gain for Approach 2, whereas they remain critical for Approach 1.

7 Conclusion and Future Directions

We present **EgoBlind-RA**, a risk-adaptive framework for egocentric BLV visual assistance. The system combines a CLIP-based urgency classifier (test ROC-AUC = 0.905) with two urgency-conditioned LoRA-fine-tuned variants of Kimi-VL-A3B-Instruct: a bifurcated design with separate urgent and non-urgent adapters, and a unified design conditioned via prompt tags.

In the bifurcated-adapter setting (Approach 1), oracle routing achieves a roughly $4\times$ improvement in similarity over the baseline, and the trained CLIP classifier closes 88% of the gap between naïve zero-shot routing and oracle routing, yielding a deployable pipeline within 1.3% of the oracle ceiling. Unlike Approach 2, DPO is partially productive: it degrades the urgent adapter but improves the non-urgent one, and the best overall result mixes SFT and DPO adapters. This asymmetry suggests that preference-pair alignment requires regime-homogeneous training data, a condition satisfied by bifurcated adapters but not by unified training.

In the unified-adapter setting (Approach 2), urgency-conditioned SFT against carefully selected reference targets reduces composite loss by 38.1% over a uniform-policy baseline, with mean output length decreasing from 84.1 to 2.2 words and similarity increasing from 0.350 to 0.589. In contrast, applying DPO on top of SFT consistently degrades performance across pair-scoring strategies and initialization schemes. Our analysis attributes this regression to a uniform verbosity drift across question types and to a tension between the similarity and latency objectives that DPO's pairwise ranking cannot resolve at the per-example level.

Together, these results suggest that for short-form egocentric VQA with strong reference supervision, carefully constructed SFT objectives outperform preference-pair-based alignment, and that both alignment method and classifier-robustness requirements depend strongly on adapter architecture.

Limitations. Both similarity metrics we report — token-overlap F1 and sentence-transformer cosine — are coarse proxies for semantic correctness; metrics like BERTScore or LLM-based judges would be more reliable but were prohibitive at our scale. The CLIP classifier uses standard binary cross-entropy and a fixed 0.5 threshold, despite false negatives being more costly than false positives in this setting. Our evaluation also lacks human-in-the-loop studies with BLV users, which would be needed to validate real-world utility.

Approach 1's primary limitation is classifier dependence: $\sim$30% of urgent queries are misrouted to the non-urgent adapter, where the system degrades gracefully but with added latency. The bifurcated design also doubles adapter storage. Approach 2 exhibits the opposite limitation: it handles classifier noise gracefully at inference, but the latency penalty biases SFT toward terse outputs (mean 2.2 words), with a non-trivial fraction of *"I don't know"* refusals. While the composite loss may reward this behaviour, BLV users may not.

Future directions. For Approach 1, threshold tuning to favour recall would directly improve urgent-query handling, since false negatives are more costly than false positives in this setting. Multi-frame inference at the response generator (matching the classifier's four-frame setup) may improve grounding. Finally, GRPO [17], which maintains per-example supervision, may offer a more principled alternative to DPO for the urgent regime.

For Approach 2, two extensions are most relevant. First, the DPO regression points toward methods that retain per-example supervision rather than pairwise preference. Candidates include reward-weighted SFT and teacher-model distillation. Second, while the unified adapter is robust to noisy classifier routing at deployment, training-time label noise hurts more substantially: an ablation in which SFT trains on CLIP-predicted rather than ground-truth urgency tags incurs an 18.7% composite-loss regression. Effort spent improving the urgency classifier therefore yields larger returns at training time than at inference time, the opposite of Approach 1, where deployment-time recall is the dominant bottleneck.

8 Code and Reproducibility

The implementation of **EgoBlind-RA** is publicly available on GitHub at github.com/juliavekim/EgoBlind-RA. Individual course repositories, containing course assignments and the EgoBlind-RA final project, are also available: Julia's and Xander's. LoRA adapters for Approach 1 are available on HuggingFace at xabackus/egoblind-ra-adapters.

References

[1] Rupert R.A. Bourne et al. Magnitude, temporal trends, and projections of the global prevalence of blindness and distance and near vision impairment: a systematic review and meta-analysis. *The Lancet Global Health*, 5(9):e888–e897, 2017.

[2] Junbin Xiao et al. Egoblind: Towards egocentric visual assistance for the blind. *arXiv preprint arXiv:2503.08221*, 2025.

[3] Jeffrey P. Bigham et al. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 23nd Annual ACM Symposium on User Interface Software and Technology*, 2010.

[4] Haotian Liu et al. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[5] Wenliang Dai et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[6] Danna Gurari et al. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[7] Kristen Grauman et al. Ego4d: Around the world in 3,000 hours of egocentric video. *arXiv preprint arXiv:2110.07058*, 2021.

[8] Abhishek Das et al. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[9] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[10] Moonshot AI. Kimi-vl technical report. 2025. `https://github.com/MoonshotAI/Kimi-VL`.

[11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[12] Edward J. Hu et al. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2022.

[13] Long Ouyang et al. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*, 2022.

[14] Rafael Rafailov et al. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*, 2023.

[15] Kawin Ethayarajh, Winnie Xu, Niklas Muennighoff, Dan Jurafsky, and Douwe Kiela. Kto: Model alignment as prospect theoretic optimization. *arXiv preprint arXiv:2402.01306*, 2024.

[16] Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. Rlaif: Scaling reinforcement learning from human feedback with ai feedback. *arXiv preprint arXiv:2309.00267*, 2023.

[17] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y.K. Li, Y. Wu, and Daya Guo. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[18] Spencer Whitehead et al. Reliable visual question answering: Abstain rather than answer incorrectly. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

[19] Lorenz Kuhn et al. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

[20] Tal Schuster et al. Confident adaptive language modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[21] Yaowei Zheng et al. Llama-factory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*, 2024.

A Baseline Prompt

The uniform-policy baseline uses a single system prompt for every test question, regardless of the GPT-5.2-assigned urgency label. The prompt is adapted from the EgoBlind benchmark's evaluation setup [2].

```
I want you to act as a blind person. I will provide you with
a question raised by a blind person from their first-person
perspective in the current scene. Your task is to answer the blind
person's question. The answer needs to be based on the content of
the picture and the objective characteristics that the blind person
cannot see. If the question cannot be answered, you can say I don't
know. Do not include Chinese characters in your response. The
question is: {question}
```

B Urgency Annotation Prompt

We annotate the EgoBlind dataset with urgency labels using GPT-5.2 (`gpt-5.2-2025-12-11`, `reasoning_effort=xhigh`). For each question, the model receives five evenly-spaced video frames alongside the question text, and is prompted to return either `urgent` or `not_urgent`. The complete system prompt is reproduced below.

```
You are a classifier for the EgoBlind dataset, which contains
egocentric video from blind or visually impaired people navigating
the real world. You must classify each question as 'urgent' or
'not_urgent'.

A question is URGENT if reducing response latency would significantly
improve safety, prevent physical harm, or prevent irreversible
loss or missed opportunity -- meaning speed is more important than
perfect accuracy in this situation.

Classify as URGENT when:

  • The person is about to take an immediate physical action
    (stepping, crossing, touching, boarding, pouring, cutting,
    exiting a vehicle, etc.)
  • The answer will immediately influence the person's next physical
    movement or decision
  • The environment is dynamic or changing (moving vehicles, closing
    doors, approaching obstacles, active traffic)
  • There is imminent risk of injury (traffic, falling, tripping,
    burns, sharp objects)
  • There is risk of losing something or missing a time-sensitive
    opportunity (vehicle leaving, item left behind, boarding ending)

Classify as NOT_URGENT when:

  • Accuracy is more important than speed
  • The question is descriptive or exploratory
  • It involves reading text, identifying medications or products,
    or verifying information where correctness matters more than
    immediacy
  • It concerns planning or orientation without imminent movement or
    time pressure
  • Delay would not meaningfully change the outcome

Use the video context to determine whether the situation involves
immediate action, dynamic change, or time-sensitive consequences.

Respond with ONLY one word: urgent or not_urgent.
```

The user-side prompt accompanying each set of frames is:

```
These are frames from an egocentric video recorded by a blind
person (start time: {start_time}s). They are asking the following
question:

Question: {question}

Is this person in an immediate action-risk situation where delay
increases danger, or is this a time-sensitive question where a slow
response makes the answer useless? Respond with ONLY: urgent or
not_urgent.
```

C Loss Components

The closed form of the latency penalty referenced in Section 4.3 is

$$\mathcal{L}_{\text{lat}}(y) = \begin{cases} 0 & \text{if } |y| \le \tau_{\min} \\ \left(\dfrac{|y| - \tau_{\min}}{\tau_{\max} - \tau_{\min}} \right)^{p} & \text{if } \tau_{\min} < |y| \le \tau_{\max} \\ 1 + \kappa \left(|y| - \tau_{\max} \right) & \text{if } |y| > \tau_{\max} \end{cases}, \tag{6}$$

where $p \ge 1$ controls curvature in the transition zone and $\kappa > 0$ is the overflow slope. The grid-searched values used throughout the paper are $\tau_{\min} = 8$, $\tau_{\max} = 40$, $p = 3$, $\kappa = 0.3$.

D Tables

Table 6: Uniform-policy baseline (`moonshot-v1-32k-vision-preview`) performance on the test set under the constrained-optimal grid-search configuration.

Metric	Value
N	1,283
Composite loss	0.607
Mean output length	84.1 words
Similarity (token F1)	0.350

Table 7: Per-category composite loss under the constrained-optimal grid-search configuration ($\alpha = 0.5, \beta = 0.3, \gamma = 0.2, \tau_{\min} = 8, \tau_{\max} = 40, p = 3, \kappa = 0.3$; baseline, test set).

Question Type	Loss	Similarity	N
Safety warnings	1.118	0.118	946
Tool use	0.948	0.139	211
Navigation	0.787	0.117	363
Other resources	0.501	0.112	222
Information reading	0.489	0.089	1,873
Social communication	0.459	0.065	37
Communication & interaction	0.450	0.065	81

E Approach 1 Training Ablations

Table 8 summarizes the progression of Approach 1 training configurations. The first three runs used text-only inputs; all produced degenerate outputs despite low training loss. SFT v1 used the default `kimi_vl` template, which injects thinking tokens (◁`think`▷/◁`/think`▷) that corrupted generation. SFT v2–v3 fixed the template but retained aggressive LoRA settings (rank 16, all 7 projection targets), causing catastrophic overfitting. SFT v4 adopted the lighter configuration from Approach 2 (rank 8, `q_proj+v_proj`, 1 epoch) and produced coherent text-only outputs, but 71% of urgent predictions collapsed to "Yes." or "No."; pattern-matching without visual grounding. Vision-conditioned SFT

resolved all issues. The non-urgent adapter additionally required a $5\times$ learning-rate reduction ($1 \times 10^{-4} \rightarrow 2 \times 10^{-5}$); at the higher rate, gradient norms diverged to NaN within the first epoch.

Table 8: Approach 1 SFT ablation summary. All runs use Kimi-VL-A3B-Instruct.

Run	**Template**	**LoRA**	**Vision**	**Epochs**	**Outcome**
v1	`kimi_vl`	rank 16, all 7	No	5	Thinking-token garbage
v2	`nothink`	rank 16, all 7	No	2	Repetition loops
v3	`nothink`	rank 16, all 7	No	2	Still degenerate
v4	`nothink`	rank 8, q+v	No	1	Yes/No collapse (71%)
Final	`nothink`	rank 8, q+v	**Yes**	1	**Coherent (0.505 sim)**