Using Hierarchical Controlled Vocabularies to Understand CLIP Retrieval Failures in Historical Photo Collections

Ratan Sebastian[1,2] [0000−0002−8034−3382], Anett Hoppe[3,4,1] [0000−0002−1452−9509], Christoph Rippe[5] [0000−0003−3209−2504], and Ralph Ewerth[3,4,1,2] [0000−0003−0918−6297]

[1] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[2] L3S Research Center, Leibniz University Hannover, Germany
[3] Marburg University, Germany
[4] Hessian Center for Artificial Intelligence (hessian.AI), Marburg/Darmstadt, Germany
[5] Goethe University Frankfurt, University Library Frankfurt, Germany

Abstract. GLAM institutions (Galleries, Libraries, Archives, and Museums) organise image access using controlled vocabularies such as the Art and Architecture Thesaurus (AAT). For content-based image retrieval in these settings, vision-language models like CLIP are increasingly used, but their performance varies. This variation is known to relate to measures like concept abstraction and concept frequency. However, no prior work explains this variation in terms of the structural properties of vocabularies like the AAT that GLAM professionals already use. The AAT groups concepts into broad facets (*Objects*, *Activities*, *Agents*, etc.) and arranges terms hierarchically within them. In this paper, we ask whether two structural properties (root facet type and hierarchy depth) explain where CLIP retrieval succeeds and fails, and where fine-tuning helps. Across three historical photographic collections annotated with AAT terms, we examine visual coherence (whether a term's photographs cluster in CLIP's embedding space), text-image alignment (whether its label is near that cluster), and standard retrieval measures, which conflate the two. We find that visual coherence and text-image alignment are nearly uncorrelated across terms and jointly separate distinct failure modes. Terms whose photographs cluster tightly but whose label is distant from the cluster retrieve poorly in every collection, in two of three collections even worse than terms that fail on both metrics. We also show that while retrieval metrics do not correlate significantly with either structural property, root facet type does significantly separate categories with varying visual coherence. Finally, we find that fine-tuning improves retrieval overall, but its gains favour shallower terms in the hierarchy, where text-image alignment improves most, beyond what concept frequency explains. Together, these results make CLIP's failure more interpretable through the structure of a domain-specific vocabulary like the AAT.

Keywords: CLIP · image retrieval · visual coherence · Art and Architecture Thesaurus · historical photographic collections · domain adaptation

1 Introduction

GLAM institutions are deploying vision-language models like CLIP (Contrastive Language-Image Pre-training) for content-based image retrieval in photographic collections [15]. Unlike general web image retrieval, archival retrieval is organised around controlled vocabularies [10]. GLAM professionals assign controlled vocabulary terms to describe the content of photographs, and users rely on these terms as their primary search entry points [10]. CLIP enables zero-shot retrieval (querying without collection-specific training data), which is especially attractive in GLAM institutions where metadata is often incomplete, inconsistent, or missing entirely [15,3]. But the domain shift between CLIP's training domain which is mostly general web images and the historical archives domain leads to issues like the model not understanding domain specific terms or visual concepts [1].

Previous work has shown that CLIP performance varies by concept abstraction level [14,17]. Conceptual queries such as *family vacations* or *working class life* retrieve far fewer relevant results than descriptive ones such as *people chopping vegetables* [14]. It also varies by concept frequency [12,2]. Low-frequency terms such as *battle of the bulge*, *war bonds*, and *sailboats* are retrieved poorly by zero-shot CLIP as opposed to high frequency ones such as *aircraft*, *medical service* and *airstrike* in historical photographic collections [2]. These accounts explain failure in terms of WordNet position [14] and pretraining corpus frequency [2]. But no work so far has captured its variation in terms of a controlled vocabulary such as the Art and Architecture Thesaurus (AAT) [6] that GLAM professionals already use to organise image access. Further, previous work has characterized CLIP failure only in terms of retrieval metrics such as Recall@K and NDCG@K (Normalised Discounted Cumulative Gain), which confound whether photographs in a category cluster visually and whether the category label is recognised by CLIP's text encoder.

In this paper, we address these two gaps. We use the AAT, which organises concepts into broad facets and arranges terms hierarchically within them, to correlate CLIP behaviour (measured by visual coherence, text-image alignment and retrieval performance) with two structural properties (**root facet type** and **hierarchy depth**) across three historical photographic archives. Following Leemann et al. [11], we measure intra-category **visual coherence** (d_{mean}): the mean pairwise cosine similarity of CLIP image embeddings within a vocabulary category. A high value means photographs cluster tightly in CLIP's visual space. Separately, we measure **text-image alignment**: the cosine similarity between the CLIP text embedding of the category label and the centroid of its image embeddings. A high value means the label lands near the image cluster. Their joint distribution identifies three distinct sources of CLIP retrieval failure, detailed in Section 5.2: photographs that are visually incoherent, labels that land far from a coherent cluster, and labels that land near a cluster too diffuse for precise

retrieval. The distinction matters because it lets us ask which of these sources of difficulty fine-tuning strategies actually repair. We find, for instance, that fine-tuning influences mainly text-image alignment and not visual coherence. **Root facet types** are the AAT's [6] major concept classes which are ordered from abstract to concrete: *Associated Concepts* (abstract ideas), *Styles and Periods* (art historical styles and periods), *Agents* (people and organizations), *Activities* (actions and processes), *Materials* (physical substances), and *Objects* (tangible, human-made things). **Hierarchy depth** measures how many levels below the root facet a term sits: shallower terms name broad categories, deeper terms name specific concepts that are a kind of their broader context. These two properties correspond to the concreteness and specificity dimensions along which CLIP performance is known to vary, as discussed earlier in this section. To understand the stability of our results across photographic collections we apply our analysis to three of them from different contexts: colonial-era German photos [16]), early 20th-c. Catalan press photos [5], and FAIR Photos [19] which contain postwar Dutch press photos.

Concretely, our research questions are:

RQ1 Do AAT **root facet type** and **hierarchy depth** correlate with **visual coherence**, **text-image alignment**, and standard retrieval metrics (Recall@K and NDCG@K)?

RQ2 Do AAT **root facet type** and **hierarchy depth** correlate with fine-tuning improvement?

The following are some key findings from our study. The correlation between **visual coherence** and **text-image alignment** is close to zero, both across all collections and within each collection, meaning that these two factors that lead to retrieval failures are independent. For example, in the *Materials* facet photographs cluster tightly, yet labels are the worst-aligned. This independence property helps us decompose retrieval failure and fine-tuning successes to explain which of these two dimensions is most affected. We can see, for example, that in the zero-shot setting the structural properties of AAT terms are more correlated with **visual coherence** than overall retrieval performance. For instance, deeper AAT terms form tighter image clusters than shallow, broad ones, a relationship that holds up across collections, yet this same ranking by **hierarchy depth** shows essentially no relationship to NDCG@10. Interestingly, we see that terms whose photographs cluster tightly but whose label is far from that cluster retrieve worse than terms that are both diffuse and poorly aligned. Also, as previously stated, we can see that fine-tuning mainly improves text-image alignment, not visual coherence.

2 Related Work

We review CLIP-based retrieval in cultural heritage settings, image indexing research that grounds the AAT's structural properties as predictors of retrieval behaviour, and accounts of CLIP's known limits, which our work extends from linguistic and data properties to vocabulary structure.

2.1 CLIP in GLAM

Multimodal retrieval is changing how digital humanities research engages with photographic collections, enabling exploration and analysis at scales previously impractical [15]. CLIP retrieves via text queries without collection-specific training, letting GLAM institutions with incomplete, inconsistent, or absent metadata make collections discoverable without manual cataloguing [3]. Deploying generic CLIP in archival settings still has challenges. The visual domain shift between web-trained features and historical photographic material causes issues: halftone printing, physical degradation, and microfilm intermediaries produce visual characteristics generic CLIP has not encountered, leading to poor retrieval [1]. Specialised domain vocabulary also creates text-image alignment failures [20], and fine-grained categories require representations generic CLIP does not provide [4].

Fine-tuning on domain-specific data addresses these problems: adapting CLIP to historical archives with long-tailed distributions improves retrieval, with progressive unfreezing raising Recall@1 from 15% to 68% [2]. It remains unclear, however, whether these gains are uniform across low-frequency categories or concentrated in types meaningful to collection users, the gap this work addresses.

2.2 Controlled vocabularies for images

Image indexing in library and information science (LIS) organises image content along two dimensions: ontological type (what kind of thing is depicted) and abstraction level (how far a description is from a concrete visual referent). This understanding derives from three influential frameworks: Jörgensen's Pyramid of Visual Descriptors [8], which takes ontological category as its primary axis; the Visual Information Vocabulary [7], which formalises this into facets such as Natural Objects, Living Beings, and Abstract Concepts; and Westman [18], who approaches the same space from the user side, categorizing information needs as generic, specific, or abstract, and further into objects and activities. *Objects* have stable visual forms while *Activities* and abstract concepts do not.

The AAT [6] is an authority file maintained by the Getty Vocabulary Program, used at cataloguing time across GLAM institutions to assign controlled terms to collection objects, including photographs, promoting consistency and cross-collection retrieval. Its structure maps onto both dimensions identified above: its seven root facets (introduced in Section 1) encode ontological type, ordered from abstract to concrete, and within each facet the hierarchy encodes genus/species relationships, so depth measures conceptual specificity.

2.3 Limits of CLIP

CLIP [13] learns joint image-text representations from web image-caption pairs via a contrastive objective, retrieving via cosine similarity between a text query and image embeddings. Previous work shows a fivefold drop in Recall@1 from caption-style to conceptual queries [14], with WordNet hyponymy depth as

the primary predictor: concepts closer to the WordNet root are harder to retrieve. Concept frequency also affects retrieval performance: concepts underrepresented in web training data are retrieved less accurately, regardless of abstraction level [12,2]. Term visualness (whether a word evokes a mental image) predicts retrieval quality [17], and category visual coherence (the degree to which a category's photographs cluster in CLIP's embedding space) correlates with human judgements of concept coherence [11].

Across these accounts, CLIP failure is explained by linguistic properties (WordNet depth, visualness) or data properties (concept frequency), never by the structural properties of a vocabulary GLAM institutions already use, such as the AAT. We do so here: if CLIP's success and failure correlate with the AAT's structural divisions, GLAM professionals gain a way to anticipate retrieval problems using a framework they already understand.

3 Datasets

We study three historical photographic collections, each annotated with AAT terms and spanning different geographic regions, periods, and subject matter. All three are black-and-white photographs but differ in era, origin, and digitisation practice, so findings that hold across all three are unlikely to be artefacts of any one collection. Each is described below and summarised in Table 1.

3.1 University Library Frankfurt Collection from Colonial Contexts

This dataset draws from the Image Archive from Colonial Contexts, University Library Frankfurt [16] (henceforth abbreviated *GermanColonial*). The collection contains over 50,000 photographs of the German Colonial Society and subsequent colonial organisations. We created AAT links by mapping German content keywords from the original metadata to AAT term IDs. The metadata contains 8,124 unique keywords across 62,519 images. Matching proceeds in two stages: (1) exact case-insensitive matching against the AAT (193 keywords matched, covering 26.4% of images); (2) spaCy German lemmatization (`de_core_news_lg`), which maps inflected forms to their canonical base form and allows matching across morphological variants. Together the two stages yield 1,429 matched keywords (17.6% of unique keywords), covering 35,159 images (56.2% of the archive). The remaining 43.8% have no AAT match and are excluded. A manual check of a random sample of 50 matched keywords identified only 3 errors: one instance of metadata noise (keywords such as *s*) and two instances of ambiguous matches (e.g. *Schiff* matching both *ship* and *nave* when only one sense applies). While we know that the mapping procedure is not perfect, the low error rate is sufficient to test correlations across hundreds of keywords. The final linked datasets is dominated by *Objects* which comprise 81% of terms (Table 1).

Table 1. Collection overview and AAT root facet distribution. AAT terms are those with ≥ 20 images in that collection (Section 4). "—" ⇒ facet absent.

	GermanColonial	Sagarra	FAIR Photos
Photographs	34,882	6,274	17,605
AAT terms	378	111	85
Depth (median)	7	7	6
Depth range	3–13	3–13	1–11
Facet distribution (% of terms)			
Objects	81%	70%	38%
Agents	8%	16%	11%
Activities	2%	11%	40%
Assoc. Concepts	3%	3%	9%
Materials	3%	—	2%
Styles & Periods	2%	—	—
Physical Attributes	0%	—	—

3.2 Photographs from Catalan Photographer Josep Maria Sagarra

The Sagarra dataset [5] comprises 8,536 historical photographs taken by Josep Maria Sagarra, one of the most prominent Catalan press photographers of the first half of the twentieth century, held by the Ajuntament de Girona and retrieved via the Europeana API(henceforth abbreviated *Sagarra*). AAT identifiers are embedded directly in the Europeana metadata as Getty vocabulary URIs, requiring no mapping step and yielding near-complete coverage (99.7%). The 111 analysed terms span a more balanced facet distribution than GermanColonial, with a higher proportion of *Agents* (16%) and *Activities* (11%).

3.3 FAIRPhotos

FAIR Photos [19] is a cross-institutional collection of 18,143 historical photographs aggregated from multiple European cultural heritage repositories under open licenses. As with Sagarra, AAT identifiers originate from Getty vocabulary URIs embedded in the source metadata. FAIR Photos has the most diverse facet distribution of the three archives with *Activities* accounting for 40% of terms which is comparable to *Objects* (38%) making it a useful contrast to GermanColonial's object-dominated vocabulary.

4 Design of CLIP Analysis based on Vocabulary Structure

This section explains the extraction of AAT structural properties, CLIP fine-tuning and evaluation metrics (Figure 1).

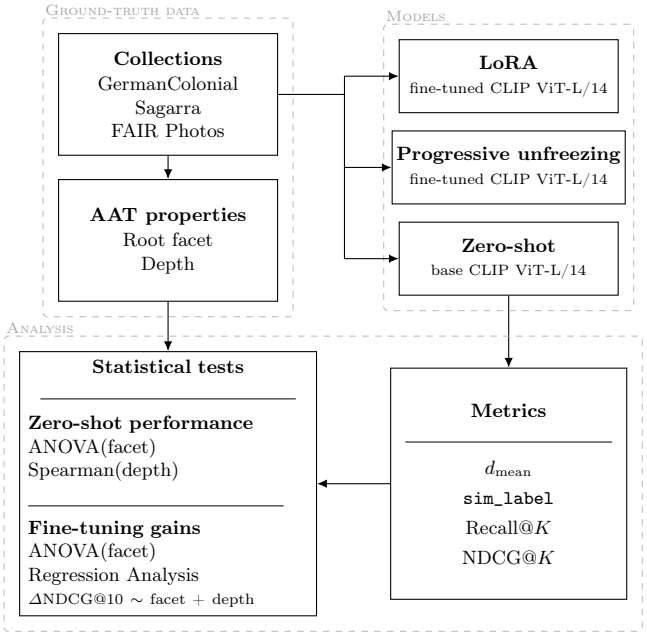

Fig. 1. Methodology overview. Ground-truth data (collections + AAT properties), three model conditions (zero-shot, LoRA, progressive unfreezing), and the analysis pipeline (metrics + statistical tests). Details in Section 4.

4.1 AAT structural properties

Each photograph is associated with a set of AAT terms. For Sagarra and FAIR Photos these come from existing metadata. For GermanColonial labels in the metadata are connected to the AAT as described in Section 3.1. For each term, we extract two structural properties: the **root facet** (one of: *Objects*, *Activities*, *Agents*, *Materials*, *Associated Concepts*, *Styles and Periods*, *Physical Attributes*, *Brand Names*) and the **depth** (number of levels below the root facet). **Root facet** aligns with the ontological type used in image classification systems [8], and **depth** aligns with term abstraction, a known factor in CLIP performance [14,17]. Depth ranges from 1 to 13 across the three collections, with the majority of terms at depths 5–9 and the distribution of facets is shown in Table 1. A term enters an analysis only if at least 20 images contribute to it: the zero-shot diagnostic analyses use terms with at least 20 images in the analysed collection, while the retrieval and fine-tuning-gain analyses use the terms covered by the held-out train/test split, which applies the same threshold to each term's images pooled across collections.

4.2 Fine-tuning CLIP

We evaluate two fine-tuning strategies applied to CLIP ViT-L/14 following previous work on fine-tuning CLIP for historical photographic collections [2]. **Low-**

Rank Adaptation (LoRA) ($r = 16$, $\alpha = 32$) applies low-rank adapters to all attention and feed-forward projection layers (`q_proj`, `v_proj`, `out_proj`, `fc1`, `fc2`) and the visual and text projection heads, training for five epochs with a symmetric contrastive objective. **Progressive unfreezing** trains in five phases, sequentially unfreezing transformer blocks from the top layer downward, with phase transitions driven by loss plateau detection, for approximately 45,000 steps total. Meta-parameters for both strategies (LoRA rank, learning rate, epoch count, plateau thresholds) were set via a small number of informal preliminary runs rather than a systematic search on a separate validation split, and chosen to match the configuration reported by prior work on fine-tuning CLIP for historical photographic collections [2]. During training, each photograph is paired with a text consisting of the term's preferred English label concatenated with a paraphrase of the term's AAT scope note. Since AAT scope notes often exceed CLIP's 77-token context limit, they were paraphrased into single short descriptive sentences using *Qwen3-8B*. The paraphrase is used only as training-time augmentation. At query time we always use the AAT's preferred English label matching the task setting in prior work [2]. Improvement due to fine-tuning (ΔNDCG@10) is computed as the difference between the fine-tuned model and the zero-shot baseline on the held out 20% test split.

4.3 Metrics

We compute four metrics per term. Two that diagnose the source of retrieval difficulty and two that measure end-to-end retrieval performance.

Visual coherence (d_{mean}) is the mean pairwise cosine similarity of image embeddings within a term [11]. Let C denote the set of n images assigned to an AAT term and $e_1, \ldots, e_n$ their CLIP image embeddings:

$$d_{\text{mean}}(C) = \frac{1}{\binom{n}{2}} \sum_{i<j} \cos(e_i, e_j)$$

For AAT terms with more than 300 photographs we use a random sample of 300 for computational efficiency.

Text-image centroid similarity (`sim_label`) is the cosine similarity between the CLIP text embedding of the term's preferred English label and the centroid of its image embeddings. The zero-shot diagnostic analyses compute both metrics over all of a term's images in the analysed collection; comparisons between fine-tuned models and the zero-shot baseline compute all metrics on the held-out test split.

Recall@K and **NDCG@K** follow [2], with the preferred English label as the query ranked against the shared pool of all test photographs. Both retrieval metrics are computed at $K \in \{1, 5, 10\}$, and we report NDCG@10 in the paper body, since it correlates with NDCG@5 ($\rho = 0.89$–0.93 across the zero-shot, LoRA, and progressive-unfreezing conditions) and NDCG@1 ($\rho = 0.61$–0.70)

as well as Recall ($\rho = 0.78$–0.80). Full results are given in the code repository accompanying this paper.[6]

Together, these four measures capture different aspects of retrieval difficulty. d_{mean} reflects whether photographs cluster visually, independent of the text encoder and of category size. `sim_label` reflects whether the category label reaches that cluster, isolating the text-side failure mode. Recall@K and NDCG@K measure end-to-end retrieval performance as a user would experience it, but conflate both failure modes and are sensitive to category size. The two diagnostic measures allow us to distinguish between a category that retrieves poorly because its photographs have no visual structure and one where the structure exists but the label does not reach it. Recall and NDCG also allow us to compare our results to previous work on CLIP retrieval in historical photographic collections [2] which only reports those metrics.

4.4 Statistical analysis

Facet and depth correlations. For zero-shot performance, one-way ANOVA tests whether root facet explains variance in d_{mean} and `sim_label`. Effect sizes are reported as η^2. Spearman rank correlation tests the relationship between depth and each metric. Since deeper terms tend to have fewer images, we additionally compute partial Spearman correlations controlling for log image count. All tests are run per collection and are two-tailed. To check the monotonicity assumption behind Spearman ρ, we fit depth + depth^2 OLS models per collection per metric (6 tests), giving little evidence of curvature (all $p \geq 0.24$ after correction). All reported p-values are unadjusted for multiple comparisons. Given the number of tests in Tables 2 and 3, we anchor conclusions on the pooled, archive-controlled models and on patterns consistent in direction across collections, and treat isolated per-collection significances as descriptive.

Fine-tuning gains. For each fine-tuned model we compute the per-term gain $\Delta m = m_{\text{fine-tuned}} - m_{\text{zero-shot}}$, where m is a metric value (NDCG@10, `sim_label`, or d_{mean}) for a given AAT term. Concept frequency n_f is the number of train-split photographs for a given term. We use $\log n_f$ throughout because concept frequency spans several orders of magnitude across AAT terms. Three analyses are run on ΔNDCG@10.

First, Spearman(ΔNDCG@10, $\log n_f$) tests whether concept frequency alone explains gains, following [2] who find that categories with more training photographs improve more after fine-tuning.

Second, one-way ANOVA(ΔNDCG@10 $\sim$ facet) with η^2 is run separately per collection to check whether the facet effect holds in each collection individually (an effect found only in one collection could reflect that collection's subject matter rather than a structural property of the AAT).

Third, ordinary least squares (OLS) regression ΔNDCG@10 $\sim \log n_f +$ depth + facet tests whether facet and depth explain variation in gains beyond

[6] https://github.com/rxvl-d/clip-aat-historical-photos-tpdl26

what concept frequency alone accounts for. This is necessary because depth, facet, and frequency are correlated, so a bivariate correlation with gains could reflect the frequency effect rather than a structural one.

5 Experimental Results

We present our results in three stages, following our two research questions. We first test whether root facet type and hierarchy depth correlate with visual coherence, text-image alignment, and end-to-end retrieval in the zero-shot setting (RQ1). We then decompose retrieval failure into the three modes this joint distribution defines. Finally, we test whether the same two properties predict fine-tuning improvement (RQ2).

5.1 Zero-shot visual coherence and text-image alignment

RQ1 asks whether AAT root facet type and hierarchy depth correlate with visual coherence, text-image alignment, and end-to-end retrieval. Across all three collections, in the zero-shot setting, the two structural properties predict visual coherence most strongly and text-image alignment and end-to-end retrieval less.

Visual coherence. Table 2 shows that root facet type correlates with d_{mean} in GermanColonial and Sagarra, but not in FAIR Photos, where the effect size is essentially zero ($\eta^2 = 0.008$). In GermanColonial, the facet effect is attributable entirely to *Activities*, a small facet ($n = 9$), and in Sagarra entirely to *Agents* ($n = 18$). Excluding either facet renders the remaining facets statistically indistinguishable ($p = 0.22$ and $p = 0.36$, respectively), matching the pattern observed across all facets in FAIR Photos, where no facet functions as a comparable outlier.

Table 3 shows that depth correlates positively with d_{mean} in GermanColonial ($\rho = +0.26$), FAIR Photos ($\rho = +0.27$), and the pooled analysis, but is absent in Sagarra ($\rho = -0.01$). Closer examination reveals that shallow terms in Sagarra are already clustering nearly as tightly as the deep ones (mean d_{mean} differs by only 0.005 between terms with the top and bottom three depths in the distribution). For GermanColonial and FAIR Photos the differences are 0.032 and 0.025 respectively. This could be a collection-specific effect since Sagarra's photographs are all press photos from a single photographer, while the other two collections are aggregated from multiple sources. There could also be a qualitative difference. The shallow terms in the Sagarra collections are mostly *Agents* (*workers*, *students*, *scholars*, *societies*) and a small cluster of terms relating to religion such as *religious holidays*, *religious art*, *religions*. Whereas in GermanColonial the shallow terms are mostly *Objects* (*portraits*, *sculpture*, *furniture*, *tools*), and in FAIR Photos mostly *Activities* (*advertising*, *exhibitions*, *elections*, *parades*).

We also examine these correlations for category-size effects but found that partial correlations controlling for log image count are essentially unchanged

Table 2. One-way ANOVA with root facet as predictor (F, p, η^2), for zero-shot metrics and fine-tuning gains (Δ). **Bold** $\Rightarrow p < 0.05$. **Pooled** $\Rightarrow$ All collections together.

		Pooled			GermanColonial			Sagarra			FAIR Photos		
Variable	Setting	F	p	η^2	F	p	η^2	F	p	η^2	F	p	η^2
d_{mean}	zero-shot	**2.76**	**0.012**	**0.029**	**4.44**	**<0.001**	**0.056**	**3.54**	**0.017**	**0.090**	0.21	0.890	0.008
	Δ LoRA	0.52	0.796	0.006	0.93	0.464	0.011	0.19	0.904	0.007	0.60	0.618	0.027
	Δ progressive	1.75	0.107	0.019	1.65	0.145	0.020	1.97	0.125	0.069	0.41	0.743	0.019
`sim_label`	zero-shot	0.74	0.620	0.008	0.78	0.561	0.010	1.51	0.216	0.041	1.27	0.289	0.046
	Δ LoRA	1.33	0.242	0.014	0.77	0.569	0.010	**4.94**	**0.003**	**0.156**	0.60	0.615	0.027
	Δ progressive	**2.72**	**0.013**	**0.029**	1.40	0.222	0.017	**7.74**	**<0.001**	**0.225**	0.73	0.539	0.032
NDCG@10	zero-shot	1.94	0.073	0.021	1.07	0.374	0.013	1.93	0.131	0.068	1.43	0.241	0.062
	Δ LoRA	1.78	0.101	0.019	**2.86**	**0.015**	**0.034**	0.30	0.828	0.011	0.08	0.970	0.004
	Δ progressive	1.89	0.080	0.020	2.08	0.067	0.025	2.30	0.084	0.079	1.62	0.194	0.070

(pooled $\rho = +0.23$, GermanColonial $+0.24$, FAIR Photos $+0.21$). Since GermanColonial contributes the most terms and is the most *Objects*-heavy collection (81% of terms against 70% and 38%), we additionally checked whether the pooled depth correlation is carried by *Objects* terms. It is not: pooled across collections, the correlation is stronger among non-*Objects* terms ($\rho = +0.33$) than among *Objects* terms ($\rho = +0.15$).

Text-image alignment. Both properties predict `sim_label` far more weakly than they predict d_{mean}. Facet type does not explain variance in `sim_label` in any collection (all $p \geq 0.21$). Depth correlates positively with `sim_label` in all three collections, but more weakly than for d_{mean} ($\rho \approx 0.07$–0.22). Interestingly, the pattern is the reverse of the coherence result. Sagarra is the only collection where depth reliably correlated with `sim_label`, rising from 0.206 (shallow) to 0.219 (mid) to 0.223 (deep), while GermanColonial (0.196 to 0.201) and FAIR Photos (0.216 to 0.214) are relatively flat. In contrast the coherence result above, Sagarra's photographs are non-uniformly aligned, while GermanColonial's and FAIR Photos' are non-uniformly coherent but uniformly aligned. No collection shows a significant depth correlation for both metrics at once. Visual coherence and text-image alignment thus need to be examined together, which we do in Section 5.2.

Retrieval metrics. Neither structural property carries over to end-to-end zero-shot retrieval. The facet effect on zero-shot NDCG@10 is not significant in any collection (Tables 2). Depth is uncorrelated in the pooled analysis and significant only in Sagarra (Table 3). The Sagarra exception is essentially the same effect as from `sim_label`. Controlling for `sim_label`, the depth–NDCG@10 correlation becomes insignificant ($r = 0.14$, $p = 0.19$), confirming it is inherited from the alignment effect already reported above rather than a separate structural relationship.

5.2 Analysis of failure modes

The joint distribution of d_{mean} and `sim_label` shows the three failure modes (Fig. 2). Each term falls into one of four quadrants, split at the per-collection

Table 3. Spearman ρ between AAT hierarchy depth and each correlated variable. In the case of the fine-tuning strategies, the variable is the Δ over the zero-shot metrics. **Bold** $\Rightarrow p < 0.05$.

		Pooled		GermanColonial		Sagarra		FAIR Photos	
Variable	Setting	ρ	p	ρ	p	ρ	p	ρ	p
d_{mean}	zero-shot	**+0.249**	**<0.001**	**+0.261**	**<0.001**	-0.007	0.938	**+0.269**	**0.013**
	Δ LoRA	+0.082	0.050	+0.024	0.634	+0.065	0.558	-0.116	0.336
	Δ progressive	**+0.150**	**<0.001**	+0.025	0.616	+0.205	0.061	+0.010	0.935
`sim_label`	zero-shot	+0.037	0.390	+0.069	0.181	**+0.218**	**0.021**	+0.190	0.082
	Δ LoRA	**-0.164**	**<0.001**	**-0.117**	**0.018**	**-0.221**	**0.043**	-0.168	0.161
	Δ progressive	**-0.166**	**<0.001**	**-0.098**	**0.048**	**-0.239**	**0.028**	-0.224	0.060
NDCG@10	zero-shot	+0.003	0.936	+0.016	0.747	**+0.244**	**0.025**	+0.154	0.200
	Δ LoRA	**-0.100**	**0.018**	-0.067	0.176	+0.110	0.321	-0.079	0.513
	Δ progressive	**-0.134**	**0.001**	-0.004	0.931	-0.086	0.434	**-0.252**	**0.034**

medians of the two metrics, abbreviated by whether d_{mean} and `sim_label` are high or low (HH, HL, LH, LL). We would expect retrieval success to occur when a term's images form a tight cluster and the label lands near it (i.e. the HH quadrant). The other three are the three failure modes:

High d_{mean}, low `sim_label` (HL). Images cluster visually but the label lands far from the cluster. This quadrant contains both domain-specific object vocabulary (*featherwork*, *mineheads*, *fishing tackle*) and activity terms (*religious holidays*). The *Materials* facet illustrates this pattern at the facet level. Pooled across collections, its zero-shot d_{mean} (0.716) is the highest of any facet, while its `sim_label` (0.192) is the lowest.

Low d_{mean}, high `sim_label` (LH). The label lands near the image centroid but the photographs scatter. This quadrant contains broad agent categories (*generals*, *military personnel*, *studio portraits*) and activity terms (*exhibitions*, *automobile racing*). The problem here is the visual heterogeneity of the category rather than the text encoder.

Low d_{mean}, low `sim_label` (LL). Neither photographs cluster nor does the label reach them. The LL quadrant captures AAT terms ranging from highly domain-specific object vocabulary (*drumheads*) to broad activity categories (*elections*, *societies*, *broadcasting*).

The two metrics are nearly uncorrelated pooled across collections (Spearman $\rho = -0.03$, $p = 0.53$) and at most weakly correlated within them ($\rho \leq +0.25$). Zero-shot NDCG@10 differs by quadrant. In GermanColonial and Sagarra, HL retrieves worst of all four quadrants (0.027 and 0.034), below LL (0.051 and 0.044). In FAIR Photos, HL (0.162) is slightly better than LL (0.114). LH is worse than HH in Sagarra (0.071 vs. 0.233) and FAIR Photos (0.315 vs. 0.444), but not in GermanColonial (0.126 vs. 0.144) (Full per-quadrant counts and NDCG@10 are in the previously linked repository).

That a category can retrieve worse with one intact metric can be explained by the retrieval mechanism. A tight cluster that sits far from its label misses the top ranks uniformly, while the scattered photographs of an LL term still place a

few images near the label by chance. A visually coherent but unaligned cluster thus manifests as a wrong ranking as opposed to an accidentally right one.

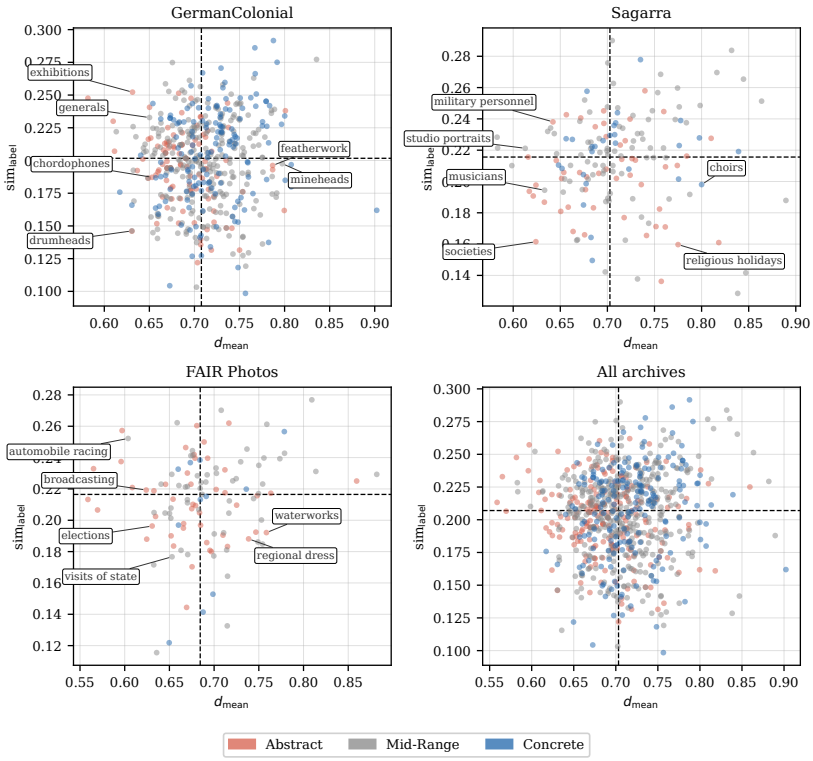

Fig. 2. Joint distribution of d_{mean} and `sim_label`. *Abstract* (depth ≤ 5), *Mid-Range* (depth 6–8), *Concrete* (depth ≥ 9). Dashed lines mark the per-collection medians. Annotated examples illustrate the three failure modes described in Section 5.2.

5.3 Fine-tuning gains

Tables 2 and 3 report the facet and depth effects on NDCG@10 gains under fine-tuning. Zero-shot NDCG@10 averages 0.114 across 563 AAT terms. LoRA fine-tuning raises this to 0.252 ($\Delta = +0.138$) and progressive unfreezing to 0.259 ($\Delta = +0.145$).

Alignment gains vs. coherence gains. Depth predicts gains in `sim_label` but not in d_{mean}. Shallower AAT terms improve alignment more under both LoRA ($\rho = -0.164$, $p < 0.001$) and progressive unfreezing ($\rho = -0.166$, $p < 0.001$). Facet explains no coherence-gain variance in any collection ($p \geq 0.107$ throughout). Depth's coherence-gain correlation is significant only across collections and only for progressive unfreezing ($\rho = +0.150$). Fine-tuning thus mostly improves

text-image alignment where it was weakest (i.e. in shallower terms), with no comparable predictable effect on visual coherence.

Frequency vs. facet as predictors. Concept frequency ($\log n_f$) correlates with NDCG@10 gains for LoRA ($\rho = +0.217$, $p < 0.001$) but not for progressive unfreezing ($\rho = +0.027$, $p = 0.529$), replicating [2] only partially. Controlling for frequency in an OLS model ($\Delta\,\text{NDCG@10} \sim \log n_f + \text{depth} + \text{facet}$), the facet joint F-test is significant for both LoRA ($F = 5.67$, $p < 0.001$) and progressive unfreezing ($F = 8.55$, $p < 0.001$; $R^2 = 0.098$ and 0.104 respectively). This means that facet explains gain variance that frequency alone does not. This holds specifically for alignment. The facet effect on Δ `sim_label` is significant when pooled across collections for progressive unfreezing ($p = 0.013$) and within Sagarra under both strategies (Table 2). The size of the effect is collection-dependent, the facet × collection interaction is significant for progressive unfreezing ($F = 2.25$, $p = 0.029$). Controlling for collection weakens the facet effect on ΔNDCG@10 to marginal for both strategies ($F = 1.78$, $p = 0.10$; $F = 1.89$, $p = 0.08$).

6 Discussion

The central finding of this study is that visual coherence and text-image alignment are nearly uncorrelated across AAT terms (Section 5.2), so a retrieval score such as NDCG@10, which depends on both, does not show which of the two failed. In answer to RQ1, root facet and depth correlate with visual coherence, each significantly in two of three collections and in the pooled analysis. Their correlation with text-image alignment is weak and inconsistent, and neither property correlates with retrieval performance at all. This follows from the near-zero correlation between the two metrics. A structural property that predicts only one of them shows little relationship with a score that depends on both.

Low alignment is the more consistent failure factor across collections: HL terms rank worst or second-worst of the four quadrants everywhere, often below LL terms that fail on both metrics, since a mislabeled tight cluster ranks consistently low (Section 5.2). Earlier accounts that explain variance in Recall@K through abstraction level [14] or concept frequency [12,2] therefore mix failure modes with distinct causes and, as argued below, distinct remedies.

The facet effect anticipated by LIS frameworks that treat ontological type as a primary axis of image content [8,18] is not strongly present here, and holds in each collection only through a single facet (*Activities* in GermanColonial, *Agents* in Sagarra; Section 5.1). The depth effect is more robust, but its absence in Sagarra suggests visual coherence is determined jointly by vocabulary granularity and collection composition, not by the vocabulary alone.

Text-image alignment shows the opposite pattern. In the zero-shot setting, neither facet nor depth predicts it, yet alignment is what fine-tuning changes, improving most for shallow terms where it was weakest, without affecting coherence. Fine-tuning with an image-text contrastive objective therefore addresses the coherent-but-misaligned (HL) failure mode but not visual scatter (LH).

At the observed effect sizes, neither facet nor depth reliably predicts retrieval behaviour for an individual AAT term. They can instead point to parts of a vocabulary that are likely to contain retrieval problems (e.g. shallow, broad terms), whose terms are then evaluated with the two diagnostic metrics, and the resulting failure mode indicates a response. For visually incoherent categories, retrieval-agnostic access (faceted browsing, seed-image similarity search) avoids relying on a label to retrieve a cluster that does not exist. For coherent but misaligned categories, domain adaptation can move the label embedding toward the existing cluster, as the fine-tuning results indicate, and query expansion may identify phrasings that lie closer to the image cluster than the archival term. For categories whose label is recognised but whose photographs scatter, the standard image-text contrastive loss is unlikely to help, since it pulls images toward a label embedding that is already near the centroid without reducing scatter. A supervised contrastive objective over image embeddings [9], treating photographs of the same AAT term as positives, would instead reduce intra-category scatter and increase d_{mean}.

Limitations and Future Work. As the effect sizes above show, the structural properties explain only modest variance, and the OLS models for gain prediction reach $R^2 \approx 0.10$, meaning most of the variation across AAT terms remains unaccounted for. The GermanColonial links are automatically derived and cover 56% of images and 17.6% of the collection's keyword vocabulary. The mapped subset is skewed toward concepts with clean German-to-AAT string matches, which plausibly also over-represents concepts in CLIP's training distribution. We plan to investigate the suggested practical responses to each failure mode, and whether they are effective in improving retrieval.

7 Conclusion

We investigated whether AAT root facet type and hierarchy depth explain where CLIP retrieval fails and where fine-tuning helps, across three historical photographic collections. Visual coherence and text-image alignment are nearly uncorrelated across terms, so a single retrieval score conflates failure modes with distinct causes, and the worst-performing terms are often those that are coherent but misaligned rather than those failing on both metrics. Vocabulary structure predicts coherence but not alignment or zero-shot retrieval, while fine-tuning improves alignment for shallow, broad terms and leaves coherence largely unchanged. These results point toward failure-mode-specific remedies: alignment fixes such as query expansion for coherent-but-misaligned terms and contrastive training for visually scattered ones, as the next step toward improving retrieval.

Acknowledgments. This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), project number: 531013489 (VABiKo).

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Aissi, M., Giardinetti, M., Bloch, I., et al.: Computer vision and halftone visual culture: improving similarity search for historical photographs. Multimedia Tools and Applications **84**, 39451–39471 (2025). https://doi.org/10.1007/s11042-025-20855-6
2. Alijani, F., Late, E., Kumpulainen, S.: Historyclip: Adaptive multi-modal retrieval of imbalanced long-tailed archival data. In: Balke, W., Golub, K., Manolopoulos, Y., Stefanidis, K., Zhang, Z. (eds.) Linking Theory and Practice of Digital Libraries - 29th International Conference on Theory and Practice of Digital Libraries, TPDL 2025, Tampere, Finland, September 23-26, 2025, Proceedings. pp. 245–262. Lecture Notes in Computer Science, Springer (2025), https://doi.org/10.1007/978-3-032-05409-8_15
3. Aske, K., Giardinetti, M.: (mis)matching metadata: Improving accessibility in digital visual archives through the eycon project. Journal on Computing and Cultural Heritage **16**(4) (Nov 2023), https://doi.org/10.1145/3594726
4. Balauca, A.A., Paudel, D.P., Toutanova, K., Van Gool, L.: Taming clip for fine-grained and structured visual understanding of museum exhibits. In: Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part LXXVI. p. 377–394. Springer-Verlag, Berlin, Heidelberg (2024), https://doi.org/10.1007/978-3-031-73116-7_22
5. Europeana: Over 6,700 Photographs from Catalan Photographer Josep Maria Sagarra (2024), https://pro.europeana.eu/data/over-6-700-photographs-from-catalan-photographer-josep-maria-sagarra
6. Getty Vocabulary Program: Art & architecture thesaurus. https://www.getty.edu/research/tools/vocabularies/aat/ (2024)
7. Jörgensen, C.: The visual indexing vocabulary: Developing a thesaurus for indexing images across diverse domains. Proceedings of the American Society for Information Science and Technology **41**(1), 287–293 (2004), https://asistdl.onlinelibrary.wiley.com/doi/abs/10.1002/meet.1450410134
8. Jörgensen, C., Jaimes, A., Benitez, A.B., Chang, S.F.: A conceptual framework and empirical research for classifying visual descriptors. Journal of the American Society for Information Science and Technology **52**(11), 938–947 (2001), https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.1161
9. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., Lin, H. (eds.) Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual (2020), https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html
10. Late, E., Ruotsalainen, H., Kumpulainen, S.: Image searching in an open photograph archive: search tactics and faced barriers in historical research. Int. J. Digit. Libr. **25**(4), 715–728 (Jan 2024), https://doi.org/10.1007/s00799-023-00390-1
11. Leemann, T., Rong, Y., Kraft, S., Kasneci, E., Kasneci, G.: Coherence evaluation of visual concepts with objects and language. In: ICLR Workshop on Objects, Structure and Causality (2022), https://openreview.net/forum?id=rRzSSS_Ucg9
12. Parashar, S., Lin, Z., Liu, T., Dong, X., Li, Y., Ramanan, D., Caverlee, J., Kong, S.: The neglected tails in vision-language models. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12988–12997 (2024), https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.01234

13. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (2021), http://proceedings.mlr.press/v139/radford21a.html
14. Rodriguez, J.M., Tavassoli, N., Levy, E., Lederman, G.S., Sivov, D., Lissandrini, M., Mottin, D.: Does the performance of text-to-image retrieval models generalize beyond captions-as-a-query? In: European Conference on Information Retrieval (2024), https://doi.org/10.1007/978-3-031-56066-8_15
15. Smits, T., Wevers, M.: A multimodal turn in digital humanities. using contrastive machine learning models to explore, enrich, and analyze digital visual historical collections. Digital Scholarship in the Humanities **38**(3), 1267–1280 (09 2023), https://doi.org/10.1093/llc/fqad008
16. University Library Frankfurt: Image Archive from Colonial Contexts, https://sammlungen.ub.uni-frankfurt.de/kolonialesbildarchiv
17. Verma, G., Rossi, R.A., Tensmeyer, C., Gu, J., Nenkova, A.: Learning the visualness of text using large vision-language models. In: Conference on Empirical Methods in Natural Language Processing (2023), https://doi.org/10.18653/v1/2023.emnlp-main.147
18. Westman, S.: Image Users' Needs and Searching Behaviour, chap. 4, pp. 63–83. John Wiley & Sons, Ltd (2009), https://doi.org/10.1002/9780470033647.ch4
19. van Wissen, L., Vriend, N., Nijssen, A., Vereecken, L., den Engelse, M.: FAIR Photos - CLARIAH FAIR Data Call 2023 (2024), https://doi.org/10.17026/SS/7VD3ME
20. Yuan, J., Zhang, J., Wu, F., Lu, H., Lu, D., Wang, Q.: Towards cross-modal retrieval in chinese cultural heritage documents: Dataset and solution. In: Yin, X., Karatzas, D., Lopresti, D. (eds.) Document Analysis and Recognition – ICDAR 2025. Lecture Notes in Computer Science, vol. 16026. Springer, Cham (2026). https://doi.org/10.1007/978-3-032-04627-7_33