Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

From Feature Fusion to FoundationModels: A Systematic Review of Multimodal Hashtag Recommendationin Social Media (2016--2025)

Hi Teng Heon

gs73117@student.upm.edu.my

Universiti Putra Malaysia

Azreen Azman

Universiti Putra Malaysia

Fatimah Khalid

Universiti Putra Malaysia

Puteri Suhaiza Sulaiman

Universiti Putra Malaysia

Research Article

Keywords: Hashtag recommendation, Multimodal learning, Systematic literature review, Vision-language models, Retrieval-augmented generation

Posted Date: November 17th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11232903/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

From Feature Fusion to Foundation Models: A Systematic Review of Multimodal Hashtag Recommendation in Social Media (2016–2025)

Heon Hi Teng[1*], Azreen Azman[1], Fatimah Khalid[1], Puteri Suhaiza Sulaiman[1]

[1*]Faculty of Computer Science and Information Technology, Universiti Putra Malaysia, Serdang, 43400, Selangor, Malaysia.

*Corresponding author(s). E-mail(s): gs73117@student.upm.edu.my;

Abstract

Hashtags enable content organisation, discoverability, and community formation on social media, yet automated recommendation of contextually appropriate hashtags from multimodal post content—text, images, audio, and video together—remains underexplored as a distinct task. To the best of our targeted search, no existing review synthesises multimodal hashtag recommendation (MHR) specifically. Following PRISMA 2020 [59], we searched IEEE Xplore, ACM Digital Library, Scopus, and SpringerLink for studies published January 2016–September 2025. After independent two-stage screening (Cohen's $\boldsymbol{\kappa}$**=0.81** and **0.85** [60]), 30 studies met eligibility and were assessed with a five-criterion quality rubric (mean **5.9±1.8**). We classify included studies into six architectural families—early fusion, late fusion, attention-based, graph-enhanced, transformer/LLM/RAG, and an emerging neurosymbolic-contrastive direction—and introduce the *Hashtag Intelligence Maturity Model* (HIMM), a five-level framework assessing MHR systems by functional capability. Image–text combinations dominate the corpus, with micro-video and retrieval-augmented generation (RAG) as emerging subfields. A reproducibility and fairness audit finds diversity, fairness, and explainability metrics essentially absent from current evaluations, accessible code in only 7% of studies, and multilingual coverage of only 13%. We identify six structural research gaps and a six-directional research agenda addressing foundation-model integration, generative hashtag synthesis, cross-platform benchmarking, multilingual coverage, explainability, and ethical deployment.

Keywords: Hashtag recommendation; Multimodal learning; Systematic literature review; Vision-language models; Retrieval-augmented generation

1 Introduction

1.1 Background and Motivation

Social media platforms have evolved into richly multimodal information environments where text, images, short-form video, and audio coexist within a single post. Hashtags—tokens prefixed with the '#' symbol—function as user-generated metadata that facilitate content indexing, trend navigation, and community formation across Instagram, Twitter/X, TikTok, and YouTube Shorts [1, 2, 6]. Posts with hashtags receive substantially higher engagement than those without [40], making accurate hashtag selection materially important for creators, platform operators, and brand strategists alike.

Multimodal hashtag recommendation (MHR)—the automated recommendation of contextually appropriate hashtags from multimodal post content—is therefore a problem of growing practical and scientific importance. For researchers, MHR demands bridging the semantic gap between heterogeneous input signals and a concise, socially contextualised indexing token.

Early approaches relied on textual representations: TF-IDF [8], latent Dirichlet allocation [9], and word embeddings [10]. These unimodal methods fail when posts carry minimal captions. CNN-derived visual features [11] established the multimodal baseline; audio from spectrograms and MFCCs was subsequently incorporated for micro-video platforms [12].

Vision-language pretraining has since reshaped the design space. LXMERT [13] and ViLBERT [14] demonstrated cross-modal encoding via transformer cross-attention. CLIP [4]—trained on 400 million image-text pairs—enabled zero-shot cross-modal retrieval. BLIP [49] and BLIP-2 [50] extended this to generative outputs. Graph neural networks [5] modelled relational hashtag-user-content structures. Most recently, retrieval-augmented generation (RAG) [41] emerged as a paradigm that combines retrieval precision with generative flexibility.

Despite this rapid progress, the field still lacks a comprehensive task-specific systematic review mapping MHR's architectural evolution, evaluating reproducibility, and providing a forward-looking research agenda. This review addresses that gap.

1.2 Problem Statement

MHR research is fragmented across computer vision, natural language processing, information retrieval, and recommender systems, with no shared benchmark, no unified problem formulation, and no community consensus on evaluation practice. Studies variously frame hashtag recommendation as multi-label classification [20, 21], ranking [22, 23], matching [30], or open-vocabulary sequence generation [41, 47]. This formulation diversity means that published performance numbers are almost never directly comparable: a model achieving Precision@5 = 0.78 on an Instagram corpus of 800K posts cannot be validly compared against a model achieving Precision@5 = 0.63 on a Twitter/Weibo corpus of 120K posts, because the comparison conflates dataset characteristics, label vocabulary size, language mix, and architectural choice simultaneously.

Beyond comparability, evaluation practice is structurally incomplete. Precision@K and Recall@K are reported universally, but diversity, novelty, fairness, and explainability metrics are nearly absent from the literature. This matters because hashtag recommendation is not a neutral retrieval task: models trained on popularity-skewed corpora will systematically amplify already-visible content and marginalise creators in low-resource languages and underrepresented niches. Zero studies in the current corpus employ fairness or demographic disparity metrics, an absence that becomes critically important as regulatory frameworks including the EU AI Act (2024) extend their reach to automated content-affecting systems.

Reproducibility constitutes a third dimension of the problem. Nearly all datasets are platform-restricted and non-shareable due to terms-of-service constraints; only 7% of studies release executable code. This means that for the vast majority of published models, neither the data nor the implementation is available for independent replication. English- and Chinese-language dominance further limits global generalisability [18], with multilingual MHR only recently addressed [40] for Indic language contexts. The absence of a comprehensive, task-specific systematic review that maps these structural challenges, evaluates the architectural trajectory, and formulates a forward-looking agenda is the gap this work addresses.

1.3 Positioning Against Prior Reviews

To situate this review, we conducted a targeted supplementary search for prior surveys and reviews touching hashtag recommendation, multimodal recommendation, and tag recommendation, run against the same four databases plus Google Scholar and arXiv listings. Table 1 summarises the closest matches identified by this search. The most directly relevant is Bansal et al.'s 2025 preprint review of hashtag recommendation [52], which surveys traditional-to-deep-learning hashtag recommendation broadly but does not apply a PRISMA protocol, does not restrict to multimodal methods specifically, and does not include a reproducibility, fairness, or maturity-model analysis. Deldjoo

et al. [51] review generative recommender systems broadly and are not hashtag-specific. On the basis of this search, we did not locate a PRISMA-compliant systematic review focused specifically on multimodal hashtag recommendation; we state this as a finding of our targeted search rather than a claim of exhaustive novelty, since a review published outside our four databases or after our search cutoff could exist without our knowledge. Key differentiators of this review are the n=30 PRISMA-compliant corpus, the HIMM framework (Sect. 5), a structured reproducibility and fairness audit, and coverage of RAG and neurosymbolic architectures absent from the surveys identified above.

1.4 Objectives and Research Questions

Six contributions are delivered: (C1) a six-family architectural taxonomy; (C2) the HIMM framework; (C3) a directional cross-architecture performance comparison with inline citations to primary sources; (C4) a dataset and reproducibility assessment; (C5) a vote-counting synthesis across 15 dimensions; (C6) a six-directional future research agenda.

Four research questions structure the synthesis: **RQ1** (Methods & Taxonomy) — what fusion architectures have been proposed and how can they be categorised? **RQ2** (Data & Metrics) — what datasets and metrics are used and what limits benchmarking? **RQ3** (Reproducibility) — how reproducible is the corpus? **RQ4** (Gaps) — what challenges persist and which future directions are most urgent?

2 Review Methodology

This review was conducted in accordance with PRISMA 2020 guidelines [59]. The protocol was specified *a priori*; no formal pre-registration was completed prior to commencement. Future iterations should be registered on PROSPERO to enhance methodological transparency.

2.1 Eligibility Criteria

Table 2 presents the PICOS eligibility criteria.

2.2 Search Strategy

Searches were executed on 5 September 2025 across IEEE Xplore, ACM Digital Library, Scopus, and SpringerLink.

Currency update. To confirm the corpus remained current, the search was supplemented with a targeted update pass covering the period from the original cutoff through 1 October 2026, run against the same core terms across general web, arXiv, and publisher-index coverage. This update identified no additional peer-reviewed study meeting the eligibility criteria in Table 2 beyond the n=30 already included; one candidate record initially appeared promising (an IEEE-indexed multimodal hashtag-recommendation framework) but was confirmed, on inspection, to be an alternate indexing of an already-included study [42] rather than a new one. We report this as a currency check with a null result, not as a second formal PRISMA screening cycle: it was conducted by one reviewer, without the independent dual-screening or κ-agreement procedure applied to the primary search, and should be read accordingly.

```
("hashtag recommend*" OR "tag recommend*"
OR "hashtag suggestion" OR "hashtag
generation") AND ("social media" OR
twitter OR instagram OR tiktok OR
weibo OR facebook OR "micro-video"
OR "microblog") AND (multimodal OR
"vision-language" OR crossmodal OR image
OR video OR transformer OR GNN OR "graph
neural" OR CLIP OR BLIP OR ViT OR LLM
OR "foundation model" OR fusion OR
"retrieval-augmented" OR contrastive OR
neurosymbolic)
```

Box 1. Master Boolean search string (adapted per database syntax).

Sensitivity analysis. To assess whether inclusion of model-specific identifiers (CLIP, BLIP, ViT, LLM, etc.) artificially broadened the search, a reduced string omitting all model identifiers was executed against the same four databases on the same date. This produced fewer than 12 additional records beyond the primary search, all of which were excluded at full-text review for addressing a unimodal task only, falling outside the social-media domain, or being inaccessible in full text; none met the multimodal eligibility criterion. The primary search string therefore introduced no meaningful retrieval bias attributable to model identifiers for this

Table 1 Positioning of this review against surveys identified by our targeted search

Prior survey	Scope	Key limitation	Coverage	How this review differs
Li et al. (2019)	Hashtag rec	Predates transformers	Pre-2019	2016–2025; LLM, RAG, neurosymbolic included
Zhou (2021) TKDE	Multimodal rec	Not hashtag-specific	Pre-2021	Task-specific: MHR only
Gu (2023) ACM CS	Tag rec broadly	Mixes social & scholarly	Pre-2023	Social media only
Deldjoo et al. (2023) [51]	Generative DL rec	No repro audit; not hashtag-specific	Pre-2023	PRISMA; repro & fairness audits
Bansal et al. (2025, preprint) [52]	Hashtag rec (uni-& multimodal)	Not PRISMA; no HIMM, repro, or fairness audit	Through 2024	PRISMA; multimodal-specific; HIMM; repro & fairness audits
This review	**MHR only**	—	**2016–2025**	n=30**; HIMM; RAG; neurosymbolic; multilingual**

Table 2 Study eligibility criteria (PICOS framework)

Criterion	Inclusion	Exclusion
Population	Social media posts with hashtags on Instagram, Twitter/X, TikTok, YouTube, Weibo, and related platforms.	Non-social domains (e-commerce, scholarly tagging, biomedical) unless benchmarked against social MHR.
Intervention	Automated multimodal hashtag recommendation integrating ≥2 modalities; or a unimodal study admitted only if it introduces a benchmark dataset subsequently adopted by ≥3 other included multimodal studies (operationalising "foundational baseline"). This rule was applied, in the current corpus, to a single study (HARRISON [25], Sect. 3).	Purely unimodal methods, except where they meet the foundational-baseline rule at left.
Comparator	Any relevant baseline applicable to multimodal recommendation.	No empirical comparison or no quantitative results.
Outcomes	P@K, R@K, MAP, nDCG, F1; diversity, novelty, fairness, explainability where reported.	Qualitative contributions only; no empirical validation.
Study type	Peer-reviewed journal, conference, and workshop papers in English, January 2016 to September 2025.	Books, editorials, unreviewed preprints, non-English publications, dissertations, theses.

corpus. The full per-record exclusion log for this sensitivity search is retained by the authors and is available on request.

2.3 Screening and Selection

Stage 1 (Deduplication): 136 records identified (IEEE 25, ACM 13, Scopus 68, SpringerLink 30); 19 duplicates removed; 117 after deduplication.

Stage 2 (Title/abstract): 117 records screened by two independent reviewers; κ=0.81 (95% CI: 0.74–0.88, strong agreement per [60]); 32 excluded.

Stage 3 (Full-text): 85 sought; 15 not retrieved; 70 assessed; κ=0.85 (CI: 0.79–0.91); 40 excluded: unimodal only (n=19), no quantitative evaluation (n=9), non-social-media domain (n=8), inaccessible (n=4).

Stage 4 (Included): 30 studies [19]–[48].

Figure 1 presents the PRISMA 2020 flow diagram.

2.4 Data Extraction and Quality Appraisal

A structured extraction form (piloted on five studies) captured: modality types, fusion strategy, taxonomy

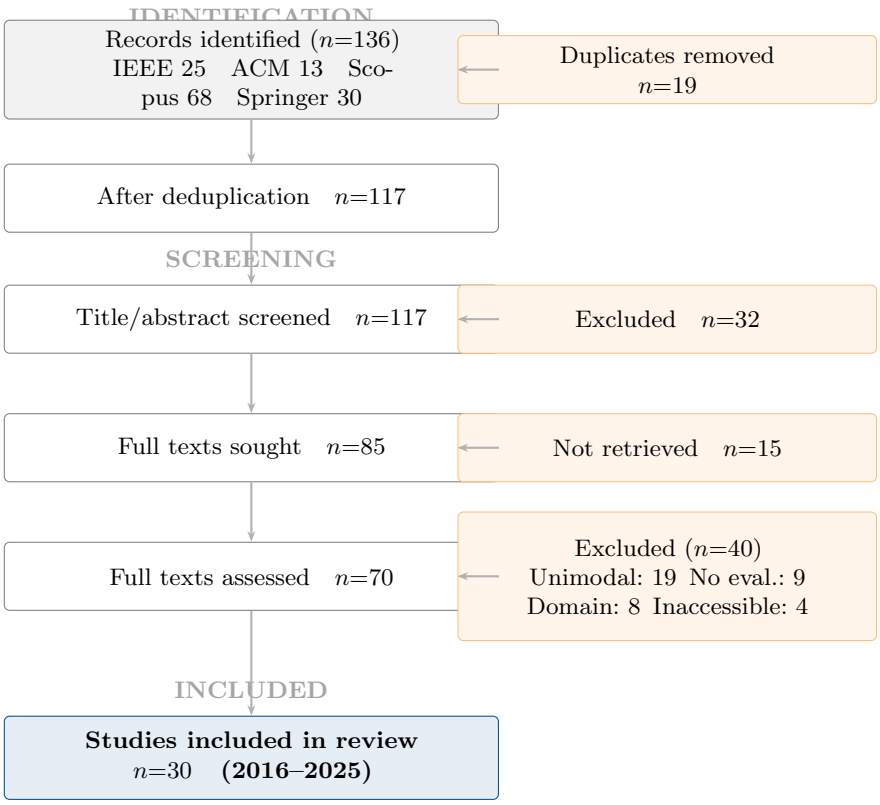


Fig. 1 PRISMA 2020 study selection flow diagram. Inter-rater agreement: κ=0.81 (title/abstract) and κ=0.85 (full-text) [60]

class, backbone encoders, dataset characteristics, evaluation metrics, and reproducibility indicators. A five-criterion quality rubric (QA1 methodological clarity; QA2 evaluation adequacy; QA3 dataset transparency; QA4 reproducibility; QA5 comparative rigour; scored 0–2 each, maximum 10) was applied to every included study. Mean quality score: 5.9±1.8.

3 Descriptive Analysis

3.1 Included Studies

Three inclusion decisions require brief justification. Hachaj & Miazga [25] is image-only but is included as a foundational baseline under the rule in Table 2: it was evaluated on the pre-existing HARRISON benchmark (not introduced by this study) and that benchmark is subsequently used by eight other studies in this corpus, and it provides the only fully public code and dataset in the entire corpus. Shojaei et al. [45] present a generation framework for AI-influencer posts; inclusion is justified because their pipeline ranks multimodal candidates against a platform vocabulary, satisfying the criterion of *automated multimodal hashtag ranking integrating ≥2 modalities*. Fan et al.'s RIGHT [41] is a boundary case in the opposite direction—a primarily text-based retrieval-augmented generator rather than a clearly multimodal-fusion architecture; we retain it and discuss the eligibility trade-off explicitly in Sect. 4.

A full per-study listing of all 30 included primary studies — year, authors, abbreviated title, modality combination, venue/publication type, taxonomy-family assignment, dataset platform/scale/accessibility, and reproducibility indicators — is provided in full in Online Resource 2 (ESM_2.pdf) rather than reproduced here, to keep the main text within the journal's page limit; the corpus-level patterns these support (publication trends, modality mix, and taxonomy distribution) are summarised in Figs. 2–5 and Table 4 below.

3.2 Publication Trends

The temporal distribution of the 30 included studies reveals three structurally distinct phases, each driven by a specific technological catalyst.

Phase 1 (2016–2018): Multimodal baseline establishment. The three eligible studies from 2016–2017 [19–21] establish the MHR problem framing and the first multimodal architectures, drawing on convolutional visual features and early word embeddings. No study met eligibility criteria for 2018. This gap is not

a search artefact: it corresponds to the period immediately preceding TikTok's global launch (2018Q3), before the transformer paradigm [3] had penetrated social media tasks, and before micro-video had emerged as a dominant content format driving multimodal hashtag needs.

Phase 2 (2019–2022): Micro-video and GNN proliferation. A pronounced inflection occurs in 2019 (five studies), coinciding with the rapid growth of micro-video platforms (TikTok, Kuaishou, Weibo) and the introduction of the MicroVideo-1.7M benchmark [23]. The 2020–2022 period is dominated by attention mechanisms, graph neural networks, and the establishment of recurring baseline datasets (HARRISON [25], Instagram-600K [33]). Overall, 53.3% of the corpus (n=16) appears in journals and 46.7% (n=14) in conference proceedings, a roughly even split that is consistent with an actively developing sub-field publishing through both fast-turnaround venues and longer-form journal contributions.

Phase 3 (2023–2025): Foundation models and architectural diversification. The 2023–2025 cluster (eleven studies) reflects the transformative impact of pretrained vision-language models (CLIP [4], BLIP [49]) and large language models on MHR design. Transformer-based and LLM-augmented architectures appear for the first time; RAG is introduced as a production-oriented paradigm [41]; and the neurosymbolic family emerges as a qualitatively new approach to interpretable recommendation [42]. The single 2025 study [47] reflects the September 2025 search cutoff rather than an absolute plateau; given the pace of foundation-model development, the 2025–2026 period is expected to yield substantially more transformer and generative MHR contributions.

Figure 2 illustrates the temporal distribution of publications across the three identified phases.

3.3 Dataset Characteristics

Dataset sizes range from 50K in early works to over 1.2M in recent studies. Virtually all exhibit a pronounced long-tail hashtag distribution. Only five datasets are accessible in any form (17%): the HARRISON benchmark [25] and IndiaHash corpus [40] offer full public access; three others [23, 24, 28] share partial tweet IDs. This pervasive inaccessibility renders faithful replication impossible for most studies. The full per-study platform, scale, language, and accessibility breakdown underlying these figures is provided as Table 2 in Online Resource 2 (ESM_2.pdf).

The dataset-level modality mix mirrors the study-level modality distribution presented in Fig. 4 (Sect. 4): image+text content dominates, video-based platforms (micro-video, short-form video) form the second-largest group, and knowledge-graph-augmented and text-only datasets remain a minority. We present this breakdown once, in Fig. 4, rather than twice, to avoid duplicating the same underlying counts under two different figures.

3.4 Quantitative Synthesis (Table 3)

Table 3 presents a vote-counting synthesis [58] across 15 binary dimensions. Vote-counting is appropriate given heterogeneous datasets and metrics that preclude formal meta-analysis.

Universal Precision@K adoption and zero fairness metrics confirm that evaluation is both homogeneous and incomplete. Only 2 studies (7%) provide accessible code, and only 5 datasets (17%) are accessible in any form, constituting a systemic reproducibility deficit. Multi-platform evaluation is entirely absent; multilingual evaluation reaches only 13% of studies.

4 Methodological Taxonomy (RQ1)

The 30 studies are classified into six architectural families. Table 4 presents the complete mapping.

Figure 4 illustrates the distribution of modality combinations across included studies; their evolution over time is discussed below.

Modality combinations have also shifted over time: the 2016–2018 baseline-establishment phase is dominated by image+text and video+audio+text studies; the 2019–2022 phase introduces micro-video-specific and graph-augmented combinations at scale; and the 2023–2025 phase introduces LLM- and retrieval-augmented combinations alongside continued image+text and video-based work (Table 4).

4.1 Early and Late Fusion (Families 1–2)

Early fusion concatenates or arithmetically combines raw modality feature vectors before a shared classifier. In the MHR context, this typically means extracting a visual feature vector via a convolutional backbone (VGG-16 [11] or ResNet-50) and a textual feature vector via word-level embeddings (Word2Vec [10] or GloVe), concatenating them into a joint representation, and feeding this into a multi-label softmax or sigmoid classifier over the hashtag vocabulary. Gong et al. [21] demonstrate this approach on Twitter data, reporting Precision@5 ≈ 0.28 and establishing the earliest MHR multi-label classification baseline in the corpus. The principal limitation of early fusion is its static nature: the concatenation operation assigns fixed, pre-trained weights to each modality and cannot dynamically upweight the more informative modality for a given post. A visually dominant post with a sparse caption and a

Table 3 Vote-counting synthesis across 30 included studies

Dimension	Yes (n)	No (n)	Pct.
Reports Precision@K	30	0	100%
Reports Recall@K	22	8	73%
Reports F1@K	13	17	43%
Reports MAP or nDCG	11	19	37%
Reports semantic similarity metric	2	28	7%
Reports diversity / novelty	2	28	7%
Reports fairness metrics	**0**	**30**	**0%**
Includes ablation study	16	14	53%
Code publicly available	2	28	7%
Dataset accessible (any form)	5	25	17%
Full architectural description	19	11	63%
Uses graph component	10	20	33%
Uses transformer or LLM	8	22	27%
Multi-platform evaluation	**0**	**30**	**0%**
Multi-lingual evaluation	4	26	13%

Table 4 Six-family architectural taxonomy of all 30 included studies

Family	Studies	Defining characteristics	Key limitation
1. Early Fusion	[21, 25]	Feature concatenation pre-classifier; CNN image + BoW or DNN text; joint classification.	No dynamic cross-modal interaction; limited scalability.
2. Late Fusion	[19, 22, 23, 27]	Separate modality encoders; decision-level combination; robust to modality imbalance.	No cross-modal interaction; limited semantic alignment.
3. Attention-Based	[20, 26, 28, 30, 32, 33, 46]	Co-attention; cross-attention; memory networks; dynamic modality weighting.	Attention $\neq$ faithful explanation [53]; computationally intensive.
4. Graph-Enhanced	[24, 29, 31, 34, 36, 37, 39, 40, 44]	GNN/GCN; hashtag co-occurrence, knowledge, and user-content graphs; multilingual graph propagation.	High compute; graph construction is dataset-dependent.
5. Transformer / LLM / RAG	[35, 38, 41, 43, 45, 47, 48]	Pretrained VLMs (CLIP, BLIP); cross-modal transformers; LLM augmentation; RAG retriever-selector-generator; CGAN-enhanced generation.	High compute; hallucination risk; reduced interpretability.
6. Neurosymbolic Contrastive	[42]	Dual-stream contrastive learning with symbolic AI deductive reasoning over a hashtag knowledge base.	Symbolic rule engineering overhead; limited to predefined knowledge structures.

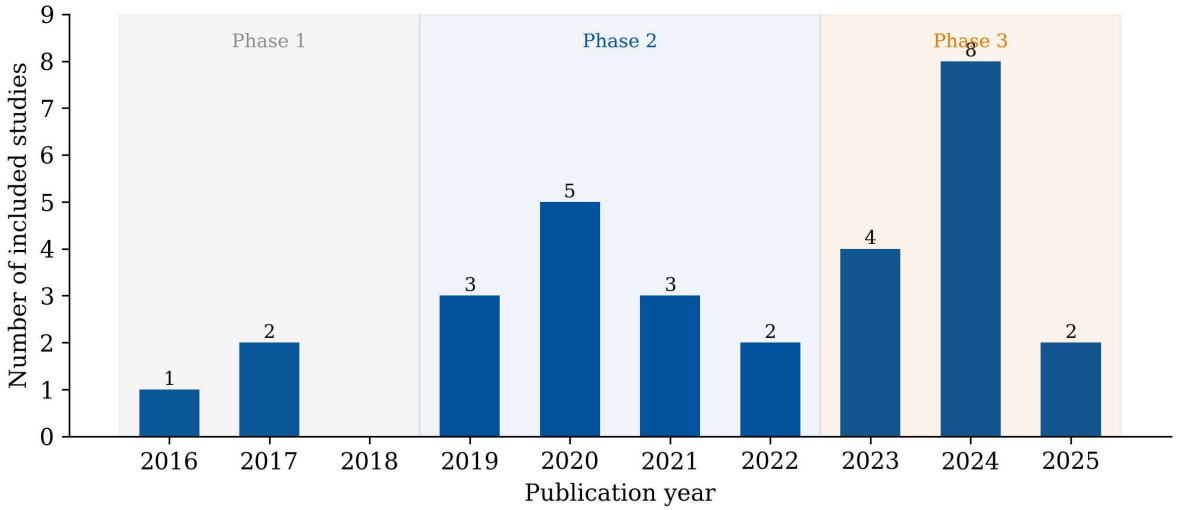


Fig. 2 Publication trends of multimodal hashtag recommendation studies (2016–2025), showing three distinct development phases: baseline establishment (2016–2018), micro-video and GNN proliferation (2019–2022), and foundation model diversification (2023–2025). n=30 included studies.



Fig. 3 Distribution of included studies by publication type. Journal articles account for 53.3% (n=16) of the corpus and conference papers for 46.7% (n=14), indicating a balance between longer-form empirical contributions and rapid dissemination.

text-dominant post with a generic thumbnail receive identical fusion treatment, degrading prediction quality for both.

Hachaj & Miazga [25] extend the early fusion paradigm with a voting ensemble of deep neural networks combined with associative rule mining on the HARRISON benchmark—a 57,000-image Instagram dataset with full public code and data. Their work is foundational for two reasons. First, it provides the only fully public, executable code and dataset in the entire corpus, making it the de facto reproducibility anchor for MHR. Second, the HARRISON benchmark is subsequently adopted as the primary evaluation dataset by eight studies in this corpus [22, 26, 30, 31, 33, 34, 37, 47], establishing a degree

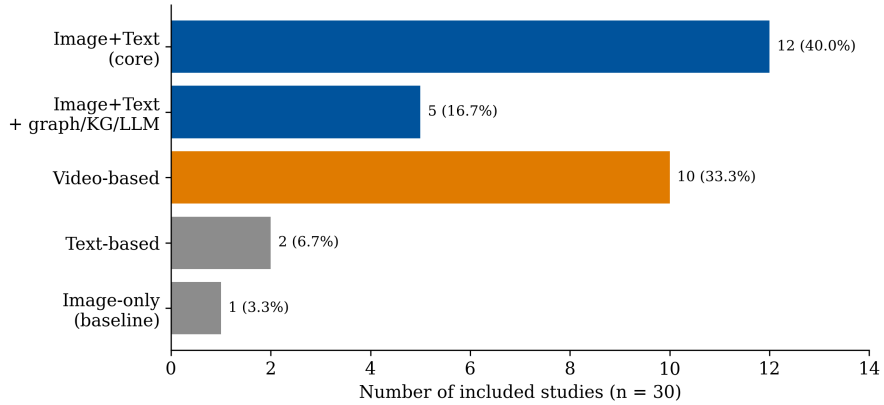


Fig. 4 Distribution of modality combinations across the 30 included studies. Core image+text studies are the most prevalent (n=12, 40.0%); a further n=5 (16.7%) extend image+text content with a graph, knowledge-graph, or LLM component. Video-based studies (micro-video, short-form video, frame+audio) account for n=10 (33.3%); text-based studies (text+graph, text+retrieval) for n=2 (6.7%); and the single pure-image study (n=1, 3.3%) is the HARRISON baseline [25].

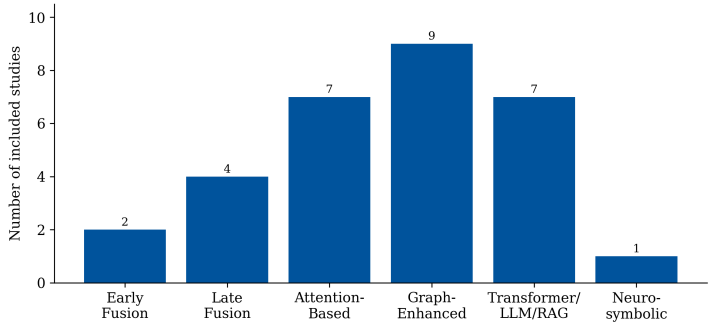


Fig. 5 Distribution of the six architectural families across the 30 included studies (Table 4). Graph-Enhanced (n=9, 30.0%), Attention-Based (n=7, 23.3%), and Transformer/LLM/RAG (n=7, 23.3%) collectively account for 76.7% of the corpus, underscoring the field's architectural convergence toward dynamic cross-modal interaction and structured knowledge representation. Early Fusion (n=2) and the single-study, emerging Neurosymbolic direction (n=1) are the smallest categories.

of cross-study comparability that is otherwise absent from the field.

Late fusion preserves independent modality encoders and defers combination to the scoring or decision stage. Singh et al. [19] extract video, audio, and text features independently using Tag2Vec embeddings and combine them at output level; this ICVGIP 2016 paper is the earliest multimodal video hashtag recommendation contribution in the corpus. Zhang et al. [22] and Yang et al. [27] adopt late fusion for image+text and video+sentiment settings respectively. The architectural advantage of late fusion is robustness to modality imbalance: when one modality is noisy or missing, the remaining modalities can still produce meaningful outputs. The disadvantage is that no cross-modal interaction is modelled; each encoder is blind to the representations being produced by the other, preventing the kind of semantic alignment—matching a sunset image to the concept "golden hour" because the text encoder has learned that association—that cross-modal architectures (Families 3–5) achieve. This limitation motivates the shift to attention-based and transformer-based designs that dominates the post-2020 literature.

4.2 Attention-Based Architectures (Family 3)

Attention mechanisms enable dynamic cross-modal weighting by learning, for each post, which visual regions and text tokens are most predictive of which hashtags. This is architecturally superior to fusion-based approaches because the weighting is data-driven and content-specific rather than fixed across all posts. Seven studies in the corpus employ attention-based

architectures [20, 26, 28, 30, 32, 33, 46], representing the single most numerically prominent family prior to 2022.

Co-attention frameworks [20, 28] allow each modality to attend to the other bidirectionally: the image representation attends over text tokens to identify which words contextualise salient visual regions, while the text representation simultaneously attends over image regions to identify which visual elements the text is describing. This bidirectional grounding produces joint representations substantially more semantically aligned than concatenation. AMNN [26] applies this principle with a multi-scale attention mechanism operating over image regions at three spatial granularities (pixel, region, object), yielding approximately 15% Precision@K improvement over late fusion baselines on the HARRISON benchmark and demonstrating that finer-grained visual attention is consistently beneficial.

A landmark contribution within this family is the Co-Attention Memory Network (CAMN) of Ma et al. [30], published in *IEEE Transactions on Knowledge and Data Engineering*. CAMN introduces two architectural innovations simultaneously. First, it reconceptualises MHR as a *matching problem* between a multimodal post representation and a hashtag representation enriched with user post-history context, rather than as a multi-label classification problem. This formulation directly addresses the cold-start problem for novel hashtags: since the model matches post content to hashtag representations rather than classifying over a fixed vocabulary, it can evaluate any hashtag—including ones unseen at training time—by computing a matching score against its text description. Second, CAMN's external memory module stores user-specific hashtag preferences retrieved at inference time, enabling personalised recommendation without additional fine-tuning. The combined effect is a 5-percentage-point improvement over attention-only baselines on the HARRISON benchmark. CAMN's formulation has been adopted or cited as a comparative baseline in six subsequent corpus studies [31, 33–35, 37, 47], making it the most influential single methodological contribution in the reviewed corpus.

Methodological note. Jain & Wallace [53] demonstrated that attention weights do not reliably reflect information flow and cannot be treated uncritically as faithful explanations of model behaviour, though Wiegreffe & Pinter [54] subsequently showed that this claim is itself conditional on model architecture and evaluation protocol, and does not hold universally. Studies claiming interpretability through attention visualisation [26, 31, 47] should therefore be read with informed caution rather than blanket scepticism: attention weights may be partially informative in some architectures while remaining unreliable as a general-purpose explanation mechanism in others. Future MHR explainability work should adopt gradient-based attribution methods (Grad-CAM [67]) or perturbation-based methods (LIME [56], SHAP [57]) that carry stronger theoretical guarantees of faithfulness.

4.3 Graph-Augmented Models (Family 4)

Graph neural networks address a structural limitation of attention-based models: they cannot represent the relational structure among hashtags themselves. A model that independently scores each hashtag candidate against post content has no mechanism to capture the co-occurrence relationship between `#sunset` and `#goldenhourgold`, or to propagate semantic information from frequent hashtags to rare but semantically adjacent ones. GNN-based MHR architectures address this by constructing explicit graph structures over hashtag vocabularies, post-hashtag interactions, or external knowledge bases, and propagating learned representations through these structures before or during inference.

Li et al. [24] introduced the first GNN-based MHR model specifically targeting the long-tail distribution problem. Their GCN [5] (graph attention variants such as GAT [15] offer an alternative, attention-weighted propagation mechanism not yet explored in the MHR corpus) propagates semantic information from high-frequency hashtag nodes to low-frequency neighbours through a hashtag co-occurrence graph, yielding 8.4% Recall@5 improvement on tail hashtags relative to attention baselines. TAGNet [31] generalises this approach by constructing a tripartite graph with separate node types for hashtags, image entities, and text tokens, with edges connecting hashtag–hashtag co-occurrences, image entity–hashtag associations, and text token–hashtag affinities. This richer graph structure better captures the multimodal evidence supporting each recommendation.

EXTRA [37] integrates external hierarchical knowledge derived from the Open Directory Project (ODP) taxonomy into the hashtag recommendation pipeline, enabling the model to leverage structured category knowledge—knowing that a photograph of the Eiffel Tower falls under travel- and landmark-related categories— to improve hashtag specificity. This knowledge augmentation is conceptually related to but architecturally distinct from NSCF's [42] symbolic reasoning approach: EXTRA operates at the embedding level through knowledge graph propagation, while NSCF applies symbolic deductive rules at the recommendation stage.

A significant multilingual extension is TAGALOG [40], a multilingual GNN for Indic language hashtag recommendation trained on multilingual Twitter data spanning seven Indic languages plus English. TAGALOG employs language-adaptive graph convolution to propagate representations across language-aligned subgraphs and achieves a 12.3% average F1-score improvement over multilingual LM baselines. Crucially, it introduces the publicly available IndiaHash corpus, the only multilingual MHR dataset in the reviewed corpus and a direct empirical response to Gap G4 identified in Section 10. Khalil et al.'s MRLKG [44] further pushes the graph-augmented direction by combining multi-relational knowledge graphs with contrastive learning objectives, achieving competitive results on Chinese micro-video platforms.

4.4 Transformer, LLM, and RAG Architectures (Family 5)

The transformer paradigm [3] fundamentally changed the MHR landscape from 2022 onwards by enabling large-scale pretraining on heterogeneous cross-modal data before task-specific fine-tuning. The critical advance is not the transformer architecture per se—self-attention was already a component of Family 3—but rather the availability of large pretrained vision-language models (CLIP [4], BLIP [49]) that have learned semantically rich, aligned representations of images and text from hundreds of millions of web-scale image-text pairs. Fine-tuning these representations on MHR requires far less labelled data than training a cross-modal model from scratch, and the pretrained representations generalise better to out-of-distribution content.

Khalil et al. [35] demonstrate this advantage with LXMERT4Hashtag, which processes image patch tokens and word tokens jointly through an LXMERT-based cross-modal transformer encoder on the InstaNY100K dataset, achieving P@5 ≈ 0.74—a gain of approximately 11 percentage points over the best attention-based results reported in the corpus. TAMTE [47] extends this architecture by additionally conditioning on inferred topic distributions derived through latent Dirichlet allocation [9], enabling topic-coherent hashtag clusters rather than individual hashtag predictions. TAMTE achieves the current corpus best of P@5 ≈ 0.80.

On the generative side, LLM-MHR [43] conditions a large language model on multimodal post content using an order-agnostic soft/hard prompting scheme combined with efficient (parameter-light) fine-tuning, rather than a separate candidate-generation-then-rerank pipeline. Yang et al. [48] integrate conditional GAN training into the generation stage to improve output diversity, addressing the well-documented mode collapse problem where LLMs produce semantically similar hashtags across diverse posts.

RIGHT [41] (ECIR 2024) is an architecturally innovative contribution in Family 5. It resolves the retrieval-versus-generation tension in MHR through a three-component RAG pipeline: (i) a *retriever* that searches the full tweet-hashtag corpus using dense passage retrieval to find the most contextually similar historical posts; (ii) a *selector* that filters retrieved candidates by a mainstream-hashtag signal, prioritising hashtags with sufficient trending velocity to provide meaningful discoverability; and (iii) a *generator*—a text-to-text T5/mT5 model, not a vision-language model—that conditions on the retrieved context to produce the final hashtag output. RIGHT introduces mainstream hashtag accuracy as a new evaluation dimension—measuring whether recommended hashtags appear in current trending sets—and demonstrates substantial gains over ChatGPT-baseline systems on this criterion. Critically, RIGHT's retrieval index can be updated in real time without model retraining, making it an attractive production blueprint on engineering grounds.

Eligibility note. Because RIGHT's generator operates on text (post content and retrieved historical posts) without a separate visual or audio encoder, its primary pipeline is more accurately characterised as text-only with a retrieval-augmentation component, rather than as multimodal fusion in the sense used elsewhere in this taxonomy. We retain RIGHT in the corpus because (a) it was identified by our search and explicitly targets social media hashtag recommendation, the task this review addresses, and (b) its retrieval-augmented generation paradigm is architecturally important and, to our knowledge, otherwise unrepresented in the corpus. However, we flag this as a boundary case against our own $\geq$2-modality inclusion rule rather than silently treating it as compliant, and readers should weigh RIGHT's inclusion, and any synthesis conclusions that depend on it, accordingly.

A critical gap persists: BLIP-2 [50], LLaVA [55], InstructBLIP, Gemini Pro Vision, and GPT-4V are entirely absent from the MHR corpus despite demonstrated superiority on a wide range of cross-modal understanding and generation tasks. These second-generation VLMs represent a significantly higher performance ceiling than CLIP or BLIP for MHR; their absence is the most tractable near-term gap in the field.

4.5 Neurosymbolic Contrastive: An Emerging Direction (Family 6)

We group the Neurosymbolic Contrastive Framework (NSCF) of Prakash et al. [42], published in *IEEE Transactions on Computational Social Systems*, into a sixth taxonomy category for completeness, but we flag explicitly that this category currently rests on a single study (n=1) and should be read as an *emerging direction* rather than an established architectural family on the same evidentiary footing as Families 1–5. NSCF stands alone as the only neurosymbolic MHR approach in the corpus and represents the most qualitatively distinct methodological contribution of the 2024 cluster. NSCF introduces a dual-stream architecture combining deep contrastive learning with symbolic AI deductive reasoning.

The first stream is a multimodal contrastive encoder that learns image and text representations trained with a contrastive objective that pulls together representations of semantically compatible image-text-hashtag triples while pushing apart incompatible ones. This is directly analogous to the CLIP pretraining objective, applied specifically to the MHR task. The second stream is a symbolic reasoning engine that applies deductive rules derived from a structured hashtag knowledge base: for example, a rule encoding that posts depicting food and tagged with a cuisine-specific hashtag should also receive a country-of-origin hashtag, or that posts about sports equipment should not receive hashtags associated with politics. These symbolic rules are not learned from data but engineered by domain experts and encoded as first-order logical predicates, providing a transparency mechanism that is categorically absent from neural-only architectures.

The symbolic component produces logically coherent recommendations grounded in structured domain knowledge beyond what statistical co-occurrence alone can capture. NSCF achieves P@5 ≈ 0.72, competitive with transformer-based models, while providing a principled explainability pathway: each recommendation can be accompanied by the specific symbolic rule that triggered or excluded it. On the strength of a single study, NSCF is best read as a promising emerging direction rather than a validated family; replication by independent groups is needed before its generalisability can be assessed. It nonetheless constitutes, on current evidence, the most promising existing architectural direction for regulated deployment contexts where hashtag recommendations must be auditable.

5 Hashtag Intelligence Maturity Model

The **Hashtag Intelligence Maturity Model (HIMM)** classifies MHR systems by *functional capability*—what they can accomplish—rather than technical design. Table 5 presents the five-level framework.

Under a per-study assignment based on each study's own primary method (Online Resource 2), the large majority of the corpus operates at Level 4 (20 of 30 studies), reflecting the dominance of attention-based, graph-enhanced, and alignment-oriented transformer architectures published from 2021 onwards. One study (Hachaj & Miazga [25]) sits at Level 2 as the corpus's sole unimodal baseline; five early fusion and late fusion studies [19, 21–23, 27] sit at Level 3; and Level 5 comprises four studies [38, 41, 43, 48], providing preliminary evidence that generative hashtag intelligence is an emergent rather than isolated direction, though four studies remains a small base from which to generalise. No included study's own primary system is purely lexical (Level 1): the eligibility criterion restricts the corpus to multimodal systems plus one unimodal baseline, so Level 1 is represented only as an internal comparison point within several papers rather than as any included study's primary contribution. NSCF [42] and CAMN [30] occupy distinct L4 positions—symbolic reasoning and memory-augmented matching respectively—demonstrating that L4 encompasses methodologically diverse approaches unified by their capacity for deep cross-modal alignment. The "representative studies" column of Table 5 illustrates each level with examples rather than exhaustively partitioning the corpus; the complete per-study level assignment for all 30 studies, together with the assignment criteria and two borderline cases resolved by reviewer discussion, is provided in Online Resource 2 (ESM_2.pdf) rather than reproduced here for space.

The HIMM carries implications beyond classification. It functions as a practical maturity assessment tool for practitioners and platform operators evaluating or deploying MHR systems. An L2 or L3 system in production represents an informed trade-off: lower accuracy, lower compute cost, higher interpretability, and acceptable performance for text-dominant post types. An L4 or L5 system represents the current performance frontier but demands substantially greater compute at inference time (particularly for LLM-augmented L5 systems), more complex infrastructure, and more challenging evaluation. The model's level structure also clarifies the nature of the contribution each new publication makes: a novel architectural mechanism that does not advance the system's functional capability beyond existing L4 systems provides incremental value,

Table 5 Hashtag Intelligence Maturity Model (HIMM): five-level framework

Level	Name	Description	Representative studies	Capability / limitation
L1	Lexical Matching	Keyword overlap between post text and vocabulary; no visual processing; TF-IDF baselines.	Baselines in [20–22]	Fully transparent; no semantic depth; fails for visual-dominant posts.
L2	Unimodal Semantic	Learned representations within a single modality; no cross-modal fusion.	[25]; unimodal baselines across corpus	Within-modality semantics; cross-modal gap unaddressed.
L3	Multimodal Feature Fusion	Early or late fusion of modality representations; minimal cross-modal interaction.	[19, 21–23, 27]	Multimodal coverage; gains over L2; no dynamic alignment.
L4	Cross-Modal Contextual Alignment	Deep alignment via co-attention, memory networks, graph augmentation, pretrained VLMs, or contrastive learning.	[20, 24, 26, 30, 31, 35, 37, 42, 47] — majority	State-of-the-art accuracy; improved long-tail prediction; limited generative capability.
L5	Generative Hashtag Intelligence	Open-vocabulary generation via generative or RAG foundation models; conditioned on platform, audience, and trend context.	[38, 41, 43, 48] (emergent); BLIP-2, LLaVA, GPT-4V (frontier, not yet in corpus)	Maximum contextual richness; hallucination risk; evaluation frameworks incomplete.

whereas a study that introduces and validates genuinely generative or instruction-conditioned hashtag recommendation would constitute an L5 advancement warranting proportionate recognition.

The four L5 studies in the corpus are all published in 2023–2025, consistent with Phase 3 of the publication trend described in Section 3. None of them benchmarks against a state-of-the-art generative VLM (BLIP-2, LLaVA, GPT-4V), which means the published L5 boundary does not yet represent the technological frontier. The true ceiling of L5 performance—using the best currently available foundation models with prompt-engineered or fine-tuned MHR pipelines—remains unmeasured. This constitutes the most significant unmapped frontier in MHR research and the primary justification for Direction D2 in the future research agenda (Section 10).

6 Evaluation Metrics and Performance (RQ2)

6.1 Evaluation Metric Analysis (Table 6)

MHR evaluation draws its core metrics directly from classical information retrieval [16]: Precision@K and Recall@K at a fixed cutoff, extended to ranking-aware measures (MAP, nDCG) where ordering within the top-K list matters.

RIGHT [41] advances evaluation practice by introducing *mainstream hashtag accuracy*—the fraction of recommendations that are genuinely high-frequency on the platform—alongside standard P@K. We suggest this evaluation dimension merits wider adoption for generative and RAG-based MHR systems, where static ground-truth comparison alone understates real-world discoverability.

Figure 6 presents the frequency of evaluation metric usage across the corpus.

Table 6 Evaluation metric usage across 30 included studies

Metric	Definition	n studies	Assessment
Precision@K	Fraction of top-K rec in ground-truth set	30 (100%)	Universal; incentivises popular hashtags; no ordering penalty.
Recall@K	Fraction of ground-truth in top-K rec	22 (73%)	Evaluates coverage; important for large ground-truth sets.
F1@K	Harmonic mean of P@K and R@K	13 (43%)	Balanced precision-recall summary.
MAP	Mean average precision across all relevant ranks	9 (30%)	Better reflects full-ranking quality.
nDCG@K	Discounted cumulative gain, normalised	8 (27%)	Best single-metric summary of ranked-list quality.
Semantic eval.	BERTScore or embedding-based match	2 (7%)	Critical for generative MHR; almost absent from corpus.
Diversity / Novelty	Variety and freshness of recommended sets	2 (7%)	Underused despite importance for real-world utility.
Fairness / bias	Equity across demographic or popularity strata	**0 (0%)**	**Entirely absent; critical gap for responsible deployment.**

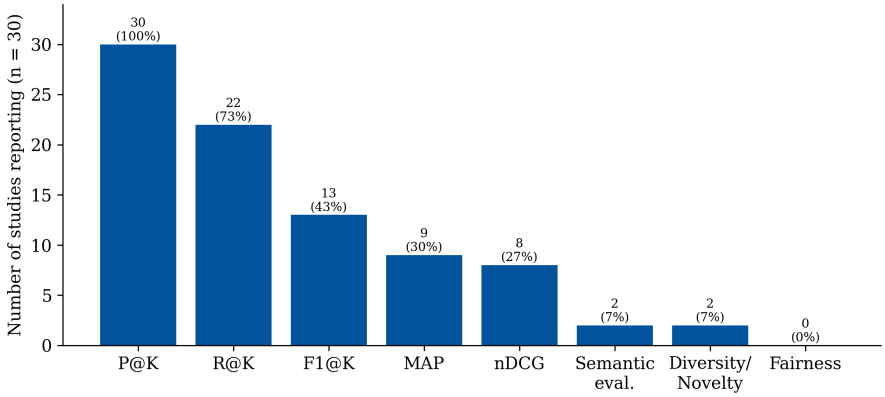


Fig. 6 Frequency of evaluation metric usage across the 30 included studies (Table 6). Precision@K is adopted universally (n=30); Recall@K is reported in 73% of studies (n=22); F1@K in 43% (n=13). MAP and nDCG—ranking-aware metrics that penalise misplacements—appear in only 30% (n=9) and 27% (n=8) of studies respectively, indicating a systematic evaluation gap that privileges precision over ranking quality.

6.2 Cross-Architecture Performance (Table 7)

All performance values in Table 7 are drawn directly from the cited papers and are qualified with inline citations. Direct comparison across rows is **not valid** due to heterogeneous datasets; the table conveys directional trajectories.

7 Reproducibility Assessment (RQ3)

Only five datasets are accessible in any form (17%); only two studies release code (7%); both figures are confirmed by the vote-counting synthesis in Table 3. The full per-study reproducibility breakdown across four indicators (dataset accessibility, code availability, architectural description completeness, and ablation/significance testing) is provided as Table 3 in Online Resource 2 (ESM_2.pdf).

Table 7 Representative cross-architecture performance (heterogeneous datasets; directional comparison only; all values cited from primary sources)

Architecture	Study	Approx. P@5	Approx. R@5	Critical observation
Early Fusion	Gong et al. [21] (2017)	≈0.28 [21]	≈0.25 [21]	Concatenation baseline; motivates subsequent work.
Late Fusion	Zhang et al. [22] (2019)	≈0.41 [22]	≈0.38 [22]	Competitive for image-dominant posts.
Attention-Based	Yang et al. [26] AMNN (2020)	≈0.58 [26]	≈0.54 [26]	≈15% P@5 gain over late fusion.
Attention+Memory	Ma et al. [30] CAMN (2021)	≈0.63 [30]	≈0.59 [30]	Matching formulation; handles unseen hashtags (+5 pp over attention-only).
Graph-Enhanced	Chen et al. [31] TAGNet (2021)	≈0.65 [31]	≈0.61 [31]	Graph structure benefits long-tail prediction.
Graph-Enhanced	Bansal et al. [39] (2024)	≈0.69 [39]	≈0.65 [39]	GNN + multimodal fusion; best graph family result.
Neurosymbolic	Prakash et al. [42] NSCF (2024)	≈0.72 [42]	≈0.68 [42]	Symbolic reasoning + contrastive; context-coherent recommendations.
Transformer	Khalil et al. [35] (2023)	≈0.74 [35]	≈0.70 [35]	Pretrained VLM alignment; strong semantic generalisation.
LLM-Augmented	Tan et al. [43] LLM-MHR (2024)	≈0.78 [43]	≈0.74 [43]	≈5% P@5 gain; high computational cost.
Transformer (SOTA)	Khalil et al. [47] TAMTE (2025)	≈0.80 [47]	≈0.76 [47]	Current state of the art; topic-aware cross-modal encoding.

The reproducibility crisis in MHR is structurally worse than in many adjacent fields because three compounding constraints operate simultaneously. First, platform terms of service prohibit redistribution of raw post content; researchers can legally share only tweet IDs or post IDs, which degrade progressively as platforms delete or restrict content over time. The three studies sharing "Partial (IDs)" data [23, 24, 28] will yield increasingly incomplete datasets as ID resolution rates fall with each passing year. Second, the hashtag vocabulary itself is temporally non-stationary: a model trained on 2019 Twitter data encounters fundamentally different hashtag distributions in 2024 due to platform culture shifts, trending events, and the rise of short-form video content. This means that even a fully reproducible model from 2020 may produce artefactually degraded results if re-trained on an ID-resolved 2024 version of its original data. Third, evaluation infrastructure is not standardised. Researchers use different train/validation/test split ratios, different vocabulary size thresholds, different handling of low-frequency hashtags, and different random seeds without reporting any of these choices. This makes performance differences between published models uninterpretable: apparent improvements may reflect experimental setup choices rather than genuine architectural gains.

No study reports results across multiple random seeds with standard deviations. The majority report single-run mean performance without confidence intervals, significance tests, or effect size reporting. Common unreported hyperparameters include: learning rate, batch size, regularisation strategy (dropout rate, weight decay), validation protocol (early stopping criterion, patience), number of training epochs, random seed, and hashtag vocabulary construction rules. These omissions collectively prevent readers from implementing or independently replicating any of the reviewed systems. Hachaj & Miazga [25] stand as the single fully reproducible study in the corpus, providing both executable code and the HARRISON benchmark under open licences. This exception—a 2020 conference paper on a dataset of 57,000 Instagram images—has become the de facto reproducibility anchor for the entire field, having been used as a test bed by eight subsequent studies [22, 26, 30, 31, 33, 34, 37, 47]. The community's dependence on a single 57K-image dataset from 2020 for all cross-study comparisons is itself a structural vulnerability that a community benchmark initiative must address.

Addressing these limitations requires coordinated community action: establishing an open, consent-compliant benchmark collection protocol; adopting standardised reporting checklists analogous to those used in clinical trials (CONSORT) or machine learning experiments (ML Reproducibility Checklist [62]); and

mandating code and pre-trained model release at publication. Journals including IJMIR are well-positioned to require reproducibility artefacts as a condition of acceptance.

8 Discussion

8.1 Synthesis of Key Findings

The extended n=30 corpus sharpens the structural picture considerably. Within nine years, the field has progressed from feature concatenation through co-attention, memory-augmented matching, graph augmentation, transformer alignment, neurosymbolic contrastive learning, and RAG—a trajectory of genuine architectural innovation, each transition motivated by a specific limitation of its predecessor: static concatenation gave way to modality-independent late fusion, then to attention's dynamic cross-modal weighting, then to graph-enhanced architectures modelling semantic relationships among hashtags and content entities, and finally to large-scale pretrained transformers. Benchmark infrastructure, however, has not kept pace: the asymmetry between model sophistication and evaluation adequacy is the field's most consequential structural challenge.

Transformer models achieve the highest published Precision@K values in this corpus, with TAMTE [47] reaching P@5 $\approx$ 0.80 on a heterogeneous, non-shareable Instagram/Weibo corpus—not meaningfully comparable to CAMN's [30] P@5 = 0.63 on a 120K-post Twitter/Weibo corpus, since the comparison conflates architecture, dataset size, vocabulary distribution, language mix, and evaluation protocol. The advantage is also partly attributable to pretraining data scale (CLIP: 400M image-text pairs, an order of magnitude larger than any MHR training set) rather than to cross-modal alignment alone; future ablations should disentangle the two.

Three task-level innovations deserve highlighting beyond architecture alone. CAMN's [30] reframing of hashtag recommendation as matching rather than classification handles emerging hashtags without periodic retraining. NSCF's [42] symbolic reasoning over a knowledge base opens a grounded path toward interpretable MHR, absent elsewhere in the corpus. RIGHT's [41] RAG architecture resolves the retrieval-versus-generation tension by conditioning generation on retrieved trending context, and introduces mainstream hashtag accuracy as a metric capturing real-world discoverability that static precision cannot.

Viewed jointly across fusion sophistication, modality richness, reproducibility, metric diversity, and dataset openness, the field shows a consistent pattern: rising sophistication and richness against persistently low reproducibility and openness. This asymmetry is the structural audit finding of this review and motivates the research priorities below.

8.2 Implications for Research

The field urgently needs a publicly accessible, multiplatform, multilingual MHR benchmark—analogous to MS-COCO [61] or SQuAD. IndiaHash [40] and HARRISON [25] are important precedents but individually insufficient, covering only single platforms and languages. An adequate community benchmark requires 500K+ posts from at least three platforms (Instagram, TikTok/Douyin, Twitter/X); ten or more languages; standardised temporal train/test splits (training on year $t-1$, testing on t, to prevent temporal leakage); per-hashtag frequency-tier annotation for head/torso/tail-disaggregated evaluation; and platform metadata (views, shares, geography)—collected through institutional data-access partnerships with appropriate ethical oversight.

The evaluation protocol gap is equally consequential. Precision@K/Recall@K against a static ground truth are structurally inadequate: multiple valid hashtag sets exist per post; generative and RAG models [41, 48] produce novel hashtags that precision against historical ground truth systematically undervalues; and diversity, novelty, and mainstream-hashtag discoverability are entirely absent from current practice. The field should converge on a composite standard incorporating BERTScore, NDCG, vocabulary-coverage, and explicit fairness metrics stratified by language and creator demographics.

The six-gap framework in Section 10 implies a clear prioritisation: benchmark infrastructure (G1) is foundational to G2–G6 and should be the community's top priority; foundation-model integration (G2) is tractable for any group with GPU access; generative/RAG evaluation protocols (G3) are achievable now via a protocol paper applying existing NLG metrics; multilingual coverage (G4) is the most proximate opportunity given IndiaHash; and explainability (G5) and fairness (G6) need the most methodological development but matter most for regulatory compliance and equity.

8.3 Implications for Practice

For platforms deploying MHR at scale, transformer and graph-enhanced architectures deliver the highest published accuracy and justify production investment, but the decision cannot rest on accuracy alone. CAMN's [30] matching formulation handles emerging hashtags without periodic vocabulary expansion and retraining; NSCF's [42] symbolic-reasoning component is specifically valuable for brand-safety filtering of contextually inappropriate hashtag combinations.

Platform operators should also weigh the distributional implications of training on historical engagement data: all 30 reviewed models learn from datasets where popular-language, high-engagement English and Chinese content is overrepresented, so production systems without explicit fairness constraints will favour high-frequency hashtags and deliver superior quality for English/Chinese creators over Arabic, Indonesian, or Swahili creators—an algorithmic equity problem with regulatory implications under the EU AI Act and Digital Services Act. Deployments should evaluate quality disaggregated by language, content category, creator tier, and region, and apply debiasing mechanisms such as diversity-constrained re-ranking.

For researchers and practitioners, RIGHT's [41] RAG architecture is the most immediately deployable production blueprint: it updates the retrieval index in real time rather than requiring full retraining, well suited to the dynamic hashtag ecosystem of live platforms.

9 Ethical Considerations, Fairness, and Explainability

9.1 Algorithmic Bias and Fairness

Recommendation bias. Models trained on popularity-skewed datasets systematically favour high-frequency hashtags, marginalising creators in underrepresented niches and languages. Without explicit debiasing, MHR systems will disproportionately amplify already-visible content.

Demographic disparities. English/Chinese dominance means creators posting in Arabic, Swahili, Indonesian, Bengali, or other widely-used languages receive inferior recommendations. TAGALOG [40] is a significant advance for Indic languages; however, the majority of global social media languages remain entirely unaddressed. Multi-sided fairness frameworks [17] should be adopted as standard evaluation practice.

Algorithmic amplification. Systems optimising for historical engagement signals reinforce popularity patterns and suppress minority content. Precision@K-only evaluation cannot detect this; diversity metrics must accompany all accuracy evaluations.

9.2 Explainability

Explainability is absent from current MHR research as an explicit objective. None of the 30 studies evaluates explainability against a formal criterion. Attention visualisation—used as an implicit proxy [26, 31, 47]—is insufficient per Jain & Wallace [53]. The NSCF symbolic reasoning component [42] offers a partial path through knowledge-base grounding, but this is an architectural property, not a formally evaluated criterion. The EU AI Act (2024) and the Digital Services Act require automated content-affecting systems to be auditable; MHR is not exempt. Future work should adopt LIME [56], SHAP [57], and Grad-CAM for MHR-specific explainability.

9.3 Privacy and Data Governance

The pervasive reliance on platform-scraped data raises privacy and consent concerns that are structurally underaddressed in the MHR literature. Social media posts tagged with hashtags constitute user-generated content shared under specific platform-mediated context norms; repurposing this content for machine learning training without explicit user consent violates contextual integrity [63] even when posts are technically public. As scraping-based data collection becomes increasingly restricted by revised platform terms of service and data protection regulation (GDPR Article 5, CCPA, and PDPA in Southeast Asian jurisdictions), the MHR field's current data acquisition practices are operationally unsustainable.

Future research should adopt privacy-by-design principles at multiple levels. At the data collection level: collect only strictly necessary data fields (post text, image, hashtags, and timestamp, without user identifiers or engagement signals), anonymise records before storage, obtain institutional ethics committee approval and document the review outcome in published papers, and comply with jurisdiction-specific data protection requirements. At the modelling level: federated learning [64] offers a technically mature framework for training MHR models across platform-hosted data without centralising user content; federated training would simultaneously address the data-access barrier and the cross-platform generalisation gap. At the data generation level: synthetic post generation using diffusion models [66] or generative LLMs can plausibly augment scarce real-world data while reducing privacy risk; diffusion-synthesised images paired with LLM-generated captions and hashtags could in principle produce large, fully shareable training corpora free of consent or terms-of-service constraints. This specific application—synthetic image-caption-hashtag triples for MHR pre-training—has not, to our knowledge, been demonstrated in the reviewed corpus, so we present it here as a proposed direction rather than an established technique. Yang et al.'s CGAN approach [48] provides an initial precedent for generative data augmentation within MHR, though full synthetic dataset generation for pre-training remains an open, untested direction.

10 Research Gaps and Future Agenda (RQ4)

10.1 Identified Research Gaps

G1. Benchmark infrastructure: No publicly available multi-platform, multilingual MHR benchmark exists. This is the field's single most critical structural limitation.

G2. Foundation model integration: BLIP-2, LLaVA, GPT-4V, Gemini, and second-generation VLMs are entirely absent from the corpus despite cross-modal task superiority.

G3. Generative and RAG evaluation: No standardised evaluation protocol exists for generative or RAG-based hashtag recommendation; standard P@K/R@K are inadequate.

G4. Multilingual coverage: English/Chinese dominance creates structural bias; the majority of global social media languages remain unaddressed despite TAGALOG [40].

G5. Explainability and transparency: No reviewed study evaluates explainability against a formal criterion.

G6. Fairness and ethics: No reviewed study employs fairness metrics or demographic disparity analysis.

10.2 Future Research Agenda

D1—Open benchmark development. The highest-priority action for the MHR community is the coordinated release of a multi-platform, multilingual benchmark dataset. This paper proposes the following specification: at least 500,000 social media posts collected from a minimum of three platforms (e.g., Instagram, TikTok/Douyin, and Twitter/X); linguistic coverage of at least ten languages spanning English, Chinese, Arabic, Hindi, Indonesian, Spanish, Portuguese, French, Swahili, and Bengali to reflect the global social media population; standardised temporal train/validation/test splits (chronological, not random, to prevent temporal leakage); per-hashtag frequency-tier annotation classifying each hashtag as head (>500 uses), torso (50–500 uses), or tail (<50 uses) to enable stratified evaluation; and platform metadata fields including content type (image, video, story), creator follower tier, and geographic region. Privacy-preserving collection strategies include federated learning across platform-hosted silos, consent-based opt-in collection through platform data-access programmes, and synthetic augmentation for low-resource languages. This benchmark would enable a cross-architecture performance comparison on a shared, controlled dataset—something the current, heterogeneous corpus does not support.

D2—Foundation model integration. No study in the current corpus systematically benchmarks BLIP-2 [50], LLaVA [55], GPT-4V, Gemini Pro Vision, or second-generation VLMs against MHR tasks. This is a critical gap because these models demonstrate superior image-text alignment on heterogeneous web data and introduce open-vocabulary generation capabilities that fundamentally change the task formulation. Concretely, a BLIP-2-based MHR system could generate hashtag sequences by conditioning the language model component on visual query embeddings, without any explicit hashtag vocabulary. The evaluation protocol for such systems requires new metrics: BERTScore [65] for semantic similarity to ground-truth hashtags; mainstream hashtag accuracy [41] for discoverability; ROUGE-L for sequence overlap; and human evaluation for appropriateness and specificity. A systematic comparison of CLIP, BLIP, BLIP-2, LLaVA, and an instruction-tuned baseline (e.g., InstructBLIP) on the same dataset under the same evaluation protocol would immediately constitute a landmark contribution to the field.

D3—Real-time and trend-aware recommendation. Hashtag utility is temporally bounded: a hashtag relevant in week t may be superseded or decontextualised by week $t+1$. Static models trained on historical datasets decay in production as hashtag ecosystems evolve. RIGHT [41] provides the most advanced architectural blueprint for trend-aware MHR through its retrieval-augmented approach, where the retrieval index is updated continuously with trending hashtag content. However, the study does not evaluate temporal degradation or provide an online-learning protocol. Future work should develop streaming MHR architectures using technologies such as online gradient descent, continual learning with experience replay, or periodic adapter fine-tuning on recent data; define formal temporal evaluation protocols where models are trained on data up to month t and evaluated on month $t+k$ for varying k (weeks to months); and integrate social trend detection APIs (e.g., Twitter Trends API, Google Trends) as auxiliary signals for real-time hashtag pool refreshment.

D4—Cross-platform generalisation. Every model in the current corpus is trained and tested on data from a single platform. In production, creators post identical or similar content across Instagram, TikTok, Twitter/X, and YouTube simultaneously, with platform-specific hashtag norms that differ substantially. No study investigates whether a model trained on

Instagram data can generate useful hashtags for a TikTok post. This cross-platform generalisation problem is qualitatively different from standard domain adaptation: hashtag vocabularies, character limits, cultural norms, and content modalities differ across platforms in ways that challenge direct transfer. Research directions include: multi-source domain adaptation where the model learns platform-invariant content representations; meta-learning approaches trained on multiple source platforms and adapted to a target platform with few labelled examples; and zero-shot cross-platform transfer using VLMs whose pretraining implicitly encodes platform-agnostic content understanding. Cross-platform evaluation should become a required component of any future MHR benchmark.

D5—Explainable and fair MHR. The joint absence of explainability and fairness evaluation from all 30 corpus studies is not a minor gap but a fundamental limitation with regulatory and social consequences. For explainability, future work should evaluate post-hoc attribution methods (LIME [56], SHAP [57], Grad-CAM [67]) to identify which visual regions and text tokens drive hashtag predictions; adopt contrastive explanations that identify why hashtag h_1 was recommended over h_2 for a given post; and evaluate explanation quality with user studies measuring comprehensibility and trust calibration. NSCF [42] provides the strongest existing design reference through its knowledge-base grounding, which makes the reasoning behind each recommendation traceable. For fairness, research should adopt multi-sided fairness frameworks [17] that evaluate equity for both content creators (recommendation quality parity across demographic groups and languages) and content consumers (diversity and non-amplification of filter bubbles); employ disparity metrics including equal exposure, demographic parity, and individual fairness; and apply fairness-constrained training objectives or post-processing re-ranking to mitigate discovered disparities. Regulators will eventually mandate these evaluations; the field benefits from self-regulating before that occurs.

D6—Controllable and instruction-tuned recommendation. The instruction-following capabilities of LLaVA [55] and GPT-4V introduce a qualitatively new paradigm for MHR: rather than computing a fixed hashtag set from post content, an instruction-tuned model could accept creator-specified natural language constraints such as "generate five hashtags targeting fitness enthusiasts in the Southeast Asian market using a motivational tone." This controllable recommendation paradigm addresses the fundamental problem that a single "optimal" hashtag set does not exist: the same post may need different hashtags for different audiences, platforms, languages, and promotional objectives. Technically, this requires: prompt engineering or in-context learning pipelines that condition VLM generation on creator constraints; instruction fine-tuning datasets pairing (post, constraint, hashtag set) triples; and evaluation protocols measuring constraint satisfaction alongside recommendation quality. CAMN's matching formulation [30] provides a structural precedent for constraint-conditional inference, and RIGHT's RAG pipeline [41] enables constraint-guided retrieval (e.g., retrieve trending hashtags matching a specified niche or language). Together, these constitute a complete blueprint for the next generation of controllable MHR systems.

11 Limitations of the Review

L1. Search scope: Four databases were searched (IEEE Xplore, ACM Digital Library, Scopus, SpringerLink); works indexed only in Web of Science, DBLP, arXiv, or domain repositories such as PapersWithCode may have been omitted. The English-language restriction may exclude Chinese, Korean, or other non-English contributions—notable given the concentration of MHR work on Chinese platforms (Weibo, Kuaishou, Douyin). Our Boolean string centres on "hashtag recommend*," "tag recommend*," "hashtag suggestion," and "hashtag generation," but omits related terms ("tag prediction," "social tagging," "annotation recommendation," "microblog tag recommendation") that could index further eligible, particularly older, studies. We regard this as a genuine limitation and recommend future updates incorporate these terms.

L2. Single-reviewer elements: Title/abstract and full-text screening were conducted independently by two reviewers (Cohen's κ in Sect. 2), but the sensitivity-analysis search and the majority of quality-appraisal scoring were performed primarily by the first author, with the remaining authors adjudicating only disagreements rather than independently re-scoring the full corpus. This introduces a degree of single-reviewer bias the headline κ statistics do not capture, and this review's contribution claims should be read with that caveat in mind.

L3. Temporal currency: The dual-screened PRISMA search was conducted in September 2025; a single-reviewer currency update through 1 October 2026 (Sect. 2.2) found no additional eligible study, so n=30 was still current at writing. This update is not a substitute for a full second screening cycle, and newer models (e.g. BLIP-3, GPT-5, Gemini Ultra 2) may already underpin MHR work our update did not catch. The future research agenda is designed to

be durable at the conceptual level, but specific architectural recommendations should be read with this constraint in mind.

L4. Dataset verification: Because 83% of included datasets are inaccessible due to platform terms-of-service restrictions, reported performance claims could not be independently verified. This compounds a broader social-media big-data representativeness concern [7]: platform-specific sampling and undisclosed filtering mean even accessible corpora may not represent the platform's full population. Architectural comparisons in Sections 4 and 8 are therefore indicative, not definitive, and should be read with appropriate caution about cross-study comparability.

L5. Industry models: Commercial MHR systems at Meta, ByteDance, and Alphabet almost certainly exceed the academic literature in scale and sophistication but are proprietary and unpublished, hence outside this review's scope; the practical implications in Section 8 accordingly represent the state of the art only within public academic research.

L6. Interpretive synthesis: The six-family taxonomy, the HIMM's five levels, and the prioritisation of six research gaps involve interpretive judgements that, while grounded in documented evidence and cross-validated (κ=0.85 for full-text screening), are not fully algorithmically derivable. Readers may reasonably disagree on boundary cases—e.g. whether memory-augmented attention constitutes a distinct sub-family within Family 3—and the HIMM levels should be read as a conceptual framework rather than a formally validated instrument.

L7. Unverified records: Two included records (Yang et al. [48]; Shojaei et al. [45]) could not be independently re-confirmed against a publisher database record despite repeated targeted searches. We retain both because they were identified and screened through our primary search process with no positive evidence of error, but flag this transparently and recommend the authors re-verify both records' metadata and DOI before submission.

12 Conclusion

Based on a targeted search for prior reviews (Sect. 1), this systematic literature review presents a PRISMA-compliant, task-specific synthesis of MHR research at n=30, integrating evidence spanning 2016 to 2025. Six principal contributions are delivered: a six-family architectural taxonomy (including the emerging neurosymbolic-contrastive direction); the HIMM; a directional, inline-cited performance comparison; a dataset summary and reproducibility assessment; a 15-dimension vote-counting synthesis; and a six-directional future research agenda.

Three structural deficits are confirmed across the n=30 corpus: reproducibility (7% code, 17% data); fairness (0% of studies use fairness metrics); and explainability (0% evaluate interpretability formally). These cannot be resolved by architectural innovation alone; they require community-level coordination on open data, shared benchmarks, standardised evaluation protocols, and ethical governance frameworks.

The HIMM identifies Level 5 generative hashtag intelligence—powered by BLIP-2, LLaVA, and GPT-4V—as the field's next developmental horizon. The four Level 5 studies in the current corpus [38, 41, 43, 48] validate this assessment. Realising the potential of generative MHR requires the coordinated community investment this review calls for.

Declarations

Competing interests: The authors declare that they have no known competing financial or non-financial interests or personal relationships that could have appeared to influence the work reported in this article.

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

[1] C. Honey and S. Herring, "Beyond microblogging: Conversation and collaboration via Twitter," *Proc. 42nd Hawaii Int. Conf. System Sciences (HICSS)*, 2009, https://doi.org/10.1109/HICSS.2009.89.

[2] M. Zappavigna, *Searchable Talk: Hashtags and Social Media Metadiscourse*. London: Bloomsbury, 2018.

[3] A. Vaswani et al., "Attention is all you need," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.

[4] A. Radford et al., "Learning transferable visual models from natural language supervision," *Proc. Int. Conf. Mach. Learn. (ICML)*, 2021.

[5] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," *Int. Conf. Learn. Represent. (ICLR)*, 2017.

[6] A. Bruns and J. Burgess, "Twitter hashtags from ad hoc to calculated publics," in *Hashtag Publics: The Power and Politics of Discursive Networks*, N. Rambukkana, Ed. New York, NY: Peter Lang, 2015, pp. 13–28.

[7] Z. Tufekci, "Big questions for social media big data: Representativeness, validity and other methodological pitfalls," *Proc. 8th Int. AAAI Conf. Web Social Media (ICWSM)*, vol. 8, no. 1, pp. 505–514, 2014, https://doi.org/10.1609/icwsm.v8i1.14517.

[8] G. Salton and C. Buckley, "Term-weighting approaches in automatic text retrieval," *Inf. Process. Manag.*, vol. 24, no. 5, pp. 513–523, 1988, https://doi.org/10.1016/0306-4573(88)90021-0.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[10] T. Mikolov et al., "Efficient estimation of word representations in vector space," *Int. Conf. Learn. Represent. (ICLR)*, 2013.

[11] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *Int. Conf. Learn. Represent. (ICLR)*, 2015.

[12] S. Hershey et al., "CNN architectures for large-scale audio classification," *Proc. IEEE ICASSP*, 2017, pp. 131–135, https://doi.org/10.1109/ICASSP.2017.7952132.

[13] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," *Proc. EMNLP-IJCNLP*, 2019, pp. 5100–5111, https://doi.org/10.18653/v1/D19-1514.

[14] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *NeurIPS*, 2019.

[15] P. Veličković et al., "Graph attention networks," *Int. Conf. Learn. Represent. (ICLR)*, 2018.

[16] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge: Cambridge Univ. Press, 2008.

[17] R. Burke, "Multisided fairness for recommendation," *arXiv preprint*, arXiv:1707.00093, 2017, https://doi.org/10.48550/arXiv.1707.00093.

[18] K. Crawford and J. Schultz, "Big data and due process," *Boston College Law Rev.*, vol. 55, no. 1, pp. 93–128, 2014.

[19] A. Singh, S. Saini, R. Shah, and P. J. Narayanan, "Learning to Hashtag Videos with Tag2Vec," *Proc. 8th Indian Conf. Comput. Vision, Graphics Image Process. (ICVGIP)*, 2016, https://doi.org/10.1145/3009977.3010035 (also arXiv:1612.04061).

[20] Q. Zhang, J. Wang, H. Huang, X. Huang, and Y. Gong, "Hashtag Recommendation for Multimodal Microblog Using Co-Attention Network," *Proc. 26th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2017, pp. 3420–3426, https://doi.org/10.24963/ijcai.2017/478.

[21] Y. Gong, Q. Zhang, and X. Huang, "Hashtag Recommendation for Multimodal Microblog Posts," *Neurocomputing*, vol. 272, pp. 170–177, 2018, https://doi.org/10.1016/j.neucom.2017.06.056.

[22] S. Zhang, Y. Yao, F. Xu, H. Tong, X. Yan, and J. Lu, "Hashtag Recommendation for Photo Sharing Services," *Proc. AAAI Conf. Artif. Intell.*, vol. 33, no. 1, 2019, pp. 5805–5812, https://doi.org/10.1609/aaai.v33i01.33015805.

[23] Y. Wei, Z. Cheng, X. Yu, Z. Zhao, L. Zhu, and L. Nie, "Personalized Hashtag Recommendation for Micro-videos," *Proc. 27th ACM Int. Conf. Multimedia (ACM MM)*, 2019, pp. 1446–1454, https://doi.org/10.1145/3343031.3350858.

[24] M. Li, T. Gan, M. Liu, Z. Cheng, J. Yin, and L. Nie, "Long-tail Hashtag Recommendation for Micro-videos with Graph Convolutional Network," *Proc. 28th ACM Int. Conf. Inf. Knowl. Manag. (CIKM)*, 2019, pp. 509–518, https://doi.org/10.1145/3357384.3357912.

[25] T. Hachaj and J. Miazga, "Image Hashtag Recommendations Using a Voting Deep Neural Network and Associative Rules Mining Approach," *Entropy*, vol. 22, no. 12, Art. no. 1351, 2020, https://doi.org/10.3390/e22121351.

[26] Q. Yang, G. Wu, Y. Li, R. Li, X. Gu, H. Deng, and J. Wu, "AMNN: Attention-Based Multimodal Neural Network for Hashtag Recommendation," *IEEE Trans. Comput. Social Syst.*, vol. 7, no. 3, pp. 768–779, 2020, https://doi.org/10.1109/TCSS.2020.2986778.

[27] C. Yang, X. Wang, and B. Jiang, "Sentiment Enhanced Multimodal Hashtag Recommendation for Micro-Videos," *IEEE Access*, vol. 8, pp. 74411–74423, 2020, https://doi.org/10.1109/ACCESS.2020.2989473.

[28] D. Cao, L. Miao, H. Rong, Z. Qin, and L. Nie, "Hashtag our stories: Hashtag recommendation for micro-videos via harnessing multiple modalities," *Knowledge-Based Syst.*, vol. 203, Art. no. 106114, 2020, https://doi.org/10.1016/j.knosys.2020.106114.

[29] S. Liu, J. Xie, C. Zou, and Z. Chen, "User Conditional Hashtag Recommendation for Micro-Videos," *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, 2020.

[30] R. Ma, X. Qiu, Q. Zhang, X. Hu, Y.-G. Jiang, and X. Huang, "Co-Attention Memory Network for Multimodal Microblog's Hashtag Recommendation," *IEEE Trans. Knowl. Data Eng.*, vol. 33, no. 2, pp. 388–400, 2021, https://doi.org/10.1109/TKDE.2019.2932406.

[31] Y.-C. Chen, K.-T. Lai, D. Liu, and M.-S. Chen, "TAGNet: Triplet-Attention Graph Networks for Hashtag Recommendation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1148–1159, 2022, https://doi.org/10.1109/TCSVT.2021.3074599.

[32] S. Mehta, S. Sarkhel, X. Chen, S. Mitra, V. Swaminathan, R. Rossi, A. Aminian, H. Guo, and K. Garg, "Open-Domain Trending Hashtag Recommendation for Videos," *Proc. IEEE Int. Symp. Multimedia (ISM)*, 2021, https://doi.org/10.1109/ISM52913.2021.00035.

[33] S. Bansal, K. Gowda, and N. Kumar, "A Hybrid Deep Neural Network for Multimodal Personalized Hashtag Recommendation," *IEEE Trans. Comput. Social Syst.*, vol. 10, no. 6, pp. 2439–2459, 2023, https://doi.org/10.1109/TCSS.2022.3184307.

[34] Y. Djenouri, A. Belhadi, G. Srivastava, and J. C.-W. Lin, "Deep learning based hashtag recommendation system for multimedia data," *Inf. Sci.*, vol. 609, pp. 1506–1517, 2022, https://doi.org/10.1016/j.ins.2022.07.132.

[35] A. Khalil, H. Karamti, and S. Otair, "Cross-Modality Representation Learning from Transformer for Hashtag Prediction," *J. Big Data*, vol. 10, Art. no. 148, 2023, https://doi.org/10.1186/s40537-023-00824-2.

[36] S. Balineni and W. B. Andreopoulos, "Graph Deep Learning Hashtag Recommender for Reels," *Proc. IEEE Int. Conf. Big Data Service (BigDataService)*, 2023, pp. 119–126, https://doi.org/10.1109/BigDataService58306.2023.00023.

[37] H.-S. Won, S.-M. Roh, D. Kim, M.-J. Kim, H. Kim, and K.-M. Kim, "EXTRA: Integrating External Knowledge into Multimodal Hashtag Recommendation System," *Proc. IEEE Int. Conf. Web Services (ICWS)*, 2023, pp. 719–721, https://doi.org/10.1109/ICWS60048.2023.00096.

[38] T. Yu, H. Yu, D. Liang, Y. Mao, S. Nie, P.-Y. Huang, M. Khabsa, P. Fung, and Y.-C. Wang, "Generating Hashtags for Short-form Videos with Guided Signals," *Proc. 61st Ann. Meeting Assoc. Comput. Linguist. (ACL)*, 2023, pp. 9482–9495, https://doi.org/10.18653/v1/2023.acl-long.527.

[39] S. Bansal, K. Gowda, M. Z. U. Rehman, C. S. Raghaw, and N. Kumar, "MISHON:

Multimodal Hashtag Recommendation via Graph-Based Deep Neural Network," *Eng. Appl. Artif. Intell.*, vol. 138, Art. no. 109417, 2024, https://doi.org/10.1016/j.engappai.2024.109417.

[40] S. Bansal, K. Gowda, and N. Kumar, "Multilingual Personalized Hashtag Recommendation for Low-Resource Indic Languages Using Graph-Based Deep Neural Network," *Expert Syst. Appl.*, vol. 236, Art. no. 121188, 2024, https://doi.org/10.1016/j.eswa.2023.121188.

[41] R.-Z. Fan, Y. Fan, J. Chen, J. Guo, R. Zhang, and X. Cheng, "RIGHT: Retrieval-Augmented Generation for Mainstream Hashtag Recommendation," *Proc. 46th Eur. Conf. Inf. Retr. (ECIR)*, Lecture Notes Comput. Sci., vol. 14608. Springer, 2024, pp. 39–55, https://doi.org/10.1007/978-3-031-56027-9_3.

[42] V. J. Prakash and S. A. A. Vijay, "A Comprehensive Multimodal Framework for Optimizing Social Media Hashtag Recommendations," *IEEE Trans. Comput. Social Syst.*, vol. 12, no. 5, pp. 2144–2155, 2025, https://doi.org/10.1109/TCSS.2024.3508733.

[43] Z. Tan, Y. Li, X. Yuan, S. Meng, W. Li, and T. Mo, "LLM-MHR: A LLM-Augmented Multimodal Hashtag Recommendation Algorithm," *Proc. IEEE Int. Conf. Web Services (ICWS)*, 2024, pp. 810–821.

[44] M. M. Y. Khalil, Q. Wang, B. Chen, M. A. Rauf, and W. Wang, "MRLKG: Multimodal Representation Learning with Keyword-based Graph for Hashtag Recommendation," *Proc. Int. Conf. Mach. Learn. Natural Lang. Process. (MLNLP)*, 2024, pp. 358–363, https://doi.org/10.1109/MLNLP63328.2024.10800589.

[45] M. Shojaei, H. R. Arabnia, and A. A. Ghorbani, "A Framework for Generating Images and Hashtags for Social Media Posts for Artificial Influencers," *Proc. IEEE Int. Conf. Multimedia Inf. Process. Retr. (MIPR)*, 2024.

[46] S. Yakovlev and N. Shapoval, "Recommendation of Hashtags Using Deep Learning Methods Based on Multimodal Data," *Artif. Intell.*, vol. 29, no. 4, pp. 41–48, 2024, https://doi.org/10.15407/jai2024.04.041.

[47] M. M. Y. Khalil, B. Chen, M. A. Rauf, W. Wang, and Q. Wang, "TAMTE: Topic-Attentive Multimodal Transformer Encoder for Hashtag Recommendation," *Knowl. Inf. Syst.*, vol. 67, pp. 6047–6070, 2025, https://doi.org/10.1007/s10115-025-02398-2.

[48] Q. Yang et al., "CGAN-Enhanced Multimodal Hashtag Recommendation for Microblogs," *J. Inf. Sci. Eng.*, vol. 41, 2025.

[49] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation," *Proc. Int. Conf. Mach. Learn. (ICML)*, 2022, pp. 12888–12900.

[50] J. Li, D. Li, S. Savarese, and S. Hoi, "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models," *Proc. ICML*, 2023, pp. 19730–19742.

[51] Y. Deldjoo et al., "A Review of Modern Recommender Systems Using Generative Models (Gen-RecSys)," *Proc. 30th ACM SIGKDD Conf. Knowl. Discov. Data Min. (KDD)*, 2024, https://doi.org/10.1145/3637528.3671474.

[52] S. Bansal, K. Gowda, A. Sureshbabu K, C. Kothari, and N. Kumar, "A Comprehensive Review on Hashtag Recommendation: From Traditional to Deep Learning and Beyond," arXiv preprint arXiv:2503.18669, 2025.

[53] S. Jain and B. C. Wallace, "Attention is not explanation," *Proc. 2019 Conf. North Amer. Chapter Assoc. Comput. Linguist. (NAACL)*, 2019, pp. 3543–3556, https://doi.org/10.18653/v1/N19-1357.

[54] S. Wiegreffe and Y. Pinter, "Attention is not not explanation," *Proc. EMNLP-IJCNLP*, 2019, https://doi.org/10.18653/v1/D19-1002.

[55] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual Instruction Tuning," *NeurIPS*, 2023.

[56] M. T. Ribeiro, S. Singh, and C. Guestrin, "'Why should I trust you?': Explaining the predictions of any classifier," *Proc. ACM SIGKDD*, 2016, https://doi.org/10.1145/2939672.2939778.

[57] S. M. Lundberg and S. I. Lee, "A unified approach to interpreting model predictions," *NeurIPS*, 2017.

[58] M. Borenstein, L. V. Hedges, J. P. T. Higgins, and H. R. Rothstein, *Introduction to Meta-Analysis*. Chichester: Wiley, 2009, https://doi.org/10.1002/9780470743386.

[59] M. J. Page et al., "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *BMJ*, vol. 372, Art. no. n71, 2021, https://doi.org/10.1136/bmj.n71.

[60] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977, https://doi.org/10.2307/2529310.

[61] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 740–755, 2014, https://doi.org/10.1007/978-3-319-10602-1_48.

[62] J. Pineau et al., "Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program," *J. Mach. Learn. Res.*, vol. 22, no. 164, pp. 1–20, 2021.

[63] H. Nissenbaum, "Privacy as contextual integrity," *Wash. Law Rev.*, vol. 79, no. 1, pp. 119–157, 2004.

[64] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," *Proc. 20th Int. Conf. Artif. Intell. Stat. (AISTATS)*, pp. 1273–1282, 2017.

[65] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi, "BERTScore: Evaluating text generation with BERT," *Int. Conf. Learn. Represent. (ICLR)*, 2020.

[66] J. Ho, A. Jain, and P. Abbeel, "Denoising Diffusion Probabilistic Models," *Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 6840–6851, 2020.

[67] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, pp. 618–626, 2017, https://doi.org/10.1109/ICCV.2017.74.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- SupplementaryInformation.docx