Before Training, Which Multimodal Data Will Help? Resolution Limits Across Granularities and Learners

Anonymous authors
Paper under double-blind review

ABSTRACT

Choosing a multimodal training corpus poses a catch: we learn whether it helps only after paying for the training run. When there are more corpora than we can try, can frozen representations tell us which ones are worth that cost? We test this question on three paired audio and text datasets for sentiment and emotion. Our sealed evaluation fixes each decision before candidate utility is revealed and repeats the comparison with four downstream learners. The answer depends on selection granularity. With logistic regression, centroid distance chooses the oracle source in 44 of 45 decisions, yet fails to rank buckets from that source. The separation between sources is 22.3 times the typical within-source spread; matching bucket size and class balance does not recover the ranking. Utility also changes with the learner and fusion rule, while candidate-specific versions of LEEP and LogME remain unreliable as direct bucket rankers. We use these signals to form a shortlist rather than replace downstream evaluation. Before opening a new repartition family, we fix a procedure that routes by distance, ranks by adapted LEEP, evaluates the top two candidates with the full downstream recipe, and keeps the target-only model if neither helps. Across three targets, it improves validation macro F1 by 0.01084 on average, with a 95% hierarchical bootstrap interval of [0.00293, 0.01957], while avoiding 87.5% of full candidate runs. The result locates a practical boundary: prospective scores can guide a training budget, but they are not universal measures of dataset quality.

1 INTRODUCTION

An audio and text project may have access to many corpora but enough compute for only a few transfer runs. The first decision is then about data, not model architecture: which corpora should receive the available budget? A source may help the target model, have no effect, or hurt it because its labels and features follow different distributions. Trying every source settles the question, but only after consuming the budget that motivated selection in the first place.

Intrinsic data quality does not settle this decision. Clean labels and aligned modalities matter, but they do not establish whether a source will help a particular target model. We define utility accordingly as $\mathcal{U}(\mathcal{S}; \mathcal{T}, M, R)$: the source $\mathcal{S}$, target $\mathcal{T}$, model M, and training recipe R together determine the quantity to be predicted. A corpus can thus help one downstream system and hurt another.

We examine two ways in which a promising selector can fail. It may distinguish widely separated corpora but be unable to order similar subsets of the chosen corpus. It may also learn a ranking that holds for one downstream learner and reverses for another. Our protocol seals each decision before evaluation: routes, bucket rankings, and abstention rules are fixed on held-out targets or repartitions before candidate utility is disclosed. We report abstention together with coverage because a rule that rejects every candidate avoids harm without solving the selection problem.

Our study yields four main findings:

- We formulate multimodal dataset utility with an explicit information boundary and evaluate it with a deterministic five-seed protocol using common frozen text and audio representations.
- For logistic regression, source routing is stable across repeated partitions, succeeding in 44 of 45 decisions over three independent target corpora. Bucket ranking and calibrated absten-

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.

tion are not stable. Split-half, equal-size, and class-balanced controls trace this difference to the much larger separation between sources than between buckets from one source.

- Repeating the same comparisons with GaussianNB, a neural MLP, and learned gated fusion shows that utility is learner dependent. The matched ablations identify two sources of reversal: audio moment shifts erase text gains under GaussianNB, while adaptive fusion changes the preferred source.
- Candidate-trained adaptations of LEEP and LogME do not solve direct bucket ranking under the same seal. A prespecified alternative uses them only to form a shortlist and then evaluates two candidates exactly. On a fresh family, this procedure improves every target mean and avoids 87.5% of full candidate runs.

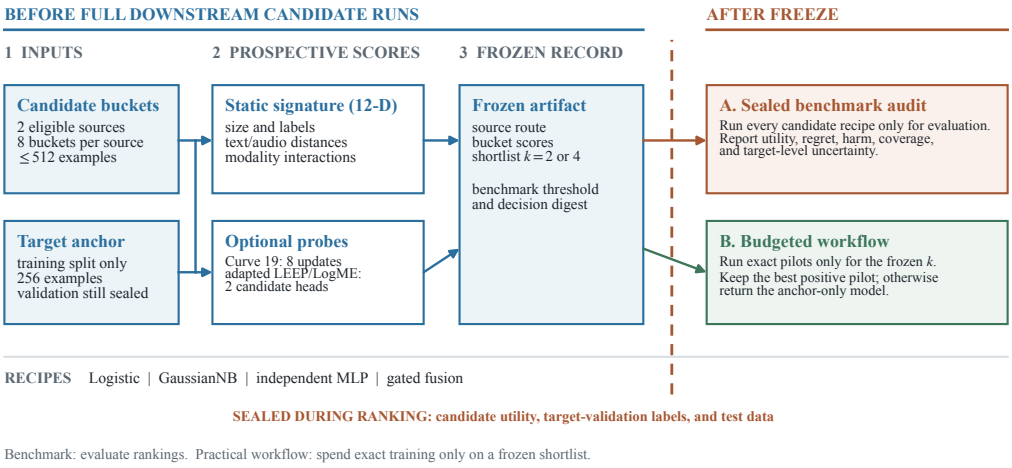

Figure 1: **Information boundary and evaluation flow.** Candidate buckets and target training anchors are used to compute prospective scores. The route, ranking, rejection threshold, and shortlist are recorded before candidate utility or validation labels are accessed. The benchmark then trains every candidate for audit. The budgeted procedure evaluates only the frozen shortlist and keeps either the best positive pilot or the anchor-only model.

2 RELATED WORK

Transferability estimation. LEEP, LogME, and TransRate estimate how well a pretrained model or representation will transfer without exhaustive fine tuning (Nguyen et al., 2020; You et al., 2021; Huang et al., 2022). These methods rank source models or feature spaces. Our candidates are training-data subsets under a shared representation, and their value is the change they induce in a specified downstream learner. Canonical LEEP uses predictions from a source-trained classifier on labeled target examples, whereas LogME measures the evidence of target labels under a supplied representation. If the target representation is fixed, neither method assigns different scores to different candidate buckets. We therefore distinguish their canonical use from our candidate-trained probe versions throughout the paper. Recent work also shows that static SITE leaderboards can reverse under realistic changes in architecture, task, and training protocol (Singh et al., 2026). We bring the same question to candidate data by separating corpus routing from within-corpus ranking and repeating both decisions across learners and repartitions. Every reported utility is tied to the downstream recipe that produces it.

Data valuation and task aware selection. Data Shapley assigns value through marginal downstream utility, whereas datamodels learn a mapping from training subsets to model predictions (Ghorbani & Zou, 2019; Ilyas et al., 2022). DsDm uses this model aware view for language model dataset selection, and LESS uses optimizer aware gradient similarity for targeted instruction tuning (Engstrom et al., 2024; Xia et al., 2024). These approaches make utility conditional on the target or learner, but obtain their scores from subset-training supervision, warm-up training, or gradients. Our

static signatures are computed before candidate training. We treat short trained probes as a separate class of method rather than call them training free.

Multimodal curation. DataComp standardizes image and text filtering under fixed compute, while JEST selects jointly learnable multimodal batches using learner and reference losses (Gadre et al., 2023; Evans et al., 2024). Recent preprints further combine intrinsic quality, relevance, diversity, validation feedback, capability specific raters, or one step proxy updates (Cheng et al., 2026; Sahi et al., 2026; Jing et al., 2026). Most of this work targets large-scale vision-language pretraining or multimodal instruction tuning. We instead ask whether entire audio-text sources, and then subsets within them, can be screened before running a fixed target-augmentation recipe.

Training free supervision data generalization. DataProphet is the closest precedent. It predicts influence rankings for vision-language supervision datasets from perplexity, similarity, and diversity without candidate-specific training (Qi et al., 2026). Our question is not whether training-free multimodal selection is possible in general. It is whether the same premise survives a finer selection granularity and a change of downstream learner. We test this with sealed natural repartitions, uncertainty over target corpora, and an abstention rule calibrated on a separate family.

3 PROBLEM FORMULATION

Let $D_{\mathcal{T}}^{\text{tr}}$ and $D_{\mathcal{T}}^{\text{val}}$ be the target training and validation sets, and let $\mathcal{S}_i$ be a candidate source block. For a fixed model and recipe (M, R), the utility realized under seed s is

$$u_{i,s} = q\big(M_R(D_{\mathcal{T}}^{\text{tr}} \cup \mathcal{S}_i; s), D_{\mathcal{T}}^{\text{val}}\big) - q\big(M_R(D_{\mathcal{T}}^{\text{tr}}; s), D_{\mathcal{T}}^{\text{val}}\big)\,, \tag{1}$$

where q denotes macro F1. We take ground-truth utility to be $\mathcal{U}_i = |\mathcal{K}|^{-1} \sum_{s \in \mathcal{K}} u_{i,s}$ over five locked seeds. Positive utility means that adding the candidate improves validation performance on average.

The prospective stage can access the candidate data, target training anchors, model metadata, and features computed from them. Candidate-specific $u_{i,s}$, target validation labels, and target test data remain hidden. Given scores $\hat{u}_i$, a forced selector chooses $\arg\max_i \hat{u}_i$. We report the selected utility and top-one regret

$$\text{Regret} = \max_i \mathcal{U}_i - \mathcal{U}_{\hat{\imath}}. \tag{2}$$

A selective rule can abstain, which gives it zero realized utility. Its coverage is reported alongside accepted harm, defined by $\mathcal{U}_{\hat{\imath}} < 0$. The budgeted procedure introduced later has a different purpose. Once a small shortlist has been frozen, it pays for exact downstream runs and uses validation utility to keep a positive pilot or fall back to the anchor-only model.

4 EVALUATION FRAMEWORK

4.1 DATA AND COMMON REPRESENTATIONS

Our multimodal panel contains CMU-MOSI (Zadeh et al., 2016), MELD (Poria et al., 2019), and CH-SIMS (Yu et al., 2020). Only audited training and validation splits are used. We map sentiment labels into negative, neutral, and positive classes. The frozen, 384-dimensional `paraphrase-multilingual-MiniLM-L12-v2` sentence transformer represents text (Reimers & Gurevych, 2019; Wang et al., 2020); the frozen, 384-dimensional `whisper-tiny` encoder represents audio (Radford et al., 2022). Model revisions are fixed in the released manifests. We align candidate and target identifiers before running transfer experiments. Target anchors contain 256 examples and candidate blocks contain at most 512 examples.

The development partition keeps groups intact by hashing video or dialogue identifiers. The three confirmation families instead hash salted sample identifiers, which makes them deterministic but not group preserving. A source example may move between salts and appear again in another family. We therefore treat salts and families as repeated measurements, not independent datasets.

For each target, a set of 256 training anchors defines the baseline. We append candidate data while capping their total source weight at the anchor weight. Separate text and audio heads each produce probabilities for three classes, and we average their outputs. Utility is the change in validation macro F1 relative to the anchor-only baseline, averaged over five fixed seeds.

4.2 PROSPECTIVE SIGNATURES AND HIERARCHICAL DECISIONS

The 12-dimensional static signature contains candidate size, label entropy, text and audio centroid distances, label alignment, cross-modal alignment, and class-conditional distances. An eight-update probe adds seven learning-curve features from support and query splits of the target training data, producing the 19-dimensional Curve 19 signature. Although the probe never sees target validation data or realized candidate utility, it does perform training.

Selection has two stages. The source route is the corpus with the smallest mean text and audio centroid distance over its eight buckets. Within the routed source, Curve 19 uses ridge regression on the 19-dimensional probe signature to rank buckets. A balanced logistic classifier supplies a second ranking by predicting whether $\mathcal{U}_i > 0$. We fit both selectors in leave-one-target-out folds. Features are standardized within each target and source stratum without using utility labels.

4.3 CANDIDATE TRAINED TRANSFERABILITY PROBES

LEEP and LogME enter the same sealed evaluation after a necessary change in scope. For every candidate bucket, we fit class-balanced logistic heads to its frozen text and audio representations. Adapted LEEP averages their aligned probabilities on the labeled target anchor. Source-head LogME treats the concatenated text and audio log probabilities as a six-dimensional target representation. We average both scores over five anchor seeds. Because each method fits a small model for every candidate, we call them probe baselines rather than static selectors. We evaluate rankings both before and after the frozen distance route. Appendix B distinguishes the canonical estimators from these adaptations, including their label access and candidate-specific computation.

4.4 SEALED CONFIRMATION AND UNCERTAINTY

Each repartition family contains 240 candidates: three targets, two eligible sources per target, five salts, and eight buckets. The three selector-audit families follow separate protocols but share samples. Prediction artifacts are written before utilities are generated. We choose rejection thresholds on a dedicated calibration family and evaluate them once on the final family.

After the exploratory analyses, we locked an unused fourth prefix together with the distance route, adapted LEEP ranking, shortlist sizes $k \in \{2, 4\}$, fallback rule, and success criterion. We wrote hashed route and score artifacts while the corresponding utility artifact was absent. The new five-seed panel was generated only after this freeze.

Partition decisions share corpora, so they are not independent replicates. For the fresh workflow, we form 10,000 hierarchical bootstrap draws, first over targets and then over salts. Earlier comparisons are averaged within targets and also use all 2^3 target-level sign flips. With only three target clusters, the intervals are descriptive. Treating salts as independent would exaggerate the effective sample size.

4.5 DOWNSTREAM LEARNERS

The primary recipe fits class-balanced logistic regression heads to text and audio. In matched panels, we replace both heads with GaussianNB or independent two-layer MLPs while keeping the candidates, weights, representations, seeds, and validation metric fixed. Each MLP has a 64-unit ReLU hidden layer and is trained for 80 deterministic, full-batch AdamW updates. A fourth panel learns fusion jointly. Modality-specific 384-to-64 ReLU branches feed a two-way softmax gate and a shared three-class head, also trained for 80 updates. The gate ablation compares fixed equal weights, one learned global weight, and the original sample-dependent gate. Every other parameter and update is held fixed. GaussianNB, independent MLP, and gated fusion anchor macro F1 ranges are 0.442 to 0.532, 0.437 to 0.540, and 0.423 to 0.543, respectively. None of the comparisons therefore starts from a completely degenerate anchor model.

4.6 COMPUTE ACCOUNTING

Table 1 reports exact training counts for the primary sealed panels and keeps machine-dependent timings separate. The four logistic families, including the fresh workflow confirmation, and the

Table 1: **Computation in the audited primary panels.** Counts are exact for the listed sealed decision and model panels. One block denotes a target, source, salt, and bucket combination and receives five seeded downstream evaluations. Head-based panels fit separate text and audio models. Preliminary panels and auxiliary diagnostics are excluded.

Panel	Blocks	Seeded evaluations	Fitted models or updates
Four logistic families	960	4,800	9,720 heads
Matched GaussianNB	240	1,200	2,430 heads
Matched neural MLP	240	1,200	2,430 heads
Head based subtotal	1,440	7,200	14,580 heads
Three gate variants	720	3,600	3,645 models / 291,600 updates

Table 2: Error decomposition on the initial three-target panel. Oracle diagnostics cannot be deployed. The low regret of the routed-source oracle places the logistic model's remaining error in within-source bucket selection.

Decision rule	Mean utility ↑	Regret ↓	Positive targets
Flat static signature (12-D)	-0.02800	0.04278	0/3
Routed static signature (12-D)	-0.01252	0.02730	1/3
Routed-source oracle bucket	+0.01417	0.00060	3/3

matched GaussianNB and MLP panels total 7,200 candidate-seed evaluations. Including shared anchor baselines, they require 14,580 fitted modality heads. The three matched gate variants add 3,600 candidate-seed evaluations and 3,645 jointly trained models, equivalent to 291,600 full-batch updates. This ledger excludes preliminary utility panels and auxiliary probe and diagnostic fits. In a deterministic rerun of one 240-candidate panel, Curve 19 screening takes 14.46 seconds and full logistic utility generation takes 17.88 seconds. The 19.1% reduction is modest because both procedures are cheap in this classical setting. Two independent MLP executions reproduce all 1,200 rows exactly.

For the motivating example of selecting 10 candidates from 100, exhaustive five-seed evaluation with two heads requires 1,000 candidate head fits. Evaluating only the selected ten requires 100, which avoids 90% of candidate fits, or 89.1% after including a shared ten-fit baseline. This is a conditional count, not a measured speedup or evidence that the selected ten are optimal. We do not measure end-to-end savings for neural training.

5 RESULTS

5.1 SOURCE ROUTING IS EASIER THAN BUCKET SELECTION

Table 2 shows where selection fails. Adding source routing to the original bucket score reduces regret from 0.04278 to 0.02730, but mean utility remains negative. If an oracle chooses the bucket after the same route, utility is positive on all three targets and regret falls to 0.00060. The route therefore reaches a source that contains useful buckets under logistic regression. The remaining error lies in choosing among those buckets.

Bucket size and class proportion do not account for this error. On the sealed safety family, the median gap between two sources is 22.33 times the standard deviation of distances among buckets from one source. Two independent halves select the same source in all 15 decisions. Within a source, however, split-half reliability is $\tau = -0.152$. Distance has almost no association with utility ($\tau = 0.005$), even after equal-size subsampling ($\tau = -0.010$) or class-balanced centroid estimation ($\tau = -0.038$). The distance signal is thus large and reproducible at the source level, but the smaller bucket differences are neither stable nor useful for ranking.

Table 3: Mechanism checks and transferability probes on the sealed safety family. Utility is the change in macro F1. Route accuracy is computed over 15 target-partition decisions.

Analysis	Result	Control or cost
Distance scale	source gap / bucket spread: 22.33	split half route: 15/15
Bucket rank	raw / equal size / balanced τ: 0.005 / -0.010 / -0.038	split half $\tau = -0.152$
GaussianNB	positive text / audio / fusion: 54.2 / 0.0 / 3.75%	smoothing sweep: 3.3 to 5.0%
Gate route	fixed / global / sample: 11 / 10 / 8 of 15	same branches and updates
Adapted LEEP	flat / routed utility: -0.01540 / 0.00497	harm: 66.7 / 46.7%
Source head LogME	flat / routed utility: -0.00350 / -0.00092	harm: 53.3 / 53.3%

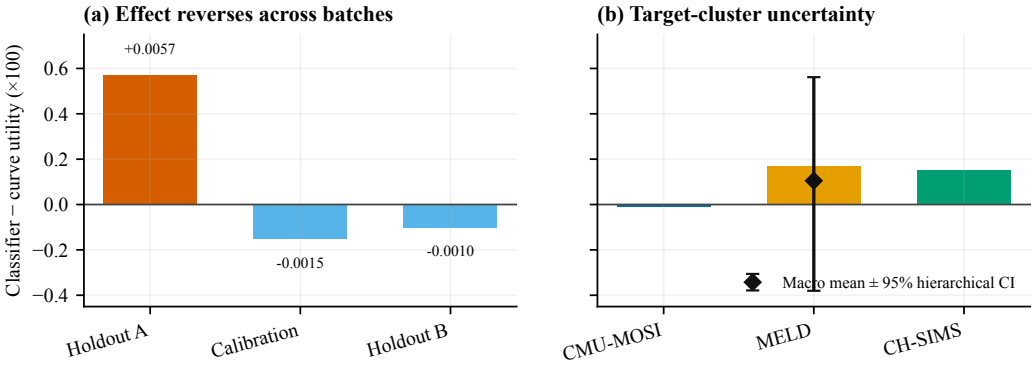

Figure 2: **Selector gains change across repartition families.** (a) The positive-utility classifier exceeds Curve 19 on the first sealed family but not on the calibration or final family. (b) Target-level differences and the hierarchical bootstrap interval, which includes zero.

5.2 APPARENT GAINS REVERSE ACROSS REPARTITIONS

The positive-utility classifier appears promising on the first sealed family. It reduces regret from 0.01780 to 0.01211 relative to Curve 19 and chooses a beneficial bucket in 10 of 15 decisions. That advantage does not repeat. On the next two families, its utility relative to Curve 19 is -0.00153 and -0.00102 (Figure 2). Over all 45 decisions, the mean difference is 0.001047, with a 95% hierarchical interval of $[-0.003804, 0.005616]$ and an exact target-level p value of 0.50.

A later audit provides a simpler reference point. Uniform bucket selection has expected utility 0.00077 and regret 0.01810. The corresponding values are 0.00178 and 0.01727 for the classifier, and 0.00073 and 0.01832 for Curve 19 (Table 4). Curve 19 is therefore no better than uniform selection on these aggregate measures. The uniform baseline was added after the primary outcomes were inspected, so the classifier's small margin over it is descriptive rather than confirmatory.

5.3 CLASSIFIER CONFIDENCE DOES NOT SUPPORT SAFE ABSTENTION

No probability threshold qualifies on the calibration family. From 0.50 to 0.99, accepted harm ranges from 50% to 75%, and realized utility remains negative at every grid point (Figure 3). We therefore lock the fallback threshold at 1.0. On the final family it rejects every decision, avoiding accepted harm only because both coverage and utility are zero.

5.4 UTILITY RANKINGS DEPEND ON THE DOWNSTREAM LEARNER

Logistic regression and GaussianNB produce moderately similar rankings on the matched panel, but they disagree on whether many candidates help (Figure 4). The beneficial share drops from 28.75%

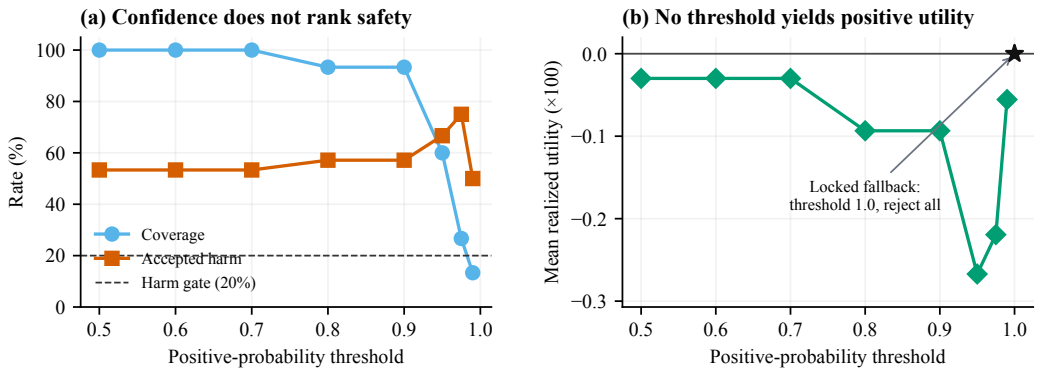

Figure 3: **Classifier probability does not yield a useful rejection threshold.** A higher positive-class threshold reduces coverage without controlling accepted harm or producing positive utility. The locked fallback rejects every candidate.

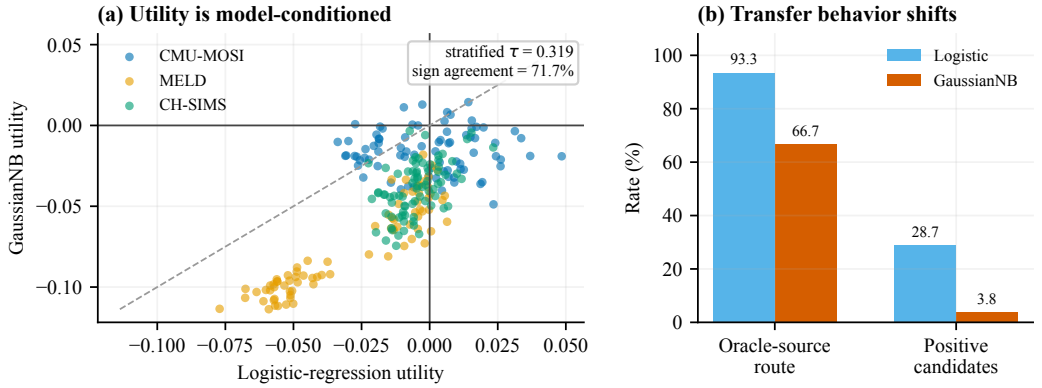

Figure 4: **The downstream learner changes utility and routing.** (a) Candidate utilities retain moderate rank agreement but shift downward under GaussianNB. (b) Oracle source routing and the proportion of beneficial candidates both decline on the matched panel.

under logistic regression to 3.75% under GaussianNB, and route accuracy drops from 14/15 to 10/15. Both frozen bucket selectors choose a harmful GaussianNB candidate in every decision (Table 4).

The loss under GaussianNB is concentrated in audio and is not an artifact of one variance-smoothing value. Across values from 10^{-11} to 10^{-3}, only 3.3% to 5.0% of fused candidates are beneficial. At the locked value, 54.2% of text candidates help, whereas no audio candidate does. Mean audio utility is -0.0884, -0.0806, and -0.0792 on the three targets, so equal-weight fusion can lose even when its text branch improves. Class-conditional audio variance shift is negatively associated with audio utility ($\rho = -0.448$); candidate size is nearly unrelated ($\rho = 0.016$). These checks do not establish Gaussian independence assumptions as the sole cause. They do rule out one unlucky smoothing value and connect the loss to the audio branch and cross-corpus variance shifts.

The MLP retains some source-level structure but little agreement within a source. The distance route matches the oracle source in 12/15 decisions. Logistic and MLP utilities have candidate-level Kendall $\tau = 0.3183$ and 67.92% sign agreement, but Kendall agreement within the routed source is only $\tau = 0.1619$. The frozen positive-utility classifier reaches mean utility $+0.00755$ with 26.7% harm. Its three target means are $+0.02754$, -0.00356, and -0.00132, showing that one target accounts for the aggregate gain. Two CPU runs reproduce all 1,200 utilities exactly, taking 83.55 and 87.86 seconds.

The gate ablation offers a second example of learner dependence. With equal weights, the distance route matches the oracle source in 11 of 15 decisions. Accuracy falls to 10/15 with a learned global

Table 4: Frozen selector decisions across downstream models. Logistic results pool three sealed partition families (45 decisions). GaussianNB, the independent-head neural MLP, and learned gated fusion use the matched final family (15 decisions each). Utility is the change in validation macro F1. Route is oracle-source accuracy and is shared by methods within a model. Expected random and size-only results (†) average ties analytically and are post-hoc descriptive baselines.

Model	Bucket rule	Utility ↑	Positive % ↑	Harm % ↓	Regret ↓	Route % ↑
Logistic	Positive classifier	+0.00178	51.1	48.9	0.01727	97.8
Logistic	Curve-19	+0.00073	40.0	60.0	0.01832	97.8
Logistic	Uniform random†	+0.00077	43.1	56.9	0.01810	97.8
Logistic	Size-only†	-0.00050	40.8	59.2	0.01937	97.8
GaussianNB	Positive classifier	-0.03734	0.0	100.0	0.02773	66.7
GaussianNB	Curve-19	-0.03744	0.0	100.0	0.02783	66.7
Neural MLP	Positive classifier	+0.00755	73.3	26.7	0.01405	80.0
Neural MLP	Curve-19	+0.00486	66.7	33.3	0.01674	80.0
Gated fusion	Positive classifier	-0.00027	46.7	53.3	0.02075	53.3
Gated fusion	Curve-19	-0.00413	26.7	73.3	0.02460	53.3

Table 5: Matched utility comparisons for 240 candidates, with logistic regression as the reference. Candidate-level agreement is higher than the within-source agreement relevant to bucket selection.

Diagnostic	GaussianNB	Neural MLP	Gated fusion
Candidate Kendall τ	0.4222	0.3183	0.3503
Source-stratified Kendall τ	0.3190	0.1619	0.1452
Utility-sign agreement	71.67%	67.92%	63.75%
Positive candidates	3.75%	51.67%	45.00%
Positive candidates, logistic		28.75%	

gate and 8/15 with a sample-dependent gate. Over the same sequence, mean selected utility for the frozen positive-utility selector changes from $+0.00511$ to $+0.00182$ and then -0.00027, while harm rises from 26.7% to 53.3%. Mean absolute gate movement from its anchor value is zero, 0.0074, and 0.0901. The branches, shared head, optimizer, and number of updates are identical across variants. The comparison isolates the change in gate parameterization, although it cannot separate adaptivity from the extra gate parameters. One interpretation is that the adaptive gate changes how source data alter the fitted decision rule, whereas the frozen route measures geometry before that adaptation. Because both encoders remain frozen, this evidence does not extend to end-to-end tuning.

5.5 TRANSFERABILITY BASELINES AND A PRACTICAL COMPROMISE

Without a candidate-specific source model, canonical LEEP and LogME cannot distinguish our buckets. Their candidate-trained versions are also weak flat rankers. On the safety family, adapted LEEP has source-stratified $\tau = 0.098$, utility -0.01540, and 66.7% harm; source-head LogME has $\tau = 0.119$, utility -0.00350, and 53.3% harm. Routing raises adapted LEEP utility to $+0.00497$ with 46.7% harm, while LogME remains negative. These observations informed the staged rule but do not count as confirmation evidence.

On the fresh family, distance selects the oracle source in 14 of 15 decisions. Directly taking adapted LEEP's top bucket gives utility $+0.00272$ and harms 5 decisions. The $k = 2$ procedure reaches safe utility $+0.01084$, with interval $[0.00293, 0.01957]$, and encounters harm once before fallback. Every target mean is positive. It includes 5 global oracles, has regret 0.00417, and avoids 87.5% of full candidate runs, meeting the frozen criterion. At $k = 4$, utility is $+0.01320$, oracle coverage is 10, regret is 0.00181, and 75% of runs are avoided. Exact pilots remain necessary, but only for shortlisted candidates.

6 DISCUSSION AND LIMITATIONS

Prospective scores have a limited but useful role here. Domain proximity can remove an implausible source, but should not choose among nearby buckets, and classifier confidence is not a safety check. A practitioner should first fix the learner and recipe, then route only if split-half geometry is stable and source gaps dominate within-source spread. Pilot the top two candidates and keep the anchor unless validation improves. Four rather than two pilots cost another 12.5 percentage points of runs, but raise oracle coverage from 5/15 to 10/15 and reduce regret from 0.00417 to 0.00181.

These recommendations are bounded by three affective target corpora and frozen representations. The MLP and fusion panels do not substitute for end-to-end tuning. Utility is validation macro F1 for one augmentation recipe, and changes with the model, metric, fusion rule, or budget. Repartitions reuse source examples. The fresh family was utility blind at protocol freeze but still uses the same corpora, so it cannot establish generalization to an unseen corpus.

The evidence does not extend to images or geospatial time series. Sparse video frames retain the affective domain, whereas geospatial data change the prediction units, labels, and temporal dependence. Either extension needs a new benchmark with the same sealed evaluation and matched fusion controls.

7 CONCLUSION

Multimodal dataset utility depends on the learner and selection granularity. For logistic regression, centroid distance routes between corpora but cannot rank buckets within one. GaussianNB and gate ablations show why routes need not transfer across recipes. A practical compromise is to route statically, then train a short list while retaining the option to add no data. Results on the fresh family support this rule at two budgets, but limited oracle coverage makes exact pilots necessary.

AI USE STATEMENT

Generative AI tools were used during hypothesis and method refinement, literature search, translation, initial code and figure drafting, artifact formatting, result interpretation, and language editing. They were not used to create training or evaluation examples, ground-truth labels, mathematical claims, or proofs. The authors checked every citation against a primary record, ran and tested all generated code locally, and matched reported numbers to versioned, machine-readable artifacts. The authors reviewed the assisted work and take responsibility for the final text, claims, code, and artifacts.

ETHICS STATEMENT

This study analyzes existing multimodal sentiment and emotion corpora and recruits no new participants. Raw speech, video, and transcripts remain subject to the licenses and access conditions of the original corpora. Supplementary materials will include only code and derived artifacts allowed under those terms. Affective labels and representations may encode demographic, linguistic, cultural, and annotator biases. A utility ranking is not a judgment about the quality or worth of the people represented in a corpus, and it should not be used to justify high-stakes affect inference. Our conclusions apply only to the audited tasks, representations, models, metrics, and training recipes.

REPRODUCIBILITY STATEMENT

Section 3 defines the prediction problem and its information boundary. Section 4 gives the datasets, representations, signatures, partitions, seeds, downstream recipes, decision rules, uncertainty analysis, and compute counts. Versioned, machine-readable artifacts generate all figures and tables. Subject to corpus licenses, the anonymized supplement will contain code, analysis scripts, and artifact hashes.

REFERENCES

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*, pp. 2236–2246, 2018. doi: 10.18653/v1/P18-1208.

Guanjie Cheng, Boyi Li, Lingyu Sun, Mengying Zhu, Yangyang Wu, Xinkui Zhao, and Shuiguang Deng. TADS: Task-aware data selection for multi-task multimodal pre-training. *arXiv preprint arXiv:2602.05251*, 2026. URL https://arxiv.org/abs/2602.05251. Preprint.

Logan Engstrom, Axel Feldmann, and Aleksander Madry. DsDm: Model-aware dataset selection with datamodels. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12491–12526, 2024.

Talfan Evans, Nikhil Parthasarathy, Hamza Merzić, and Olivier J. Hénaff. Data curation via joint example selection further accelerates multimodal learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-4485.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. DataComp: In search of the next generation of multimodal datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-1179.

Amirata Ghorbani and James Zou. Data shapley: Equitable valuation of data for machine learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2242–2251, 2019.

Long-Kai Huang, Junzhou Huang, Yu Rong, Qiang Yang, and Ying Wei. Frustratingly easy transferability estimation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9201–9225, 2022.

Andrew Ilyas, Sung Min Park, Logan Engstrom, Guillaume Leclerc, and Aleksander Madry. Datamodels: Understanding predictions with data and data with predictions. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 9525–9587, 2022.

Jinhao Jing, Qiannian Zhao, Chao Huang, and Zhan Su. Efficient data selection for multimodal models via incremental optimization utility. *arXiv preprint arXiv:2605.07488*, 2026. URL https://arxiv.org/abs/2605.07488. Preprint.

Cuong V. Nguyen, Tal Hassner, Matthias Seeger, and Cédric Archambeau. LEEP: A new measure to evaluate transferability of learned representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7294–7305, 2020.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 527–536, 2019. doi: 10.18653/v1/P19-1050.

Xuan Qi, Luxi He, Dan Roth, and Xingyu Fu. Demystifying supervision data generalization in multimodal LMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/edcd1aa172dceda2ea9d45a48f25d3e3-Abstract-Conference.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. URL https://arxiv.org/abs/2212.04356.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019. doi: 10.18653/v1/D19-1410.

Naveen Sahi, Jeremy Dohmann, Armen Aghajanyan, and Akshat Shrivastava. SkillRater: Untangling capabilities in multimodal data. *arXiv preprint arXiv:2602.11615*, 2026. URL https://arxiv.org/abs/2602.11615. Preprint.

Prabhant Singh, Sibylle Hess, and Joaquin Vanschoren. How NOT to benchmark your SITE metric: Beyond static leaderboards and towards realistic evaluation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/f0ebc318e2df08360b2df559e81602e5-Abstract-Conference.html.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Mengzhou Xia, Sadhika Malladi, Suchin Gururangan, Sanjeev Arora, and Danqi Chen. LESS: Selecting influential data for targeted instruction tuning. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 54104–54132, 2024.

Kaichao You, Yong Liu, Jianmin Wang, and Mingsheng Long. LogME: Practical assessment of pre-trained models for transfer learning. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12133–12143, 2021.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3718–3727, 2020. doi: 10.18653/v1/2020.acl-main.343.

Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6): 82–88, 2016. doi: 10.1109/MIS.2016.94.

A AUXILIARY FOUR CORPUS TEXT PANEL

An auxiliary text-only experiment adds CMU-MOSEI (Bagher Zadeh et al., 2018) as a fourth target and source. The panel contains 96 target-candidate pairs and 480 utility evaluations. The full text signature reaches target-level Kendall $\tau = 0.1486$, with a descriptive 95% target-cluster interval of $[-0.1884, 0.4855]$ (Table 6). Text centroid distance has the larger point estimate, $\tau = 0.2572$, while the full signature is negatively associated with utility within a source ($\tau = -0.1369$). This is not a multimodal panel, but the additional target does not recover within-source ranking.

Table 6: Auxiliary four-corpus, text-only, leave-one-target-out panel with CMU-MOSEI as a fourth independent target. Brackets give exploratory 95% target-cluster bootstrap intervals over four targets. Lower regret is better.

Rule	Target Kendall τ	Top-1 regret	Within-source τ
Size-only	-0.3233 [-0.5695, -0.0764]	0.03573 [0.02466, 0.04681]	-0.0952
Length-only	+0.1449 [-0.0797, +0.3043]	0.02255 [0.01098, 0.03766]	+0.0536
Text-centroid-only	+0.2572 [+0.0362, +0.4783]	0.02183 [0.00801, 0.03564]	-0.1369
Full text signature	+0.1486 [-0.1884, +0.4855]	0.00583 [0.00180, 0.01022]	-0.1369

B TRANSFERABILITY BASELINE SCOPE

Table 7 shows why canonical LEEP, LogME, and TransRate do not rank buckets when representations are fixed. It also records the extra computation required by each adapted comparator.

Table 7: Scope and cost of prospective scores. Canonical transferability estimators rank models or representations and cannot distinguish buckets that share frozen encoders. The adapted probes change the scored object by fitting heads to each candidate.

Method	Object scored	Candidate fit	Bucket distinction in this benchmark
Static 12-D	candidate data	none	yes
Curve 19	candidate data	40 short updates	yes
Canonical LEEP	pretrained source model	none here	no candidate model is supplied
Canonical LogME	fixed representation	none here	constant across buckets
Canonical TransRate	fixed representation	none here	constant across buckets
Adapted LEEP	candidate-trained heads	2 heads	yes; uses labeled target anchors
Source-head LogME	candidate-trained heads	2 heads	yes; uses labeled target anchors
Exact pilot	downstream candidate	10 heads	yes; five-seed realized utility