NOT EVERY NODE BENEFITS FROM EXPENSIVE RECOVERY: RISK-AWARE ROUTING FOR INCOMPLETE MULTIMODAL GRAPH LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Nodes with missing modalities have limited information for downstream prediction in multimodal graphs, prompting the development of increasingly sophisticated and computationally expensive recovery methods, which are often applied uniformly to affected nodes. However, our analysis shows that expensive recovery can turn correct predictions into errors even when it improves overall performance, while it can still correct cheap-path errors when it degrades overall performance. These motivate selecting nodes for expensive recovery by balancing potential correction against the risk of introducing errors. We propose Risk-aware Selective Modality Recovery (R-SMR), which formulates this selection as budget-constrained routing between cheap and expensive recovery endpoints. Using only cheap-path information at inference, R-SMR estimates the probabilities of four joint correctness states and scores each switch by balancing predicted correction against harm. It selects the highest-scoring nodes with positive utility within a budget for expensive recovery calls, leaving budget unused when further switches are predicted to be unfavorable. Across six datasets and five recovery methods, R-SMR consistently outperforms the stronger endpoint up to 4.10%, with improvements persisting in both globally beneficial and harmful recovery regimes even at a 5% budget. Comparisons with random routing and ablations demonstrate the importance of balancing correction against harm and rejecting unfavorable switches.

1 INTRODUCTION

Multimodal graph learning combines relational structure with complementary node attributes, such as text and images, to support downstream tasks (Zhu et al., 2025). In practice, nodes may lack one or more modalities; recovery methods exploit observed attributes and graph structures to compensate for missing modalities (Hu et al., 2026). Recent advances, such as propagation-based imputation (Um et al., 2025), graph denoising diffusion (Liu et al., 2026), and retrieval-enhanced completion (Li et al., 2026), provide increasingly elaborate recovery methods, yet their aggregate performance does not indicate whether replacing a cheap recovery endpoint benefits each individual node. Recent work on multimodal graph learning shows that for some nodes with missing modalities, simple and cheap recovery approaches using observed modality information and neighborhood context can still achieve great performance (Sun et al., 2026; Um et al., 2025). The elaborate and expensive recovery must therefore be assessed by its effect on downstream tasks. This raises a question: **which nodes benefit from replacing a cheap recovery with a more expensive one?**

Our node-level analysis reveals that *expensive recovery can correct cheap-path errors while also turning correct predictions into errors.* Figure 1 illustrates how net accuracy varies across recovery methods and soft-homophily bins (defined in Section 5.1.1) with results for the remaining four datasets shown in Figure 7 (Appendix A). Beyond these net changes, Figure 2 shows that *correction and harm coexist* within the same bins. Consequently, recovery methods that improve overall accuracy can still harm individual nodes, whereas recovery methods that reduce overall accuracy can still benefit a subset of nodes. Applying expensive recovery uniformly may therefore introduce avoidable errors, while rejecting it entirely may overlook useful correction opportunities. These observations **motivate node-specific recovery decisions that account for both the potential to correct errors and the risk of introducing new ones.**

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.

We formulate this decision as *budget-constrained node routing between a predefined cheap and expensive recovery method*. For each node with a missing modality, the policy chooses whether to retain the cheap-path prediction or invoke expensive recovery, to improve overall performance without exceeding the expensive recovery calls budget. Making these routing decisions requires estimating whether a node will benefit from expensive recovery before it is invoked, without access to ground-truth labels or expensive-recovery outputs at inference. Our diagnostic comparing feature propagation (FP) with the more expensive RITR recovery method (Tu et al., 2025) reveals only a weak association between reconstruction improvement and downstream task benefit in the examined setting (Section 4.2). This mismatch motivates *estimating the downstream task gains, rather than the reconstruction benefit*, when switching to expensive recovery from cheap-path information. Meanwhile, ranking nodes solely by correction probability overlooks potential harm, while treating the budget as a quota can admit expensive calls with no expected net benefit. **Effective routing must therefore balance potential *correction* against *harm* and invoke expensive recovery only when the expected gains outweigh the risks.**

We propose Risk-aware Selective Modality Recovery (R-SMR), which predicts the utility of switching to expensive recovery from cheap-path information alone. During training, predictions from a recovery method are compared with ground-truth labels to define four joint correctness states: *both correct*, *expensive only correct*, *cheap only correct*, and *both wrong*. The router learns these state probabilities from cheap-path features. At inference, the calibrated probabilities yield a risk-adjusted score—the predicted probability of correction minus the weighted probability of harm. R-SMR discards nodes with non-positive scores and routes the highest-scoring remaining nodes to expensive recovery up to the call budget; all other nodes retain their cheap-path predictions.

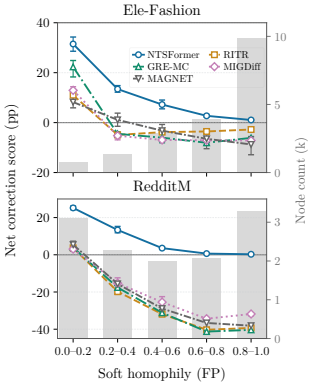


Figure 1: Net correction score of five expensive recovery methods vs. cheap recovery.

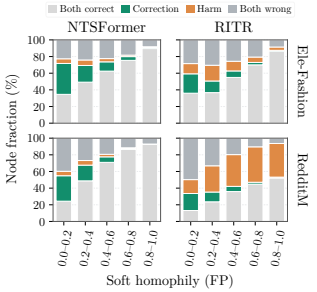


Figure 2: Correction and harm coexist within soft-homophily bins. Full diagnostics appear in Figure 8.

Contributions. Our contributions can be summarized as the following three points:

- We conduct an empirical analysis of modality recovery, showing that node-level correction and harm coexist whether expensive recovery improves or degrades overall classification performance. An FP–RITR diagnostic further shows that reconstruction gains are only weakly associated with downstream task benefit, motivating task-aware node selection.
- We formulate selective modality recovery as budget-constrained node routing and propose R-SMR, which predicts four joint correctness probabilities from cheap-path information alone at inference. R-SMR balances predicted correction against harm, allocating expensive recovery to selected nodes without necessarily exhausting the budget.
- We evaluate R-SMR across five recovery methods, six datasets, and five call budgets. R-SMR outperforms the stronger recovery endpoint by up to 4.10% and improves on cheap recovery even at a 5% call budget. Matched-call comparisons and ablations further support harm-aware node selection and positive-utility filtering.

2 Related Work

We briefly discuss the most relevant work here and provide more detailed reviews in Appendix B.

Recovery for Incomplete Multimodal Graphs. Learning with missing modalities leverages observed node attributes and graph structure for downstream tasks. Propagation-based recovery includes Feature Propagation (Rossi et al., 2022) and extensions incorporating confidence or feature injection (Um et al., 2023; 2025). R-SMR adopts masked neighbor-mean propagation as its cheap recovery path. More elaborate methods use iterative refinement (Tu et al., 2025), graph denoising

diffusion (Liu et al., 2026), or retrieval-enhanced modality completion (Li et al., 2026), offering candidate expensive recovery endpoints. However, additional computation is worthwhile only when it improves node-level predictions. Aggregate performance cannot reveal this benefit: expensive recovery may correct a cheap-path error or turn a correct prediction into an error. R-SMR therefore learns when switching to expensive recovery is likely to improve a node's prediction.

Budget-Constrained Node Routing. Adaptive prediction, cascades, and learning to defer route inputs among models or experts to balance prediction cost and accuracy (Nan & Saligrama, 2017; Enomoro & Eda, 2021; Mozannar & Sontag, 2020). Related work on graph architectures and multimodal mixture-of-experts adapts computation to node characteristics and modality availability (Wang et al., 2022; 2023; Yun et al., 2024). A closely related method, GLANCE (Loveland et al., 2026), uses inexpensive node features to decide when to query an LLM to refine GNN predictions. R-SMR addresses this allocation problem for predefined recovery endpoints: it separately models the chances of correcting an error and harming a correct prediction, then selects switches with positive estimated utility subject to an upper limit on expensive calls.

3 PRELIMINARIES

Let $G = (V, E)$ denote a multimodal graph with node set V, edge set $E \subseteq V \times V$, and modality set $\mathcal{M}$. Each node v has modality-specific features $\{x_v^m\}_{m\in\mathcal{M}}$, whose availability is indicated by $o_v^m \in \{0, 1\}$: $o_v^m = 1$ denotes an observed modality, whereas $o_v^m = 0$ denotes a missing modality. We assume that each node lacks at most one modality, i.e., $\sum_{m\in\mathcal{M}}(1 - o_v^m) \leq 1$ for all $v \in V$. The task is to predict node labels from graph structure, observed features, and modality availability.

Modality Recovery Endpoints. A recovery endpoint pairs modality recovery with a downstream classifier to predict node labels. We consider two predefined endpoints: a cheap endpoint C that provides the default prediction and an expensive endpoint E that uses additional recovery computation. For node v, endpoint $k \in \{C, E\}$ produces logits z_v^k and prediction $\hat{y}_v^k = \arg\max_c z_{v,c}^k$. This formulation accommodates different recovery methods and classifiers; "expensive" denotes higher computational cost without implying greater accuracy. Both endpoints remain frozen during router training, so the router learns only whether to replace the cheap prediction with the expensive one.

Selective Missing-Modality Recovery. The routing candidate set consists of all nodes with exactly one missing modality and is defined as

$$V_{\text{miss}} = \{v \in V : \sum_{m\in\mathcal{M}} (1 - o_v^m) = 1\}, \qquad N = |V_{\text{miss}}|.$$

For each candidate, the policy chooses $a_v \in \{0, 1\}$ to retain the cheap prediction ($a_v = 0$) or switch to the expensive prediction ($a_v = 1$). Given a call-rate budget $b \in [0, 1]$, it seeks to improve classification performance subject to $\sum_{v\in V_{\text{miss}}} a_v \leq \lceil bN \rceil$. At inference, routing uses only information available before expensive recovery, excluding test labels and expensive-endpoint outputs. The router may select any number of nodes up to this limit.

4 WHICH NODES BENEFIT FROM EXPENSIVE RECOVERY?

We first examine how aggregate endpoint performance obscures node-level correction and harm. We then test whether reconstruction improvements identify beneficial switches. Finally, we assess the potential gains from task-aware node selection while holding the calls of expensive recovery fixed.

4.1 UNIFORM RECOVERY OBSCURES NODE-LEVEL CORRECTION AND HARM

We retrospectively compare predictions from the fixed cheap and expensive endpoints with ground-truth labels for test nodes with missing modalities. Switching to the expensive endpoint yields correction when it fixes a cheap-path error and harm when it turns a correct prediction into an error. Correctness remains unchanged when both endpoints are correct or both are wrong. For each dataset with N candidates, the net correction score is

$$\text{NCS} = \frac{N_{\text{correction}} - N_{\text{harm}}}{N}.$$

Findings. Correction and harm coexist across all five operators (Table 1). In the aggregated results, NTSFormer achieves positive NCS yet harms 2.82% of candidate records, whereas the other four operators correct 5.78–6.96% of records despite negative NCS. Correction and harm also coexist within groups defined by node properties (Figure 2; complete diagnostics in Figure 8). These findings show that aggregate endpoint performance alone is insufficient to guide individual recovery decisions, motivating **node selection that balances correction opportunities against harm risks**.

Table 1: Correction and harm across five recovery operators, computed from missing-modality test records aggregated across six datasets.

Operator	Correction	Harm	NCS (pp)
NTSFormer	8.57%	2.82%	+5.75
RITR	5.78%	21.55%	-15.77
GRE-MC	6.38%	21.91%	-15.53
MIGDiff	6.68%	18.51%	-11.83
MAGNET	6.96%	22.73%	-15.77

4.2 Does Better Reconstruction Identify Beneficial Switches?

We test whether reconstruction improvement identifies beneficial switches by comparing feature propagation (FP) with RITR (Tu et al., 2025) under the same frozen graph convolutional network (GCN) (Kipf & Welling, 2017). For each missing-modality test candidate v, we define

$$B_{\mathrm{rec}}(v) = \cos(x_v^\star, \hat{x}_v^E) - \cos(x_v^\star, \hat{x}_v^C), \qquad B_{\mathrm{task}}(v) = \mathrm{CE}_C(v) - \mathrm{CE}_E(v),$$

where $x_v^\star$ is the true missing feature and C/E denote FP/RITR, respectively. Both scores require ground-truth features or labels and expensive outputs, so they serve as retrospective diagnostics. Negative B_{task} indicates increased cross-entropy, not necessarily a correct-to-wrong switch.

Findings. At a 20% selection rate, the overlap between reconstruction- and task-ranked selections ranges from 19.9% to 27.9%, while 56.0–71.6% of reconstruction-selected candidates have negative task benefit (Table 2). Within this FP–RITR diagnostic, **reconstruction improvement does not reliably identify switches that reduce classification loss, even when measured against the true missing features**. Additional budgets and correctness-state diagnostics appear in Appendix C.

Table 2: Reconstruction-based selection at a 20% selection rate.ρ_b denotes Spearman correlation within the selected Top-k set.

Metric	Grocery	Toys	Movies
Top-k ρ_b	0.10 ± 0.01	0.11 ± 0.02	0.04 ± 0.01
Overlap (%)	27.90 ± 3.60	27.50 ± 2.50	19.90 ± 3.60
Neg. task (%)	67.60 ± 2.30	71.60 ± 3.20	56.00 ± 1.20
Task gain	62.92 ± 5.61	78.10 ± 8.35	67.37 ± 3.11
Rec. gain	-3.48 ± 1.61	-8.53 ± 2.95	-1.87 ± 3.84

4.3 Opportunities for Task-Aware Node Selection

We evaluate the potential benefit of task-aware selection using the same FP–RITR pair. A label-aware Oracle selects missing-modality node candidates with the highest B_{task}, while Random selects the same number uniformly at random. For each policy, we replace the FP features of selected nodes with RITR features and rerun the shared frozen GCN on the full graph.

Table 3: Label-aware Oracle versus Random for FP $\rightarrow$ RITR at a 20% selection rate. Accuracy and Macro-F1 are reported in percentages; Oracle–Random differences are in percentage points (pp).

Dataset	FP Acc.	Oracle Acc.	Random Acc.	FP F1	Oracle F1	Random F1	Oracle−Random Acc./F1 (pp)
RedditS	92.67	93.50	92.66	86.52	87.81	86.50	+0.84 / +1.31
Grocery	76.24	76.83	76.14	61.81	62.45	61.63	+0.69 / +0.82
Toys	76.97	77.73	76.91	71.81	72.64	71.69	+0.82 / +0.95
Movies	49.07	50.28	48.99	31.65	32.35	31.01	+1.30 / +1.34
RedditM	68.13	69.49	68.29	61.91	63.42	62.16	+1.20 / +1.26
Ele-Fashion	80.11	80.80	80.08	56.16	57.08	55.93	+0.72 / +1.14

Findings. At a 20% selection rate, the Oracle outperforms matched Random on all six datasets by up to 1.30% in Accuracy and 1.34% in Macro-F1 (Table 3). Results for additional budgets appear in Appendix D. These gains **demonstrate the value of task-aware node selection**. Because the Oracle relies on labels, this comparison does not establish whether a deployable router can achieve the same gains. Together, these diagnostics motivate learning to estimate correction and harm from cheap-path information before invoking expensive recovery.

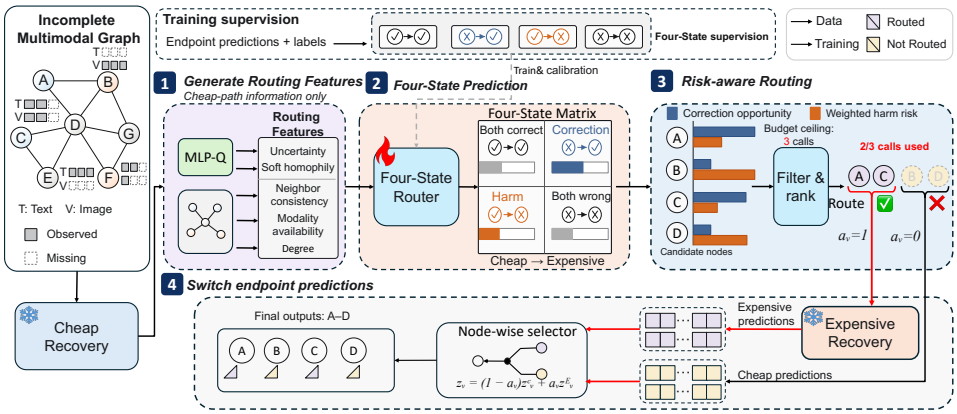


Figure 3: **R-SMR Overview.** *Step 1:* R-SMR constructs routing features from cheap-path information. *Step 2:* The router estimates four joint correctness states of the cheap and expensive endpoints. *Step 3:* Nodes are ranked by correction probability minus weighted harm probability, and positive-score candidates are selected up to the budget ceiling. *Step 4:* Selected nodes use expensive-endpoint predictions; unselected nodes retain cheap predictions.

5 How to Route: Risk-Aware Selective Modality Recovery

We propose Risk-aware Selective Modality Recovery (R-SMR), which selectively switches nodes between cheap and expensive recovery endpoints under an expensive-call budget. As illustrated in Figure 3, R-SMR leverages cheap-path information to estimate the endpoints' joint correctness and allocates expensive calls by balancing predicted correction opportunities against harm risks.

5.1 Components of R-SMR

We detail the four-step workflow below, from routing feature construction to endpoint switching. Router training and calibration are described in Section 5.2.

5.1.1 Step 1: Routing Feature Construction

Cheap-path information. Cheap recovery uses masked neighbor-mean propagation to produce features $\mathbf{x}_v^C$ for default prediction and routing. From these features, the downstream classifier generates default logits z_v^C, while a separately fitted MLP-Q predictor estimates class posteriors $\mathbf{p}_v^Q$. Together with graph structure and modality-observation masks, the recovered features and MLP-Q posteriors provide the information needed to construct routing features before expensive recovery.

Routing Features. From this cheap-path information, we construct routing features that capture predictive uncertainty, neighborhood agreement, feature consistency, modality availability, and degree. We first quantify uncertainty and neighborhood agreement using predictive entropy and soft homophily derived from the MLP-Q posteriors. With $\mathcal{N}(v)$ denoting the one-hop neighborhood and $d_v = \max\{|\mathcal{N}(v)|, 1\}$, these quantities are defined as:

$$H_v = -\sum_c p_{v,c}^Q \log p_{v,c}^Q, \qquad h_v = (\mathbf{p}_v^Q)^\top \left(\frac{1}{d_v} \sum_{u \in \mathcal{N}(v)} \mathbf{p}_u^Q \right).$$

Beyond soft homophily, we also measure local consistency in the recovered feature space using mean neighbor similarity:

$$s_v = \frac{1}{d_v} \sum_{u \in \mathcal{N}(v)} (\tilde{\mathbf{x}}_u^C)^\top \tilde{\mathbf{x}}_v^C, \qquad \tilde{\mathbf{x}}_v^C = \frac{\mathbf{x}_v^C}{\max\{\|\mathbf{x}_v^C\|_2, \epsilon\}},$$

where $\epsilon > 0$ prevents division by zero. For a missing modality m, the fraction of neighbors with that modality originally observed is

$$\rho_{v,m} = \frac{1}{d_v} \sum_{u \in \mathcal{N}(v)} o_u^m.$$

For each candidate v, let m_v denote its unique missing modality; we use $\rho_v = \rho_{v,m_v}$. With missingness type τ_v, the raw routing vector is

$$\phi_v^{\text{raw}} = \big[h_v, H_v, s_v, d_v, \log(1+d_v), \rho_v, \text{onehot}(\tau_v)\big].$$

We obtain ϕ_v by median-imputing continuous variables, adding missing-value indicators, and standardizing the resulting numerical features. All preprocessing parameters are fitted on router-fitting data and then frozen, while every input is available before expensive recovery. Appendix E.1 provides the cheap-path configuration and preprocessing details.

5.1.2 Step 2: Four-State Prediction

Using the routing features ϕ_v, we estimate whether switching to the expensive endpoint will correct an error, introduce an error, or leave correctness unchanged. Following Section 4.1, we describe these outcomes through the joint correctness of the cheap and expensive endpoints: $\mathcal{T} = \{(1,1),(0,1),(1,0),(0,0)\}$, where the first and second entries indicate cheap- and expensive-endpoint correctness, respectively. The four states correspond to both correct, correction, harm, and both wrong. Modeling them jointly separates correction opportunities from harm risks while accounting for switches that do not change correctness.

To estimate these outcomes before expensive recovery, a multinomial logistic-regression router maps ϕ_v to four logits, $\ell_v = W\phi_v + \beta$. We then apply temperature scaling (Guo et al., 2017) with a scalar $T > 0$ to obtain the state probabilities:

$$\widehat{P}_v(t) = \frac{\exp(\ell_{v,t}/T)}{\sum_{t' \in \mathcal{T}} \exp(\ell_{v,t'}/T)}.$$

The router and temperature are fitted using development data and frozen before inference. The resulting correction and harm probabilities provide the basis for the switching utility defined next.

5.1.3 Step 3: Risk-Aware Routing

Switching Utility. To turn state probabilities into routing decisions, we assign a unit reward to correction, a penalty $\lambda \geq 0$ to harm, and zero utility to unchanged correctness:

$$U_\lambda(v) = \mathbb{I}[t_v = (0,1)] - \lambda \mathbb{I}[t_v = (1,0)].$$

Since the realized state is unknown at inference, we estimate its expected utility using the calibrated probabilities:

$$\widehat{u}_v(\lambda) = \widehat{P}_v((0,1)) - \lambda \widehat{P}_v((1,0)).$$

At $\lambda = 1$, the score estimates the net change in correctness; larger values penalize harm more strongly. This utility concerns correctness rather than the cross-entropy benefit in Section 4.2.

Budget-Constrained Selection. We use these scores to allocate expensive recovery calls among the N candidates in V_{miss}. With each selected node counted as one call, a nominal call-rate limit b gives the ceiling $K_b = \lceil bN \rceil$:

$$\max_{\substack{a_v \in \{0,1\} \\ \sum_{v \in V_{\text{miss}}} a_v \leq K_b}} \sum_{v \in V_{\text{miss}}} a_v \widehat{u}_v(\lambda).$$

The solution selects up to K_b nodes with the highest positive scores, breaking ties by ascending node identifier. Thus, the budget remains partly unused when fewer candidates have positive scores; if none do, all cheap predictions are retained. The selected set $S_b = \{v : a_v = 1\}$ is fixed before expensive outputs are accessed. This rule optimizes predicted utility without guaranteeing realized gains or directly optimizing Macro-F1.

5.1.4 Step 4: Endpoint Prediction Switching

Once S_b is determined, R-SMR uses expensive-endpoint predictions for selected nodes and retains cheap predictions for the remaining nodes. Writing this decision as $a_v = \mathbb{I}[v \in S_b]$, we obtain the final logits and predicted label:

$$z_v = (1 - a_v) z_v^C + a_v z_v^E, \qquad \hat{y}_v = \arg\max_c z_{v,c}.$$

Because a_v is binary, each node uses the logits of a single endpoint rather than a mixture of both. This preserves the cheap prediction whenever a node is not selected, without introducing a learned fusion head. The rule specifies which predictions are replaced, while the computation required to obtain expensive predictions remains endpoint-specific.

5.2 Router Training and Calibration

Four-State Supervision. We construct router supervision by comparing the two fixed endpoints' predictions with the label y_v of each router-training candidate:

$$t_v = \big(\mathbb{I}[\hat{y}_v^C = y_v], \mathbb{I}[\hat{y}_v^E = y_v]\big).$$

Router Fitting. Using these targets, we fit W and β by minimizing L_2-regularized multiclass cross-entropy on router-training data, while keeping both endpoints and the separately fitted MLP-Q fixed. Preprocessing parameters are fitted on the same data and reused unchanged for calibration and evaluation.

Calibration and Policy Selection. With the fitted router fixed, we estimate $T > 0$ by minimizing state negative log-likelihood on a reserved development calibration subset. We then select λ using development data, keeping the positive-score threshold at zero. All models, preprocessing parameters, T, and λ are frozen before final evaluation, so test-time selection relies solely on cheap-path information without accessing test labels or expensive outputs before selecting nodes. Appendix E.2 provides optimization settings, calibration details, and policy-selection criteria. The inference-time and space complexity of routing are analyzed in Appendix E.3.

6 Empirical Analysis

We evaluate R-SMR across recovery operators and datasets, first comparing selective recovery with the fixed endpoints and matched-count random routing. We then examine performance across routing budgets and assess the contributions of the routing components and features.

6.1 Experimental Setup

Datasets and Missingness Settings. We evaluate node classification on six multimodal graphs in different domains: Grocery, Toys, Movies, RedditS, and RedditM from MAGB (Yan et al., 2025), and Ele-Fashion from MM-Graph (Zhu et al., 2025). The MAGB datasets use label-stratified 60%/20%/20% training/validation/test splits, while Ele-Fashion retains its provided split of approximately 60%/10%/30%. For each graph, we randomly mask textual and visual features on disjoint node subsets at a missing rate of 40% per modality. Each node therefore lacks at most one modality, and nodes missing exactly one modality form the routing candidate set. Dataset statistics and split details are provided in Appendix F.

Baselines. We pair R-SMR with five advanced expensive recovery operators: NTSFormer (Hu et al., 2026), RITR (Tu et al., 2025), GRE-MC (Li et al., 2026), MAGNET (Tabakhi et al., 2025), and MIGDiff (Liu et al., 2026). For each endpoint pair, All-cheap and All-expensive assign all node candidates to the cheap and expensive endpoints, respectively. Both random-routing baselines sample candidates uniformly: Rand-b_{eff} matches R-SMR's realized call count, whereas Rand-$b_{\max}$ exhausts the nominal budget.

Training and Evaluation. We follow the prescribed data splits, using training data for fitting and validation data for calibration and hyperparameter selection of R-SMR. For R-SMR, routing policies share the same splits and frozen endpoints. We evaluate Accuracy and Macro-F1 on test nodes missing either modality and report means across 5 seeds, with standard deviations. Cross-dataset summaries equally weight the dataset means. The main comparison uses a 50% nominal call-budget cap. Additional budgets, random-sampling procedures and implementation details are provided in Appendix G. The code will be open-sourced after the review process.

Table 4: Node classification performance (%) under a 50% nominal cap on expensive recovery calls. b_{eff} is the mean percentage of missing-modality test candidates selected for expensive recovery. For b_{eff}, red indicates reaching the budget cap, whereas green indicates there is still remaining budget.

Operator	Dataset	All-cheap		All-expensive		R-SMR		
		Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1	b_{eff} (%)
NTSFormer	Ele-Fashion	79.29 ± 0.20	61.64 ± 0.39	83.68 ± 0.23	67.54 ± 0.49	**84.63 ± 0.20**	**67.96 ± 0.25**	50.00
	Grocery	74.76 ± 0.81	61.90 ± 0.93	80.54 ± 0.48	70.97 ± 1.66	**82.70 ± 0.56**	**72.95 ± 1.93**	50.01
	Movies	49.54 ± 1.18	35.54 ± 2.81	51.73 ± 1.53	44.56 ± 2.10	**52.69 ± 1.23**	**46.32 ± 2.38**	50.01
	RedditM	66.64 ± 0.68	61.67 ± 0.86	75.88 ± 0.31	72.14 ± 0.31	**76.86 ± 0.22**	**73.79 ± 0.21**	50.00
	RedditS	90.77 ± 0.25	84.26 ± 0.64	93.66 ± 0.47	88.73 ± 1.03	**93.77 ± 0.47**	**88.75 ± 0.91**	50.01
	Toys	75.18 ± 1.22	70.86 ± 1.16	78.81 ± 0.92	75.29 ± 1.01	**79.34 ± 0.93**	**75.94 ± 1.18**	50.00
GRE-MC	Ele-Fashion	79.30 ± 0.14	61.75 ± 0.48	74.16 ± 0.78	50.04 ± 0.30	**81.20 ± 0.01**	**63.58 ± 0.46**	12.88
	Grocery	75.20 ± 0.64	61.98 ± 1.42	53.23 ± 2.23	39.98 ± 1.62	**77.38 ± 0.51**	**64.01 ± 1.33**	3.77
	Movies	49.91 ± 0.25	36.51 ± 0.46	39.27 ± 1.01	8.47 ± 1.81	**51.88 ± 0.24**	**37.43 ± 0.49**	3.23
	RedditM	66.49 ± 0.74	61.56 ± 0.90	42.27 ± 0.07	41.81 ± 0.63	**69.59 ± 0.47**	**64.13 ± 0.66**	15.90
	RedditS	90.68 ± 0.15	83.88 ± 0.22	58.01 ± 0.88	55.27 ± 0.43	**91.19 ± 0.37**	**84.52 ± 0.95**	2.36
	Toys	75.64 ± 0.13	71.34 ± 0.91	44.27 ± 1.00	39.72 ± 1.14	**77.59 ± 0.31**	**73.25 ± 1.06**	3.84
MAGNET	Ele-Fashion	80.18 ± 0.07	62.10 ± 0.55	74.62 ± 3.24	47.19 ± 9.81	**81.36 ± 0.71**	**63.52 ± 1.79**	17.23
	Grocery	75.20 ± 0.64	61.98 ± 1.42	27.21 ± 22.60	13.39 ± 20.90	**76.82 ± 0.64**	**63.98 ± 1.42**	2.71
	Movies	49.86 ± 0.24	36.54 ± 1.74	32.93 ± 0.00	2.48 ± 0.00	**51.97 ± 0.28**	**38.67 ± 1.72**	3.41
	RedditM	68.61 ± 0.36	63.69 ± 0.28	47.39 ± 1.08	45.45 ± 0.64	**72.71 ± 0.09**	**67.70 ± 0.24**	18.93
	RedditS	91.48 ± 0.77	85.69 ± 0.47	56.68 ± 2.67	51.01 ± 1.22	**92.32 ± 0.44**	**86.31 ± 0.35**	3.63
	Toys	76.40 ± 0.29	71.80 ± 0.64	56.69 ± 1.42	47.63 ± 1.54	**78.75 ± 0.37**	**73.72 ± 0.64**	4.81
RITR	Ele-Fashion	79.18 ± 0.10	61.13 ± 0.76	76.54 ± 0.34	52.80 ± 0.86	**80.71 ± 0.17**	**62.47 ± 0.53**	7.59
	Grocery	75.08 ± 0.60	61.98 ± 0.79	41.86 ± 1.11	26.48 ± 1.10	**77.66 ± 0.65**	**63.21 ± 0.73**	4.43
	Movies	49.89 ± 0.83	36.61 ± 1.28	40.53 ± 0.34	9.04 ± 0.33	**51.91 ± 0.82**	**37.43 ± 1.29**	3.88
	RedditM	66.78 ± 0.63	61.80 ± 0.99	42.45 ± 0.69	42.41 ± 0.69	**69.73 ± 0.40**	**64.64 ± 0.72**	13.88
	RedditS	90.82 ± 0.38	84.52 ± 0.69	37.04 ± 3.38	19.90 ± 1.42	**92.49 ± 0.43**	**85.23 ± 0.71**	3.05
	Toys	76.32 ± 0.61	72.15 ± 0.83	47.17 ± 1.80	34.86 ± 1.89	**77.03 ± 0.62**	**74.02 ± 0.83**	3.42
MIGDiff	Ele-Fashion	79.45 ± 0.32	62.99 ± 0.23	73.36 ± 0.42	48.25 ± 0.48	**80.17 ± 0.29**	**63.20 ± 0.04**	9.61
	Grocery	75.38 ± 0.34	61.87 ± 1.66	60.35 ± 3.16	44.88 ± 2.10	**75.97 ± 0.30**	**62.43 ± 1.67**	2.45
	Movies	49.74 ± 0.34	36.57 ± 0.56	41.63 ± 0.52	12.10 ± 0.71	**50.81 ± 0.34**	**37.48 ± 0.60**	5.03
	RedditM	67.46 ± 0.64	62.37 ± 0.91	47.70 ± 0.02	44.20 ± 0.45	**70.52 ± 0.27**	**65.14 ± 0.55**	12.58
	RedditS	91.50 ± 0.61	84.98 ± 1.15	76.64 ± 1.36	66.72 ± 1.17	**92.78 ± 0.35**	**86.08 ± 0.92**	2.40
	Toys	75.47 ± 1.79	71.45 ± 1.33	63.13 ± 1.44	55.16 ± 1.41	**75.69 ± 1.80**	**72.11 ± 1.28**	2.12

6.2 Can R-SMR Outperform Both Endpoints?

Under a 50% nominal budget cap on expensive recovery calls, **R-SMR consistently outperforms both frozen endpoints across all 30 operator–dataset combinations** (Table 4). Relative to the globally stronger frozen endpoint in each combination, R-SMR **achieves improvements of up to 4.10% and 4.01% in Accuracy and Macro-F1, respectively**. Since both endpoints remain frozen, these gains indicate that R-SMR exploits node-level predictive complementarity that aggregate endpoint rankings fail to capture. This advantage holds across both globally beneficial and harmful expensive endpoints, demonstrating that selective allocation surpasses the stronger uniform recovery policy while constraining expensive calls. Per-dataset results across budgets are provided in Appendix H.

6.3 How Does R-SMR Respond to Different Budget Caps?

For the ***globally beneficial endpoint***, such as NTSFormer, R-SMR yields monotonic performance improvements as the nominal expensive-call cap increases from 5% to 50%, with realized call rates closely tracking the designated cap (Table 9 in Appendix H). At the 50% cap, **R-SMR outperforms the All-expensive baseline by 0.95% and 1.08% in Accuracy and Macro-F1, respectively, while fully utilizing the expensive-call budget across all six datasets** (Table 4). This selective allocation strategy thus achieves superior performance with approximately half the expensive calls, underscoring its efficacy even when the endpoint confers universal benefits.

For the ***globally harmful endpoints***, such as MIGDiff, GRE-MC, MAGNET, and RITR, **R-SMR consistently outperforms the All-cheap baseline across all prespecified budgets. Under the 50% cap, realized call rates span merely 2.12%–18.93% across all 24 operator–dataset combinations**, with each combination surpassing All-cheap (Table 4). Both dataset-averaged performance and call rates largely plateau beyond a 20% cap (Table 9), indicating that these gains are robust to expanded budget availability. R-SMR thus extracts localized benefits from endpoints that are deleterious in aggregate, while leaving substantial budget capacity unutilized.

6.4 Can R-SMR Outperform Random Routing across Budgets?

We compare R-SMR with call-matched Rand-b_{eff} and budget-exhausting Rand-$b_{\max}$ across five operators, six datasets, and five budgets. **R-SMR consistently outperforms both call-matched Rand-b_{eff} and budget-exhausting Rand-$b_{\max}$ across all operator–dataset–budget combinations.** Relative to Rand-b_{eff}, R-SMR achieves gains of 8.10% in Accuracy and 7.62% in Macro-F1; against Rand-$b_{\max}$, the improvements reach 28.57% and 26.52%, respectively. The matched-call comparisons demonstrate that R-SMR allocates expensive recovery operations more efficiently under identical frozen endpoints, thereby delivering superior performance at equivalent expensive-call counts. Selected Accuracy results appear in Figure 4, and complete results appear in Appendix I.

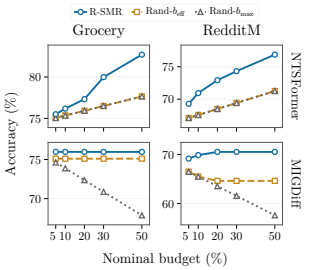


Figure 4: Accuracy of NTSFormer and MIGDiff on Grocery and RedditM across five nominal budgets.

6.5 Which Routing Components and Features Matter?

Risk-aware components. At a 20% nominal expensive-call cap, **removing the harm penalty from the four-state router and using a binary correction-only predictor incur performance losses of 3.72% and 3.67%**, respectively, averaged across five operators, six datasets. Both variants underperform Full R-SMR in all 60 comparisons. Retaining positive-score filtering improves mean performance by 2.23%, with gains in 58 of the 60 comparisons. These results demonstrate the value of balancing correction against harm and rejecting switches with non-positive predicted utility (Figure 5; complete results in Appendix J, Figure 14).

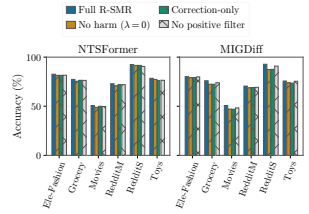


Figure 5: Accuracy ablations for NTSFormer and MIGDiff across six datasets at 20% budget.

Routing features. At a 50% nominal expensive-call cap, we remove one numerical feature family at a time and refit the router, retaining the missing-modality indicator and frozen endpoints. **Full R-SMR exceeds all five feature-removal variants** across the four settings in Figure 6, with gains ranging from 0.83% to 3.45%. Removing soft homophily yields the largest decline in every displayed setting, reducing Accuracy on RedditM by 2.33% for NTSFormer and 3.45% for MIGDiff. The complete results in Appendix K include smaller and reversed effects, showing that feature contributions depend on the operator and dataset.

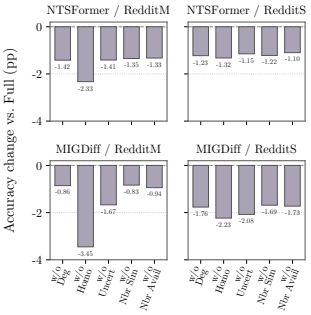


Figure 6: Routing-feature sensitivity at 50% call ceiling.

7 Conclusion

In this work, we investigated how expensive recovery can be selectively applied to improve incomplete multimodal graph learning. Through node-level analysis, we showed that correction and harm coexist under expensive recovery, revealing recovery opportunities obscured by aggregate performance. Building on this finding, we proposed R-SMR, which leverages cheap-path information to balance predicted correction against harm, selecting promising nodes and adapting actual budget use to their expected benefit. Experiments across six datasets and five advanced recovery methods demonstrate that R-SMR outperforms both frozen endpoints across all 30 combinations under a 50% nominal call-budget cap, with gains over the stronger endpoint reaching 4.10%. Even for globally harmful endpoints, R-SMR improves predictions while invoking expensive recovery on only 2.12%–18.93% of candidate nodes. Our findings establish risk-aware selection and adaptive budget allocation as an effective complement to recovery-model development, showing that existing recovery methods deliver greater predictive value when applied to the right nodes.

AI USE STATEMENT

We used generative AI tools to assist with translation, language editing, and manuscript formatting. The authors reviewed the AI-assisted manuscript and take full responsibility for the final content of this work, including its text, claims, and presentation.

ETHICS STATEMENT

Our study uses publicly available multimodal graph datasets from existing benchmarks. We develop a selective modality recovery framework that balances potential prediction gains against the risk of introducing errors under a limited recovery budget. Our evaluation examines both beneficial and harmful effects of recovery, with experimental protocols and implementation details provided in the main paper and appendix. The method and planned code release are intended to support academic research on incomplete multimodal graph learning. Applications involving personal data or consequential decisions would require further assessment of privacy, fairness, and reliability beyond the scope of our benchmark evaluation.

REPRODUCIBILITY STATEMENT

We describe the proposed framework, implementation details, and evaluation protocol in the main paper and appendix. To facilitate reproducibility, we will publicly release our code after the review process has concluded.

REFERENCES

Xu Chen, Siheng Chen, Jiangchao Yao, Huangjie Zheng, Ya Zhang, and Ivor W. Tsang. Learning on attribute-missing graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(2):740–757, 2022. doi: 10.1109/TPAMI.2020.3032189.

Shohei Enomoro and Takeharu Eda. Learning to cascade: Confidence calibration for improving the accuracy and computational cost of cascade inference systems. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7331–7339, 2021.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Jun Hu, Yufei He, Yuan Li, Bryan Hooi, and Bingsheng He. Ntsformer: A self-teaching graph transformer for multimodal isolated cold-start node classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 14856–14864, 2026.

Xinke Jiang, Zidi Qin, Jiarong Xu, and Xiang Ao. Incomplete graph learning via attribute-structure decoupled variational auto-encoder. In *Proceedings of the 17th ACM international conference on web search and data mining*, pp. 304–312, 2024.

Jiwan Kim, Hongseok Kang, Sein Kim, Kibum Kim, and Chanyoung Park. Disentangling and generating modalities for recommendation in missing modality scenarios. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1820–1829, 2025.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. URL https://openreview.net/forum?id=SJU4ayYgl.

Yuan Li, Jun Hu, Jiaxin Jiang, Bryan Hooi, and Bingsheng He. Robust multimodal recommendation via graph retrieval-enhanced modality completion. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 946–957, New York, NY, USA, 2026. Association for Computing Machinery. ISBN

9798400725999. doi: 10.1145/3805712.3809607. URL https://doi.org/10.1145/3805712.3809607.

Ye Liu, Yang Chen, and Hongmin Cai. Migdiff: Multi-attributes imputations for attribute-missing graphs via graph denoising diffusion model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15369–15376, 2026.

Donald Loveland, Yao-An Yang, and Danai Koutra. Glance for context: Learning when to leverage llms for node-aware gnn-llm fusion. In *International Conference on Learning Representations*, volume 2026, pp. 90402–90428, 2026.

Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Tommaso Di Noia, and Fragkiskos D Malliaros. Do we really need to drop items with missing modalities in multimodal recommendation? In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3943–3948, 2024.

Daniele Malitesta, Emanuele Rossi, Claudio Pomo, Tommaso Di Noia, and Fragkiskos D. Malliaros. Training-free graph-based imputation of missing modalities in multimodal recommendation. *IEEE Transactions on Knowledge and Data Engineering*, 38(5):3250–3263, 2026. doi: 10.1109/TKDE.2026.3667005.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7076–7087, 2020. URL https://proceedings.mlr.press/v119/mozannar20b.html.

Feng Nan and Venkatesh Saligrama. Adaptive classification for prediction under a budget. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/d9ff90f4000eacd3a6c9cb27f78994cf-Abstract.html.

Emanuele Rossi, Henry Kenlay, Maria I Gorinova, Benjamin Paul Chamberlain, Xiaowen Dong, and Michael M Bronstein. On the unreasonable effectiveness of feature propagation in learning on graphs with missing node features. In *Learning on graphs conference*, pp. 11–1. PMLR, 2022.

Yuanfu Sun, Kang Li, Pengkang Guo, Jiajin Liu, and Qiaoyu Tan. Mario: Multimodal graph reasoning with large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19219–19228, 2026.

Sina Tabakhi, Chen (Cherise) Chen, and Haiping Lu. Missing-modality-aware graph neural network for cancer classification. *arXiv preprint arXiv:2506.22901*, 2025. doi: 10.48550/arXiv.2506.22901. URL https://arxiv.org/abs/2506.22901.

Wenxuan Tu, Sihang Zhou, Xinwang Liu, Yue Liu, Zhiping Cai, En Zhu, Changwang Zhang, and Jieren Cheng. Initializing then refining: A simple graph attribute imputation network. In Luc De Raedt (ed.), *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pp. 3494–3500. ijcai.org, 2022. doi: 10.24963/IJCAI.2022/485. URL https://doi.org/10.24963/ijcai.2022/485.

Wenxuan Tu, Bin Xiao, Xinwang Liu, Sihang Zhou, Zhiping Cai, and Jieren Cheng. Revisiting initializing then refining: An incomplete and missing graph imputation network. *IEEE Transactions on Neural Networks and Learning Systems*, 36(2):3244–3257, 2025. doi: 10.1109/TNNLS.2024.3349850.

Daeho Um, Jiwoong Park, Seulki Park, and Jin Young Choi. Confidence-based feature imputation for graphs with partially known features. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=YPKBIILy-Kt.

Daeho Um, Sunoh Kim, Jiwoong Park, Jongin Lim, Seong Jin Ahn, and Seulki Park. Propagate and inject: Revisiting propagation-based feature imputation for graphs with partially observed features. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 60530–60560. PMLR, 13–19 Jul 2025.

Haotao Wang, Ziyu Jiang, Yuning You, Yan Han, Gaowen Liu, Jayanth Srinivasa, Ramana Kompella, Zhangyang Wang, et al. Graph mixture of experts: Learning on large-scale graphs with explicit diversity modeling. *Advances in Neural Information Processing Systems*, 36:50825–50837, 2023.

Zhen Wang, Zhewei Wei, Yaliang Li, Weirui Kuang, and Bolin Ding. Graph neural networks with node-wise architecture. In *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining*, pp. 1949–1958, 2022.

Felix Wu, Amauri Souza, Tianyi Zhang, Christopher Fifty, Tao Yu, and Kilian Weinberger. Simplifying graph convolutional networks. In *International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6861–6871, 2019. URL https://proceedings.mlr.press/v97/wu19e.html.

Huangbiao Xu, Huanqi Wu, Xiao Ke, Junyi Wu, Rui Xu, and Jinglin Xu. Mcmoe: Completing missing modalities with mixture of experts for incomplete multimodal action quality assessment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 11241–11249, 2026.

Hao Yan, Chaozhuo Li, Jun Yin, Zhigang Yu, Weihao Han, Mingzheng Li, Zhengxin Zeng, Hao Sun, and Senzhang Wang. When graph meets multimodal: Benchmarking and meditating on multimodal attributed graph learning. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 5842–5853. Association for Computing Machinery, 2025. doi: 10.1145/3711896.3737404. URL https://doi.org/10.1145/3711896.3737404.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:98782–98805, 2024.

Zihao Zeng, Yibo Miao, Hongcheng Gao, Hao Zhang, and Zhijie Deng. Adamoe: Token-adaptive routing with null experts for mixture-of-experts language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 6223–6235, 2024.

Wenqian Zhao, Kai Yang, Peijin Ding, Ce Na, and Wen Li. Graph attention contrastive learning with missing modality for multimodal recommendation. *Knowledge-Based Systems*, 311:113035, 2025. doi: 10.1016/j.knosys.2025.113035.

Jing Zhu, Yuhang Zhou, Shengyi Qian, Zhongmou He, Tong Zhao, Neil Shah, and Danai Koutra. Mosaic of modalities: A comprehensive benchmark for multimodal graph learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14215–14224. IEEE, 2025.

A Complete Six-Dataset Motivation Diagnostics

Figure 1 presents five recovery methods on Ele-Fashion and RedditM to illustrate how NCS varies across soft-homophily bins. Figure 2 shows the four-state decomposition for NTSFormer and RITR on the same two datasets. Figure 7 complements Figure 1 with NCS diagnostics on Grocery, Movies, RedditS, and Toys. Figure 8 reports the four-state decomposition over all six datasets, with correction minus harm matching the corresponding NCS in each soft-homophily bin.

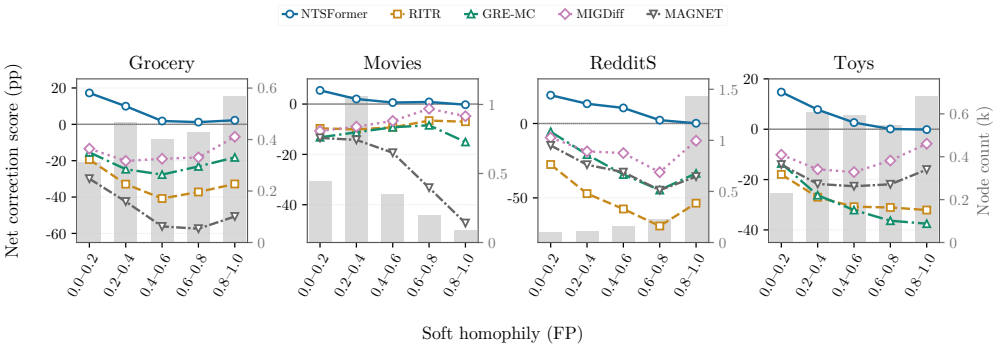


Figure 7: Operator-conditional routing utility across soft-homophily bins on Grocery, Movies, RedditS, and Toys. Curves show seed-mean NCS in percentage points.



Figure 8: Complete four-state decomposition across six datasets (rows) and five recovery methods (columns). Each bar gives seed-mean node fractions within the indicated soft-homophily bin, on a common 0–100% scale. Green denotes `expensive_only` correction and orange denotes `cheap_only` harm; gray segments denote jointly correct or jointly wrong predictions.

B Detailed Related Work

We expand the two research themes discussed in the main text, detailing recovery mechanisms and their relationship to budget-constrained node routing.

B.1 Recovery for Incomplete Multimodal Graphs

Multimodal graph learning combines relational structure with complementary node attributes, including textual and visual information. Mosaic of Modalities provides a unified benchmark for studying these information sources (Zhu et al., 2025). Missing modalities motivate methods that exploit available attributes and graph context, including approaches for multimodal isolated cold-start node classification (Hu et al., 2026) and recommendation with incomplete item modalities (Malitesta et al., 2024). Existing solutions include training-free graph-based imputation (Malitesta et al., 2026) and completion combined with attention-based fusion and contrastive learning (Zhao et al., 2025).

Propagation-based recovery exploits graph connectivity to transfer information from observed attributes to missing entries. Feature Propagation derives a diffusion procedure from Dirichlet energy (Rossi et al., 2022). Confidence-based imputation incorporates channel-level confidence into propagation (Um et al., 2023), while Propagate and Inject address low-variance channels through synthetic feature injection (Um et al., 2025). R-SMR adopts the same broad principle of propagating available information, using a simple masked neighbor-mean operator for its cheap recovery path. It performs two propagation rounds independently for text and visual features while keeping observed blocks fixed. A shared SGC-style then processes the recovered features with a classifier (Wu et al., 2019) to produce the default prediction. These propagation studies provide the methodological background; the cheap endpoint in R-SMR consists of the specified propagation operator and its downstream classifier.

More elaborate methods introduce representation learning, iterative refinement, generation, or retrieval. Existing approaches include distribution matching between structure and attributes (Chen et al., 2022), initialization followed by dynamic-affinity refinement (Tu et al., 2022; 2025), and attribute–structure decoupled variational modeling (Jiang et al., 2024). MIGDiff uses graph denoising diffusion for attribute imputation (Liu et al., 2026), while DGMRec generates missing modalities from shared and modality-specific factors (Kim et al., 2025). GRE-MC retrieves modality-aware subgraphs to support completion (Li et al., 2026). These mechanisms offer alternatives to simple propagation paths. The cheap–expensive distinction depends on the selected endpoints and their computational requirements, rather than implying a universal ordering of predictive quality.

Recovery research studies how to construct useful information from incomplete inputs. R-SMR studies when to replace the cheap-path prediction. With the endpoints fixed, expensive recovery may correct a cheap-path error or corrupt a correct cheap prediction. Neither modality missingness nor aggregate endpoint performance identifies which outcome will occur for a particular node. R-SMR therefore learns a node-level switching policy from the joint correctness of the two endpoints.

B.2 Budget-Constrained Node Routing

Conditional computation provides a broader context for allocating limited computation across inputs. Adaptive classification jointly learns a low-cost predictor and a gate that limits access to a high-cost model under a prediction-cost constraint (Nan & Saligrama, 2017). Learning to defer studies whether a model should predict or delegate to an expert, with consistent surrogate losses derived through a reduction to cost-sensitive learning (Mozannar & Sontag, 2020). These approaches establish input-dependent prediction allocation as an existing research direction. R-SMR studies a specific instance with fixed recovery endpoints: it learns their joint correctness states and uses predicted correction and weighted harm to select nodes under an expensive-call ceiling. Cascade inference uses confidence calibration to balance accuracy and computational cost (Enomoro & Eda, 2021). On graphs, node-wise architectures adapt depth, aggregation, and resolution (Wang et al., 2022), while graph mixture-of-experts models provide diverse computations associated with graph characteristics (Wang et al., 2023). Adaptive expert activation can also reduce unnecessary computation: AdaMoE introduces null experts to permit a variable number of active experts per token (Zeng et al., 2024). These approaches motivate input-dependent computation, although their actions and constraints differ from selecting recovery calls under a node-level budget.

Expert routing also supports incomplete multimodal inputs. Flex-MoE accommodates arbitrary observed modality combinations (Yun et al., 2024), and MCMoE uses expert-based generation to complete missing modalities (Xu et al., 2026). MAGNET uses missingness-aware modality attention and a patient-similarity graph for cancer classification with incomplete multiomics data (Tabakhi et al., 2025). These approaches connect routing to modality availability, representation construction, or expert specialization. R-SMR focuses on the task consequence of switching between fixed recovery endpoints. Cheap-path uncertainty alone does not specify this consequence: a cheap-path error may remain incorrect after switching, while a correct prediction may be corrupted.

GLANCE (Loveland et al., 2026) is particularly related in its use of inexpensive node information to decide where additional computation is beneficial. It constructs routing features from node representations, uncertainty, estimated homophily, and degree. Selected nodes receive LLM processing, and a trainable refiner combines GNN and LLM embeddings. Its routing objective incorporates the prediction-loss difference between the original and refined paths together with a query-cost penalty. GLANCE therefore already learns from the relative value of additional computation.

R-SMR specializes the invocation decision to missing-modality recovery and uses a different supervision target. Predictions from fixed cheap and expensive endpoints define four joint correctness states on labeled development nodes. The router estimates their probabilities from cheap-path information, allowing correction and harm to be weighted separately. The execution rule also differs: GLANCE (Loveland et al., 2026) uses fixed top-k selection within each batch, whereas R-SMR filters out non-positive risk-adjusted scores before selecting within a call ceiling. This permits unused capacity when further calls are not predicted to justify their risk, while unselected nodes retain their cheap-path predictions.

C Supplementary Reconstruction and Task-Utility Diagnostics

We supplement the reconstruction-selection analysis in Section 4.2 with budget-specific results and a correctness-state breakdown. The diagnostic uses Grocery, Toys, and Movies at budgets of 5%, 10%, and 20%.

Budget-specific selection diagnostics. For each dataset, seed, and budget b, we selected the $k = \lceil bN \rceil$ test candidates with the largest reconstruction benefit, where N is the number of test candidates. We computed

$$\rho_b = \mathrm{Spearman}\big(\{B_{\mathrm{rec}}(v)\}_{v \in S_b^{\mathrm{rec}}}, \{B_{\mathrm{task}}(v)\}_{v \in S_b^{\mathrm{rec}}}\big),$$

where S_b^{rec} is the reconstruction-ranked Top-k set. Table 5 reports means and sample standard deviations over evaluation seeds. Unlike correlation over all test candidates, ρ_b measures ranking agreement within the selected set and can vary with budget. It does not measure whether selected switches have positive task benefit; the negative-benefit fractions and task-gain sums address that question separately.

Table 5: Reconstruction-based selection across budgets. ρ_b is the Spearman correlation between B_{rec} and B_{task} within the reconstruction-ranked Top-k set. Overlap is the intersection of reconstruction- and task-ranked Top-k sets divided by k. Neg. task is the fraction of reconstruction-selected nodes with $B_{\mathrm{task}} < 0$. Task and Rec. gains sum B_{task} over their respective selections. Values are the mean $\pm$ sample standard deviation over evaluation seeds.

Dataset	Budget	Top-k ρ_b	Overlap (%)	Neg. task (%)	Task gain	Rec. gain
Grocery	5%	0.104 ± 0.176	8.1 ± 4.6	57.4 ± 5.7	46.28 ± 6.37	-0.55 ± 1.76
Grocery	10%	0.045 ± 0.102	11.7 ± 3.0	59.5 ± 3.7	56.64 ± 6.15	-0.87 ± 1.97
Grocery	20%	0.103 ± 0.069	27.9 ± 3.6	67.6 ± 2.3	62.92 ± 5.61	-3.48 ± 1.61
Toys	5%	0.052 ± 0.133	5.3 ± 1.8	61.1 ± 6.6	57.58 ± 6.03	-0.80 ± 2.46
Toys	10%	0.077 ± 0.072	12.3 ± 2.8	66.2 ± 5.4	71.65 ± 7.48	-2.08 ± 2.37
Toys	20%	0.106 ± 0.056	27.5 ± 2.5	71.6 ± 3.2	78.10 ± 8.35	-8.53 ± 2.95
Movies	5%	0.027 ± 0.176	3.6 ± 2.5	52.5 ± 5.1	39.09 ± 3.32	-1.87 ± 3.53
Movies	10%	0.001 ± 0.080	8.7 ± 2.0	54.1 ± 3.2	52.99 ± 3.81	-1.08 ± 3.90
Movies	20%	0.040 ± 0.083	19.9 ± 3.6	56.0 ± 1.2	67.37 ± 3.11	-1.87 ± 3.84

Correction and harm under reconstruction-based selection. Across the 15 dataset–seed runs, reconstruction-ranked selections have correction/harm rates of 2.07/3.06%, 1.99/2.35%, and 1.94/2.10% at the three budgets, yielding NCS values of -0.99, -0.36, and -0.16 percentage points. Direct task-benefit ranking yields $+7.21$, $+6.95$, and $+5.22$ points. A negative B_{task} denotes an increase in continuous cross-entropy loss; `cheap_only` harm is the correctness-state event defined in Section 4.1. These are distinct outcomes. The negative NCS of reconstruction-ranked selections shows that, in this diagnostic, corrections do not offset the errors introduced by switching.

D Supplementary Task-Aware Oracle Results

Table 6 expands the FP and Oracle comparison in Section 4.3 across three budgets on Grocery, Toys, and Movies. The 20% Accuracy and Macro-F1 means reproduce the corresponding entries in Table 3; the additions are the 5% and 10% budgets, standard deviations, and cross-entropy. For each selected set, FP features are replaced with dense RITR features and the frozen GCN is rerun on the full graph. Accuracy and Macro-F1 increase and mean CE decreases in all 45 dataset–seed–budget cells. These results supplement the same label-aware Oracle diagnostic; they are neither deployable routing results nor full-graph outcomes of reconstruction-ranked selection.

Table 6: Cross-budget FP and Task-benefit Oracle results supplementing Table 3. Each entry gives All-FP $\rightarrow$ Task-benefit Oracle under the same frozen GCN. Accuracy and Macro-F1 are percentages; CE is mean cross-entropy. Values are mean $\pm$ standard deviation over evaluation seeds. This is a label-aware diagnostic, not a deployable router.

Dataset	Budget	Accuracy: FP $\rightarrow$ Oracle	Macro-F1: FP $\rightarrow$ Oracle	CE: FP $\rightarrow$ Oracle
Grocery	5%	$76.24 \pm 1.29 \rightarrow 76.56 \pm 1.20$	$61.81 \pm 1.32 \rightarrow 62.19 \pm 1.19$	$0.9300 \pm 0.0585 \rightarrow 0.9045 \pm 0.0605$
	10%	$76.24 \pm 1.29 \rightarrow 76.67 \pm 1.30$	$61.81 \pm 1.32 \rightarrow 62.30 \pm 1.22$	$0.9300 \pm 0.0585 \rightarrow 0.8994 \pm 0.0607$
	20%	$76.24 \pm 1.29 \rightarrow 76.83 \pm 1.25$	$61.81 \pm 1.32 \rightarrow 62.45 \pm 1.31$	$0.9300 \pm 0.0585 \rightarrow 0.8963 \pm 0.0606$
Toys	5%	$76.97 \pm 0.82 \rightarrow 77.18 \pm 0.81$	$71.81 \pm 1.62 \rightarrow 72.04 \pm 1.57$	$0.8489 \pm 0.0270 \rightarrow 0.8232 \pm 0.0268$
	10%	$76.97 \pm 0.82 \rightarrow 77.48 \pm 0.83$	$71.81 \pm 1.62 \rightarrow 72.33 \pm 1.59$	$0.8489 \pm 0.0270 \rightarrow 0.8172 \pm 0.0269$
	20%	$76.97 \pm 0.82 \rightarrow 77.73 \pm 0.77$	$71.81 \pm 1.62 \rightarrow 72.64 \pm 1.56$	$0.8489 \pm 0.0270 \rightarrow 0.8144 \pm 0.0277$
Movies	5%	$49.07 \pm 1.22 \rightarrow 49.54 \pm 1.12$	$31.65 \pm 3.91 \rightarrow 32.00 \pm 3.83$	$1.5748 \pm 0.0440 \rightarrow 1.5517 \pm 0.0451$
	10%	$49.07 \pm 1.22 \rightarrow 49.91 \pm 1.22$	$31.65 \pm 3.91 \rightarrow 32.17 \pm 3.80$	$1.5748 \pm 0.0440 \rightarrow 1.5446 \pm 0.0452$
	20%	$49.07 \pm 1.22 \rightarrow 50.28 \pm 1.12$	$31.65 \pm 3.91 \rightarrow 32.35 \pm 3.63$	$1.5748 \pm 0.0440 \rightarrow 1.5379 \pm 0.0453$

E Implementation Details of R-SMR

This section supplements the main method with cheap-path configurations, feature preprocessing, and router optimization. Dataset splits, missingness settings, evaluation metrics, and recovery operator training are described in Section 6.1 and Appendix G.

E.1 Cheap-Path Configuration and Feature Preprocessing

Cheap recovery, default classifier, and MLP-Q. Cheap recovery uses a fixed masked neighbor-mean propagation operator with two propagation steps and no trainable parameters. The recovered representations support both the default classifier and routing feature construction.

The default classifier produces the cheap-endpoint logits z_v^C, whereas the separately fitted MLP-Q provides class posteriors $\mathbf{p}_v^Q$ for predictive entropy and soft homophily. The default classifier applies two steps of symmetrically normalized graph propagation with self-loops, followed by a two-layer ReLU classification head with hidden dimension 128. These propagation steps are separate from masked feature recovery. The classifier is fitted once on the full feature matrix using only training-node labels, with full-batch cross-entropy, Adam, a learning rate of 0.01, weight decay of 5×10^{-4}, and 200 epochs. The final-epoch parameters are frozen during router fitting, calibration, and evaluation; cheap predictions use the recovered features.

MLP-Q uses a two-layer ReLU network with hidden dimension 128. Training uses Adam with learning rate 0.01, weight decay 5×10^{-4}, and 200 epochs. The fitted predictor remains fixed during router fitting, calibration, and evaluation.

Feature provenance. The router uses six numerical variables organized into five families: degree and log-degree, available-neighbor fraction, predictive entropy, soft homophily, and mean neighbor-feature similarity. Missing-modality type is included as a categorical variable.

Predictive entropy uses MLP-Q posteriors, while soft homophily combines these posteriors with graph adjacency. Neighbor-feature similarity uses cheap-recovered representations and graph adjacency. Degree is computed from graph adjacency; modality availability and missingness type use the original modality-observation masks. Recovering a missing feature therefore does not change its observation indicator.

Preprocessing. Numerical features are imputed using the median of each feature, augmented with binary indicators of missing values, and then standardized. Missing-modality type is represented using one-hot encoding, with previously unseen categories mapped to an all-zero vector. Imputation statistics, scaling parameters, and categorical mappings are fitted exclusively on router-fitting data and kept fixed during calibration and evaluation.

Ground-truth labels, realized correctness states, expensive-endpoint predictions, realized switching benefits, and label-based local homophily are excluded from routing inputs. Labels and endpoint outcomes are used to construct supervision, not the feature vector.

E.2 Router Optimization and Calibration

Router configuration. The four-state router is a multinomial logistic-regression model trained with L_2-regularized multiclass cross-entropy. We use the L-BFGS solver with a maximum of 500 iterations and inverse regularization strength $C = 1$. Preprocessing and the classifier are fitted as a single pipeline, while both endpoints and MLP-Q remain fixed.

Temperature calibration. After router fitting, we freeze its parameters and fit one scalar temperature using the reserved development calibration subset $\mathcal{D}_{\mathrm{cal}}$:

$$T^\star = \arg\min_{T\in[0.05,20]} -\frac{1}{|\mathcal{D}_{\mathrm{cal}}|}\sum_{v\in\mathcal{D}_{\mathrm{cal}}} \log \frac{\exp(\ell_{v,t_v}/T)}{\sum_{t'\in\mathcal{T}}\exp(\ell_{v,t'}/T)}.$$

We perform bounded optimization in $\log T$. Calibration adjusts the state probabilities without updating the router or either endpoint.

Policy selection. The harm weight is selected from

$$\lambda \in \{1.0, 1.5, 2.0, 3.0\}$$

using validation Accuracy. Ties are resolved by higher Macro-F1, fewer expensive calls, and then larger λ. The positive-score threshold is fixed at zero. At inference, candidates are ranked by descending positive score, with ties broken by ascending node identifier.

Frozen inference configuration. Before final evaluation, we freeze endpoint identities, MLP-Q, feature order, preprocessing parameters, router parameters, temperature, harm weight, and decision threshold. Test labels are used only to compute evaluation metrics. Expensive-endpoint outputs do not enter routing features or determine the selected set.

E.3 Inference Complexity

Let $n = |V|$, $e = |E|$, d be the cheap-recovered feature dimension, c the number of classes, h the MLP-Q hidden width, and $r = \dim(\phi_v)$; $N \le n$ denotes the number of routing candidates. We analyze one inference pass with frozen models, sparse adjacency, and a fixed number of modalities. Given cheap-recovered features, evaluating the two-layer MLP-Q on all nodes costs $O(n(dh + hc))$. Entropy and soft homophily require $O((n + e)c)$ time, while feature normalization and mean neighbor similarity require $O((n + e)d)$. Degree and modality-availability aggregation cost $O(n + e)$. Frozen preprocessing and four-state logistic regression cost $O(Nr)$; temperature scaling, utility computation, and positive-score filtering cost $O(N)$. Sorting candidates with deterministic tie-breaking costs $O(N \log N)$. Thus, the routing time is

$$T_{\mathrm{route}} = O(nh(d + c) + (n + e)(d + c) + Nr + N \log N)\,.$$

The two-step masked neighbor-mean recovery adds $O((n+e)d)$ time if it has not already been computed by the cheap endpoint, without changing this bound. Endpoint classifier evaluation and expensive recovery are separate from T_{route}.

Using sparse matrix aggregation or bounded edge batches avoids storing feature vectors for every edge and admits a total storage bound of

$$S_{\text{route}} = O(e + n(d + h + c) + Nr + h(d + c)),$$

including graph and feature storage, MLP-Q activations and parameters, routing features, and sorting workspace. Materializing all edge-wise feature or posterior messages instead adds $O(e(d+c))$ temporary space. For fixed dimensions ($h = 128$ here), routing therefore takes $O(n+e+N\log N)$ time and admits $O(n+e)$ space.

F Dataset Descriptions and Statistics

We evaluate node classification on six text–image graphs spanning E-commerce and social media. Grocery, Toys, Movies, RedditS, and RedditM come from MAGB (Yan et al., 2025); Ele-Fashion comes from MM-Graph (Zhu et al., 2025). Table 7 summarizes the data files used in our experiments before modality masking.

Table 7: Statistics of the six experimental datasets. Edges are unique undirected node pairs after removing self-loops; reverse and duplicate entries are counted once. Classes count distinct labels present in the data. All datasets contain textual and visual node attributes.

Dataset	Domain	Nodes	Edges	Classes
Grocery	E-commerce	17,074	71,131	20
Toys	E-commerce	20,695	56,701	18
Movies	E-commerce	16,672	80,401	20
RedditS	Social media	15,894	141,540	20
RedditM	Social media	99,638	291,797	50
Ele-Fashion	E-commerce	97,766	199,586	11

Amazon product graphs. Grocery, Toys, and Movies represent products from the Grocery & Gourmet Food, Toys & Games, and Movies & TV categories, respectively. Edges encode co-purchase or co-view relationships. Product titles and descriptions provide textual attributes, product images provide visual attributes, and product categories define the classification targets (Yan et al., 2025). We report the statistics of the loaded Grocery graph (17,074 nodes), rather than those of a different released version.

Reddit post graphs. RedditS and RedditM contain image–caption posts derived from RedCaps. Two posts are connected when a user has commented on both, and the target is the post's subreddit category (Yan et al., 2025). RedditM extends the classification setting to 50 classes and a larger graph than RedditS.

Ele-Fashion. Ele-Fashion represents Amazon products with co-purchase links, textual product titles, and product images (Zhu et al., 2025). Its loaded labels contain 11 distinct categories, with identifiers $\{0, 1, 2, 3, 4, 6, 7, 8, 9, 10, 11\}$. The experimental loader retains these identifiers and sets the output dimension to $\max(y)+1 = 12$; the table reports the number of observed categories rather than this output dimension.

Statistical and split conventions. The edge counts describe graph connectivity, not the number of entries in message-passing tensors, which may retain repeated or reverse edges. The MAGB loader uses label-stratified 60%/20%/20% train/validation/test node splits. For Ele-Fashion, we preserve the supplied split of 58,659/9,777/29,330 nodes. These graph-level counts are distinct from the missing-modality candidate counts used to define routing call rates.

G Training and Evaluation Details

Missingness and evaluation candidates. For each dataset, textual and visual modalities are randomly masked on disjoint node subsets at a rate of 0.4 per modality. Thus, approximately 40% of nodes lack text, 40% lack visual features, and 20% retain both modalities, subject to integer rounding. Each node lacks at most one modality, consistent with the problem definition in Section 3; routing candidates lack exactly one modality. Routing comparisons are evaluated on

$$V_{\text{miss}}^{\text{test}} = V_{\text{miss}} \cap V_{\text{test}}, \qquad N_{\text{test}} = |V_{\text{miss}}^{\text{test}}|.$$

This candidate set determines both the evaluation population and the denominator of the realized call rate.

Training and evaluation controls. Models are fitted on the training split. Probability calibration and hyperparameter selection use validation data, while the test split is reserved for final evaluation. All fitted models, preprocessing parameters, and routing settings are frozen before testing. Within each dataset–seed run and fixed endpoint pair, compared routing policies use the same data splits, modality masks, evaluation candidates, and endpoint predictions. Router fitting and calibration are detailed in Appendix E.

Nominal budgets and realized calls. The main endpoint comparison uses a nominal call-budget cap of 50%. Budget analyses evaluate the prespecified caps $b \in \{0.05, 0.10, 0.20, 0.30, 0.50\}$. For the selected set S_b, the call ceiling and realized call rate are

$$K_b = \lceil bN_{\text{test}} \rceil, \qquad |S_b| \le K_b, \qquad b_{\text{eff}} = \frac{|S_b|}{N_{\text{test}}}.$$

Positive-score filtering can leave the budget partly unused. When the ceiling is reached, integer rounding can make b_{eff} slightly exceed the nominal fraction b, by less than $1/N_{\text{test}}$. Each selected node counts as one expensive call; the computation required to obtain expensive predictions remains operator-specific. Component ablations use a 20% cap, while routing-feature ablations use a 50% cap.

Random-routing baselines. For each dataset–seed–budget combination, Rand-b_{eff} uniformly selects exactly $|S_b|$ candidates, matching R-SMR's realized call count. Rand-$b_{\max}$ uniformly selects K_b candidates, exhausting the nominal budget. Both policies sample without replacement and average the evaluation metrics over 100 random draws within each run. The endpoint predictions and evaluation candidates remain fixed across draws. These comparisons assess routing policies within each endpoint pair.

Metrics and aggregation. Accuracy and Macro-F1 are computed separately for each dataset–seed run before aggregation. Compared policies use the same evaluation candidates and the same class set for Macro-F1. For random policies, the within-run values are the averages over the random draws described above. Per-dataset tables report means and sample standard deviations across the evaluation seeds. Cross-dataset summaries then equally average the six dataset means, rather than pooling predictions across datasets. Call rates are computed within each run and follow the same aggregation order. The policy-comparison figures report seed means within each dataset.

Seeds for random-routing comparisons. The random-routing comparisons in Figure 4 and Appendix I use five evaluation seeds for NTSFormer and RITR, and three for GRE-MC, MAGNET, and MIGDiff.

Separate diagnostic conventions. Table 1 pools candidate-record counts across its diagnostic runs before computing correction, harm, and NCS; repeated observations across seeds are counted as separate records. Selected-node NCS uses the number of selected nodes as its denominator. The full-graph FP–RITR Oracle diagnostic follows the separately specified procedure in Appendix D.

Hyperparameter selection. Table 8 lists the search domains for the five recovery operators. Discrete sets enumerate candidate values; $\mathrm{U}[a, b]$ and $\mathrm{LogU}[a, b]$ denote uniform and log-uniform sampling, respectively. Operator hyperparameters are selected using validation performance before the

endpoints are frozen. The dataset splits, missingness settings, and reporting convention follow Section 6.1.

Table 8: Hyperparameter search domains for recovery operators. Fixed architectural and training settings are reported below.

Operator	Hyperparameter	Search domain
NTSFormer	Learning rate	$\mathrm{LogU}[5\times10^{-4}, 5\times10^{-3}]$
	Weight decay	$\mathrm{LogU}[10^{-4}, 10^{-1}]$
	Global dropout	$\mathrm{U}[0.3, 0.7]$
	Input dropout	$\mathrm{U}[0, 0.5]$
	Group dropout	$\mathrm{U}[0, 0.4]$
	Global-input dropout	$\mathrm{U}[0, 0.4]$
	Group-output dropout	$\mathrm{U}[0, 0.3]$
	Propagation depth (`pre_k`)	$\{1, 2, 3, 4\}$
	Attention heads	$\{2, 4, 8\}$
	Loss (implementation option)	`ce`, `loge_ce`
RITR	Scale σ	$\{0.05, 0.1, 0.2, 0.5\}$
	Order (`r_order`)	$\{1, 2, 3\}$
	Learning rate	$\{3\times10^{-4}, 10^{-3}, 3\times10^{-3}\}$
MIGDiff	Latent dimension	$\{64, 96, 128, 192, 256\}$
	GNN layers	$\{1, 2, 3\}$
	Neighbor fanouts	$\{(8, 4), (8, 8), (16, 8), (24, 12), (32, 16)\}$
	Diffusion steps	$\{5, 10, 15, 20, 30\}$
	Learning rate	$\mathrm{LogU}[10^{-4}, 5\times10^{-3}]$
	Weight decay	$\mathrm{LogU}[10^{-7}, 5\times10^{-4}]$
	Reconstruction weight	$\mathrm{LogU}[0.05, 8]$
	Diffusion samples	$\{1, 3, 5, 8\}$
GRE-MC	Anchor count	$\{5, 10, 20\}$
	Codebook size	$\{10, 50, 100\}$
	Usage weight	$\{0.5, 1, 2\}$
	Load-balancing weight	$\{0.5, 1, 2\}$
	Learning rate	$\{10^{-4}, 3\times10^{-4}, 10^{-3}, 3\times10^{-3}, 10^{-2}\}$
	Weight decay	$\{10^{-6}, 10^{-5}, 10^{-4}\}$
MAGNET	Retrieved neighbors k	$\{8, 16, 32\}$
	Learning rate	$\mathrm{LogU}[10^{-4}, 3\times10^{-3}]$
	Weight decay	$\mathrm{LogU}[10^{-6}, 10^{-3}]$
	Dropout	$\{0, 0.2, 0.5\}$

Operator search procedures and fixed settings. NTSFormer uses Bayesian search to maximize validation teacher accuracy. To fit within 24 GB of GPU memory, we use one Transformer layer instead of the two layers in the original configuration; learning-rate decay power is fixed to 1. Feature-projection and feed-forward layer widths are $[512, 128]$, and the classification head has a 128-dimensional hidden layer. The input MoE uses one shared expert; the number of routed experts equals the number of input feature groups, rather than being fixed to six. The self-teaching loss weight is 1.0, and the MoE auxiliary-loss coefficient is 0.1. For RITR, we retain the original fixed loss weights $\alpha = \beta = 10$, encoder hidden widths of 64 and 64, dropout 0.5, and Adam weight decay 5×10^{-4}. In our experimental pipeline, we perform coordinate search in the order σ, `r_order`, and learning rate over the domains in Table 8, selecting configurations by validation classification cross-entropy. Each candidate recovery model is trained for 300 epochs during search. After selection, the recovery model is trained from scratch for 1,000 epochs with the selected hyperparameters and then frozen. These are our recovery-training budgets, separate from downstream classifier training. MIGDiff initializes 40 configurations with a scrambled Sobol sequence and uses successive halving at 40, 80, and 160 epochs, retaining 20 and then 10 configurations. Its structural search uses one diffusion sample; the sample-count search subsequently considers $\{1, 3, 5, 8\}$. Selection uses

validation representation-interface accuracy on Grocery and the two-interface accuracy criterion on the other datasets. The batch size is 64; the final MIGDiff endpoint is trained for 200 epochs and then frozen. GRE-MC uses a Bayesian sweep minimizing validation reconstruction MSE. Its fixed settings are hidden dimension 128, two layers, four attention heads, top-$p = 4$, temperature 0.5, positional dimension 20, ten MAGE iterations, dropout 0.5, and batch size 512; training uses at most 100 epochs with patience 10. MAGNET uses 20 TPE trials per dataset, maximizing validation accuracy, with ties resolved by macro-F1, lower query FLOPs, and earlier best epoch. Its fixed settings are hidden dimension 128, four attention heads, two GNN layers, KL weight 0.1, candidate multiplier 8, and HNSW parameters $M = 32$ and `efConstruction`$= 200$.

R-SMR implementation. Appendix E specifies the cheap-path configuration, feature preprocessing, router optimization, and calibration. For random routing, effects are computed within each dataset–seed cell before averaging across seeds and balancing across datasets.

H Cross-Budget Results

H.1 Cross-Dataset Summary

Table 9 summarizes performance and realized expensive-call rates across five recovery operators and five nominal budget caps. Each value equally averages the six dataset means following aggregation across evaluation seeds within each dataset.

Table 9: Complete cross-budget five-operator comparison. Values are dataset- and seed-balanced means (%); b_{eff} is the percentage of missing-modality test candidates selected for expensive recovery, averaged with the same weights.

Operator	Budget	b_{eff} (%)	All-cheap		All-expensive		R-SMR	
			Acc.	Macro-F1	Acc.	Macro-F1	Acc.	Macro-F1
NTSFormer	5%	5.01	72.70	62.65	**77.38**	**69.87**	74.10	64.74
	10%	10.01	72.70	62.65	**77.38**	**69.87**	74.82	65.62
	20%	20.01	72.70	62.65	**77.38**	**69.87**	75.76	66.80
	30%	30.01	72.70	62.65	**77.38**	**69.87**	76.64	67.82
	50%	50.01	72.70	62.65	77.38	69.87	**78.33**	**70.95**
GRE-MC	5%	3.76	72.87	62.84	51.87	39.22	**74.52**	**64.20**
	10%	5.53	72.87	62.84	51.87	39.22	**74.68**	**64.25**
	20%	7.00	72.87	62.84	51.87	39.22	**74.81**	**64.49**
	30%	7.00	72.87	62.84	51.87	39.22	**74.81**	**64.49**
	50%	7.00	72.87	62.84	51.87	39.22	**74.81**	**64.49**
MAGNET	5%	4.10	73.62	63.63	49.25	34.53	**75.34**	**65.11**
	10%	5.76	73.62	63.63	49.25	34.53	**75.55**	**65.34**
	20%	8.02	73.62	63.63	49.25	34.53	**75.64**	**65.62**
	30%	8.45	73.62	63.63	49.25	34.53	**75.66**	**65.65**
	50%	8.45	73.62	63.63	49.25	34.53	**75.66**	**65.65**
RITR	5%	4.03	73.01	63.03	47.60	30.92	**74.72**	**64.30**
	10%	5.40	73.01	63.03	47.60	30.92	**74.93**	**64.50**
	20%	6.04	73.01	63.03	47.60	30.92	**74.92**	**64.50**
	30%	6.04	73.01	63.03	47.60	30.92	**74.92**	**64.50**
	50%	6.04	73.01	63.03	47.60	30.92	**74.92**	**64.50**
MIGDiff	5%	3.67	73.17	63.37	60.47	45.22	**74.09**	**64.23**
	10%	5.23	73.17	63.37	60.47	45.22	**74.21**	**64.33**
	20%	5.70	73.17	63.37	60.47	45.22	**74.32**	**64.41**
	30%	5.70	73.17	63.37	60.47	45.22	**74.32**	**64.41**
	50%	5.70	73.17	63.37	60.47	45.22	**74.32**	**64.41**

H.2 Per-Dataset Results

Table 10 provides the corresponding per-dataset results and standard deviations for all 150 operator–dataset–budget combinations.

Table 10: Per-dataset cross-budget results (%). Values are mean ± standard deviation; b_{eff} is the realized R-SMR call rate. For b_{eff}, red indicates reaching the budget cap, whereas green indicates there is still remaining budget.

Budget	b_{eff} (%)	All-cheap Accuracy	All-cheap Macro-F1	All-expensive Accuracy	All-expensive Macro-F1	R-SMR Accuracy	R-SMR Macro-F1
NTSFormer / Ele-Fashion							
5%	5.00 ± 0.00	79.29 ± 0.20	61.64 ± 0.39	**83.68 ± 0.23**	**67.54 ± 0.49**	80.73 ± 0.27	61.95 ± 0.27
10%	10.00 ± 0.00	79.29 ± 0.20	61.64 ± 0.39	**83.68 ± 0.23**	**67.54 ± 0.49**	81.64 ± 0.29	62.00 ± 0.30
20%	20.00 ± 0.00	79.29 ± 0.20	61.64 ± 0.39	**83.68 ± 0.23**	**67.54 ± 0.49**	82.61 ± 0.23	62.14 ± 0.26
30%	30.00 ± 0.00	79.29 ± 0.20	61.64 ± 0.39	**83.68 ± 0.23**	**67.54 ± 0.49**	83.27 ± 0.31	63.05 ± 0.20
50%	50.00 ± 0.00	79.29 ± 0.20	61.64 ± 0.39	83.68 ± 0.23	67.54 ± 0.49	**84.63 ± 0.20**	**67.96 ± 0.25**
NTSFormer / Grocery							
5%	5.01 ± 0.01	74.76 ± 0.81	61.90 ± 0.93	**80.54 ± 0.48**	**70.97 ± 1.66**	75.50 ± 0.87	63.35 ± 1.40
10%	10.02 ± 0.01	74.76 ± 0.81	61.90 ± 0.93	**80.54 ± 0.48**	**70.97 ± 1.66**	76.18 ± 0.94	64.78 ± 1.95
20%	20.02 ± 0.01	74.76 ± 0.81	61.90 ± 0.93	**80.54 ± 0.48**	**70.97 ± 1.66**	77.33 ± 0.81	66.75 ± 2.08
30%	30.01 ± 0.01	74.76 ± 0.81	61.90 ± 0.93	**80.54 ± 0.48**	**70.97 ± 1.66**	79.98 ± 0.61	69.45 ± 1.81
50%	50.01 ± 0.01	74.76 ± 0.81	61.90 ± 0.93	80.54 ± 0.48	70.97 ± 1.66	**82.70 ± 0.56**	**72.95 ± 1.93**
NTSFormer / Movies							
5%	5.02 ± 0.01	49.54 ± 1.18	35.54 ± 2.81	**51.73 ± 1.53**	**44.56 ± 2.10**	50.07 ± 1.24	38.22 ± 2.70
10%	10.02 ± 0.01	49.54 ± 1.18	35.54 ± 2.81	**51.73 ± 1.53**	**44.56 ± 2.10**	50.21 ± 1.33	39.00 ± 2.93
20%	20.02 ± 0.01	49.54 ± 1.18	35.54 ± 2.81	**51.73 ± 1.53**	**44.56 ± 2.10**	50.73 ± 1.16	40.40 ± 2.45
30%	30.02 ± 0.01	49.54 ± 1.18	35.54 ± 2.81	**51.73 ± 1.53**	**44.56 ± 2.10**	51.17 ± 1.38	41.30 ± 2.21
50%	50.01 ± 0.01	49.54 ± 1.18	35.54 ± 2.81	51.73 ± 1.53	44.56 ± 2.10	**52.69 ± 1.23**	**46.32 ± 2.38**
NTSFormer / RedditM							
5%	5.00 ± 0.00	66.64 ± 0.68	61.67 ± 0.86	**75.88 ± 0.31**	**72.14 ± 0.31**	69.27 ± 0.63	64.45 ± 0.80
10%	10.00 ± 0.00	66.64 ± 0.68	61.67 ± 0.86	**75.88 ± 0.31**	**72.14 ± 0.31**	70.94 ± 0.62	66.42 ± 0.75
20%	20.00 ± 0.00	66.64 ± 0.68	61.67 ± 0.86	**75.88 ± 0.31**	**72.14 ± 0.31**	72.92 ± 0.51	68.84 ± 0.58
30%	30.00 ± 0.00	66.64 ± 0.68	61.67 ± 0.86	**75.88 ± 0.31**	**72.14 ± 0.31**	74.31 ± 0.38	70.40 ± 0.40
50%	50.00 ± 0.00	66.64 ± 0.68	61.67 ± 0.86	75.88 ± 0.31	72.14 ± 0.31	**76.86 ± 0.22**	**73.79 ± 0.21**
NTSFormer / RedditS							
5%	5.02 ± 0.01	90.77 ± 0.25	84.26 ± 0.64	**93.66 ± 0.47**	**88.73 ± 1.03**	91.72 ± 0.26	86.66 ± 0.43
10%	10.01 ± 0.02	90.77 ± 0.25	84.26 ± 0.64	**93.66 ± 0.47**	**88.73 ± 1.03**	92.22 ± 0.35	87.29 ± 0.75
20%	20.01 ± 0.01	90.77 ± 0.25	84.26 ± 0.64	**93.66 ± 0.47**	**88.73 ± 1.03**	92.50 ± 0.43	87.66 ± 0.75
30%	30.03 ± 0.01	90.77 ± 0.25	84.26 ± 0.64	**93.66 ± 0.47**	**88.73 ± 1.03**	92.53 ± 0.46	87.72 ± 0.91
50%	50.01 ± 0.01	90.77 ± 0.25	84.26 ± 0.64	93.66 ± 0.47	88.73 ± 1.03	**93.77 ± 0.47**	**88.75 ± 0.91**
NTSFormer / Toys							
5%	5.01 ± 0.01	75.18 ± 1.22	70.86 ± 1.16	**78.81 ± 0.92**	**75.29 ± 1.01**	77.33 ± 1.13	73.79 ± 1.29
10%	10.01 ± 0.01	75.18 ± 1.22	70.86 ± 1.16	**78.81 ± 0.92**	**75.29 ± 1.01**	77.73 ± 1.12	74.25 ± 1.28
20%	20.02 ± 0.01	75.18 ± 1.22	70.86 ± 1.16	**78.81 ± 0.92**	**75.29 ± 1.01**	78.45 ± 1.04	74.99 ± 1.20
30%	30.01 ± 0.01	75.18 ± 1.22	70.86 ± 1.16	**78.81 ± 0.92**	**75.29 ± 1.01**	78.60 ± 1.01	75.02 ± 1.27
50%	50.00 ± 0.01	75.18 ± 1.22	70.86 ± 1.16	78.81 ± 0.92	75.29 ± 1.01	**79.34 ± 0.93**	**75.94 ± 1.18**
GRE-MC / Ele-Fashion							
5%	4.84 ± 0.29	79.30 ± 0.14	61.75 ± 0.48	74.16 ± 0.78	50.04 ± 0.30	**80.29 ± 0.10**	**62.15 ± 0.26**
10%	10.00 ± 0.00	79.30 ± 0.14	61.75 ± 0.48	74.16 ± 0.78	50.04 ± 0.30	**80.57 ± 0.04**	**62.28 ± 0.44**
20%	12.88 ± 3.39	79.30 ± 0.14	61.75 ± 0.48	74.16 ± 0.78	50.04 ± 0.30	**81.20 ± 0.01**	**63.58 ± 0.46**
30%	12.88 ± 3.39	79.30 ± 0.14	61.75 ± 0.48	74.16 ± 0.78	50.04 ± 0.30	**81.20 ± 0.01**	**63.58 ± 0.46**
50%	12.88 ± 3.39	79.30 ± 0.14	61.75 ± 0.48	74.16 ± 0.78	50.04 ± 0.30	**81.20 ± 0.01**	**63.58 ± 0.46**
GRE-MC / Grocery							
5%	3.27 ± 1.34	75.20 ± 0.64	61.98 ± 1.42	53.23 ± 2.23	39.98 ± 1.62	**77.34 ± 0.54**	**64.01 ± 1.39**
10%	3.77 ± 0.72	75.20 ± 0.64	61.98 ± 1.42	53.23 ± 2.23	39.98 ± 1.62	**77.38 ± 0.51**	**64.01 ± 1.33**
20%	3.77 ± 0.72	75.20 ± 0.64	61.98 ± 1.42	53.23 ± 2.23	39.98 ± 1.62	**77.38 ± 0.51**	**64.01 ± 1.33**
30%	3.77 ± 0.72	75.20 ± 0.64	61.98 ± 1.42	53.23 ± 2.23	39.98 ± 1.62	**77.38 ± 0.51**	**64.01 ± 1.33**
50%	3.77 ± 0.72	75.20 ± 0.64	61.98 ± 1.42	53.23 ± 2.23	39.98 ± 1.62	**77.38 ± 0.51**	**64.01 ± 1.33**
GRE-MC / Movies							
5%	3.23 ± 0.14	49.91 ± 0.25	36.51 ± 0.46	39.27 ± 1.01	8.47 ± 1.81	**51.88 ± 0.24**	**37.43 ± 0.49**
10%	3.23 ± 0.14	49.91 ± 0.25	36.51 ± 0.46	39.27 ± 1.01	8.47 ± 1.81	**51.88 ± 0.24**	**37.43 ± 0.49**
20%	3.23 ± 0.14	49.91 ± 0.25	36.51 ± 0.46	39.27 ± 1.01	8.47 ± 1.81	**51.88 ± 0.24**	**37.43 ± 0.49**
30%	3.23 ± 0.14	49.91 ± 0.25	36.51 ± 0.46	39.27 ± 1.01	8.47 ± 1.81	**51.88 ± 0.24**	**37.43 ± 0.49**
50%	3.23 ± 0.14	49.91 ± 0.25	36.51 ± 0.46	39.27 ± 1.01	8.47 ± 1.81	**51.88 ± 0.24**	**37.43 ± 0.49**
GRE-MC / RedditM							
5%	5.01 ± 0.00	66.49 ± 0.74	61.56 ± 0.90	42.27 ± 0.07	41.81 ± 0.63	**68.84 ± 0.66**	**63.84 ± 0.71**
10%	10.00 ± 0.00	66.49 ± 0.74	61.56 ± 0.90	42.27 ± 0.07	41.81 ± 0.63	**69.48 ± 0.55**	**64.02 ± 0.71**
20%	15.90 ± 1.97	66.49 ± 0.74	61.56 ± 0.90	42.27 ± 0.07	41.81 ± 0.63	**69.59 ± 0.47**	**64.13 ± 0.66**
30%	15.90 ± 1.97	66.49 ± 0.74	61.56 ± 0.90	42.27 ± 0.07	41.81 ± 0.63	**69.59 ± 0.47**	**64.13 ± 0.66**
50%	15.90 ± 1.97	66.49 ± 0.74	61.56 ± 0.90	42.27 ± 0.07	41.81 ± 0.63	**69.59 ± 0.47**	**64.13 ± 0.66**
GRE-MC / RedditS							
5%	2.36 ± 0.15	90.68 ± 0.15	83.88 ± 0.22	58.01 ± 0.88	55.27 ± 0.43	**91.19 ± 0.37**	**84.52 ± 0.95**
10%	2.36 ± 0.15	90.68 ± 0.15	83.88 ± 0.22	58.01 ± 0.88	55.27 ± 0.43	**91.19 ± 0.37**	**84.52 ± 0.95**
20%	2.36 ± 0.15	90.68 ± 0.15	83.88 ± 0.22	58.01 ± 0.88	55.27 ± 0.43	**91.19 ± 0.37**	**84.52 ± 0.95**
30%	2.36 ± 0.15	90.68 ± 0.15	83.88 ± 0.22	58.01 ± 0.88	55.27 ± 0.43	**91.19 ± 0.37**	**84.52 ± 0.95**
50%	2.36 ± 0.15	90.68 ± 0.15	83.88 ± 0.22	58.01 ± 0.88	55.27 ± 0.43	**91.19 ± 0.37**	**84.52 ± 0.95**

Continued on next page

Table 10: Per-dataset cross-budget results (continued).

Budget	b_{eff} (%)	All-cheap Accuracy	All-cheap Macro-F1	All-expensive Accuracy	All-expensive Macro-F1	R-SMR Accuracy	R-SMR Macro-F1
GRE-MC / Toys							
5%	3.84 ± 0.59	75.64 ± 0.13	71.34 ± 0.91	44.27 ± 1.00	39.72 ± 1.14	**77.59 ± 0.31**	**73.25 ± 1.06**
10%	3.84 ± 0.59	75.64 ± 0.13	71.34 ± 0.91	44.27 ± 1.00	39.72 ± 1.14	**77.59 ± 0.31**	**73.25 ± 1.06**
20%	3.84 ± 0.59	75.64 ± 0.13	71.34 ± 0.91	44.27 ± 1.00	39.72 ± 1.14	**77.59 ± 0.31**	**73.25 ± 1.06**
30%	3.84 ± 0.59	75.64 ± 0.13	71.34 ± 0.91	44.27 ± 1.00	39.72 ± 1.14	**77.59 ± 0.31**	**73.25 ± 1.06**
50%	3.84 ± 0.59	75.64 ± 0.13	71.34 ± 0.91	44.27 ± 1.00	39.72 ± 1.14	**77.59 ± 0.31**	**73.25 ± 1.06**
MAGNET / Ele-Fashion							
5%	5.00 ± 0.00	80.18 ± 0.07	**62.10 ± 0.55**	74.62 ± 3.24	47.19 ± 9.81	**80.99 ± 0.07**	61.69 ± 0.46
10%	10.00 ± 0.00	80.18 ± 0.07	**62.10 ± 0.55**	74.62 ± 3.24	47.19 ± 9.81	**81.21 ± 0.32**	62.06 ± 0.30
20%	15.53 ± 3.87	80.18 ± 0.07	62.10 ± 0.55	74.62 ± 3.24	47.19 ± 9.81	**81.29 ± 0.54**	**63.38 ± 1.11**
30%	17.23 ± 6.81	80.18 ± 0.07	62.10 ± 0.55	74.62 ± 3.24	47.19 ± 9.81	**81.36 ± 0.71**	**63.52 ± 1.79**
50%	17.23 ± 6.81	80.18 ± 0.07	62.10 ± 0.55	74.62 ± 3.24	47.19 ± 9.81	**81.36 ± 0.71**	**63.52 ± 1.79**
MAGNET / Grocery							
5%	2.71 ± 0.13	75.20 ± 0.64	61.98 ± 1.42	27.21 ± 22.60	13.39 ± 20.90	**76.82 ± 0.64**	**63.98 ± 1.42**
10%	2.71 ± 0.13	75.20 ± 0.64	61.98 ± 1.42	27.21 ± 22.60	13.39 ± 20.90	**76.82 ± 0.64**	**63.98 ± 1.42**
20%	2.71 ± 0.13	75.20 ± 0.64	61.98 ± 1.42	27.21 ± 22.60	13.39 ± 20.90	**76.82 ± 0.64**	**63.98 ± 1.42**
30%	2.71 ± 0.13	75.20 ± 0.64	61.98 ± 1.42	27.21 ± 22.60	13.39 ± 20.90	**76.82 ± 0.64**	**63.98 ± 1.42**
50%	2.71 ± 0.13	75.20 ± 0.64	61.98 ± 1.42	27.21 ± 22.60	13.39 ± 20.90	**76.82 ± 0.64**	**63.98 ± 1.42**
MAGNET / Movies							
5%	3.41 ± 0.94	49.86 ± 0.24	36.54 ± 1.74	32.93 ± 0.00	2.48 ± 0.00	**51.97 ± 0.28**	**38.67 ± 1.72**
10%	3.41 ± 0.94	49.86 ± 0.24	36.54 ± 1.74	32.93 ± 0.00	2.48 ± 0.00	**51.97 ± 0.28**	**38.67 ± 1.72**
20%	3.41 ± 0.94	49.86 ± 0.24	36.54 ± 1.74	32.93 ± 0.00	2.48 ± 0.00	**51.97 ± 0.28**	**38.67 ± 1.72**
30%	3.41 ± 0.94	49.86 ± 0.24	36.54 ± 1.74	32.93 ± 0.00	2.48 ± 0.00	**51.97 ± 0.28**	**38.67 ± 1.72**
50%	3.41 ± 0.94	49.86 ± 0.24	36.54 ± 1.74	32.93 ± 0.00	2.48 ± 0.00	**51.97 ± 0.28**	**38.67 ± 1.72**
MAGNET / RedditM							
5%	5.01 ± 0.00	68.61 ± 0.36	63.69 ± 0.28	47.39 ± 1.08	45.45 ± 0.64	**71.21 ± 0.30**	**66.29 ± 0.27**
10%	10.00 ± 0.00	68.61 ± 0.36	63.69 ± 0.28	47.39 ± 1.08	45.45 ± 0.64	**72.25 ± 0.21**	**67.32 ± 0.20**
20%	18.00 ± 1.94	68.61 ± 0.36	63.69 ± 0.28	47.39 ± 1.08	45.45 ± 0.64	**72.67 ± 0.12**	**67.65 ± 0.21**
30%	18.93 ± 3.45	68.61 ± 0.36	63.69 ± 0.28	47.39 ± 1.08	45.45 ± 0.64	**72.71 ± 0.09**	**67.70 ± 0.24**
50%	18.93 ± 3.45	68.61 ± 0.36	63.69 ± 0.28	47.39 ± 1.08	45.45 ± 0.64	**72.71 ± 0.09**	**67.70 ± 0.24**
MAGNET / RedditS							
5%	3.63 ± 0.92	91.48 ± 0.77	85.69 ± 0.47	56.68 ± 2.67	51.01 ± 1.22	**92.32 ± 0.44**	**86.31 ± 0.35**
10%	3.63 ± 0.92	91.48 ± 0.77	85.69 ± 0.47	56.68 ± 2.67	51.01 ± 1.22	**92.32 ± 0.44**	**86.31 ± 0.35**
20%	3.63 ± 0.92	91.48 ± 0.77	85.69 ± 0.47	56.68 ± 2.67	51.01 ± 1.22	**92.32 ± 0.44**	**86.31 ± 0.35**
30%	3.63 ± 0.92	91.48 ± 0.77	85.69 ± 0.47	56.68 ± 2.67	51.01 ± 1.22	**92.32 ± 0.44**	**86.31 ± 0.35**
50%	3.63 ± 0.92	91.48 ± 0.77	85.69 ± 0.47	56.68 ± 2.67	51.01 ± 1.22	**92.32 ± 0.44**	**86.31 ± 0.35**
MAGNET / Toys							
5%	4.81 ± 0.12	76.40 ± 0.29	71.80 ± 0.64	56.69 ± 1.42	47.63 ± 1.54	**78.75 ± 0.37**	**73.72 ± 0.64**
10%	4.81 ± 0.12	76.40 ± 0.29	71.80 ± 0.64	56.69 ± 1.42	47.63 ± 1.54	**78.75 ± 0.37**	**73.72 ± 0.64**
20%	4.81 ± 0.12	76.40 ± 0.29	71.80 ± 0.64	56.69 ± 1.42	47.63 ± 1.54	**78.75 ± 0.37**	**73.72 ± 0.64**
30%	4.81 ± 0.12	76.40 ± 0.29	71.80 ± 0.64	56.69 ± 1.42	47.63 ± 1.54	**78.75 ± 0.37**	**73.72 ± 0.64**
50%	4.81 ± 0.12	76.40 ± 0.29	71.80 ± 0.64	56.69 ± 1.42	47.63 ± 1.54	**78.75 ± 0.37**	**73.72 ± 0.64**
RITR / Ele-Fashion							
5%	4.37 ± 0.88	79.18 ± 0.10	61.13 ± 0.76	76.54 ± 0.34	52.80 ± 0.86	**80.15 ± 0.14**	**61.76 ± 0.55**
10%	7.59 ± 2.36	79.18 ± 0.10	61.13 ± 0.76	76.54 ± 0.34	52.80 ± 0.86	**80.71 ± 0.17**	**62.47 ± 0.53**
20%	7.59 ± 2.36	79.18 ± 0.10	61.13 ± 0.76	76.54 ± 0.34	52.80 ± 0.86	**80.71 ± 0.17**	**62.47 ± 0.53**
30%	7.59 ± 2.36	79.18 ± 0.10	61.13 ± 0.76	76.54 ± 0.34	52.80 ± 0.86	**80.71 ± 0.17**	**62.47 ± 0.53**
50%	7.59 ± 2.36	79.18 ± 0.10	61.13 ± 0.76	76.54 ± 0.34	52.80 ± 0.86	**80.71 ± 0.17**	**62.47 ± 0.53**
RITR / Grocery							
5%	4.43 ± 0.30	75.08 ± 0.60	61.98 ± 0.79	41.86 ± 1.11	26.48 ± 1.10	**77.66 ± 0.65**	**63.21 ± 0.73**
10%	4.43 ± 0.30	75.08 ± 0.60	61.98 ± 0.79	41.86 ± 1.11	26.48 ± 1.10	**77.66 ± 0.65**	**63.21 ± 0.73**
20%	4.43 ± 0.30	75.08 ± 0.60	61.98 ± 0.79	41.86 ± 1.11	26.48 ± 1.10	**77.66 ± 0.65**	**63.21 ± 0.73**
30%	4.43 ± 0.30	75.08 ± 0.60	61.98 ± 0.79	41.86 ± 1.11	26.48 ± 1.10	**77.66 ± 0.65**	**63.21 ± 0.73**
50%	4.43 ± 0.30	75.08 ± 0.60	61.98 ± 0.79	41.86 ± 1.11	26.48 ± 1.10	**77.66 ± 0.65**	**63.21 ± 0.73**
RITR / Movies							
5%	3.88 ± 0.36	49.89 ± 0.83	36.61 ± 1.28	40.53 ± 0.34	9.04 ± 0.33	**51.91 ± 0.82**	**37.43 ± 1.29**
10%	3.88 ± 0.36	49.89 ± 0.83	36.61 ± 1.28	40.53 ± 0.34	9.04 ± 0.33	**51.91 ± 0.82**	**37.43 ± 1.29**
20%	3.88 ± 0.36	49.89 ± 0.83	36.61 ± 1.28	40.53 ± 0.34	9.04 ± 0.33	**51.91 ± 0.82**	**37.43 ± 1.29**
30%	3.88 ± 0.36	49.89 ± 0.83	36.61 ± 1.28	40.53 ± 0.34	9.04 ± 0.33	**51.91 ± 0.82**	**37.43 ± 1.29**
50%	3.88 ± 0.36	49.89 ± 0.83	36.61 ± 1.28	40.53 ± 0.34	9.04 ± 0.33	**51.91 ± 0.82**	**37.43 ± 1.29**
RITR / RedditM							
5%	5.00 ± 0.00	66.78 ± 0.63	61.80 ± 0.99	42.45 ± 0.69	42.41 ± 0.69	**69.08 ± 0.51**	**64.16 ± 0.86**
10%	10.00 ± 0.00	66.78 ± 0.63	61.80 ± 0.99	42.45 ± 0.69	42.41 ± 0.69	**69.78 ± 0.48**	**64.61 ± 0.87**
20%	13.88 ± 2.79	66.78 ± 0.63	61.80 ± 0.99	42.45 ± 0.69	42.41 ± 0.69	**69.73 ± 0.40**	**64.64 ± 0.72**
30%	13.88 ± 2.79	66.78 ± 0.63	61.80 ± 0.99	42.45 ± 0.69	42.41 ± 0.69	**69.73 ± 0.40**	**64.64 ± 0.72**
50%	13.88 ± 2.79	66.78 ± 0.63	61.80 ± 0.99	42.45 ± 0.69	42.41 ± 0.69	**69.73 ± 0.40**	**64.64 ± 0.72**

Continued on next page

Table 10: Per-dataset cross-budget results (continued).

Budget	b_{eff} (%)	All-cheap Accuracy	All-cheap Macro-F1	All-expensive Accuracy	All-expensive Macro-F1	R-SMR Accuracy	R-SMR Macro-F1
RITR / RedditS							
5%	3.05 ± 0.10	90.82 ± 0.38	84.52 ± 0.69	37.04 ± 3.38	19.90 ± 1.42	$\mathbf{92.49 \pm 0.43}$	$\mathbf{85.23 \pm 0.71}$
10%	3.05 ± 0.10	90.82 ± 0.38	84.52 ± 0.69	37.04 ± 3.38	19.90 ± 1.42	$\mathbf{92.49 \pm 0.43}$	$\mathbf{85.23 \pm 0.71}$
20%	3.05 ± 0.10	90.82 ± 0.38	84.52 ± 0.69	37.04 ± 3.38	19.90 ± 1.42	$\mathbf{92.49 \pm 0.43}$	$\mathbf{85.23 \pm 0.71}$
30%	3.05 ± 0.10	90.82 ± 0.38	84.52 ± 0.69	37.04 ± 3.38	19.90 ± 1.42	$\mathbf{92.49 \pm 0.43}$	$\mathbf{85.23 \pm 0.71}$
50%	3.05 ± 0.10	90.82 ± 0.38	84.52 ± 0.69	37.04 ± 3.38	19.90 ± 1.42	$\mathbf{92.49 \pm 0.43}$	$\mathbf{85.23 \pm 0.71}$
RITR / Toys							
5%	3.42 ± 0.15	76.32 ± 0.61	72.15 ± 0.83	47.17 ± 1.80	34.86 ± 1.89	$\mathbf{77.03 \pm 0.62}$	$\mathbf{74.02 \pm 0.83}$
10%	3.42 ± 0.15	76.32 ± 0.61	72.15 ± 0.83	47.17 ± 1.80	34.86 ± 1.89	$\mathbf{77.03 \pm 0.62}$	$\mathbf{74.02 \pm 0.83}$
20%	3.42 ± 0.15	76.32 ± 0.61	72.15 ± 0.83	47.17 ± 1.80	34.86 ± 1.89	$\mathbf{77.03 \pm 0.62}$	$\mathbf{74.02 \pm 0.83}$
30%	3.42 ± 0.15	76.32 ± 0.61	72.15 ± 0.83	47.17 ± 1.80	34.86 ± 1.89	$\mathbf{77.03 \pm 0.62}$	$\mathbf{74.02 \pm 0.83}$
50%	3.42 ± 0.15	76.32 ± 0.61	72.15 ± 0.83	47.17 ± 1.80	34.86 ± 1.89	$\mathbf{77.03 \pm 0.62}$	$\mathbf{74.02 \pm 0.83}$
MIGDiff / Ele-Fashion							
5%	5.00 ± 0.00	79.45 ± 0.32	62.99 ± 0.23	73.36 ± 0.42	48.25 ± 0.48	$\mathbf{80.16 \pm 0.21}$	$\mathbf{63.23 \pm 0.02}$
10%	9.36 ± 0.97	79.45 ± 0.32	62.99 ± 0.23	73.36 ± 0.42	48.25 ± 0.48	$\mathbf{80.17 \pm 0.34}$	$\mathbf{63.20 \pm 0.07}$
20%	9.61 ± 1.26	79.45 ± 0.32	62.99 ± 0.23	73.36 ± 0.42	48.25 ± 0.48	$\mathbf{80.17 \pm 0.29}$	$\mathbf{63.20 \pm 0.04}$
30%	9.61 ± 1.26	79.45 ± 0.32	62.99 ± 0.23	73.36 ± 0.42	48.25 ± 0.48	$\mathbf{80.17 \pm 0.29}$	$\mathbf{63.20 \pm 0.04}$
50%	9.61 ± 1.26	79.45 ± 0.32	62.99 ± 0.23	73.36 ± 0.42	48.25 ± 0.48	$\mathbf{80.17 \pm 0.29}$	$\mathbf{63.20 \pm 0.04}$
MIGDiff / Grocery							
5%	2.45 ± 1.03	75.38 ± 0.34	61.87 ± 1.66	60.35 ± 3.16	44.88 ± 2.10	$\mathbf{75.97 \pm 0.30}$	$\mathbf{62.43 \pm 1.67}$
10%	2.45 ± 1.03	75.38 ± 0.34	61.87 ± 1.66	60.35 ± 3.16	44.88 ± 2.10	$\mathbf{75.97 \pm 0.30}$	$\mathbf{62.43 \pm 1.67}$
20%	2.45 ± 1.03	75.38 ± 0.34	61.87 ± 1.66	60.35 ± 3.16	44.88 ± 2.10	$\mathbf{75.97 \pm 0.30}$	$\mathbf{62.43 \pm 1.67}$
30%	2.45 ± 1.03	75.38 ± 0.34	61.87 ± 1.66	60.35 ± 3.16	44.88 ± 2.10	$\mathbf{75.97 \pm 0.30}$	$\mathbf{62.43 \pm 1.67}$
50%	2.45 ± 1.03	75.38 ± 0.34	61.87 ± 1.66	60.35 ± 3.16	44.88 ± 2.10	$\mathbf{75.97 \pm 0.30}$	$\mathbf{62.43 \pm 1.67}$
MIGDiff / Movies							
5%	5.03 ± 0.01	49.74 ± 0.34	36.57 ± 0.56	41.63 ± 0.52	12.10 ± 0.71	$\mathbf{50.81 \pm 0.34}$	$\mathbf{37.48 \pm 0.60}$
10%	5.03 ± 0.01	49.74 ± 0.34	36.57 ± 0.56	41.63 ± 0.52	12.10 ± 0.71	$\mathbf{50.81 \pm 0.34}$	$\mathbf{37.48 \pm 0.60}$
20%	5.03 ± 0.01	49.74 ± 0.34	36.57 ± 0.56	41.63 ± 0.52	12.10 ± 0.71	$\mathbf{50.81 \pm 0.34}$	$\mathbf{37.48 \pm 0.60}$
30%	5.03 ± 0.01	49.74 ± 0.34	36.57 ± 0.56	41.63 ± 0.52	12.10 ± 0.71	$\mathbf{50.81 \pm 0.34}$	$\mathbf{37.48 \pm 0.60}$
50%	5.03 ± 0.01	49.74 ± 0.34	36.57 ± 0.56	41.63 ± 0.52	12.10 ± 0.71	$\mathbf{50.81 \pm 0.34}$	$\mathbf{37.48 \pm 0.60}$
MIGDiff / RedditM							
5%	5.00 ± 0.00	67.46 ± 0.64	62.37 ± 0.91	47.70 ± 0.02	44.20 ± 0.45	$\mathbf{69.15 \pm 0.46}$	$\mathbf{64.05 \pm 0.64}$
10%	10.00 ± 0.00	67.46 ± 0.64	62.37 ± 0.91	47.70 ± 0.02	44.20 ± 0.45	$\mathbf{69.86 \pm 0.42}$	$\mathbf{64.68 \pm 0.57}$
20%	12.58 ± 2.55	67.46 ± 0.64	62.37 ± 0.91	47.70 ± 0.02	44.20 ± 0.45	$\mathbf{70.52 \pm 0.27}$	$\mathbf{65.14 \pm 0.55}$
30%	12.58 ± 2.55	67.46 ± 0.64	62.37 ± 0.91	47.70 ± 0.02	44.20 ± 0.45	$\mathbf{70.52 \pm 0.27}$	$\mathbf{65.14 \pm 0.55}$
50%	12.58 ± 2.55	67.46 ± 0.64	62.37 ± 0.91	47.70 ± 0.02	44.20 ± 0.45	$\mathbf{70.52 \pm 0.27}$	$\mathbf{65.14 \pm 0.55}$
MIGDiff / RedditS							
5%	2.40 ± 1.47	91.50 ± 0.61	84.98 ± 1.15	76.64 ± 1.36	66.72 ± 1.17	$\mathbf{92.78 \pm 0.35}$	$\mathbf{86.08 \pm 0.92}$
10%	2.40 ± 1.47	91.50 ± 0.61	84.98 ± 1.15	76.64 ± 1.36	66.72 ± 1.17	$\mathbf{92.78 \pm 0.35}$	$\mathbf{86.08 \pm 0.92}$
20%	2.40 ± 1.47	91.50 ± 0.61	84.98 ± 1.15	76.64 ± 1.36	66.72 ± 1.17	$\mathbf{92.78 \pm 0.35}$	$\mathbf{86.08 \pm 0.92}$
30%	2.40 ± 1.47	91.50 ± 0.61	84.98 ± 1.15	76.64 ± 1.36	66.72 ± 1.17	$\mathbf{92.78 \pm 0.35}$	$\mathbf{86.08 \pm 0.92}$
50%	2.40 ± 1.47	91.50 ± 0.61	84.98 ± 1.15	76.64 ± 1.36	66.72 ± 1.17	$\mathbf{92.78 \pm 0.35}$	$\mathbf{86.08 \pm 0.92}$
MIGDiff / Toys							
5%	2.12 ± 0.19	75.47 ± 1.79	71.45 ± 1.33	63.13 ± 1.44	55.16 ± 1.41	$\mathbf{75.69 \pm 1.80}$	$\mathbf{72.11 \pm 1.28}$
10%	2.12 ± 0.19	75.47 ± 1.79	71.45 ± 1.33	63.13 ± 1.44	55.16 ± 1.41	$\mathbf{75.69 \pm 1.80}$	$\mathbf{72.11 \pm 1.28}$
20%	2.12 ± 0.19	75.47 ± 1.79	71.45 ± 1.33	63.13 ± 1.44	55.16 ± 1.41	$\mathbf{75.69 \pm 1.80}$	$\mathbf{72.11 \pm 1.28}$
30%	2.12 ± 0.19	75.47 ± 1.79	71.45 ± 1.33	63.13 ± 1.44	55.16 ± 1.41	$\mathbf{75.69 \pm 1.80}$	$\mathbf{72.11 \pm 1.28}$
50%	2.12 ± 0.19	75.47 ± 1.79	71.45 ± 1.33	63.13 ± 1.44	55.16 ± 1.41	$\mathbf{75.69 \pm 1.80}$	$\mathbf{72.11 \pm 1.28}$

I PER-DATASET POLICY COMPARISONS

Figures 9–13 report complete classifier-interface policy comparisons for all five operators, six datasets, five nominal budgets, and both evaluation metrics. Each point is the mean for that operator–dataset–budget cell; datasets are not pooled in these figures. The panels display means without error bars. Rand-b_{eff} matches R-SMR's realized call count, and Rand-$b_{\max}$ exhausts the nominal budget; the random-policy values average 100 draws within each seed. Lines connect evaluated budgets, and vertical scales vary across panels. Figure 4 selects the Accuracy panels for NTSFormer and MIGDiff on Grocery and RedditM.

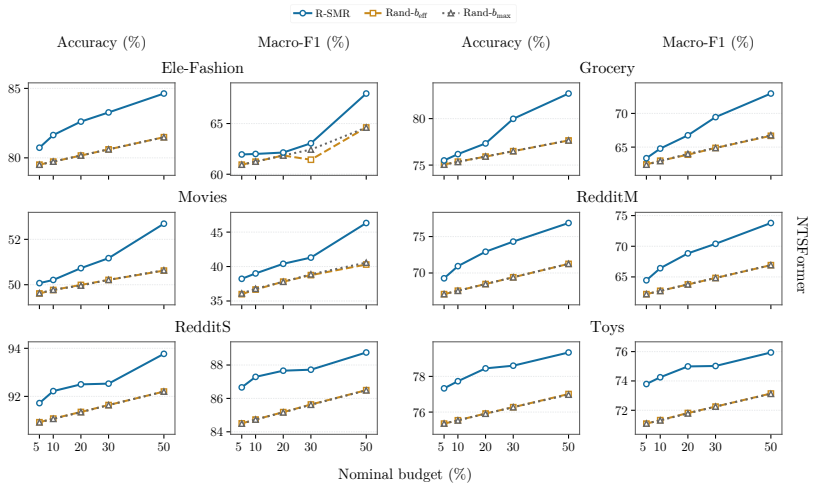


Figure 9: Per-dataset policy comparison for NTSFormer.

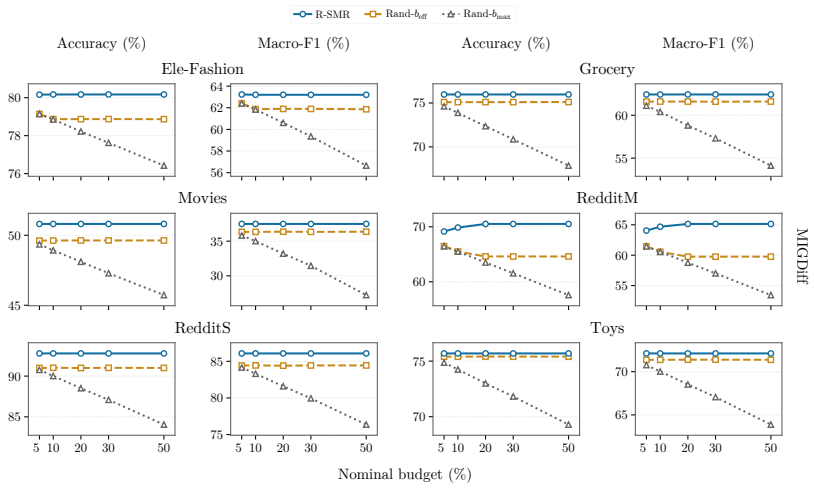


Figure 10: Per-dataset policy comparison for MIGDiff.

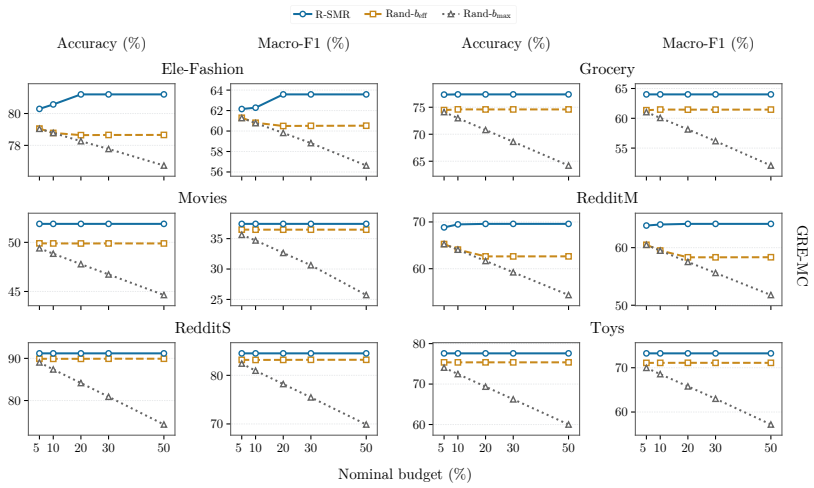


Figure 11: Per-dataset policy comparison for GRE-MC.

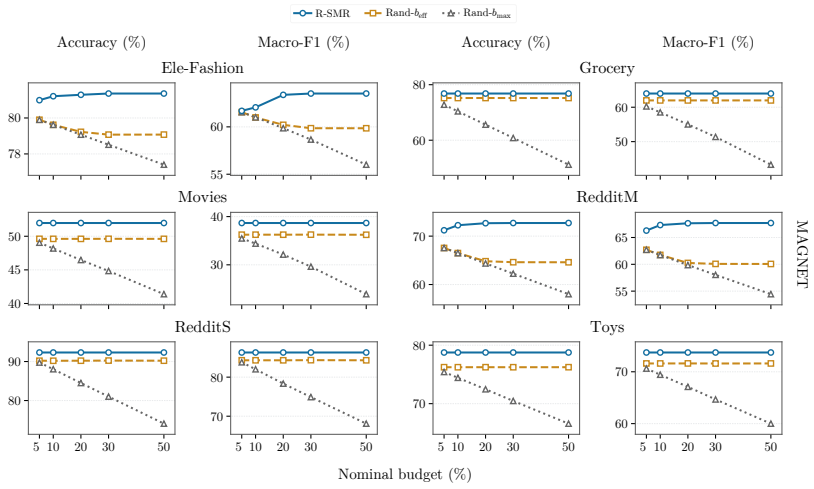


Figure 12: Per-dataset policy comparison for MAGNET.

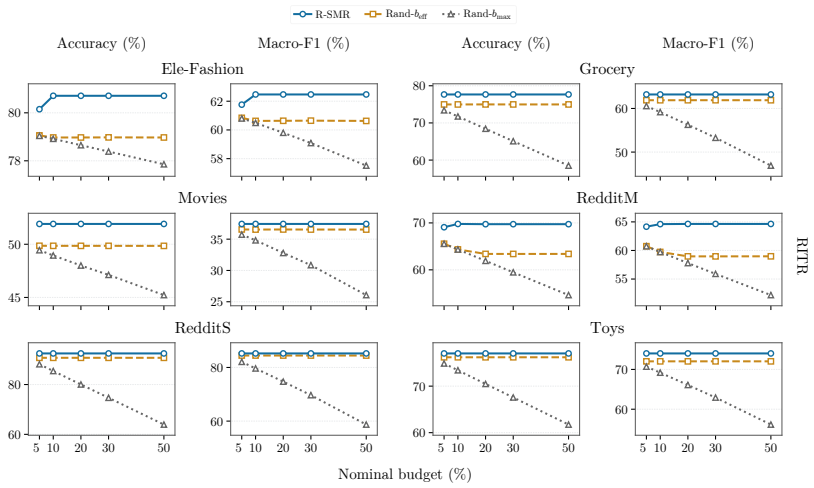


Figure 13: Per-dataset policy comparison for RITR.

J Additional Component Ablations

J.1 Component Ablation Definitions and Protocol

All component ablations use exactly the same cheap- and expensive-endpoint logits, candidate node sets, and random seeds as Full. The endpoint models remain fixed. Each variant is evaluated after refitting its router and recalibrating its probabilities using the same training and validation partitions as Full. For variants that retain the harm penalty, λ is reselected on validation data under that variant's routing rule. No harm penalty and correction-only fix $\lambda = 0$ rather than tuning it. All variants use the same 20% nominal budget cap and are evaluated on the same test nodes, following the data-split protocol in the experimental setup.

Table 11 specifies the four variants. No harm penalty still learns all four correctness states using a multinomial objective; only its decision score omits the harm term. Correction-only instead learns a binary logistic-regression model with $\mathbb{I}[t_v = (0, 1)]$ as the target, pooling both-correct, harm, and both-wrong nodes into the negative class. Its separately calibrated correction probability is denoted by $\widehat{q}_v$. Thus, the two variants differ in supervision and probability estimation, although both rank nodes by a predicted correction probability.

Table 11: Definitions of the component ablations. All variants share the endpoint logits, node sets, seeds, and nominal budget. Positive-score filtering selects at most K_b eligible nodes; removing it selects the highest-scoring K_b candidates even when some scores are nonpositive.

Variant	Prediction target	Routing score	Eligible nodes
Full	Four correctness states	$\widehat{P}_v((0,1)) - \lambda\widehat{P}_v((1,0))$	Positive score
No harm penalty	Four correctness states	$\widehat{P}_v((0,1))$; $\lambda = 0$	Positive score
Correction-only	Correction vs. all other states	$\widehat{q}_v$; $\lambda = 0$	Positive score
No positive-score filtering	Four correctness states	$\widehat{P}_v((0,1)) - \lambda\widehat{P}_v((1,0))$	All candidates

J.2 Complete Component Ablation Results

Figure 14 reports complete component ablations for all five operators and both metrics; Figure 5 shows the Accuracy panels for NTSFormer and MIGDiff.

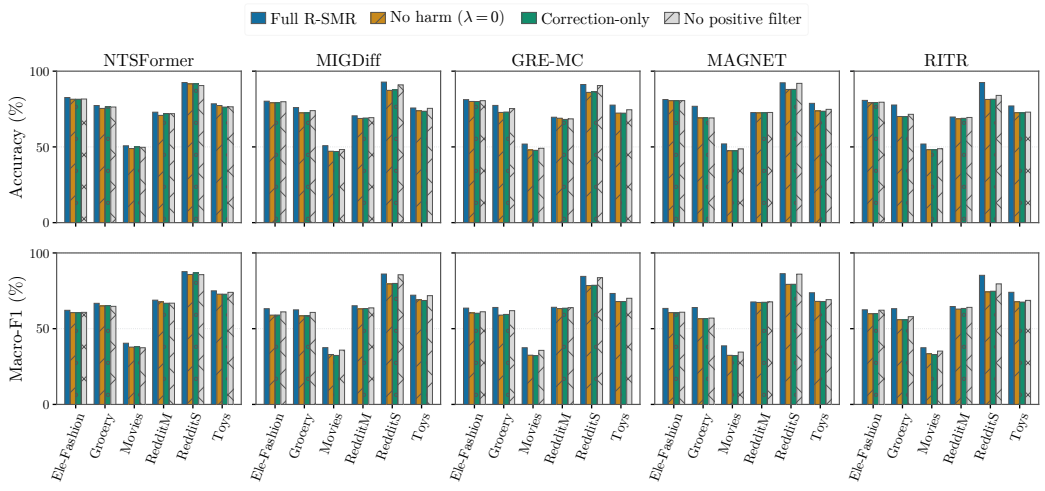


Figure 14: Core-component ablations for all five operators across six datasets at the classifier interface and a 20% call ceiling. Bars compare full R-SMR, no harm penalty ($\lambda = 0$), correction-only routing, and no positive-score filtering. Rows show Accuracy and Macro-F1; Figure 5 selects the NTSFormer and MIGDiff Accuracy panels.

K Additional Routing Feature Sensitivity Results

We report all remaining panels of the 50%-ceiling feature analysis in Section 6.5. The main text selects NTSFormer and MIGDiff on RedditM and RedditS for visible differences; the panels below retain the other settings, including small or reversed effects. Each router is refitted after removing degree and log-degree together, soft homophily, cheap predictive entropy, cheap-neighbor feature similarity, or available-neighbor fraction. The missing-modality indicator is retained. Recovery operators and endpoint predictions remain frozen. A 50% ceiling is used both when selecting the harm penalty on validation data and when routing evaluation nodes. Nonpositive scores are rejected, so realized calls may be lower.

The following figures report changes in mean Accuracy and Macro-F1 relative to Full R-SMR for each operator and dataset. Negative values indicate lower performance after feature removal, whereas positive values indicate higher performance. Numbers at the bar ends give these differences in percentage points. Deg, Homo, Uncert, Nbr Sim, and Nbr Avail denote removal of degree (including log-degree), soft homophily, uncertainty, neighbor similarity, and available-neighbor frac-

tion, respectively. The complete results include smaller and reversed effects outside the settings highlighted in the main text, indicating that feature contributions depend on the operator, dataset, and metric.

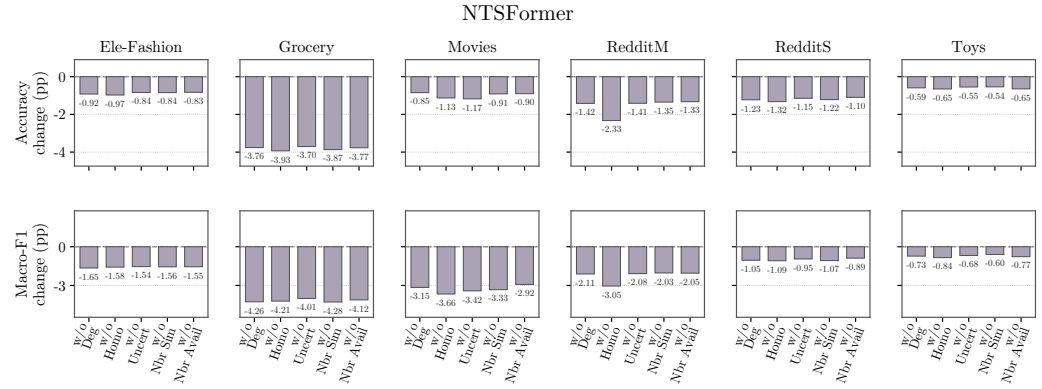


Figure 15: NTSFormer routing-feature sensitivity at a 50% call ceiling. Bars show changes in mean performance relative to Full R-SMR (pp). Each row shares a vertical scale.

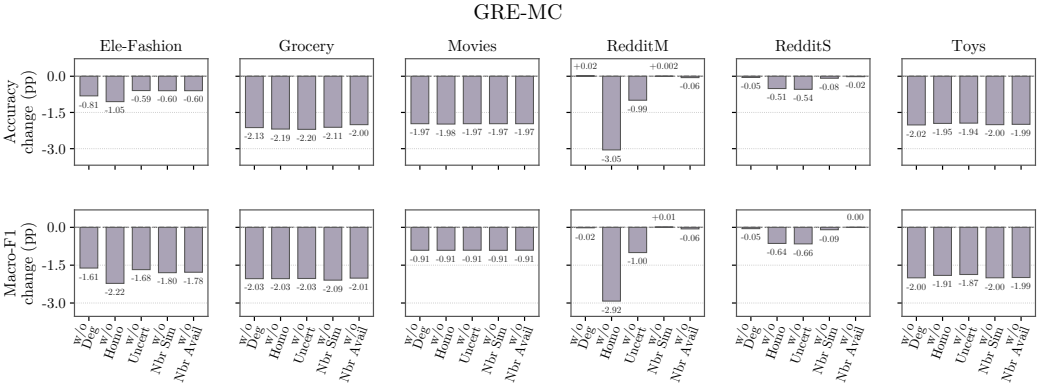


Figure 16: GRE-MC routing-feature sensitivity at a 50% call ceiling. Bars show changes in mean performance relative to Full R-SMR (pp). Each row shares a vertical scale.

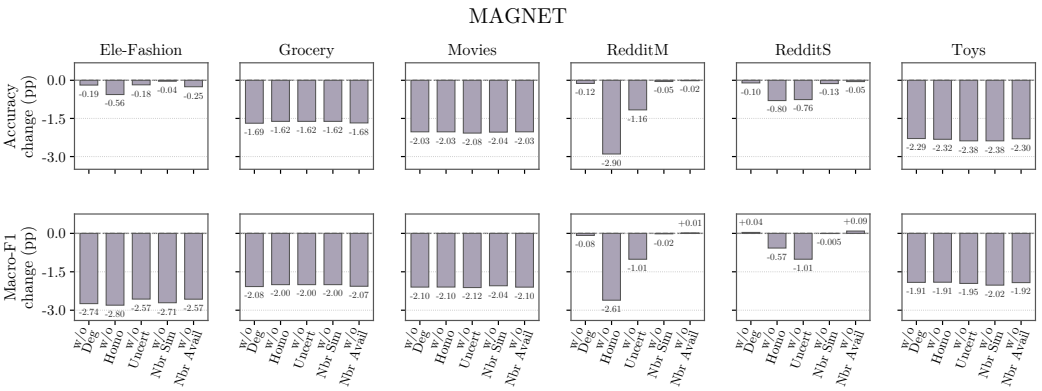


Figure 17: MAGNET routing-feature sensitivity at a 50% call ceiling. Bars show changes in mean performance relative to Full R-SMR (pp). Each row shares a vertical scale.

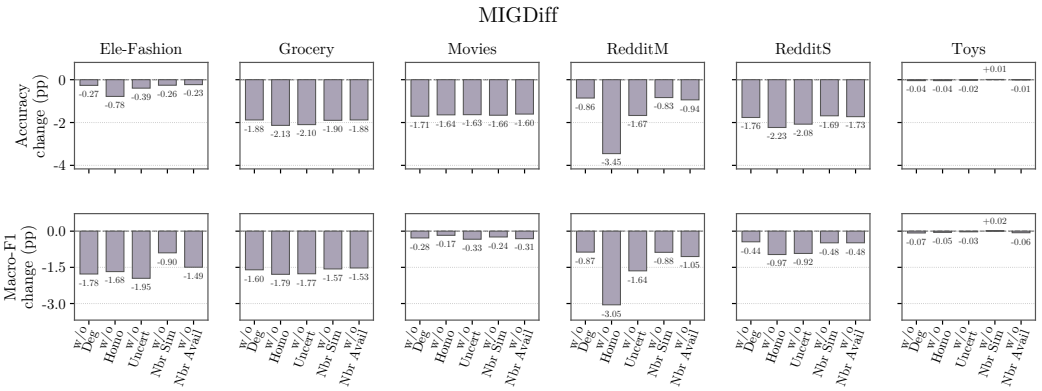


Figure 18: MIGDiff routing-feature sensitivity at a 50% call ceiling. Bars show changes in mean performance relative to Full R-SMR (pp). Each row shares a vertical scale.

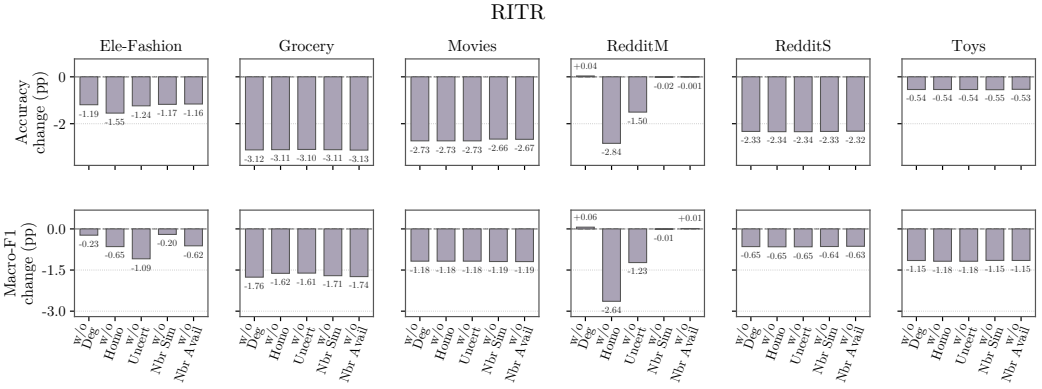


Figure 19: RITR routing-feature sensitivity at a 50% call ceiling. Bars show changes in mean performance relative to Full R-SMR (pp). Each row shares a vertical scale.