Engineering and Technology Journal e-ISSN: 2456-3358

Volume 11 Issue 08 August-2026, Page No.- 11339- 11350

DOI: 10.47191/etj/v11i08.22, I.F. – 8.595

How Much Does Each Modality Matter? A Cross-Study Empirical Comparison of Visual and Textual Signal Contributions in Multimodal Recommendation on E-Commerce Data

Redwood Meridian College

School of Computing and Information Sciences | Undergraduate Program in Computer Science | Portland, Oregon, USA

ABSTRACT: Multimodal recommenders routinely report accuracy gains over interaction-only models, yet the field rarely quantifies how much of that gain is attributable to the visual channel and how much to the textual channel. This paper reports a cross-study empirical comparison built on 54 published result cells drawn from three peer-reviewed studies that share an identical evaluation protocol on three Amazon categories: the same 5-core splits, the same released 4096-dimensional visual and 384-dimensional textual features, the same 8:1:1 per-user partition, and the same all-ranking evaluation. No new architecture is proposed. Three findings emerge. The apparent value of multimodal content depends heavily on the interaction-only reference point: 40 of 42 multimodal cells improve on a matrix-factorisation baseline by 65.91% on average, while only 30 of 42 improve on a tuned LightGCN baseline, with a mean gain of 10.29%. A published per-modality ablation shows that one content channel captures most of the available benefit, the second channel adding 3.44% on average. Published conclusions about which channel dominates conflict on identical data. Modality contribution is best read as backbone-conditional rather than as a property of the data.

KEYWORDS: multimodal recommendation; modality contribution; reproducibility; empirical comparison

1. INTRODUCTION

1.1. Content Signals in Recommendation

Recommender systems for e-commerce platforms have access to far more than click logs. Product photographs, titles, category paths, and free-text descriptions accompany almost every catalogue entry, and a decade of work has sought to convert this material into better rankings. The visual Bayesian personalised ranking method established the template that most later work follows, appending features from a pre-trained convolutional network to item identifier embeddings inside a matrix-factorisation objective [1]. Graph-based successors replaced the flat fusion step with modality-specific message passing over user-item bipartite graphs, allowing preferences to be modelled separately for each channel before aggregation [2]. Surveys of the resulting literature now catalogue dozens of published architectures organised by how they encode, interact with, enhance, and optimise content features [3]. Deployment has followed the research: product imagery and descriptive text are routinely pushed through frozen pre-trained encoders and cached as fixed vectors, an arrangement that makes the choice of which channel to maintain an operational question with a recurring extraction and storage cost attached to it.

Growth in the number of methods has not been matched by growth in the understanding of what the content actually contributes. Published papers almost always report a single aggregate comparison against interaction-only baselines, treating the multimodal signal as one undifferentiated block. Recent critical work questions whether the reported advantages survive careful re-evaluation, arguing that content helps mainly under sparse interaction regimes and that the relative usefulness of each channel is task-specific rather than universal [4]. A parallel line of enquiry contrasts identifier-based against modality-based representations and finds the ordering between them sensitive to the experimental regime [5].

1.2. Quantifying Per-Channel Contribution

The gap this paper addresses is narrow and measurable. Existing evaluations answer whether content helps in aggregate; they seldom answer how much of the benefit each channel supplies, whether the two channels are complementary or largely redundant, and whether any of the answers hold across encoders trained on the same data.

A. The Unresolved Attribution Question

Three attribution questions remain open. The magnitude question asks how large the content contribution is once it is measured against a competently tuned interaction-only reference rather than a weak one. The decomposition question asks how the contribution splits between the visual and textual channels, and how much the second channel adds beyond the first. The stability question asks whether the answers to the first two travel across recommendation backbones, across catalogue categories, and across independent research groups analysing identical inputs. Published studies address these questions only incidentally,

inside ablations designed to justify a proposed architecture rather than to characterise the data.

B. Research Questions and Contributions

This work poses three research questions matching those gaps and answers them by re-analysing published evidence rather than by proposing a new method. The first asks how the measured contribution of content changes with the choice of interaction-only reference. The second asks how the contribution decomposes into visual and textual components. The third asks how stable the decomposition is across backbones, categories, and source studies. The contribution is threefold: a traceable evidence base of published result cells collected under one shared protocol, a set of contribution indices computed uniformly over that base, and a documented conflict between two independent studies that reach opposite conclusions about channel dominance from identical features on identical splits.

2. RELATED WORK

2.1. Fusion Strategies for Content-Aware Recommendation

A. Graph-Based Structure Exploitation

After modality-specific bipartite propagation was introduced, refinement of the interaction graph itself became the dominant design. One influential method scores user-item edges by the affinity between user preference and item content, pruning likely false-positive feedback before propagation [6]. A second family shifted attention from the user-item graph to item-item relations, learning a modality-aware latent structure among catalogue entries and injecting the resulting affinities into identifier embeddings rather than into the features themselves [7]. An extension added contrastive fusion across the modality-specific graphs, arguing that linear combination discards fine-grained cross-channel agreement [8]. A later analysis showed the learned structure to be largely dispensable: freezing an item-item graph built directly from raw features before training, while denoising the interaction graph by degree-sensitive edge pruning, improves accuracy by 19.07% on average over the learned-structure predecessor and cuts memory cost substantially [9]. A further method separates content and behaviour into distinct views, purifying content features under behavioural guidance and fusing them with adaptive per-user channel weights [10].

B. Self-Supervised and Lightweight Designs

A second design current applies self-supervision to the content channels. One approach constructs feature-dropout, masking, and coarse-to-fine augmentations to build a node self-discrimination task over multimodal patterns, treating each channel as a view to be contrasted against the others [11]. A lighter variant removes negative sampling and target networks entirely, perturbing representations through dropout and aligning channels with a reconstruction objective, which cuts training time by a large factor while remaining competitive on accuracy [12]. These designs matter for the present analysis because they change where content enters the model. In one family the features reach the preference predictor directly; in another they only shape a graph over which identifier embeddings are propagated. Any per-channel attribution has to be read relative to whichever mechanism produced it.

2.2. How Per-Channel Evidence Is Currently Reported

Per-channel evidence appears in the literature almost exclusively as a by-product. Ablation tables are built to defend a proposed component, so they typically drop one channel at a time from the authors' own architecture and report the resulting loss, without a matched comparison against a tuned interaction-only reference and without repeating the exercise on a competing architecture. Some papers report a tuned scalar weighting between the two channels and read channel importance off that scalar; others report a single-channel variant of the full model. The two readings are not commensurable, since one measures the contribution of a channel to a similarity structure and the other its contribution to a ranking score. Neither convention is wrong, and the resulting numbers are rarely compared across papers.

2.3. Positioning of This Analysis

The present work occupies the space between the architecture papers, which supply per-channel numbers without cross-study comparison, and the reproducibility literature, which supplies cross-study comparison at the level of whole methods rather than channels. It proposes nothing new and retrains nothing. It collects the per-channel evidence that already exists within one shared evaluation protocol, converts it to a common set of indices, and asks whether the resulting attributions agree.

2.4. Broader AI Systems and Applied Contexts

Beyond recommendation-specific architectures, recent work has built realistic web, desktop, embodied, and tool-use environments for evaluating agents, alongside prompting, reflection, planning, and search strategies for improving deliberate action. [21–23, 36–39, 41–43, 46–47, 52–54, 61–62, 69–70, 72]

A parallel applied literature uses machine learning, graph methods, anomaly detection, and explainability across conflict warning, cyber-threat analysis, industrial control, medical imaging, data systems, payroll migration, infrastructure monitoring, and advertising-fraud detection. [24–27, 33–35, 40, 49–51, 55–60]

Related studies extend these techniques to community banking, credit-risk monitoring, financial contagion, capital planning, renewable-energy decisions, carbon-aware scheduling, hospital operations, and AI-assisted drug discovery. [28–32, 44–45, 48, 63–68, 71]

Within Chuan Wu's publication record, the broader methodological context also includes road-scene captioning, traffic reconstruction, vehicle retrieval, GAN-based image translation, hospital resource forecasting, supply-chain credit

risk, and graph/reinforcement-learning methods for drug discovery. [66, 68, 71, 73–78]

Additional work in Haojun Weng's publication record covers healthcare billing anomalies, repository-context retrieval for code completion, banking customer segmentation, deepfake detection, renewable-aware distributed training, industrial IoT anomaly detection, hospital resource forecasting, programmatic advertising, and consumer-credit feature selection. [66, 79–87]

3. DATA AND ANALYSIS DESIGN

3.1. Evaluation Substrate and Methodological Stance

The analysis is a secondary study. No models were retrained; every input value is a number printed in a published table, transcribed with its source recorded at cell level. The design follows from the target quantity: the object of interest is a contrast between published attributions, and re-running ten architectures under a fresh tuning budget would add a new source of variance rather than remove the one under examination. The methodological stance is inherited from the reproducibility literature in recommendation, which established that reported neural gains often shrink or vanish once baselines are tuned with comparable effort [13], formalised those criteria over a larger sample of published work [14], and produced evaluation toolkits that fix splitting, search, and significance testing so that comparisons rest on shared machinery [15]. That literature narrowed its object of study to whole methods; the present analysis narrows it further, to individual content channels.

The substrate is the Amazon review collection in the three categories on which the multimodal recommendation literature has converged. All source studies apply 5-core filtering to both users and items, use the released 4096-dimensional visual features extracted by a pre-trained convolutional network, use 384-dimensional sentence-transformer textual embeddings built from the concatenation of title, description, category, and brand, split each user's history 80/10/10 into training, validation, and test partitions, fix the embedding width at 64, optimise a pairwise ranking objective, and evaluate against the full catalogue rather than a sampled candidate set. Table 1 records the specifications, with densities recomputed from the published user, item, and interaction counts rather than copied across.

Table 1. Specifications of the three evaluation datasets.

Counts are as printed in FREEDOM Table 3 and MGCN Table 1, which agree exactly. Density is recomputed as interactions divided by the user-item product.

Dataset	Users	Items	Interactions	Density (%)	Visual dim.	Text dim.
Baby	19,445	7,050	160,792	0.1173	4,096	384
Sports and Outdoors	35,598	18,357	296,337	0.0453	4,096	384
Clothing, Shoes and Jewelry	39,387	23,033	278,677	0.0307	4,096	384

One discrepancy is recorded rather than reconciled. The earliest of the three source studies prints different interaction counts for the same three categories, namely 139,110, 256,308, and 237,488, and uses a 1024-dimensional textual encoder in place of the 384-dimensional one. That study is kept in a separate protocol group, and its values are never pooled with the other two inside any statistic reported below.

3.2. Evidence Base Construction

A. Source Selection and Transcription

Three studies qualify on three criteria: they evaluate on the categories above, they publish complete result tables covering both interaction-only and multimodal methods, and two of the three run every reported model inside a single open-source multimodal recommendation toolbox, which removes independent re-implementation as a source of divergence [16]. Transcription produced 54 main-result cells covering ten distinct recommendation methods, of which 42 are multimodal and 12 are interaction-only, plus a nine-cell per-channel ablation and a nine-cell backbone sweep. Every cell carries the source paper, the source table number, the dataset, the method, and both Recall@20 and NDCG@20, so any figure quoted later can be traced to a printed table. Values were transcribed twice and compared, and no value was estimated, interpolated, or read off a plot.

Table 2. Composition of the transcribed evidence base.

Source study	Venue	Table used	Cells	Role in the analysis
LATTICE	ACM MM 2021	Tables 2, 3	9	Backbone sweep: three collaborative filtering models with and without appended features

Source study	Venue	Table used	Cells	Role in the analysis
FREEDOM	ACM MM 2023	Tables 3, 4	27	Main comparison under the shared toolbox; tuned visual weighting
MGCN	ACM MM 2023	Tables 1, 2, 3	27 + 9	Main comparison; per-channel ablation isolating each modality

B. Contribution Indices

Four indices are computed uniformly over the base. The weak-reference gain expresses each multimodal result as a relative change over the matrix-factorisation result on the same dataset in the same table. The strong-reference gain does the same against the tuned LightGCN result, which is the most competitive interaction-only model any of the source studies report. The single-channel lift expresses a visual-only or text-only variant as a relative change over the behaviour-only variant of the identical architecture. The second-channel gain expresses the both-channel variant as a relative change over the stronger of the two single-channel variants, isolating what an additional modality contributes once one is already present. All four are ratios of published point estimates, and each is computed within a single source table so that no index mixes tuning regimes.

3.3. Consistency and Statistical Treatment

A. Cross-Study and Cross-Dataset Agreement

Two of the source studies report five methods in common on all three datasets, which permits a direct reproducibility check: the same method, the same dataset, the same toolbox, two independent research groups. Agreement is summarised by the mean and the maximum absolute relative difference in Recall@20 across those shared cells. Rank stability across catalogue categories is measured with Kendall's rank correlation computed over the multimodal methods shared between each pair of datasets, which separates the question of whether method ordering travels from the question of whether channel attribution travels.

B. Limits of the Statistical Claims

The design imposes real limits, stated here rather than deferred. Only three datasets exist within the shared protocol, so any correlation involving dataset-level properties rests on three points and is reported with its p-value rather than interpreted as an effect. The source studies publish point estimates without variance for most cells, so every dispersion figure below reflects spread across datasets and methods, not run-to-run noise. The per-channel decomposition rests on a single published ablation, which bounds how far it can be generalised; the tuned channel weighting from a second study serves as an independent qualitative cross-check, because its underlying sweep is published only as a line plot and its point values are not machine-readable and were left untranscribed. Feature extraction is held fixed across every cell, so channel quality and encoder quality cannot be separated here, a confound that a controlled extraction pipeline would be needed to break [17]. Claims are calibrated to match, and directional statements are preferred over precise effect sizes wherever the supporting sample is a single study.

4. RESULTS

4.1. Magnitude of the Content Contribution

A. Dependence on the Interaction-Only Reference

The size of the multimodal advantage is governed by the reference point. Measured against the matrix-factorisation baseline, 40 of the 42 multimodal cells improve, with a mean relative gain of 65.91% in Recall@20. Measured against the tuned LightGCN baseline on the same rows, only 30 of 42 improve, the mean gain falls to 10.29%, and the spread runs from −37.64% to 79.66%. The median strong-reference gain is 11.52%, close to the mean, indicating that the reduction is a level shift rather than the work of a few outliers.

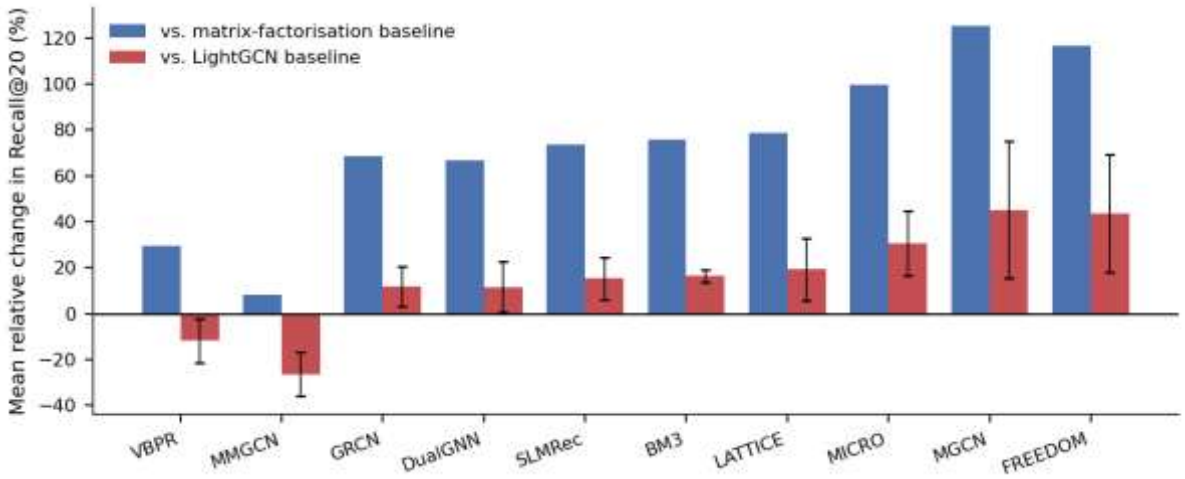

Figure 1. Mean relative change in Recall@20 for each multimodal method against two interaction-only references.

Each method is averaged over the three datasets and both source studies. Against the matrix-factorisation reference every method except MMGCN shows a large positive change, ranging up to 125.24% for MGCN. Against the LightGCN reference the same methods separate sharply: VBPR (−11.83%) and MMGCN (−26.42%) fall below the reference, the graph-refinement and self-supervised group clusters between 11.54% and 19.26%, and only MGCN (45.17%) and FREEDOM (43.53%) retain a large advantage. Error bars show the standard deviation of the strong-reference gain across datasets and studies.

B. Which Methods Survive the Stronger Reference

All twelve negative cells belong to exactly two methods, and both fail on every dataset in both studies: the matrix-factorisation-style method that concatenates features into item embeddings, and the earliest modality-specific graph convolution method. Methods that route content through a refined interaction graph, a frozen item-item graph, or a self-supervised auxiliary objective remain positive throughout. The pattern indicates that the deciding factor is not the presence of content but the mechanism by which it enters the model.

Table 3. Mean relative change in Recall@20 against each interaction-only reference.
Averaged across the three datasets and, where a method appears in both source studies, across studies.

Method	Family	vs. matrix factorisation (%)	vs. LightGCN (%)
VBPR	flat fusion	29.69	−11.83
MMGCN	modality-specific graph	8.08	−26.42
DualGNN	auxiliary user graph	66.70	11.54
GRCN	refined interaction graph	68.58	11.80
SLMRec	self-supervised	73.65	15.26
BM3	self-supervised	75.89	16.45
LATTICE	learned item-item graph	78.56	19.26
MICRO	contrastive item-item graph	99.69	30.50
FREEDOM	frozen item-item graph	116.62	43.53
MGCN	purified multi-view	125.24	45.17

The auxiliary user-graph method sits in the middle of the surviving group, consistent with its position in both source rankings [18].

4.2. Decomposition into Visual and Textual Channels

The published per-modality ablation isolates each channel against a behaviour-only variant. On Baby the text-only

variant reaches 0.0857 and the visual-only variant 0.0939 in Recall@20, against 0.0754 for behaviour alone, giving single-channel lifts of 13.66% and 24.54%. On Sports the corresponding values are 0.1017 and 0.1055 against 0.0864, giving 17.71% and 22.11%. On Clothing they are 0.0902 and 0.0919 against 0.0526, giving 71.48% and 74.71%.

The decisive quantity is what the second channel adds. Combining both channels reaches 0.0964, 0.1106, and 0.0945, which exceeds the better single channel by 2.66%, 4.83%, and 2.83%, a mean of 3.44%. Set against single-channel lifts averaging 40.45% for vision and 34.28% for text, the arithmetic implies that one content channel captures the large majority of the accessible benefit and the second contributes a modest increment. The two channels behave as substitutes far more than as complements on this data.

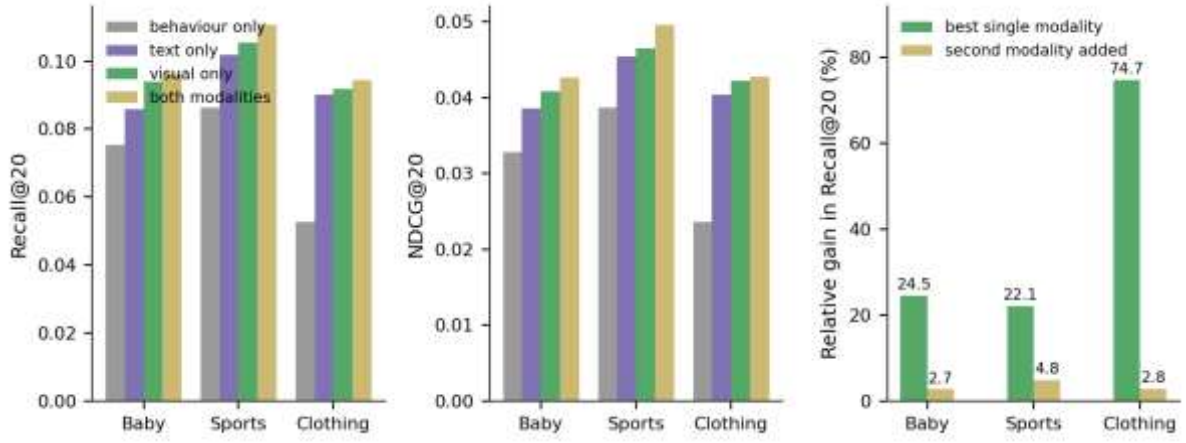

Figure 2. Decomposition of the content contribution into single-channel and second-channel components.

Panels (a) and (b) report Recall@20 and NDCG@20 for the behaviour-only, text-only, visual-only, and both-channel variants on each dataset; the visual-only bar exceeds the text-only bar on all three datasets and both metrics. Panel (c) contrasts the lift of the better single channel over behaviour alone (24.54%, 22.11%, and 74.71%) with the further gain obtained by adding the second channel (2.66%, 4.83%, and 2.83%), making the substitution pattern visible: the second channel recovers a small fraction of what the first already supplied.

Table 4. Per-channel decomposition of the content contribution, Recall@20.

Absolute values are as printed in the source ablation; lifts and gains are computed against the behaviour-only column and the stronger single channel respectively.

Dataset	Behaviour only	Text only	Visual only	Both	Text lift (%)	Visual lift (%)	Second-channel gain (%)
Baby	0.0754	0.0857	0.0939	0.0964	13.66	24.54	2.66
Sports	0.0864	0.1017	0.1055	0.1106	17.71	22.11	4.83
Clothing	0.0526	0.0902	0.0919	0.0945	71.48	74.71	2.83

4.3. Stability of the Attribution

A. A Documented Conflict Over Channel Dominance

The ablation above places the visual channel ahead of the textual channel on all three datasets, and its authors attribute this to shoppers responding to appearance while descriptions carry distracting material. A second study, working on the same three datasets with the same released features inside the same toolbox, tunes the weight of the visual channel in its item-item graph and settles on 0.1, leaving 0.9 to text, concluding that textual features are the more informative of the two. The two conclusions are directly opposed, and neither can be dismissed on data grounds. The mechanisms differ: one weights channels inside a fused representation consumed by a preference predictor, the other weights them only when building a similarity graph over items. Channel dominance emerges as a property of the fusion mechanism at least as much as of the catalogue, which is the central caution this analysis supports.

B. Backbone Dependence and Cross-Study Agreement

The backbone sweep sharpens the same point. Appending raw features to a matrix-factorisation backbone raises Recall@20 by 84.94% on average across the three datasets; the identical operation on a graph-based backbone yields 10.65%, and on a tuned LightGCN backbone only 3.84%,

with one of the nine cells negative, where 0.0782 falls to 0.0754. The measured worth of content shrinks by more than an order of magnitude as the interaction-only backbone strengthens.

Reproducibility across research groups is respectable but imperfect. For the five methods reported by both studies on all three datasets, mean absolute disagreement in Recall@20 is 2.35%, 2.43%, and 3.24%, with maxima of 6.82%, 4.87%, and 9.14%; the same method accounts for the largest disagreement on every dataset. Method ranking, by contrast, is stable: Kendall correlations between dataset pairs are 0.810, 0.714, and 0.905, all significant at the 5% level, averaging 0.810. Ranking which method wins is a reliable exercise; attributing the win to a channel is not.

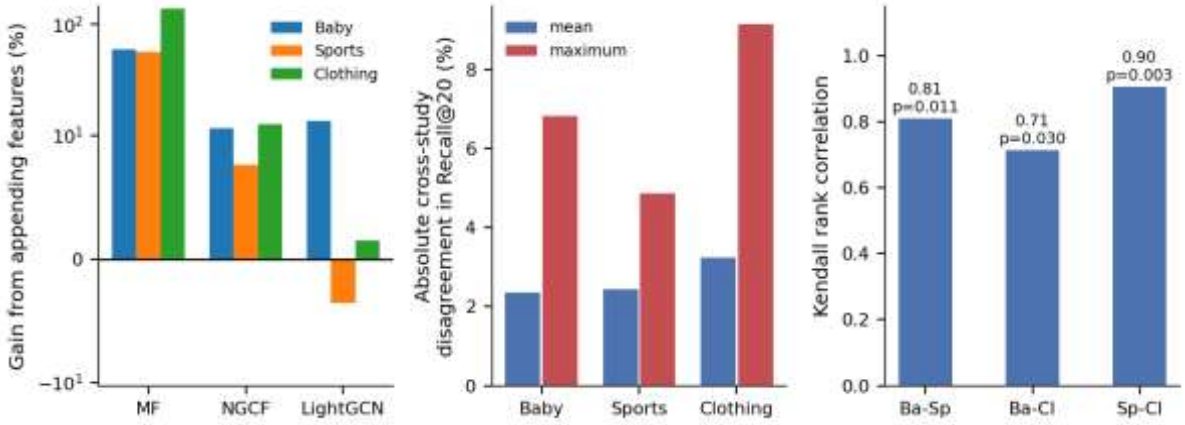

Figure 3. Backbone dependence and cross-study consistency.

Panel (a) shows the gain from appending raw content features to three collaborative filtering backbones on a symmetric-log scale, falling from a mean of 84.94% on matrix factorisation to 3.84% on LightGCN and turning negative in one cell. Panel (b) shows mean and maximum absolute disagreement in Recall@20 between the two source studies for the five methods they share, peaking at 9.14% on Clothing. Panel (c) shows Kendall rank correlations between dataset pairs, all between 0.714 and 0.905 with p-values below 0.05.

Table 5. Consistency statistics.

Quantity	Baby	Sports	Clothing
Mean cross-study disagreement, Recall@20 (%)	2.35	2.43	3.24
Maximum cross-study disagreement (%)	6.82	4.87	9.14
Mean multimodal gain vs. LightGCN (%)	6.59	6.09	18.18
Dataset density (%)	0.1173	0.0453	0.0307

The sparsest category shows the largest mean advantage, at 18.18% against 6.59% and 6.09%, matching the expectation that content compensates where interactions are thin. With three datasets the association is not statistically resolvable, at $r = -0.632$ and $p = 0.564$, and is reported as a direction rather than an effect. Independent work using hypergraph fusion reports the same directional pattern across interaction-sparsity strata [19], and content-rich benchmarks in the short-video domain have been released partly to test whether the pattern survives outside e-commerce [20].

5. DISCUSSION AND FUTURE WORK

5.1. Interpretation and Practical Reading

Three readings follow from the numbers above, each deliberately modest. The first concerns baselines. A practitioner deciding whether to build a content pipeline should compare against the strongest interaction-only model they can tune, not against matrix factorisation. The difference between those two reference points is the difference between an apparent 65.91% gain and a real 10.29% one, and for two of the ten methods examined it is the difference between a gain and a loss. Published improvement percentages are not portable across baseline configurations.

The second concerns channel selection. If one content channel supplies a lift of roughly 20% to 75% over behaviour alone and the second adds around 3%, then the engineering decision is which single channel to extract, not whether to extract both. Extraction, storage, and refresh costs for product imagery are substantially higher than for text, and a 3% increment rarely justifies them. The stated reservation is that this decomposition rests on one published ablation on three catalogue categories.

The third concerns attribution itself. Two competent studies reached opposite conclusions about channel dominance from identical inputs, while method ranking stayed stable across categories at an average rank correlation of 0.810. The honest summary is that a modality-contribution claim is a claim about a specific fusion mechanism operating on a specific catalogue, and reporting it without that qualification overstates what the evidence supports.

5.2. Limitations and Future Work

The limitations are structural. This is a secondary analysis of published point estimates, so it inherits whatever tuning asymmetries the source studies contain and cannot supply confidence intervals for individual cells. The shared protocol covers three catalogue categories from one retailer, which bounds external validity. The per-channel decomposition draws on a single ablation, and the cross-check against a second study is qualitative because the underlying sweep was published as a figure. Cold-start and long-tail slices, where content is widely expected to matter most, are not separable from the published aggregates, since the source studies report only catalogue-wide averages and do not release the per-user or per-item breakdowns that would allow those slices to be reconstructed after the fact.

Four extensions would tighten the picture. Re-running the decomposition across several backbones under one tuning budget would convert the observed conflict from a documented disagreement into a measured interaction effect. Repeating it with modern encoders in place of the ageing convolutional and sentence-transformer features would separate channel value from encoder quality. Extending to a domain where imagery plausibly dominates, such as short-video catalogues, would test whether the substitution pattern is a property of e-commerce or of the method family. Reporting per-channel ablations against a tuned interaction-only reference, as a routine element of multimodal recommendation papers, would make the attribution question answerable from the literature itself rather than through reconstruction.

REFERENCES

1. He, R., & McAuley, J. (2016). VBPR: Visual Bayesian personalized ranking from implicit feedback. *Proceedings of the AAAI Conference on Artificial Intelligence*, *30*(1), 144–150.
2. Wei, Y., Wang, X., Nie, L., He, X., Hong, R., & Chua, T.-S. (2019). MMGCN: Multi-modal graph convolution network for personalized recommendation of micro-video. *Proceedings of the 27th ACM International Conference on Multimedia*, 1437–1445.
3. Liu, Q., Hu, J., Xiao, Y., Zhao, X., Gao, J., Wang, W., Li, Q., & Tang, J. (2024). Multimodal recommender systems: A survey. *ACM Computing Surveys*, *57*(2), 1–17.
4. Zhou, H., Zhang, Y., Sun, A., & Shen, Z. (2025). Does multimodality improve recommender systems as expected? A critical analysis and future directions. *arXiv preprint arXiv:2508.05377*.
5. Yuan, Z., Yuan, F., Song, Y., Li, Y., Fu, J., Yang, F., Pan, Y., & Ni, Y. (2023). Where to go next for recommender systems? ID- vs. modality-based recommender models revisited. *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2639–2649.
6. Wei, Y., Wang, X., Nie, L., He, X., & Chua, T.-S. (2020). Graph-refined convolutional network for multimedia recommendation with implicit feedback. *Proceedings of the 28th ACM International Conference on Multimedia*, 3541–3549.
7. Zhang, J., Zhu, Y., Liu, Q., Wu, S., Wang, S., & Wang, L. (2021). Mining latent structures for multimedia recommendation. *Proceedings of the 29th ACM International Conference on Multimedia*, 3872–3880.
8. Zhang, J., Zhu, Y., Liu, Q., Zhang, M., Wu, S., & Wang, L. (2023). Latent structure mining with contrastive modality fusion for multimedia recommendation. *IEEE Transactions on Knowledge and Data Engineering*, *35*(9), 9154–9167.
9. Zhou, X., & Shen, Z. (2023). A tale of two graphs: Freezing and denoising graph structures for multimodal recommendation. *Proceedings of the 31st ACM International Conference on Multimedia*, 935–943.
10. Yu, P., Tan, Z., Lu, G., & Bao, B.-K. (2023). Multi-view graph convolutional network for multimedia recommendation. *Proceedings of the 31st ACM International Conference on Multimedia*, 6576–6585.
11. Tao, Z., Liu, X., Xia, Y., Wang, X., Yang, L., Huang, X., & Chua, T.-S. (2023). Self-supervised learning for multimedia recommendation. *IEEE Transactions on Multimedia*, *25*, 5107–5116.
12. Zhou, X., Zhou, H., Liu, Y., Zeng, Z., Miao, C., Wang, P., You, Y., & Jiang, F. (2023). Bootstrap

latent representations for multi-modal recommendation. *Proceedings of the ACM Web Conference 2023*, 845–854.

13. Ferrari Dacrema, M., Cremonesi, P., & Jannach, D. (2019). Are we really making much progress? A worrying analysis of recent neural recommendation approaches. *Proceedings of the 13th ACM Conference on Recommender Systems*, 101–109.
14. Ferrari Dacrema, M., Boglio, S., Cremonesi, P., & Jannach, D. (2021). A troubling analysis of reproducibility and progress in recommender systems research. *ACM Transactions on Information Systems, 39*(2), 1–49.
15. Anelli, V. W., Bellogín, A., Ferrara, A., Malitesta, D., Merra, F. A., Pomo, C., Donini, F. M., & Di Noia, T. (2021). Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2405–2414.
16. Zhou, X. (2023). MMRec: Simplifying multimodal recommendation. *Proceedings of the 5th ACM International Conference on Multimedia in Asia Workshops*, 1–2.
17. Malitesta, D., Gassi, G., Pomo, C., & Di Noia, T. (2023). Ducho: A unified framework for the extraction of multimodal features in recommendation. *Proceedings of the 31st ACM International Conference on Multimedia*, 9668–9671.
18. Wang, Q., Wei, Y., Yin, J., Wu, J., Song, X., & Nie, L. (2023). DualGNN: Dual graph neural network for multimedia recommendation. *IEEE Transactions on Multimedia, 25*, 1074–1084.
19. Guo, Z., Li, J., Li, G., Wang, C., Shi, S., & Ruan, B. (2024). LGMRec: Local and global graph learning for multimodal recommendation. *Proceedings of the AAAI Conference on Artificial Intelligence, 38*(8), 8454–8462.
20. Ni, Y., Cheng, Y., Liu, X., Fu, J., Li, Y., He, X., Zhang, Y., & Yuan, F. (2025). A content-driven micro-video recommendation dataset at scale. *Proceedings of the 34th ACM International Conference on Information and Knowledge Management*.
21. Zhou, S., Xu, F. F., Zhu, H., Zhou, X., Lo, R., Sridhar, A., Cheng, X., Ou, T., Bisk, Y., Fried, D., Alon, U., & Neubig, G. (2024). WebArena: A realistic web environment for building autonomous agents. International Conference on Learning Representations (ICLR 2024).
22. Xie, T., Zhang, D., Chen, J., Li, X., Zhao, S., Cao, R., Hua, T. J., Cheng, Z., Shin, D., Lei, F., Liu, Y., Xu, Y., Zhou, S., Savarese, S., Xiong, C., Zhong, V., & Yu, T. (2024). OSWorld: Benchmarking multimodal agents for open-ended tasks in real computer environments. Advances in Neural Information Processing Systems 37 (NeurIPS 2024), Datasets and Benchmarks Track.
23. Liu, X., Yu, H., Zhang, H., Xu, Y., Lei, X., Lai, H., Gu, Y., Ding, H., Men, K., Yang, K., Zhang, S., Deng, X., Zeng, A., Du, Z., Zhang, C., Shen, S., Zhang, T., Su, Y., Sun, H., ... Tang, J. (2024). AgentBench: Evaluating LLMs as agents. International Conference on Learning Representations (ICLR 2024).
24. Shang, W., Xu, W., & Zhou, Y. (2026). Feature Weight Optimization in Machine Learning Classifiers for Conflict Escalation Early Warning: Evidence from Diplomatic Signals and News Text. Journal of Sustainability, Policy, and Practice, 2(3), 26–37.
25. Shang, W., Zhao, F., & Liu, M. (2026). A Comparative Study of Spatiotemporal Clustering and Classification Approaches for Security Incident Risk Assessment in UN Peacekeeping Operations. Journal of Sustainability, Policy, and Practice, 2(3), 1–11.
26. Chen, Y., & Tang, T. (2026). Evaluating Prompt Engineering Strategies for Few-Shot Cyber Threat Intelligence Entity and Relation Extraction from Multi-Source Reports. Journal of Science, Innovation & Social Impact, 2(2), 153–164.
27. Li, Y., & Ling, Z. (2026). Real-Time Multi-Risk Early Warning for Community Banks: An Application of Ensemble Anomaly Detection and Explainable Artificial Intelligence. Journal of Advanced Computing Systems, 6(2), 15–27.
28. Li, Y., & Zhang, S. (2025). Machine Learning-Based Credit Risk Early Warning System for Small and Medium-Sized Financial Institutions: An Ensemble Learning Approach with Interpretable Risk Indicators. Journal of Science, Innovation & Social Impact, 1(1), 372–383.
29. Li, Y. (2026). Enhancing Financial Compliance Transparency through Automated Data Governance and Intelligent Risk Reporting. Journal of Science, Innovation & Social Impact, 2(1), 299–313.
30. Li, Y., Zhang, X., & Hao, C. (2026). Graph Neural Network-Based Cross-Market Risk Contagion Analysis between US Equity Sectors and Treasury Yields during the 2023 Regional Banking Stress. Journal of Sustainability, Policy, and Practice, 2(4), 154–168.

31. Li, Y., & Fu, X. (2026). Comparative Evaluation of Graph Neural Networks for Cross-Market Risk Contagion Path Identification in Multi-Layer Financial Networks. Journal of Sustainability, Policy, and Practice, 2(3), 1–14.
32. Li, Y., Zhao, F., & Hu, J. (2026). Identifying Cross-Market Risk Contagion Amplifiers via Graph Attention Networks: Empirical Evidence from US Financial Stress Periods. Journal of Computing Innovations and Applications, 4(1), 164–175.
33. Chen, Y., & Hu, J. (2026). Graph Neural Network-Based Cascading Disruption Path Identification in Multi-Tier Rare Earth Processing Networks. Journal of Global Engineering Review, 4(1), 99–112.
34. Chen, Y., & Lai, J. (2026). Multi-Metric Trustworthiness Evaluation of AI-Assisted Medical Imaging Diagnosis: Integrating Confidence Calibration and Distribution Shift Detection. Journal of Global Engineering Review, 4(1), 113–126.
35. Chen, Y. (2026). Multi-Dimensional Training Data Bias Detection and Fairness-Aware Augmentation for Equitable AI-Assisted Medical Imaging Diagnosis. Journal of Sustainability, Policy, and Practice, 2(4), 1–15.
36. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. International Conference on Learning Representations (ICLR 2023).
37. Shinn, N., Cassano, F., Berman, E., Gopinath, A., Narasimhan, K., & Yao, S. (2023). Reflexion: Language agents with verbal reinforcement learning. Advances in Neural Information Processing Systems 36 (NeurIPS 2023).
38. Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., & Lim, E.-P. (2023). Plan-and-Solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL 2023), 2609–2634.
39. Yao, S., Yu, D., Zhao, J., Shafran, I., Griffiths, T. L., Cao, Y., & Narasimhan, K. (2023). Tree of Thoughts: Deliberate problem solving with large language models. Advances in Neural Information Processing Systems 36 (NeurIPS 2023).
40. Chen, Y. (2024). Explainable Attack Path Reasoning for Industrial Control Network Security Based on Knowledge Graphs. Journal of Computing Innovations and Applications, 2(1), 128–139.
41. Wang, G., Xie, Y., Jiang, Y., Mandlekar, A., Xiao, C., Zhu, Y., Fan, L., & Anandkumar, A. (2024). Voyager: An open-ended embodied agent with large language models. Transactions on Machine Learning Research.
42. Zheng, B., Gou, B., Kil, J., Sun, H., & Su, Y. (2024). GPT-4V(ision) is a generalist web agent, if grounded. Proceedings of the 41st International Conference on Machine Learning (ICML 2024).
43. He, H., Yao, W., Ma, K., Yu, W., Dai, Y., Zhang, H., Lan, Z., & Yu, D. (2024). WebVoyager: Building an end-to-end web agent with large multimodal models. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024).
44. Cao, H., Hu, J., & Luo, C. (2025). Behavioural Feature Analysis for Anomalous Click Detection in Mobile Advertising Environments: Toward In-App Browser-Specific Detection. Journal of Computing Innovations and Applications, 3(2), 96–105.
45. Cao, H. (2026). An Empirical Analysis of Click Temporal Features for Automated Ad Fraud Detection in Mobile In-App Browser Environments. Journal of Sustainability, Policy, and Practice, 2(4), 142–153.
46. Liu, E. Z., Guu, K., Pasupat, P., Shi, T., & Liang, P. (2018). Reinforcement learning on web interfaces using workflow-guided exploration. International Conference on Learning Representations (ICLR 2018).
47. Mialon, G., Fourrier, C., Swift, C., Wolf, T., LeCun, Y., & Scialom, T. (2024). GAIA: A benchmark for general AI assistants. International Conference on Learning Representations (ICLR 2024).
48. Peng, Y., & Gao, S. (2026). An Empirical Comparison of XGBoost and LightGBM for Capital Expenditure Deviation Prediction in US FERC-Regulated Rate-Base Transmission Projects. Journal of Sustainability, Policy, and Practice, 2(4), 90–102.
49. Hu, J., Wang, X., & Lai, J. (2026). Benchmarking Learned Cardinality Estimation Techniques for Analytical Query Processing in Data Warehouses. Journal of Computer Technology and Applied Mathematics, 3(3), 1–8.
50. Hu, J., & Long, X. (2024). Graph Learning-Based Behavioral Detection for Software Supply Chain Attacks. Journal of Advanced Computing Systems, 4(4), 49–60.
51. Wen, S., & Tang, T. (2025). A Comparative Evaluation of URL-Sharing, Content Similarity, and Temporal Synchronicity Signals for Detecting Coordinated Inauthentic Behavior in Multilingual Political Discourse. Journal of Global Engineering Review, 3(2), 69–78.
52. Koh, J. Y., Lo, R., Jang, L., Duvvur, V., Lim, M. C., Huang, P.-Y., Neubig, G., Zhou, S., Salakhutdinov,

R., & Fried, D. (2024). VisualWebArena: Evaluating multimodal agents on realistic visual web tasks. Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024), 881–905.

53. Deng, X., Gu, Y., Zheng, B., Chen, S., Stevens, S., Wang, B., Sun, H., & Su, Y. (2023). Mind2Web: Towards a generalist agent for the web. Advances in Neural Information Processing Systems 36 (NeurIPS 2023), Datasets and Benchmarks Track.
54. Yao, S., Chen, H., Yang, J., & Narasimhan, K. (2022). WebShop: Towards scalable real-world web interaction with grounded language agents. Advances in Neural Information Processing Systems 35 (NeurIPS 2022).
55. Cao, H., & Shi, W. (2026). Statistical Anomaly Detection Approach for Field Mapping Validation in Enterprise Payroll Data Migration. Journal of Computing Innovations and Applications, 4(1), 137–153.
56. Cao, H., & Liu, M. (2026). Empirical Evaluation of Constraint Discovery Techniques for Business Logic Consistency Verification in Payroll Data Migration. Journal of Sustainability, Policy, and Practice, 2(3), 92–103.
57. Cao, H., & Long, L. (2026). Empirical Evaluation of Multi-Source Monitoring Signal Effectiveness and Lead Time for Performance Degradation Prediction in Kubernetes-Based Microservices. Journal of Advanced Computing Systems, 6(4), 15–26.
58. Long, X., Hu, J., & Ling, Z. (2026). A Comparative Analysis of Telemetry-Driven Anomaly Detection Approaches for Dual-Purpose Operational and Security Optimization in Edge Computing Infrastructure. Journal of Computing Innovations and Applications, 4(1), 79–88.
59. Cao, H. (2024). Privacy-Preserving Click Pattern Anomaly Detection for Mobile In-App Browser Advertising Fraud. Journal of Computing Innovations and Applications, 2(2), 151–161.
60. Cao, H. (2024). Detecting Fraudulent Click Patterns in Mobile In-App Browsers: A Multi-dimensional Behavioral Analysis Approach. Artificial Intelligence and Machine Learning Review, 5(2), 130–142.
61. Qin, Y., Liang, S., Ye, Y., Zhu, K., Yan, L., Lu, Y., Lin, Y., Cong, X., Tang, X., Qian, B., Zhao, S., Hong, L., Tian, R., Xie, R., Zhou, J., Gerstein, M., Li, D., Liu, Z., & Sun, M. (2024). ToolLLM: Facilitating large language models to master 16000+ real-world APIs. International Conference on Learning Representations (ICLR 2024).
62. Cemri, M., Pan, M. Z., Yang, S., Agrawal, L., Chopra, B., Tiwari, R., Keutzer, K., Parameswaran, A., Klein, D., Ramchandran, K., Gonzalez, J. E., Zaharia, M., & Stoica, I. (2025). Why do multi-agent LLM systems fail? arXiv preprint arXiv:2503.13657.
63. Chen, Y., & Chen, Z. (2025). Multi-Objective Deep Reinforcement Learning for Carbon-Aware Spatiotemporal Workload Scheduling in Geo-Distributed Data Centers. Journal of Advanced Computing Systems, 5(10), 18–30.
64. Long, L., & Hu, J. (2026). Multi-Objective Particle Swarm Optimization for Site Selection and Policy Subsidy Maximization of Foreign Renewable Energy Enterprises in the United States. Artificial Intelligence and Machine Learning Review, 7(2), 54–69.
65. Li, Y., & Long, L. (2026). Lightweight AI-Driven Stress Testing for Small and Medium Financial Institutions: A Variational Autoencoder Approach with Extreme Value Theory for Macroeconomic Scenario Generation. Artificial Intelligence and Machine Learning Review, 7(1), 108–119.
66. Wu, C., Guan, H., & Weng, H. (2024). Forecasting Hospital Resource Demand Using Gradient Boosting: An Operational Analytics Approach for Bed Allocation and Patient Flow Management. Journal of Computing Innovations and Applications, 2(1), 74–85.
67. Chen, Y., Chen, Z., & Zou, D. (2025). CarbonShift: Harnessing Grid Carbon Variability for Geo-Distributed Workload Scheduling. Artificial Intelligence and Machine Learning Review, 6(4), 18–31.
68. Wei, C., & Wu, C. (2024). Credit Risk Transmission Mechanism and Prevention Strategies in Supply Chain Finance: A Core Enterprise Perspective. Artificial Intelligence and Machine Learning Review, 5(2), 101–115.
69. Shridhar, M., Yuan, X., Côté, M.-A., Bisk, Y., Trischler, A., & Hausknecht, M. (2021). ALFWorld: Aligning text and embodied environments for interactive learning. International Conference on Learning Representations (ICLR 2021).
70. Patil, S. G., Zhang, T., Wang, X., & Gonzalez, J. E. (2024). Gorilla: Large language model connected with massive APIs. Advances in Neural Information Processing Systems 37 (NeurIPS 2024).
71. Wu, C., & Pan, Z. (2024). An Integrated Graph Neural Network and Reinforcement Learning Framework for Intelligent Drug Discovery. Journal of Advanced Computing Systems, 4(6), 19–29.

72. Furuta, H., Lee, K.-H., Nachum, O., Matsuo, Y., Faust, A., Gu, S. S., & Gur, I. (2024). Multimodal web navigation with instruction-finetuned foundation models. International Conference on Learning Representations (ICLR 2024).
73. Li, Y., Wu, C., Li, L., Liu, Y., & Zhu, J. (2021). Caption generation from road images for traffic scene modeling. IEEE Transactions on Intelligent Transportation Systems, 23(7), 7805–7816.
74. Li, L., Li, Y., Wu, C., Dong, H., Jiang, P., & Wang, F. (2021). Detail fusion GAN: High-quality translation for unpaired images with GAN-based data augmentation. 2020 25th International Conference on Pattern Recognition (ICPR), 1731–1736.
75. Huo, H., Li, Y., Wu, C., Wu, X., Tian, Z., & Liu, Y. (2019). Semantic segmentation and scene reconstruction for traffic simulation using CNN. 2019 2nd China Symposium on Cognitive Computing and Hybrid Intelligence.
76. Wu, X., Li, Y., Hao, Z., Wu, C., Wang, X., & Liu, Y. (2019). Image style transformation based on structure GAN. 2019 Chinese Automation Congress (CAC), 2002–2007.
77. Wu, C., Li, Y., Li, L., Wang, L., & Liu, Y. (2020). Caption generation from road images for traffic scene construction. 2020 IEEE Intelligent Vehicles Symposium (IV), 1271–1276.
78. Wu, X., Li, Y., Liu, Y., Pang, S., Wang, L., Wu, C., & Huo, H. (2019). Jointly detecting and retrieving vehicles from road image sequences based on CNN. 2019 IEEE Intelligent Vehicles Symposium (IV), 2530–2535.
79. Shi, X., & Weng, H. (2024). Comparative analysis of unsupervised learning approaches for anomalous billing pattern detection in healthcare payment integrity. Journal of Computing Innovations and Applications, 2(1), 111–127.
80. Weng, H., Yang, Q., Fu, H., Pan, H., & Lv, X. (2026). When retrieval hurts code completion: A diagnostic study of stale repository context. arXiv preprint arXiv:2605.14478.
81. Weng, H. (2025). Deep embedding clustering with adaptive feature selection for banking customer segmentation. Spectrum of Research, 5(2).
82. Weng, H., & Lei, Y. (2024). Cross-modal artifact mining for generalizable deepfake detection in the wild. Journal of Computing Innovations and Applications, 2(2), 78–87.
83. Weng, H., & Li, X. (2024). Renewable-aware cooperative scheduling for distributed AI training across geo-distributed data centers. Artificial Intelligence and Machine Learning Review, 5(2), 91–100.
84. Zhang, S., Wang, Y., & Weng, H. (2024). Industrial IoT anomaly detection using improved autoencoder architecture. Artificial Intelligence and Machine Learning Review, 5(1), 67–78.
85. Weng, H., Wang, H., & Wei, C. (2024). Adaptive bidding strategies for hybrid auction mechanisms in programmatic advertising. Journal of Advanced Computing Systems, 4(4), 13–25.
86. Weng, H., Zhang, S., & Min, S. (2024). Multi-constraint optimization for real-time bidding: A reinforcement learning approach. Artificial Intelligence and Machine Learning Review, 5(1), 93–104.
87. Weng, H. (2026). Predictive power versus feature stability: An empirical comparison of filter, embedded, and wrapper feature selection methods for consumer credit pre-approval. Academia Nexus Journal, 5(1).