LOOM: LOW-LATENCY ONLINE OBSTACLE MONITORING FOR EGOCENTRIC AUDIO-VISUAL INTRUSION ANTICIPATION

Anonymous authors
Paper under double-blind review

ABSTRACT

For wearable assistive systems designed for people with visual impairments, timely alerting users to surrounding hazards, critical events, and key objects is paramount to ensuring safe navigation and independent mobility. We formulate egocentric audio-visual near-body intrusion anticipation: predicting whether an object will newly enter a region extending 0.40 m beyond the body within three seconds, using historical monocular RGB and simulated two-channel audio. We introduce LOOM, a lightweight framework combining visual motion and expansion statistics, a learned acoustic sequence model, development-selected probability fusion, and stateful warning generation. We construct ReactionGap, comprising 150 simulated episodes across ten motion families, with fixed wearer-centered viewpoints, synchronized audio-visual observations, and geometric trajectories. Its entering and non-entering scenes provide 3,900 labeled queries for training and template-disjoint evaluation of future-entry prediction and sequential warnings. Across three seeds, LOOM achieves pooled average precision of 0.5438, compared with 0.5234 for vision and 0.4776 for audio. Median computation takes 8.8 ms, compared with 54.1–72.4 ms for the evaluated temporal and fusion adaptations. Warning replay yields one-second recall of 0.4178 at 0.7289 non-useful alarms per episode; vision retains stronger time-matched discrimination. These results highlight a critical discrepancy between pure model prediction and actionable warning performance, establishing ReactionGap as a standardized benchmark to advance robust, real-time wearable assistance.

1 INTRODUCTION

Wearable cameras and microphones offer a means of sensing nearby hazards for people with visual impairments. Effective assistance requires more than recognizing an obstacle: a system must anticipate whether it will approach the wearer and issue a warning while time remains to respond. A moving door, rolling load, or passing vehicle can be visually prominent without entering the space immediately surrounding the body. Conversely, a brief approach may demand a timely warning even when the object is difficult to identify. Clinical work on wearable collision warnings highlights the importance of warning timing (Pundlik et al., 2021), while recent egocentric assistance systems address scene understanding, walking guidance, and trajectory prediction (Haghighi et al., 2024; Yuan et al., 2025; Xiao et al., 2025). Together, these applications motivate studying how historical sensory evidence can support near-body event anticipation.

We formulate this problem as *egocentric audio-visual near-body intrusion anticipation*. Given historical monocular RGB and two-channel audio, the model predicts whether an object will newly enter a body-adjacent region during a future interval. This formulation connects a geometric event to an actionable temporal objective: the prediction must precede the entry, and a delivered warning must provide sufficient lead time. It complements accident anticipation, which forecasts scene-level events (Zhao et al., 2025; Patera et al., 2026), and time-to-collision estimation, which models the temporal relationship between an observer and approaching surfaces (Manglik et al., 2019; Badki et al., 2021). In our formulation, an object's current presence near the body and its future entry are distinct. This distinction permits direct evaluation of whether a warning anticipates an event, arrives late, or repeats information already conveyed.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The central challenge is to interpret incomplete and heterogeneous sensory evidence. Visual changes reveal movement and apparent expansion, but lateral passage, changing visibility, and approach can produce similar image differences. Sound can provide evidence about motion and contact, yet level alone also depends on the source and recording conditions. Evidence from acoustic looming and auditory collision judgments motivates including temporal sound cues (Neuhoff, 1998; Oberfeld et al., 2022). A multimodal predictor must therefore combine branches whose reliability can vary across scene types. The downstream objective adds another challenge: improved ranking of isolated queries need not yield earlier or less repetitive warnings. Closely spaced predictions concern overlapping future intervals, so an event-level evaluation must account for their sequential use and computational delay.

We introduce LOOM to connect efficient audio-visual prediction with sequential warning generation. The visual branch summarizes the spatial support, location, and expansion of image changes; a convolutional recurrent audio branch represents spectral evolution and stereo level dynamics. Independently trained branches produce probabilities combined with a development-selected weight. A stateful alarm layer converts the fused scores into warnings. Compact visual statistics and a small acoustic model reduce computation along this prediction-to-warning path. Our experiments use a 0.40 m body-surface margin and a three-second forecast; Section 3 gives the task definition and input constraints.

To support this task, we construct ReactionGap: 150 simulated episodes across ten motion families, pairing RGB from fixed wearer-centered viewpoints and simulated two-channel audio with geometric trajectories. The dataset includes entering and non-entering scenes, enabling models to learn the distinction between future intrusion and nearby motion that stays outside. Geometry-derived event times provide 3,900 query labels and support evaluating warning lead time and burden. Template-disjoint partitions test transfer to unseen motion templates. Across three seeds, LOOM achieves pooled average precision (AP) of 0.5438, versus 0.5234 for vision and 0.4776 for audio. Its median computation time is 8.8 ms on an RTX 4090, versus 54.1–72.4 ms for the evaluated temporal and fusion adaptations. Vision remains stronger under time-matched discrimination, so we report query-level and event-level performance together. The main contributions of this work are summarized as follows:

- We formulate egocentric audio-visual near-body intrusion anticipation for visual impairment assistance, linking historical monocular RGB and two-channel audio to future entry events and advance warnings.
- We develop LOOM, a lightweight multimodal framework combining visual motion and expansion, a learned acoustic sequence model, development-selected probability fusion, and stateful warnings. Its 8.8 ms median computation time gives a 6.1–8.2$\times$ speedup over the evaluated temporal and fusion adaptations under the same timing protocol, supporting prompt prediction updates.
- We construct ReactionGap, a dataset of 150 simulated audio-visual episodes with synchronized observations, geometric trajectories, and entry-event annotations. Entering and non-entering scenes, together with template-disjoint partitions, support training intrusion predictors, testing transfer across motion templates, and evaluating advance warnings. We will publicly release the dataset and code upon acceptance.

The rest of this paper is organized as follows. Section 2 reviews related work. Section 3 presents the task formulation, architecture, and key components of LOOM. Section 4 describes the experimental setup and evaluates prediction performance, ablations, warning behavior, and computational efficiency on ReactionGap. Finally, Section 5 concludes the paper.

2 Related Work

Visual impairment assistance and proximity warnings. Egocentric assistance encompasses understanding the environment, anticipating motion, and communicating relevant hazards. HEADS-UP studies head-mounted trajectory prediction for blind assistance (Haghighi et al., 2024); WalkVLM addresses walking assistance with vision-language models (Yuan et al., 2025); and EgoBlind provides a benchmark for egocentric visual assistance (Xiao et al., 2025). These settings establish the relevance of wearer-centered perception, while our prediction target specifies a future crossing

of a body-adjacent boundary. ARAware identifies critical moving objects using object-dependent collision zones (Surougi et al., 2024), and wearable collision-warning studies evaluate time-based alerts in everyday use (Pundlik et al., 2021). Acoustic sensing offers an additional route: PAWS uses wearable microphones for pedestrian safety (de Godoy et al., 2018). Perceptual studies of rising sounds and quiet-vehicle time-to-collision judgments further connect acoustic dynamics with perceived approach (Neuhoff, 1998; Oberfeld et al., 2022). We use these observations to motivate temporal audio modeling alongside visual motion, and evaluate its contribution through predictions and delivered warnings.

Temporal anticipation and motion evidence. Accident anticipation methods learn to recognize evidence preceding a hazardous outcome. TOP models temporal occurrence (Zhao et al., 2025), RiskProp uses collision-anchored risk propagation (Zou et al., 2026), and TSD studies lightweight spatio-temporal modeling through temporally shifted distillation (Patera et al., 2026). Action anticipation similarly depends on extracting predictive information from past observations: AGA uses predicted action sequences to guide temporal attention (Tai et al., 2026). TRN models relations among sparse frames at multiple temporal scales (Zhou et al., 2018). StillMix and TRoVe examine static-feature biases in temporal representations (Li et al., 2023; Varma et al., 2025), motivating scrutiny of what a predictive score captures. More geometric approaches estimate time to collision from monocular video (Manglik et al., 2019) or classify temporal geofences (Badki et al., 2021). Bio-inspired looming models emphasize expansion and collision-relevant motion (Zhao et al., 2023; Liu & Fu, 2025), while WAFT estimates dense optical flow through warping (Wang & Deng, 2026). Our visual branch uses compact image-change and expansion statistics as evidence for supervised entry prediction. Time-matched evaluation complements this design by comparing queries at the same observation time, reducing the influence of temporal prevalence on discrimination scores.

Audio-visual representation learning and fusion. Audio-visual learning can combine complementary observations at the representation or prediction level. Tensor Fusion Networks (TFN) explicitly form cross-modal interactions through outer products (Zadeh et al., 2017), while low-rank multimodal fusion (LMF) factorizes these interactions for efficiency (Liu et al., 2018). MMTM instead transfers information through joint squeeze and channel excitation (Vaezi Joze et al., 2020). CAV-MAE combines contrastive learning with masked reconstruction (Gong et al., 2023), and teacher-guided dual-path learning addresses semantic noise in audio-visual representations (Wang et al., 2026). For anticipation, AFFT learns to combine multimodal features through a transformer (Zhong et al., 2023). Fusion is also sensitive to modality quality: QMF analyzes dynamic fusion for low-quality inputs (Zhang et al., 2023), and PDF develops predictive dynamic fusion (Cao et al., 2024). These methods motivate treating the contribution of each modality as a modeling decision rather than assuming that concatenation is sufficient. LOOM uses independently supervised branches and a fixed probability weight selected on development data. We compare this prediction-level combination with learned tensor, low-rank, and channel-transfer fusion under the same event definition. Our experiments examine how these choices affect both query-level ranking and the subsequent warning sequence.

3 METHOD

3.1 TASK FORMULATION

Let $\mathcal{B}$ denote the modeled body and define its surrounding region as

$$\mathcal{R}_m = \{x : \mathrm{dist}(x, \mathcal{B}) \le m\}, \qquad m = 0.40\,\mathrm{m}. \tag{1}$$

The distance is measured from the body surface. Ground-truth object trajectories determine the set $\mathcal{E}$ of outside-to-inside entry times. At query time t_0, the target is

$$y(t_0) = \mathbf{1}\{\exists e \in \mathcal{E} : t_0 < e \le t_0 + 3\,\mathrm{s}\}. \tag{2}$$

This event definition distinguishes anticipation from occupancy: an object already inside the region has a negative label when there is no new entry in the forecast interval. It also accommodates repeated entries following an exit.

ReactionGap stores stereo RGB observations, while all visual models evaluated here use only the monocular left-view stream; the right view is not a model input. Each observation comprises 15

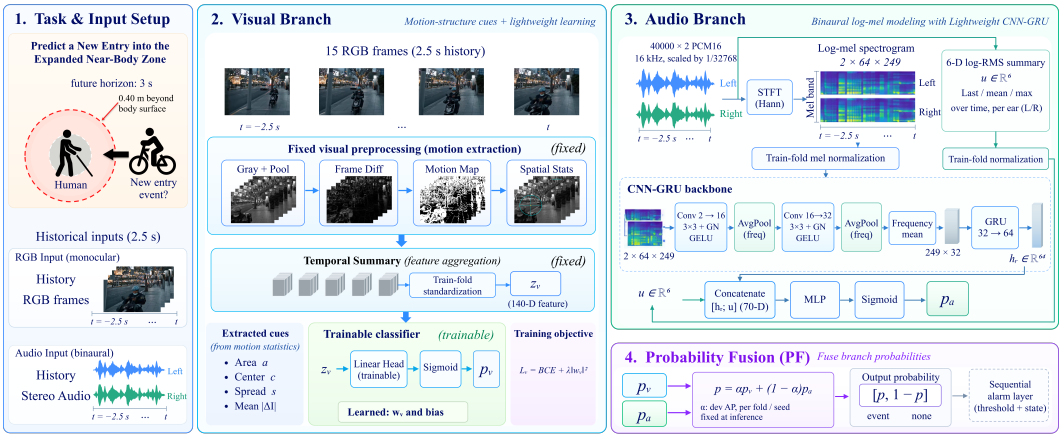


Figure 1: LOOM for near-body intrusion anticipation. Two independently trained branches predict future entry from historical monocular left-view RGB and simulated two-channel audio. Development data select a probability-fusion weight, fixed during inference, and a sequential warning policy. Geometry supplies labels and evaluation events. Predictor state is local to each historical window; alarm state persists across queries. Input thumbnails and the body-zone sketch are illustrative.

left-view RGB frames at 448×448 resolution and 40,000 samples per channel of simulated 16 kHz two-channel audio from the preceding 2.5 s. Frame selection enforces exposure and availability constraints, so every retained observation is available by t_0. Audio is converted from PCM16 by division by 32,768, with no segment-wise gain normalization. Timestamps and validity flags accompany the inputs. Geometry provides supervision and evaluation events; inference uses only historical RGB, PCM, timestamps, and validity. We choose 0.40 m to reflect blind participants' reported 38–41 cm mean shoulder clearances (Kolarik et al., 2017) while preserving entry and clear-pass episodes in a geometry-only check (Appendix A.5). The three-second prediction horizon is motivated by the TTC threshold in a wearable collision-warning trial (Pundlik et al., 2021); actual warning lead is evaluated separately.

3.2 Visual Motion and Expansion

The visual branch represents how image changes are distributed and evolve over the observation window. RGB values are first scaled to $[0, 1]$. Grayscale conversion with weights $(0.299, 0.587, 0.114)$ and average pooling by four produce 112×112 frames G_k. For each of the 14 adjacent pairs, we compute

$$D_k = |G_{k+1} - G_k|, \qquad M_k = \mathbf{1}[D_k > 0.04], \qquad W_k = D_k M_k. \tag{3}$$

The binary support M_k identifies changing pixels, while W_k weights their spatial distribution by change magnitude. Eight statistics describe each pair: active-pixel fraction, mean absolute difference, weighted horizontal and vertical centroids, their spatial standard deviations, and mean difference in central and lower image regions. The centroids and spreads retain coarse spatial information beyond the amount of image change.

To summarize expansion across pairs, we append differences of log support area and log spatial-spread product. Each contributes 13 values. Two additional values compare the first and last three-pair means of these quantities. Together with the 14×8 pair statistics, they form a 140-dimensional descriptor v. These summaries provide evidence about apparent expansion without requiring dense optical flow or object tracking. They can also respond to lateral movement and visibility changes; supervised learning determines their relationship to future entry in the training data.

A linear readout converts the standardized descriptor into an entry probability:

$$\tilde{v} = \frac{v - \mu_V}{\max(s_V, 10^{-6})}, \qquad p_V = \sigma(w_V^\top \tilde{v} + b_V). \tag{4}$$

The mean μ_V and standard deviation s_V are estimated on the training fold. The 141-parameter readout minimizes mean binary cross-entropy plus $0.01\|w_V\|_2^2$. This compact branch, denoted LOOM-V, preserves an explicit connection between the input motion statistics and the predicted event.

3.3 Learned Acoustic Sequence Model

The acoustic branch, denoted A25 for its 2.5 s audio history, models spectral evolution together with stereo level dynamics. A two-channel log-mel frontend produces a $2 \times 64 \times 249$ tensor using 20 ms periodic-Hann windows, a 10 ms hop, a 320-point FFT, and 64 mel filters from 50 to 8,000 Hz. The representation retains the temporal ordering of sounds throughout the same history used by the visual branch.

Two convolutional blocks map channels $2 \to 16 \to 32$. Each applies group normalization, GELU, and pooling along frequency. Frequency averaging then produces a 249-step sequence of 32-dimensional vectors. A unidirectional GRU summarizes it in a 64-dimensional final state. This sequence model can represent changing acoustic patterns beyond an overall sound-level summary.

Six log-RMS statistics retain the last-frame, temporal-mean, and temporal-maximum level in each channel. Concatenating them with the recurrent state gives a 70-dimensional representation, which a projection and prediction MLP map to p_A. Log-mel and level statistics are normalized using training-fold estimates. We train A25 independently with binary cross-entropy on the same entry labels as LOOM-V. Appendix B specifies the layers, normalization, and parameter accounting. Independent supervision produces two directly comparable event probabilities before fusion.

3.4 Probability Fusion and Warning Generation

The fused entry probability is a convex combination of the branch outputs:

$$p_{\mathrm{AV}} = \alpha p_V + (1-\alpha)p_A, \qquad \alpha \in \{0, 0.25, 0.5, 0.75, 1\}. \tag{5}$$

For each fold and seed, we select α by development AP, breaking ties toward the smaller visual weight. The selected weight remains fixed during inference. This separates the fitting of each sensory branch from the decision about their relative contribution. The candidate endpoints allow development data to select a single branch. Probability fusion (PF) uses the branch probabilities directly; calibrated logit fusion (CLF) is evaluated separately.

The warning layer processes queries at nominal 100 ms intervals and delivers each prediction after its measured computation time. Queries arriving while the processor is busy are skipped. A development set selects the threshold and one of three policies: thresholding with cooldown, rearming after an alarm, or confirmation by two consecutive processed observations. All policies enforce a 200 ms minimum inter-alarm interval. Selection maximizes event recall with at least one second of warning lead, subject to a development alarm-burden cap. Thus query ranking determines the fusion weight, while sequential warning utility determines the operating policy. Appendix B.3 specifies the state transitions and tie rules. Ground-truth event matching is confined to offline evaluation.

4 Experiments

4.1 Experimental Setup

ReactionGap comprises 150 eight-second simulated episodes across ten generation families, with stereo RGB, simulated 16 kHz two-channel audio, and geometric trajectories from stationary wearer-centered rigs. All visual models use only monocular left-view RGB. At 0.40 m, 75 episodes contain entries and 75 remain outside. Subsets A and B contain 63 and 87 episodes, respectively. We intentionally train and evaluate on their union to expose models to different rendering and acoustic conventions, with a shared task specification and entry-label definition (Section 3). Their generation and quality-control protocols differ (Appendix A). Two independent human evaluators inspected all 150 episodes, and all passed. Queries at $t_0 = 2.5, 2.6, \ldots, 5.0$ s yield 3,900 examples: 1,169 positive and 2,731 negative, associated with 75 entry events. Trajectories supply evaluation targets and event timing, not model inputs.

Three outer folds separate ten evaluation template groups, with every episode tested exactly once. All queries from an episode stay in one partition. Training/development/test episode counts are 90/15/45,

Table 1: Main comparison at 0.40 m, averaged over seeds 42/44/46. Higher AP, TM-AUC, and R_{1s} are better; lower burden B_{test} and latency are better. Bold marks the best value per column. Warning policies use development budget $B_{\text{dev}} \leq 0.5$ and median-cost replay. Time is the P1-path median in ms, excluding disk I/O. External methods are task adaptations. AGA-AV-PF and TRN-AV-PF fuse the corresponding EgoVLP predictor with A25.

System	Pooled AP	Fold AP	TM-AUC	R_{1s}	B_{test}	ms
LOOM (ours)	**0.5438**	0.5394	0.7257	0.4178	0.7289	8.8
LOOM-V	0.5234	0.5547	**0.7793**	0.4044	0.7956	**2.1**
A25	0.4776	0.4997	0.6653	0.3378	0.5911	6.9
LOOM-AV-CLF	0.4743	0.5305	0.7266	**0.4400**	0.8200	8.6
AGA-EgoVLP	0.3998	0.5252	0.6529	0.0889	0.2978	72.4
AGA-AV-PF	0.3892	0.5188	0.6537	0.2222	0.5822	71.4
MMTM-AV	0.4541	0.5659	0.6994	0.0889	**0.2267**	56.6
LMF-AV	0.3802	0.4890	0.5787	0.2178	1.1711	55.7
TFN-AV	0.4272	0.5488	0.6440	0.2311	1.4756	54.7
TRN-EgoVLP	0.4474	0.5574	0.6477	0.1111	0.3644	54.1
TRN-AV-PF	0.4203	**0.5756**	0.6912	0.2889	0.6511	54.6
TIME	0.4388	0.4424	0.5000	–	–	–

75/15/60, and 90/15/45; each development set contains one template group. We train task-specific predictors with seeds 42, 44, and 46 using AdamW, learning rate 3×10^{-4}, weight decay 0.01, batch size 16, gradient clipping at 1, and 20 scheduled epochs: 2,940, 2,440, and 2,940 updates across folds. Computation uses FP32 with TF32 disabled. Models share partitions and batch schedules; selection retains the earliest checkpoint attaining the best development AP.

Table 1 compares unimodal branches, calibrated logit fusion (LOOM-AV-CLF), and two families of task adaptations. Temporal models AGA (Tai et al., 2026) and TRN (Zhou et al., 2018) operate on frozen EgoVLP sequences (Lin et al., 2022); AGA-AV-PF and TRN-AV-PF combine their probabilities with A25 using our development-selected fusion rule. Representation-level models combine frozen visual and acoustic features through MMTM channel transfer (Vaezi Joze et al., 2020), rank-four LMF (Liu et al., 2018), or full TFN outer products (Zadeh et al., 2017). TIME uses only the query clock. Table 2 compares audio-level statistics and early-fusion readouts. All share the query, label, partition, and warning protocol; Appendix B specifies architecture, initialization, and loss differences.

Prediction metrics are pooled out-of-fold AP, the unweighted mean of fold APs, and time-matched AUC (TM-AUC). The latter computes AUC within each fold and query-time slice containing both labels, then weights by slice size, controlling for query time and fold-specific score scales. Metrics are computed per seed and averaged; inference uses one model instance. Pairwise 95% intervals use 10,000 template-group bootstrap draws with seeds averaged inside each draw. Resampling ten groups preserves within-template dependence; intervals are not adjusted for multiple comparisons.

4.2 Prediction Performance

LOOM attains the highest pooled AP, 0.5438. Its gain over audio is 0.0662, with paired interval $[0.0098, 0.1916]$; the gain over vision is 0.0204, with interval $[-0.0623, 0.0662]$. Thus the evidence is stronger relative to audio. LOOM also has lower mean NLL and Brier score than vision (0.5223/0.1754 versus 0.5300/0.1796), complementing ranking with probability quality. TIME reaches AP 0.4388 but TM-AUC 0.5000, highlighting the predictive structure of the query schedule.

The strongest system depends on the evaluation objective. TRN-AV-PF has the highest fold AP, 0.5756, followed by MMTM-AV at 0.5659; LOOM reaches 0.5394. TRN-AV-PF's fold APs are 0.4831, 0.3647, and 0.8790, compared with LOOM's 0.5157, 0.5233, and 0.5792. Its advantage is concentrated in the slider/wheeled fold. LOOM-V leads TM-AUC at 0.7793 versus LOOM's 0.7257, and exceeds LOOM in all nine fold–seed comparisons. Pooled ranking, within-fold ranking, and time-matched discrimination therefore capture distinct aspects of transfer. Appendix C gives complete fold and seed results.

Among learned representation-level fusion models, MMTM-AV reaches pooled AP 0.4541, TFN-AV 0.4272, and LMF-AV 0.3802. LOOM's gain over LMF is 0.1636 with interval $[0.0017, 0.2748]$. Its

Table 2: Branch and fusion-design ablations under the same settings as Table 1. LOOM is repeated for reference; bold marks the best value per column within this table. ALS-linear uses audio-level statistics; LOOM-EF concatenates these with visual statistics in a linear readout; LOOM-EF-MLP replaces that readout with an MLP; LOOM-EF-cal applies development-set calibration.

System	Pooled AP	Fold AP	TM-AUC	R_{1s}	B_{test}	ms
LOOM (ours)	**0.5438**	0.5394	0.7257	0.4178	0.7289	8.8
LOOM-EF	0.4992	0.5363	0.7607	0.5156	0.9711	6.9
ALS-linear	0.3173	0.3399	0.4707	**0.6844**	4.1044	**4.9**
LOOM-EF-MLP	0.5073	**0.5477**	**0.7655**	0.2400	**0.4156**	6.9
LOOM-EF-cal	0.5244	0.5363	0.7607	0.5333	0.8978	6.8

intervals against MMTM, TFN, both AGA variants, and both TRN variants include zero. For example, the LOOM–TFN difference is 0.1166 with interval $[-0.0494, 0.2448]$, and LOOM–TRN-AV-PF is 0.1235 with interval $[-0.0646, 0.2435]$. The point estimates favor LOOM in pooled AP, with substantial uncertainty across the ten template groups.

The subsets' different positive-query rates (40.48% and 22.37%) make subset membership informative about labels. We therefore condition AUC jointly on fold, query time, and subset. LOOM-V (0.7526), LOOM (0.6921), and A25 (0.6307) exceed metadata-only diagnostics (0.5000), covering 93.28% of queries. Thus discrimination extends beyond these metadata, although generation-related cues may persist within subsets. Fusion's pooled-AP advantage may partly reflect between-subset differences. These diagnostics do not establish causality or resolve missing template–subset combinations; Appendix C provides definitions, coverage, and complete results.

4.3 Ablation and Fusion Analysis

The learned acoustic sequence model outperforms the audio-level-statistics baseline. A25 reaches AP 0.4776 and TM-AUC 0.6653, versus 0.3173 and 0.4707 for ALS-linear. This comparison evaluates the complete learned audio branch rather than isolating recurrence. LOOM-V is the strongest unimodal predictor by pooled AP and TM-AUC, motivating a separate evaluation of how audio should be combined with vision.

Probability fusion gives higher pooled AP than the early-fusion variants in Table 2: 0.5438 for LOOM versus 0.4992 for the linear readout, 0.5073 for its MLP replacement, and 0.5244 after calibration. Nonlinearity alone does not recover the late-fusion result. Calibration changes pooled AP while preserving rounded fold AP and TM-AUC, illustrating the effect of cross-fold score scales. LOOM-AV-CLF reaches 0.4743, below LOOM in each seed; calibrating branch logits does not necessarily improve their combined ranking.

The probability rule is also representation-dependent. Adding A25 lowers pooled AP from 0.3998 to 0.3892 for AGA and from 0.4474 to 0.4203 for TRN, although TRN's fold AP rises from 0.5574 to 0.5756 and TM-AUC from 0.6477 to 0.6912. For LOOM, development selection retains both modalities in six of nine instances and selects audio alone in three. Its visual weights are $0, 0.25, 0$ in fold 0; $0.5, 0, 0.25$ in fold 1; and 0.75 for all seeds in fold 2. These selections reflect development-template performance and need not transfer uniformly to held-out templates.

4.4 Warning Performance and Efficiency

An event contributes to R_{1s} when its first valid warning arrives at least one second before entry. All 75 events enter the denominator. Alarm burden counts false, duplicate, and late alarms per episode:

$$B = \frac{N_{\text{false}} + N_{\text{duplicate}} + N_{\text{late}}}{150}. \tag{6}$$

Policies are selected under $B_{\text{dev}} \leq 0.5$; test burden is measured, not constrained. Matching and computation-delay replay are identical across systems.

LOOM achieves recall 0.4178 at burden 0.7289, versus 0.4044/0.7956 for vision and 0.3378/0.5911 for audio. Its recall gain over vision has interval $[-0.2267, 0.2835]$. LMF and TFN achieve

0.2178/1.1711 and 0.2311/1.4756: LOOM's mean operating point provides greater advance coverage with fewer non-useful alarms. Their recall-difference intervals still include zero. TRN-AV-PF instead has lower recall and burden, 0.2889/0.6511. ALS-linear reaches the highest recall, 0.6844, but incurs burden 4.1044; LOOM-EF-MLP reduces burden to 0.4156 at recall 0.2400. These are development-selected operating points, not comparisons at matched test burden.

LOOM's mean P1-path median is 8.8 ms and mean P95 is 11.5 ms on the RTX 4090. Adapted temporal and fusion models require 54.1–72.4 ms at the median, giving LOOM a 6.1–8.2$\times$ speedup; LOOM-V and A25 require 2.1 and 6.9 ms. The comparison includes complete input processing, frozen encoders where used, prediction, and alarm updates. Disk I/O is separate and zero-weight branches are skipped. Timing is standardized to P1, while replay uses the development-selected policy. The query schedule allows a pre-entry query for 74 events and one second of lead for 57, giving a 76% opportunity ceiling. Appendix C reports P95, probability quality, and additional warning comparisons.

These experiments assess transfer across simulated motion templates with a fixed camera per episode. The two generation subsets differ in rendering and acoustic scaling, and each fold uses one development template. Moving-camera recordings and wearer studies would further assess sensor motion, environmental variability, and warning usability.

5 Conclusion

We presented LOOM for egocentric audio-visual near-body intrusion anticipation in visual impairment assistance. Compact visual motion statistics, a learned acoustic sequence model, probability fusion, and stateful warnings yield an 8.8 ms median computation time, a 6.1–8.2$\times$ speedup over the evaluated temporal and fusion adaptations. We constructed ReactionGap with synchronized audio-visual observations, geometric trajectories, and entry annotations to support training and evaluation across motion templates. On this dataset, LOOM improves pooled AP over its individual branches, while vision retains stronger time-matched discrimination. Sequential replay shows how prediction scores translate into advance coverage and notification burden. Together, the framework and dataset support research on efficient multimodal anticipation and timely assistive warnings.

Reproducibility Statement

Appendix A describes the ReactionGap generation pipelines, task construction, episode checks, and asset provenance. Appendix B specifies input processing, partitions, networks, training, fusion selection, and warning accounting. Appendix C provides seed-level results, fusion weights, probability quality, latency, paired comparisons, and generation-subset diagnostics. The accompanying manuscript sources include machine-readable result and data inventories supporting the tables and figures. Code and data will be made publicly available upon acceptance.

AI Use Statement

Generative AI tools assisted with task formulation, experiment planning, implementation and debugging of training and evaluation software, result interpretation, evidence organization, literature search, manuscript drafting and editing, figure construction, and numerical and document checks. AI assistance also supported preparation of the dataset inventory and description. Episodes were produced by the simulation and audio pipelines described in Appendix A; Subset B uses recorded audio sources. The two independent human evaluations of the episodes are separate from these AI-assisted activities.

References

Abhishek Badki, Orazio Gallo, Jan Kautz, and Pradeep Sen. Binary TTC: A temporal geofence for autonomous navigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021. URL https://arxiv.org/abs/2101.04777.

Bing Cao, Yinan Xia, Yi Ding, Changqing Zhang, and Qinghua Hu. Predictive dynamic fusion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 5608–5628, 2024.

Daniel de Godoy, Bashima Islam, Stephen Xia, Md Tamzeed Islam, Rishikanth Chandrasekaran, Yen-Chun Chen, Shahriar Nirjon, Peter R. Kinget, and Xiaofan Jiang. PAWS: A wearable acoustic system for pedestrian safety. In *IEEE/ACM International Conference on Internet-of-Things Design and Implementation*, pp. 237–248, 2018. doi: 10.1109/IOTDI.2018.00031.

Yuan Gong, Andrew Rouditchenko, Alexander H. Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James Glass. Contrastive audio-visual masked autoencoder. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=QPtMRyk5rb.

Yasaman Haghighi, Celine Demonsant, Panagiotis Chalimourdas, Maryam Tavasoli Naeini, Jhon Kevin Munoz, Bladimir Bacca, Silvan Suter, Matthieu Gani, and Alexandre Alahi. HEADS-UP: Head-mounted egocentric dataset for trajectory prediction in blind assistance systems. *arXiv preprint arXiv:2409.20324*, 2024. URL https://arxiv.org/abs/2409.20324.

Andrew J. Kolarik, Amy C. Scarfe, Brian C. J. Moore, and Shahina Pardhan. Blindness enhances auditory obstacle circumvention: Assessing echolocation, sensory substitution, and visual-based navigation. *PLOS ONE*, 12(4):e0175750, 2017. doi: 10.1371/journal.pone.0175750.

Haoxin Li, Yuan Liu, Hanwang Zhang, and Boyang Li. Mitigating and evaluating static bias of action representations in the background and the foreground. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 19911–19923, 2023.

Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z. Xu, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/31fb284a0aaaad837d2930a610cd5e50-Abstract-Conference.html.

Renyuan Liu and Qinbing Fu. Attention-driven LPLC2 neural ensemble model for multi-target looming detection and localization. In *International Joint Conference on Neural Networks*, 2025. URL https://arxiv.org/abs/2504.04477.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2247–2256, 2018. doi: 10.18653/v1/P18-1209. URL https://aclanthology.org/P18-1209/.

Aashi Manglik, Xinshuo Weng, Eshed Ohn-Bar, and Kris M. Kitani. Forecasting time-to-collision from monocular video: Feasibility, dataset, and challenges. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 8081–8088, 2019. doi: 10.1109/IROS40897.2019.8967730.

John G. Neuhoff. Perceptual bias for rising tones. *Nature*, 395(6698):123–124, 1998. doi: 10.1038/25862.

Daniel Oberfeld, Marlene Wessels, and David Büttner. Overestimated time-to-collision for quiet vehicles: Evidence from a study using a novel audiovisual virtual-reality system for traffic scenarios. *Accident Analysis & Prevention*, 175:106778, 2022. doi: 10.1016/j.aap.2022.106778.

Patrik Patera, Yie-Tarng Chen, and Wen-Hsien Fang. Lightweight spatio-temporal modeling via temporally shifted distillation for real-time accident anticipation. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8zzfTSVds2.

Claire E. Pittet, Eduardo Villar Ortega, Maël Fabien, Mark T. Wallace, and Micah M. Murray. Efficacy of electronic travel aids for the blind and visually impaired during wayfinding. *Scientific Reports*, 16:6423, 2026. doi: 10.1038/s41598-026-37578-9.

Shrinivas Pundlik, Vilte Baliutaviciute, Mojtaba Moharrer, Alex R. Bowers, and Gang Luo. Home-use evaluation of a wearable collision warning device for individuals with severe vision impairments: A randomized clinical trial. *JAMA Ophthalmology*, 139(9):998–1005, 2021. doi: 10.1001/jamaophthalmol.2021.2624.

Chiara F. Sambo and Gian Domenico Iannetti. Better safe than sorry? the safety margin surrounding the body is increased by anxiety. *Journal of Neuroscience*, 33(35):14225–14230, 2013. doi: 10.1523/JNEUROSCI.0706-13.2013.

Hadeel Surougi, Cong Zhao, and Julie A. McCann. ARAware: Assisting visually impaired people with real-time critical moving object identification. *Sensors*, 24(13):4282, 2024. doi: 10.3390/s24134282.

Tsung-Ming Tai, Sofia Casarin, Andrea Pilzer, Werner Nutt, and Oswald Lanz. Action-guided attention for video action anticipation. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2603.01743.

Hamid Reza Vaezi Joze, Amirreza Shaban, Michael L. Iuzzolino, and Kazuhito Koishida. MMTM: Multimodal transfer module for CNN fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13289–13299, 2020. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Joze_MMTM_Multimodal_Transfer_Module_for_CNN_Fusion_CVPR_2020_paper.html.

Maya Varma, Jean-Benoit Delbrouck, Sophie Ostmeier, Akshay Chaudhari, and Curtis Langlotz. TRoVe: Discovering error-inducing static feature biases in temporal vision-language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0335.

Linge Wang, Yingying Chen, Bingke Zhu, Lu Zhou, and Jinqiao Wang. Semantic noise reduction via teacher-guided dual-path audio-visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32005–32014, 2026.

Yihan Wang and Jia Deng. WAFT: Warping-alone field transforms for optical flow. In *International Conference on Learning Representations*, 2026. URL https://arxiv.org/abs/2506.21526.

Junbin Xiao, Nanxin Huang, Hao Qiu, Zhulin Tao, Xun Yang, Richang Hong, Meng Wang, and Angela Yao. EgoBlind: Towards egocentric visual assistance for the blind. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0886. Datasets and Benchmarks Track.

Zhiqiang Yuan, Ting Zhang, Yeshuang Zhu, Jiapei Zhang, Ying Deng, Zexi Jia, Peixiang Luo, Xiaoyue Duan, Jie Zhou, and Jinchao Zhang. WalkVLM: Aid visually impaired people walking by vision language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9845–9854, 2025.

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103–1114, 2017. doi: 10.18653/v1/D17-1115. URL https://aclanthology.org/D17-1115/.

Qingyang Zhang, Haitao Wu, Changqing Zhang, Qinghua Hu, Huazhu Fu, Joey Tianyi Zhou, and Xi Peng. Provable dynamic fusion for low-quality multimodal data. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 41753–41769, 2023.

Junyu Zhao, Shengkai Xi, Yan Li, Aike Guo, and Zhihua Wu. A fly inspired solution to looming detection for collision avoidance. *iScience*, 26(4):106337, 2023. doi: 10.1016/j.isci.2023.106337.

Tianhao Zhao, Yiyang Zou, Zihao Mao, Peilun Xiao, Yulin Huang, Hongda Yang, Yuxuan Li, Tracy Li, Guobin Wu, and Yutian Lin. Accident anticipation via temporal occurrence prediction. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0307.

Zeyun Zhong, David Schneider, Michael Voit, Rainer Stiefelhagen, and Jürgen Beyerer. Anticipative feature fusion transformer for multi-modal action anticipation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 6068–6077, 2023.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 803–818, 2018. URL https://openaccess.thecvf.com/content_ECCV_2018/html/Bolei_Zhou_Temporal_Relational_Reasoning_ECCV_2018_paper.html.

Yiyang Zou, Tianhao Zhao, Peilun Xiao, Hongyu Jin, Longyu Qi, Yuxuan Li, Liyin Liang, Yifeng Qian, Chunbo Lai, Yutian Lin, Zhihui Li, and Yu Wu. RiskProp: Collision-anchored self-supervised risk propagation for early accident anticipation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2768–2777, 2026.

A REACTIONGAP DATASET AND TASK CONSTRUCTION

A.1 DATASET COMPOSITION AND ENTRY LABELS

ReactionGap comprises Subset A (63 episodes) and Subset B (87 episodes), constructed with two image and sound generation pipelines. Their disjoint episode IDs cover ten balanced generation families (Table 3); Subset A contains no F02 or F08 examples. Generator identifiers are retained in the accompanying reproducibility documentation.

Table 3: Generation families in ReactionGap. Counts are episodes, not query labels or entry events.

ID	Generation family	Subset A	Subset B	Total
F01	Large-ball free flight	2	13	15
F02	Large-ball bounce and roll	0	15	15
F03	Hinged-door sweep	9	6	15
F04	Sliding door or panel	8	7	15
F05	Four-wheel push cart	8	7	15
F06	Two-wheel or rolling load	8	7	15
F07	Large rigid-body tip	1	14	15
F08	Large rigid-body slide and drop	0	15	15
F09	Rail vehicle, lateral approach	14	1	15
F10	Road vehicle, slow lateral approach	13	2	15
	Total	63	87	150

Generation families and evaluation template groups are related but not identical. F01 is divided into `throw` (12) and `push_off` (3), whereas F05 and F06 share `wheeled` (30). The other families map to `bounce_roll`, `hinge`, `slider`, `tip`, `slide_drop`, `rail`, and `road`. These ten unequal-sized groups define the folds and bootstrap.

Inputs and labels are constructed from full 8 s media and trajectories using 2.5 s of history, a 3 s forecast, and a 0.40 m body margin (Section 3). Supervision uses geometric entries rather than cropped preview windows, physical-event categories, or acoustic-label metadata. ReactionGap provides 3,900 queries: 1,169 positive and 2,731 negative, associated with 75 geometric entries. Subset A contributes 1,638 queries, 663 positives, and 36 entries; Subset B contributes 2,262 queries, 506 positives, and 39 entries. The body is represented by a capsule of radius 0.25 m with axis endpoints $(0, 0.25, 0)$ and $(0, 1.45, 0)$ m. Boundary crossings are computed from 960 Hz geometric trajectories with interpolation. Full eight-second support covers the forecast interval of every query.

A.2 PHYSICAL CONSTRUCTION AND IMAGE FORMATION

A task specifies an object family, a mechanism, an intended margin-entry or clear-pass outcome, and an anchor lead band. The generator adjusts recorded initial timing to place the anchor. Kinematic actors initiate pushes, releases, or tips; rigid bodies, hinges, constraints, and motors govern subsequent motion. The shared solver is Blender RigidBodyWorld (Bullet),[1] with 960 Hz trajectories, 60 solver iterations, split impulse, and gravity of magnitude $9.81\ \mathrm{m/s^2}$. Subset A records Blender 5.2.1 LTS, build `9e2066aef7ef`; Subset B records Blender 4.5.14 LTS, build `62c1db4208e8`.

Subset A was additionally solved at 1920 Hz. Its acceptance rule required anchor and relevant contact-time differences of at most 1 ms and minimum body-gap differences of at most 2 mm. Recorded maximum differences are 0.781 ms, 0.733 ms, and 0.474 mm, respectively. Subset B uses a separate visual/physics quality protocol without certification of these step-halving tolerances. Contacts are extracted from solved surface distances; Blender's unexposed contact-impulse field is not treated as measured ground truth. The wearer capsule is a geometric reference, not a collider, and does not emit a body-collision sound.

Both subsets use one stationary stereo camera rig per episode, centered at $(0, 1.60, 0)$ m, with a 65 mm baseline and 70° horizontal field of view. Yaw and pitch are selected per scene to retain the

[1] Software references: `https://www.blender.org/features/simulation/` and `https://github.com/bulletphysics/bullet3`. Version strings are recorded in episode metadata.

target and its cause. Each 8 s clip stores 240 frames per view at 30 Hz. Both left and right RGB streams are retained; the experiments use only the left view. A custom ModernGL 5.12 rasterizer produces the images; Subset B uses a fixed opaque ambient/diffuse profile.

The Subset A RGB pipeline stores lossless 896×896 FFV1 masters and derives 448×448 images by decoding sRGB, averaging 2×2 blocks in linear light, and re-encoding to 8-bit sRGB. Its depth reduction keeps the nearest subpixel only when all four instance IDs agree; zero depth denotes invalidity. Subset B instead renders directly at 448×448, storing lossless FFV1 and uint8 pixel arrays, with IDs and metric depth from the same evaluated surfaces. Both use simplified community meshes and primitive backgrounds. Twelve key objects in Subset A use size-adjusted class substitutes; Subset B also permits proxy colliders and declared substitutions. These choices produce controlled scenes with simplified appearance and object geometry.

A.3 Clock Alignment and Acoustic Construction

Video, trajectories, and audio share an ideal simulation timeline, not a measured hardware clock. A nominal exposure starts at $i/30$ s and lasts 1 ms. The preprocessing procedure handles both integer-microsecond and exact fractional timestamps before selecting legal observations. For LOOM, exposure start must lie within the preceding 2.5 s, and exposure completion and availability must not exceed the query time. Fifteen rounded, evenly spaced indices are selected from that legal-frame list. Audio contains 128,000 samples per channel over the full clip; each model query takes the preceding 40,000.

Subset A: library impacts and procedural tracks. Nineteen episodes have a Kenney Impact Sounds library event, 27 have procedural vehicle drive/rolling audio, and 17 have a silent target track. All add a procedural ambience bed with nominal RMS 0.010 Pa. A contact sample is scaled by a shared library gain, material factor, and contact speed. A point source at $x(t_e)$ reaches microphone m after geometric delay and inverse-distance scaling:

$$t_a = t_e + \frac{\|x(t_e) - m\|}{343}, \qquad p_m(t_a) = \frac{p_{\mathrm{src}}(t_e)}{\max(\|x(t_e) - m\|, 0.30)}. \tag{7}$$

The Subset A gain convention uses 20 Pa full scale,

$$Q(p) = \mathrm{rint}(\mathrm{clip}(32767p/20, -32767, 32767)). \tag{8}$$

The same gain applies to the mixture, target, background, and leave-one-target-out tracks, without per-clip/per-ear normalization or dither. These pressure and ambience labels are simulation conventions, not measured SPL calibration, room acoustics, or HRTF measurements. Acoustic-validity flags in Subset A are false and are not used as LOOM targets.

Subset B: recorded audio sources and fixed digital gain. The Subset B pipeline uses licensed recordings, including recorded backgrounds, rather than procedural or generative source audio. It crops and causally resamples donors to 16 kHz, propagates them using emission-time geometry at 343 m/s, and mixes target, background, causal non-target, and nuisance roles. Short continuous textures use cross-faded recorded segments at their original speed. The audio collection contains 32 uniquely identified recordings with file hashes. The common digital gain is 0.125, followed by deterministic round-to-nearest-even PCM16 quantization without dither; overflow is rejected. Per-source fixed gains can differ, but there is no per-episode or per-role normalization. The receiver is an identity software reference, not a fitted physical microphone. Some donors are public MP3 previews, and donor room responses remain in the sound. Material-level Foley approximations are documented and are not claimed to be recordings of the exact simulated object.

Neither subset establishes measured acoustic calibration, and their gain conventions are not physically interchangeable. The experiment preserves the generated PCM and applies only PCM16/32,768 before its training-fold feature normalization. It does not equalize the two subsets to a common SPL.

A.4 Quality Checks and Asset Provenance

Automated checks passed for all 63 episodes in Subset A and all 87 episodes in Subset B under their respective quality protocols. Checks cover disjoint episode IDs, decoded-media integrity, file

hashes, causal audio propagation, and full/leave-one-out mixture closure. All 150 geometric-label file hashes and the episode index were verified against the experiment manifest. Two independent human evaluators inspected all 150 episodes, and all passed. These inspections complement computational verification of the 0.40 m entry labels. The numerical step-halving criteria above apply to Subset A; Subset B uses its separately documented criteria.

Kenney provides the CC0 impact library and multiple low-poly asset packs.[2] Attributed CC BY 3.0 meshes include the basketball by Zer0_Time, soccer ball by Poly by Google, and volleyball by Zsky.[3] The Subset B audio attribution file identifies Freesound and OpenGameArt recordings with their authors, source URLs, hashes, and recorded CC0/CC BY 4.0 licenses. The manuscript source package includes an anonymous data inventory, asset attribution file, and figure provenance records. Individual licenses and attribution requirements are retained in these records.

A.5 Spatial Margin and Observation Support

The spatial margin specifies a body-surface clearance. Published studies provide context for its scale while measuring different quantities. Kolarik et al. (2017) report mean shoulder clearance of 41 cm under auditory guidance and 38 cm under tactile sensory substitution in eight blind non-echolocators. Sambo & Iannetti (2013) identify a defensive-response boundary between 20 and 40 cm from the face in 15 healthy participants. ARAware uses user-centered collision-zone radii of 0.75, 1.25, and 3 m for pedestrians, bicycles, and cars (Surougi et al., 2024); the BuzzClip description in Pittet et al. (2026) gives a 1–3 m sensing range. These behavioral clearances, center radii, and detection ranges motivate near-body sensing at different scales, rather than specifying one universal safety distance. The three-second TTC threshold in Pundlik et al. (2021) additionally motivates considering warning time alongside spatial proximity.

Before training at 0.40 m, a geometry-only scan evaluated eleven candidate margins on the 63 episodes in Subset A: 0.20, 0.30, 0.40, 0.50, 0.60, 0.75, 0.80, 1.00, 1.25, 1.50, and 2.00 m. The analysis used 1,638 query labels and no prediction scores. At 0.40 m, all 36 entry events and 27 never-entering episodes remained. At 0.50 m, Subset A's only push_off episode entered before the first query. At 1 m, only eight never-entering episodes remained, and the road and rail development groups lost their never-entering episodes. The chosen 0.40 m boundary retains these negative controls while maintaining observation support in Subset A.

Table 4: Geometry-only margin diagnostic on the 63 episodes in Subset A. "Early" counts entries at or before the first query; "≥ 1 s" counts events with an available query offering one second of lead at zero computation cost.

Margin (m)	Events	Never-entering	Positive / negative	Early	≥ 1 s
0.20	36	27	726 / 912	0	32
0.40	36	27	663 / 975	0	32
0.50	36	27	633 / 1005	1	31
1.00	55	8	800 / 838	2	37

The complete 150-episode dataset contains both query classes in every training, development, and test partition at 0.40 m. Relative to 0.20 m, 173 query labels change: 171 positive-to-negative and two negative-to-positive. The former are already inside the expanded region with no subsequent entry in the forecast interval. In total, 649 queries are occupied but negative. Two repeated-entry cases merge into longer occupancy intervals, reducing the event count from 77 to 75. The same 75 episodes contain events at both margins, while the other 75 remain outside. The margin changes timing and interval structure without adding event-bearing episodes. Of the 75 events, 74 have a pre-entry query and 57 permit at least one second of lead. These full-dataset counts determine the warning denominator and opportunity ceiling used in Section 4.

[2] `https://kenney.nl/assets/impact-sounds`; individual packs and donor pages are listed in the accompanying asset attribution file.

[3] `https://poly.pizza/m/wNXXIhP0ip`, `https://poly.pizza/m/fnDmeCySRhV`, and `https://poly.pizza/m/LjWmmH9CG7`.

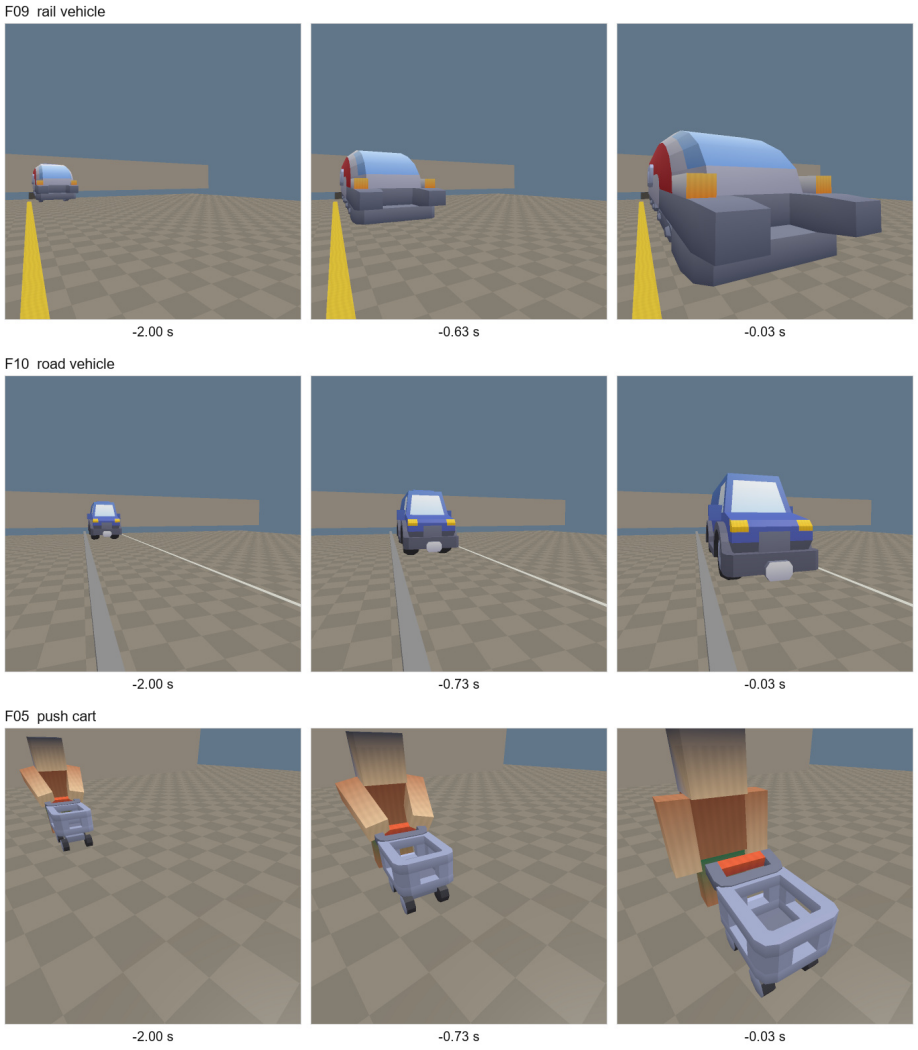


Figure 2: Left-view RGB examples from Subset A: rail vehicle `RG150-125`, road vehicle `RG150-139`, and cart `RG150-065`. Each row shows three frames from a 2 s window preceding the scenario anchor, with timestamps relative to that anchor rather than a LOOM query cutoff. The examples have the largest target-area growth within these three families; area increases and median depth decreases across each row. They illustrate approach and rendering quality rather than random sampling or model accuracy. Episode IDs and selection details accompany the figure provenance records.

B Implementation and Evaluation Details

B.1 Partitions and Optimization

Table 5 gives query counts for the template-disjoint folds. Fold 0 tests hinge, road, and bounce_roll; fold 1 tests push_off, throw, tip, rail, and slide_drop; fold 2 tests slider and wheeled. Each training partition contains the remaining groups after excluding development and test. The ten template groups are also the resampling units for confidence intervals.

Table 5: Fixed partitions. Entries are query counts with positive counts in parentheses. Every episode contributes 26 queries to a single partition in each fold.

Fold	Development group	Train	Development	Test
0	rail	2340 (634)	390 (156)	1170 (379)
1	road	1950 (626)	390 (145)	1560 (398)
2	road	2340 (632)	390 (145)	1170 (392)

Normalization uses training data only. For the LOOM readouts and A25, output initialization also uses the training labels. Standard deviations are floored at 10^{-6}, and the output bias is initialized from the smoothed training-label prior, with zero output weights. The visual model adds coefficient 0.01 times the sum of squared linear weights to mean BCE, in addition to AdamW weight decay. A25 uses mean BCE. Evaluation occurs at updates 0, 8, 16, 32, 64, 128, each scheduled epoch end, and the final update. Ties retain the earliest checkpoint achieving the best development AP. Optimization schedules and checkpoint selection are shared across comparisons. All external adaptations use two-class cross-entropy without the output-prior override. LMF factors use Xavier-normal initialization; other trainable modules use their implementation defaults.

B.2 Descriptors, Networks, and Fusion Comparisons

The visual descriptor concatenates 112 pair statistics, 13 log-area differences, 13 log-spread-product differences, and two endpoint summaries. Coordinates are normalized to $[0, 1]$. Central and lower regions are pooled-image indices $[28 : 84, 28 : 84]$ and $[56 : 112, :]$, respectively. Log-area and log-spread offsets are 10^{-4} and 10^{-6}. Differences are computed between adjacent selected frames without division by their temporal separation. The visual frontend has no learned encoder or camera-motion compensation.

The acoustic frontend uses natural logarithms of power, a periodic Hann window, no centered padding, and the HTK mel transform without area normalization. The first convolution has $2 \to 16$ channels with four GroupNorm groups; the second has $16 \to 32$ channels with eight groups. Both use 3×3 kernels, unit stride, padding one, GELU, and average pooling $(2, 1)$. Frequency averaging supplies a single-layer GRU with input size 32 and hidden size 64. The 70-dimensional recurrent/level representation is projected to 128 dimensions with GELU and dropout 0.1. For compatibility with the evaluated implementation, the prediction MLP receives this representation concatenated with a fixed 128-dimensional zero slot, applies a $256 \to 128$ layer, GELU, dropout 0.1, and a $128 \to 1$ output. A25 has 65,969 registered trainable parameters, including weights connected to that inactive slot; the slot carries no visual input.

The audio-level descriptor used by ALS-linear and the early-fusion comparisons has 57 dimensions. It aggregates the newest 240 log-mel frames into ten temporal segments, summarizing broadband log power in each channel and their difference, and appending first differences. LOOM-EF concatenates these 57 values with the 140 visual values and uses a linear readout. LOOM-EF-MLP uses a 32-dimensional hidden layer with GELU and dropout 0.1. Both use training-fold standardization and the same linear-weight regularization rule as LOOM-V.

For LOOM-EF-cal, a one-dimensional logistic fit maps development logits z to $az + b$, with inverse regularization strength $C = 1$. A nonpositive fitted slope is replaced by zero and an intercept corresponding to the development-label prevalence. CLF fits this monotone transformation to each branch independently, then computes

$$p_{\mathrm{CLF}} = \sigma[\alpha(a_V z_V + b_V) + (1 - \alpha)(a_A z_A + b_A)] . \tag{9}$$

The weight uses the same five-point grid and development-AP selection rule as LOOM. Calibration, fusion, and warning parameters remain fixed on the test partition. LOOM applies no Platt transformation.

AGA-EgoVLP and AGA-AV-PF. The AGA adaptation uses the official implementation at commit `cb46630694443f8224d1948957e2c2b1a8945b13`, with a frozen EgoVLP global patch-mean sequence of shape 15×256. The model has hidden dimension 2,048, two output classes, memory order 15, dropout and attention dropout 0.6, and moving-average recurrent queries with ratio 0.8. The input gate is enabled and recurrent hidden-state feedback is disabled. Memory resets at every query; no ground-truth action labels are supplied to the model. Training applies cross-entropy only to the final sequence output. Input channels are standardized with training-fold statistics. AGA-AV-PF applies the same development-selected probability rule as Eq. 5 to AGA-EgoVLP and the independently trained A25 branch. EgoVLP may mix frames within the permitted historical window; the prediction uses no observation after the query cutoff.

MMTM-AV. We use the official MMTM channel-recalibration module at commit `1c81cfefad5532cfb39193b8af3840ac3346e897`, retaining its squeeze, excitation, and sigmoid operations. The visual input is the same frozen 15×256 EgoVLP sequence. For audio, the same-fold, same-seed A25 CNN–GRU provides all 249 recurrent states. Each state is concatenated with the six normalized level statistics and passed through the frozen $70 \to 128$ projection, producing 249×128 features. A25's classifier is not used in this prediction path. Both sequences are standardized from the training partition, converted to channel-first layout, and passed through one MMTM stage with ratio 4. The recalibrated streams are temporally averaged, concatenated, and classified by a $384 \to 128 \to 2$ head with ReLU and dropout 0.1. All encoders remain in evaluation mode. This evaluates a single-stage adaptation on the entry task, whereas the original method supports transfer modules at multiple network depths.

LMF-AV and TFN-AV. Both use the same frozen visual and audio sequences as MMTM. Training-fold channel standardization precedes temporal averaging. Each modality then passes through batch normalization, dropout 0.1, and three Linear–ReLU layers of width 128. LMF uses rank-four multiplicative fusion with modality-specific factors and two output logits. Factors are registered trainable parameters and initialized with Xavier-normal draws. TFN appends a constant to each modality vector and computes the full 129×129 outer product, including unimodal and joint terms. Its flattened 16,641-dimensional tensor passes through dropout 0.1 and a $16{,}641 \to 128 \to 128 \to 2$ head with ReLU between linear layers, totaling 2,263,170 trainable parameters. Both adapt the MultiBench fusion implementation at commit `49f9be224f342de005b7f001281f33df4301eb22` to two modalities and binary entry prediction; no text input is used.

TRN-EgoVLP and TRN-AV-PF. The visual adapter uses the official multi-scale Temporal Relation Network at commit `060509c4a25bb83d9dfa7c3444e850a49b8bed88`. The normalized 15×256 EgoVLP sequence supplies relation scales $k = 15, 14, \ldots, 2$. Each scale uses ReLU, a $256k \to 256$ layer, ReLU, and a $256 \to 2$ layer. Training uses the unique 15-frame relation and three sampled combinations without replacement at every smaller scale, summing 40 relation outputs. Inference fixes these combinations using random seed 20260926, shared across folds and seeds; predictions therefore do not depend on batch order. The head has 7,809,564 trainable parameters and resets for every query. TRN-AV-PF combines its probability with A25 using the same five-point development grid as LOOM. Neither TRN nor the other external prediction heads carries state across queries; sequential state resides in the warning layer.

B.3 Warning Policies and Event Accounting

P1 emits when $p > \theta$, subject to a 0.2 s cooldown. P2 additionally requires an armed state; it rearms when $p \leq \theta/2$ or 0.5 s has elapsed since the last alarm. P3 requires the preceding processed query also to exceed threshold, with at most 0.5 s between those queries, and retains the cooldown. Predictors reset between historical windows; only the alarm layer carries state across queries.

Thresholds are $0, 0.002, \ldots, 1$, plus 2 as a silence option. For each development burden budget in $\{0.25, 0.5, 1\}$, selection maximizes $R_{1\mathrm{s}}$, then prefers smaller burden, larger threshold, and the fixed order P1, P2, P3. Main results use budget 0.5. A delivered alarm is matched to the earliest event

strictly after its query and within the three-second forecast horizon. If no such event exists, it is false; if delivery is at or after the event, it is late; if that event already received a valid warning, it is duplicate. An already matched earliest event is not skipped to claim a later event. The first valid delivery determines lead-time credit. These rules apply uniformly to all 75 events and 150 episodes.

B.4 Latency and Statistical Evaluation

The evaluation platform uses Python 3.12.10, CUDA runtime 12.8, and an NVIDIA GeForce RTX 4090. Core branches and AGA variants use profiles collected for the AGA comparison; MMTM, LMF, TFN, and TRN variants use their corresponding complete-input profiles. The four early-fusion/audio-statistics ablations retain their original profiles. The 135 system/fold/seed profiles comprise 99 main-comparison and 36 ablation profiles, all using the same sampling and timing procedure. Each system/fold/seed latency profile samples 151 queries: one deterministic query per episode and one extra query. Five warmups precede three repetitions. Measurements include raw RGB/PCM preprocessing, prediction, fusion where applicable, and the alarm update; disk I/O is measured separately. Branches with zero contribution are skipped. Reported medians and P95 values follow the profile aggregation used for the evaluated model instances, rather than an ensemble of seeds.

Warning replays use fixed median or P95 delays and skip queries arriving during computation. Event matching occurs only in the evaluator. Intervals are the 2.5th and 97.5th percentiles of 10,000 paired bootstrap draws over the ten template groups, averaging seeds within each draw. Group sampling preserves within-template correlations.

C Additional Results

Table 6: Core systems by seed. AP, TM-AUC, and R_{1s} are higher-is-better; B_{test} is lower-is-better. Recall uses development budget 0.5 and median-cost replay. Burden counts false, duplicate, and late alarms per episode.

System	Seed	Pooled AP	TM-AUC	R_{1s}	B_{test}
LOOM	42	0.5001	0.6913	0.4667	0.8067
LOOM	44	0.5713	0.7513	0.3733	0.7333
LOOM	46	0.5601	0.7345	0.4133	0.6467
LOOM-V	42	0.5269	0.7831	0.3867	0.6867
LOOM-V	44	0.5080	0.7801	0.4000	0.8333
LOOM-V	46	0.5352	0.7748	0.4267	0.8667
A25	42	0.4081	0.6228	0.4267	0.6600
A25	44	0.5213	0.6879	0.2667	0.6733
A25	46	0.5035	0.6851	0.3200	0.4400
LOOM-AV-CLF	42	0.4538	0.6927	0.4800	0.8800
LOOM-AV-CLF	44	0.4777	0.7480	0.4267	0.9000
LOOM-AV-CLF	46	0.4913	0.7390	0.4133	0.6800

Table 6 reports the three independently trained seeds for the primary fusion comparison. The paired template-bootstrap interval for LOOM minus A25 pooled AP is $[0.0098, 0.1916]$; for LOOM minus LOOM-V it is $[-0.0623, 0.0662]$. The corresponding one-second recall intervals are $[-0.0147, 0.2014]$ and $[-0.2267, 0.2835]$. These intervals summarize variation across template resampling, with seeds averaged within each draw.

Table 7: Development-selected LOOM visual weights α; audio receives weight $1 - \alpha$. Each weight is fixed for its fold and seed.

Fold	Seed 42	Seed 44	Seed 46
0	0	0.25	0
1	0.5	0	0.25
2	0.75	0.75	0.75

Six configurations combine visual and acoustic evidence, while three select audio alone (Table 7). The selected endpoints are included in both prediction and latency summaries. Across seeds, LOOM obtains mean negative log-likelihood 0.5223 and Brier score 0.1754, compared with 0.5300 and 0.1796 for LOOM-V. The accompanying machine-readable results file contains all sixteen systems and 48 system–seed records.

The external adaptations have higher mean NLL and Brier score than LOOM (Table 9). Together with the different pooled and fold AP rankings, these probability measures help characterize how learned scores transfer across held-out templates. They do not isolate the effects of the frozen representation, the adaptation head, and optimization.

The LOOM–LMF pooled-AP interval lies above zero; all six other adapted-comparator intervals include zero. Warning recall gains over LMF and TFN are 0.2000 and 0.1867, with intervals $[-0.0133, 0.4658]$ and $[-0.0595, 0.4685]$. Their mean burden differences are -0.4422 and -0.7467. Against TRN-EgoVLP, recall improves by 0.3067 with interval $[0.1000, 0.5524]$, while burden increases by 0.3644. Against TRN-AV-PF, the corresponding recall difference is 0.1289 with interval $[-0.0093, 0.3063]$ and burden increase 0.0778. These intervals describe development-selected operating points, not matched-test-burden comparisons.

The higher mean fold AP of TRN-AV-PF is concentrated in fold 2, which tests slider and wheeled templates (Table 11). TFN also performs substantially better on this fold than on the other two. These differences emphasize reporting held-out-template results alongside pooled scores.

Table 8: Adapted comparison systems by seed, using the same metrics, development budget, and median-cost replay as Table 6. AGA-AV-PF and TRN-AV-PF combine the corresponding visual predictor with A25.

System	Seed	Pooled AP	TM-AUC	R_{1s}	B_{test}
AGA-EgoVLP	42	0.4008	0.6392	0.1067	0.2667
AGA-EgoVLP	44	0.4207	0.6687	0.0000	0.0000
AGA-EgoVLP	46	0.3778	0.6508	0.1600	0.6267
AGA-AV-PF	42	0.3995	0.6261	0.2533	0.4467
AGA-AV-PF	44	0.3902	0.6842	0.2533	0.6733
AGA-AV-PF	46	0.3778	0.6508	0.1600	0.6267
MMTM-AV	42	0.4030	0.6434	0.0800	0.2467
MMTM-AV	44	0.4407	0.7161	0.1467	0.4133
MMTM-AV	46	0.5185	0.7385	0.0400	0.0200
LMF-AV	42	0.3864	0.5897	0.1333	0.5733
LMF-AV	44	0.3130	0.5436	0.3733	2.3733
LMF-AV	46	0.4412	0.6027	0.1467	0.5667
TFN-AV	42	0.3837	0.6586	0.1200	0.5200
TFN-AV	44	0.3777	0.5627	0.3733	3.1533
TFN-AV	46	0.5201	0.7108	0.2000	0.7533
TRN-EgoVLP	42	0.4871	0.6547	0.1600	0.5333
TRN-EgoVLP	44	0.4596	0.6420	0.1200	0.4467
TRN-EgoVLP	46	0.3956	0.6464	0.0533	0.1133
TRN-AV-PF	42	0.4347	0.6940	0.2933	0.5733
TRN-AV-PF	44	0.4280	0.6861	0.2667	0.8267
TRN-AV-PF	46	0.3982	0.6934	0.3067	0.5533

Table 9: Probability quality and complete-input computation time, averaged over three seeds. Lower NLL, Brier score, and latency are better. Median and P95 refer to the P1 path in milliseconds, excluding disk I/O. The timing profiles are those used for Table 1.

System	NLL	Brier	Median (ms)	P95 (ms)
LOOM (ours)	0.522	0.175	8.8	11.5
LOOM-V	0.530	0.180	2.1	2.9
A25	0.605	0.192	6.9	11.6
LOOM-AV-CLF	0.595	0.208	8.6	11.1
AGA-EgoVLP	2.731	0.304	72.4	85.2
AGA-AV-PF	2.445	0.287	71.4	82.2
MMTM-AV	0.927	0.222	56.6	72.1
LMF-AV	2.373	0.274	55.7	64.6
TFN-AV	2.410	0.264	54.7	57.4
TRN-EgoVLP	4.303	0.294	54.1	57.9
TRN-AV-PF	2.700	0.258	54.6	60.1
TIME	1.260	0.254	–	–
LOOM-EF	0.548	0.186	6.9	9.1
ALS-linear	0.607	0.209	4.9	6.2
LOOM-EF-MLP	0.568	0.188	6.9	9.2
LOOM-EF-cal	0.571	0.188	6.8	8.7

Table 10: Paired pooled-AP differences: LOOM minus each adapted comparator. Intervals use 10,000 template-group bootstrap draws, averaging seeds inside each draw.

Comparator	AP difference	95% interval
AGA-EgoVLP	0.1440	$[-0.0652,\ 0.2707]$
AGA-AV-PF	0.1546	$[-0.0508,\ 0.2798]$
MMTM-AV	0.0897	$[-0.0187,\ 0.1947]$
LMF-AV	0.1636	$[0.0017,\ 0.2748]$
TFN-AV	0.1166	$[-0.0494,\ 0.2448]$
TRN-EgoVLP	0.0964	$[-0.0854,\ 0.2390]$
TRN-AV-PF	0.1235	$[-0.0646,\ 0.2435]$

Table 11: Per-fold AP, averaged over three seeds. The corresponding fold partitions are listed in Table 5. Every row uses the same episodes and queries.

System	Fold 0	Fold 1	Fold 2
LOOM (ours)	0.5157	0.5233	0.5792
LOOM-V	0.6140	0.4956	0.5545
A25	0.4818	0.5171	0.5004
LOOM-AV-CLF	0.4818	0.5404	0.5694
AGA-EgoVLP	0.5087	0.3213	0.7456
AGA-AV-PF	0.4894	0.3213	0.7456
MMTM-AV	0.4467	0.4534	0.7976
LMF-AV	0.3973	0.2946	0.7750
TFN-AV	0.4776	0.3495	0.8194
TRN-EgoVLP	0.4286	0.3647	0.8790
TRN-AV-PF	0.4831	0.3647	0.8790
TIME	0.4371	0.4390	0.4511
LOOM-EF	0.5670	0.4959	0.5462
ALS-linear	0.3954	0.2813	0.3429
LOOM-EF-MLP	0.5671	0.5158	0.5604
LOOM-EF-cal	0.5670	0.4959	0.5462

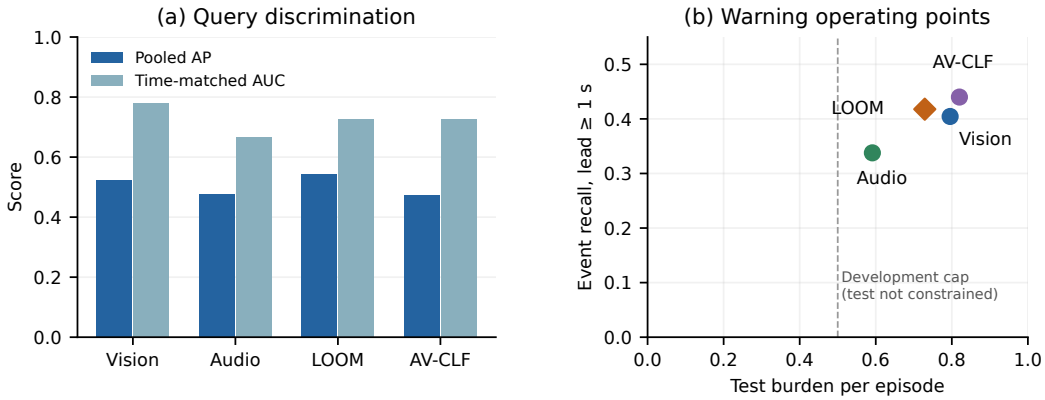


Figure 3: Core branch and fusion outcomes averaged over three seeds. Left: pooled AP and time-matched AUC. Right: one-second event recall and observed test alarm burden for development-selected policies. Points are individual operating configurations, not a threshold-swept frontier; bars and points denote means rather than confidence intervals.

Generation-subset diagnostics. The following supplementary analyses were added in response to review and do not change the original models, partitions, or metrics. They use the existing out-of-fold predictions, episode-to-subset mapping, and training-fold labels. Subset A has 663 positive queries out of 1,638 (40.48%), whereas Subset B has 506 out of 2,262 (22.37%). The subsets also differ in rendering, acoustic construction, and motion-family composition (Appendix A). Consequently, a subset-dependent score can contain label information even without observing motion or sound.

Table 12 evaluates every existing system separately within each subset. For the core branches, LOOM obtains AP 0.5466/0.5320 in A/B, LOOM-V obtains 0.5108/0.5693, and A25 obtains 0.5181/0.3806. TIME obtains 0.5293/0.3800. These are means of the three seed-specific APs. Differences between subsets combine changes in class prevalence, temporal label structure, template composition, and generation pipeline; they are not controlled estimates of a pipeline effect.

Table 12: Existing systems by generation subset, averaged over seeds 42/44/46. Higher AP and TM-AUC and lower NLL and Brier are better; bold marks the best displayed value per column. A contains 63 episodes and B contains 87. Within-subset TM-AUC uses eligible fold–time cells: 77 cells/1,614 queries in A and 72 cells/2,024 queries in B.

	Subset A				Subset B			
System	AP	TM-AUC	NLL	Brier	AP	TM-AUC	NLL	Brier
LOOM (ours)	0.5466	0.6570	0.630	0.221	0.5320	0.7200	**0.444**	**0.142**
LOOM-V	0.5108	0.7017	**0.626**	0.224	**0.5693**	**0.7933**	0.461	0.147
A25	0.5181	0.6543	0.684	0.232	0.3806	0.6119	0.547	0.163
LOOM-AV-CLF	0.4963	0.6563	0.685	0.248	0.4759	0.7228	0.530	0.180
AGA-EgoVLP	0.5291	0.6618	3.353	0.350	0.3236	0.6714	2.280	0.271
AGA-AV-PF	0.5158	0.6679	2.906	0.325	0.3123	0.6650	2.111	0.259
MMTM-AV	0.5864	**0.7392**	1.108	0.241	0.3382	0.6571	0.796	0.207
LMF-AV	0.5931	0.6688	3.378	0.308	0.2854	0.6096	1.645	0.248
TFN-AV	**0.6150**	0.7077	3.396	0.306	0.2904	0.6674	1.696	0.234
TRN-EgoVLP	0.5437	0.6410	6.320	0.381	0.3774	0.7122	2.842	0.232
TRN-AV-PF	0.4893	0.6565	3.668	0.329	0.3573	0.7576	1.998	0.206
TIME	0.5293	0.5000	1.198	0.252	0.3800	0.5000	1.305	0.256
LOOM-EF	0.5764	0.6859	0.629	0.222	0.4370	0.7801	0.490	0.161
ALS-linear	0.4551	0.5954	0.707	0.255	0.3013	0.5862	0.534	0.175
LOOM-EF-MLP	0.5208	0.6562	0.699	0.233	0.4988	0.7911	0.474	0.155
LOOM-EF-cal	0.5804	0.6859	0.663	**0.215**	0.4780	0.7801	0.505	0.168

We construct three metadata diagnostics using training labels only. TRAIN-PRIOR assigns the overall training-fold positive rate; BATCH assigns the rate for the episode's generation subset; BATCH+TIME conditions on both subset and the exact query time. Every rate uses add-one smoothing, $(n_+ + 1)/(n + 2)$. An absent subset–time cell falls back to its training-subset rate, then to the overall training rate if the subset is absent. These fallbacks are specified for completeness; all queried cells have training support in this dataset. The diagnostics involve no neural fitting, development-label selection, or test-label estimation. Their test inputs are metadata rather than RGB or audio, and they are not inputs to LOOM. TIME retains its original fixed decreasing query-clock score.

Table 13 reports both original metrics and an additional conditional AUC (C-AUC). The latter computes AUC separately in each *outer-fold, query-time, generation-subset* cell containing both labels, then averages by the number of queries in each eligible cell. It therefore compares scores within a subset as well as within the original fold–time strata. Of 156 possible cells, 149 are eligible, covering 3,638/3,900 queries (93.28%); seven single-class cells with 262 queries are excluded. The same cells and weights apply to every system. Subset A contributes 1,614/1,638 queries (98.53%), and B contributes 2,024/2,262 (89.48%). C-AUC is supplementary and does not replace TM-AUC, which covers all 3,900 queries here.

BATCH reaches TM-AUC 0.6293, demonstrating that the original time-matching operation does not eliminate information associated with subset membership. BATCH+TIME reaches pooled AP 0.4678, while all metadata diagnostics have C-AUC 0.5000 because their scores are constant within each conditional cell. LOOM, LOOM-V, and A25 obtain C-AUC 0.6921, 0.7526, and 0.6307. Their within-cell discrimination cannot be explained by these metadata alone. It does not identify the

physical or generation-related cues used; fusion's pooled-AP advantage may partly reflect between-subset differences. Differences between C-AUC and TM-AUC also reflect the excluded single-class cells and changed weighting.

Table 13: Supplementary metadata diagnostics and conditional discrimination. Higher values are better; bold marks the best displayed value per column. Existing systems are averaged over their three seeds. TRAIN-PRIOR, BATCH, and BATCH+TIME are deterministic training-fold estimates and are evaluated once. Original AP, fold AP, and TM-AUC are retained; C-AUC additionally conditions on subset with the coverage stated in the text.

System	Pooled AP	Fold AP	TM-AUC	C-AUC
LOOM (ours)	**0.5438**	0.5394	0.7257	0.6921
LOOM-V	0.5234	0.5547	**0.7793**	**0.7526**
A25	0.4776	0.4997	0.6653	0.6307
LOOM-AV-CLF	0.4743	0.5305	0.7266	0.6933
AGA-EgoVLP	0.3998	0.5252	0.6529	0.6671
AGA-AV-PF	0.3892	0.5188	0.6537	0.6663
MMTM-AV	0.4541	0.5659	0.6994	0.6935
LMF-AV	0.3802	0.4890	0.5787	0.6359
TFN-AV	0.4272	0.5488	0.6440	0.6853
TRN-EgoVLP	0.4474	0.5574	0.6477	0.6806
TRN-AV-PF	0.4203	**0.5756**	0.6912	0.7127
TIME	0.4388	0.4424	0.5000	0.5000
LOOM-EF	0.4992	0.5363	0.7607	0.7383
ALS-linear	0.3173	0.3399	0.4707	0.5903
LOOM-EF-MLP	0.5073	0.5477	0.7655	0.7312
LOOM-EF-cal	0.5244	0.5363	0.7607	0.7383
TRAIN-PRIOR	0.2796	0.3047	0.5000	0.5000
BATCH	0.3539	0.3606	0.6293	0.5000
BATCH+TIME	0.4678	0.4735	0.5891	0.5000

Tables 14 and 15 describe the support available to these analyses. Subset A contains no bounce_roll or slide_drop episodes; other template counts are also imbalanced. Subset stratification cannot recover these missing combinations or separate template from generation effects. The accompanying machine-readable diagnostic results include every subset–template–partition cell, including zeros, together with fold estimates, per-seed metrics, probability-quality scores, and coverage. These records support reproducing the diagnostics without treating them as a batch-balanced experiment or a causal intervention.

Table 14: Generation-subset support by evaluation template. Columns give episodes (Ep.), queries (Q), positive queries (+), and entry events (Ev.). Counts refer to each episode once, before fold assignment; zeros denote absent template–subset combinations.

	Subset A				Subset B			
Template	Ep.	Q	+	Ev.	Ep.	Q	+	Ev.
bounce_roll	0	0	0	0	15	390	88	7
hinge	9	234	115	6	6	156	31	2
push_off	1	26	3	1	2	52	16	1
rail	14	364	156	8	1	26	0	0
road	13	338	145	7	2	52	0	0
slide_drop	0	0	0	0	15	390	72	7
slider	8	208	0	0	7	182	112	7
throw	1	26	0	0	11	286	48	6
tip	1	26	11	1	14	364	92	7
wheeled	16	416	233	13	14	364	47	2

Table 15: Subset composition in each fold and partition. Positive rates are percentages of queries; event counts include all entries in the corresponding episodes. Every episode belongs to one partition per fold, and exactly one test partition across folds.

Fold	Partition	Subset	Episodes	Queries	Positive	Rate (%)	Events
0	Train	A	27	702	247	35.19	15
0	Train	B	63	1638	387	23.63	30
0	Development	A	14	364	156	42.86	8
0	Development	B	1	26	0	0.00	0
0	Test	A	22	572	260	45.45	13
0	Test	B	23	598	119	19.90	9
1	Train	A	33	858	348	40.56	19
1	Train	B	42	1092	278	25.46	18
1	Development	A	13	338	145	42.90	7
1	Development	B	2	52	0	0.00	0
1	Test	A	17	442	170	38.46	10
1	Test	B	43	1118	228	20.39	21
2	Train	A	26	676	285	42.16	16
2	Train	B	64	1664	347	20.85	30
2	Development	A	13	338	145	42.90	7
2	Development	B	2	52	0	0.00	0
2	Test	A	24	624	233	37.34	13
2	Test	B	21	546	159	29.12	9