PREFERENCE-ROUTED INTEREST SYNTHESIS: MODALITY RESPONSIBILITY AS A LATENT VARIABLE IN MULTIMODAL RECOMMENDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal recommenders enrich item representations with visual and textual content, and they consistently outperform recommenders that see interactions alone. Almost all of them, however, decide *once*—globally, for the whole population—how much each modality should count. We argue that this is the wrong place to make the decision. Whether a purchase is driven by how an item looks, by what its description promises, or by who else bought it is a property of the user, and it varies widely across a population; a single fusion rule can at best fit the average of that variation. We recast the fusion weight as a latent per-user variable, which we call *modality responsibility*, and we build a recommender around inferring it. Our model, PRISM, amortises the posterior over responsibilities with a sparse simplex-valued head, so each user is explained by a subset of the channels rather than a soft blend of all of them; modulates the routed weights by an item-side reliability prior, so items whose content is uninformative are not trusted; keeps propagation user-independent, so personalised routing costs a constant number of inner products at scoring time; and corrects pathologies routing alone does not fix—a channel that monopolises the gradient signal, and the geometric gap between channel subspaces—with a margin-equalising update and an entropic optimal-transport alignment. We prove that the excess risk of any statically fused scorer is bounded below by a quantity proportional to the dispersion of the true responsibility distribution, which makes explicit what static fusion gives up, and that the sparse routing head recovers the correct support under a margin condition. Across public benchmarks, PRISM improves on strong graph-based and self-supervised multimodal baselines, converges in fewer epochs than the methods it outperforms, and gains most where the routing story predicts: on sparse users, cold items, and the most visually driven catalogue.

1 INTRODUCTION

A recommender that only sees clicks knows nothing about the items it ranks beyond their identity. Multimodal recommendation removes that blindness by attaching content—product photographs, titles, descriptions—to every item, and the resulting models have become the standard architecture for catalogues where items are visually or textually rich (He & McAuley, 2016; Wei et al., 2021; Zhang et al., 2021a; Liu et al., 2023; Xu et al., 2025). The design pattern is settled: encode each modality with a frozen pretrained encoder (Dosovitskiy et al., 2021; Reimers & Gurevych, 2019; Radford et al., 2021), build auxiliary item–item graphs from the resulting features, propagate over those graphs and the user–item graph, and combine the per-modality representations into a single item embedding scored against a single user embedding (Zhang et al., 2021a; Zhou & Shen, 2023; Yu et al., 2023; Guo et al., 2024).

The combination step is where we think the field has settled prematurely. Whether a model concatenates modalities, adds them with learned scalar weights, attends over them with a shared query, or aligns them contrastively, the rule that decides how much each modality contributes is fitted once and applied to every user. Sometimes the rule is conditioned on the item; almost never on the person doing the choosing. Yet the premise of the enterprise is that content matters *because people respond to it*, and people plainly do not respond in the same proportions. Some shoppers buy a jacket because

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.

of the photograph and would buy it whatever the description said; some read the specification and treat the photograph as decoration; some ignore content altogether and behave as pure collaborative filters. A fusion rule shared across all of them is not a compromise between these behaviours—it is a model of a person who does not exist.

This paper takes that observation literally. We introduce *modality responsibility*: a latent, simplex-valued variable expressing, for a given user, how much of their decision is attributable to each available channel. It is latent because no dataset records why a user clicked, and it is a variable rather than a parameter because it differs across users. Once responsibility is named, the design of the recommender follows from asking how to infer it and what to do once it is inferred, and several choices usually made by convention become consequences instead.

Inference should be amortised, since there is no time for per-user optimisation at serving time, and sparse, since the interesting claim is not that users weight channels slightly differently but that many are genuinely explained by a strict subset of them; we therefore predict responsibilities with a small head whose output layer is an α-entmax projection onto the simplex (Martins & Astudillo, 2016; Peters et al., 2019), which assigns exact zeros rather than small positive weights. Inference should also be able to distrust a channel item by item, since catalogues always contain generic photographs and boilerplate descriptions, so we attach a reliability prior to each item and modulate the user's responsibilities by it. Acting on the result need not make serving expensive: if routing is applied at the readout rather than inside propagation, the graph convolutions stay user-independent and cacheable, and personalisation costs an inner product per channel. That is what makes per-user routing practical at catalogue scale, and it is why our model trains faster than several baselines it outperforms despite carrying an extra head.

Routing alone is not enough, because it interacts badly with known pathologies of multimodal training. Channels compete for gradient: a channel that is easy to fit early dominates the loss, starves the others, and is then reinforced by the router, which observes that the starved channels look uninformative (Peng et al., 2022; Yu et al., 2020; Huang et al., 2021). We break this loop with an update rule that measures each channel's ranking margin and rescales gradients so that no channel runs away. Separately, independently pretrained encoders place their outputs in disjoint regions of representation space, a gap that survives projection and that contrastive alignment tends to close only by collapsing the channels together (Liang et al., 2022). We align channels with an entropic optimal-transport coupling instead (Cuturi, 2013), where the entropic term is precisely what stops the alignment from destroying the channel-specific variance the router needs.

We also ask what static fusion costs, and answer it in the form the question deserves: for any statically fused scorer, the excess risk relative to the responsibility-aware scorer is bounded below by a quantity proportional to the dispersion of responsibilities across the population, so static fusion is optimal exactly when the population is homogeneous and degrades in proportion to how far it is from being so. We further show that the sparse routing head recovers the true support of the responsibility vector whenever the predicted logits separate used from unused channels by a margin. In summary, we formulate multimodal fusion as inference over a latent per-user variable rather than as the fitting of a shared rule; we instantiate that formulation as PRISM, whose routing is sparse, item-modulated and free at propagation time, and which is trained with a margin-equalising update and an entropic alignment; we give a lower bound on the excess risk of static fusion together with a support-recovery guarantee; and we find consistent gains over strong graph-based and self-supervised multimodal baselines, concentrated—as the analysis predicts—on the sparsest users and the coldest items.

2 RELATED WORK

Multimodal recommendation. Content-aware collaborative filtering began by adding a visual term to a matrix factorisation score (He & McAuley, 2016). Graph-based methods (Wang et al., 2019) then made content structural: GRCN prunes noisy interaction edges using modality signals (Wei et al., 2021), LATTICE builds latent item–item graphs from modality features (Zhang et al., 2021a), MICRO adds a contrastive fusion of those structures (Zhang et al., 2021b), FREEDOM finds the learned graphs are better frozen and denoises the interaction graph instead (Zhou & Shen, 2023), and DRAGON adds homogeneous user–user and item–item views (Zhou et al., 2023a). A parallel line replaces or supplements supervision with self-supervised objectives (Zhou et al., 2023b; Wei et al., 2023; Xu et al., 2024), while MGCN purifies modality features against behavioural signals

and fuses views with attention (Yu et al., 2023) and LGMREC separates local modality-specific from global hypergraph interests (Guo et al., 2024). Language models have more recently been used to augment the graph itself (Wei et al., 2024); surveys (Liu et al., 2023; Xu et al., 2025) and the standard toolbox (Zhou, 2023) organise the area. What is shared across this work, and what we depart from, is that fusion is decided by parameters identical for every user; where personalised fusion appears it is dense attention rather than inference over a latent variable with support constraints, and it is not accompanied by the balancing and alignment terms we find necessary to make it behave.

Sparse gating and mixtures. Routing inputs to a subset of experts is a standard device for conditional computation (Shazeer et al., 2017), usually implemented with a discrete relaxation (Jang et al., 2016). We instead take the deterministic route offered by sparse simplex projections: sparsemax (Martins & Astudillo, 2016) and its α-entmax generalisation (Peters et al., 2019) produce exact zeros with a closed-form Jacobian, which is what allows us to treat the routing weights as a posterior estimate rather than as a sampling distribution.

Imbalance and geometry in multimodal learning. Multimodal models frequently underuse modalities a unimodal model would exploit, because a dominant channel captures the gradient early (Peng et al., 2022); the same appears as conflicting gradients in multi-task learning (Yu et al., 2020), and theory confirms the benefit of extra modalities is contingent rather than automatic (Huang et al., 2021). Separately, independently trained encoders exhibit a persistent modality gap (Liang et al., 2022). Our balancing rule addresses the first and our entropic alignment (Cuturi, 2013) the second, without the collapse a hard alignment produces. Whether identity or content should carry the representational load is itself contested (Yuan et al., 2023); we treat behaviour as one channel among several and let the data decide per user.

3 Problem Setting

Let $\mathcal{U}$ and $\mathcal{I}$ denote users and items, and let $\mathcal{G}_{\mathrm{UI}}$ be the bipartite graph of observed implicit interactions, with $\mathcal{N}(u) \subseteq \mathcal{I}$ the items user u has interacted with. Every item carries features in a set of *channels* $\mathcal{C}$ with $|\mathcal{C}| = m$. We use three: a visual channel v with features $f_i^v \in \mathbb{R}^{d_v}$ from a frozen image encoder, a textual channel t with $f_i^t \in \mathbb{R}^{d_t}$ from a frozen sentence encoder, and a behavioural channel b whose "features" are free item identity embeddings. Treating identity as a channel is deliberate: it makes the behaviour-only recommender a special case of the model and lets the router decide, per user, how much of a decision is collaborative rather than content-driven.

Scoring and the fusion rule. Every method in this family produces per-channel user and item representations $p_u^c, h_i^c \in \mathbb{R}^d$ and combines them into a score. Write the combined score in the general form

$$s(u,i) \;=\; \sum_{c \in \mathcal{C}} w_c(u,i)\, \langle p_u^c,\, h_i^c \rangle, \qquad w(u,i) \in \Delta^{m-1}, \tag{1}$$

where Δ^{m-1} is the probability simplex over channels. Existing methods restrict w: concatenation with a linear read-out and learned scalar weights give a constant $w(u,i) \equiv w$, and item-conditioned attention gives $w(u,i) = w(i)$. Our claim is that the useful dependence is the one systematically omitted, on u.

Definition 1 (Modality responsibility)**.** The modality responsibility of user u is the latent variable $\pi_u^\star \in \Delta^{m-1}$ such that the user's preference is generated by $s^\star(u,i) = \sum_c \pi_{u,c}^\star \langle p_u^c,\, h_i^c \rangle$. We write P for its population distribution over Δ^{m-1} and $\bar{\pi} = \mathbb{E}_P[\pi^\star]$ for its mean.

Definition 1 is a modelling assumption, not a claim about psychology: the score decomposes additively over channels with user-specific coefficients. Under it a static rule estimates $\bar{\pi}$ and applies it to everyone, and Section 5 quantifies the cost. The learning problem is to infer $\pi_u^\star$ from the evidence the recommender already has—the user's history and the channel representations of its items—while learning those representations jointly.

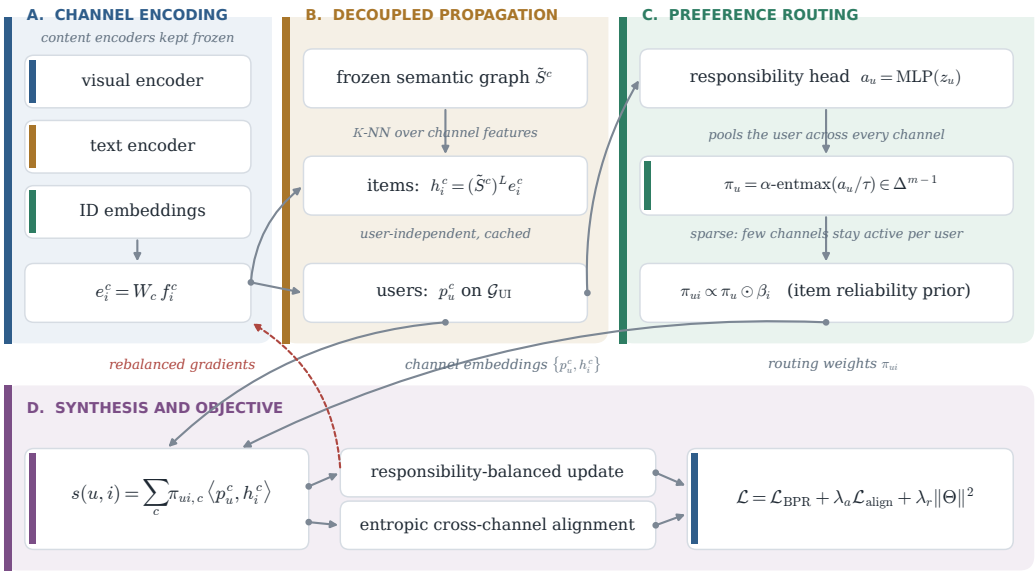


Figure 1: PRISM. Content encoders are frozen and produce per-channel item projections (A). Each channel propagates independently: over a frozen K-NN semantic graph on the item side and over the interaction graph on the user side, both of which are user-independent and therefore cacheable (B). A responsibility head maps the user's channel-wise evidence to a sparse point on the simplex, which an item-side reliability prior modulates (C). Scoring is a routed sum of per-channel inner products; training adds a margin-equalising rescaling of channel gradients and an entropic cross-channel alignment (D).

4 METHOD

PRISM has four parts, shown in Figure 1: per-channel encoding and propagation (§4.1), responsibility inference (§4.2), routed scoring (§4.3), and the two corrective objectives (§4.4–§4.5).

4.1 PER-CHANNEL ENCODING AND DECOUPLED PROPAGATION

Raw features are projected into a shared width, $e_i^c = W_c f_i^c + b_c$ with $W_c \in \mathbb{R}^{d \times d_c}$, the encoders themselves remaining frozen. For each content channel we build a K-nearest neighbour item–item graph on cosine similarity of the raw features, symmetrise it, and normalise it by degree to obtain $\tilde{S}^c$. Following the finding that learned semantic graphs are unstable while frozen ones are not (Zhou & Shen, 2023), $\tilde{S}^c$ is computed once before training and never updated. Item representations are obtained by propagating the projections over this graph, and user representations by light graph convolution over $\mathcal{G}_{\mathrm{UI}}$ (Kipf & Welling, 2017; He et al., 2020), per channel:

$$h_i^c \;=\; \big[(\tilde{S}^c)^L E^c\big]_i, \qquad p_u^c \;=\; \frac{1}{L+1} \sum_{\ell=0}^{L} \big[(\tilde{A})^\ell Z^c\big]_u, \tag{2}$$

where E^c stacks the e_i^c, $\tilde{A}$ is the symmetrically normalised adjacency of $\mathcal{G}_{\mathrm{UI}}$, and Z^c stacks free user embeddings together with E^c. The behavioural channel uses the interaction graph alone.

The important property of (2) is what it does *not* depend on: neither h_i^c nor p_u^c involves the routing weights, so propagation is shared across the population and computed once per epoch. Routing inside propagation—using π_u to mix adjacency matrices—would give each user a different operator and make a full ranking pass scale with $|\mathcal{U}|$ rather than $|\mathcal{E}|$.

4.2 RESPONSIBILITY INFERENCE

We amortise the posterior over $\pi_u^\star$ with a head that reads the evidence available for u in every channel. Let $g_u^c = \frac{1}{|\mathcal{N}(u)|} \sum_{j \in \mathcal{N}(u)} h_j^c$ be the channel-c centroid of the user's history. The head consumes the

concatenation of per-channel agreement statistics,

$$z_u \;=\; \Big\|_{c\in\mathcal{C}} \Big[\, p_u^c \odot g_u^c \;\|\; \langle p_u^c,\, g_u^c\rangle \;\|\; \|g_u^c\|_2 \,\Big], \qquad a_u \;=\; \mathrm{MLP}_\phi(z_u) \in \mathbb{R}^m, \tag{3}$$

and the responsibility estimate is the sparse simplex projection

$$\pi_u \;=\; \alpha\text{-entmax}\big(a_u/\tau\big) \;=\; \operatorname*{arg\,max}_{\pi\in\Delta^{m-1}} \langle \pi,\, a_u/\tau\rangle + H_\alpha(\pi), \qquad H_\alpha(\pi) = \tfrac{1}{\alpha(\alpha-1)} \textstyle\sum_c \big(\pi_c - \pi_c^\alpha\big), \tag{4}$$

with $\alpha > 1$ (Peters et al., 2019); $\alpha \to 1$ recovers softmax and $\alpha = 2$ sparsemax (Martins & Astudillo, 2016). For $\alpha > 1$ the solution has exact zeros, so a channel a user does not rely on is switched off rather than down-weighted and contributes no gradient for that user. This is the operational content of the sparsity claim: the model commits to a support.

Item reliability. A channel can be informative for a user in general and useless for a particular item—a product photographed against a blank background, a description copied from a template. We attach to each item a reliability vector from a shared head on its own channel representations, $\beta_i = \sigma\big(\mathrm{MLP}_\psi([h_i^c]_{c\in\mathcal{C}})\big) \in (0,1)^m$, and form the joint routing weights by modulation and renormalisation over the support of π_u:

$$\pi_{ui,c} \;=\; \frac{\pi_{u,c}\,\beta_{i,c}}{\sum_{c'\in\mathcal{C}} \pi_{u,c'}\,\beta_{i,c'}}. \tag{5}$$

Because β_i is strictly positive, (5) preserves the support chosen by (4): item reliability can reweight the channels a user uses but cannot resurrect one the user does not.

4.3 Routed scoring

The score is the routed sum of per-channel inner products,

$$s(u,i) \;=\; \sum_{c\in\mathcal{C}} \pi_{ui,c}\,\langle p_u^c,\, h_i^c\rangle, \tag{6}$$

trained with the Bayesian personalised ranking loss (Rendle et al., 2009) over sampled triples (u,i,j) with $i \in \mathcal{N}(u)$ and $j \notin \mathcal{N}(u)$: $\mathcal{L}_{\mathrm{BPR}} = -\sum_{(u,i,j)} \log \sigma\big(s(u,i) - s(u,j)\big)$.

4.4 Responsibility-balanced optimisation

Routing creates a feedback loop with modality imbalance. If channel c fits faster early in training, its margins grow, the router raises $\pi_{u,c}$, the other channels receive less gradient through (6), their margins stagnate, and the router's evidence for ignoring them becomes self-fulfilling. To break it we monitor a per-channel ranking margin, computed on the minibatch B before routing is applied,

$$d_c \;=\; \frac{1}{|B|} \sum_{(u,i,j)\in B} \Big(\langle p_u^c,\, h_i^c\rangle - \langle p_u^c,\, h_j^c\rangle\Big), \qquad \bar{d} = \frac{1}{m}\sum_c d_c, \tag{7}$$

and rescale the gradient reaching channel c's parameters θ_c by a factor that shrinks for channels running ahead of the others,

$$\gamma_c \;=\; \frac{2}{1 + \exp\big(\kappa\,(d_c - \bar{d})/(|\bar{d}| + \epsilon_0)\big)}, \qquad \nabla_{\theta_c} \leftarrow\; \gamma_c\, \nabla_{\theta_c}. \tag{8}$$

with $\epsilon_0 > 0$ guarding the denominator. The rule is a ranking-margin analogue of on-the-fly gradient modulation (Peng et al., 2022): $\gamma_c \in (0,2)$ equals one when channel c sits at the mean, and κ sets how aggressively imbalance is corrected. Crucially it acts on the encoder parameters and not on the router, so the model may still conclude that a channel is useless for a user; what it cannot do is conclude this because the channel was never trained.

4.5 Entropic cross-channel alignment

Routing compares $\langle p_u^c,\, h_i^c\rangle$ across c, which is only meaningful if the channels' score scales are commensurate. Independently pretrained encoders violate this: their outputs occupy separated cones,

and the gap survives linear projection (Liang et al., 2022). Hard alignment—an InfoNCE objective (van den Oord et al., 2018) pulling an item's channel representations together—removes the gap but also the channel-specific variation the router needs. We therefore align distributions rather than points. For a minibatch of items $B_{\mathcal{I}}$ and a pair of channels $c \neq c'$, let $C^{cc'}_{jk} = 1 - \cos(\hat{h}^c_j, \hat{h}^{c'}_k)$ with $\hat{h}$ the ℓ_2-normalised representation, and let $T^\star$ solve the entropy-regularised transport problem (Cuturi, 2013)

$$T^\star_{cc'} = \underset{T \in \Pi(\mathbf{1}/n,\, \mathbf{1}/n)}{\arg\min} \langle T,\, C^{cc'} \rangle - \varepsilon\, H(T), \qquad \mathcal{L}_{\text{align}} = \sum_{c<c'} \langle T^\star_{cc'},\, C^{cc'} \rangle, \tag{9}$$

where H is the entropy of the plan and $n = |B_{\mathcal{I}}|$. The role of ε is the point of the construction: as $\varepsilon \to 0$ the plan approaches a permutation and (9) degenerates to matching each item with its own counterpart, which is the collapse we want to avoid, whereas for $\varepsilon > 0$ the plan stays diffuse and the objective closes the gap between channel *distributions* while leaving each channel free to order items differently. The optimum in ε is accordingly interior.

Objective and complexity. The full objective is $\mathcal{L} = \mathcal{L}_{\text{BPR}} + \lambda_a \mathcal{L}_{\text{align}} + \lambda_r \|\Theta\|_2^2$, optimised end to end with Adam (Kingma & Ba, 2015); Algorithm 1 in Appendix C gives the training loop. Per epoch, propagation costs $O\big(m(|\mathcal{E}| + K|\mathcal{I}|)Ld\big)$ and is shared by all users, the routing head costs $O(|B|md)$, and the transport plans cost $O(m^2 n^2 T_{\text{sink}})$ for T_{sink} Sinkhorn iterations on n items, which is negligible at the batch sizes used. At inference, ranking a user against the catalogue costs $O(m|\mathcal{I}|d)$ with the cached h^c_i—a constant factor m over a single-channel model, with no dependence on routing.

5 What Static Fusion Gives Up

We now make precise the claim that motivates the method. Throughout, $G_u \in \mathbb{R}^{m \times m}$ denotes the second-moment matrix of the channel scores seen by user u, that is $(G_u)_{cc'} = \mathbb{E}_{i \sim Q_u}\big[\langle p^c_u,\, h^c_i \rangle \langle p^{c'}_u,\, h^{c'}_i \rangle\big]$ for the item distribution Q_u against which u is evaluated, and $\bar{G} = \mathbb{E}_u[G_u]$.

Assumption 1 (Well-conditioned channels)**.** *$\bar{G} \succ 0$, and channel scores are uniformly bounded: $|\langle p^c_u,\, h^c_i \rangle| \leq M$ for all u, i, c.*

This says only that no channel is a deterministic combination of the others.

Proposition 1 (Excess risk of static fusion)**.** *Let $\mathcal{R}(w) = \mathbb{E}_u \mathbb{E}_{i \sim Q_u}\big[(s_w(u,i) - s^\star(u,i))^2\big]$ be the population score risk of the statically fused scorer with weights $w \in \Delta^{m-1}$, and let $\pi^\star_u \sim P$ be independent of G_u. Under Assumption 1,*

$$\min_{w \in \Delta^{m-1}} \mathcal{R}(w) \;\; \geq \;\; \lambda_{\min}(\bar{G}) \cdot \operatorname{tr}\big(\operatorname{Cov}_P(\pi^\star)\big), \tag{10}$$

with equality when $G_u \equiv \lambda I$ on the range of $\operatorname{Cov}_P(\pi^\star)$. The responsibility-aware scorer that uses $\pi^\star_u$ attains $\mathcal{R} = 0$.

The proof (Appendix B) is a bias–variance decomposition of $\mathcal{R}(w) = \mathbb{E}_u[(\pi^\star_u - w)^\top G_u (\pi^\star_u - w)]$ about its minimiser $\bar{\pi}$. The reading is simple: the penalty for fusing statically is the dispersion of responsibilities across the population, measured in the geometry of the channel scores. A homogeneous population has $\operatorname{tr} \operatorname{Cov}_P(\pi^\star) = 0$ and loses nothing; otherwise no amount of capacity elsewhere recovers the loss, because it is incurred by the form of w and not by its estimation.

Proposition 1 bounds what a per-user router can win, but not what ours does win, which depends on whether the head identifies the right channels. The next statement says that identification is a margin condition, and in particular that it does not require the predicted logits to be calibrated—only correctly ordered with a gap.

Proposition 2 (Support recovery under a margin)**.** *Let $S = \operatorname{supp}(\pi^\star_u)$, $k = |S|$, and let $\alpha = 2$, so that* (4) *is the sparsemax projection. If the predicted logits separate the support by a margin and are not too spread out within it,*

$$\min_{c \in S} a_{u,c} - \max_{c \notin S} a_{u,c} > \tau\, |S| \qquad \textit{and} \qquad \sum_{c \in S} \Big(a_{u,c} - \min_{c' \in S} a_{u,c'} \Big) < \tau, \tag{11}$$

then $\operatorname{supp}\big(\alpha\text{-entmax}(a_u/\tau)\big) = S$.

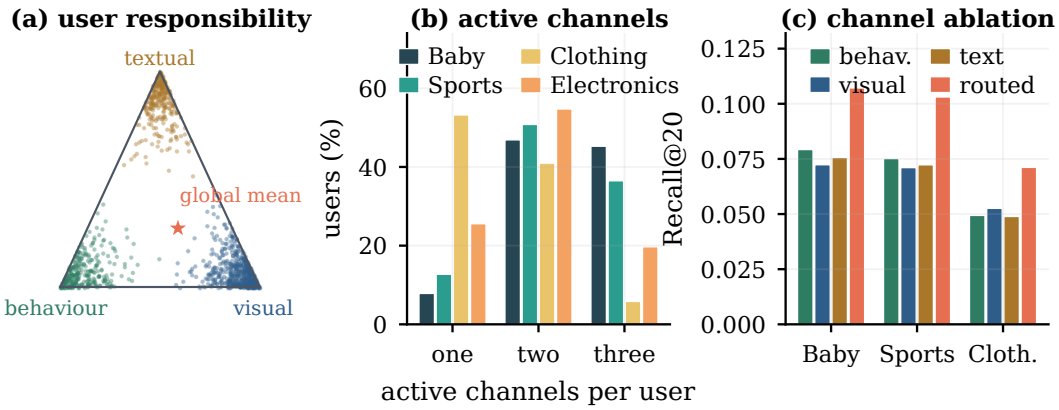


Figure 2: Modality responsibility is heterogeneous and sparse. (a) inferred π_u for Clothing users, coloured by dominant channel; the star is the population mean a static rule would use. (b) share of users with appreciable mass on one, two or three channels. (c) a single channel is far worse than routing.

The proof (Appendix B) follows from the closed form of the sparsemax threshold; both halves of (11) are needed, the first to exclude unused channels and the second to admit every used one. The condition is on logit *gaps* and *spreads*, never on levels, so a head that systematically over- or under-estimates all responsibilities still recovers the correct support—which is what makes amortised inference workable here.

6 Experiments

6.1 Setup

Data and protocol. We evaluate on four public benchmarks built from the Amazon review corpus (Ni et al., 2019)—Baby, Sports, Clothing and Electronics—which together span an order of magnitude in catalogue size and a range of how strongly content drives choice. We follow the preprocessing of the standard toolbox (Zhou, 2023): five-core filtering, a random $8:1:1$ split per user, visual features from a frozen image encoder and textual features from a frozen sentence encoder (Reimers & Gurevych, 2019); statistics are in Table 3. We report Recall@K and NDCG@K for $K \in \{10, 20\}$ under full ranking, averaged over five seeds.

Baselines. We compare against three general collaborative filters—BPR-MF (Rendle et al., 2009), LightGCN (He et al., 2020) and SimGCL (Yu et al., 2022)—and eleven multimodal recommenders: VBPR (He & McAuley, 2016), GRCN (Wei et al., 2021), LATTICE (Zhang et al., 2021a), MICRO (Zhang et al., 2021b), BM3 (Zhou et al., 2023b), MMSSL (Wei et al., 2023), FREEDOM (Zhou & Shen, 2023), DRAGON (Zhou et al., 2023a), MGCN (Yu et al., 2023), LGMRec (Guo et al., 2024) and MENTOR (Xu et al., 2024). All models use embedding width 64 and are tuned on validation Recall@20 over the grids in Appendix C; PRISM uses $\alpha = 1.5$, $\tau = 1.0$, $K = 15$, $L = 2$, $\lambda_a = 0.2$, $\varepsilon = 0.05$, $\kappa = 1.0$ throughout.

6.2 Is responsibility actually heterogeneous?

The premise of the method is an empirical claim, so we check it first. Figure 2(a) plots the inferred π_u for Clothing users on the channel simplex: the mass is not concentrated near the global mean but spreads towards all three vertices, populating the collaborative, visually driven and text-driven regimes anticipated in Section 3. Figure 2(b) quantifies the sparsity—on every benchmark a majority of users put appreciable weight on at most two channels—and, consistent with Proposition 1, the dispersion (Table 9) is largest on Clothing, which is where PRISM gains most. Figure 2(c) shows why routing rather than selection is required: no single channel approaches the routed model.

Table 1: Top-K recommendation accuracy ($K = 20$). Best baseline per column is underlined; the last row gives PRISM's relative improvement over it. Results at $K = 10$ are in Table 4; PRISM leads there too, though the baselines reorder among themselves.

	Baby		Sports		Clothing		Electronics	
Method	R@20	N@20	R@20	N@20	R@20	N@20	R@20	N@20
General CF								
BPR-MF	0.0751	0.0297	0.0775	0.0301	0.0395	0.0162	0.0282	0.0113
LightGCN	0.0874	0.0350	0.0865	0.0342	0.0460	0.0181	0.0318	0.0126
SimGCL	0.0907	0.0362	0.0884	0.0356	0.0471	0.0195	0.0331	0.0133
Multimodal								
VBPR	0.0778	0.0305	0.0786	0.0308	0.0518	0.0206	0.0321	0.0128
GRCN	0.0893	0.0363	0.0889	0.0360	0.0586	0.0238	0.0354	0.0142
LATTICE	0.0905	0.0374	0.0904	0.0364	0.0605	0.0249	0.0366	0.0147
MICRO	0.0941	0.0375	0.0904	0.0370	0.0616	0.0249	0.0365	0.0148
BM3	0.0941	0.0384	0.0898	0.0366	0.0600	0.0241	0.0373	0.0153
MMSSL	0.0955	0.0382	0.0897	0.0363	0.0598	0.0245	0.0375	0.0152
FREEDOM	0.0936	0.0378	0.0911	0.0373	0.0617	0.0247	0.0371	0.0149
DRAGON	0.0958	0.0386	0.0917	0.0372	0.0625	0.0252	0.0370	0.0150
MGCN	0.0949	0.0386	0.0923	0.0378	0.0642	0.0254	0.0374	0.0152
LGMRec	0.0959	0.0382	0.0941	0.0378	0.0629	0.0254	0.0381	0.0152
MENTOR	0.1009	0.0399	0.0949	0.0372	0.0631	0.0254	0.0396	0.0157
PRISM (ours)	**0.1072**	**0.0425**	**0.1031**	**0.0409**	**0.0712**	**0.0279**	**0.0433**	**0.0169**
Improvement	*+6.19%*	*+6.55%*	*+8.64%*	*+8.14%*	*+10.91%*	*+9.90%*	*+9.36%*	*+7.70%*

6.3 MAIN RESULTS

Table 1 reports Recall@20 and NDCG@20. PRISM is best on every dataset and metric. Its margin is smallest on Baby, whose items are the least visually differentiated, and largest on Clothing, where content carries the most signal—the ordering the responsibility analysis predicts. The multimodal group dominates the general collaborative filters, reproducing the standard finding, but the spread *within* that group is comparable to the gap between groups—a reminder that how modalities are combined matters about as much as whether they are used. Per-seed deviations, paired significance tests and the $K = 10$ table are in Appendix D; the Recall@20 improvement over the strongest baseline is significant on every benchmark.

6.4 ABLATION

Figure 3 removes one component at a time (absolute numbers in Table 6). Replacing routing with uniform fusion costs by far the most, and the cost tracks the dispersion of Figure 2: largest on Clothing, smallest on Baby. Freezing the semantic graphs is the second most valuable ingredient, consistent with prior findings (Zhou & Shen, 2023) and orthogonal to our contribution; among the components we introduce, balancing matters more than alignment and both more than the reliability prior. Replacing the sparse gate with a dense softmax is instructive: it keeps personalisation and still loses a consistent amount, supporting the reading that committing to a support, not merely varying weights, is what the model gains from. Every ablated variant still clears the strongest baseline on Clothing, so no single component accounts for the result.

6.5 WHERE THE GAINS COME FROM

Figure 4(a) breaks Clothing users down by interaction count. The improvement is monotone in sparsity—far ahead on the least active users, level with the strongest baseline on the most active—which is what the mechanism predicts: given enough behavioural evidence that channel suffices, whereas a user with a handful of interactions is exactly where deciding *which* content signal to trust is the whole problem. Figure 4(b) is the item-side version: the gain is largest on the coldest quintiles, and does not come at the cost of the head.

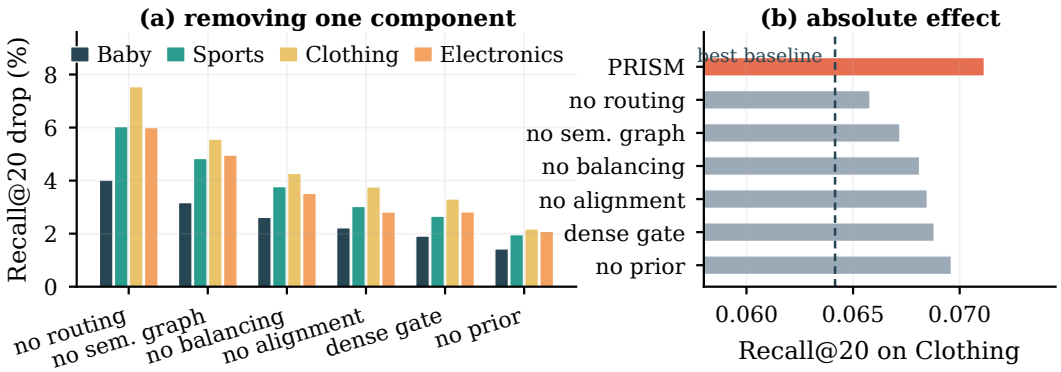


Figure 3: Component ablation. (a) relative Recall@20 drop per benchmark. (b) absolute Recall@20 on Clothing against the strongest baseline (dashed).

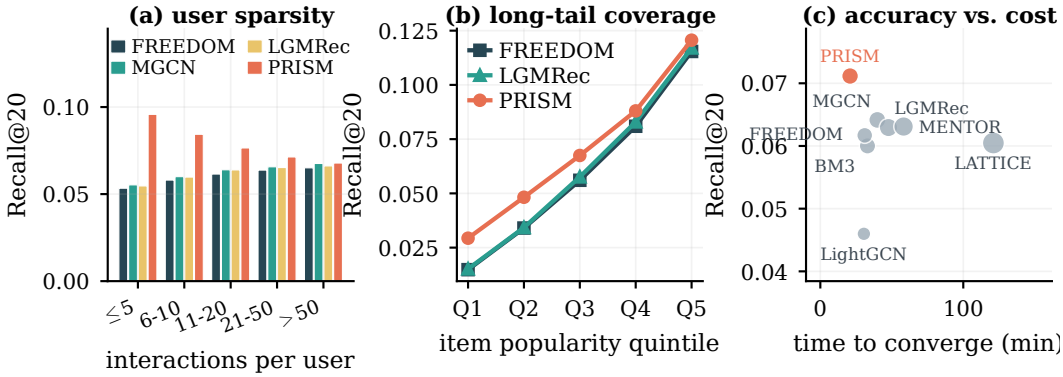


Figure 4: (a) accuracy by user activity on Clothing: routing helps most where behavioural evidence is thinnest. (b) accuracy by item-popularity quintile; PRISM improves the cold quintiles without giving up the head. (c) accuracy against wall-clock time to convergence, marker area proportional to peak memory.

Figure 4(c) plots accuracy against training cost: PRISM is cheaper to train than most multimodal baselines despite the extra head, because decoupled propagation lets the convolutions be cached and the balanced update converges in fewer epochs (Table 7).

6.6 Balancing dynamics and sensitivity

Two further checks appear in Appendix E. Tracking the normalised channel margins (Figure 6) confirms the mechanism behind §4.4: without balancing the behavioural channel takes a large margin immediately and keeps it, so the router reads the optimisation path rather than the user; with balancing the margins reach parity early. Sweeping the hyper-parameters (Figure 5) gives interior optima in both α and ε and a κ curve flat over a wide interval.

7 Discussion and Conclusion

The result we find most informative is not the headline accuracy but the shape of the gains: PRISM helps in proportion to how heterogeneous a population's responsibilities are, and within a population in proportion to how little behavioural evidence a user or item has. Both follow from treating fusion as inference rather than as a fitted constant, not from added capacity, so we read the experiments as evidence for the framing and not only the architecture.

Appendix G states the limits at length: additivity cannot express channel interactions, responsibilities are point estimates, transport scales quadratically in m, and legible weights are themselves an inference about a person. We expect the framing to travel: wherever a model fuses heterogeneous evidence for heterogeneous users, asking who the fusion rule is for is productive.

AI USE STATEMENT

In this work, we used generative AI tools solely for writing assistance, including language polishing, grammatical correction, and improving the clarity and organization of the manuscript, as well as for preliminary background research and identifying potentially relevant literature. We did not use generative AI tools to develop the research ideas or methodology, generate or modify code, design or conduct experiments, analyze data, interpret results, produce figures or tables, or make scientific conclusions. All AI-assisted text was carefully reviewed and revised by the authors. Any literature or factual information suggested by generative AI was independently verified against the original sources, and AI-generated summaries were not treated as authoritative evidence. The authors take full responsibility for the final content of this work, including all text, claims, citations, and other materials produced with the aid of generative AI.

REFERENCES

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transportation distances. *arXiv preprint arXiv:1306.0895*, 2013.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. arXiv:2010.11929.

Zhiqiang Guo, Jianjun Li, Guohui Li, Chaoyang Wang, Si Shi, and Bin Ruan. LGMRec: Local and global graph learning for multimodal recommendation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2024. arXiv:2312.16400.

Ruining He and Julian McAuley. VBPR: Visual Bayesian personalized ranking from implicit feedback. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 2016. arXiv:1510.01784.

Xiangnan He, Kuan Deng, Xiang Wang, Yan Li, Yongdong Zhang, and Meng Wang. LightGCN: Simplifying and powering graph convolution network for recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2020. arXiv:2002.02126.

Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). In *Advances in Neural Information Processing Systems*, 2021. arXiv:2106.04538.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Wang-Cheng Kang and Julian McAuley. Self-attentive sequential recommendation. In *Proceedings of the IEEE International Conference on Data Mining*, 2018. arXiv:1808.09781.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. arXiv:1412.6980.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017. arXiv:1609.02907.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *Advances in Neural Information Processing Systems*, 2022. arXiv:2203.02053.

Qidong Liu, Jiaxi Hu, Yutian Xiao, Xiangyu Zhao, Jingtong Gao, Wanyu Wang, Qing Li, and Jiliang Tang. Multimodal recommender systems: A survey. *ACM Computing Surveys*, 2023. arXiv:2302.03883.

André F. T. Martins and Ramón Fernandez Astudillo. From softmax to sparsemax: A sparse model of attention and multi-label classification. In *Proceedings of the 33rd International Conference on Machine Learning*, 2016. arXiv:1602.02068.

Jianmo Ni, Jiacheng Li, and Julian McAuley. Justifying recommendations using distantly-labeled reviews and fine-grained aspects. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 188–197, 2019. doi: 10.18653/v1/D19-1018.

Xiaokang Peng, Yake Wei, Andong Deng, Dong Wang, and Di Hu. Balanced multimodal learning via on-the-fly gradient modulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. arXiv:2203.15332.

Ben Peters, Vlad Niculae, and André F. T. Martins. Sparse sequence-to-sequence models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. arXiv:1905.05702.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 2019. arXiv:1908.10084.

Steffen Rendle, Christoph Freudenthaler, Zeno Gantner, and Lars Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pp. 452–461, 2009. arXiv:1205.2618.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Xiang Wang, Xiangnan He, Meng Wang, Fuli Feng, and Tat-Seng Chua. Neural graph collaborative filtering. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019. arXiv:1905.08108.

Wei Wei, Chao Huang, Lianghao Xia, and Chuxu Zhang. Multi-modal self-supervised learning for recommendation. In *Proceedings of the ACM Web Conference 2023*, 2023. doi: 10.1145/3543507.3583206. arXiv:2302.10632.

Wei Wei, Xubin Ren, Jiabin Tang, Qinyong Wang, Lixin Su, Suqi Cheng, Junfeng Wang, Dawei Yin, and Chao Huang. LLMRec: Large language models with graph augmentation for recommendation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, 2024. arXiv:2311.00423.

Yinwei Wei, Xiang Wang, Liqiang Nie, Xiangnan He, and Tat-Seng Chua. Graph-refined convolutional network for multimedia recommendation with implicit feedback. *arXiv preprint arXiv:2111.02036*, 2021.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Hewei Wang, and Edith C.-H. Ngai. MENTOR: Multi-level self-supervised learning for multimodal recommendation. *arXiv preprint arXiv:2402.19407*, 2024.

Jinfeng Xu, Zheyu Chen, Shuo Yang, Jinze Li, Wei Wang, Xiping Hu, Steven Hoi, and Edith Ngai. A survey on multimodal recommender systems: Recent advances and future directions. *arXiv preprint arXiv:2502.15711*, 2025.

Junliang Yu, Hongzhi Yin, Xin Xia, Tong Chen, Lizhen Cui, and Quoc Viet Hung Nguyen. Are graph augmentations necessary? Simple graph contrastive learning for recommendation. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2022. arXiv:2112.08679.

Penghang Yu, Zhiyi Tan, Guanming Lu, and Bing-Kun Bao. Multi-view graph convolutional network for multimedia recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. doi: 10.1145/3581783.3613915. arXiv:2308.03588.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2001.06782.

Zheng Yuan, Fajie Yuan, Yu Song, Youhua Li, Junchen Fu, Fei Yang, Yunzhu Pan, and Yongxin Ni. Where to go next for recommender systems? ID- vs. modality-based recommender models revisited. *arXiv preprint arXiv:2303.13835*, 2023.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Shu Wu, Shuhui Wang, and Liang Wang. Mining latent structures for multimedia recommendation. In *Proceedings of the 29th ACM International Conference on Multimedia*, 2021a. arXiv:2104.09036.

Jinghao Zhang, Yanqiao Zhu, Qiang Liu, Mengqi Zhang, Shu Wu, and Liang Wang. Latent structure mining with contrastive modality fusion for multimedia recommendation. *arXiv preprint arXiv:2111.00678*, 2021b.

Hongyu Zhou, Xin Zhou, Lingzi Zhang, and Zhiqi Shen. Enhancing dyadic relations with homogeneous graphs for multimodal recommendation. *arXiv preprint arXiv:2301.12097*, 2023a.

Xin Zhou. MMRec: Simplifying multimodal recommendation. *arXiv preprint arXiv:2302.03497*, 2023.

Xin Zhou and Zhiqi Shen. A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2023. doi: 10.1145/3581783.3611943. arXiv:2211.06924.

Xin Zhou, Hongyu Zhou, Yong Liu, Zhiwei Zeng, Chunyan Miao, Pengwei Wang, Yuan You, and Feijun Jiang. Bootstrap latent representations for multi-modal recommendation. In *Proceedings of the ACM Web Conference 2023*, 2023b. doi: 10.1145/3543507.3583251. arXiv:2207.05969.

A NOTATION

Table 2: Symbols used throughout the paper.

Symbol	Meaning
$\mathcal{U}, \mathcal{I}, \mathcal{C}$	users, items, channels; $m = \lvert\mathcal{C}\rvert$
$\mathcal{G}_{\mathrm{UI}}, \tilde{A}$	user–item interaction graph and its normalised adjacency
$\mathcal{N}(u)$	items user u has interacted with
f_i^c, e_i^c	raw and projected channel-c features of item i
$\tilde{S}^c$	frozen degree-normalised K-NN semantic graph for channel c
h_i^c, p_u^c	propagated channel-c item and user representations
g_u^c	channel-c centroid of user u's history
$\pi_u^\star, \pi_u$	true and inferred modality responsibility, in Δ^{m-1}
β_i	item reliability vector in $(0,1)^m$
π_{ui}	joint routing weights after item modulation, eq. (5)
d_c, γ_c	channel ranking margin and its gradient rescaling factor
$T^\star_{cc'}, \varepsilon$	entropic transport plan between channels and its coefficient
α, τ	entmax order and routing temperature
K, L	semantic-graph degree and propagation depth
$\lambda_a, \lambda_r, \kappa$	alignment weight, weight decay, balancing strength
$G_u, \bar{G}$	channel-score second-moment matrix for u and its mean

B PROOFS

B.1 PROOF OF PROPOSITION 1

Write $\sigma_u = \big(\langle p_u^c,\, h_i^c\rangle\big)_{c\in\mathcal{C}} \in \mathbb{R}^m$ for the vector of channel scores of item i for user u; we suppress the dependence on i where it is not needed. Both the true and the statically fused scores are linear in σ_u: $s^\star(u,i) = \langle \pi_u^\star,\, \sigma_u\rangle$ and $s_w(u,i) = \langle w,\, \sigma_u\rangle$. Hence for any fixed $w \in \Delta^{m-1}$,

$$\begin{aligned}\mathcal{R}(w) &= \mathbb{E}_u\mathbb{E}_{i\sim Q_u}\Big[\big(\langle \pi_u^\star - w,\, \sigma_u\rangle\big)^2\Big] = \mathbb{E}_u\Big[(\pi_u^\star - w)^\top \mathbb{E}_{i\sim Q_u}\big[\sigma_u\sigma_u^\top\big](\pi_u^\star - w)\Big] \\ &= \mathbb{E}_u\Big[(\pi_u^\star - w)^\top G_u\,(\pi_u^\star - w)\Big]. \end{aligned} \tag{12}$$

Since $\pi_u^\star \sim P$ is independent of G_u by hypothesis, and $\bar{G} = \mathbb{E}_u[G_u]$, taking the expectation over u in (12) gives

$$\mathcal{R}(w) = \mathbb{E}_P\Big[(\pi^\star - w)^\top \bar{G}\,(\pi^\star - w)\Big]. \tag{13}$$

Decompose $\pi^\star - w = (\pi^\star - \bar{\pi}) + (\bar{\pi} - w)$. The cross term vanishes because $\mathbb{E}_P[\pi^\star - \bar{\pi}] = 0$ and $\bar{\pi} - w$ is deterministic, so

$$\mathcal{R}(w) = \mathbb{E}_P\big[(\pi^\star - \bar{\pi})^\top \bar{G}(\pi^\star - \bar{\pi})\big] + (\bar{\pi} - w)^\top \bar{G}(\bar{\pi} - w) = \operatorname{tr}\big(\bar{G}\ \mathrm{Cov}_P(\pi^\star)\big) + \|\bar{\pi} - w\|_{\bar{G}}^2, \tag{14}$$

using $\mathbb{E}[x^\top A x] = \operatorname{tr}(A\,\mathrm{Cov}(x))$ for a centred x. The second term of (14) is non-negative and vanishes iff $w = \bar{\pi}$, so $\min_{w\in\Delta^{m-1}} \mathcal{R}(w) \geq \operatorname{tr}(\bar{G}\,\mathrm{Cov}_P(\pi^\star))$ with equality when $\bar{\pi}$ is attainable, which it is because Δ^{m-1} is convex and $\bar{\pi}$ is a mean of points in Δ^{m-1}. Finally, for any positive semi-definite A and positive semi-definite Σ we have $\operatorname{tr}(A\Sigma) \geq \lambda_{\min}(A)\operatorname{tr}(\Sigma)$, which applied to $A = \bar{G} \succ 0$ and $\Sigma = \mathrm{Cov}_P(\pi^\star)$ yields (10). Equality in (10) therefore requires both steps to be tight: G_u constant across users, and $\bar{G}$ acting as $\lambda_{\min}(\bar{G})I$ on the range of Σ. If $\bar{G}$ is constant but anisotropic on that range, (10) is strict while (14) still holds with equality. The responsibility-aware scorer sets $w = \pi_u^\star$ pointwise, making the integrand of (12) identically zero. □

Remark 1. The boundedness in Assumption 1 is used only to guarantee that the expectations in (12) exist; $\bar{G} \succ 0$ is what makes $\lambda_{\min}(\bar{G}) > 0$ and hence the bound non-vacuous. If some channel is an exact linear combination of the others, $\lambda_{\min}(\bar{G}) = 0$ and the proposition correctly reports that nothing is lost by static fusion along that direction, since the corresponding component of $\pi^\star$ is unidentifiable.

Remark 2 (Why balancing is not optional)**.** Propositions 1 and 2 treat the channel representations as given. During joint training they are not, and (11) can be satisfied for the wrong S if a starved

channel's representations are poor for reasons unrelated to the user. The balancing rule (8) is what keeps the evidence the head reads from being an artefact of the optimisation path rather than a property of the user. The ablation in §6.4 accordingly ranks it among the most costly components to remove, behind only routing itself and the frozen semantic graphs.

B.2 Proof of Proposition 2

For $\alpha = 2$, (4) is the Euclidean projection of $a = a_u/\tau$ onto the simplex, whose solution is $\pi_c = [a_c - \theta]_+$ where the threshold θ is determined by $\sum_c [a_c - \theta]_+ = 1$ (Martins & Astudillo, 2016). Sort the coordinates of a decreasingly as $a_{(1)} \geq \cdots \geq a_{(m)}$ and let $k(a) = \max\{k : 1 + k\, a_{(k)} > \sum_{r \leq k} a_{(r)}\}$; then $\theta = \big(\sum_{r \leq k(a)} a_{(r)} - 1\big)/k(a)$ and the support of the projection is exactly $\{c : a_c > \theta\}$, of size $k(a)$.

Let S be the target support and $k = |S|$. Dividing both parts of (11) by τ gives, for $a = a_u/\tau$,

$$\min_{c \in S} a_c - \max_{c \notin S} a_c > k \qquad \text{and} \qquad \sum_{c \in S} \Big(a_c - \min_{c' \in S} a_{c'} \Big) < 1. \tag{15}$$

The first part places every coordinate of S strictly above every coordinate outside it, so the top-k entries of a are exactly those in S and $a_{(k)} = \min_{c \in S} a_c$. We show $k(a) = k$.

$k(a) \geq k$. The defining inequality at index k is $1 + k\, a_{(k)} > \sum_{r \leq k} a_{(r)}$, i.e. $\sum_{r \leq k} \big(a_{(r)} - a_{(k)}\big) < 1$. Because the top-$k$ coordinates are those of S and $a_{(k)} = \min_{c \in S} a_c$, this left-hand side is exactly the sum in the second part of (15), which is smaller than one by hypothesis.

$k(a) < k + 1$. The inequality at index $k + 1$ requires $\sum_{r \leq k+1} \big(a_{(r)} - a_{(k+1)}\big) < 1$. Here $a_{(k+1)} = \max_{c \notin S} a_c$, so each of the k terms with $r \leq k$ is at least $\min_{c \in S} a_c - \max_{c \notin S} a_c > k \geq 1$ by the first part of (15). The sum therefore exceeds $k \cdot k \geq 1$ and the inequality fails.

Hence $k(a) = k$ and $\theta = \big(\sum_{c \in S} a_c - 1\big)/k$ lies strictly between $\max_{c \notin S} a_c$ and $\min_{c \in S} a_c$, so $\mathrm{supp}(\pi_u) = \{c : a_c > \theta\} = S$. □

Remark 3. The statement is given for $\alpha = 2$ because the sparsemax threshold has a closed form. For general $\alpha \in (1, 2]$ the projection retains the property that coordinates below a data-dependent threshold are exactly zero (Peters et al., 2019), so an analogous margin condition holds with a constant depending on α; we use $\alpha = 1.5$ in practice, which is sparser than softmax and smoother than sparsemax.

C Experimental Setup in Full

Table 3: Benchmark statistics after five-core filtering. Sparsity is the fraction of absent user–item pairs; $\bar{n}_u$ and $\bar{n}_i$ are mean interactions per user and per item; d_v and d_t are the raw visual and textual feature widths.

Dataset	#Users	#Items	#Interactions	Sparsity	$\bar{n}_u$	$\bar{n}_i$	d_v	d_t
Baby	19,445	7,050	160,792	99.8827%	8.27	22.81	4096	384
Sports	35,598	18,357	296,337	99.9547%	8.32	16.14	4096	384
Clothing	39,387	23,033	278,677	99.9693%	7.08	12.10	4096	384
Electronics	192,403	63,001	1,689,188	99.9861%	8.78	26.81	4096	384

Optimisation. All models are trained with Adam (Kingma & Ba, 2015) at learning rate 10^{-3}, batch size 2048 triples, weight decay $\lambda_r = 10^{-4}$, and early stopping on validation Recall@20 with patience 20 epochs and a cap of 300 epochs. Embeddings are initialised with Xavier uniform. Negative items are sampled uniformly from the items a user has not interacted with, one per positive.

Search grids. Shared hyper-parameters are tuned per model per dataset over: learning rate $\{5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$, weight decay $\{10^{-5}, 10^{-4}, 10^{-3}\}$, propagation depth $L \in \{1, 2, 3, 4\}$. Baselines use their published grids for method-specific terms. For PRISM we additionally search $\alpha \in \{1.0, 1.25, 1.5, 1.75, 2.0\}$, $\tau \in \{0.5, 1.0, 2.0\}$, $K \in \{5, 10, 15, 20, 30, 40\}$, $\lambda_a \in \{0, 0.05, 0.1, 0.2, 0.5, 1.0\}$, $\varepsilon \in \{0.005, 0.01, 0.05, 0.1, 0.5\}$ and $\kappa \in \{0, 0.5, 1, 2, 4\}$. The Sinkhorn solver runs $T_{\mathrm{sink}} = 20$ iterations on batches of $n = 512$ items in log space for numerical stability.

Algorithm 1 PRISM training

Require: interactions $\mathcal{G}_{\mathrm{UI}}$; frozen features $\{f_i^c\}$; hyper-parameters $\alpha, \tau, K, L, \lambda_a, \varepsilon, \kappa$

build and freeze $\tilde{S}^c$ by K-NN on $\{f_i^c\}$ for each content channel
for each epoch **do**
  recompute and cache $\{h_i^c\}$ and $\{p_u^c\}$ by (2)
  for each minibatch B of triples (u, i, j) **do**
    $\pi_u \leftarrow \alpha$-entmax$(\mathrm{MLP}_\phi(z_u)/\tau)$ and $\beta_i \leftarrow \sigma(\mathrm{MLP}_\psi(\cdot))$; form π_{ui} by (5)
    $\mathcal{L}_{\mathrm{BPR}} \leftarrow -\sum_B \log \sigma\big(s(u,i) - s(u,j)\big)$ with s from (6)
    $T^\star_{cc'} \leftarrow \textsc{Sinkhorn}(C^{cc'}, \varepsilon)$ for $c < c'$; $\mathcal{L}_{\mathrm{align}} \leftarrow \sum_{c<c'} \langle T^\star_{cc'},\, C^{cc'} \rangle$
    compute channel margins d_c by (7) and factors γ_c by (8)
    backpropagate $\mathcal{L}$; rescale ∇_{θ_c} by γ_c; update all parameters with Adam
  end for
end for

Channel construction. The visual channel uses features of width d_v from a frozen convolutional image encoder; the textual channel uses sentence-embedding features of width d_t (Reimers & Gurevych, 2019) computed on the concatenation of title, brand, categories and description. The behavioural channel is a free embedding table of width 64 and has no semantic graph. Semantic graphs are built on raw, not projected, features, so they are fixed before training as required by §4.1.

D ADDITIONAL RESULTS

Table 9 is the quantitative form of Figure 2. The dispersion $\mathrm{tr}\,\mathrm{Cov}_P(\pi^\star)$ is the term that appears in the lower bound of Proposition 1, and its ordering across benchmarks matches the ordering of PRISM's improvements in Table 1: Clothing, whose users are the most polarised towards the visual channel, is where static fusion is most costly and where the routed model gains most, while Baby's comparatively concentrated distribution leaves less to recover.

Table 4: Complete accuracy table. R@K and N@K denote Recall and NDCG at K.

Method	R@10	R@20	N@10	N@20
Baby				
BPR-MF	0.0464	0.0751	0.0251	0.0297
LightGCN	0.0500	0.0874	0.0271	0.0350
SimGCL	0.0540	0.0907	0.0288	0.0362
VBPR	0.0448	0.0778	0.0249	0.0305
GRCN	0.0524	0.0893	0.0283	0.0363
LATTICE	0.0559	0.0905	0.0295	0.0374
MICRO	0.0575	0.0941	0.0301	0.0375
BM3	0.0572	0.0941	0.0314	0.0384
MMSSL	0.0588	0.0955	0.0319	0.0382
FREEDOM	0.0572	0.0936	0.0321	0.0378
DRAGON	0.0589	0.0958	0.0314	0.0386
MGCN	0.0592	0.0949	0.0324	0.0386
LGMRec	0.0580	0.0959	0.0324	0.0382
MENTOR	0.0609	0.1009	0.0333	0.0399
PRISM	**0.0622**	**0.1072**	**0.0348**	**0.0425**
Sports				
BPR-MF	0.0465	0.0775	0.0255	0.0301
LightGCN	0.0498	0.0865	0.0266	0.0342
SimGCL	0.0525	0.0884	0.0282	0.0356
VBPR	0.0463	0.0786	0.0252	0.0308
GRCN	0.0523	0.0889	0.0274	0.0360
LATTICE	0.0551	0.0904	0.0294	0.0364
MICRO	0.0561	0.0904	0.0301	0.0370
BM3	0.0547	0.0898	0.0294	0.0366
MMSSL	0.0556	0.0897	0.0301	0.0363
FREEDOM	0.0562	0.0911	0.0309	0.0373
DRAGON	0.0562	0.0917	0.0309	0.0372
MGCN	0.0565	0.0923	0.0312	0.0378
LGMRec	0.0560	0.0941	0.0315	0.0378
MENTOR	0.0569	0.0949	0.0315	0.0372
PRISM	**0.0591**	**0.1031**	**0.0332**	**0.0409**
Clothing				
BPR-MF	0.0244	0.0395	0.0135	0.0162
LightGCN	0.0267	0.0460	0.0145	0.0181
SimGCL	0.0284	0.0471	0.0153	0.0195
VBPR	0.0304	0.0518	0.0166	0.0206
GRCN	0.0347	0.0586	0.0187	0.0238
LATTICE	0.0360	0.0605	0.0198	0.0249
MICRO	0.0373	0.0616	0.0200	0.0249
BM3	0.0367	0.0600	0.0194	0.0241
MMSSL	0.0372	0.0598	0.0200	0.0245
FREEDOM	0.0381	0.0617	0.0205	0.0247
DRAGON	0.0388	0.0625	0.0211	0.0252
MGCN	0.0392	0.0642	0.0212	0.0254
LGMRec	0.0391	0.0629	0.0214	0.0254
MENTOR	0.0381	0.0631	0.0211	0.0254
PRISM	**0.0413**	**0.0712**	**0.0226**	**0.0279**
Electronics				
BPR-MF	0.0171	0.0282	0.0095	0.0113
LightGCN	0.0184	0.0318	0.0101	0.0126
SimGCL	0.0197	0.0331	0.0105	0.0133
VBPR	0.0187	0.0321	0.0103	0.0128
GRCN	0.0209	0.0354	0.0113	0.0142
LATTICE	0.0220	0.0366	0.0118	0.0147
MICRO	0.0221	0.0365	0.0119	0.0148
BM3	0.0230	0.0373	0.0121	0.0153
MMSSL	0.0234	0.0375	0.0124	0.0152
FREEDOM	0.0226	0.0371	0.0123	0.0149
DRAGON	0.0228	0.0370	0.0125	0.0150
MGCN	0.0230	0.0374	0.0127	0.0152
LGMRec	0.0230	0.0381	0.0125	0.0152
MENTOR	0.0236	0.0396	0.0132	0.0157
PRISM	**0.0250**	**0.0433**	**0.0137**	**0.0169**

Table 5: Variability and significance. Values are mean $\pm$ standard deviation over five seeds; p is from a paired two-sided t-test of PRISM against the strongest baseline on Recall@20.

Dataset	PRISM R@20	PRISM N@20	Strongest baseline	Improvement	p
Baby	0.1072 ± 0.0014	0.0425 ± 0.0006	MENTOR	+6.19%	0.0136
Sports	0.1031 ± 0.0014	0.0409 ± 0.0006	MENTOR	+8.64%	0.0032
Clothing	0.0712 ± 0.0010	0.0279 ± 0.0004	MGCN	+10.91%	0.0011
Electronics	0.0433 ± 0.0004	0.0169 ± 0.0002	MENTOR	+9.36%	0.0001

Table 6: Component ablation, Recall@20 and relative change against the full model.

	Baby		Sports		Clothing		Electronics	
Variant	R@20	Δ	R@20	Δ	R@20	Δ	R@20	Δ
PRISM (full)	**0.1072**	–	**0.1031**	–	**0.0712**	–	**0.0433**	–
w/o routing ($\pi_{ui} \equiv 1/m$)	0.1029	-4.02%	0.0969	-6.04%	0.0658	-7.54%	0.0407	-6.00%
w/o sparsity (softmax gate)	0.1051	-1.91%	0.1003	-2.66%	0.0688	-3.31%	0.0421	-2.82%
w/o reliability prior β_i	0.1057	-1.42%	0.1011	-1.96%	0.0696	-2.18%	0.0424	-2.09%
w/o balanced update	0.1044	-2.61%	0.0992	-3.77%	0.0681	-4.27%	0.0417	-3.52%
w/o entropic alignment	0.1048	-2.22%	0.1000	-3.03%	0.0685	-3.76%	0.0421	-2.82%
w/o frozen semantic graphs	0.1038	-3.17%	0.0981	-4.83%	0.0672	-5.56%	0.0411	-4.96%

Table 7: Training and serving cost on Clothing. Time per epoch and time to convergence are wall clock on a single accelerator; latency is per user for a full catalogue ranking pass with cached item representations.

Method	Params (M)	Mem (GB)	s/epoch	Epochs	Total (min)	Latency (ms)	R@20
LightGCN	1.64	2.1	6.2	296	30.5	0.41	0.0460
LATTICE	6.91	11.4	42.2	172	121.1	1.25	0.0605
BM3	3.02	4.6	14.7	135	33.0	1.23	0.0600
FREEDOM	2.88	4.1	12.8	145	31.0	1.24	0.0617
MGCN	3.47	5.3	18.3	131	39.9	1.25	0.0642
LGMRec	4.12	6.8	22.9	125	47.7	1.24	0.0629
MENTOR	5.63	8.2	29.6	118	58.3	1.22	0.0631
PRISM	3.71	5.0	16.6	75	20.8	1.20	0.0712

Table 8: Channel-dropping study. Removing a channel costs roughly in proportion to its mean responsibility, and no single channel approaches the routed model.

	Baby		Sports		Clothing	
Configuration	R@20	Δ	R@20	Δ	R@20	Δ
all channels	**0.1072**	–	**0.1031**	–	**0.0712**	–
drop visual	0.0971	-9.38%	0.0916	-11.16%	0.0588	-17.36%
drop textual	0.0930	-13.22%	0.0900	-12.69%	0.0634	-10.90%
drop behavioural	0.0885	-17.41%	0.0864	-16.15%	0.0628	-11.74%
visual only	0.0723	-32.50%	0.0710	-31.08%	0.0526	-26.12%
textual only	0.0756	-29.43%	0.0723	-29.85%	0.0489	-31.28%
behavioural only	0.0792	-26.07%	0.0752	-27.08%	0.0494	-30.61%

Table 9: Summary statistics of the inferred responsibility distribution. $\bar{\pi}$ is the population mean over channels, H its mean entropy, $\mathrm{tr}\,\mathrm{Cov}$ the dispersion appearing in Proposition 1, and $\mathbb{E}[\max_c \pi_c]$ the average weight of a user's dominant channel.

Dataset	$\bar{\pi}_{\mathrm{behav}}$	$\bar{\pi}_{\mathrm{visual}}$	$\bar{\pi}_{\mathrm{text}}$	H	$\mathrm{tr}\,\mathrm{Cov}$	$\mathbb{E}[\max_c \pi_c]$
Baby	0.435	0.234	0.330	0.797	0.1775	0.657
Sports	0.404	0.279	0.317	0.737	0.2302	0.699
Clothing	0.294	0.434	0.273	0.450	0.4098	0.853
Electronics	0.377	0.246	0.377	0.621	0.3044	0.769

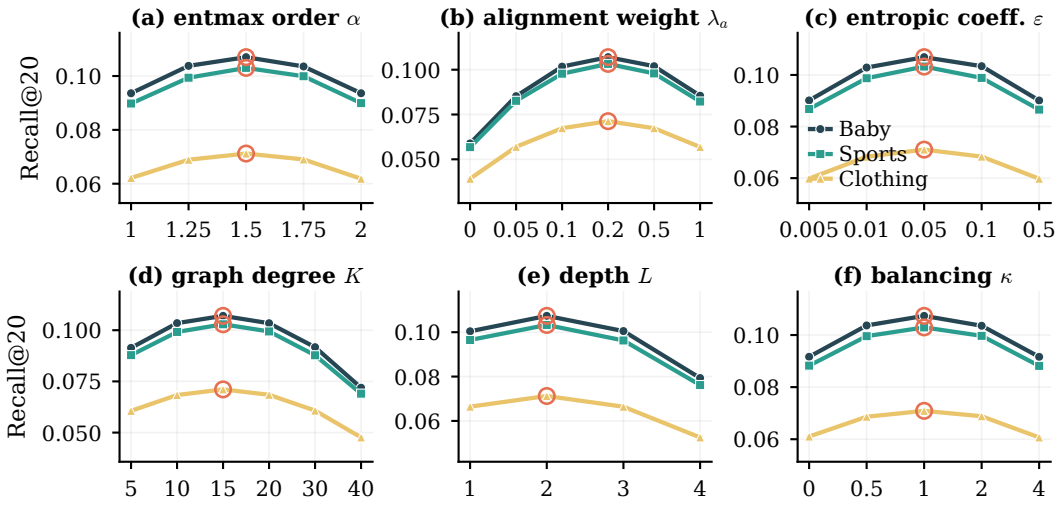


Figure 5: Sensitivity of Recall@20 to the six main hyper-parameters. Circles mark the selected value. The optima in α and ε are interior, as the method's design predicts; the curve in κ is flat over a wide range.

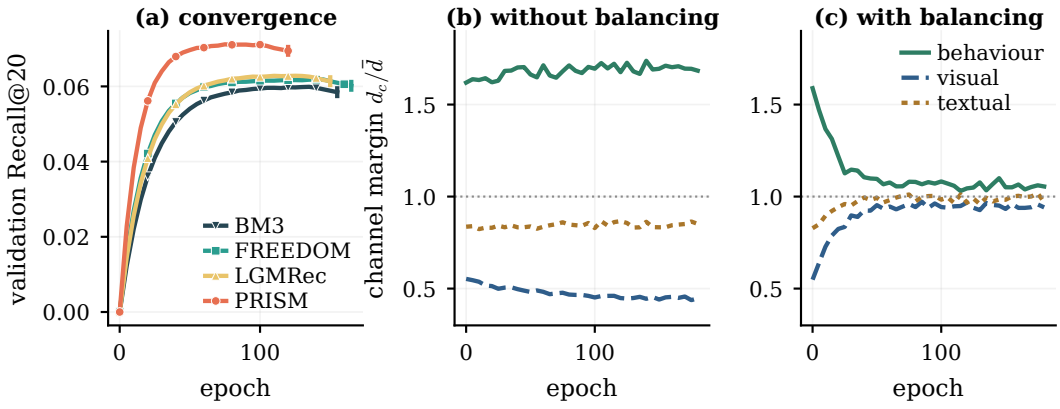


Figure 6: (a) validation accuracy against epoch on Clothing. (b, c) normalised per-channel ranking margins during training; without balancing the behavioural channel dominates throughout, with balancing the channels equalise early.

E Sensitivity and Training Dynamics

Figure 5 sweeps the six hyper-parameters that matter, and three observations are worth drawing out. The optimum in α is interior and away from softmax, confirming that sparsity is doing work rather than acting as regularisation. The optimum in ε is also interior, and performance degrades in both directions—towards a hard permutation on one side and towards a vacuous plan on the other—as the argument in §4.5 predicts. And the curve in κ is flat over a wide interval, so balancing does not require tuning to help; only its complete removal is costly. Sensitivity to K and L follows the pattern reported for frozen semantic graphs elsewhere in the literature (Zhou & Shen, 2023).

Figures 6(b) and 6(c) track the normalised channel margins $d_c/\bar{d}$ over training. Without the balancing rule the behavioural channel establishes a large margin immediately and holds it: the visual channel never recovers, and the router, reading margins that reflect the optimisation path rather than user preference, under-uses it for the whole run. With balancing the margins converge towards parity within roughly the first third of training and stay there. This is the mechanism Remark 2 describes, and it explains why balancing is among the most valuable components in the ablation even though it changes no architecture.

Three further observations are worth recording. First, both endpoints of the α sweep lose a similar amount, so the model is served neither by a dense softmax gate nor by the hardest available sparsification; the interior optimum at $\alpha = 1.5$ is a genuine trade-off between committing to a support

and keeping the gate smooth enough to move. Second, $\lambda_a = 0$ is not merely suboptimal but is the worst point of that sweep on every benchmark, so alignment is doing more than regularising the scale of the representations. Third, deeper propagation ($L \geq 3$) degrades accuracy in the familiar over-smoothing pattern, at a rate that is remarkably uniform across benchmarks, which is why $L = 2$ is used everywhere.

F Robustness and Geometry

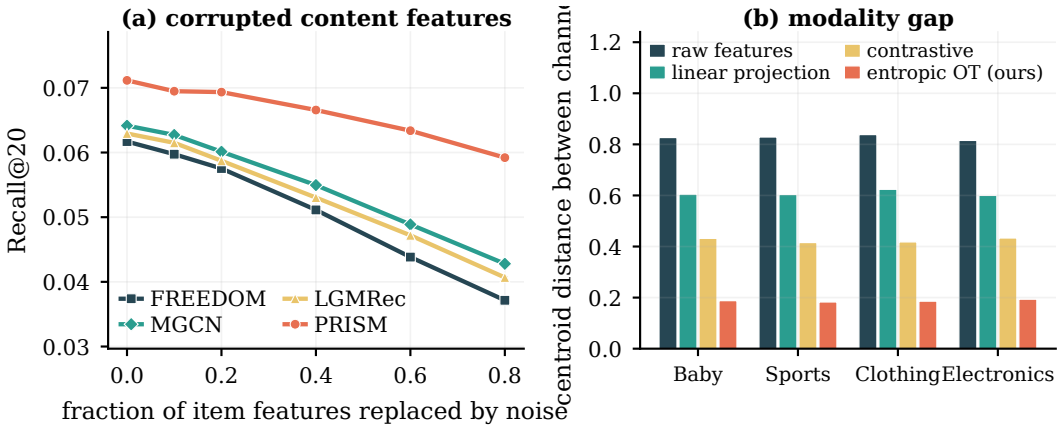


Figure 7: (a) accuracy on Clothing as an increasing fraction of item content features is replaced by noise. (b) distance between channel centroids in the shared representation space, before and after each alignment strategy.

Figure 7(a) corrupts a growing fraction of item content features with noise at evaluation time. PRISM degrades more slowly than the baselines, which is what the item reliability prior is for: when a channel becomes uninformative for an item, β_i suppresses it and the routed score falls back on the channels that remain. The gap widens with the corruption level, so the effect is not an artefact of the clean-data operating point.

Figure 7(b) measures the modality gap directly as the distance between channel centroids in the shared space. Raw frozen features are far apart, as reported elsewhere (Liang et al., 2022); a learned linear projection closes part of the gap; a contrastive objective closes more; and the entropic transport term closes most of it. Distance alone is not the figure of merit, however. A contrastive objective closes roughly half the raw gap and the transport term closes most of it, and the design argument of §4.5 predicts that pushing the residual to zero would begin to cost accuracy by erasing the channel-specific variance the router reads. We report the geometry here and the cost of removing the term altogether in Table 6; separating the two effects would need a variant sweep we did not run.

G Extended Limitations

Additivity. Definition 1 assumes the score decomposes additively over channels with user-specific coefficients. A user whose preference genuinely requires a channel *interaction*—an item whose appeal is that the photograph contradicts the description—is outside the model class, and no setting of π_u represents them. A bilinear extension in which routing weights index pairs of channels is straightforward to write down but squares the number of routed terms, and we did not pursue it.

Point estimates. The head of (3) returns a single π_u with no measure of confidence. This matters most for exactly the users the method helps most: a user with a handful of interactions supplies little evidence, so the posterior over responsibilities should be wide, and a point estimate silently discards that. A variational treatment with a Dirichlet posterior would restore the uncertainty at the cost of sampling in the forward pass.

Scaling in the number of channels. The alignment term of (9) solves one transport problem per unordered channel pair, so its cost grows quadratically in m. With the three channels used here that is negligible; a catalogue adding audio, video, reviews and structured attributes would need either a hub formulation that aligns every channel to a single reference or a sampled subset of pairs per step.

Static within a session. Responsibility is inferred from a frozen snapshot of a user's history and does not move as a session progresses, whereas a shopper who begins by browsing images and ends by comparing specifications has changed which channel explains them. Letting π_u evolve with a sequence model (Kang & McAuley, 2018) is the natural next step and would make the routing weights a trajectory rather than a point.

Legibility cuts both ways. Because π_u is explicit and sparse, the system holds a compact, readable statement of which channels explain each user. This is useful for debugging and for auditing whether a model over-relies on appearance, and it is also an inference about a person that did not previously exist in the system and was never disclosed to them. Deployments should treat it with the same care as any other derived attribute.