LISTEN FIRST, LOOK AS NEEDED: UNCERTAINTY-AWARE ASYMMETRIC ROUTING FOR AUDIO-VISUAL RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Efficient audio-visual recognition should preserve accuracy while activating vision only when required. Existing static, visual-centric dynamic, and audio-preview methods allocate fixed budgets, route after visual processing begins, or use audio to rank visual segments, but do not calibrate audio-only acceptance before activating vision. We propose Uncertainty-Aware Asymmetric Routing (UAAR), a "listen first, look as needed" framework separating pre-visual acceptance from progressive multimodal verification. An empirical selective-risk criterion governs zero-frame exits, and an entropy-based reliability test disables audio guidance under degradation. Unresolved samples acquire sparse clips and autoregressively accumulate cross-modal evidence under three-, six-, or nine-frame budgets. On Kinetics-Sounds and AVE, UAAR achieves 83.35% and 85.82% accuracy while reducing activated computation by 57.3% and 53.4%, respectively, relative to full nine-frame inference. Under progressive audio corruption, UAAR suppresses unsafe audio-only exits and adaptively requests additional visual evidence. Code will be made publicly available upon publication.

1 INTRODUCTION

The central objective of efficient audio-visual recognition is to preserve accuracy while activating only the sensory computation each sample requires. The visual budget should therefore vary with evidence demand: acoustically distinctive events may require no vision, whereas ambiguous or corrupted audio requires progressively stronger visual verification. A fixed budget cannot accommodate this heterogeneity; full visual processing wastes computation on audio-sufficient samples, while uniform reduction withholds evidence from difficult ones. Sample-adaptive visual activation is therefore necessary for a favorable accuracy–computation trade-off. In continuous, resource-constrained settings, audio provides inexpensive, always-on context, whereas visual encoding constitutes the dominant sensing cost (Chalk et al., 2024; Geng et al., 2025; Piergiovanni et al., 2024). This modality asymmetry motivates a natural design principle: **first determine whether the audio evidence is reliable and sufficient, and activate progressively more visual computation only when additional evidence is required.**

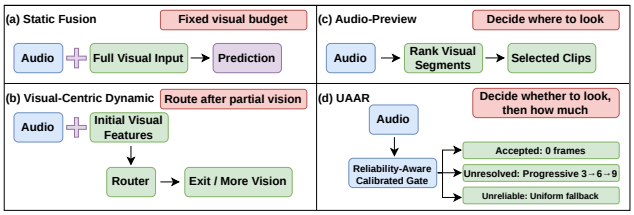

Figure 1: Visual-acquisition paradigms. UAAR calibrates visual necessity before activation, then progressively allocates evidence to unresolved samples and invokes a fallback for unreliable audio.

Existing methods, summarized in Figure 1, reduce computation at different stages of the inference pipeline. Static designs compress the network or employ lightweight asymmetric backbones (Chen

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.

et al., 2021; Chalk et al., 2024; Liu et al., 2025), but retain a fixed sensing budget. Visual-centric dynamic methods reduce computation through early exits, frame or token selection, or adaptive backbones (Ghodrati et al., 2021; Wang et al., 2022; Han et al., 2023; Cao et al., 2024; Luo et al., 2025). Audio-preview methods use acoustic cues to rank candidate visual locations before encoding them (Gao et al., 2020; Nugroho et al., 2023). Adaptive modality selection also allows expensive branches to be skipped: HCMS uses audio and historical states to gate appearance and motion processing (Weng et al., 2023), while AdaMML learns per-segment modality choices with a lightweight multimodal policy network (Panda et al., 2021). These methods establish the value of audio-first processing and input-dependent modality participation.

Yet learning to reduce modality usage does not explicitly calibrate *which audio predictions can be accepted without visual correction*. HCMS and AdaMML optimize classification and computation-related objectives, rather than selecting audio-only acceptance from held-out selective error. An incorrect audio-only exit forfeits complementary visual evidence; rejecting that exit alone is also insufficient if uncertain audio still guides the system toward uninformative visual locations. Conversely, requiring vision does not imply requiring the full visual budget. The central challenge is therefore to couple *selective audio acceptance* with *adaptive visual verification*: determine when audio can stand alone, when its sampling guidance should be discarded, and when the accumulated visual evidence is sufficient to stop.

To address this challenge, we propose Uncertainty-Aware Asymmetric Routing (UAAR), which factorizes dynamic audio-visual inference into two distinct decision levels. The first is a *pre-activation participation decision* that determines whether vision is needed before any visual feature is computed. The second is a *post-activation evidence-allocation process* that progressively acquires visual clips only for unresolved samples. A held-out empirical acceptance rule makes the audio-exit operating point explicit without requiring visual preview features. An uncertainty-triggered fallback governs both acceptance and sampling, while fusion-based stopping prevents visual activation from committing every unresolved sample to the full budget. This factorization connects the decision to invoke vision with the decision to stop acquiring visual evidence, supporting zero-, three-, six-, or nine-frame inference.

Concretely, UAAR selects the audio-only acceptance threshold after training using the empirical error among accepted predictions on held-out data. An entropy threshold h_{deaf} disables both audio-only acceptance and audio-guided sampling when predictive uncertainty is high, switching to uniformly sampled visual clips. Otherwise, unresolved samples use audio saliency to prioritize sparse clips. The Autoregressive Fusion Module accumulates audio-conditioned visual evidence through Dimension-wise Gated Fusion (DGF) and exposes an intermediate prediction after each clip, enabling three-, six-, or nine-frame execution. Together, these designs address two sources of unnecessary computation: activating vision for audio-sufficient samples and continuing visual processing after the fused prediction meets the stopping criterion. Calibration is empirical, and the fallback targets high predictive uncertainty rather than guaranteeing detection of every audio failure.

In summary, our main contributions are as follows:

- **Selective Audio-Only Acceptance:** An audio-only exit eliminates the opportunity for visual correction, making acceptance quality an explicit routing concern. UAAR selects its exit threshold from the empirical error among accepted predictions on held-out data, separating acceptance calibration from representation learning and making the exit operating point explicitly adjustable after training.
- **Uncertainty-Gated Acceptance and Acquisition:** Audio influences both whether vision is activated and where visual evidence is acquired. Rejecting an uncertain audio prediction alone leaves its sampling guidance unchecked. UAAR therefore uses a shared entropy-triggered fallback to disable both audio-only acceptance and audio-guided sampling, replacing the latter with uniform temporal coverage.
- **Progressive Visual Verification:** Failure to accept an audio prediction does not imply that the full visual budget is necessary. UAAR couples pre-visual acceptance with post-fusion stopping: autoregressive DGF accumulates complementary evidence and exposes intermediate predictions that determine whether further clips are needed. This connects the decision to invoke vision with the decision to stop acquiring visual evidence.

2 Related Work

Audio-Visual Representation Learning. Audio-visual recognition learns complementary acoustic and visual representations through cross-modal correspondence or contrastive objectives (Arandjelovic & Zisserman, 2017; Chen et al., 2021), attention and temporal interaction (Nagrani et al., 2021; Chalk et al., 2024), and modality-aware or hierarchical collaboration (Wei et al., 2025; Meng et al., 2026). These methods improve fusion after both streams have been encoded, but typically use a fixed visual budget and therefore optimize representation quality without deciding whether visual sensing is required.

Dynamic Multimodal Inference. Dynamic inference reduces cost through early exits, frame or token selection, and adaptive backbones (Ghodrati et al., 2021; Wang et al., 2022; Han et al., 2023; Luo et al., 2025); audio preview prioritizes visual locations (Gao et al., 2020; Nugroho et al., 2023), while quality-aware fusion adjusts modality contributions (Zhang et al., 2023; Cao et al., 2024). More directly related, HCMS (Weng et al., 2023) uses audio and historical states to gate appearance and motion networks through hierarchical LSTMs. AdaMML (Panda et al., 2021) learns per-segment modality choices with a lightweight multimodal policy network, including visual preview features, and aggregates segment predictions. Both optimize classification and modality-use objectives. UAAR shares their goal of avoiding unnecessary computation, but separates post-training selection of the audio acceptance threshold from representation learning. It combines this held-out empirical calibration with an entropy-triggered veto of both acceptance and audio-guided sampling, followed by stopping based on progressively fused predictions. The focus is therefore the coordination of acceptance, sampling, and stopping rather than modality skipping alone.

Robust Multimodal Learning and Risk Calibration. Robust multimodal methods reweight, reconstruct, or route already observed features under modality mismatch and sensor failure (Mckinzie et al., 2023; Park et al., 2025; Zhang et al., 2025). Selective prediction and CRC instead calibrate decision risk (Geifman & El-Yaniv, 2017; 2019; Angelopoulos et al., 2024), including recent representation calibration for low-quality multimodal inputs (Jiang et al., 2026). UAAR connects these objectives by extending risk calibration from prediction confidence to sensing allocation: a calibrated audio decision governs visual acquisition, a corruption-aware fallback overrides unreliable audio exits, and progressive fusion verifies unresolved samples.

3 Methodology

3.1 Problem Formulation and Framework Overview

Given an audio Mel spectrogram $X_a \in \mathbb{R}^{F\times T_a}$ and its accompanying video stream X_v, UAAR predicts an event label with an input-dependent budget of zero, three, six, or nine visual frames (Figure 2). Audio-only acceptance is evaluated before visual feature extraction; unresolved samples proceed to multimodal fusion.

The audio stream is always processed because it provides the low-cost routing prior. MobileOne (Vasu et al., 2023b) maps X_a to a temporal representation $A_{\text{seq}} \in \mathbb{R}^{T_s\times d_a}$; temporal pooling gives the global prior $A \in \mathbb{R}^{d_a}$, and an audio classifier produces $P_a \in \mathbb{R}^C$. These outputs serve two distinct purposes: P_a determines whether audio alone is reliable and sufficient, while A_{seq} identifies candidate visual timestamps when verification is needed. Only the selected clips are then processed by a FastViT–TSM visual encoder (Vasu et al., 2023a; Lin et al., 2019). A clip c_t containing three frames yields $v_t \in \mathbb{R}^{d_v}$, and at most three clips are acquired sequentially.

3.2 Uncertainty-Aware Asymmetric Routing

The router sequentially evaluates audio reliability, audio-only acceptance, and visual acquisition.

Stage 1: Acoustic-Reliability Fallback. We first use predictive entropy as a lightweight operational indicator of audio reliability:

$$h_{audio} = -\sum_{c=1}^{C} p_{a,c}\log p_{a,c}\,. \tag{1}$$

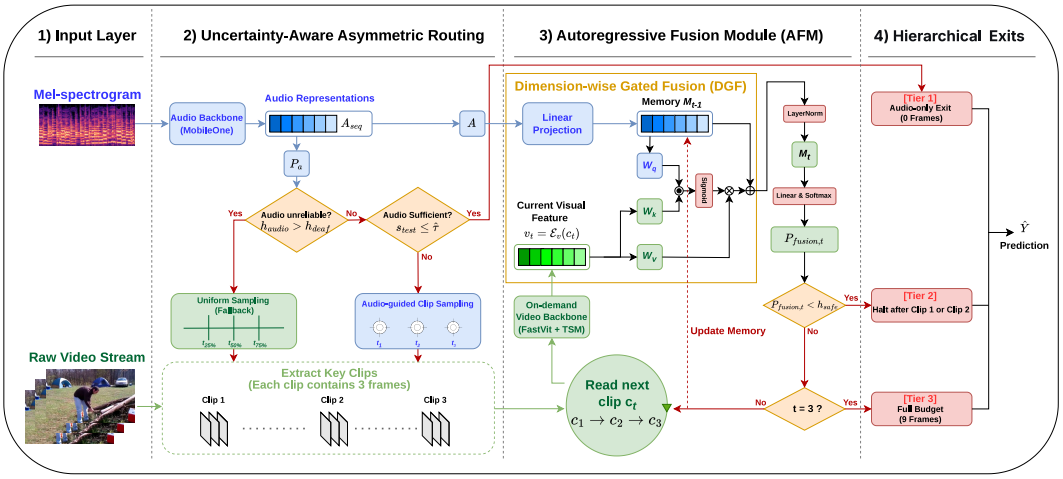

Figure 2: Overall architecture of the Uncertainty-Aware Asymmetric Routing (UAAR). An entropy-based fallback treats high-entropy audio predictions as unreliable; held-out selective-risk calibration determines $\hat{\tau}$ for a 0-frame exit (Tier-1). Otherwise, the Autoregressive Fusion Module (AFM) integrates sparse visual clips through Dimension-wise Gated Fusion (DGF), stopping when $h_{\text{fusion},t} < h_{\text{safe}}$ (Tier-2) or after all sampled clips are processed (Tier-3).

If $h_{audio} > h_{deaf}$, the router requests visual verification at $\mathcal{T}_{\text{uni}} = \{t_{25\%}, t_{50\%}, t_{75\%}\}$ instead of using audio saliency; otherwise, it proceeds to Stage 2. The operational threshold h_{deaf} depends on the class space and calibration and is specified in the experimental configuration. Predictive entropy does not identify every confidently incorrect prediction or provide general-purpose out-of-distribution detection.

Stage 2: Held-Out Calibration of Audio-Only Acceptance. Following selective prediction (Geifman & El-Yaniv, 2017), we select the audio-exit threshold after training using held-out empirical accepted error.

Given $\mathcal{D}_{cal} = \{(X_i, y_i)\}_{i=1}^n$, let $s_i = 1 - \max_c p_{a,i,c}$ and $\hat{y}_i = \arg\max_c p_{a,i,c}$. For a candidate threshold τ, define the accepted-sample count $m(\tau) = \sum_{i=1}^n \mathbb{I}(s_i \le \tau)$ and the corresponding error count $e(\tau) = \sum_{i=1}^n \mathbb{I}(s_i \le \tau)\mathbb{I}(\hat{y}_i \neq y_i)$. The empirical selective error is

$$\hat{R}(\tau) = \frac{e(\tau)}{m(\tau)}, \qquad m(\tau) > 0. \tag{2}$$

Coverage is $m(\tau)/n$. Calibration counts confidence-selected samples without the entropy veto; deployment additionally applies Stage 1.

For a calibration target ϵ, the empirical rule searches the observed candidate scores $\mathcal{T} = \{s_i\}_{i=1}^n$:

$$\hat{\tau} = \max\big(\{-1\} \cup \{\tau \in \mathcal{T} : m(\tau) > 0, \hat{R}(\tau) \le \epsilon\}\big). \tag{3}$$

Including all tied scores, the largest feasible threshold maximizes calibration coverage; $\hat{\tau} = -1$ disables exits if none is feasible. Unlike formal CRC (Angelopoulos et al., 2024), ϵ is an empirical target, not a test-error bound. This rule applies to KS; dataset-specific routing settings are given in Sec. 4.1.

With $\hat{\tau}$ frozen for testing, a **Tier-1 audio-only exit** requires $h_{audio} \le h_{deaf}$ and $s_{\text{test}} \le \hat{\tau}$. High-entropy samples use Stage 1's uniform fallback; the remaining non-accepted samples proceed to Stage 3. Test selective error and coverage are measured on deployed exits.

Stage 3: Audio-Guided Visual Acquisition. For samples reaching Stage 3, candidate timestamps come directly from A_{seq}, without a visual proposal pass. Channel-wise root mean square defines temporal saliency:

$$S_t = \sqrt{\frac{1}{d_a}\sum_{j=1}^{d_a} A_{\text{seq}}(t,j)^2}\,. \tag{4}$$

Under a weak synchronization assumption that prominent acoustic events are coarsely associated with informative visual moments (Gao et al., 2020), 1D non-maximum suppression removes neighboring peaks and the three highest remaining responses define

$$\mathcal{T}_{active} = \underset{t\in\{1,\dots,T_s\}}{\arg\operatorname{topk}} \; S_t \,. \tag{5}$$

The corresponding clips are processed only when requested and passed sequentially to the autoregressive fusion module.

3.3 Autoregressive Fusion Module with DGF

A shared autoregressive DGF cell processes the requested clips sequentially. Its memory is initialized as $M_0 = AW_p \in \mathbb{R}^d$, with $W_p \in \mathbb{R}^{d_a \times d}$. At step $t \in \{1, 2, 3\}$, the audio-conditioned state M_{t-1} and new visual feature v_t determine a dimension-wise gate:

$$G_t = \sigma\left(\frac{(M_{t-1}W_q) \odot (v_t W_k)}{\sqrt{d}}\right) \in \mathbb{R}^d \,, \tag{6}$$

where $W_q \in \mathbb{R}^{d\times d}$, $W_k \in \mathbb{R}^{d_v \times d}$, $\odot$ denotes element-wise multiplication, and σ is sigmoid. The gated visual feature is

$$F_t = G_t \odot (v_t W_v) \,. \tag{7}$$

with $W_v \in \mathbb{R}^{d_v \times d}$. Memory is updated residually:

$$M_t = \text{LayerNorm}(M_{t-1} + F_t) \,. \tag{8}$$

All projections are shared across the three steps.

After each update, $W_{\text{cls}} \in \mathbb{R}^{d\times C}$ produces $P_{\text{fusion},t} = \text{Softmax}(M_t W_{\text{cls}})$ with entropy $h_{\text{fusion},t}$. The validation-selected h_{safe} trades accuracy against clip count, rather than calibrating selective error. If $h_{\text{fusion},t} < h_{\text{safe}}$ at $t \in \{1, 2\}$, a **Tier-2 fused exit** skips the remaining clips; otherwise, processing continues to M_3 for the **Tier-3 full-budget prediction**.

3.4 Progressive Supervision and Inference Cost

Training evaluates the audio branch and all three fusion steps so that every inference route has a directly supervised classifier:

$$\mathcal{L} = \text{CE}(P_a, Y) + \sum_{t=1}^{3} \lambda_t \,\text{CE}(P_{\text{fusion},t}, Y), \tag{9}$$

where $\lambda = \{0.5, 0.8, 1.0\}$ places increasing weight on predictions with more visual evidence. The routing thresholds are not optimized by this loss: $\hat{\tau}$ is estimated on $\mathcal{D}_{cal}$ after training, whereas h_{deaf} and h_{safe} are selected on $\mathcal{D}_{val}$ as described above.

Let $K \in \{0, 1, 2, 3\}$ be the number of three-frame clips processed at inference and let $\pi_k = \Pr(K = k)$. If $\mathcal{C}_A$ includes the audio encoder, classifier, and routing operations, $\mathcal{C}_V$ is the cost of encoding one visual clip, and $\mathcal{C}_F$ is one fusion update, the expected activated computation is

$$\mathbb{E}[\mathcal{C}] = \mathcal{C}_A + \sum_{k=0}^{3} \pi_k \, k(\mathcal{C}_V + \mathcal{C}_F). \tag{10}$$

The corresponding average visual budget is $3\mathbb{E}[K]$ frames. We report both quantities from the routes actually activated on the test set, rather than inferring them from the model parameter count.

4 Experiments

4.1 Datasets and Experimental Setup

Datasets and Partitioning. We evaluate sample-level event classification, rather than temporal localization, on Kinetics-Sounds (Arandjelovic & Zisserman, 2017) and Audio-Visual Event (AVE)

(Tian et al., 2018), with separate training, validation ($\mathcal{D}_{val}$), calibration ($\mathcal{D}_{cal}$), and test partitions. Network fitting is separated from post-training routing-parameter selection. The dataset-specific threshold procedures are described below; their empirical calibration does not establish a finite-sample test-risk guarantee.

Architecture and Input Protocol. We use an asymmetric lightweight architecture: *MobileOne-s0* processes audio Mel spectrograms, and *FastViT-t8* with a Temporal Shift Module (TSM) encodes 224×224 visual frames. Forced full-power inference uses the maximum budget of 9 frames per video: 3 salience-guided clips, each containing 3 consecutive frames. Audio segments span 10 seconds, synchronized with the visual temporal window.

Optimization. The original comparison models are trained for up to 40 epochs with AdamW (Loshchilov & Hutter, 2019), validation-based early stopping, weight decay 10^{-2}, and cosine learning-rate decay, using batch sizes of 8 and 16 on Kinetics-Sounds and AVE, respectively. The best-validation checkpoint is retained for test evaluation. The AFM learning rates are 5×10^{-5} and 10^{-4}, while both pretrained sensory backbones use one tenth of the corresponding rate. Network parameters are frozen before threshold selection. The additional dynamic-baseline training settings are given in Appendix A.

Empirical Routing Parameters. The reported configurations use $(\hat{\tau}, h_{\text{safe}}, h_{\text{deaf}}) = (0.0499, 0.7550, 3.0)$ on Kinetics-Sounds and $(0.0240, 0.8, 3.0)$ on AVE. On Kinetics-Sounds, the empirical selective-error search at $\epsilon = 0.05$ yields 0.0499; h_{safe} is chosen from a calibration-set fusion-entropy percentile sweep. On AVE, h_{safe} maximizes validation accuracy minus 0.05 times the average number of visual clips over $\{0.1, 0.2, \ldots, 3.0\}$. Both configurations use the fixed entropy veto 3.0. These procedures select empirical operating points and do not certify the test-time error of accepted predictions.

Computation Metrics. End-to-end GFLOPs are measured with 224×224 visual inputs after MobileOne structural reparameterization. For dynamic methods, we average only the audio, visual, and fusion operations actually activated by each test sample; hence the reported cost reflects input-dependent execution rather than the full-network theoretical bound.

4.2 Comparison with Existing Methods

We compare adapted ARL (Wei et al., 2025), CLCR (Meng et al., 2026), AVGN (Nugroho et al., 2023), and HCMS (Weng et al., 2023) using MobileOne/FastViT recognition backbones, alongside controlled fusion and uniform-sampling baselines. AVGN retains temporal glance but replaces spatial patch selection with the shared frame processor. HCMS provides an audio-first dynamic comparison; its Kinetics-Sounds configuration is selected by validation accuracy, with the AVE operating point reported alongside it. These are adapted implementations, not original-paper results. Method-specific temporal inputs, shared pretrained initializations, and all KS configurations are detailed in Appendix A.

Classification and Efficiency Trade-offs. UAAR-static attains the highest accuracies in Table 1, while UAAR-dynamic trades accuracy for lower computation through sample-dependent acquisition. Relative to adapted AVGN, it gains 3.06 points for 0.24 additional GFLOPs on KS, rather than imposing a fixed three-frame budget. On AVE, it exceeds the reported HCMS operating point by 1.24 points while using 41.7% less computation (2.77 vs. 4.75 GFLOPs).

Dynamic Baselines and Pareto Trade-offs. The lowest-cost tested HCMS configuration on KS (Appendix A) achieves 84.03% at 2.501 GFLOPs, versus UAAR's 83.35% at 2.192: 12.36% less computation with a 0.67-point accuracy reduction. HCMS processes fewer recognition frames (3.099 vs. 3.553), illustrating that frame counts omit audio and policy costs. UAAR remains non-dominated among the evaluated points in Figure 3 and Table A1; this does not establish dominance over complete method frontiers.

4.3 Ablation Study

Table 2 evaluates UAAR on 1,634 Kinetics-Sounds test samples. Routing variants share the trained backbone and differ only in inference policy; the parallel-fusion counterpart is trained separately and compared with full-compute AFM using identical nine-frame evidence.

Table 1: Efficiency and classification performance on Kinetics-Sounds and AVE.

Method	Kinetics-Sounds				AVE			
	Params (M)	Avg. Frames	Avg. GFLOPs	Acc. (%)	Params (M)	Avg. Frames	Avg. GFLOPs	Acc. (%)
Audio-only	1.09	0.00	0.27	43.33	1.08	0.00	0.27	67.91
Visual-only	3.28	9.00	4.85	72.09	3.28	9.00	4.85	66.92
Concat	5.39	9.00	5.13	82.13	4.39	9.00	5.13	85.57
Cross-Attention	7.13	9.00	5.74	81.64	6.13	9.00	5.72	82.84
ARL (ICCV'25)	5.39	9.00	5.13	80.60	4.39	9.00	5.13	85.32
Uniform Sampling	6.65	**3.00**	**1.89**	79.74	6.90	**3.00**	**1.90**	81.34
AVGN (ICCV'23)	4.97	3.00	1.95	80.29	4.95	3.00	1.95	81.84
CLCR (CVPR'26)	10.57	9.00	5.17	82.56	10.56	9.00	5.17	84.08
HCMS (TOMM'23)	10.20	6.62	4.40	83.11	10.22	6.54	4.75	84.58
UAAR-static	6.65	9.00	5.13	**84.64**	6.90	9.00	5.94	**86.82**
UAAR-dynamic	6.65	3.55	2.19	83.35	6.90	3.12	2.77	85.82

Note: Params and GFLOPs are measured using the deployment models after MobileOne structural reparameterization. For dynamic methods, Avg. GFLOPs is averaged over the actually activated inference routes. For AVGN, Avg. Frames counts frames processed by FastViT, while its lightweight glance overhead is included in GFLOPs. HCMS is an audio–RGB adaptation with shared recognition-backbone architectures, not an original-paper result. AdaMML results, including policy-preview costs, are reported in Appendix A.

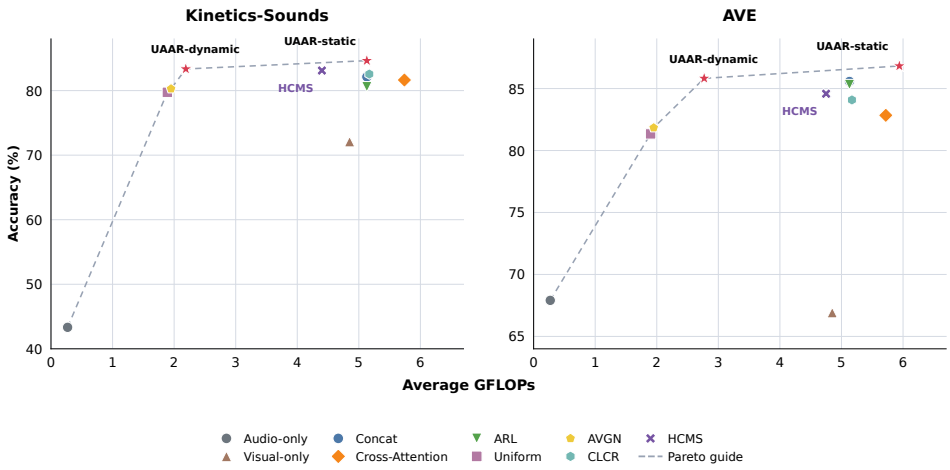

Figure 3: Accuracy versus average GFLOPs on Kinetics-Sounds and AVE. The light dashed polyline connects the empirically non-dominated operating points only as a visual guide.

Table 2: Component ablation on Kinetics-Sounds. The fixed-confidence variant removes selective-risk calibration and adopts a conventional maximum-softmax threshold of 0.9. Accuracy uncertainty is the binomial standard error on the 1,634-sample held-out test set, rather than variation across independent training runs.

Variant	Params (M)	Avg. Frames	Avg. GFLOPs	Acc. (% ± SE)
Full UAAR	6.65	3.553	2.192	83.35 ± 0.92
Fixed Confidence (0.9)	6.65	3.325	2.069	82.99 ± 0.93
w/o Audio Prior	6.65	4.399	2.648	83.90 ± 0.91
w/o AFM (Parallel)	6.39	9.000	5.132	82.99 ± 0.93
w/o Fallback	6.65	3.442	2.132	83.17 ± 0.93
w/o Dynamic Routing	6.65	9.000	5.131	84.64 ± 0.89

Efficiency Contributions (Prior, Routing, and Calibration). Disabling dynamic routing forces nine-frame processing, increasing computation by 134% (to 5.131 GFLOPs) for a 1.29-point accuracy gain. Removing the audio prior both disables zero-frame exits and replaces audio-guided sampling with uniform acquisition, raising cost from 2.192 to 2.648 GFLOPs (20.8%) for a 0.55-point gain. This joint ablation assesses the combined audio-prior policy, rather than isolating the savings from zero-frame acceptance alone.

At nearly matched budgets, UAAR achieves 83.35% at 2.191597 GFLOPs versus 82.93% at 2.188623 GFLOPs for entropy routing, whereas validation-tuned confidence routing matches UAAR (Appendix B). The approximately 0.42-point entropy comparison is a single-run observation, not established statistical superiority. Empirical calibration therefore supplies an explicit held-out accepted-error criterion, without demonstrating an advantage over all conventional threshold tuning.

Fusion Quality and Reliability. At nearly identical nine-frame cost (5.131 vs. 5.132 GFLOPs), full-compute AFM achieves 84.64% accuracy versus 82.99% for parallel fusion, a 1.65-point gain consistent with a benefit from recurrent audio-conditioned fusion. Removing fallback changes clean accuracy from 83.35% to 83.17%; we next examine its behavior under corruption. These single-run comparisons do not establish statistical significance or training-seed robustness.

4.4 Robustness under Audio Corruption

UAAR's entropy gate jointly blocks audio-only acceptance and replaces audio-guided sampling with uniform acquisition. We test this policy by progressively perturbing the normalized audio Mel spectrogram:

$$\widetilde{X}_a = \mathrm{clip}\left(X_a + \sigma\ \mathrm{std}(X_a)\,\xi, 0, 1\right), \quad \xi \sim \mathcal{N}(0, I), \tag{11}$$

where $\sigma \in \{0, 0.05, 0.10, 0.20, 0.40, 0.80\}$. Thresholds are calibrated on clean data, frozen during testing, and results are averaged over three corruption seeds.

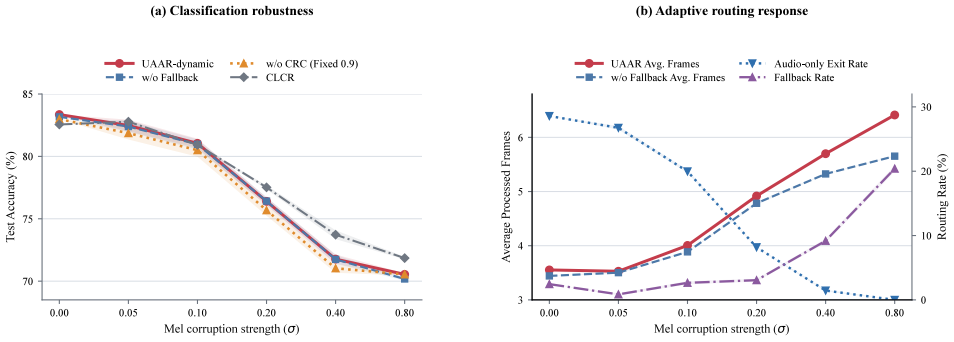

Figure 4: Robustness under progressive Mel-spectrogram corruption. (a) Accuracy of UAAR, its ablations, and CLCR. (b) Adaptive routing response to audio degradation.

Empirically Calibrated Exit. At $\sigma \in [0.05, 0.40]$, the calibrated exit exceeds fixed confidence (0.9) in mean accuracy by up to 0.74 points (Figure 4(a)). This compares frozen operating points, not budget-matched threshold tuning.

Adaptive Computational Scaling. As σ increases to 0.80, audio-only exits fall from 28.6% to 0%, fallback rises from 2.4% to 20.4%, and the visual budget grows from 3.553 to 6.413 frames, showing increased reliance on vision under the tested corruption. Low-entropy incorrect audio can still evade fallback (see Limitations).

Robustness–Efficiency Trade-off. At $\sigma = 0.80$, fallback increases observed accuracy from 70.18% to 70.54%, without isolating its two branches. UAAR uses 3.735 GFLOPs versus CLCR's 5.17 at 71.85% accuracy: 27.8% less computation for a 1.31-point accuracy reduction. These results characterize the trade-off under Gaussian Mel-spectrogram corruption; three corruption seeds do not establish training-seed robustness or statistical significance.

4.5 Hyperparameter Sensitivity and Pareto Efficiency

Sensitivity to the Selective-Risk Tolerance. Figure 5 sweeps h_{safe} after estimating KS audio thresholds on $\mathcal{D}_{cal}$ for $\epsilon \in \{0.01, 0.05, 0.10\}$. This empirical calibration target controls audio acceptance: smaller values send more samples to vision, while larger values permit more zero-frame exits.

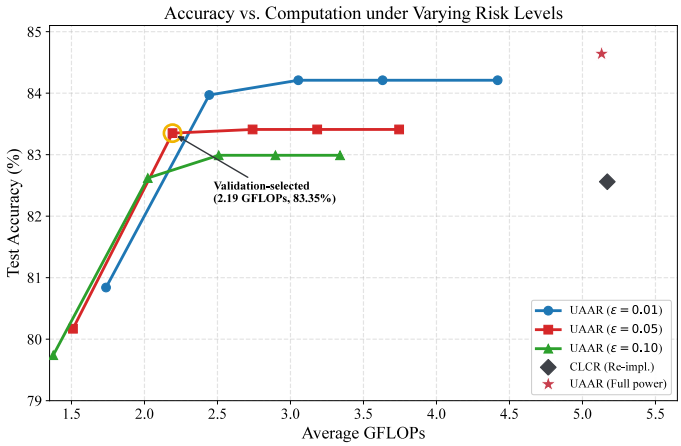

Figure 5: Accuracy–computation trade-offs on Kinetics-Sounds under empirical selective-risk tolerances. Each curve sweeps h_{safe}, and the highlighted point is the reported configuration.

Fusion Threshold & Diminishing Returns. At $\epsilon = 0.05$, a permissive $h_{\text{safe}} = 2.787$ yields 80.17% accuracy, versus 83.35% at 0.755. Tightening it to 0.165 increases the budget from 3.553 to 4.57 frames for only a 0.06-point gain, indicating diminishing returns from additional visual processing in these settings.

Routing Distribution. At the reported operating point, 78.4% of samples avoid full visual processing. Tier-1 accepts 28.58% with 5.78% selective error, versus 34.46% coverage and 7.64% error for fixed confidence (0.9): lower observed accepted error at reduced coverage. Matched-budget threshold comparisons appear in Appendix B.

5 Conclusion

UAAR combines empirically calibrated audio-only acceptance, reliability-gated fallback, and autoregressive visual verification. On Kinetics-Sounds and AVE, it reduces computation by 57.3% and 53.4% relative to full-compute UAAR, with 1.29- and 1.00-point accuracy reductions. These trade-offs support deciding per sample both whether to activate vision and when to stop acquiring evidence.

Limitations. UAAR assumes continuously available, low-cost audio. Empirical calibration does not guarantee test risk under distribution shift, and low-entropy but incorrect audio predictions may evade fallback. Our tests omit reverberation, temporal misalignment, and nonstationary noise; GFLOPs do not capture device-specific latency or energy consumption.

AI use statement

Generative AI tools were used to assist with code optimization and refactoring, and literature search. They also assisted with manuscript polishing, reference formatting, and LaTeX preparation. The authors are responsible for the final content of this work, including any text, code, claims, and artifacts prepared with AI assistance.

Reproducibility statement

The technical appendix documents dataset partitions (Appendix E), model and optimization settings (Appendix G), threshold calibration (Appendix H), computing infrastructure (Appendix I), and evaluation metrics and run accounting (Appendix J). Full corruption and routing results appear in Appendix K. The accompanying anonymous code supplement provides training, calibration, evaluation, and ablation scripts, dependency requirements, and data-layout instructions. Dataset content and pretrained checkpoints are not included in the code archive; their availability and exclusions are described in its README.

References

Anastasios Nikolas Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

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.

Jacob Chalk, Jaesung Huh, Evangelos Kazakos, Andrew Zisserman, and Dima Damen. TIM: A time interval machine for audio-visual action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18153–18163, 2024.

Yanbei Chen, Yongqin Xian, A. Sophia Koepke, Ying Shan, and Zeynep Akata. Distilling audio-visual knowledge by compositional contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7016–7025, 2021.

Ruohan Gao, Tae-Hyun Oh, Kristen Grauman, and Lorenzo Torresani. Listen to look: Action recognition by previewing audio. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10457–10467, 2020.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yonatan Geifman and Ran El-Yaniv. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159, 2019.

Tiantian Geng, Jinrui Zhang, Qingni Wang, Teng Wang, Jinming Duan, and Feng Zheng. LongVALE: Vision-audio-language-event benchmark towards time-aware omni-modal perception of long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18959–18969, 2025.

Amir Ghodrati, Babak Ehteshami Bejnordi, and Amirhossein Habibian. FrameExit: Conditional early exiting for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15608–15618, 2021.

Yizeng Han, Dongchen Han, Zeyu Liu, Yulin Wang, Xuran Pan, Yifan Pu, Chao Deng, Junlan Feng, Shiji Song, and Gao Huang. Dynamic perceiver for efficient visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5992–6002, 2023.

Xun Jiang, Yufan Gu, Disen Hu, Yuqing Hou, Yazhou Yao, Fumin Shen, Heng Tao Shen, and Xing Xu. Multimodal learning on low-quality data with conformal predictive self-calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23041–23050, 2026.

Ji Lin, Chuang Gan, and Song Han. TSM: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, 2019.

Liang Liu, Shuaiyong Li, and Yongqiang Zhu. Audio-visual semantic graph network for audio-visual event localization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23957–23966, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Liwei Luo, Shuaitengyuan Li, Dongwei Ren, Qilong Wang, Pengfei Zhu, and Qinghua Hu. Decoupled multi-predictor optimization for inference-efficient model tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3628–3638, 2025.

Brandon Mckinzie, Vaishaal Shankar, Joseph Yitan Cheng, Yinfei Yang, Jonathon Shlens, and Alexander T. Toshev. Robustness in multimodal learning under train-test modality mismatch. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 24291–24303, 2023.

Chunlei Meng, Guanhong Huang, Rong Fu, Runmin Jian, Zhongxue Gan, and Chun Ouyang. CLCR: Cross-level semantic collaborative representation for multimodal learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1606–1615, 2026.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *Advances in Neural Information Processing Systems*, volume 34, pp. 14200–14213, 2021.

Muhammad Adi Nugroho, Sangmin Woo, Sumin Lee, and Changick Kim. Audio-visual glance network for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10150–10159, 2023.

Rameswar Panda, Chun-Fu Richard Chen, Quanfu Fan, Ximeng Sun, Kate Saenko, Aude Oliva, and Rogerio Feris. AdaMML: Adaptive multi-modal learning for efficient video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7576–7585, 2021.

Konyul Park, Yecheol Kim, Daehun Kim, and Jun Won Choi. Resilient sensor fusion under adverse sensor failures via multi-modal expert fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6720–6729, 2025.

AJ Piergiovanni, Isaac Noble, Dahun Kim, Michael S. Ryoo, Victor Gomes, and Anelia Angelova. Mirasol3B: A multimodal autoregressive model for time-aligned and contextual modalities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26804–26814, 2024.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European Conference on Computer Vision*, pp. 252–268, 2018. doi: 10.1007/978-3-030-01216-8_16.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. FastViT: A fast hybrid vision transformer using structural reparameterization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5785–5795, 2023a.

Pavan Kumar Anasosalu Vasu, James Gabriel, Jeff Zhu, Oncel Tuzel, and Anurag Ranjan. MobileOne: An improved one millisecond mobile backbone. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7907–7917, 2023b.

Yulin Wang, Yang Yue, Yuanze Lin, Haojun Jiang, Zihang Lai, Victor Kulikov, Nikita Orlov, Humphrey Shi, and Gao Huang. AdaFocus V2: End-to-end training of spatial dynamic networks for video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20062–20072, 2022.

Shicai Wei, Chunbo Luo, and Yang Luo. Improving multimodal learning via imbalanced learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2250–2259, 2025.

Zejia Weng, Zuxuan Wu, Hengduo Li, Jingjing Chen, and Yu-Gang Jiang. HCMS: Hierarchical and conditional modality selection for efficient video recognition. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 20(2):35:1–35:18, 2023. doi: 10.1145/3572776.

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.

Zhihui Zhang, Luanyuan Dai, Qika Lin, Yunfeng Diao, Guangyin Jin, Yufei Guo, Jing Zhang, and Xiaoshuai Hao. Synergistic prompting for robust visual recognition with missing modalities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1881–1890, 2025.

A Additional Dynamic Baselines

We evaluate independently trained audio–RGB adaptations of HCMS and AdaMML on the same Kinetics-Sounds partitions. Recognition networks are initialized from the same pretrained MobileOne-s0 and FastViT-t8 checkpoints, not from UAAR's fused checkpoint. These controlled adaptations differ from the original published systems and are not exact reproductions of their reported results.

Both adaptations process three chronological temporal locations (25%, 50%, and 75%), with three RGB frames per selected clip. Audio uses three temporal thirds of the Mel image, each resized to 224×224, rather than UAAR's single global audio pass. HCMS retains hierarchical audio/visual recurrent states and conditional visual execution, omitting the motion branch. Its audio/visual projections have dimensions 512/1024 and recurrent hidden dimensions 128/512. Training combines classification loss with twice the squared deviation of per-video RGB usage from the target budget.

AdaMML retains MobileNetV2 policy encoders, a joint MLP, a 256-unit policy LSTM, and independent audio/RGB selection, while replacing recognition backbones. Each segment uses a 160×160 RGB preview and a three-channel Mel-image policy input. Preview computation is always charged, even when RGB recognition is skipped. Evaluation uses deterministic argmax gates. The correctness-gated, cost-weighted usage penalty is computed per sample; these input and training choices are adaptations of the original implementation.

All six configurations use 40 epochs, training seed 0, AdamW with weight decay 10^{-4}, learning rates 5×10^{-6} for recognition backbones and 5×10^{-5} for new modules, cosine decay, batch size 8 with four-step gradient accumulation, and gradient clipping at 5. After five full-modality warmup epochs, HCMS trains jointly; AdaMML alternates recognition and policy passes, freezing its policy for the final five epochs. HCMS usage targets are $\{0.25, 0.50, 0.75\}$ and AdaMML penalty coefficients are $\{0.05, 0.10, 0.20\}$. Checkpoints are selected by validation accuracy within each configuration; the main table selects configurations by validation accuracy as well, not test accuracy.

Final evaluation disables TF32 and checks sparse execution against dense gated replay on all 2,311 validation and 1,634 test samples. HCMS runs resumed after a numerical parity failure; legacy checkpoints did not retain all random-number and mixed-precision state, so continuation is not bitwise identical to uninterrupted training. GFLOPs use one multiply–accumulate as one operation, include recognition, policy previews, projections, and recurrent matrix products, and do not represent measured latency. Unsupported elementwise operations in the profiler are not exhaustively counted. Avg. Frames counts RGB recognition frames, not AdaMML policy previews. All results are single-training-seed measurements.

Table A1: All predefined dynamic-baseline configurations on Kinetics-Sounds. The main-table HCMS configuration is selected by validation accuracy.

Method	Setting	Val. Acc.	Test Acc.	Avg. Frames	GFLOPs
HCMS	budget 0.25	83.86	84.03	3.099	2.501
HCMS	budget 0.50	84.42	83.35	5.231	3.653
HCMS	budget 0.75	85.24	83.11	6.621	4.405
AdaMML	penalty 0.05	85.94	85.07	8.846	6.973
AdaMML	penalty 0.10	85.76	85.62	9.000	7.070
AdaMML	penalty 0.20	85.55	85.07	9.000	7.127

The validation-selected AdaMML configuration (penalty 0.05) achieves 85.07% at 6.973 GFLOPs. Its higher accuracy than UAAR-dynamic comes with higher computation. The evaluated AdaMML checkpoints retain nearly all visual recognition frames and do not cover UAAR's low-budget operating point; these results do not establish matched-budget dominance. AdaMML Avg. Frames excludes policy previews, whose cost is included in GFLOPs.

B MATCHED-BUDGET AUDIO-EXIT THRESHOLD COMPARISONS

We hold recognition weights and the remaining routing policy fixed. The conventional confidence threshold is selected by enumerating distinct validation scores and minimizing the GFLOPs difference from UAAR, with ties resolved by validation accuracy and then the smaller threshold. The threshold is frozen for testing. This yields the same test operating point as UAAR. The entropy-threshold rerun is approximately budget matched: its computation differs by 0.14% and its accuracy is approximately 0.42 points lower. These single-run observations do not establish statistical significance. Compared with fixed confidence 0.9, UAAR uses 0.123 additional GFLOPs and gains approximately 0.36 points; this is a more conservative operating point, not a matched-budget comparison. Empirical calibration supplies an explicit held-out accepted-error criterion, not a new confidence ranking.

Table A2: Kinetics-Sounds threshold comparisons with fixed recognition weights. The entropy operating point is approximately, not exactly, budget matched.

Audio-exit rule	Avg. GFLOPs	Acc. (%)
Fixed confidence (0.9)	2.068716	82.99
Budget-matched confidence	2.191597	83.35
Near-budget-matched entropy	2.188623	82.93
UAAR empirical calibration	2.191597	83.35

C CLEAN-AUDIO SEMANTIC MISMATCH STRESS TEST

To test a failure mode distinct from additive noise, we fix all 1,634 KS test videos and their labels and replace each audio input with clean audio from a different class. A seeded permutation (seed 2027) uses every test audio exactly once. We randomly order class groups and samples within groups, then cyclically shift the concatenated indices by the largest class size; every resulting pair is checked to have different class labels. Pairing does not depend on predictions or confidence. This preserves the audio marginal distribution while disrupting audio–video correspondence. Model weights and thresholds $(0.0499, 0.755, 3.0)$ remain frozen.

Table A3: Synthetic semantic mismatch, evaluated against the original video labels. The no-zero-exit comparison uses more computation.

Policy	Acc. (%)	Avg. Frames	GFLOPs
UAAR	47.67	4.089	2.481
Zero-frame exits disabled	61.57	5.400	3.188

The audio-branch logits exactly match those of the corresponding clean donor inputs. Zero-frame coverage remains 467/1,634 (28.58%), but all 467 accepted predictions are incorrect relative to the video label. The high-entropy rate remains 2.45%. Thus, audio classification confidence and entropy do not establish cross-modal semantic relevance: clear but unrelated audio can cause an incorrect exit before visual correction is possible. Disabling zero-frame exits improves accuracy but increases cost and does not remove the audio input from fusion or sampling. This single-pairing synthetic stress test is not an estimate of natural-shift performance; it identifies an applicability limit of the reported routing policy.

D TECHNICAL SUPPLEMENT

This supplement provides implementation and evaluation details that are omitted from the main paper because of the page limit. It documents the data partitions, input construction, optimization and calibration settings, computing environment, metric definitions, number of runs, and complete audio-corruption results. The method definition and the evidence needed to assess the central claims remain in the main paper.

E Datasets and Partitioning

We use Kinetics-Sounds (33 classes) and AVE (28 classes) for sample-level audio-visual event classification. All split files contain sample identifiers and class labels; class indices are assigned by sorting the training-set class names. Table A4 reports the exact split sizes used by every method.

Table A4: Dataset partitions used in all experiments.

Dataset	Train	Validation	Calibration	Test
Kinetics-Sounds	18,593	2,311	2,354	1,634
AVE	3,339	201	201	402

Kinetics-Sounds. The official training pool is divided class-wise into training, validation, and calibration subsets using an 80/10/10 allocation and random seed 42. Integer rounding is performed independently within each class, which accounts for the small difference between the two held-out subset sizes. The official test partition is retained as the final test set and is never used for checkpoint or threshold selection.

AVE. We retain the official training set and official test set. The 402-sample official validation set is divided equally into validation and calibration subsets with `random_state=42`. The validation subset selects checkpoints and the fused-exit threshold; the calibration subset estimates the audio-only acceptance threshold. The four partitions are mutually exclusive for both datasets.

F Input Construction and Sampling

Audio. The models consume precomputed 10-second Mel-spectrogram images. During training and evaluation, each spectrogram is converted to RGB, resized to 224×224, and converted to a tensor in $[0, 1]$ before normalization by the ImageNet mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. Kinetics-Sounds additionally applies random erasing during training with probability 0.5, erased area in $[0.02, 0.10]$, and aspect ratio in $[0.1, 10]$. No waveform-level transformation is performed inside the recognition training loop.

Video. Frames are resized to 256×256 and cropped to 224×224. Kinetics-Sounds uses a random crop and horizontal flip for training and a center crop for evaluation; AVE uses the center crop in the reported configuration. The full visual budget contains three clips centered at relative timestamps 0.25, 0.50, and 0.75. Each clip contains three local frames, giving nine physical frames. On AVE, the local frames are located at approximately -0.5, 0, and $+0.5$ seconds around the clip center. On Kinetics-Sounds, they are the adjacent stored frames around the selected center.

Dynamic acquisition. For reliable but audio-insufficient samples, the temporal audio representation ranks candidate timestamps; one-dimensional non-maximum suppression removes adjacent peaks and the top three remaining locations define the visual clips. Clips are decoded only as they are requested. If the audio reliability test triggers, the saliency ranking is disabled and the three uniform locations are used. Consequently, UAAR activates exactly 0, 3, 6, or 9 frames. The Uniform Sampling baseline instead uses the beginning, midpoint, and end frames and therefore always activates three frames.

G Model and Optimization Details

Both datasets use MobileOne-s0 for the audio stream and FastViT-t8 with a Temporal Shift Module for the visual stream. Their output dimensions are 1,024 and 768, respectively. AFM projects both modalities to a 512-dimensional state and shares its DGF projections and classifier across the three recurrent steps. Every exit is trained with cross entropy. The audio and visual auxiliary losses have unit weight, while the three fused losses use $\lambda = (0.5, 0.8, 1.0)$.

Table A5: Optimization settings used across the compared models. LR denotes the initial learning rate.

Setting	Kinetics-Sounds	AVE
Epoch budget	40	40
Batch size	8	16
Gradient accumulation	4	1
Effective batch size	32	16
AFM/head LR	5×10^{-5}	1×10^{-4}
Backbone LR	5×10^{-6}	1×10^{-5}
Optimizer	AdamW	AdamW
Weight decay	10^{-2}	10^{-2}
LR schedule	cosine	cosine
Data-loader workers	8	4
Reproduction seed	0	3407

The sensory backbones are initialized from the corresponding unimodal checkpoints. We select the checkpoint with the highest validation accuracy, freeze all learned parameters, and only then select routing thresholds. Concat, Cross-Attention, ARL, and CLCR retain the same sensory backbones, spatial resolution, and nine-frame evaluation protocol. All compared models follow the dataset-specific settings in Table A5: AdamW optimization for at most 40 epochs, weight decay 10^{-2}, cosine learning-rate decay, and validation-based early stopping. Trainable fusion and classification layers use the listed AFM/head rate, while pretrained sensory backbones use one tenth of that rate. All comparisons retain the validation-selected checkpoint and evaluate the held-out test set once. For the adapted AVGN models, the best-validation checkpoints occur at epoch 2; subsequent epochs do not improve validation accuracy, and training terminates under the same validation-based early-stopping protocol.

H Threshold Calibration and Inference

Threshold selection never accesses test labels. Let $s = 1 - \max_c p_{a,c}$ be the audio nonconformity score. For each candidate τ on the calibration split, we compute the finite-sample-corrected selective risk defined in the main paper and retain the largest threshold whose corrected risk is at most ϵ. We evaluate $\epsilon \in \{0.01, 0.05, 0.10\}$ and use $\epsilon = 0.05$ for the principal model. The fused threshold h_{safe} is selected independently on the validation split by sweeping prediction-entropy cutoffs and choosing the validation accuracy–computation operating point. The reliability threshold is the 99th percentile of clean validation audio entropies; the deployed value is rounded to 3.0. Table A6 lists the frozen inference settings.

Table A6: Deployed routing thresholds. None is recalibrated on the test set or on corrupted inputs.

Setting	Kinetics-Sounds	AVE
Selective-risk tolerance ϵ	0.05	0.05
Audio threshold $\hat{\tau}$	0.0499	0.0240
Reliability threshold h_{deaf}	3.000	3.000
Fused threshold h_{safe}	0.755	0.800
Maximum clips / frames	3 / 9	3 / 9

At test time, the reliability test precedes audio-only acceptance. A sample with $h_{\mathrm{audio}} > h_{\mathrm{deaf}}$ uses the uniform visual fallback. Otherwise, $s \le \hat{\tau}$ produces the zero-frame exit. All remaining samples acquire clips sequentially and halt after step $t \in \{1, 2\}$ when $h_{\mathrm{fusion},t} < h_{\mathrm{safe}}$; samples that remain unresolved return the step-3 prediction.

I Computing Infrastructure

Experiments were executed on a shared Ubuntu 18.04.6 LTS server with two Intel Xeon Silver 4314 processors (64 logical CPUs), 251 GiB of system memory, and nine NVIDIA GeForce RTX 4090 GPUs (24,564 MiB each). Each reported training or evaluation process used one selected GPU. The NVIDIA driver was 535.86.05. The software environment used Python 3.11.13, PyTorch 2.5.1+cu121, torchvision 0.20.1+cu121, timm 1.0.25, NumPy 2.3.5, Pillow 11.3.0, OpenCV 4.13.0, and fvcore 0.1.5.post20221221. PyTorch was built against CUDA 12.1.

J Evaluation Metrics and Run Accounting

For a test set of size N, accuracy is $N^{-1}\sum_i \mathbb{I}(\hat{y}_i = y_i)$. Where the ablation table reports standard error (SE), it is the binomial estimate

$$\mathrm{SE}(\hat{p}) = 100\sqrt{\frac{\hat{p}(1-\hat{p})}{N}} \tag{A1}$$

in percentage points; it is not the standard deviation across independent training runs. The principal comparison and ablation entries each use one training run per independently trained model, followed by one evaluation of the validation-selected checkpoint. Inference-only ablations reuse the same trained UAAR checkpoint and change only the routing rule.

Parameter counts are measured after MobileOne structural reparameterization and include every module required by the deployed method. FLOPs are profiled with fvcore using one multiply–accumulate as one FLOP. For a dynamic method, we sum the audio, visual, and fusion operations actually activated for each test sample and average over the test set. If $K_i \in \{0, 1, 2, 3\}$ is the number of acquired clips,

$$\mathrm{AvgFrames} = \frac{3}{N}\sum_{i=1}^{N} K_i, \tag{A2}$$

$$\mathrm{AvgGFLOPs} = \frac{1}{N}\sum_{i=1}^{N} \left(\mathcal{C}_A + K_i(\mathcal{C}_V + \mathcal{C}_F)\right). \tag{A3}$$

The Kinetics-Sounds UAAR profile uses $\mathcal{C}_A = 0.27417$ GFLOPs, $\mathcal{C}_V = 1.61712$ GFLOPs per three-frame clip, and $\mathcal{C}_F = 0.00173$ GFLOPs per fusion update, with a 0.00052-GFLOP memory-initialization cost included once when vision is activated.

Table A7: Kinetics-Sounds accuracy under Mel-spectrogram corruption. Values are mean ± standard deviation across three corruption seeds; the deterministic clean point is shown without a deviation.

Method	$\sigma = 0$	0.05	0.10	0.20	0.40	0.80
UAAR-dynamic	83.35	82.48±0.47	81.05±0.37	76.40±0.31	71.77±0.37	70.54±0.09
w/o Fallback	83.17	82.39±0.43	80.97±0.40	76.42±0.31	71.75±0.35	70.18±0.13
Fixed Confidence (0.9)	82.99	81.86±0.51	80.50±0.52	75.68±0.43	71.03±0.36	70.54±0.09
CLCR	82.56	82.78±0.15	80.93±0.04	77.52±0.15	73.73±0.35	71.85±0.11

K Additional Experimental Results

K.1 Complete Audio-Corruption Results

Gaussian noise is applied to the $[0, 1]$-scaled Mel-spectrogram tensor before ImageNet normalization as $\widetilde{X}_a = \mathrm{clip}(X_a + \sigma\ \mathrm{std}(X_a)\xi, 0, 1)$. For every nonzero severity, we use corruption seeds 17, 29, and 43 and report the population mean and standard deviation across the three corrupted test sets. The clean point is deterministic. All thresholds remain fixed at their clean-data values.

Table A8: Routing response of UAAR-dynamic under the same corruption protocol. Values are mean ± standard deviation across seeds.

Metric	$\sigma = 0$	0.05	0.10	0.20	0.40	0.80
Avg. Frames	3.553	3.528±0.023	4.005±0.014	4.917±0.033	5.698±0.015	6.413±0.022
Avg. GFLOPs	2.192	2.178±0.013	2.436±0.007	2.928±0.018	3.350±0.008	3.735±0.012
Audio-only Exit (%)	28.58	26.74±0.28	19.97±0.23	8.18±0.07	1.45±0.14	0.00
Fallback (%)	2.45	0.86±0.16	2.65±0.25	3.06±0.21	9.18±0.32	20.40±0.46

K.2 Extended Routing Analysis

The complete results clarify how risk-calibrated routing responds to progressive acoustic degradation. From $\sigma = 0$ to 0.80, the audio-only exit rate decreases from 28.58% to zero, while the fallback rate rises from 2.45% to 20.40%. The average visual budget therefore increases from 3.553 to 6.413 frames and from 2.192 to 3.735 GFLOPs. UAAR does not preserve a rigid computation budget; it allocates additional visual evidence when the clean-audio routing prior becomes unreliable.

The calibrated exit also improves over fixed confidence throughout the moderate corruption range, with gains of 0.62, 0.55, 0.72, and 0.74 percentage points at $\sigma \in \{0.05, 0.10, 0.20, 0.40\}$, respectively. At $\sigma = 0.80$, the fallback recovers 0.36 points over the variant without fallback. CLCR retains the highest accuracy at the strongest corruptions, but always executes its full 5.17-GFLOP route; UAAR instead exposes an input-dependent robustness–efficiency trade-off.