How AI Gets to See: Preparing Community Categories for AI-Assisted Tagging on Wikimedia Commons

Anonymous ACL submission

Abstract

Wikimedia Commons contains decades of community-built categories that support search, navigation, and reuse, but these category strings are difficult for software systems to interpret as machine-readable supervision. As Commons transitions toward Wikidata-linked *Depicts* metadata through Structured Data on Commons, legacy categories offer potential candidate-level signals for AI-assisted tagging: not final labels to apply automatically, but training signals for learning which entities to suggest for contributor review. We study when category-derived strings can support Wikidata *Depicts* grounding. Using learned weighting as a diagnostic probe across two festival/tradition datasets and an auxiliary five-domain collection, we find that short-tag count provides the most robust structural boundary: one short tag yields no learning gain, while the clearest gains appear with three or more short tags. Raw diversity is less stable, predicting gain in a smaller mixed-domain dataset but breaking down under broader heterogeneity and within coarse domain bags. These findings reframe category reuse as upstream data preparation for participatory ML: systems should prioritize images with sufficient short-tag structure, interpret diversity within domain bags, and involve contributors and domain communities in deciding which category-derived signals should inform AI-assisted tagging.

1 Introduction

Wikimedia Commons is one of the world's largest free image repositories. For more than two decades, Commons contributors have annotated and organized its media files through a community-maintained category system: an English-dominant broad folksonomy that supports search, navigation, and reuse across Wikimedia projects and internet users worldwide. These category strings are typically single-language, lack semantic typing, and do not conform to a shared data schema (Yu and McDonald, 2022). Like other user-generated image tags, they can be noisy, inconsistent, incomplete, or ambiguous (Wang et al., 2012), while also reflecting different contributor preferences, tendencies, and assumptions (Thornton and McDonald, 2012; Sen et al., 2006). As a result, software systems cannot reliably interpret category meanings or link them to common concepts. This limits semantic search, discoverability, and multilingual support on Commons.

To address these limitations and support Commons as a machine-readable visual knowledge infrastructure, the Wikimedia Foundation launched the Structured Data on Commons (SDC) project[1], which introduced Wikidata-linked metadata for media files. The Computer-Aided Tagging (CAT) tool[2] was an early attempt to support this transition by using AI to suggest Wikidata entities as *Depicts* statements for contributor review. CAT was deactivated in 2023 after community pushback. Prior qualitative work (Yu and McDonald, 2026) shows that contributors objected not only to model errors, but also to a top-down workflow that appeared to bypass the existing category system rather than building on it. Contributors did not simply reject automation. They wanted a role in deciding how category knowledge should be prepared, interpreted, and incorporated into structured *Depicts* metadata. The issue for AI-assisted tagging on Commons is therefore not to invent another tagging system, but to determine how participatory ML can prepare legacy category knowledge for contributor-reviewed, Wikidata-linked *Depicts* suggestions.

This paper responds from the data-preparation side. If future AI-assisted tagging systems should

[1] https://commons.wikimedia.org/wiki/Commons:Structured_data

[2] https://commons.wikimedia.org/wiki/Commons:Structured_data/Computer-aided_tagging

build on Commons' legacy category knowledge, a prior question must be answered: when can category-derived strings provide candidate-level supervision for Wikidata *Depicts* grounding? We use candidate supervision to mean selected category-derived strings that are not final labels, but training signals for learning which possible Wikidata entities should be suggested for contributor review.

We study this question through weakly supervised entity grounding. Using only each Commons image's category-derived tags, we ask when a lightweight learned weighting model can improve over simple aggregation in retrieving the image's gold Wikidata *Depicts* entities. We use learned weighting as a diagnostic probe rather than as a production model, so that stronger architectures do not obscure the upstream preparation problem. We find that short-tag count provides the most robust structural boundary: one short tag provides no alternative for learned selection, while the clearest gains appear with three or more short tags. Diversity is more conditional: it can track useful alternatives in a smaller mixed-domain dataset, but becomes unstable under broader heterogeneity and within coarse domain bags.

These findings contribute a training-image selection logic for participatory ML on Commons. Rather than ranking images by global diversity, systems should prioritize images with sufficient short-tag structure, interpret tags within domain bags, and separate heterogeneous domains by target type before using legacy categories as candidate supervision. This reframes participation as upstream data preparation: contributors and domain communities are needed not only to verify model outputs, but also to decide which category-derived signals should inform AI-assisted tagging systems in the first place.

2 Related Work

2.1 Participatory ML

Participatory ML involves affected communities in the design, auditing, adaptation, and governance of ML systems, rather than treating labels, features, or model behavior as engineer-defined objects alone (Sloane et al., 2022; Delgado et al., 2023; Birhane et al., 2022; Martin Jr et al., 2020). In open knowledge work, Wikipedia systems such as ORES (Halfaker and Geiger, 2020; Smith et al., 2020) and LiftWing[3] show how communities can negotiate, audit, and adapt ML tools around local editorial needs and values. These systems have often been organized around language editions, whose governance structures provide natural units for participatory model development. Commons presents a different problem. Contributors from many language editions share a single visual repository without comparable linguistic segmentation (Yu and McDonald, 2022). Prior qualitative work on Computer-Aided Tagging found that Commons contributors were willing to participate in domain-specific preparation of category knowledge for structured data, and suggested content domains as a possible starting point for participatory ML on Commons (Yu and McDonald, 2026). We test this premise empirically by asking whether domain organization clarifies the structure of category-derived alternatives and the conditions under which legacy categories can function as candidate supervision.

2.2 Weak Supervision

Our task is related to weakly supervised entity grounding, where training signal comes from noisy or indirect sources rather than clean annotations. Distant supervision (Mintz et al., 2009) established the use of existing knowledge bases to generate noisy labels for relation extraction. Later work extends weak or indirect supervision to entity linking (Wu et al., 2020), visual grounding (Rohrbach et al., 2016), and multimodal retrieval (Faghri et al., 2017; Jia et al., 2021). A recurring finding is that the structure of the weak supervision source, not only its size, shapes what a model can learn from it. Folksonomies and community tagging systems have supported tag recommendation and image retrieval (Sigurbjörnsson and Van Zwol, 2008; Liu et al., 2009). This work shows that community-produced tags can support learning, but also that tag noise, ambiguity, and inconsistency require careful handling. Commons categories share these problems while adding further complexity: they are not designed as depiction tags, they mix multiple semantic roles, and they are embedded in a large navigational hierarchy. Preparing them for supervised learning therefore requires more than noise filtering. It requires role assignment and domain structuring.

[3] https://wikitech.wikimedia.org/wiki/Machine_Learning/LiftWing

2.3 Data Preparation

Data-centric AI shifts attention from model architecture to data quality, selection, and documentation (Zha et al., 2025; Sambasivan et al., 2021). Related work on programmatic labeling (Ratner et al., 2017), confident learning (Northcutt et al., 2021), curriculum learning (Bengio et al., 2009), and dataset documentation (Gebru et al., 2021; Holland et al., 2020) shows that labels, examples, provenance, and intended use shape downstream model behavior. We apply this perspective to inherited community metadata: short-tag count acts as a structural analogue of data selection, while domain structure and target-type consistency determine whether category-derived strings can function as candidate supervision before model training or suggestion generation. Concurrent work audits the complementary question of how much multilingual Commons captions and descriptions remain recoverable from current category-and-entity metadata (Anonymous, 2026). We instead study how legacy categories can be prepared as candidate supervision for future Depicts suggestions.

3 Data

All datasets are built from Wikimedia Commons files with community-maintained categories[4] and Wikidata *Depicts* statements. We treat Commons categories as inherited community metadata and *Depicts* statements as gold target entities. Category strings are parsed into category-derived tags as described in Section 4. Our main analyses focus on short tags, defined as category-derived tags containing at most three words, with fallback to the full tag set when no short tag is available. This cutoff is a practical operationalization of compact candidate-like strings: stricter cutoffs become sparse, while looser cutoffs admit longer compound category strings that are less likely to function as direct candidate supervision. Sensitivity checks appear in Appendix B. All data items contain at least one category-derived tag and at least one *Depicts* entity. The two main datasets have similar gold-set densities: images contain an average of 1.64 Wikidata *Depicts* entities in the smaller mixed-domain dataset and 1.63 in the larger heterogeneous extension. We therefore treat the larger extension primarily as a heterogeneity stress test rather than as a setting with substantially denser gold targets.

[4] https://commons.wikimedia.org/wiki/Commons:Categories

The smaller mixed-domain dataset contains 1,007 Commons images from multiple festival- and tradition-related domains, including Chinese New Year, Diwali, Day of the Dead, Halloween, and related cultural practices. We use it as the first main setting for testing when learned weighting can improve over simple heuristics in a realistic weak-supervision setting. The larger heterogeneous extension contains 4,931 images across 17 festival- and tradition-related domains, including Christmas, Easter, Onam, Navaratri, Lantern Festivals, Carnival of Rio de Janeiro, Day of the Dead, Chinese New Year variants, and other culturally grounded celebrations. It partially overlaps with the smaller dataset, sharing 576 images, but is not a strict superset. We retain this overlap intentionally: the smaller dataset serves as the initial mixed-domain setting where the diversity hypothesis is most interpretable, while the larger extension tests whether the same patterns survive broader domain heterogeneity. We therefore do not treat the two datasets as independent benchmarks. We use them as paired diagnostic settings with different heterogeneity profiles.

The auxiliary five-domain collection contains 2,462 featured images from animals, plants, people, places, and natural phenomena. It is not used as a third main benchmark. Instead, it supports the five-domain diagnostics in Section 5.6, where we examine whether domain organization makes raw diversity reliable or reveals different structures of alternatives. For this collection, we additionally compute tag-gold overlap, coarse lexical type counts, and per-domain diagnostics. The lexical type-count method is described in Appendix A.

4 Method

We formulate the task as weakly supervised entity grounding from inherited Commons categories. For each Commons image, using only its category-derived tags, the goal is to retrieve Wikidata entities corresponding to the image's gold *Depicts* statements. We use frozen text embeddings and lightweight aggregation methods so that the analysis focuses on when category-derived tags provide learnable structure, rather than on optimizing a new grounding architecture.

4.1 Task and Candidates

For each image, we clean Commons category strings by removing the `Category:` prefix, normal-

izing surface forms, dropping empty strings, and deduplicating the resulting category-derived tags. As introduced in Section 3, short tags are category-derived tags containing at most three words. If an image has no short tags, aggregation methods fall back to the full category-derived tag set so that the image remains evaluable. Unless otherwise stated, short-tag count is measured before fallback.

For each image, the gold set consists of its Wikidata *Depicts* entities. Because an image may depict multiple entities, evaluation is multi-label. We compute Recall@5 against the full gold set for each image, not only against gold entities that appear in a pruned candidate pool. We build a global candidate inventory from the unique Wikidata *Depicts* labels observed in each dataset. Category-derived tags and candidate labels are encoded with the same frozen sentence encoder, and candidates are ranked by cosine similarity between candidate-label embeddings and the tag-set representation produced by each aggregation method.

4.2 Aggregation and Weighting

We compare four aggregation strategies. *Short-tag aggregation* averages the embeddings of short tags, with fallback to the full category-derived tag set when no short tag is available. This is the main heuristic baseline for learning-gain analyses. *Equal aggregation* averages all category-derived tag embeddings without short-tag filtering or learned weights. *Oracle single-tag* retrieves candidates from each individual tag alone and reports the highest per-image Recall@5 against the full gold set. Because it requires knowing the gold set, it is used only as an upper-bound diagnostic for recoverable signal.

Learned weighting uses the same short tags and fallback as short-tag aggregation, but assigns learned weights before aggregation. We encode tags and candidate Wikidata labels with frozen `sentence-transformers/all-MiniLM-L6-v2` embeddings (Reimers and Gurevych, 2019; Wang et al., 2020). A lightweight two-layer MLP predicts a scalar score for each tag, normalizes scores with a softmax over the image's tag set, and constructs a weighted tag representation for candidate reranking. The weighted representation is added to the mean tag representation so that the learned model remains close to the short-tag baseline. We train the MLP with Adam, learning rate 10^{-3}, for 5 epochs. The model is intentionally small and standard: its role is to test whether category-derived tags contain learnable selection structure, not to propose a new grounding architecture.

For learned weighting, we construct an image-specific pruned candidate pool by retrieving the top 100 entities most similar to the short-tag aggregation representation. Gold entities are not forcibly inserted: only gold *Depicts* labels that survive pruning are used as positives during training, while the remaining candidates act as implicit negatives. At evaluation time, Recall@5 is still computed against the image's full gold set. Learned weighting therefore functions as a diagnostic reranker within a plausible candidate space, not as an end-to-end retriever over the full Wikidata universe.

4.3 Metrics and Diagnostics

Because the downstream setting is AI-assisted candidate suggestion, we report Precision@5 and Recall@5 in overall comparisons, and use Recall@5 as the primary metric in diagnostic analyses. *Learning gain* is learned-weighting Recall@5 minus short-tag aggregation Recall@5. Because learned weighting reranks a pruned pool while short-tag aggregation retrieves over the global inventory, learning gain is interpreted diagnostically rather than as a direct architecture comparison. *Oracle gap* is oracle single-tag Recall@5 minus short-tag aggregation Recall@5. It measures recoverable headroom that may or may not be exploited by learned weighting.

We use diversity narrowly as an operational proxy for differences among short tags, not as a normative or demographic concept. We compute three proxies: *embedding diversity*, the mean pairwise cosine distance among short-tag embeddings; *lexical diversity*, the mean pairwise Jaccard distance among tokenized short tags; and *IDF spread*, the variance of tag-level inverse document frequency scores. We group images by short-tag count and diversity tertiles, then use controlled two-tag comparisons, regressions, oracle-gap analysis, and the auxiliary five-domain collection to test whether diversity remains informative beyond short-tag count and across domains.

5 Results

5.1 Overall Performance

We begin with overall retrieval performance across the two main datasets. This comparison establishes whether category-derived tags contain signal that

Dataset	Method	P@5	R@5
Small	Short-tag aggregation	0.1408	0.5526
Small	Equal aggregation	0.1692	0.6666
Small	Learned weighting	0.1611	0.6618
Small	Oracle single-tag	0.2139	0.8291
Large	Short-tag aggregation	0.1115	0.4182
Large	Equal aggregation	0.1209	0.4510
Large	Learned weighting	0.1198	0.4612
Large	Oracle single-tag	0.1551	0.5846

Table 1: Overall baseline performance.

Dataset	Group	N	Short-tag R@5	Learned R@5	Gain
Small	1 tag	445	0.5074	0.5074	0.0000
Small	2 tags	273	0.6205	0.7730	0.1525
Small	$\geq$3 tags	289	0.5582	0.7946	0.2364
Large	1 tag	2568	0.4303	0.4303	0.0000
Large	2 tags	1141	0.5165	0.5318	0.0153
Large	$\geq$3 tags	1222	0.3010	0.4602	0.1592

Table 2: Performance grouped by short-tag count.

can be selected or weighted.

Table 1 shows that learned weighting improves over short-tag aggregation in both datasets, but does not consistently outperform equal aggregation. Oracle single-tag performance is higher than all deployable methods, indicating that individual tags sometimes contain strong recoverable signal. The gap between oracle and learned weighting suggests that this signal is not automatically exploitable. The rest of the results therefore ask when short tags provide useful structure to select among.

5.2 Short Tags Enable Selection

We next ask when learned weighting has any meaningful choice to make. Images with only one short tag provide no alternative for the model to upweight or downweight, so learned weighting and short-tag aggregation receive the same effective signal. This one-tag case establishes a boundary condition: learned weighting can only help when an image has more than one short tag.

Table 2 shows this boundary across both main datasets. One-tag cases yield no learning gain. Once multiple short tags are available, learned weighting can improve, with the clearest gains in the three-or-more-tag group: +0.2364 in the smaller mixed-domain dataset and +0.1592 in the larger heterogeneous extension.

Short-tag count therefore marks whether selection is possible, but not whether the candidate set is easy or high quality. In the larger heterogeneous extension, the three-or-more-tag group has a lower short-tag baseline than the one-tag group (0.3010 vs. 0.4303), even though learned weighting produces a much larger gain. Richer tag sets may therefore occur in harder cases, with more complex images, larger gold sets, or more heterogeneous domain contexts. Multiple short tags are necessary for learned weighting to improve, but not sufficient. The next sections ask whether diversity captures candidate quality and whether its meaning remains stable across heterogeneous category data.

Proxy	Group	N	Short-tag R@5	Learned R@5	Gain
Embedding	LOW	336	0.3595	0.3595	0.0000
Embedding	MEDIUM	336	0.8100	0.9065	0.0965
Embedding	HIGH	335	0.4882	0.7197	0.2315
Lexical	LOW	336	0.3595	0.3595	0.0000
Lexical	MEDIUM	336	0.7495	0.8906	0.1411
Lexical	HIGH	335	0.5489	0.7356	0.1867
IDF spread	LOW	336	0.5070	0.5070	0.0000
IDF spread	MEDIUM	336	0.6074	0.7195	0.1121
IDF spread	HIGH	335	0.5435	0.7593	0.2158

Table 3: Diversity-group analysis in the smaller mixed-domain dataset.

5.3 Diversity Helps in the Small Dataset

Having established that multiple short tags are necessary for learned weighting to help, we next ask whether differences among those tags matter. We test whether diversity in a short-tag set predicts learning gain in the smaller mixed-domain dataset, using embedding diversity, lexical diversity, and IDF spread.

Table 3 shows the strongest version of the diversity pattern. Across all three proxies, LOW-diversity groups yield no learning gain, while MEDIUM and HIGH groups show positive gains. For embedding diversity, learned weighting improves over short-tag aggregation by 0.0965 in the MEDIUM group and 0.2315 in the HIGH group. Lexical diversity and IDF spread show the same broad pattern. In this smaller setting, diversity appears to track differences among short tags that learned weighting can exploit.

This result supports the diversity hypothesis in an interpretable setting, but does not establish diversity as a general selection rule. Gains also reflect baseline difficulty and available headroom, and the meaning of diversity may change under broader heterogeneity. The next section tests whether the pattern survives in the larger extension.

5.4 Diversity Becomes Unstable

The smaller mixed-domain dataset suggests that diversity can track useful alternatives in a relatively interpretable setting. We next test whether this rela-

Proxy	Group	N	Short-tag R@5	Learned R@5	Gain
Embedding	LOW	1644	0.2823	0.2823	0.0000
Embedding	MEDIUM	1644	0.5753	0.6539	0.0786
Embedding	HIGH	1643	0.3969	0.4473	0.0504
Lexical	LOW	1644	0.2823	0.2823	0.0000
Lexical	MEDIUM	1644	0.5418	0.6322	0.0905
Lexical	HIGH	1643	0.4305	0.4690	0.0385
IDF spread	LOW	1644	0.3069	0.3083	0.0014
IDF spread	MEDIUM	1644	0.5492	0.6055	0.0563
IDF spread	HIGH	1643	0.3984	0.4698	0.0713

Table 4: Diversity-group analysis in the larger heterogeneous extension.

Analysis	Result
Controlled 2-tag gain, LOW diversity	$+0.0222$
Controlled 2-tag gain, HIGH diversity	$+0.0084$
Controlled gain difference, HIGH–LOW	-0.0138
Regression short-tag count coefficient	0.0422 (95% CI $[0.0332, 0.0532]$)
Regression raw-diversity coefficient	-0.0352 (95% CI $[-0.0741, 0.0033]$)
Continuous correlation, diversity vs. gain	$r = 0.1297$, $p < 0.001$
Oracle gap, LOW diversity	0.1628
Oracle gap, MEDIUM diversity	0.1191
Oracle gap, HIGH diversity	0.2174

Table 5: Controlled, regression, and oracle diagnostics for raw embedding diversity in the larger heterogeneous extension.

tionship survives broader heterogeneity. If raw diversity were a stable proxy, higher-diversity groups should show consistently larger learning gains in the larger extension. We therefore repeat the same tertile-based analysis for embedding diversity, lexical diversity, and IDF spread.

Table 4 weakens the diversity pattern. For embedding diversity, the MEDIUM group shows the largest gain ($+0.0786$), while the HIGH group gains less ($+0.0504$). Lexical diversity shows the same non-monotonic pattern: the MEDIUM group gains $+0.0905$, while the HIGH group gains only $+0.0385$. IDF spread is more monotonic, but the overall pattern no longer supports a simple rule that higher raw diversity yields larger learning gains.

This instability suggests that diversity changes meaning under heterogeneity. High diversity may still indicate useful signal, but it may also reflect mixed target types, domain difficulty, or more complex gold targets. The large differences in short-tag baselines across tertiles reinforce this point: grouped gains reflect both candidate structure and example difficulty. Diversity therefore remains informative, but not as a global selection rule. The next section tests whether it provides stable predictive value beyond short-tag count.

5.5 Short-Tag Count Is More Robust

The grouped results suggest that raw diversity becomes unstable in the larger heterogeneous extension. Because diversity groups also differ in short-tag count, baseline performance, and likely task difficulty, we run controlled, regression, and oracle-gap diagnostics.

Table 5 shows that raw embedding diversity has a weak positive aggregate correlation with learning gain, but this relationship does not survive more controlled views. Among images with exactly two short tags, HIGH-diversity pairs gain less than LOW-diversity pairs. In regression, short-tag count remains clearly positive, while raw diversity is not clearly positive after accounting for short-tag count.

Oracle diagnostics further separate available signal from exploitable signal. The HIGH-diversity group has the largest oracle gap, indicating more available headroom, but it does not have the largest realized learning gain in Table 4. Higher diversity may therefore indicate that useful signal exists somewhere in the short-tag set, but not that the signal is easy to select. Short-tag count is the more robust structural signal: it marks whether there is enough material for selection. The next section examines what diversity is measuring across domains.

5.6 Domains Reveal Diversity

The preceding analyses show that raw diversity is not a stable global proxy. We next ask whether evaluating diversity within coarse domain bags makes it reliable. We test this with an auxiliary five-domain collection covering animals, plants, people, places, and natural phenomena. This analysis is diagnostic rather than a third benchmark: its goal is to test whether domain control rescues raw diversity, or instead reveals that the meaning of "alternative" differs across domains.

Table 6 summarizes the structure of the five domains. Animals and plants show more direct relationships between short tags and gold targets: animals have the highest tag–gold overlap and the smallest average gold set, while plants remain partly taxonomic but involve larger gold sets. People, places, and natural phenomena have lower overlap and more diffuse gold sets, suggesting that short tags more often mix entity identity, roles, events, scenes, locations, carriers, or background context. The final column is descriptive rather than independently annotated. It summarizes patterns supported by average tag count, gold-set size, tag–gold overlap, representative examples, and the rule-based lexical analysis in Appendix A.

Domain	Avg. tags	Avg. gold	Overlap	Main structure
Animals	1.96	1.51	0.38	Taxonomic alternatives
Plants	2.44	2.88	0.22	Botanical identity + context
People	3.46	2.55	0.16	Entity/role/event mixture
Places	2.70	3.50	0.16	Place identity + scene context
Natural phenomena	3.85	4.05	0.13	Phenomenon + carrier/background

Table 6: Domain structure in the auxiliary five-domain collection. Avg. tags and Avg. gold are per-sample means. Overlap is average token overlap between short tags and gold *Depicts* labels. The final column is descriptive and is not used as supervision or model input.

Domain	Gain–Tag r	Gain–Div r	Joint Tag	Joint Div	Tag–Div r
Animals	0.1733	0.0535	0.0344	-0.0659	0.6737
Plants	0.1074	0.0528	0.0106	-0.0051	0.5419
People	0.0575	0.0156	0.0061	-0.0097	0.4214
Places	0.0517	-0.0117	0.0067	-0.0340	0.5474
Natural phenomena	0.0198	-0.0096	0.0024	-0.0187	0.4044

Table 7: Within-domain diagnostics in the auxiliary five-domain collection. Gain–Tag and Gain–Div are Pearson correlations between learning gain and short-tag count or raw embedding diversity. Joint Tag and Joint Div are coefficients from domain-internal regressions of learning gain on short-tag count and raw embedding diversity. Tag–Div reports their correlation.

Table 7 shows that raw embedding diversity does not become stable within domains. Short-tag count is more consistently associated with learning gain than diversity, especially in animals and plants. Raw diversity remains weak, unstable, or negative across domains. In animals, for example, the zero-order correlation between gain and diversity is weakly positive, but the joint diversity coefficient becomes negative after controlling for short-tag count. Across all five domains, short-tag count and diversity remain strongly correlated, indicating that raw diversity still mixes the number of available tags with their structure.

These results clarify the role of domain organization. Domain bags do not rescue diversity by making it a stable selection metric. Rather, they reveal what diversity is measuring. In animals and plants, differences among short tags more often correspond to taxonomic or botanical alternatives. In people, places, and natural phenomena, similar differences may instead reflect mixtures of entities, roles, events, scenes, locations, carriers, or background context. Diversity is therefore useful only after domain structure is specified: it can index clean alternatives in some domains and signal the need for further target-type structuring in others.

6 Discussion

The results above show that category-derived tags support learned selection only under specific structural conditions: short-tag count marks when selection is possible, while raw diversity depends on domain and target-type structure. These findings change how AI-assisted image tagging on Commons should be understood. Much of the annotation and human-in-the-loop pipeline literature treats human participation as corrective: a model produces candidate labels, and people accept, reject, or repair its outputs after the fact (Settles, 2009; Monarch, 2021; Ouyang et al., 2022; Snow et al., 2008). The deactivated CAT tool followed this logic by applying a generic image-recognition system across Commons and asking contributors to review suggested labels. Commons exposes the limits of this assumption. In open knowledge infrastructures, contributor labor is not only corrective but constitutive: contributors help define what counts as a meaningful target, which legacy signals should be reused, and how AI tools should enter existing community workflows. Prior qualitative work shows that Commons contributors objected not only to suggestion quality, but also to a workflow that bypassed legacy metadata, especially categories, and treated AI deployment as a top-down tool rather than a community-negotiated process (Yu and McDonald, 2026). For Commons, the issue is therefore not simply whether humans review model outputs, but where community judgment enters the AI pipeline.

Our results reframe category reuse on Commons as an upstream data-preparation problem for participatory ML. Legacy categories contain useful signal, but they do not become useful training signal simply by being passed to a learning pipeline. The learned weighting model is intentionally small and standard: its purpose is not to maximize retrieval performance, but to reveal when category-derived tags contain enough structure for a model to select among alternatives. The contribution is therefore a division of labor: models can surface candidate images and structural signals, while contributors and domain communities decide whether those signals describe meaningful targets for AI-assisted tagging.

The clearest condition is short-tag count. Images with one short tag provide no alternative for learned selection, while images with multiple short tags can support learning gain. Across datasets and domains, the clearest gains appear when images have three or more short tags. Training-image selection should therefore prioritize images with at least three short tags, while treating two-tag cases as a

weaker minimum. This is not a universal threshold, but an inspectable operational rule for selecting images whose category-derived tags are compact enough and numerous enough to support candidate-level supervision for *Depicts* grounding. This step can be model-assisted as a first-pass filter before contributors decide whether those images belong in a domain-specific preparation workflow.

Raw diversity is diagnostic rather than a global ranking signal. High diversity often means that category-derived tags mix target types and should not be learned together without further structuring. In animals and plants, tag differences more often reflect taxonomic or botanical alternatives. In people, places, and natural phenomena, they may mix identity, role, activity, event, place, scene, carrier, or context. These cases require contributor-defined target-type bags before training-image selection: models can flag or cluster cases, but Commons contributors must decide what kind of target a bag is being prepared for before tags can be used as candidate supervision.

This interpretation leads to a conservative training-image selection pipeline for preparing category-derived signals before AI-assisted tagging. Short-tag count can support model-assisted filtering; diversity and oracle-gap diagnostics can surface cases for inspection; domain and target-type organization require contributor judgment about what kind of target is being prepared; and final reuse should remain tied to contributor-reviewed candidate suggestions rather than automatic conversion.

This upstream framing raises a second question: what unit should organize participatory ML on Commons? Prior participatory ML approaches in Wikipedia often use language editions as natural participation units: each language edition develops its own policies, workflows, and editorial norms, so language editions can provide units for defining ML goals, training data, evaluation criteria, and deployment processes (Halfaker and Geiger, 2020; Smith et al., 2020). Commons does not offer the same organizing structure. It is a multilingual media repository that supports more than 300 Wikipedia language editions, but its category system is shared across languages and organized through decentralized, cross-cutting categorization practices rather than language-edition boundaries. As a result, participatory ML on Commons cannot simply be organized by language community alone.

Following prior qualitative work (Yu and McDonald, 2026), we therefore treat domains as a possible organizing unit for participatory ML. Our results show that domain organization is useful but not definitive. Some domains, especially animals and plants, appear more internally coherent: their short tags more often correspond to taxonomic or botanical alternatives, consistent with prior contributor suggestions that taxon- and plant-identification tasks are promising domain-specific AI/ML use cases on Commons (Yu and McDonald, 2026). Other broad domains, such as people, places, and natural phenomena, are more heterogeneous. Their category-derived tags may mix identity, role, activity, event, scene, location, carrier, or background context. Existing category pages, WikiProjects, and maintenance workflows could provide venues for deciding whether a domain is coherent, whether high-diversity cases should be split by target type, and which category-derived tags should function as candidate-level supervision. The goal is to make the preparation step legible and reviewable within existing Commons governance, rather than to bypass it through automatic filtering.

The broader HCAI implication is that responsible AI-assisted tagging begins before suggestion generation. Commons categories are inherited community structures that organize different kinds of visual and contextual meaning. Preparing them for AI-assisted tagging requires role assignment rather than indiscriminate reuse. Some short tags can support candidate-level supervision for *Depicts* grounding, while others indicate that the domain, target type, or image set needs further community interpretation. The design question is therefore not only whether a model can learn from legacy categories, but who decides which category-derived signals should inform which tagging task.

7 Conclusion

This paper examined when community-maintained Commons categories can be prepared as candidate-level signals for AI-assisted *Depicts* tagging. We showed that short-tag count provides the most reliable signal, that diversity is conditional on domain and target type, and that broad domains vary in how coherently their category structures support learned selection. These findings reframe category reuse as upstream data preparation for participatory ML. Open knowledge infrastructures need role assignment, not indiscriminate reuse.

8 Limitations

This study is situated in Wikimedia Commons, whose category system is unusually large, community-maintained, English-dominant, and historically layered. The specific short-tag threshold and domain patterns may not transfer directly to other repositories. We therefore treat the three-word cutoff as an operational rule for compact category-derived tags, not as a universal boundary.

Our evaluation uses Wikidata *Depicts* statements as gold target entities. These statements provide a practical grounding target, but they are incomplete and shaped by existing Commons and Wikidata practices. Some useful category-derived signals may not correspond to current *Depicts* entities, and some meanings may be better preserved as prose, provenance, contextual metadata, or community-facing categories.

Finally, learned weighting is a diagnostic probe rather than a production grounding model. Stronger architectures or multimodal models may achieve higher retrieval performance, but our goal is to test when category-derived tags provide learnable selection structure. Future work should evaluate this preparation pipeline with Commons contributors, additional domains, and contributor-defined target-type bags.

9 Ethical Considerations

This study analyzes publicly available Wikimedia Commons metadata, including community-maintained categories and Wikidata *Depicts* statements. Wikimedia Commons and Wikidata content are publicly available under Wikimedia project licensing and reuse terms. Our derived analysis uses only metadata fields available from these public project pages and does not redistribute media files. We do not collect private personal information, interview contributors, or evaluate individual contributors. Our analyses focus on aggregate patterns in category-derived signals and their suitability as candidate-level signals for AI-assisted tagging.

Some category-derived tags and Wikidata *Depicts* labels may name identifiable public entities, including people, places, or organizations, because such names are part of public Commons and Wikidata metadata. Commons metadata may also refer to sensitive, culturally specific, historical, or potentially offensive subjects. We do not use contributor identifiers, infer attributes about individuals, analyze private persons, or redistribute media files. All analyses are conducted at aggregate level over category-derived strings and *Depicts* labels.

Although the data are public, Commons categories are produced through volunteer labor and may encode cultural, historical, curatorial, or community-specific meanings. We therefore do not argue that legacy categories should be automatically converted into *Depicts* statements or used indiscriminately as training data. Some category-derived strings may support candidate-level supervision for contributor-reviewed *Depicts* grounding, while others may be better preserved as provenance, context, maintenance information, or community-facing organization.

The proposed selection pipeline is intended to support contributor-facing review rather than automation-first conversion. Future AI-assisted tagging systems should involve domain judgment and community governance in deciding which category-derived meanings should inform contributor-reviewed *Depicts* suggestions.

References

Anonymous. 2026. What ai gets to see: Auditing the loss of multilingual community knowledge in machine-facing metadata on wikimedia commons. Under review.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th annual international conference on machine learning*, pages 41–48.

Abeba Birhane, William Isaac, Vinodkumar Prabhakaran, Mark Diaz, Madeleine Clare Elish, Iason Gabriel, and Shakir Mohamed. 2022. Power to the people? opportunities and challenges for participatory ai. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–8.

Fernando Delgado, Stephen Yang, Michael Madaio, and Qian Yang. 2023. The participatory turn in ai design: Theoretical foundations and the current state of practice. In *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–23.

Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612*.

Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. Datasheets for datasets. *Communications of the ACM*, 64(12):86–92.

Aaron Halfaker and R Stuart Geiger. 2020. Ores: Lowering barriers with participatory machine learning in wikipedia. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW2):1–37.

Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2020. The dataset nutrition label. *Data protection and privacy*, 12(12):1.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

Dong Liu, Xian-Sheng Hua, Linjun Yang, Meng Wang, and Hong-Jiang Zhang. 2009. Tag ranking. In *Proceedings of the 18th international conference on World wide web*, pages 351–360.

Donald Martin Jr, Vinodkumar Prabhakaran, Jill Kuhlberg, Andrew Smart, and William S Isaac. 2020. Participatory problem formulation for fairer machine learning through community based system dynamics. *arXiv preprint arXiv:2005.07572*.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 1003–1011.

Robert Munro Monarch. 2021. *Human-in-the-Loop Machine Learning: Active learning and annotation for human-centered AI*. Simon and Schuster.

Curtis Northcutt, Lu Jiang, and Isaac Chuang. 2021. Confident learning: Estimating uncertainty in dataset labels. *Journal of Artificial Intelligence Research*, 70:1373–1411.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alexander Ratner, Stephen H Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. 2017. Snorkel: Rapid training data creation with weak supervision. In *Proceedings of the VLDB endowment. International conference on very large data bases*, volume 11, page 269.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 3982–3992.

Anna Rohrbach, Marcus Rohrbach, Ronghang Hu, Trevor Darrell, and Bernt Schiele. 2016. Grounding of textual phrases in images by reconstruction. In *European Conference on Computer Vision*, pages 817–834. Springer.

Nithya Sambasivan, Shivani Kapania, Hannah Highfill, Diana Akrong, Praveen Paritosh, and Lora M Aroyo. 2021. "everyone wants to do the model work, not the data work": Data cascades in high-stakes ai. In *proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Shilad Sen, Shyong K Lam, Al Mamunur Rashid, Dan Cosley, Dan Frankowski, Jeremy Osterhouse, F Maxwell Harper, and John Riedl. 2006. Tagging, communities, vocabulary, evolution. In *Proceedings of the 2006 20th anniversary conference on Computer supported cooperative work*, pages 181–190.

Burr Settles. 2009. Active learning literature survey.

Börkur Sigurbjörnsson and Roelof Van Zwol. 2008. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th international conference on World Wide Web*, pages 327–336.

Mona Sloane, Emanuel Moss, Olaitan Awomolo, and Laura Forlano. 2022. Participation is not a design fix for machine learning. In *Proceedings of the 2nd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, pages 1–6.

C Estelle Smith, Bowen Yu, Anjali Srivastava, Aaron Halfaker, Loren Terveen, and Haiyi Zhu. 2020. Keeping community in the loop: Understanding wikipedia stakeholder values for machine learning-based systems. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14.

Rion Snow, Brendan O'connor, Dan Jurafsky, and Andrew Y Ng. 2008. Cheap and fast–but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 conference on empirical methods in natural language processing*, pages 254–263.

Katherine Thornton and David W McDonald. 2012. Tagging wikipedia: collaboratively creating a category system. In *Proceedings of the 2012 ACM International Conference on Supporting Group Work*, pages 219–228.

Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. 2012. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM computing surveys (CSUR)*, 44(4):1–24.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Ledell Wu, Fabio Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Scalable zero-shot entity linking with dense entity retrieval. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 6397–6407.

Yihan Yu and David W McDonald. 2022. Unpacking stitching between wikipedia and wikimedia commons: Barriers to cross-platform collaboration. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–35.

Yihan Yu and David W. McDonald. 2026. Computer-aided tagging on wikimedia commons: Designing for human-ai collaboration in open knowledge work. Accepted to CSCW 2026, to appear.

Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric artificial intelligence: A survey. *ACM Computing Surveys*, 57(5):1–42.

A Lexical Type Counts for Domain Summaries

To support the interpretive summaries in Table 6, we computed coarse lexical type counts over short tags and gold *Depicts* labels. This analysis is descriptive only: it is not used as supervision, not used as a model feature, and not treated as an independently annotated variable.

We define seven coarse lexical types: *animal*, *plant*, *person_role*, *place*, *phenomenon_time*, *specific_objectish*, and *other_context*. These categories are intended to characterize the broad structure of tags and gold labels within a domain, not to provide a fine-grained ontology.

Each phrase is classified using a deterministic surface-rule heuristic. We first remove any `Category:` prefix, lowercase the phrase, tokenize it with regular expressions, and then match the resulting tokens against hand-written keyword lists and a small number of patterns. For example, terms such as *bird*, *fish*, or *coral* are mapped to *animal*; terms such as *flower*, *tree*, or *orchidaceae* are mapped to *plant*; terms such as *woman*, *children*, or *portrait* are mapped to *person_role*; terms such as *city*, *bridge*, *bay*, or *mountain* are mapped to *place*; terms such as *sunrise*, *fog*, *reflection*, or *snow* are mapped to *phenomenon_time*; and terms such as *building*, *boat*, *hat*, or *statue* are mapped to *specific_objectish*. Phrases that do not match these rules are assigned to *other_context*.

The heuristic also includes two simple pattern checks. First, phrases that resemble Latin binomials are mapped to *animal* or *plant* when they occur in the corresponding domain. Second, phrases containing place-like constructions such as *in X*, *of X*, *from X*, or *at X*, as well as proper-name-like phrases in the places domain, are mapped to *place*.

We apply this heuristic separately to short tags and gold *Depicts* labels, then aggregate counts and proportions within each domain. The resulting file, `domain_type_counts.csv`, is used only to support the descriptive interpretation of the domain structures in Table 6. For example, in the animals domain, many retained tags and gold labels are species names, Latin binomials, or terms such as *fish*, *bird*, and *coral*. Combined with the low average filtered tag count, low average gold-set size, and high tag–gold overlap, this supports the summary that the animals domain contains relatively direct taxonomic alternatives.

This analysis has several limitations. It is not manual annotation, not a trained classifier, and not an external ontology or dictionary system. We do not claim independent inter-annotator validation for these labels. The analysis should therefore be read as a reproducible, rule-based descriptive check that helps interpret domain structure, rather than as a source of supervision or a formal categorization of the data. Its role is solely to provide a transparent, reproducible basis for the interpretive summaries in Table 6, and to make those summaries checkable against the underlying data.

	Smaller dataset			Larger extension		
Cutoff	Avg. tags	2-tag gain	≥3 gain	Avg. tags	2-tag gain	≥3 gain
≤2 words	2.72	0.1762	0.2075	1.99	0.0040	0.2080
≤3 words	2.27	0.1525	0.2364	2.06	0.0193	0.1509
≤4 words	2.56	0.1562	0.2093	2.44	0.0146	0.1368
≤5 words	2.53	0.1509	0.0805	2.68	-0.0210	0.0728

Table 8: Sensitivity of learning gain to the short-tag word-count cutoff. Gains are learned-weighting Recall@5 minus short-tag aggregation Recall@5. One-tag gain is 0.0000 under all cutoffs in both datasets because one-tag cases provide no alternative for learned selection.

B Short-Tag Cutoff Sensitivity

Our main analyses define short tags as category-derived tags containing at most three words. This threshold is not intended as a theoretical claim that three words is universally optimal. It is an interpretable operationalization of compact, candidate-like category strings. To test whether our conclusions depend on this cutoff, we repeat the main analyses using thresholds of at most two, three, four, and five words.

Table 8 shows that the main conclusions are stable across the two-, three-, and four-word cutoffs. One-tag cases always yield zero gain, and images with at least three short tags consistently show the largest gain. The smaller dataset also preserves the positive diversity pattern under these stricter cutoffs, while the larger heterogeneous extension remains unstable, with raw diversity negative or near zero depending on the cutoff. The five-word cutoff weakens the results: in the smaller dataset, the $\geq$3-tag gain drops from 0.2364 under the three-word cutoff to 0.0805, and in the larger extension the two-tag group produces negative gain.

The five-domain sensitivity analysis leads to the same interpretation. Domain control does not make raw diversity a stable proxy: only the two-word cutoff yields even one domain with a stable positive diversity effect, while the three-, four-, and five-word cutoffs yield none. Short-tag count is more robust than diversity across domains, but this pattern also weakens as the cutoff becomes looser: tag count is more explanatory than diversity in all five domains under the two- and three-word cutoffs, in four domains under the four-word cutoff, and in only two domains under the five-word cutoff.

We therefore use the three-word cutoff as a practical default rather than an optimized threshold. It balances coverage and compactness: stricter cutoffs become sparse, while looser cutoffs admit longer compound category strings that are less likely to function as direct candidate supervision and do not make raw diversity more stable. The underlying principle is not that three words is universal, but that candidate supervision should be selected from compact category-derived tags that provide inspectable alternatives, rather than from all legacy categories indiscriminately.