CONTROLLED ACQUISITION AND ABSTENTION IN THREE-CHANNEL SCORE CONFLICTS

Anonymous authors
Paper under double-blind review

ABSTRACT

When audio, video, and text disagree, prediction accuracy alone does not reveal whether another source is worth acquiring or whether the system should withhold an answer. We separate these decisions in a controlled benchmark: a policy sees two signed scores, can request a third at a cost, and may abstain. The primary evaluation uses a mechanism-specific reward: abstention is correct only for one designated ambiguity mechanism and is scored as incorrect on mixed corruption. We call accuracy under this target targeted decision accuracy. Matched controls show that the threshold policy makes the same final decisions as an always-request policy while issuing fewer requests. Its advantage over always-answer fusion depends on rewarding abstention for one specified ambiguity pattern. Under the mechanism-conditioned target, the threshold policy reaches 0.789 ± 0.006 targeted decision accuracy and 0.481 ± 0.014 utility on the partially held-out synthetic split across 83 seeds. A zero-request majority reference that receives all three scores reaches 0.626 ± 0.008 and 0.252 ± 0.016; these results use different information budgets. A matched-budget test uses one terminal rule that allows the requested score to change the action. The train-only value selector improves utility over no-query and matched random at 10% and 25% budgets, but pair uncertainty has higher utility at every budget and the learned selector reduces utility at 50% and 63.7%. At those higher budgets, the learned selector has slightly higher non-ambiguous binary accuracy, so the utility result is not an unqualified accuracy ranking. Under the all-abstentions-error target, Threshold utility falls below majority, reversing their ordering. At the central temporal setting, full-trace controls match the neural models, while position perturbations separate them. On voice and face emotion clips from held-out actors, eight-frame fusion has opposite signed accuracy differences across two frozen encoder pairs; both actor intervals include zero. Matched-request routing gains are small and uncertain. Together, these results distinguish full-modality accuracy from pre-request selection value and show that the latter depends on the request budget and the observed-pair ranking.

1 INTRODUCTION

Audio, video, and text provide complementary evidence, but their combination becomes difficult when one source is corrupted or when the available sources support incompatible explanations. A fusion model that always emits a label hides two operational decisions. It does not say whether a designated missing source is worth acquiring, and it does not say when the observations fail to identify a safe answer. These omissions matter because acquisition, incorrect answers, delays, and abstention can carry different application-dependent costs.

Recent multimodal conflict benchmarks make contradiction detection measurable through controlled perturbations and hierarchical tasks (Tian et al., 2026). Work on modality interference and credibility-aware fusion shows that source reliability varies across examples (Cai et al., 2025; Tenali et al., 2026). Omnimodal benchmarks also expose counterfactual conflicts across several channels (Mukherjee et al., 2026). Yet a high aggregate accuracy does not specify the value of a missing observation or whether a conflict is identifiable from observations alone. We therefore ask:

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.

When is a third source worth acquiring, when should a conflict trigger abstention, and can its value be predicted before a request?

We study three signed scores representing audio, video, and text. A latent binary event generates the scores before corruption modifies one or more sources. A policy observes two scores, requests the third on disagreement or weakness, and answers when two strong scores agree; otherwise it abstains. Policies never receive mechanism labels. Query and non-ambiguous abstention costs make these decisions measurable. Full-modality value denotes the held-out accuracy change with all modalities observed; pre-request value denotes the gain from selecting requests using only previously observed information. Training and evaluation use separately generated samples, with partly held-out corruption mechanisms.

1.1 Contributions

First, we define a three-score protocol that separates acquisition from the final answer or abstention decision (Figure 1). Matched controls identify when request savings leave final actions unchanged. Second, a common-rule, matched-budget test compares a train-only value selector with pair uncertainty, matched random, and a non-deployable oracle; it shows where a learned ranking helps and where it harms utility. Equal-information temporal controls identify which simulator findings survive matched representations. Third, actor-disjoint media studies separate full-fusion accuracy from selected-request value and show that the results depend on the frozen encoder pair. The evidence supports a controlled decision diagnostic, not a generally superior acquisition policy.

The reproduction code and configurations are available from the anonymous 4open.science Git repository.[1]

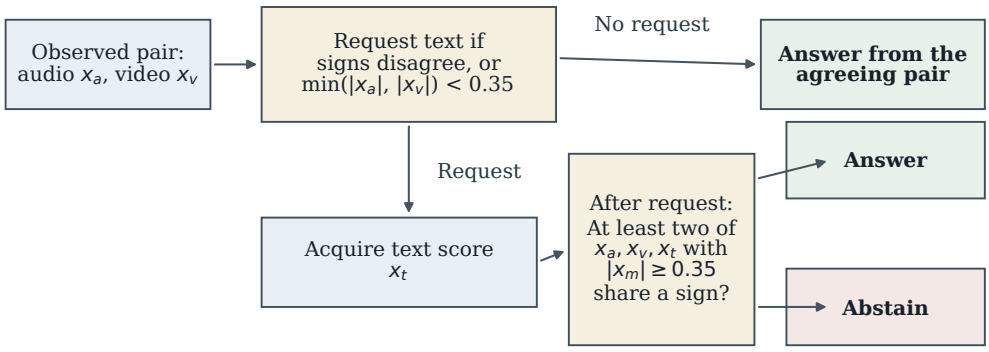


Figure 1: The fixed decision rule. The policy requests text when the audio and video signs disagree or either observed score has magnitude below 0.35. Without a request, the two strong, agreeing scores determine the answer. After acquisition, the policy answers only when at least two scores have magnitude at least 0.35 and share a sign; otherwise it abstains. Utility separately charges for a request and for non-ambiguous abstention.

2 Related Work

Multimodal learning has traditionally focused on learning a shared representation or combining modality-specific representations (Baltrušaitis et al., 2019). In conflict settings, this framing is incomplete because the system must reason about whether a source is reliable for the current example. CrossCheck-Bench studies hierarchical image–text contradiction detection in real-world media with controlled perturbations (Tian et al., 2026). AV-ConfuseBench tests audiovisual confusion created by muting visually present objects (Ye et al., 2026).

[1]Anonymous repository snapshot (direct ZIP download): `https://anonymous.4open.science/api/repo/audio-video-text-conflicts-9C15/zip`.

EmoMM steers attention over supplied audio, video, and text under emotion conflict and missingness (Sun et al., 2026).

Our study instead prices a request for a missing score and permits abstention. EmoMM evaluates media models; our primary evidence uses generated scalar scores, with bounded frozen-feature media diagnostics. These tasks and evidence differ.

Modality-interference work diagnoses misleading channels through perturbations and consistency objectives (Cai et al., 2025). Credibility-aware fusion estimates instance-level source trust in context (Tenali et al., 2026). Source weighting addresses answer formation; acquisition and rejection require additional decisions.

C^3PO studies counterfactual conflict across audio, image, video, and text (Mukherjee et al., 2026). We take a narrower experimental route to make action semantics auditable: a three-score generator, explicit mechanisms, and a partially held-out split. This choice sacrifices media realism in exchange for a clear test of whether a policy uses observations alone and whether its errors arise from mandatory answering, unnecessary requests, or non-identifiability.

Selective prediction studies the tradeoff between coverage and predictive risk; SelectiveNet integrates a reject option into a learned model (Geifman & El-Yaniv, 2019). Active feature acquisition studies purchasing missing information, including generative-surrogate policies (Li & Oliva, 2021) and cost-sensitive decision making (Li & Oliva, 2025). Probabilistic personalised cascades combine sequential feature acquisition with a rejector that weighs the cost of unobserved features; rejection can trigger another acquisition stage (Shpakova & Sokolovska, 2021). Active Acquisition for Multimodal Temporal Data (A2MT) selects modalities through time under acquisition costs and shows that non-adaptive controls can match or outperform learned agents (Kossen et al., 2023). Acquisition Conditioned Oracle provides a general non-greedy formulation based on expected downstream loss reduction (Valancius et al., 2024). GuessBench studies iterative visual question asking for active reasoning in multimodal large language models (Liu et al., 2026). Our setting fixes two observed scores and one designated unavailable score, then records the request and terminal abstention separately. Its distinct evidence is the matched action and reward-convention audit of score conflict, not a general acquisition method. Unlike standard selective risk, our synthetic targeted decision accuracy rewards abstention on a designated ambiguity mechanism; the target convention must accompany any comparison.

3 Problem Formulation

Each example has a binary latent event $y \in \{0, 1\}$ and an observation vector $x = (x_a, x_v, x_t) \in \mathbb{R}^3$ for audio, video, and text. A clean score is

$$x_m = (2y - 1)s + \epsilon_m, \qquad \epsilon_m \sim \mathcal{N}(0, \sigma^2), \tag{1}$$

where $s = 1$ and $\sigma = 0.22$ in the campaign, and write $\mu_y = (2y - 1)s$. Labels and mechanisms are sampled uniformly and independently. For uniformly drawn coordinates j and $k \neq j$, invert and swap both set $x_j = -\mu_y + \epsilon_j$; dropout sets $x_j \sim \mathcal{N}(0, \max(1.8\sigma, 0.05)^2)$; mixed sets $x_j = -\mu_y + \epsilon_j$ and $x_k \sim \mathcal{N}(0, \max(1.8\sigma, 0.05)^2)$; and burst scales the clean vector by 0.7 before setting $x_j = -2.5\mu_y + \epsilon_j$. Ambiguity randomly assigns independent draws from $\mathcal{N}(\mu_y, \sigma^2)$, $\mathcal{N}(-\mu_y, \sigma^2)$, and $\mathcal{N}(0, \max(0.6\sigma, 0.04)^2)$ to the three coordinates. The protocol retains seven mechanism labels but only six distinct score-level transformations because invert and swap use the same operation. The training and seen labels are clean, invert, swap, and dropout. The partially held-out split includes independently sampled clean cases and introduces mixed, burst, and ambiguity.

Ambiguity receives an abstention target. This is a mechanism-specific reward convention, not a general definition of non-identifiability. Mixed and ambiguity both contain positive, negative, and zero-centered scores in random positions. Their positive and negative score distributions have equal variance, so swapping the latent label leaves the observation distribution unchanged: $p(x \mid y = 0, z) = p(x \mid y = 1, z)$ for both mechanisms. With balanced label priors, neither permits observation-only binary accuracy above 50% in expectation, even with all three sources. Nevertheless, the benchmark rewards rejection only on ambiguity and charges it on mixed. Their zero-centered variances differ, creating a statistical cue for this reward preference. Results therefore measure adaptation to that preference, not universal detection of non-identifiability.

For action $a \in \{0, 1, 2\}$, where 2 denotes abstention, correctness is measured against y on non-ambiguous examples and against 2 on ambiguity examples. The reward also depends on whether the policy requests a third score, indicated by $q \in \{0, 1\}$. We define utility as

$$u(a, q) = 2\mathbb{1}[a = y^\star] - 1 - \lambda_q q \\ - \lambda_a \mathbb{1}[a = 2,\ z \neq \text{ambiguity}], \tag{2}$$

where $y^\star = 2$ for ambiguity and y otherwise, $\lambda_q = 0.10$, and $\lambda_a = 0.20$. Thus a correct action receives $+1$ before costs and an incorrect action receives -1; with the campaign costs, per-example utility lies in $[-1.30, 1.00]$. We call correctness under this mechanism-conditioned target targeted decision accuracy. We report targeted decision accuracy, coverage, selective accuracy (accuracy conditional on a non-abstaining action), query rate, abstention rate, and mean utility. Means and standard deviations are computed over independent seeds, not over individual examples pooled across seeds.

4 Protocol and Methods

4.1 Generator and splits

Each scalar run uses 6,000 training examples and 4,000 examples per evaluation split. Training and seen splits sample clean, invert, swap, and dropout labels with independent seeds. The partially held-out split samples new clean cases, mixed corruption, bursts, and ambiguity; only clean appears in training. Invert and swap implement the same score transformation, giving it twice the sampling and mechanism-macro weight.

4.2 Fixed and learned baselines

Majority predicts the sign supported by at least two scores. Median predicts the coordinate-median sign. Weighted predicts the sign of $\sum_m x_m |x_m|$. These policies receive all three scores initially, always answer, and never request a source. The multilayer perceptron (MLP) maps $[x, |x|]$ through $6 \to 32 \to 16 \to 3$ with ReLU hidden units. Gated uses a tanh-hidden $6 \to 24 \to 3$ softmax gate and a ReLU $2 \to 16 \to 3$ head over weighted mean and score spread. Both learned fusion models also receive all three scores initially and never request a source. These are full-information references, not matched acquisition baselines. Ambiguity is an evaluation-only third class, so neither learned model is fitted as an ambiguity detector. Learned acquisition maps the first two scores and their magnitudes through $4 \to 24 \to 12 \to 1$ with ReLU hidden units. Its binary target is whether a third-source request improves training utility after query cost. This target is privileged offline supervision: it is computed from the training target (including the ambiguity target) and the full three-score vector by comparing the no-request pair action with the post-request strong-score action. The scorer itself receives only the initially observed scores and magnitudes at fitting and inference, and its threshold is selected by utility on a separate train-only calibration split. Without a request it answers only when the initial signs agree, otherwise abstaining, and it applies no magnitude threshold. All scalar networks use Adam for 90 epochs, batch size 256, learning rate 0.0015, and gradient-norm clipping at 5; classifiers use cross-entropy and the query scorer uses binary cross-entropy. The scorer is fitted on the four training mechanisms, then selected by utility on 4,000 independent train-only cases over thresholds $\{0.1, 0.2, \ldots, 0.9\}$. It receives no mechanism, evaluation label, or unavailable text score; post-query decisions use the same strong-score rule as Threshold.

4.3 Threshold query-and-abstain policy

Threshold starts with the audio and video coordinates, which stand for the currently available pair. It requests the third coordinate when the two signs disagree or when either absolute score is below $\tau = 0.35$. This threshold was fixed in the original protocol before held-out evaluation; it was not selected on the partially held-out split or calibrated for optimal utility. Without a request, the policy answers using the first sign. After a request, it answers only when at least two coordinates have magnitude at least τ and share a sign; otherwise it abstains. This rule uses x only. In particular, it does not inspect z, the ambiguity indicator, or the ground-truth label.

4.4 MATCHED PROTOCOL CONTROLS

All controls start with the same two scores and pay λ_q for requesting text. Pair-selective never requests text and abstains whenever the initial pair fails the active rule's agreement/strength test. Query-majority shares the active query schedule but always answers using three-score majority after a request. Always-query-selective always acquires text and then applies the same two-strong-agree rule as THRESHOLD.

The last control has identical final actions to THRESHOLD for every input when $\tau > 0$. If the active rule requests text, both apply the same decision rule. If it does not, the initial pair already contains two strong agreeing scores, which a single third score cannot overturn. Thus their accuracies and coverage are identical, and their mean utility difference is exactly $\lambda_q(1-\mathbb{E}[q])$. This identity isolates acquisition savings from the advantage of permitting abstention.

4.5 MATCHED-BUDGET SELECTOR TEST

To test whether a request can improve the terminal action, we apply a shared sum rule before and after acquisition. For an observed set S, it answers with the sign of $\sum_{m\in S} x_m$ when the magnitude is at least $\tau = 0.35$, and abstains otherwise. Unlike THRESHOLD, the action can change when the third score is added. At each seed, a standardized ridge model ($\alpha = 1$) uses seven features of the initial pair and a training target equal to the pre-request-cost reward change between applying this rule to the full triple and applying it to the pair. Labels and the third score define this target only on the 6,000 training examples; the selector receives the initial pair at evaluation. We compare its ranking with pair uncertainty (lowest absolute pair sum first), a seed-specific random ranking, and an oracle that uses evaluation labels and the unavailable score. The oracle is an upper bound, not a deployable selector. Every method requests the exact top k of 4,000 held-out examples at budgets of 10%, 25%, 50%, and 63.7%, with the same query and abstention costs. Utility differences use 10,000 paired-seed percentile resamples with seed 92309.

5 EXPERIMENTAL SETUP

The bounded scalar campaign uses 19 one-GPU seeds and 64 disjoint CPU seeds, for 83 verified runs. Separately, the temporal factorial has nine cells formed by sequence lengths 32, 64, and 128 crossed with noise levels 0.12, 0.22, and 0.40; each cell has 19 GPU seeds and four CPU seeds, for 207 verified records. Sequence models receive the full $[3, T]$ trace, while scalar controls receive its mean over time. The temporal models use AdamW, width 96, depth 4, batch size 256, and 800 optimizer steps. The temporal evaluation additionally samples contiguous sign inversion, noise occlusion, and channel-local drift. Its seven equally likely mechanisms give an expected ambiguity fraction of one seventh, versus one quarter in the scalar evaluation; their accuracy levels are not directly comparable. Their three-class heads receive no ambiguity examples in training, so the abstention logit is unsupervised; neither temporal model abstained in the verified evaluation records. No mechanism labels or ambiguity flags are provided as inputs.

The long signal-to-noise ratio (SNR) expansion is incomplete and excluded from this paper. A bounded equal-information follow-up instead fixes $T = 64$ and evaluates central $(s, \sigma) = (1.0, 0.22)$ and hard $(0.55, 0.55)$ settings with five GPU seeds each. It compares the existing temporal fusion and gated networks with a train-only waveform template and a ridge classifier over the same full traces. It also applies a shared time permutation, independent channel permutations, and per-example circular shifts. This study tests whether the initial separation requires the fixed waveform or temporal order.

We use the same utility costs across splits and never tune the active threshold on held-out mechanisms. Acquisition policies start with two coordinates. Full-information scalar references receive three, pay no request cost, and answer every evaluated example.

The seen split tests the training mechanisms. The partially held-out split retains clean cases and adds mixed corruption, high-confidence bursts, and symmetric ambiguity. Comparisons are descriptive when methods use different information or terminal-action rules. We assess answer quality together with coverage, requests, abstention, and utility.

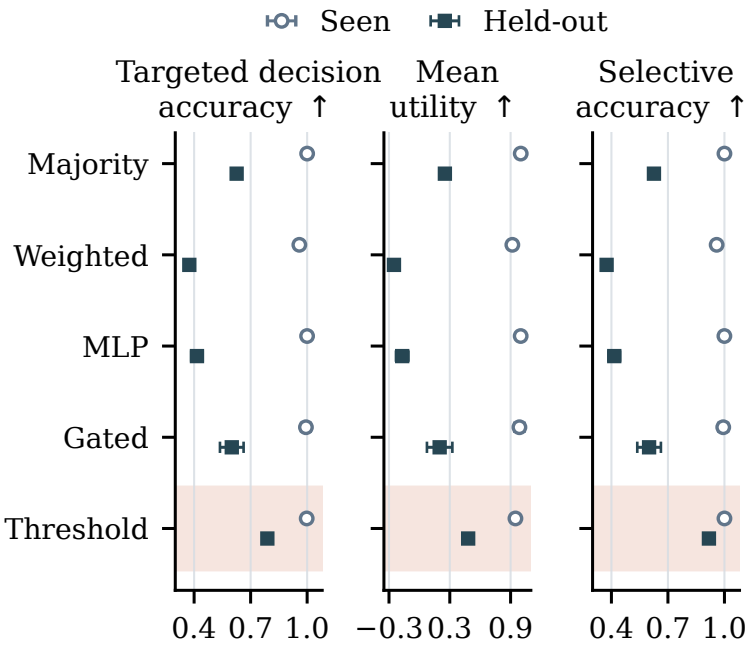


Figure 2: Mean and standard deviation across 83 verified seeds. The partially held-out split compares full-information fusion references with policies that start from two scores and may request a third. Targeted decision accuracy uses the mechanism-conditioned target defined in the text. Utility includes a 0.10 query cost and a 0.20 cost for non-ambiguous abstention.

The studies use different analysis units: synthetic seeds, temporal settings, test videos, held-out actors, training-only source checks, and meeting teams. The six actor allocations overlap; the two AMI replays score identical evaluation records under different fitted protocols. Appendix A gives the counts and denominators before the detailed media results.

6 Results

Table 12 gives the direct acquisition comparison. The threshold rule and always-query-selective make identical final decisions, so their utility difference is exactly the price of requests the threshold avoids. Figure 2 and Table 1 show the broader synthetic results. On the partially held-out split, the threshold rule reaches 0.789 ± 0.006 targeted decision accuracy and 0.481 ± 0.014 utility. The zero-request majority reference receives all three scores and reaches 0.626 ± 0.008 and 0.252 ± 0.016. This is a full-information reference comparison, not a matched acquisition-policy ranking. Table 1 places the mechanism-conditioned metric beside non-ambiguous binary accuracy, which counts every abstention as an error; on this measure, threshold reaches 0.721 ± 0.008 versus 0.834 ± 0.007 for majority. The learned classifiers receive no ambiguity examples during fitting and serve as additional full-information references.

6.1 Mechanism-level behavior

The mechanism breakdown explains the aggregate. On novel ambiguity cases, the active policy abstains at 0.992, matching the designated rejection target. On burst cases, it retains 0.977 accuracy and queries on 0.674 of examples. On mixed cases, mean accuracy is 0.187, abstention is 0.627, and utility is -0.839. Mixed examples are also label-nonidentifiable, as shown in Section 3; the low utility reflects a reward that still requires a binary answer. It is not evidence that the policy failed to discover an identifiable trusted source.

Table 1: Partially held-out results over 83 seeds. Parentheses give sample standard deviations; rates are seed means with dispersion omitted. Mixed, burst, and ambiguity mechanisms are held out. Targeted accuracy rewards abstention only for designated ambiguity; non-ambiguous binary accuracy excludes ambiguity and counts abstentions as errors. Utility follows Equation 2. Majority, median, weighted, MLP, and gated start with all three scores. Learned acquisition uses pair-only inputs, a full-score training target, and a different no-request action.

Method	Targeted acc. ↑	Non-ambig. binary acc. ↑	Utility ↑	Coverage	Query ↓	Abstain
Majority	0.626 (0.008)	0.834 (0.007)	0.252 (0.016)	1.000	0.000	0.000
Median	0.626 (0.008)	0.834 (0.007)	0.252 (0.016)	1.000	0.000	0.000
Weighted	0.375 (0.007)	0.499 (0.008)	−0.251 (0.014)	1.000	0.000	0.000
MLP	0.415 (0.031)	0.553 (0.040)	−0.170 (0.061)	1.000	0.000	0.000
Gated	0.600 (0.063)	0.800 (0.083)	0.200 (0.125)	1.000	0.000	0.000
Learned acquisition	0.746 (0.010)	0.756 (0.009)	0.417 (0.019)	0.709	0.519	0.291
THRESHOLD	0.789 (0.006)	0.721 (0.008)	0.481 (0.014)	0.590	0.637	0.410

Table 1 makes the action tradeoff explicit. The threshold rule covers 0.590 of cases and abstains on 0.410. Its accuracy conditional on an answer is 0.918 ± 0.005, while majority and median answer every case. A deterministic 10,000-resample paired-seed percentile bootstrap (seed 20270918) gives threshold-minus-majority 95% intervals of 0.163 [0.161, 0.165] for targeted decision accuracy and 0.230 [0.226, 0.233] for utility.

Excluding shared-clean cases, targeted decision accuracy is 0.718 [0.717, 0.720] for the threshold rule, 0.649 [0.647, 0.651] for learned acquisition, and 0.501 [0.499, 0.503] for majority; these are 95% paired-seed intervals. The corresponding mechanism-label macro values are 0.719, 0.649, and 0.501. This macro average weights mixed, burst, and ambiguity equally.

The reward convention changes the interpretation. In the mixed-reward variant, both mixed and ambiguity cases receive the abstention target, and the abstention cost is waived for both; the threshold rule then has 0.826 ± 0.007 utility versus 0.252 ± 0.016 for majority. In the all-abstentions-error variant, every example targets its binary label and every abstention incurs λ_a; utility is -0.064 ± 0.019 versus 0.252 ± 0.016. Thus the utility ordering is a result about the declared target and cost masks, not a general selective-risk advantage.

6.2 TEMPORAL SIMULATOR CHECK

The temporal extension replaces each scalar observation with a zero-mean two-harmonic trace. Its fixed sign-bearing waveform gives sequence models information absent from time means. The 207-record factorial is a simulator check, not a representation ranking. At length 64 and noise 0.22 on the partially held-out split (including new clean cases), temporal gated reaches 0.787 ± 0.006 targeted decision accuracy and 0.574 ± 0.011 utility; the time-mean MLP reaches 0.434 and time-mean majority reaches 0.417. Across all nine length–noise cells on this split, the cell means range from 0.786 to 0.788 for temporal gated and 0.779 to 0.787 for temporal fusion, versus 0.429 to 0.438 for the time-mean MLP and 0.407 to 0.420 for time-mean majority. These are descriptive ranges across cells, not uncertainty intervals or comparisons with an established chance level. Neither temporal network learned abstention because ambiguity is absent during training.

The equal-information follow-up changes the interpretation: Table 2 shows full-trace template and ridge matching the networks at the central setting. Both controls reach 0.786 targeted decision accuracy and 0.572 utility on the partially held-out split. A shared train–evaluation permutation preserves this result.

Independent channel permutations reduce template and ridge targeted decision accuracy to 0.428 ± 0.008 and 0.428 ± 0.015 while fusion retains 0.768 ± 0.009; per-example shifts give 0.424, 0.430, and 0.781, respectively (Appendix E). The hard identity setting barely changes utility, so it supplies no discriminating difficulty gradient. The fixed waveform explains the central and shared-permutation ranking, but the neural advantage under position perturbations remains a simulator-specific finding.

Table 2: Equal-information temporal check at $T = 64$, five seeds per setting. All methods receive full traces. Central and shared-permutation settings preserve the positional template; hard identity weakens signal and raises noise. Values are partially held-out targeted decision accuracy or utility, mean (sample standard deviation).

Method	Central targeted decision acc. ↑	Central util. ↑	Shared util. ↑	Hard util. ↑
Temporal fusion	0.780 (0.011)	0.560 (0.021)	0.571 (0.013)	0.570 (0.016)
Temporal gated	0.786 (0.007)	0.571 (0.015)	0.567 (0.018)	0.572 (0.018)
Full-trace ridge	0.786 (0.006)	0.572 (0.012)	0.572 (0.012)	0.572 (0.014)
Train-only template	0.786 (0.009)	0.572 (0.018)	0.572 (0.018)	0.574 (0.013)

6.3 Whole-policy learned acquisition diagnostic

The query scorer receives the initially available audio and video scores and their magnitudes at decision time; changing the unavailable text score leaves these inputs unchanged. Its training target is privileged: it is computed using the full three-score vector and the mechanism-conditioned utility. The 83-run rerun is separate from the earlier invalid aggregate. On seen mechanisms, learned acquisition reaches 0.998 ± 0.001 targeted decision accuracy and 0.954 ± 0.004 utility while querying on 0.425 ± 0.034 of examples. The threshold policy reaches the same targeted decision accuracy, 0.948 ± 0.002 utility, and a query rate of 0.471 ± 0.008. On the partially held-out split, the scorer queries less often (0.519 ± 0.030 versus 0.637 ± 0.007) but reaches only 0.746 ± 0.010 targeted decision accuracy and 0.417 ± 0.019 utility, compared with 0.789 ± 0.006 and 0.481 ± 0.014 for the threshold policy. Its no-request branch also changes the terminal answer rule, so this whole-policy comparison cannot isolate the value of its query schedule.

6.4 Cost sensitivity and information budget

Recomputing the 83 verified records with fixed actions gives active utility 0.481 ± 0.014 at campaign costs $(\lambda_q, \lambda_a) = (0.10, 0.20)$ and 0.192 ± 0.018 at $(0.40, 0.80)$, below majority at 0.252 ± 0.016 (Appendix D). The policy is not retuned across costs, and the sweep does not measure request latency or media preprocessing.

6.5 Matched controls and threshold sensitivity

Across five seeds (4,000 evaluation examples each), always-query-selective makes the same decisions as the active rule. Their utility gap, 0.03656, equals avoided-query cost $0.10(1 - 0.6344)$; query-majority has the same request rate but lower utility (Appendix B, Table 12).

Excluding shared-clean cases, active targeted accuracy is 0.720 ± 0.006, utility is 0.313 ± 0.013, and coverage is 0.461. A 3×3 noise–threshold grid shows that utility changes sign as the threshold rises (Appendix D).

6.6 Matched-budget selector value

Under one terminal rule, train-only value selection exceeds no-query utility by 0.0138 [0.0129, 0.0148] at 10% and 0.0041 [0.0022, 0.0060] at 25%; its differences from matched random are $+0.0249$ and $+0.0303$ (paired intervals in Table 13). Pair uncertainty has higher utility at every budget. At 50% and 63.7%, learned versus pair-uncertainty non-ambiguous binary accuracy is 0.4922

versus 0.4827 and 0.4443 versus 0.4296. Non-ambiguous binary accuracy excludes ambiguity and counts abstentions as errors; utility includes request and abstention costs; learned utility is lower at both budgets. Among requested examples, mean reward change from the third score before query cost is -0.0591 versus $+0.1756$ at 50% and -0.1343 versus $+0.0235$ at 63.7% (learned versus pair uncertainty). These are within-seed request means across 83 seeds, not independent population units; high-budget learned requests reduce reward. Learned-minus-no-query utility differences are -0.0795 [-0.0825, -0.0766] and -0.1493 [-0.1523, -0.1462]. The oracle uses held-out labels and the unavailable score (Appendix C, Table 13).

6.7 Oracle value versus pre-query cue

Across 24 runs of a two-candidate simulator, oracle gain is 0.10 at a 0.50 request rate, while cue gain over matched random ranges from 0.000019 with no cue to 0.099953 with a perfect cue (Appendix A.3). Oracle headroom alone does not establish feasible pre-request value.

On CMU-MOSEI descriptors (Bagher Zadeh et al., 2018), calibration reaches 65.17% accuracy versus 65.29% for same-request majority. Task-fit projections of released descriptors pass the head-adequacy check, but the router's error-reduction difference from the calibration-selected matched active baseline is 0.000499 (95% video-group interval [-0.000439, 0.001908]) at a 0.50 request budget, below the registered 0.005 margin; its video-macro difference is -0.000156. An online threshold reduces error from no request by 0.025227 at a 0.490754 request rate. These tests use descriptors, not raw media (Appendices A.1 and A.2).

The Perception source gate failed all 25 training-only checks; no conflict benchmark or held-out results were produced (Appendix A.4).

6.8 Natural-media diagnostics

On 1,244 actor-held-out CREMA-D clips (Cao et al., 2014), eight-frame fusion minus audio accuracy is $+5.87\%$ (95% actor interval [-0.56%, $+11.77\%$]) for wav2vec 2.0 (Baevski et al., 2020) with Vision Transformer (Dosovitskiy et al., 2021), and -2.49% ([-8.40%, $+2.81\%$]) for HiddenUnit BERT (Hsu et al., 2021) with Data-efficient Image Transformer (Touvron et al., 2021). Router minus matched-random accuracy is $+0.04\%$ at 99.36% requests and $+0.42\%$ at 71.70%; both intervals span zero (Appendix A.5, Table 7). In a two-team AMI replay, audio–video–text accuracy is 35.77% versus 44.60% for audio–video; text routing minus static audio–video accuracy is -3.69% at 15.57% requests (Appendix A.6). Neither study establishes a reliable selective-acquisition gain.

7 Conclusion and Limitations

The threshold rule saves requests under its synthetic reward. Learned selection helps only at low budgets and loses to pair uncertainty in utility. Temporal findings depend on the waveform; natural-media acquisition gains, latency, privacy, and perceptual benefit remain unconfirmed. Human evaluation is future work (Appendix E). The anonymous repository provides the reproduction code; the supplementary source ZIP contains the same MIT-licensed ConflictBench software with its configurations and setup instructions.

AI USE STATEMENT

We used AI tools to help polish the writing and find related work. Details are described in the paper.

REPRODUCIBILITY STATEMENT

Sections 4 and 5, Appendices A–E, and the anonymous repository and supplementary source ZIP provide the evaluation details, implementation, and setup instructions. The ConflictBench code in the ZIP is released under the MIT License; dataset and dependency terms remain with their respective providers.

REFERENCES

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208. URL https://aclanthology.org/P18-1208/.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2): 423–443, 2019. doi: 10.1109/TPAMI.2018.2798607.

Rui Cai, Bangzheng Li, Xiaofei Wen, Muhao Chen, and Zhe Zhao. Diagnosing and mitigating modality interference in multimodal large language models, 2025. URL https://arxiv.org/abs/2505.19616.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5(4):377–390, 2014. doi: 10.1109/TAFFC.2014.2336244. URL https://doi.org/10.1109/TAFFC.2014.2336244.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Maël Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain A. McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction: Second International Workshop, MLMI 2005*, 2005. URL https://publications.idiap.ch/publications/show/430.

Cheyney Computer Science. CREMA-D: Crowd-sourced emotional multimodal actors dataset. GitHub repository, 2014. URL https://github.com/CheyneyComputerScience/CREMA-D. Official repository commit 1658cd342dff90010aa843eaeebd53610a08b1dc.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/pdf?id=YicbFdNTTy.

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*, pp. 2151–2159, 2019. URL https://proceedings.mlr.press/v97/geifman19a.html.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *arXiv preprint arXiv:2106.07447*, 2021. URL https://arxiv.org/abs/2106.07447.

Jannik Kossen, Cătălina Cangea, Eszter Vértes, Andrew Jaegle, Viorica Patraucean, Ira Ktena, Nenad Tomasev, and Danielle Belgrave. Active acquisition for multimodal temporal data: A challenging decision-making task. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=Gbu1bHQhEL.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 6450–6459, 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Yang Li and Junier Oliva. Towards cost sensitive decision making. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, pp. 3601–3609, 2025. URL https://proceedings.mlr.press/v258/li25h.html.

Hongcheng Liu, Pingjie Wang, Yuhao Wang, Siqu Ou, Yanfeng Wang, and Yu Wang. When seeing is not enough: Revealing the limits of active reasoning in MLLMs. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27395–27422, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1264. URL https://aclanthology.org/2026.acl-long.1264/.

Swapnanil Mukherjee, Agyeya Negi, Tanuja Ganu, and Ponnurangam Kumaraguru. C^3po: Evaluating cross-modal composition and counterfactual performance in omnimodal models, 2026. URL https://arxiv.org/abs/2608.05381.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In *Advances in Neural Information Processing Systems 36*, pp. 42748–42761, 2023. doi: 10.52202/075280-1852. URL https://doi.org/10.52202/075280-1852.

Qwen Team. Qwen2.5-Omni-3B. Hugging Face model repository, 2025. URL https://huggingface.co/Qwen/Qwen2.5-Omni-3B/tree/f75b40e3da2003cdd6e1829b1f420ca70797c34e. Revision f75b40e3da2003cdd6e1829b1f420ca70797c34e.

Tatiana Shpakova and Nataliya Sokolovska. Probabilistic personalised cascade with abstention. *Pattern Recognition Letters*, 147:8–15, 2021. doi: 10.1016/j.patrec.2021.03.029. URL https://doi.org/10.1016/j.patrec.2021.03.029.

Yueru Sun, Yimeng Zhang, Haoyu Gu, Nuo Chen, Dong She, Xianrong Yao, Yang Gao, and Zhanpeng Jin. EmoMM: Benchmarking and steering MLLM for multimodal emotion recognition under conflict and missingness. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 20351–20371, San Diego, California, United States, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.findings-acl.1018. URL https://aclanthology.org/2026.findings-acl.1018/.

Pranuthi Tenali, Sahil Sidheekh, Saurabh Mathur, Erik Blasch, Kristian Kersting, and Sriraam Natarajan. Context-specific credibility-aware multimodal fusion with conditional probabilistic circuits, 2026. URL https://arxiv.org/abs/2603.26629.

Baoliang Tian, Yuxuan Si, Jilong Wang, LingYao Li, Zhongyuan Bao, Zineng Zhou, Tao Wang, Sixu Li, Ziyao Xu, Mingze Wang, Zhouzhuo Zhang, Zhihao Wang, Yi Ke Yun, Ke Tian, Ning Yang, and Minghui Qiu. Crosscheck-bench: Diagnosing compositional failures in multimodal conflict resolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 25887–25895, 2026. doi: 10.1609/aaai.v40i31.39788. URL https://ojs.aaai.org/index.php/AAAI/article/view/39788.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers and distillation through attention. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 10347–10357, 2021. URL https://proceedings.mlr.press/v139/touvron21a.html.

Michael Valancius, Maxwell Lennon, and Junier Oliva. Acquisition conditioned oracle for nongreedy active feature acquisition. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 48957–48975. PMLR, 2024. URL https://proceedings.mlr.press/v235/valancius24a.html.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*, 2025. URL https://arxiv.org/abs/2503.20215.

Qilang Ye, Wei Zeng, Meng Liu, Jie Zhang, Yupeng Hu, Zitong Yu, and Yu Zhou. When eyes and ears disagree: Can MLLMs discern audio-visual confusion? *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(14):11955–11963, March 2026. doi: 10.1609/aaai.v40i14.38183. URL https://ojs.aaai.org/index.php/AAAI/article/view/38183.

A Extended Transfer and Source Diagnostics

This appendix follows two questions: what changes when another source is available, and whether its benefit can be anticipated before requesting it. CMU-MOSEI tests the scalar rule on released descriptors and screens request policies with separate fitting groups. A controlled simulator distinguishes hidden oracle value from an observable pre-request cue. The Perception Test study stops at a training-only source-feasibility check; it produces no conflict benchmark. CREMA-D compares audiovisual fusion and matched-request routing across frozen encoder pairs, while AMI provides an exploratory three-modality replay over two evaluation teams. The later sections examine synthetic action controls, matched request budgets, and cost sensitivity. None of the media studies establishes a reliable gain from selectively requesting a source.

The study counts refer to different objects. The scalar campaign has 83 run seeds and 4,000 examples in each evaluation split per seed. The temporal factorial has 207 run records, 23 per length–noise setting across nine settings; its five-seed full-trace test is a separate follow-up. Each of four controlled pre-query simulation families has 24 seeds and six cue-strength settings, with 400,000 generated evaluation examples per seed at each setting. CMU-MOSEI scores utterances but resamples 676 test videos for the reported transfer interval. The descriptor-routing screens instead reuse the same 1,871 validation utterances from 300 videos across fitting seeds and protocols; they are not independent data replications. The Perception Test counts 100 training-only targets, 200 target–donor pairs, and 2,800 generated answers separately; the 25 failed items are prespecified checks, not 25 independent experiments. CREMA-D scores 1,244 clips and resamples 16 evaluation actors, with six overlapping role allocations used only to probe sensitivity to the allocation. AMI scores 244 donor-matchable dialogue-act records and reports an unweighted mean over two evaluation teams. Its two replays reuse identical evaluation record IDs but change the fitted heads and feature extraction. None of these media units is exchangeable with a synthetic seed or a temporal setting.

A.1 CMU-MOSEI Descriptor Transfer

We test the unchanged rule on released CMU-MOSEI computational-sequence descriptors (Bagher Zadeh et al., 2018): COVAREP audio, OpenFace2 visual, and Timestamped Word Vectors text features. For each labeled utterance, we average temporally overlapping features, standardize with training utterances, and project onto the normalized positive-minus-negative training class-mean direction. The audio, visual, and text descriptors have 74, 713, and 300 dimensions. Sentiment below zero is negative; other values are nonnegative. This is projected-descriptor sentiment, not raw-media inference or a corruption benchmark.

The official video folds yield 16,327/1,871/4,662 train/validation/test utterances shared across modalities. Two test videos are absent from the acquired feature intersection; their IDs and provenance are recorded. We verified retained binary targets against original utterance labels and excluded an invalid earlier implementation that paired whole-video features with each video's first label. Projections use no validation/test labels. Five CPU seeds use batch size 256, learning rate 0.0015, and 40 epochs. The fixed rule keeps $\tau = 0.35$ without target-domain tuning.

Table 3 shows that the synthetic ranking does not transfer. The threshold rule obtains 0.624 test accuracy, 0.944 coverage, and 0.417 query rate; majority reaches 0.653 and the multilayer perceptron (MLP) reaches 0.773 ± 0.002. A deterministic 10,000-resample bootstrap over 676 test videos gives 95% intervals of [0.597, 0.651] for the threshold rule and [0.627, 0.679] for majority; their paired difference is -0.029 $[-0.035, -0.023]$. All models receive only three projected scores, so this is not a comparison with current multimodal encoders. MLP deviation is training-seed variability on fixed folds; fixed policies do not vary across seeds. We add no synthetic ambiguity target or corruption.

Table 3: CMU-MOSEI descriptor accuracy on official folds. Five seeds; only learned models vary. These are not corruption-family splits.

Method	**Validation accuracy ↑**	**Test accuracy ↑**
Majority / median	0.631	0.653
Weighted	0.587	0.616
MLP	0.773 (0.002)	0.773 (0.002)
Gated	0.773 (0.002)	0.772 (0.002)
THRESHOLD	0.598	0.624

As a calibration check, we select each control's threshold by validation utility over 11 values from 0.01 to 2.00, breaking ties downward and leaving test labels unused. Every control selects the grid minimum, so this does not establish an optimum outside the grid. The calibrated rule reaches 65.17% test accuracy, 0.2659 utility, 0.3700 query rate, and 0.9981 coverage. The same-request majority control reaches 65.29% and 0.2689; always acquiring text with the selective rule has identical accuracy but 0.2029 utility. Calibration explains much of the fixed rule's accuracy deficit, while acquisition savings remain distinct from final decisions.

A.2 PROJECTED-DESCRIPTOR ACTION SCREEN

We also replay a separate five-seed, cost-sensitive action screen on the verified utterance-aligned projected cache. It divides official training videos into task-fit, router-fit, and calibration groups and does not use the official test fold for this screen. The three-score projection itself was fit with labels from all official training groups before that division. This is an exploratory downstream screen; its projected features do not constitute an independent feature-learning test. At a fixed 0.50 acquisition budget, an incorrect answer costs 1, abstention costs 0.25, and a request costs 0.05 in normalized loss units. Table 4 reports cost reduction relative to making no request, with 95% intervals formed by resampling validation-video groups. All five seeds have negative absolute differences for each task head, but each interval includes zero. The data therefore support neither a positive acquisition result nor a strictly harmful effect. This check uses released descriptors, not raw media or a natural conflict label.

Table 4: Projected CMU-MOSEI direct-action screen at a 0.50 request budget. Mean normalized cost reduction versus no request over five seeds; paired video-group intervals use 10,000 bootstrap draws. Positive values favor acquisition.

Task head	**Cost reduction ↑**	**95% interval**
Modality-dropout ridge	-0.0055	$[-0.0126, 0.0017]$
Separate-mask ridge	-0.0041	$[-0.0112, 0.0032]$
Shared mask-aware ridge	-0.0055	$[-0.0128, 0.0017]$

A second five-seed screen evaluates a value router that uses only available features at decision time and a calibration-fixed online threshold on the same projected cache. With one observed source and two possible sources to request, the selected router's error-reduction margin over its selected active baseline at the 0.50 target budget is 0.000748 (video-group 95% interval $[-0.000710, 0.002339]$). With two observed sources and one possible request, the margin is 0.000784 ($[-0.001203, 0.003330]$). Both miss the preregistered 0.005 margin and have intervals crossing zero. Relative to no request, the online threshold reduces error by 0.022911 at a 0.487333 realized request rate with two candidates; with one candidate, its reduction is 0.013932 at a 0.504935 rate. Neither comparison establishes an advantage over the matched active baseline. At the same 0.50 budget, a label-informed oracle reduces error by 0.074755 with two candidates and 0.041511 with one. It uses outcomes unavailable before the request. This headroom does not yield a supported allocation gain under this screen.

For the audio-observed, two-candidate decision in this projected-descriptor screen, the online router requested text 6,768 times and video 42 times across the five fits at the 0.50 target budget. The recorded realized-value sum for video requests was zero in each fit, compared with a total of 216 for text requests. These are repeated decisions on overlapping validation videos, not independent examples; this diagnostic neither establishes that video has no value in general nor describes the source-preserving study below.

We repeat the value study from pooled, utterance-aligned COVAREP audio (74 dimensions), OpenFace2 video (713), and timestamped word vectors (300). The cache stores these unprojected released descriptors. For each seed, we split videos first, then standardize and project each source to one score using a class-mean direction fitted only on task-fit videos. We fit task heads there, then train the router and select its baseline on separate groups. Five seeds each cover two observed sources with one possible request and one observed source with two possible requests. Task-head adequacy and two-candidate oracle headroom pass. For the two-candidate decision at an exact 0.50 request budget, the selected router improves error reduction over its calibration-selected active baseline by 0.000499 (95% video-group interval $[-0.000439,\ 0.001908]$). This misses the registered 0.005 margin; the video-macro difference is -0.000156. The gain by initial source is 0.001924 for audio, 0.000107 for text, and -0.000534 for video. An online threshold fitted on calibration videos reduces error against no request by 0.025227 (95% interval $[0.005862,\ 0.045641]$) at a realized 0.490754 request rate. That comparison does not show improvement over the matched active baseline. At the same exact 0.50 budget, the label-informed oracle's error reductions are 0.078675 with two candidates and 0.042188 with one; it cannot serve as a deployable pre-query policy.

The source-preserving replay therefore retains the negative selection result while removing the projected feature-learning overlap in the earlier screen. These are released computational descriptors, not a new raw-media encoding experiment.

A.3 Controlled Pre-Query Value Checks

The controlled simulator fixes one helpful and one harmful candidate at values $+0.20$ and -0.20 before costs. Its observable cue identifies the helpful candidate with probability $(1 + \rho)/2$, so the oracle has information that the policy lacks. Each of four studies uses 24 seeds and 400,000 evaluation examples per seed. The first uses no acquisition cost and an exact 0.50 query rate for the cue, random, and oracle policies. Oracle mean gain is 0.10 at every ρ; cue-minus-random mean gain is 0.000019 at $\rho = 0$, 0.039848 at $\rho = 0.4$, and 0.099953 at $\rho = 1$. This trend is induced by the cue construction and checks that the evaluator separates hidden value from pre-query information.

The second study charges both candidates 0.02. At the same exact 0.50 query rate, cue-minus-random gain is -0.000067 at $\rho = 0$ and 0.099995 at $\rho = 1$. Its risk-aware abstention rule has no exact-budget constraint; at $\rho = 1$ it queries every example and has mean net gain 0.18. The third study estimates candidate-specific lower gains from a separate 400,000-example calibration sample, with candidate costs 0.02 and 0.10. The calibrated policy queries none at $\rho = 0$, one candidate on roughly half the examples at $\rho = 0.4$ (mean query rate 0.499899, net gain 0.030002), and both candidates at $\rho = 0.8$ (query rate 1.000000, net gain 0.099993). Replacing its calibration sample with 400 examples agrees at those three registered checkpoints: query rates 0, 0.499954, and 1.000000 at $\rho = 0$, 0.4, and 0.8, with net gains 0, 0.030001, and 0.099999. Agreement does not extend across the inspected cue-strength grid. At $\rho = 0.2$, the large-calibration policy requests at rate 0.499850, whereas the 400-example policy makes no requests. At $\rho = 0.6$, their rates are 1.000000 and 0.687523, respectively; nine of the 24 finite-calibration runs request on every example, while the other 15 request on roughly half. Their mean net gains are 0.060004 and 0.053745, respectively. The calibration sizes use different seed families, so these differences are descriptive, not paired estimates of a calibration effect. These cost-gated calibration policies can exceed the 0.50 budget recorded for their random and cue comparators. The study is a simulator and implementation diagnostic, not a natural-source evaluation or a fixed-budget advantage for the calibrated policy.

The contrast is between information that exists after acquisition and a cue available beforehand. An oracle gain alone cannot justify a request policy: with an uninformative cue the same-budget policy remains near random, while the unconstrained calibrated policies can request more often than their comparators.

A.4 RAW-MEDIA FEASIBILITY GATE

We test whether Perception Test training data (Pătrăucean et al., 2023) support source-specific questions for a controlled audiovisual extension. Qwen2.5-Omni-3B revision `f75b40e3` (Xu et al., 2025; Qwen Team, 2025) answers the original choices from audio, video, or aligned audiovisual input using deterministic four-token decoding; video uses 2 frames per second (fps) and 4–32 frames. The training-only design has 100 targets and 200 target–donor pairs, split by component into 60/20/20 targets for fitting, calibration, and the untouched gate. Its decision partition has 40 pairs. Question-only and same-question shuffled-media controls test agreement without aligned input. All 2,800 outputs parse under the frozen grammar.

Before inference, the locked gate sets minima of 70% source accuracy, 50% pair support, and an absolute aligned-over-control accuracy gain of 20%. Target/donor accuracy tests unique questions; pair support requires both questions correct. Gain subtracts the larger question-only or shuffled-media agreement from aligned accuracy. These are feasibility cutoffs, not result-selected uncertainty thresholds. Table 5 shows 30–40% target accuracy, 40–55% donor accuracy, 15–25% pair support, and an absolute aligned gain of at most 10%. For 20 pairs, pair-support Wilson 95% intervals are 5.2–36.0%, 8.1–41.6%, and 11.2–46.9% at 15%, 20%, and 25%. All nine paired-component bootstrap intervals for gain include zero, with endpoints from -20% to 30% in absolute accuracy differences. All 15 source/donor/pair, nine gain, and one unseen-component minimum fail: 25 failures, including 12 unseen components versus 15 required. The stop blocks construction, validation/test access, raw routing, and conflict evaluation. This frozen-pool result is not a general Perception Test accuracy estimate.

Table 5: Training-only Perception source gate (percent; target shows Wilson 95% CI; donor/pair rates use 20 pairs). A+V is aligned input; T/S/O is target/same/opposite. Minima: 70 source accuracy, 50 pair support, and an absolute accuracy gain of 20%.

Input	**Target (%) [CI]** ↑	**Donor S/O (%)** ↑	**Pair S/O (%)** ↑	**Gain T/S/O (%)** ↑
Audio	30 [15, 52]	50 / 55	25 / 15	0 / 10 / 5
Video	40 [22, 61]	40 / 55	20 / 20	5 / 0 / 5
A+V	40 [22, 61]	40 / 55	20 / 20	5 / 0 / 5

A.5 CREMA-D NATURAL AUDIO–VIDEO EVIDENCE

CREMA-D provides paired WAV and FLV media, independent voice and face votes, and integrated emotion votes under the ODbL-1.0 database and DbCL-1.0 contents terms recorded in its official repository (Cheyney Computer Science, 2014). We verified all 14,884 staged media files against their official LFS identities before modeling. This is an audio–video representation diagnostic within an audio–video–text protocol paper, not a natural-text evaluation. We allocate its 91 actors by a fixed pre-label hash into 45 task-fit, 15 router-fit, 15 calibration, and 16 evaluation actors. The primary target uses the 7,022 singleton integrated votes. We retain the 420 tied votes for a set-valued sensitivity analysis. Clip identifiers are pairing keys only, and the fixed sentence context is not an emotion feature.

We freeze `facebook/wav2vec2-base-960h` audio features and `google/vit-base-patch16-224` visual features, fit standardized logistic heads on task-fit actors, and mean-pool uniformly sampled visual frames. The router uses only observed-modality features, is fit on router-fit actors, and chooses its threshold on calibration actors. Its matched random comparator has the same evaluation request rate p and expected correctness $p\,\mathrm{correct}_{A+V} + (1-p)\,\mathrm{correct}_{A}$. Table 6 shows that one visual frame is reliably worse than audio alone, while four and eight frames have positive but imprecise actor-cluster intervals on the initial role allocation. Across five more overlapping actor allocations, eight-frame fusion has absolute accuracy gains of 4.75–8.38%; these are split-robustness checks, not independent populations.

Table 6: CREMA-D actor-held-out singleton evaluation with frozen wav2vec 2.0 and Vision Transformer (ViT) features. A+V denotes audio–video fusion. Every condition uses the same 45/15/15/16 actor roles and 1,244 evaluation clips; only the number of uniformly sampled visual frames changes. The last column is the absolute accuracy difference (A+V minus audio) computed from unrounded scores; intervals use 10,000 paired bootstrap resamples of the 16 evaluation actors.

Visual frames	**Video acc. (%) ↑**	**A+V acc. (%) ↑**	**A+V − audio absolute diff. (%) ↑**
1	22.03	35.45	-9.65 [-13.70, -5.28]
4	37.78	48.15	$+3.05$ [-2.80, $+8.21$]
8	43.09	50.96	$+5.87$ [-0.56, $+11.77$]

To probe encoder dependence, we replace wav2vec 2.0 audio with Hidden-Unit BERT (HuBERT) and Vision Transformer (ViT) video with a Data-efficient Image Transformer (DeiT) in the eight-frame procedure. The media, actor roles, labels, prediction heads, and router protocol remain fixed. Audio is decoded as 16-kHz mono float32 and mean-pooled from the last hidden state; video frames are sampled uniformly with endpoints included, processed by the model image processor, and mean-pooled from the class-token embeddings. For wav2vec 2.0, the audio positional-convolution weights are restored from the checkpoint's legacy parameterization, and their digest is recorded by the extractor. The HuBERT–DeiT diagnostic gives 56.43% audio–video accuracy versus 58.92% audio-only accuracy, an absolute accuracy difference of -2.49% with a 16-actor bootstrap interval of [-8.40%, $+2.81\%$]. Across the same five deterministic role reallocations, absolute fused-minus-audio accuracy differences range from -6.00% to $+0.24\%$ (mean -2.35%).

This result does not support an encoder-general temporal-coverage effect under these fixed checkpoints and does not isolate which encoder property explains the difference.

Full-modality value does not establish predictable acquisition value. In the eight-frame condition, the calibrated audio-to-video router requests video for 99.36% of singleton evaluation clips. Its accuracy is 50.96%, versus 45.10% for audio alone. Subtracting the unrounded scores gives an absolute accuracy difference of $+5.87\%$. The router-minus-matched-random absolute accuracy difference at the same request rate is only $+0.04\%$, with a 16-actor bootstrap interval of [-0.30%, $+0.36\%$]. Starting from video yields the same pattern: the policy acquires audio for every evaluation clip. On all 1,300 evaluation clips, set-valued full audio–video accuracy rises from 34.95% to 48.46% in the voice–face-disagreement stratum; this post-query stratum is diagnostic only and never a router input.

For this frozen feature pair, the audiovisual accuracy difference is positive but uncertain; its calibrated router shows no selective value beyond matched acquisition. Under HuBERT–DeiT, the audio-to-video router requests video on 71.70% of clips and has an absolute accuracy difference from matched random selection of $+0.42\%$ [-1.02%, $+1.87\%$]; the reverse router acquires audio for every clip. Thus the initial-role second pair also supplies no selective-acquisition result.

Table 7 collects the frozen-pair comparisons. Each full-fusion and matched-random difference has a 95% actor interval spanning zero, so these results do not establish a reliable full-fusion or selective-acquisition gain.

An expanded 2×2 grid tests whether this boundary survives component swaps. It joins wav2vec 2.0 or HuBERT audio features with ViT or DeiT visual features and repeats the pre-label actor allocation six times. The reallocations overlap, so their means and ranges are split-robustness descriptions and not independent-population intervals. At calibration budgets of 10, 25, 50, and 75%, wav2vec 2.0–ViT has mean absolute policy-minus-matched-random accuracy differences of $+0.54\%$, $+1.24\%$, $+1.04\%$, and $+0.89\%$ at realized request rates of 5.46, 24.13, 44.23, and 65.50%, respectively.

All six allocation means are positive at 10% and 25%, but actor-bootstrap intervals span zero in five of six allocations at 10% and four of six at 25%. At 50% and 75%, one of the six allocation means is negative. The three swapped pairs have no coherent sign pattern across budgets and reallocations. Table 8 makes the pair-specific means visible: wav2vec 2.0–ViT has the largest descriptive margin at each calibration target, but realized request rates differ across pairs.

Table 7: Pinned CREMA-D comparison on 1,244 singleton-vote clips from 16 held-out actors. Pair names list audio then video encoders; A+V uses both. Audio and A+V columns report accuracy, and request is the realized audio-to-video router rate. The two difference columns report A+V minus audio and router minus matched-random accuracy, respectively, computed from unrounded scores as absolute differences (%). Brackets give 95% actor bootstrap intervals; matched random uses the same request rate.

Frozen pair	Audio (%) ↑	A+V (%) ↑	A+V − audio (%) ↑	Request (%) ↓	Router − matched random absolute diff. (%) ↑
wav2vec 2.0–ViT	45.10	50.96	+5.87 [−0.56, +11.77]	99.36	+0.04 [−0.30, +0.36]
HuBERT–DeiT	58.92	56.43	−2.49 [−8.40, +2.81]	71.70	+0.42 [−1.02, +1.87]

Table 8: CREMA-D comprehensive frozen-feature grid. Values are means over six overlapping pre-label actor reallocations. Column percentages are calibration targets, not exact evaluation request rates. Full headroom is full audio–video accuracy minus audio-only accuracy. Δ is policy accuracy minus expected matched-random accuracy at the realized request rate; all values are absolute accuracy differences (%). The reallocation means are descriptive split robustness, not independent-population estimates.

Frozen pair	Full headroom (%) ↑	10% Δ (%) ↑	25% Δ (%) ↑	50% Δ (%) ↑	75% Δ (%) ↑
wav2vec 2.0 + ViT	+6.28	+0.54	+1.24	+1.04	+0.89
wav2vec 2.0 + DeiT	+3.09	+0.01	+0.16	+0.34	+0.05
HuBERT + ViT	+0.89	+0.14	−0.13	+0.25	+0.68
HuBERT + DeiT	−2.35	+0.06	+0.34	+0.71	+0.47

The other frozen pairs mix positive and negative margins across overlapping reallocations. They do not establish an encoder-general selective-acquisition result.

Table 9 reports how often each mean difference is positive across the six overlapping actor reallocations. The wav2vec 2.0–ViT matched margin is positive in all six allocations at 10% and 25%, but in five of six at 50% and 75%; the other pairs show mixed signs across several targets. These counts describe split sensitivity and are not independent replications.

Table 9: Sign consistency of the CREMA-D comprehensive grid. Each cell gives the number of positive reallocation-level estimates out of six overlapping actor allocations. Full headroom is full audio–video accuracy minus audio-only accuracy; the four request columns give policy accuracy minus matched-random accuracy at the realized request rate. Counts summarize split sensitivity and are not confidence intervals or independent-population estimates.

Frozen pair	Full headroom	10%	25%	50%	75%
wav2vec 2.0 + ViT	6/6	6/6	6/6	5/6	5/6
wav2vec 2.0 + DeiT	5/6	3/6	3/6	5/6	5/6
HuBERT + ViT	5/6	3/6	3/6	3/6	3/6
HuBERT + DeiT	1/6	3/6	4/6	5/6	5/6

A.6 AMI AUDIO–VIDEO–TEXT PILOT

We reproduced a frozen-feature dialogue-act diagnostic from official AMI signals and annotations (Carletta et al., 2005). Ten meeting teams are assigned by role: five for task fitting, one for router fitting, one for calibration, two for evaluation, and one for same-label donor controls. Unique donor matching retains 244 of 616 requested evaluation records. The first replay corrected `static` to use the available audio–video head in place of a majority-class constant. Its two-team mean accuracy was 44.50%, versus 64.68% for the full audio–video–text head and 69.60% for a label-aware oracle choosing between them. Text alone reached 75.35%. This replay used standardized L2 logistic regression with $C = 1.0$, although the registration specified a multinomial ridge head. Its per-example timing and matched-cost fields were fixed placeholders (0 milliseconds and 1 unit); the 98.38% calibrated request rate did not establish selective value.

The evaluation record-ID sets in the two AMI replays have an exact 244-of-244 intersection. The following numbers are therefore a different analysis of the same evaluation subset, not an independent replication or a pooled estimate. We repeated frozen extraction with masked-mean text pooling, recorded observed encoder time for each modality, and fit standardized one-hot ridge heads with $\alpha = 1.0$. On the same 244 donor-matchable evaluation records, the two-team macro accuracies are 60.62% for audio, 44.60% for audio–video, 76.18% for text, and 35.77% for audio–video–text. Adding modalities to this fitted head therefore reduces accuracy; text-only accuracy does not imply a usable text-request policy. The learned router uses only the available audio–video features and head outputs, fits on router-fit teams, and selects its threshold on calibration teams subject to a 50% request cap. It requests text on 38 of 244 evaluation records (15.57%). The router-minus-static absolute difference in accuracy is -3.69%, averaged over the two evaluation teams. At exactly 38 requests, the router-minus-random and router-minus-confidence absolute differences in accuracy are -1.69% and -4.96%, respectively. A label-informed oracle at this same request count has an absolute difference in accuracy of $+15.56\%$ versus static; it is an unattainable reference, not evidence that the learned router can identify those records. Mean observed text extraction time is 4.28 ms and audio–video extraction is 1404.26 ms on these records. These modality measurements do not include complete application latency.

The macro values above average the two held-out meeting-team accuracies with equal weight: $(\text{accuracy}_{\text{ES2006}} + \text{accuracy}_{\text{ES2011}})/2$. Table 10 gives the observed class counts and all seven subset accuracies by team; Table 11 gives the selected router and its exact-count controls by team. This exposes the two terms behind each macro value; a pooled accuracy over all 244 records would weight the teams differently.

Table 10: AMI corrected replay by evaluation meeting team. Class counts follow the observed target-class order `elicit`, `minor`, `other`, `task`. Accuracy is the percentage correct within each team for the indicated feature subset; A, V, and T denote audio, video, and text. The two-team macro is the unweighted mean of the ES2006 and ES2011 values; the breakdown is descriptive because only two evaluation teams are available.

Team	Records	Class counts	Accuracy (%) ↑						
			A	V	T	A+V	A+T	V+T	A+V+T
ES2006	123	9 / 23 / 11 / 80	65.04%	27.64%	82.11%	52.85%	64.23%	34.96%	21.95%
ES2011	121	9 / 23 / 10 / 79	56.20%	19.83%	70.25%	36.36%	53.72%	37.19%	49.59%

Table 11: AMI corrected-replay accuracy by evaluation team for the available audio–video baseline and text-request policies. Each percentage is computed within the team's records. Each exact-count control uses 38 requests across all 244 records; team allocations can differ. Macro differences described in the text are unweighted means of the two team-level differences.

Team	Policy	Accuracy (%) ↑
ES2006	Audio–video static	52.85%
	Learned text router	48.78%
	Matched random	44.72%
	Matched confidence	48.78%
	Label-informed oracle	69.92%
ES2011	Audio–video static	36.36%
	Learned text router	33.06%
	Matched random	40.50%
	Matched confidence	42.98%
	Label-informed oracle	50.41%

The corrected replay is still exploratory. Only two meeting teams contribute evaluation records, making a team bootstrap too coarse for broad transfer inference. More fundamentally, the registered operating-point validator requires the static, all-request, and selective policies to have the same request rate, which is impossible by their definitions. We report actual rates and exact-count matched controls, but do not call the registered natural-source acquisition gate passed. The original and corrected replays remain separate because their task heads, text pooling, and timing semantics differ.

Across the descriptor, audiovisual, and meeting studies, full-modality accuracy does not establish a useful pre-request ranking. The studies also have different targets and independent units, so their accuracy differences should not be pooled as one transfer estimate.

B Matched Synthetic Controls

The five-seed control isolates the action identity of the fixed threshold policy. Always-query-selective uses the same final-action rule but pays for every request; query-majority keeps the request schedule and changes the post-request action. Table 12 shows that the active and always-query-selective policies make identical decisions at different request rates, while query-majority keeps the active request rate but has lower utility after changing the post-request action.

Table 12: Matched synthetic controls at $\sigma = 0.22$, $\tau = 0.35$. Targeted decision accuracy and utility are mean (sample standard deviation) over five paired seeds. All methods start with the same pair; all requests cost 0.10.

Policy	Targeted acc. ↑	Utility ↑	Query ↓	Coverage
Pair-selective	0.594 (0.006)	0.110 (0.012)	0.000	0.366
Query-majority	0.628 (0.005)	0.192 (0.010)	0.634	1.000
Always-query-selective	0.790 (0.003)	0.447 (0.007)	1.000	0.594
THRESHOLD	0.790 (0.003)	0.484 (0.006)	0.634	0.594

The matched-action identity is direct evidence for request savings. It does not imply that the initial pair predicts which third scores will change an answer: under this terminal rule, no third score changes the action when the threshold policy declines to request it.

C MATCHED-BUDGET SELECTOR VALUE

The 83-seed comparison applies one terminal rule to train-only value, pair-uncertainty, matched-random, and oracle request schedules at four budgets. The no-query reference has utility -0.0459 (sample standard deviation 0.0160) at each budget. The final column reports the train-only selector's paired mean utility difference from matched random, with a 95% paired-seed percentile interval from 10,000 resamples using seed 92309. Positive values favor learned value selection. Pair uncertainty has higher utility than the learned selector at all four budgets. The oracle uses held-out labels and the unavailable score, so it is an upper bound at each exact budget and is not a deployable policy. An exact budget forces requests even when their gain is below the request cost; the oracle utility may therefore fall as the budget grows. Table 13 shows the budget dependence: the learned selector exceeds matched random at 10% and 25%, falls below no-query utility at 50% and 63.7%, and remains below pair uncertainty in utility at every budget. The two high-budget comparisons reverse for non-ambiguous binary accuracy: learned versus pair uncertainty is 0.4922 versus 0.4827 at 50% and 0.4443 versus 0.4296 at 63.7%. This accuracy excludes ambiguity cases and scores abstentions as errors. It does not encode request or abstention cost, and targeted decision accuracy favors pair uncertainty at both budgets. Among requested examples, mean pre-query-cost reward change from the third score is -0.0591 versus $+0.1756$ at 50% and -0.1343 versus $+0.0235$ at 63.7% (learned versus pair uncertainty). This is the change between the common terminal rule's full-score and pair-only actions before request cost, averaged over requested examples within each of 83 seeds. Because the schedules request different examples, it is a conditional diagnostic, not a paired population effect or evidence that the third score is harmful for all examples; repeated requests are not independent population units.

Table 13: Matched-budget selector comparison under one shared terminal rule over 83 seeds. Cells show utility mean (sample standard deviation). The final column gives the train-only value selector's paired mean utility difference from matched random with a 95% paired-seed percentile interval. The no-query reference is -0.0459 (0.0160) at each budget.

Query budget	**Train-only value ↑**	**Pair uncertainty ↑**	**Matched random ↑**	**Oracle upper bound ↑**	**Value–random Δ U (95% CI) ↑**
10%	-0.0321 (0.0166)	-0.0304 (0.0164)	-0.0570 (0.0168)	0.1479 (0.0159)	$+0.0249$ $[0.0233, 0.0264]$
25%	-0.0418 (0.0161)	-0.0296 (0.0166)	-0.0721 (0.0164)	0.2455 (0.0151)	$+0.0303$ $[0.0280, 0.0325]$
50%	-0.1254 (0.0175)	-0.0081 (0.0147)	-0.0994 (0.0164)	0.2205 (0.0151)	-0.0260 $[-0.0292, -0.0228]$
63.7%	-0.1952 (0.0167)	-0.0946 (0.0151)	-0.1148 (0.0158)	0.2068 (0.0151)	-0.0804 $[-0.0832, -0.0776]$

The low-budget gain over matched random is limited to this shared terminal rule and synthetic population. At larger budgets the learned ranking requests examples with negative mean reward change, while pair uncertainty retains higher utility despite the learned ranking's small binary-accuracy advantage. Accuracy and costed utility answer different questions here.

D COST-SENSITIVITY DETAIL

A separate 3×3 grid varies score noise $\sigma \in \{0.11, 0.22, 0.44\}$ and threshold $\tau \in \{0.175, 0.35, 0.7\}$ without test-set threshold selection. Active utility spans -0.058 to 0.497.

At campaign noise, raising τ from 0.35 to 0.7 changes utility from 0.484 to -0.039; at $\tau = 0.35$, raising noise to 0.44 gives 0.337. These shifts reflect score magnitudes and the weak-score ambiguity marker, not robustness to arbitrary rescaling.

Figure 3 asks how much of the threshold rule's synthetic utility survives higher request prices. Its utility decreases as λ_q rises, while majority stays constant because it neither queries nor abstains. This is a re-scoring of fixed actions, not a policy refit.

Table 14 gives the companion numerical comparison. At the campaign costs $(\lambda_q, \lambda_a) = (0.10, 0.20)$, the threshold rule has 0.481 ± 0.014 mean utility versus 0.252 ± 0.016 for majority. At $(0.40, 0.80)$, its utility is 0.192 ± 0.018, below majority's unchanged 0.252 ± 0.016. The cost pair changes the preferred policy under this generator; the analysis does not estimate real acquisition latency or cost.

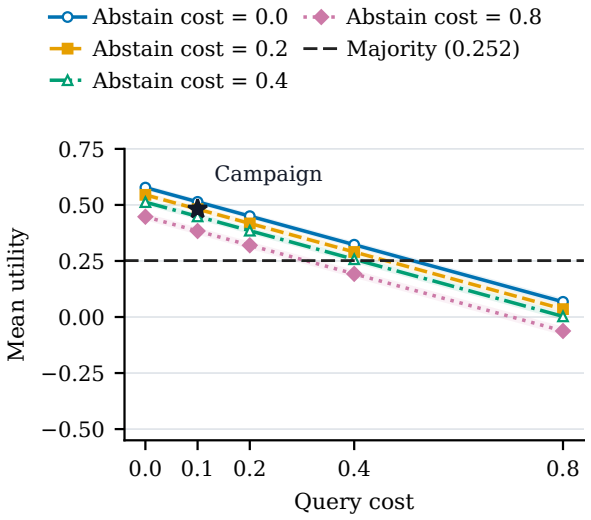


Figure 3: Cost sensitivity on the partially held-out split. Curves show the active policy's mean utility as query cost varies; shaded regions are one standard deviation across seeds. The dashed line is the majority baseline, whose query and abstention rates are zero. The star marks the campaign costs.

Table 14: Cost sensitivity from the same 83 verified runs. Policy utilities are mean (sample standard deviation); the utility-difference column is threshold utility minus majority utility. Majority is unchanged because it never requests or abstains. Costs are query λ_q and non-ambiguous abstention λ_a.

Costs (λ_q, λ_a)	**Threshold utility** ↑	**Majority utility** ↑	**Utility difference** ↑
$(0.00, 0.00)$	0.578 (0.012)	0.252 (0.016)	0.326
$(0.10, 0.20)$	0.481 (0.014)	0.252 (0.016)	0.230
$(0.20, 0.40)$	0.385 (0.015)	0.252 (0.016)	0.133
$(0.40, 0.80)$	0.192 (0.018)	0.252 (0.016)	−0.059

Figure 3 and Table 14 locate a change in policy ordering under fixed actions. They do not identify application-specific acquisition prices, and no threshold is refitted for the alternative costs.

E Extended Limitations

The study isolates a protocol, and that isolation determines what the results mean. Scores are not waveforms, frames, or language-model representations. The active policy receives a third score after a request without modeling acquisition latency, missingness, privacy, or temporal alignment.

The CREMA-D routing comparison matches request rate only; it does not measure decoding delay, monetary cost, or a deployment availability process. The synthetic generator also uses a fixed threshold and a small set of mechanisms. Invert and swap share the same score-level transformation.

The reported mechanism-label summaries retain both labels and therefore count that transformation twice. Future studies could test distinct transformations.

The temporal extension is not a natural-media evaluation. Its zero-mean marker removes information from time means while leaving a fixed waveform in full traces. The equal-information follow-up shows that a train-only template and a ridge classifier match the temporal networks and that a shared time permutation preserves the separation. The original factorial therefore does not establish a temporal-model advantage or transfer to raw audio, video, or text.

Permutation diagnostics. Under independent channel permutations, partially held-out-split targeted decision accuracies are 0.768 ± 0.009, 0.619 ± 0.009, 0.428 ± 0.015, and 0.428 ± 0.008 for temporal fusion, temporal gated, full-trace ridge, and the train-only template, respectively. Per-example shifts give 0.781 ± 0.009, 0.720 ± 0.016, 0.430 ± 0.018, and 0.424 ± 0.007 in the same order. The shared train–evaluation permutation leaves all four methods within their central-setting range because it preserves one fixed positional template. The hard identity setting uses weaker signal and stronger noise but yields similar utility across the methods (Table 2); it does not establish a general difficulty gradient.

The learned baselines use only observed scores as model inputs and are trained on the same synthetic family. Their seen performance is therefore a sanity check, not a comparison with a state-of-the-art multimodal encoder. In particular, the active rule has access to the third coordinate after a request, whereas fixed baselines always receive all three coordinates at the initial decision. The CREMA-D frozen-backbone replication also reverses the positive wav2vec 2.0–ViT fusion difference, so the natural-media result is not encoder-independent. A fair real-media comparison would need to define modality availability, request latency, preprocessing cost, and a preregistered encoder set for every method.

The designated ambiguity mechanism encodes a particular safety preference: abstention is rewarded for these examples, but not for the equally label-nonidentifiable mixed examples. Changing λ_q or λ_a changes the preferred operating point. We therefore report the underlying action rates and selective accuracy alongside utility, and we treat cost sensitivity as a required stress test before interpreting a policy as preferable.

The descriptor experiment adds real-data provenance but does not resolve the score abstraction. Validation-only calibration largely closes the fixed threshold's accuracy gap, but does not improve utility over the matched majority control. The Perception Test pilot adds raw-media inference without establishing a usable conflict benchmark: its question pool is not reliably source-grounded under the frozen evaluator, and its 20-component decision partition is too small for the pre-inference unseen-question check. The stop protects the later analysis from an outcome-selected construction, but it does not show that every Perception Test question or multimodal model lacks source-specific information. A new raw study needs a separately frozen dataset, model, and task contract. The study uses previously released benchmark data and derived descriptors; it recruits no participants, gathers no new annotations, and runs no human evaluation. It inherits the representational and consent limitations of those releases. Human evaluation is future work, subject to explicit approval; the present results do not address perceptual quality, accessibility, clinical use, or deployment readiness.