ARE DICTIONARY LEARNING METRICS FOR NEURAL CONCEPTS INTERCHANGEABLE PROXIES FOR HUMAN-INTERPRETABILITY?

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept-based methods aim to make neural networks more interpretable by representing their behavior in terms of human-meaningful concepts. As learned concept representations have become more common, recent work has proposed reconstruction fidelity, sparsity, *Representation Diversity* and *Stability* metrics to evaluate their quality. However, it remains unclear whether such metrics are actually informative of how comprehensible learned concept spaces are to humans, a prerequisite for increasing neural network interpretability. We study this question empirically by adopting an activation-space factorization view of learned concept representations, following recent work that frames concept extraction as dictionary learning. We then test whether dictionary learning metrics of the resulting concept spaces are associated with human evaluation of those spaces through a study with four complementary tasks and five measures of concept comprehension, involving 351 participants and yielding 5,616 individual responses. Our results show that different metrics exhibit substantially different relationships to human evaluation across tasks, and that these relationships are robust to the influence of individual classes. While reconstruction-based and stability metrics show no consistent connection to human judgments, *Gini Sparsity* and *Representation Diversity* display associations with linguistic comprehension measures for the studied setting. Overall, standard dictionary-learning metrics do not provide sufficient evidence for the evaluation of human-comprehensible concept spaces and should not be treated as interchangeable proxies for interpretability. However, some of them can be used as candidate indicators thereof in a systematic, human-validated analysis.

1 INTRODUCTION

Concept-based explanations aim to describe model behavior in terms of human-meaningful concepts, rather than only in terms of low-level input features or individual activations [Kim et al., 2017; Ghorbani et al., 2019; Koh et al., 2020]. This has motivated methods that extract concepts from trained models for explanation, intervention and model inspection. Because these methods are typically proposed as tools for interpretability, their value depends not only on mathematical properties of the learned concept space, but also on whether the extracted concept spaces are comprehensible for people.

A common view of learned concept extraction casts it as representation factorization, with dictionary learning (DL) as one prominent instance [Fel et al., 2023a]. This view directly motivates quantitative evaluations of concept spaces through structural properties of the resulting factorization. These metrics are valuable because they allow comparing different concept extraction approaches in a systematic and scalable way. However, prior work has emphasized that interpretability cannot be inferred from representation structure alone [Marconato et al., 2023; Ramaswamy et al., 2023; Stammer et al., 2024]. Better scores on such metrics are not sufficient if the resulting concept spaces are not meaningful to humans, leaving open whether these metrics reflect human comprehensibility. This validation gap is particularly important as concept-based methods move from predefined human annotations toward learned concept spaces derived from model representations [Yuksekgonul et al., 2023; Oikarinen et al., 2023; Schrodi et al., 2025]. We therefore take learned concept spaces, rather than individual concepts or specific explanation methods, as the object of evaluation. This raises the central empirical

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

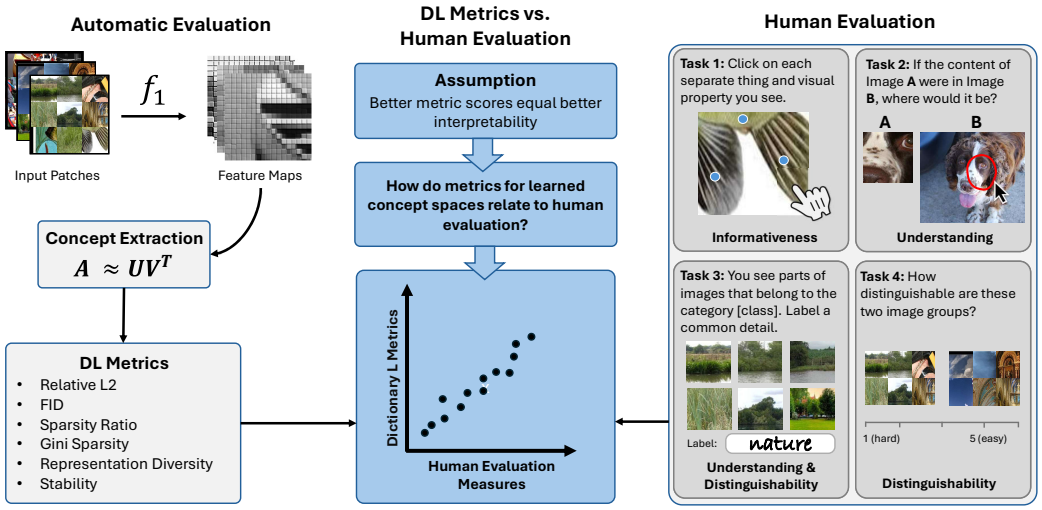

Figure 1: **Evaluation Overview.** Concept spaces are extracted from image patches and evaluated structurally via DL metrics *(left)* and by humans across four tasks *(right)*. Multiple correlation analyses investigate the relationship between the DL metrics and human comprehension *(center)*.

question we study: **to what extent are dictionary learning metrics for automatically extracted concept spaces aligned with human evaluation measures of concept-space comprehensibility?**

We study this question empirically in an activation-factorization setting, using concept spaces extracted from trained ResNet representations via non-negative matrix factorization (NMF) [Lee & Seung, 1999]. Because human comprehensibility is not a single observable quantity, we evaluate it through four complementary tasks (see Figure 1) that present learned concepts at different granularity and yield five behavioral measures. These measures capture whether concept spaces are informative, understandable, and distinguishable to human observers. In a study with 351 participants and 5,616 individual responses, we compare these human measures with metrics capturing reconstruction fidelity, sparsity, *Representation Diversity*, and *Stability*. This design allows us to test whether structural metrics align with human evaluations of concept-space comprehensibility, and whether different metric families relate to different aspects of human concept-space comprehensibility. In summary, our contributions are the following:

- We present, to our knowledge, the first systematic human study of whether dictionary learning metrics for concept-space extraction align with human evaluation measures of concept-space comprehensibility.
- We introduce four complementary human-evaluation tasks yielding five behavioral measures that probe distinct aspects of concept-space comprehensibility, including informativeness, understandability, and distinguishability.
- We show that different metrics exhibit substantially different relationships with human evaluation measures. Notably, *Gini Sparsity* and *Representation Diversity* may act as supportive indicators of concept-space comprehensibility.
- We argue that dictionary learning metrics do not provide sufficient evidence for the evaluation of human-comprehensible concept spaces, and should not be treated as interchangeable proxies for interpretability.

2 Related Work

Our work lies at the intersection of concept-based interpretability, automatic evaluation of learned concept spaces, and human-centered evaluation of explanation comprehensibility. We focus on whether DL metrics for learned concept spaces reflect human comprehension.

Concept-based approaches. Concept-based methods aim to explain model behavior using higher-level concepts rather than individual input features or activations. Early work introduced concept activation vectors to probe whether user-specified concepts are represented in trained models [Kim et al., 2017], while related post-hoc approaches analyzed learned representations by aligning individual units with semantic concepts or by extracting concepts through clustering or matrix factorization

[Bau et al., 2017; Ghorbani et al., 2019; Fel et al., 2023b]. A parallel line of work imposed concept structure directly in the predictive model through concept bottlenecks, where predictions are mediated by human-defined concept variables [Koh et al., 2020]. More recent post-hoc and label-free variants reduce the reliance on dense concept annotations by constructing concept bottlenecks from pretrained representations, language-derived concept sets, or unsupervised concept discovery [Yuksekgonul et al., 2023; Oikarinen et al., 2023; Schrodi et al., 2025]. These developments blur the distinction between post hoc concept extraction and bottleneck-style models at the level of learned representations. More specifically, while concept-based approaches differ in how learned concepts are used, they increasingly share a common representational object: learned concept-spaces extracted from intermediate model activations.

Automatic evaluation of learned concept spaces. As concept representations are increasingly learned rather than fully specified by humans, their evaluation has become a central methodological issue. One influential perspective casts automatic concept extraction as a dictionary learning problem, in which activations are approximated by a set of learned concept directions and input-specific coefficients [Fel et al., 2023a]. This view motivates structural metrics of the resulting factorization, such as reconstruction fidelity, sparsity, stability, and distributional preservation. Complementary to dictionary learning perspectives, Rocchi-Henry et al. [2025] relate concept bottleneck models and sparse autoencoders through a shared geometric view of learned concept directions, and compare concept structures using cone-based geometric criteria. However, such automatic metrics characterize technical properties of learned representations. They do not by themselves establish that the resulting concepts are meaningful or usable to humans. Indeed, prior work has shown that learned concept representations can encode unintended information, and that metrics imported from adjacent areas such as disentanglement may not transfer cleanly to concept learning [Espinosa Zarlenga et al., 2023]. Our work builds on these automatic evaluation perspectives, but asks a different question about whether quantitative metrics used to evaluate concept spaces are empirically associated with human evaluations of concept-space comprehensibility.

Human-centered perspectives on concept-space comprehensibility. Concept-based explanations are often motivated by their apparent human interpretability, but several works have emphasized that interpretability cannot be inferred from representation structure alone. From a conceptual and formal perspective, Marconato et al. [2023] argue that human-interpretable representation learning should explicitly account for the observer and their available concept vocabulary. In the context of predefined concept-based explanations, Ramaswamy et al. [2023] show that explanation usefulness depends strongly on concept and dataset design choices, and that humans may only be able to use a limited number of concepts effectively. Related human studies in prototype-based explanation settings similarly demonstrate that explanatory units that appear meaningful by construction or by automatic criteria still require behavioral validation, including in hierarchical or automatically discovered concept spaces [Davoodi et al., 2023; Parchami-Araghi et al., 2026; Hill et al., 2026]. More broadly, recent work on human-AI concept alignment frames shared conceptual structure between models and people as a prerequisite for interpretable interaction [Rane et al., 2024]. Together, these studies motivate evaluating concept explanations with human observers, but they do not directly test whether dictionary learning metrics for learned concept spaces predict human comprehension. Relative to this literature, our work targets the concept-space-level validation problem. We ask whether automatic structural metrics used to evaluate learned concept spaces align with human evaluation measures of those same spaces. Unlike prior work that uses human studies to assess specific explanation methods, modeling choices, or visualization strategies, we examine the empirical relationship between dictionary-learning-based metrics and human concept-space comprehensibility across multiple tasks and behavioral measures.

3 Problem Formulation

We formulate our study as an empirical assessment of whether quantitative metrics that capture structural properties of learned concept spaces are informative about human comprehensibility. In the following, we formalize the concept extraction setting, introduce the metrics, and define the evaluation criteria used to assess the human comprehensibility of concept spaces.

3.1 ACTIVATION-SPACE CONCEPT FACTORIZATION

We consider a dataset of images $\mathcal{X} = \{x^{(i)}\}_{i=1}^{N}$ with corresponding labels $\mathcal{Y} = \{y^{(i)}\}_{i=1}^{N}$. Each image $x^{(i)}$ is partitioned into n patches $p_1^{(i)}, \ldots, p_n^{(i)}$, which a patch encoder f maps to activation vectors $a_j^{(i)} = f\left(p_j^{(i)}\right) \in \mathbb{R}^d$. Stacking all patch activations yields the matrix $A \in \mathbb{R}^{Nn \times d}$, where each row corresponds to one patch-level activation. The labels $\mathcal{Y}$ are not used in the factorization itself, but are retained for subsequent class-level analyses and evaluation.

Following prior work, we adopt a dictionary learning view of concept representation [Fel et al., 2023a]. In line with activation-factorization approaches to concept-based explanation in vision [Fel et al., 2023b], we model patch activations through the factorization

$$A \approx UV^\top, \qquad U \in \mathbb{R}^{Nn \times K}, \quad V \in \mathbb{R}^{d \times K}.$$

Here, K denotes the number of latent factors, which we interpret as concepts. The m-th row of U, denoted $u_m \in \mathbb{R}^K$, gives the concept coefficients of the corresponding patch, while the k-th column of V represents the associated concept direction in activation space. Thus, each patch is represented as a point in a K-dimensional concept space, where each coordinate quantifies the contribution of one latent factor to the patch representation.

Because the factorization is defined at the patch level, these concept coefficients can be mapped back to specific image regions in the original dataset. Accordingly, for each concept $c \in \{1, \ldots, K\}$, one may identify a representative set of patches $\tilde{P}_c$ from patches with the largest coefficients in dimension c. This provides a direct link between the learned concept space and the visual domain, and enables inspection of the visual patterns associated with each latent factor.

This factorization-based representation defines the object of study in the remainder of this section. We next introduce dictionary learning metrics that quantify structural and reconstructive properties of such concept representations.

3.2 DICTIONARY LEARNING METRICS

We evaluate each learned concept space with metrics motivated by the dictionary learning formulation of concept extraction. Building on Fel et al. [2023a], we use *Relative L2* (i.e., pointwise reconstruction error), *FID* (i.e., distributional discrepancy between original and reconstructed patch activations), *Sparsity Ratio* (i.e., fraction of inactive coefficients), and *Stability* (i.e., consistency across repeated factorization runs). However, we omit their out-of-distribution metric, which concerns distribution-shift robustness rather than in-distribution human comprehensibility. We additionally include *Gini Sparsity*, a continuous measure of coefficient concentration satisfying sparsity criteria [Hurley & Rickard, 2009], and *Representation Diversity*, which assesses redundancy among learned concept directions [Cogswell et al., 2015]. Both are not specific to that metric protocol but are naturally defined for factorized concept spaces. We report formal definitions, aggregation details, and implementation choices in Appendix A.

3.3 COMPREHENSIBILITY OF CONCEPT SPACES

In this study, we aim to compare the metrics described in Section 3.2 with a human evaluation of concept spaces. The goal of our analysis is to find out whether human comprehension aligns with the concept space quality proposed by the metrics. We assume that a high alignment of human evaluation measures with a metric indicates a good comprehensibility of the concept space and furthermore its predictive power for human interpretability of the model output.

Although encoding higher-level semantics, concept spaces are still abstract multi-dimensional constructs and not intuitive for humans. To assess whether concept spaces are in fact comprehensible for humans, we first require measurable properties that correspond to human comprehension factors. Our five hypotheses **[H]** are based on the following properties of each concept space:

Informativeness of its concepts. For humans, the registration of single perceptual features guided by focused attention is one important building block for object recognition [Treisman & Gelade, 1980]. Due to this analytic part-based processing, we hypothesize that human comprehension of extracted concepts depends on a sufficient amount of such detectable properties. If predictive for

interpretability, we expect that metrics are correlated with the number of distinct entities and visual properties that humans can identify among the concepts in a concept space **[H1a]**.

Understandability of its concepts. This is a common measure in evaluations of AI explanations [Poeta et al., 2025] and builds on the idea that visual perception can go hand in hand with mapping stimuli to mental conceptual categories [Treisman & Gelade, 1980]. By understandability, we mean how well and consistently people can infer the meaning of all the concepts generated for each class of objects. If predictive for interpretability, we expect that metrics are correlated with understandability of the concepts in a concept space, if tested visually **[H2a]** as well as linguistically **[H2b]**.

Distinguishability of its different concepts. A concept space is defined not only by its concepts, but also by the spatial relationships between them. We use this property to measure how well people can tell the concepts apart from one another, reflecting the semantic similarity of their mental representations. This is supported by the idea that mental representations of objects contain properties that are used for discrimination [Pohl et al., 2026]. If predictive for interpretability, we expect that metrics are correlated with humans' ability to distinguish between different concepts in a concept space, if tested visually **[H3a]** and also linguistically **[H3b]**.

We introduce evaluation measures for each of these properties in Section 4. Overall, our global hypothesis can be summarized as follows: the six dictionary learning metrics each align with the five human evaluation measures. Specifically, we hypothesize that reconstruction loss metrics are negatively correlated and the others are positively correlated with human comprehensibility measures.

4 Human Evaluation Study

In this section, we assess whether dictionary-learning-based metrics align with human evaluation of concept spaces through five complementary evaluation measures. We first describe the experimental setup and then, for each of the five hypotheses motivated in the previous section, we introduce the relevant task together with the corresponding results.

4.1 Experimental Setup

Dataset. To keep the complexity of our study manageable, we use Imagenette [Howard, 2019], a ten-class subset of ImageNet [Deng et al., 2009]. To improve statistical power for our correlation analysis, we additionally include the two randomly selected ImageNet classes airliner and canoe, resulting in a total of twelve classes. Details concerning how we extract concepts from this data as the foundation for our evaluation from both technical and human perspectives, as well as, more information on participants and responses can be found in Appendix C.

Study Design. As described in Section 3.3, we operationalize our hypotheses by four complementary tasks that aim to evaluate human concept comprehension using five different measures. The participants' responses to these tasks serve as human reference measures for correlation comparisons with the dictionary learning metrics. To capture human comprehension at multiple levels of granularity, our tasks vary in how they represent a concept. Task 1 and 2 represent a concept by its most highly activated patch (top_1), while Task 3 and 4 work with a representative group $\tilde{P}_c$ of activated patches (top_6). Screenshots of the corresponding response interfaces can be found in Appendix K.

4.2 Hypothesis 1: Informativeness

To assess the richness of information within a concept we employ a click frequency task that is outlined in the following section.

4.2.1 Evaluation Measure 1: Click Frequency

Task Description. For all concepts in a concept space, participants are shown its top_1 most representative patch. As visualized in Figure 1, participants are asked to respond to the prompt: *Click on each separate thing and visual property you see.".* The responses serve as a measure for the first hypothesis: the number of distinct entities and visual properties within a concept. The number of clicks a participant makes on each image provides a direct signal of perceived informativeness, in the sense that more identified features suggest richer and potentially more interpretable concept content.

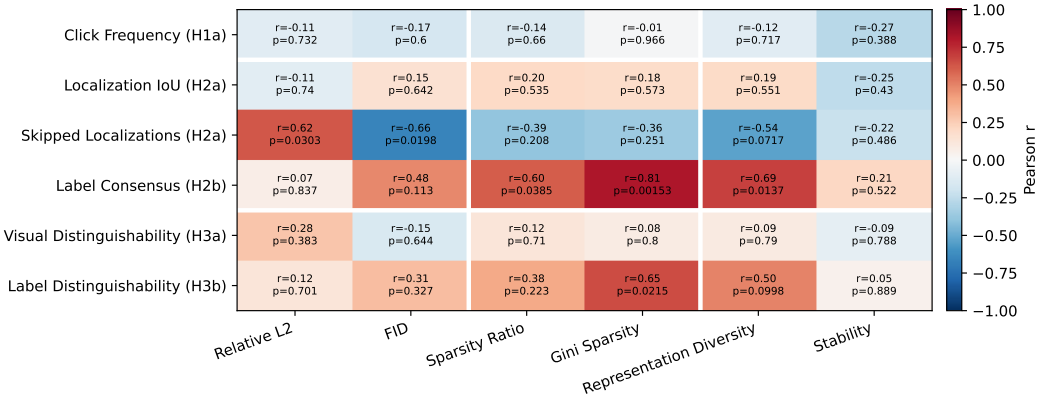

Figure 2: Correlation matrix between all tasks of the human evaluation and the dictionary learning metrics. Horizontal white dividers separate the different hypotheses (H1) Informativeness, (H2) Understanding, and (H3) Distinguishability from one another, while vertical dividers separate reconstruction-, sparsity- and other metrics.

A detailed description of outlier handling and integrity checks of the provided responses can be found in Appendix C.2.

Results. In this task, the most information rich concept space belongs to the *church* class with an average of 6.72 clicks per patch. Contrastingly, the concept space with the lowest amount of clicks is *parachute* with 3.43. The average click frequency across all concept spaces is 5.00, with the $tench$ concept space being the closest to it at 5.049. Refer to Table 4 for the full results and to Appendix I for visual concept representations. Figure 2 shows all correlations between click frequencies and dictionary learning metrics in the first row. It is evident that no significant relationship is present for the task with no correlation reaching statistical significance (all $p > 0.16$).

4.3 Hypothesis 2: Understanding

To accurately measure participants' understanding of the given concept spaces, we employ both a visual and a linguistic task. The following sections describe these tasks in detail.

4.3.1 Evaluation Measure 2a: Context Localization

Task Description. Participants are tasked with localizing the most prototypical patch, (top_1), of a concept within a target image representative of the class underlying the concept space. The goal of the task is to measure the degree of spatial agreement between the participants. Figure 1 shows that participants respond to: *"If the content of Image A were part of Image B, where would it be?"* by drawing free-hand regions onto the target image. High overlap among regions marked independently by participants indicates that the concept is readily recognizable and spatially well localized. We also recorded the number of participants who chose to skip drawing a region (skip). Low overlap and high skip rates indicate that the concept cannot be put into context easily.

Results. The mean IoU across all spaces is 18.00% with minimum and maximum scores at 9.68% and 33.20% belonging to *cassette_recorder*, *english_springer* respectively (see Appendix I). The correlations between human driven IoU scores and metrics yield no significance at ($p > 0.43$) for all pairs. Comparing the number of skips in the third row of Figure 2 with metrics, shows that *Relative L2* correlates positively, while *FID* correlates negatively ($r{=}{+}0.62$ vs. $r{=}{-}0.66$, $p{=}0.030$ and $p{=}0.020$ respectively). The moderate diverging correlations between both metrics are driven by the airliner concept space being an outlier for *Relative L2* and *FID* scores; however, the trend is still evident without taking this class into account. A leave-one-class-out analysis of all correlations is reported in Appendix F. A more detailed discussion can be found in Section 5.

4.3.2 Evaluation Measure 2b: Label Consensus

Task Description. Participants are tasked to label concept image groups: *You see parts of images that belong to the category [class]. Label a common detail* (see Figure 1). The labels provided by participants allow us to assess whether the concepts convey semantic meaning that is consistently recognized across individuals. To compare similarities between labels, we use language-embedding

distances, as described in more detail in Section C.2. If participants independently assign identical or semantically similar labels, this suggests that the concept captures a meaningful and coherent property. In contrast, highly divergent labels indicate conceptual ambiguity.

Results. As shown in Figure 2, the label consensus measure achieves correlations of r=0.81 (p=0.002) against *Gini Sparsity* and r=0.69 (p=0.014) against *Representation Diversity* metrics. This indicates significant trends with moderate correlation. For a more intuitive overview of all the surveyed labels and their semantic meanings, see Figure 14 in the Appendix.

4.4 Hypothesis 3: Distinguishability

While Evaluation Measure 2b assesses whether individual participants agree on the interpretation of a given concept, this hypothesis examines a complementary question: whether different concepts within the same concept space are distinguishable from one another. As in the previous tasks, distinguishability of concepts can be evaluated from both visual impressions and linguistic descriptions. We therefore use two separate evaluation measures, 3a and 3b, to capture these dimensions.

4.4.1 Evaluation Measure 3a: Visual Distinguishability

Task Description. Participants are presented with two geometrically similar concepts, selected based on minimal embedding distance, and are tasked to rate their visual distinguishability on a five-point ordinal Likert scale (1 = hard to distinguish to 5 = easy to distinguish), without relying on language. Participants respond to the question: *"How distinguishable are these two image groups?"*. This task provides an independent measure of concept distinguishability. Section C.2 describes how the pairwise ratings are attributed to concept spaces. Even if two concepts have semantically clear labels, they do not effectively partition the model's learned feature space if they remain visually confusable.

Results. The lowest visual distinguishability score belongs to the *golf_ball* concept space, achieving a score of 3.63, which is notably above the midpoint of the scale (3). The highest score of 4.29 belongs to the *english_springer*. An explanation of this upside bias can be found in Section 5. No correlation analysis between the ratings and a dictionary metric yields a significant statement with $p > 0.38$.

4.4.2 Evaluation Measure 3b: Label Distinguishability

Task Description. The label distinguishability measure captures the *inter-concept separation* of labels provided by participants. It uses the same task as Evaluation Measure 3a, but measures distinguishability implicitly. It also relies on the semantic embedding framework introduced there. Details on the analysis can be found in C.2.

Results. The distinguishability score achieves correlations of r=0.65 (p=0.022) against *Gini Sparsity* and r=0.50 (p=0.098) against *Representation Diversity*, marginally above the significance threshold of p=0.05. Label distinguishability therefore shows a similar overall correlation pattern with sparsity-based metrics.

5 Discussion

In this section, we interpret the main findings from Section 4, discuss their implications for practice and evaluation, outline the limitations of our study and directions for future work, and explain why our results better support a diagnostic tool rather than a benchmark. Overall, our results show that **metrics are informative about particular properties of concept spaces, but are not interchangeable proxies for human comprehension.**

5.1 Interpretation of main findings

[H1 – Click Frequency] Our study does not provide sufficient evidence for Hypothesis 1, which does not produce a measurable connection between the informativeness and the tested DL metrics. The absence of a systematic correlation between click frequencies and the metrics may also reflect individual variability in participants' clicking behavior. That is, the chosen level of granularity introduces a variance that exceeds the differences between the studied concept spaces. Concept spaces with richer hierarchical structure may be required to reveal systematic differences in this task.

[H2a – Context Localization] We did not find evidence for Hypothesis 2a analysing the IoU. However, *FID* correlates negatively and *Relative L2* positively with the skip rate. A possible

explanation is that IoU is evaluated only for trials in which participants localized the patch, whereas the skip count captures whether participants understood the content at all; IoU may therefore be a less expressive discriminator. IoU itself provides limited signal: it is sensitive to imprecision in hand-drawn boundaries—noise that does not necessarily reflect semantic understanding—and the drawn regions tend to be geometrically similar across concepts, so boundary noise stacks on comparatively coarse spatial information and may obscure the semantic variation captured by the sparsity metrics. The skip count, by contrast, is sensitive to differences in localization difficulty, though it cannot establish that participants share a common understanding—only that they understood the concept in *some* way. Why *Relative L2* and *FID*, both capturing information loss, diverge in sign remains an open question (see Section 5.3).

[H2b – Label Consensus] Our findings support Hypothesis 2b, suggesting that *Gini Sparsity* and *Representation Diversity* serve as useful indicators of concept understandability. Both correlations are consistent with the hypothesized mechanism in that sparser concept spaces, in which each concept captures a more selective combination of features, afford participants a clearer semantic signal and elicit more uniform labeling. The particularly strong performance of *Gini Sparsity* may stem from its status as one of the most comprehensive measures of sparsity [Hurley & Rickard, 2009]. This relationship holds relative to the studied concept spaces and does not imply universal guidance. The direction and rationale of all hypothesized associations are detailed in Appendix B.

[H3a – Visual Distinguishability] We did not find evidence for Hypothesis 3a: the visual distinguishability measure shows no systematic correlation with any metric, despite measurable differences in pairwise distinguishability. Ratings are inherently relative since they attribute a pair of two concepts at the same time despite each concept having both, more and less distinguishable counterparts. When aggregating the ratings across pairings per concept, extreme pairings may therefore cancel out, yielding similar patterns across concept spaces and making contrastive ratings difficult to compress into a concept-space-level metric.

[H3b – Label Distinguishability] We show that *Gini Sparsity* correlates with the linguistic distinguishability measure. This supports the assumption that sparser concept space geometries indeed tend to yield more distinguishable labels, rather than collapsing into a single umbrella term for all concepts. Thus, sparsity may indicate distinguishability to some extent, but this interpretation neglects cases where the semantics of concepts are distinguishable while not expressible in language, and ignores a potential influence of our implicit measure instead of explicit pairwise distinctions.

5.2 Implications for practice and evaluation

The trends in our results suggest two cautious and pragmatic conclusions for practical consideration. When systematically investigating extracted concepts, researchers may

- refer to *Gini Sparsity* and *Representation Diversity* as supportive indicators of semantic comprehension and discrimination of concepts. At the same time, this is a trend within the context of extracted concept spaces: no absolute sparsity values can be defined as precise targets or allow for direct rating of human comprehension.
- conduct a human evaluation of the comprehensibility of their concept-spaces when visual in-context understanding and visual distinguishability are of interest, as no general conclusions could be drawn as of now.

Consequently, dictionary metrics should not be treated as drop-in substitutes for human evaluation. Instead, they should be used as complementary diagnostics whose relevance must be validated for the specific interpretability goal. Because different metrics map to different properties of a learned concept space, multi-dimensional evaluation (including human-grounded checks) remains essential.

5.3 Limitations and future work

Several limitations constrain how far these results generalize. First, our study uses concept spaces extracted post hoc from a fixed vision model's representations via NMF. The results may therefore not transfer unchanged to other architectures, datasets, extraction methods, or concept-based methods that explicitly optimize for completeness, intervention utility or other downstream performance. Future work should investigate whether the relationship between human evaluation and *Gini Sparsity* as well as *Representation Diversity* hold across broader classes of learned concept spaces.

Second, our human-evaluation measures probe only selected aspects of concept-space comprehensibility. In particular, the visual tasks yielded weaker and less consistent correlations with metrics

than the linguistic tasks. This may reflect limited task sensitivity, insufficient visual contextualization, or a mismatch between the visual protocol and the properties captured by the metrics. Future work should develop visual evaluation protocols that better isolate concept space comprehension and build primarily on the label-based measures, which showed the clearest connections to the metrics.

Third, some tasks may require richer or hierarchical concept spaces to reveal consistent patterns across metrics. Measures such as informativeness may be difficult to capture with flat concept spaces whose concepts correspond primarily to local visual patterns. Future work should therefore investigate whether hierarchical or multi-scale concept representations yield stronger alignment with the measure informativeness.

Finally, some relationships between metrics and human evaluation measures may be sensitive to class-level variation or influential stimuli. Since our conclusions rest on only twelve concept spaces, individual classes or stimuli can drive individual correlations. To measure the effect, we conducted a leave-one-class-out analysis (Appendix F) showing that most correlations remain stable, with no class reversing the central associations. While this supports the findings of our study, future studies should aim to use larger and more diverse stimulus sets. In addition, the human evaluation measures may interact rather than provide independent probes of concept space comprehensibility, and future work should clarify whether the divergence between *Relative L2* and *FID* reflects a general difference between pixel-level reconstruction error and distributional preservation. Finally, since participants in our study serve as human reference measures rather than annotators, conventional methods of quantifying response quality, such as inter-annotator agreement statistics, do not apply. We refer to Appendix G for further discussion of this distinction.

5.4 Diagnostics tool or (not) a benchmark

The five measures serve as reference operationalizations of concept space comprehensibility in our setting and are not a universal definition of it. For this reason, we do not propose a new optimizable target for concept spaces. Turning this study into a benchmark would elevate a study-specific measurement setup into a normative optimization target. That would be methodologically inconsistent with the aim of the paper, and would also create the wrong incentives for the community. Once a fixed benchmark exists, methods are incentivized to optimize performance on that particular operationalization [Ott et al., 2022], potentially exploiting task- or dataset-specific regularities rather than yielding genuinely more human-comprehensible concept spaces. We therefore position the five evaluation measures as diagnostic tools for understanding what dictionary learning metrics capture and what they miss about human evaluation of learned concept spaces, rather than as a leaderboard-style benchmark.

5.5 Reproducibility

To ensure transparency and reproducibility, the data and code used in this work are made available to the research community. Upon publication, all relevant resources, such as datasets, preprocessing scripts, participant responses, and implementation code, will be publicly released via an accessible repository.

6 Conclusion

In this work, we studied whether metrics motivated by dictionary learning formulations are informative about how comprehensible learned concept spaces are to humans. In an NMF-based concept extraction setting, using five complementary human-evaluation measures, we find that dictionary learning metrics relate to them in substantially different ways. *Gini Sparsity* shows the most consistent relationship among the metrics we evaluate, followed by *Representation Diversity*, although both remain evaluation measure-dependent. By contrast, the reconstruction-based metrics *Relative L2* and *FID* provide conflicting evidence regarding their connection to concept-space comprehensibility. These findings indicate that DL metrics are informative about particular properties of concept spaces, but are not interchangeable proxies for human-comprehension and thus interpretability. More broadly, as concept-based methods increasingly rely on learned concept spaces, our results caution against treating distinct properties of these spaces as evidence of human comprehensibility without human-centered validation.

ETHICS STATEMENT

This work involves a human evaluation study with 351 voluntary participants yielding 5,616 individual responses across four tasks (property identification, region localization, free-text labeling, and visual distinguishability rating). We describe here how the study was conducted in accordance with the ICLR Code of Ethics.

Participant treatment. Participation was voluntary, unpaid, and based on convenience sampling through direct invitations from the authors' personal networks; participants could discontinue the study at any point. The study consisted of four short visual tasks with no deceptive elements, no potentially distressing content (all stimuli are everyday ImageNet images and patch crops), and an expected completion time of under 5 minutes. No compensation was offered or paid, which avoids false monetary incentives.

Data protection. No personally identifiable information was collected: we deliberately did not record demographic data (age, gender, location), since the research questions involve no demographic hypotheses and this choice follows the data-minimization principle. Responses are stored under a random, pseudonymous per-session identifier with no linkage to identity. Free-text labels were additionally filtered to remove accidental personal information before analysis. The Web-based study infrastructure stores responses in a cloud backend (Firebase Firestore); only pseudonymous responses and task metadata are retained.

Consent and transparency. Participants were informed of the study's purpose, the tasks involved, and the estimated completion time on the landing page before starting, and participation implied informed consent to the pseudonymous use of their responses for research purposes.

Limitations of the participant pool. Because recruitment relied on the authors' personal networks, the participant pool skews toward a Western- and Central-European population. We discuss this sampling bias and its implications for the generalizability of our findings in Section 5.3.

AI USE STATEMENT

In this work, generative AI tools were used only for tasks that fall under recommended disclosure: editing parts of the manuscript to improve readability and clarity. Generative AI tools were not used to generate synthetic data sets, develop theoretical models or proofs, formulate mathematical claims, design or conduct the research methodology or experiments, implement methods, perform data analysis, or interpret results; all of these were carried out by the authors. The study design, the human-subjects experiment, and all findings and claims are the work of the authors. All AI-assisted work was reviewed by the authors, who take full responsibility for the final content of this work, including text, claims, and artifacts.

REFERENCES

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

Michael Cogswell, Faruk Ahmed, Ross Girshick, Larry Zitnick, and Dhruv Batra. Reducing overfitting in deep networks by decorrelating representations. *arXiv preprint arXiv:1511.06068*, 2015.

Omid Davoodi, Shayan Mohammadizadehsamakosh, and Majid Komeili. On the interpretability of part-prototype based classifiers: a human centric analysis. *Scientific Reports*, 13(1):23088, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Mateo Espinosa Zarlenga, Pietro Barbiero, Zohreh Shams, Dmitry Kazhdan, Umang Bhatt, Adrian Weller, and Mateja Jamnik. Towards robust metrics for concept representation evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(10):11791–11799, Jun. 2023. doi: 10.1609/aaai.v37i10.26392. URL https://ojs.aaai.org/index.php/AAAI/article/view/26392.

Thomas Fel, Victor Boutin, Louis Béthune, Remi Cadene, Mazda Moayeri, Léo Andéol, Mathieu Chalvidal, and Thomas Serre. A holistic approach to unifying automatic concept extraction and concept importance estimation. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 54805–54818. Curran Associates, Inc., 2023a. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/abf3682c9cf9245a0294a4bebe4544ff-Paper-Conference.pdf.

Thomas Fel, Agustin Picard, Louis Bethune, Thibaut Boissin, David Vigouroux, Julien Colin, Rémi Cadène, and Thomas Serre. Craft: Concept recursive activation factorization for explainability. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2711–2721, 2023b.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/77d2afcb31f6493e350fca61764efb9a-Paper.pdf.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Oscar Hill, Mateo Espinosa Zarlenga, and Mateja Jamnik. Hierarchical concept-based interpretable models. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=h61OIERd38.

Jeremy Howard. Imagenette: A smaller subset of 10 easily classified classes from imagenet. https://github.com/fastai/imagenette, March 2019.

Niall Hurley and Scott Rickard. Comparing measures of sparsity. *IEEE Transactions on Information Theory*, 55(10):4723–4741, 2009.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie J. Cai, James Wexler, Fernanda B. Viégas, and Rory Sayres. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, 2017.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5338–5348. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/koh20a.html.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *nature*, 401(6755):788–791, 1999.

Emanuele Marconato, Andrea Passerini, and Stefano Teso. Interpretability is in the mind of the beholder: A causal framework for human-interpretable representation learning. *Entropy*, 25 (12), 2023. ISSN 1099-4300. doi: 10.3390/e25121574. URL https://www.mdpi.com/1099-4300/25/12/1574.

Tuomas Oikarinen, Subhro Das, Lam M. Nguyen, and Tsui-Wei Weng. Label-free concept bottleneck models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=FlCg47MNvBA.

Mayu Otani, Yiqin Song, Pratyusha Cha, Sanghyuk Lee, Takumi Tsuchiyama, Jongwoo Lee, Koji Yamaguchi, Yung-Hsuan Lee, Hao Shen, Xudong Wang, et al. Toward verifiable and reproducible human evaluation for text-to-image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Simon Ott, Adriano Barbosa-Silva, Kathrin Blagec, Jan Brauner, and Matthias Samwald. Mapping global dynamics of benchmark creation and saturation in artificial intelligence. *Nature Communications*, 13(1), November 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-34591-0. URL http://dx.doi.org/10.1038/s41467-022-34591-0.

Amin Parchami-Araghi, Sukrut Rao, Jonas Fischer, and Bernt Schiele. Fact: Faithful concept traces for explaining neural network decisions, 2026. URL https://arxiv.org/abs/2510.25512.

Eleonora Poeta, Gabriele Ciravegna, Eliana Pastor, Tania Cerquitelli, and Elena Baralis. Concept-based explainable artificial intelligence: A survey. *ACM Comput. Surv.*, November 2025. ISSN 0360-0300. doi: 10.1145/3774643. URL https://doi.org/10.1145/3774643.

Stephan Pohl, Edgar Y. Walker, David L. Barack, Jennifer Lee, Rachel N. Denison, Ned Block, Florent Meyniel, and Wei Ji Ma. Clarifying the conceptual dimensions of representation in neuroscience. *Nature Reviews Neuroscience*, 27(5):357–372, May 2026. ISSN 1471-0048. doi: 10.1038/s41583-026-01030-8. URL https://doi.org/10.1038/s41583-026-01030-8.

Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10932–10941, June 2023.

Sunayana Rane, Polyphony J. Bruna, Ilia Sucholutsky, Christopher Kello, and Thomas L. Griffiths. Concept alignment, 2024. URL https://arxiv.org/abs/2401.08672.

Alexandre Rocchi-Henry, Thomas Fel, and Gianni Franchi. A geometric unification of concept learning with concept cones, 2025. URL https://arxiv.org/abs/2512.07355.

S. Schrodi, J. Schur, M. Argus, and T. Brox. Selective concept bottleneck models without predefined concepts. *Transactions on Machine Learning Research (TMLR)*, May 2025. URL http://lmb.informatik.uni-freiburg.de/Publications/2025/SAB25.

Wolfgang Stammer, Antonia Wüst, David Steinmann, and Kristian Kersting. Neural concept binder, 2024. URL https://arxiv.org/abs/2406.09949.

Warren S. Torgerson. Multidimensional scaling of similarity. *Psychometrika*, 30(4):379–393, 1965. doi: 10.1007/BF02289530.

Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, 1980. ISSN 0010-0285. doi: https://doi.org/10.1016/0010-0285(80)90005-5. URL https://www.sciencedirect.com/science/article/pii/0010028580900055.

Mert Yuksekgonul, Maggie Wang, and James Zou. Post-hoc concept bottleneck models, 2023. URL https://arxiv.org/abs/2205.15480.

A METRICS

We provide formal definitions, aggregation procedures, and implementation details for the metrics introduced in Section 3.2. We formalize all metrics using the notation introduced in Section 3.1, so that each metric is defined with respect to the same factorized representation. The metric set includes metrics adapted from Fel et al. [2023a], with *Gini Sparsity* [Hurley & Rickard, 2009] and *Representation Diversity* [Cogswell et al., 2015] added as complementary metrics naturally defined for factorized concept spaces. We omit the out-of-distribution metric of Fel et al. [2023a] because our study focuses on in-distribution human evaluations of the extracted concept spaces.

Concretely, we consider the following metrics:

Relative L2 (↓). Measures reconstruction error between the original activation matrix A and its reconstruction $UV^\top$, using the normalized Frobenius norm. It is a *space-level* metric and is computed once per learned concept space.

$$\text{Relative-L2}(A, U, V) \;=\; \frac{\|A - UV^\top\|_F}{\|A\|_F}.$$

FID (↓). Measures the distributional discrepancy between two sets of patch-level features, here used to quantify how much the reconstructed representation departs from the original one. It is a *space-level* metric and is computed on the full set of patches associated with a concept space.

$$\text{FID}(A, U, V) \;=\; W_2\big(A, UV^\top\big).$$

Sparsity Ratio (↑). Measures how many concept coefficients are inactive in a patch representation. It is first computed *per patch* from the code vector $u_m \in \mathbb{R}^K$, by counting exact zeros (inactive coefficients), and then averaged over all patches in the concept space.

$$\text{SparsityRatio}(u_m) \;=\; \frac{\#\{k : u_{m,k} = 0\}}{C} \;=\; 1 - \frac{\#\{k : u_{m,k} neq 0\}}{C}.$$

Gini Sparsity (↑). Measures how concentrated the mass of a patch representation is over concept dimensions. Unlike threshold-based sparsity, it provides a continuous indicator of signal concentration. It is computed *per patch* from the normalized coefficient vector u_m, and then averaged over all patches in the concept space. We chose to include this specific metric because it satisfies all six criteria of an expressive sparsity measure as defined by Hurley & Rickard [2009].

$$G(u_m) \;=\; \frac{\sum_{i=1}^{C} \sum_{j=1}^{C} |u_{m,i} - u_{m,j}|}{2C \sum_{i=1}^{C} u_{m,i}}.$$

Representation Diversity (↑). Measures how distinct the learned concept directions are from one another. It is a *space-level* metric, computed from the set of concept vectors in U, and aggregated over all concept pairs.

$$\text{Diversity}(V) \;=\; \frac{2}{C(C-1)} \sum_{i<j} \left(1 - \frac{v_i^\top v_j}{\|v_i\|\|v_j\|}\right).$$

Stability (↓). Measures how consistent the learned concept representation is across repeated runs or perturbations of the data. It is a *space-level* metric and is aggregated across multiple factorization runs. *Stability* is computed across multiple runs using optimal matching (Hungarian) on cosine similarities; the per-run matched cost is $1-$cosine and the final *Stability* is the mean matched cost across runs.

Unless stated otherwise, per-patch metrics are averaged over all patches in a concept space, and space-level metrics are reported per concept space and then averaged across the evaluation set.

B Hypothesis Rationale and Expected Metric-Measure Associations

Table 1 summarizes, for each dictionary-learning metric family and human-evaluation measure, the direction of the association hypothesized in Section 3 and the rationale behind it.

Table 1: Hypothesized associations between dictionary-learning metric families and human-evaluation measures. Signs state the expected direction if the metric is predictive of human comprehension.

Metric family	Metric	Expected sign	Rationale
Faithfulness	Relative L2	− (all measures)	Higher pointwise reconstruction error implies less information preserved, weakening any comprehension signal.
Faithfulness	FID	− (all measures)	Larger distributional discrepancy implies less faithful concept spaces.
Reliability	Stability	− (all measures)	Consistency across factorization runs suggests robust, reproducible concepts.
Structural	Sparsity Ratio	+ (consensus, distinguishability)	Sparser spaces yield sharper, more distinct semantic signals.
Structural	Gini Sparsity	+ (consensus, distinguishability)	Most comprehensive sparsity measure [Hurley & Rickard, 2009]; stronger semantic sharpness.
Structural	Repr. Diversity	+ (consensus, distinguishability)	Less redundancy means more semantically distinct concepts.

C Study Design

C.1 Setup

Concept Extraction and Concept Evaluation We generate a separate concept space for each class. This allows us to systematically generate concept spaces of different geometry while avoiding architectural changes to the pipeline that may hinder direct comparability. To enable the discovery of multiple concepts at different locations within an image, we split the images into a set of patches and extract concepts at a patch level. Precisely, the patches P are obtained by first resizing the original image to 256×256, then extracting crops of size 64×64 with stride 80% and finally upscaling each patch to 224×224. The last upscaling step ensures that the full receptive field of the pre-trained ResNet50 that we use as a feature extractor is utilized [He et al., 2016]. The features are extracted at the penultimate layer of the network, as the penultimate layer has shown to provide the best trade-off between semantic richness and sparse class-level representations [Fel et al., 2023a]. Since the patch representations are already being computed spatially independent from one another, we can pool their internal spatial dimensions into a single vector. To keep our setup simple, we employ NMF (non-negative Matrix Factorization) as a classic linear-additive concept extractor. We fixed $k = 5$ across all classes to balance concept coverage with the feasibility of the human evaluation (targeting a participant pool of $N = 300$, i.e., at least 20 responses per stimulus across four tasks) and to ensure a controlled, comparable decomposition granularity across classes. While classes differ in visual complexity, the differences arising from applying a fixed k to classes of varying complexity are part of the phenomenon under investigation: adapting k per class would partially remove these differences and obscure whether dictionary learning metrics track human judgments of the resulting spaces. This follows Fel et al. [2023a], who selected $k = 10$ in their setting. In total this yields $12 \times 5 = 60$ concepts for the technical and the human evaluation.

Resources The concepts were extracted using a *Dell Precision 3571* laptop featuring a *12th Gen Intel(R) Core(TM) i7-12800H* processor, 32 GBs of RAM and an *Nvidia T600* GPU with 4 GB dedicated VRAM.

Participants and Responses We recruited 451 participants, of whom 351 completed the full online study and were included in the analysis. At 4 tasks per participant with 4 visual stimuli each, we were

able to take $351 \times 4 \times 4 = 5,616$ responses into account. As evident from Table 2, the amount of responses were enough for us to cover the targeted significance threshold of $N > 20$ participants per visual stimulus. This design provides a principled basis for capturing patterns in human judgments, offering insights that are indicative beyond individual perspectives. Since interpreting ImageNet concepts does not require specialized expertise, we did not restrict recruitment to a specific target population. Participants were instead recruited via convenience sampling through direct messages within the authors' personal networks (e.g., friends, family, and community contacts), which are predominantly located in Western and Central Europe. Participants could choose between an English and a German version of the study, with English as the default setting. We deliberately did not collect demographic data (e.g., age, gender, or cultural background): our hypotheses involve no demographic effects, and recording such data would conflict with data-minimization principles; as a consequence, our results cannot speak to cultural, linguistic, age-related, or expertise variation in concept comprehension (see Section 5.3). The average completion time across all participants is approximately 3.5 minutes (211 s); the per-task average response times are approximately 41.7 s (Click Frequency), 52.8 s (Region Localization), 79.1 s (Labeling), and 37.5 s (Distinguishability Rating). Participation was voluntary and uncompensated, which avoids false monetary incentives. The study complies with local law as well as the ethics guidelines of the associated university.

Control and Bias Avoidance Designing a human evaluation study is a complex and often failure prone task [Otani et al., 2023]. There are many traps to fall into, as humans are prone to biases such as the fatigue bias or priming through the order in which questions are posed. To mitigate these effects and ensure a controlled response behavior, a rule-based heuristic scheduler is employed. Its function is to dispatch the question items and matching visual stimuli to the participants such that the following rules are adhered to and the objective is optimized:

- **Randomization Objective:** The question items are assigned to a specific participant are in principle drawn at random, with the exception of the following rules.
- **Instruction Repetition Rule:** Questions are grouped by their respective task type. Questions of the same type are all shown in blocks, such that the same tasks (sharing the same instruction) are all shown in an unbroken sequence. This ensures that the participants do not need to re-read everything.
- **Reference Item Rule:** Each question block is preceded by a training item of the same task type. The sole purpose of this item is to aid as a reference anchor point to the following questions. Its responses are disregarded in the final inspection.
- **Priming Avoidance Rule:** The sequence in which the four question blocks are shown to the participants is balanced equally across all participants (i.e. across the study each task is shown equally often in the first, second, third, and fourth spot in the sequence). This behavior compensates for priming biases that arise from showing one task instruction before the other.
- **Efficiency Rule:** To avoid duplicate responses from one single participant, no participant receives a duplicate item.

For more information on the amount of participants and their responses see Table 2, randomly selected example responses for each task are exhibited in Section 4.

C.2 Detailed Processing of Responses

Click Frequency To account for outliers that stem from misinterpretations of the task, we filter for responses where the amount of clicks deviate more than 2.5 standard deviations from the mean click frequency of the respective patch. These outliers are ignored for subsequent analysis. We then mean-aggregate the filtered click frequencies per patch to achieve a score that is representative of human comprehension. We then average the scores for all patches that belong to the same concept space to quantify the perceived informativeness.

Context Localization We use the free-hand region annotations given in the responses as the basis to compute multiple agreement statistics between annotators. These include the mean IoU as a measure of overlap and the amount of skipped task responses as an indicator of cases where no reasonable localization is possible.

Label Consensus For free-text labeling, we normalize labels to lowercase formatting and clean them using a handcrafted filter list. The filter either corrects spelling mistakes, translates to English, or removes non-informative entries that do not constitute meaningful judgments (e.g., "-", ".", "...", or isolated characters). In total, no more than 5% of labels are either removed or modified. After filtering, labels are embedded into a semantic vector space using *embeddinggemma-300m*. To assess label consensus, we focus on *intra-concept similarity*. Specifically, we compute a pairwise intra-concept cosine similarity score reflecting how semantically similar the labels assigned to the same concept are.

Let $L^{(c)} = \{\ell_1, \ell_2, \ldots, \ell_{N_c}\}$ be the set of cleaned, normalized labels assigned to concept c, where N_c is the number of labels for concept c, and let $\mathbf{v}_i = \mathcal{E}(\ell_i) \in \mathbb{R}^d$ denote the embedding of label ℓ_i.

The intra-concept consensus score is the mean pairwise cosine similarity:

$$S_c = \frac{2}{N_c(N_c-1)} \sum_{1 \le i < j \le N_c} \frac{\mathbf{v}_i \cdot \mathbf{v}_j}{\|\mathbf{v}_i\| \, \|\mathbf{v}_j\|}$$

Visual Distinguishability We group ratings by concept pairs across all users to estimate their general mean distinguishability rating. We then average the pairwise judgments back onto individual concepts and concept spaces as a measure of visual distinguishability.

Let $\mathcal{P}$ be the set of all rated concept pairs within a concept space, and let $R_{c,c'} = \{r_1, r_2, \ldots, r_n\}$ be the set of ratings collected for the pair (c, c'), where each $r_i \in \{1, 2, 3, 4, 5\}$ and higher values indicate the concepts are easier to distinguish. The mean pairwise rating is:

$$\bar{r}_{c,c'} = \frac{1}{|R_{c,c'}|} \sum_{i=1}^{|R_{c,c'}|} r_i$$

The visual distinguishability score for concept c is then the mean over all pairs in which c participates:

$$V_c = \frac{1}{|\mathcal{N}(c)|} \sum_{c' \in \mathcal{N}(c)} \bar{r}_{c,c'}$$

where $\mathcal{N}(c) = \{c' \mid (c, c') \in \mathcal{P} \text{ or } (c', c) \in \mathcal{P}\}$ is the set of concepts paired with c. Higher values of V_c indicate that concept c is easier to visually distinguish from other concepts in the same space.

Label Distinguishability We reuse the postprocessing and embedding pipeline described in Evaluation Measure 3: labels are normalized, lightly filtered, embedded and compared via cosine similarity. For each concept, we compute a *concept-versus-rest distinguishability score*, measuring how well that concept's label embeddings are separated from labels assigned to other concepts within the same class or concept space. High distinguishability indicates that participants not only agree internally on what a concept represents, but also use labels that are semantically specific to that concept rather than generic or overlapping.

Let $L^{(c)} = \{\ell_1, \ldots, \ell_{N_c}\}$ be the labels for concept c and $L^{(-c)} = \{\ell'_1, \ldots, \ell'_{N_{-c}}\}$ the pooled labels of all other concepts in the same concept space. Define the pairwise distance matrix over the combined set as $D_{ij} = 1 - \cos(\mathbf{v}_i, \mathbf{v}_j)$, with group labels $y_k = 1$ for samples belonging to $L^{(c)}$ and $y_k = 0$ otherwise.

For each label ℓ_k, let $a(k)$ be its mean distance to all other labels in the same group, and $b(k)$ its mean distance to all labels in the other group. The per-sample value is:

$$s(k) = \frac{b(k) - a(k)}{\max\{a(k),\, b(k)\}}$$

The concept-level distinguishability score is the mean score over all $N_c + N_{-c}$ samples:

$$D_c = \frac{1}{N_c + N_{-c}} \sum_{k=1}^{N_c+N_{-c}} s(k) \;\in [-1, 1]$$

Higher values indicate that the labels of concept c are semantically more distinct from those of the remaining concepts in the same space.

Table 2: Overview of question coverage and participation per task. For each task, the table reports the total number of unique question statements, the average number of unique participants per question, the total number of task assignments, the number of assignments that resulted in a usable response, and the resulting drop-out rate expressed as the percentage gap between assignments and usable responses.

Task	# Questions	Avg. Participants per Question	Total Assignments	Usable Responses	Drop-out Rate (%)
Click Frequency	60	23.100	1600	1386	13.4
Context Localization	60	23.183	1604	1391	13.3
Labeling	60	23.083	1680	1385	17.6
Distinguish. Rating	42	33.190	1572	1394	11.3

D Detailed Results

This section reports detailed statistics on participation, technical metrics, and human evaluation measures. An overview of question coverage and participation is provided in Table 2. The evaluation metrics computed on individual concept spaces are summarized in Table 3. The full definitions and preprocessing details for these metrics are provided in Section A. Complementing the metric evaluation, the results of the human-centered measurements are presented in Table 4. Further details on the evaluation procedure and postprocessing of responses are described in Section C.2.

Table 3: **Overview of Metric Results.** Evaluation results of metrics computed on individual concept spaces. **Bold** values mark the best result per column; arrows indicate the direction of improvement (↑ higher is better, ↓ lower is better). The table reports reconstruction fidelity (*Relative L2*, *FID*), sparsity measures (*Sparsity Ratio*, *Gini Sparsity*), *Representation Diversity*, and *Stability* for each concept space.

Concept Space	Relative L2 ↓	FID ↓	Sparsity Ratio ↑	Gini Sparsity ↑	Repr. Diversity ↑	Stability ↓
airliner	**0.49**	695.0	0.20	0.51	**0.46**	0.027
canoe	0.55	550.0	0.19	**0.52**	0.44	0.037
cassette recorder	0.54	394.6	0.13	0.44	0.38	0.025
chainsaw	0.56	508.5	0.15	0.49	0.42	0.040
church	0.54	449.8	0.18	0.49	0.43	0.033
english springer	0.55	431.0	0.16	0.47	0.40	0.019
french horn	0.54	503.9	0.15	0.47	0.38	0.022
garbage truck	0.52	443.4	0.14	0.47	0.40	**0.013**
gas pump	0.55	**349.9**	0.13	0.44	0.36	0.021
golf ball	0.53	550.2	0.19	0.51	**0.46**	0.040
parachute	0.53	556.6	**0.20**	0.49	0.45	0.049
tench	0.52	504.5	0.14	0.48	0.40	0.064

Table 4: **Overview of Human Measurement Results.** Results from the human evaluation of concept spaces. **Bold** values mark the best result per column; arrows indicate the direction of improvement (↑ higher is better, ↓ lower is better). All values are means over the participants of the human evaluation study.

Concept Space	Click Freq. ↑	IoU ↑	Skips ↓	Label Consensus ↑	Label Disting. ↑	Visual Disting. ↑
airliner	5.50	**0.25**	4.0	0.63	0.07	3.55
canoe	5.26	0.16	10.0	**0.69**	0.10	3.83
cassette recorder	4.74	0.10	9.8	0.55	0.03	3.63
chainsaw	4.83	0.21	6.2	0.67	**0.10**	3.92
church	**6.72**	0.22	8.8	0.61	0.07	3.91
english springer	4.74	0.33	10.0	0.64	0.08	**4.20**
french horn	4.77	0.15	6.8	0.59	0.06	3.69
garbage truck	5.32	0.20	4.6	0.63	0.09	4.00
gas pump	5.39	0.13	10.4	0.59	0.05	3.49
golf ball	4.36	0.12	6.4	0.67	0.08	3.51
parachute	3.44	0.15	**3.8**	0.62	0.06	3.99
tench	5.05	0.16	7.0	0.60	0.06	3.71

D.1 Multiple-Comparison Correction

Since each evaluation measure is correlated against all dictionary learning metrics, the p-values in Figure 2 are affected by multiple testing. We therefore additionally report the *Gini Sparsity* correlations (one per evaluation measure, as the most comprehensive sparsity measure) at the concept-space level ($n = 12$), corrected using Benjamini–Hochberg FDR control. Table 5 reports the raw and corrected significance levels.

Table 5: **Multiple-comparison correction for the six primary hypothesis tests** (evaluation measure score vs. *Gini Sparsity*, concept-space level, $n = 12$). Reported are the Pearson correlation r, the raw p-value, and the Benjamini–Hochberg FDR-adjusted q-value, computed across the six tests of this family.

Human Task	r	p (raw)	q (BH)
Label Consensus	0.81	0.0015	**0.009**
Label Distinguishability	0.65	0.0215	0.064
Localizations Skips	−0.36	0.2507	0.502
Region Overlap (IoU)	0.18	0.5734	0.860
Visual Distinguishability	0.08	0.7995	0.959
Click Frequency	−0.01	0.9662	0.966

The *Label Consensus* measure stays significant ($r = 0.81$, $q = 0.009$). It also remains significant when the Benjamini–Hochberg procedure is applied globally to all 102 correlation tests computed in our analysis ($q = 0.012$). The *Label Distinguishability* correlation ($r = 0.65$, raw $p = 0.0215$) does not survive correction and should be read as suggestive rather than confirmed. The remaining correlations fail correction clearly, so their raw significance in the heatmap should not be interpreted as evidence. The concept-level relationship between *Gini Sparsity* and *Visual Distinguishability* ($r = 0.24$, $p = 0.067$, $n = 60$, Spearman $\rho = 0.22$, $p = 0.092$) likewise remains exploratory.

With only $n = 12$ concept spaces the statistical power is low, and only very strong effects ($|r| > 0.6$) can survive correction. The number of concept spaces is the main limiting factor for the borderline cases. The underlying computations for this analysis are provided in the accompanying analysis notebook (`correlation.ipynb`).

Table 6 reports the same corrections for the evaluation measure × metric correlation matrix, i.e. the 36 combinations of Figure 2, corrected within this family using Benjamini–Hochberg FDR control.

The evaluation measures form a pre-specified family, since each corresponds to one of the hypotheses formulated in Section 3.3.

Table 6: **Evaluation measure × metric correlation matrix with multiple-comparison correction.** Each cell shows the Pearson correlation r and the Benjamini–Hochberg FDR-adjusted q-value, corrected across the 36 tests of this family (all pairwise combinations of the human-evaluation measures and dictionary-learning metrics at the concept-space level, $n = 12$). No cell reaches $q < 0.05$ within this family alone. The strongest cell (*Label Consensus* × *Gini Sparsity*) reaches $q = 0.055$ here and $q = 0.012$ when corrected globally across all 102 tests computed in our analysis.

	Rel. L2	FID	Spars. Ratio	Gini Sparsity	Repr. Diversity	Stability
Click Frequency	$-0.11/0.87$	$-0.17/0.87$	$-0.14/0.87$	$-0.01/0.97$	$-0.12/0.87$	$-0.27/0.87$
Localizations Skips	$+0.62/0.22$	$-0.66/0.19$	$-0.39/0.73$	$-0.36/0.75$	$-0.54/0.37$	$-0.22/0.87$
Localization (IoU)	$-0.11/0.87$	$+0.15/0.87$	$+0.20/0.87$	$+0.18/0.87$	$+0.19/0.87$	$-0.25/0.87$
Label Consensus	$+0.07/0.89$	$+0.48/0.45$	$+0.60/0.23$	$+\mathbf{0.81}/0.055$	$+0.69/0.19$	$+0.21/0.87$
Label Disting.	$+0.12/0.87$	$+0.31/0.87$	$+0.38/0.73$	$+0.65/0.19$	$+0.50/0.45$	$+0.05/0.91$
Visual Disting.	$+0.28/0.87$	$-0.15/0.87$	$+0.12/0.87$	$+0.08/0.87$	$+0.09/0.87$	$-0.09/0.87$

E Impact of Individual Concepts on Dictionary Learning Metrics

To assess the metric scores on a more detailed level, we inspect how each individual concept impacts the *Gini sparsity* of the entire space. We do so by calculating the metric with a weighted average where the weight of each concept vector is the activation for one of the concepts.

Let $U \in \mathbb{R}_+^{M \times C}$ denote the concept-assignment matrix for one concept space, where M is the number of patches and C is the number of concepts. For a fixed concept c, the column

$$u^{(c)} = U_{:,c} = (u_{1,c}, \ldots, u_{M,c})^\top$$

contains the activation strength of concept c for every patch. We treat $\nu^{(c)}$ as a soft membership vector and use it as weights over all patches.

After normalization,

$$\bar{w}_m^{(c)} = \frac{u_{m,c}}{\sum_{r=1}^{M} u_{r,c}}, \qquad m = 1, \ldots, M,$$

the concept-level score is obtained by evaluating a metric with these weights. In abstract form,

$$S_c = \mathcal{M}\Big(Z^{(c)}, \bar{w}^{(c)}\Big),$$

where $Z^{(c)}$ is the data needed by the metric (U for sparsity metrics). The resulting scores be found in Figure 3.

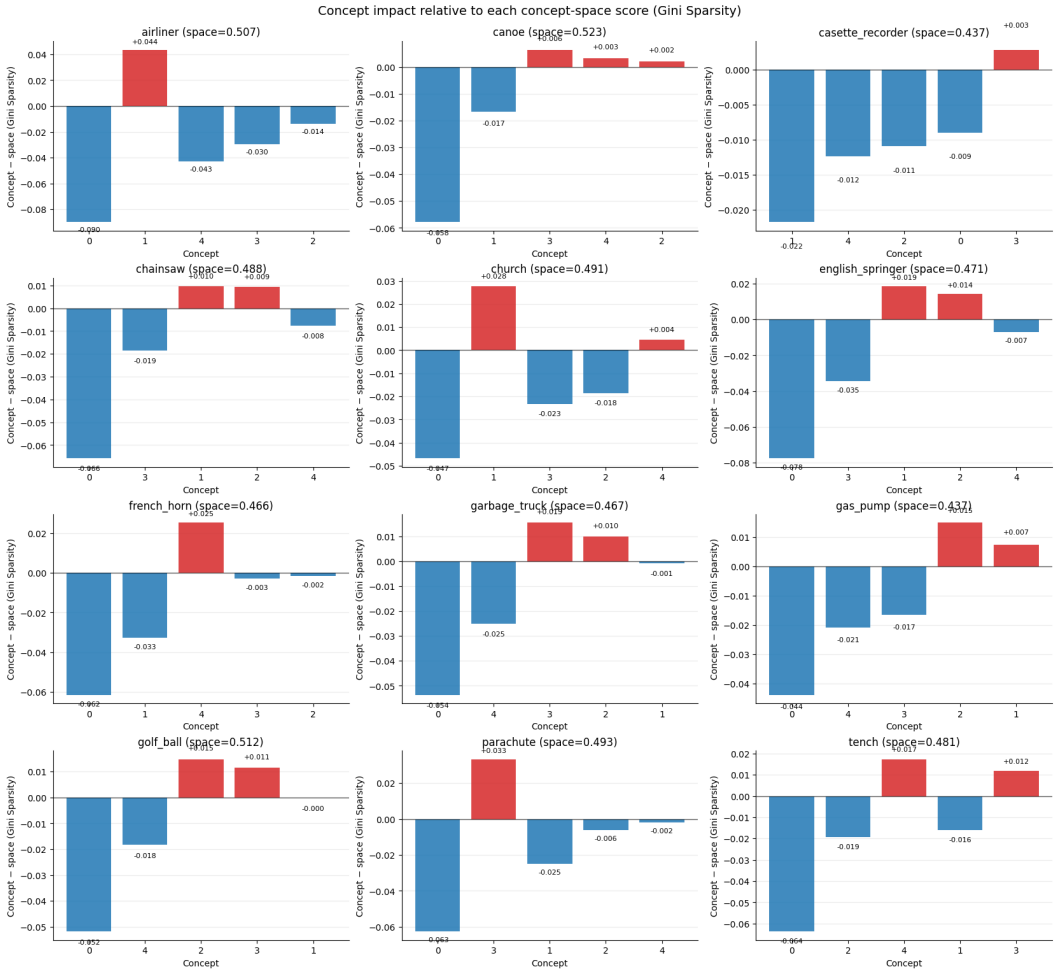

Figure 3: Impact of individual concepts for the gini sparsity measure *Gini Sparsity*

F Leave-One-Class-Out Robustness Analysis

To test the robustness of the reported correlations against class-level influences, we performed a leave-one-class-out (LOO) analysis: each of the twelve concept spaces was removed once, and all correlation coefficients between human-evaluation measures and dictionary-learning metrics were recomputed on the remaining eleven. Table 7 reports the measure-metric pairs with the largest absolute change, where r denotes the correlation over all twelve concept spaces, r_{loo} the correlation with the indicated concept space removed, and $\Delta r = r_{\mathrm{loo}} - r$.

Table 7: **Leave-One-Class-Out Correlation Analysis.** For each pair of a human-evaluation measure and a dictionary-learning metric, the correlation is recomputed after excluding one concept space at a time. The table lists the pairs with the largest absolute change in correlation. A large $|\Delta r|$ marks a concept space that strongly drives (or weakens) the respective correlation.

Human Measure	**DL Metric**	**Left out**	r	r_{loo}	Δr
Click Freq.	Sparsity Ratio	parachute	-0.14	0.20	0.34
Region Overlap (IoU)	Relative L2	airliner	-0.10	0.22	0.33
Region Overlap (IoU)	FID	english springer	0.14	0.47	0.32
Visual Distinguish.	FID	gas pump	-0.14	-0.45	-0.30
Skipped Localizations	Gini Sparsity	canoe	-0.35	-0.64	-0.28
Visual Distinguish.	Repr. Diversity	gas pump	0.08	-0.16	-0.25
Visual Distinguish.	Gini Sparsity	gas pump	0.08	-0.16	-0.24
Visual Distinguish.	Relative L2	airliner	0.27	0.03	-0.24
Click Freq.	Repr. Diversity	parachute	-0.11	0.12	0.23
Label Consensus	Stability	tench	0.20	0.44	0.23
Region Overlap (IoU)	Repr. Diversity	english springer	0.19	0.41	0.22
Region Overlap (IoU)	Gini Sparsity	cassette recorder	0.18	-0.03	-0.21

Overall, most correlations remained relatively stable, indicating that the observed relationships are not driven by a single concept space. Nevertheless, a small number of concept spaces had a noticeable influence. The largest positive change was observed for the correlation between *Click Frequency* and *Sparsity Ratio*, where excluding *parachute* increased the correlation from $r = -0.14$ to $r_{\mathrm{loo}} = 0.20$ ($\Delta r = 0.34$). Similarly, removing *airliner* strengthened the association between *Region Overlap (IoU)* and *Relative L2* from $r = -0.10$ to $r_{\mathrm{loo}} = 0.22$ ($\Delta r = 0.33$). Excluding *english springer* also increased the correlations of IoU with both *FID* ($\Delta r = 0.32$) and *Representation Diversity* ($\Delta r = 0.22$).

The most influential negative effects were primarily associated with the concept space *gas pump*: its removal decreased the correlation between *Visual Distinguishability* and *FID* from $r = -0.14$ to $r_{\mathrm{loo}} = -0.45$ ($\Delta r = -0.30$), with similar reductions for *Representation Diversity* ($\Delta r = -0.25$) and *Gini Sparsity* ($\Delta r = -0.24$); *gas pump* therefore contributes disproportionately to these relationships. Additional notable decreases were found for *Skipped Localizations* versus *Gini Sparsity* when excluding *canoe* ($\Delta r = -0.28$) and for *Visual Distinguishability* versus *Relative L2* when excluding *airliner* ($\Delta r = -0.24$). None of these changes reversed the sign of the two validated associations (*Label Consensus* and *Label Distinguishability* with *Gini Sparsity*), supporting the robustness of the paper's central conclusions while confirming that individual classes—most notably *airliner*—moderate the weaker correlation patterns discussed in Section 4.

G On Measuring Response Consistency

A distinctive property of our study is that participants are not treated as interchangeable annotators who are leveraged to synthesize an unambiguous ground-truth label (i.e., a class label). Instead, the measured quantity is the cognitive process of human comprehension itself. The variation in the responses is therefore not solely attributable to noise but rather to the genuine heterogeneity of human cognition. While it is operationally necessary to construct tasks that are familiar to conventional annotation tasks, with common objectives for all participants, the way in which their results should be interpreted differs; conventional annotation tasks are evaluated using inter-annotator agreement statistics. This is not an appropriate approach for our setting, as disagreement here partially reflects the

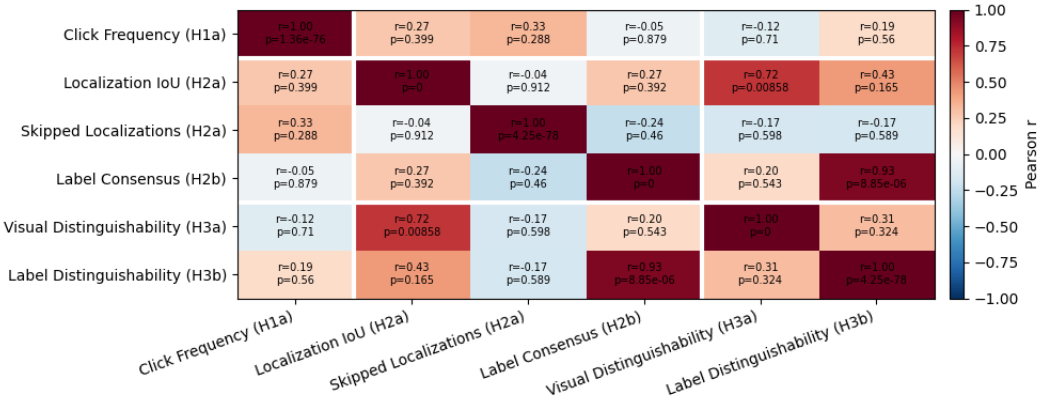

Figure 4: Correlation matrix featuring comparisons of human evaluation measures against themselves.

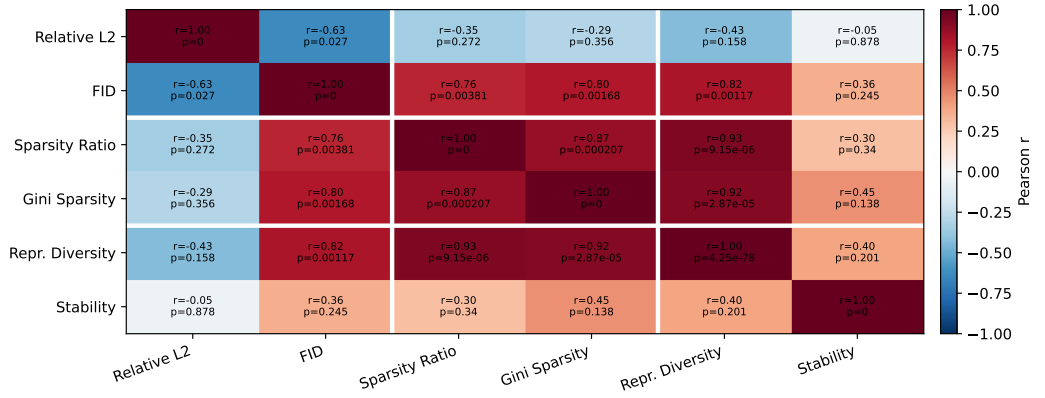

Figure 5: Correlation matrix featuring comparisons of dictionary learning metrics against themselves.

measured construct itself rather than measurement error. We therefore discuss participant variability as a limitation and avenue for future work (Section 5.3).

H Self Correlation of Dictionary Learning Metrics and Human Evaluation Measures

In addition to Figure 2, which illustrates the correlations between dictionary learning metrics and human evaluation measures, the subsequent Figures 4 and 5 present the relationships among metrics themselves and, separately, among evaluation measures themselves.

I VISUALIZATION OF SELECTED CONCEPT SPACES

This section presents visualizations of the concept spaces derived from the *golf_ball* and *church* classes using NMF, as illustrated in Figures 6 and 7, respectively. For the exact methodology of concept space generation refer to Section C.

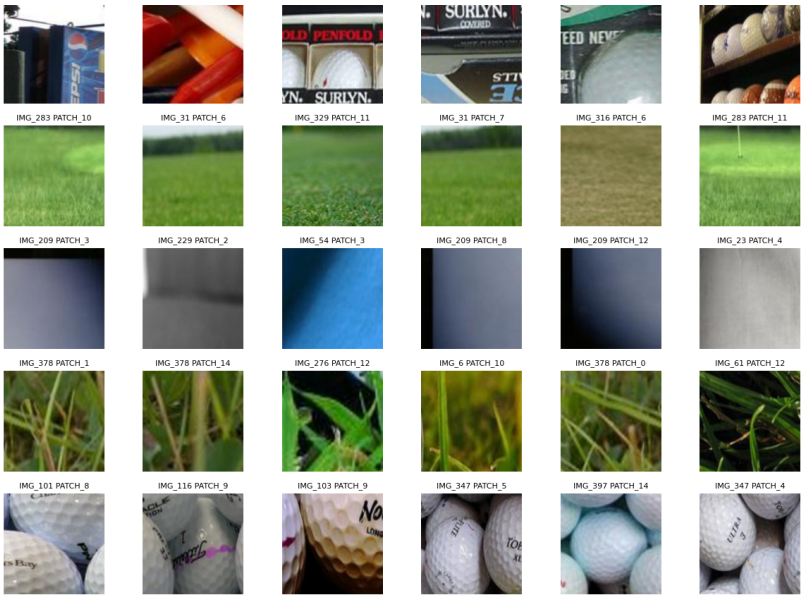

Figure 6: Concept space constructed from the *golf_ball* class: each row shows the 6 most representative patches for concepts one to five.

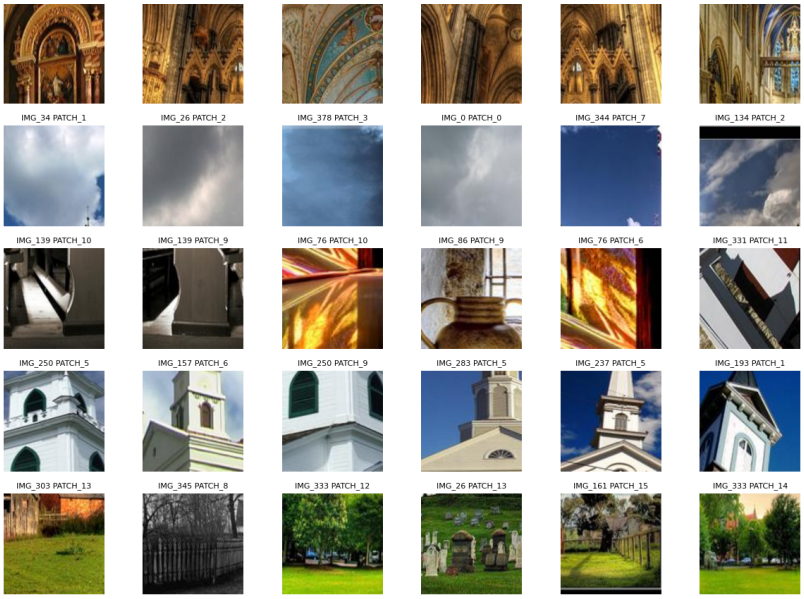

Figure 7: Concept space constructed from the *church* class: each row shows the 6 most representative patches for concepts one to five.

J EXEMPLARY PARTICIPANT RESPONSES

The following section provides representative examples of participant responses across all four tasks, illustrating how our study captures different aspects of concept comprehensibility.

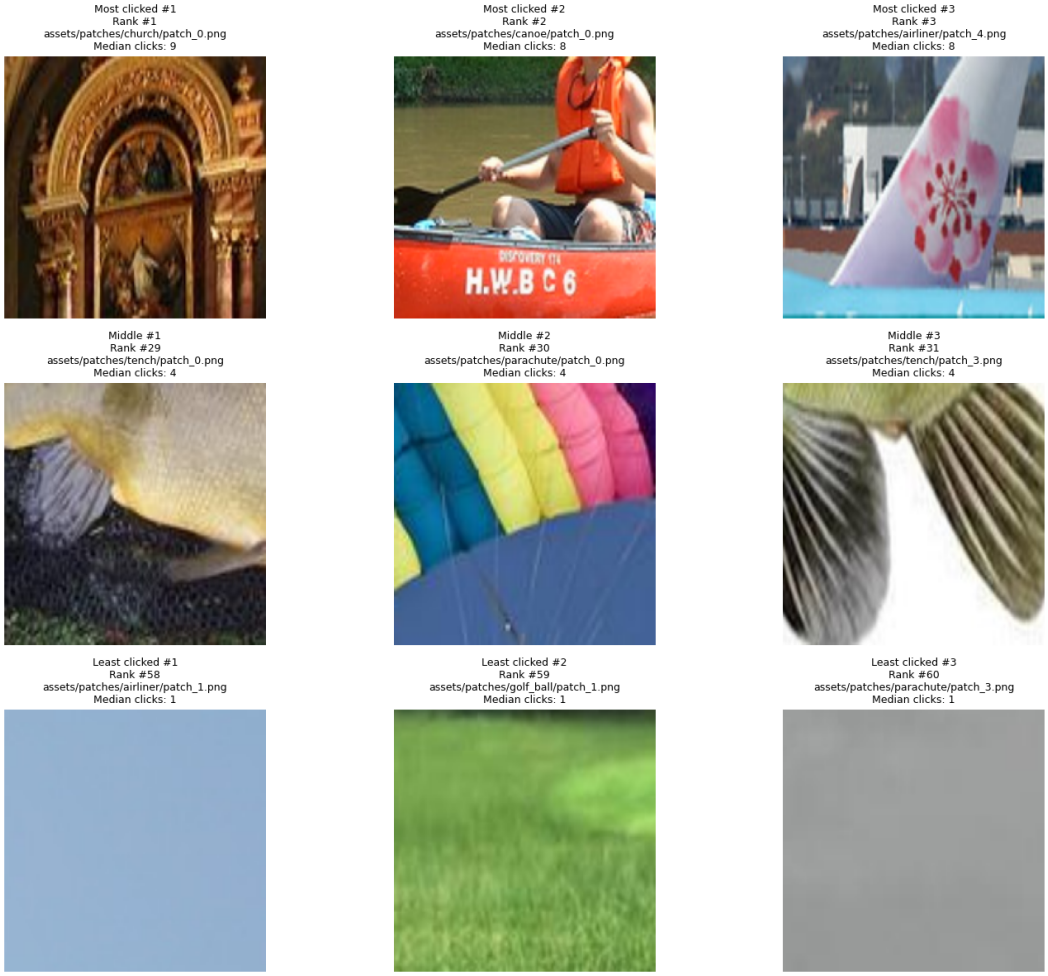

Figure 8: Each row shows patches with varying degree of informativeness. The first row shows the three most frequently clicked patches in our entire study while the second row consists out of patches that were clicked medially. The bottom row tends to be notably uninformative, with two out of the three patches being blank, displaying no recognizable content. These results can generally be interpreted as a successful implementation of our first task.

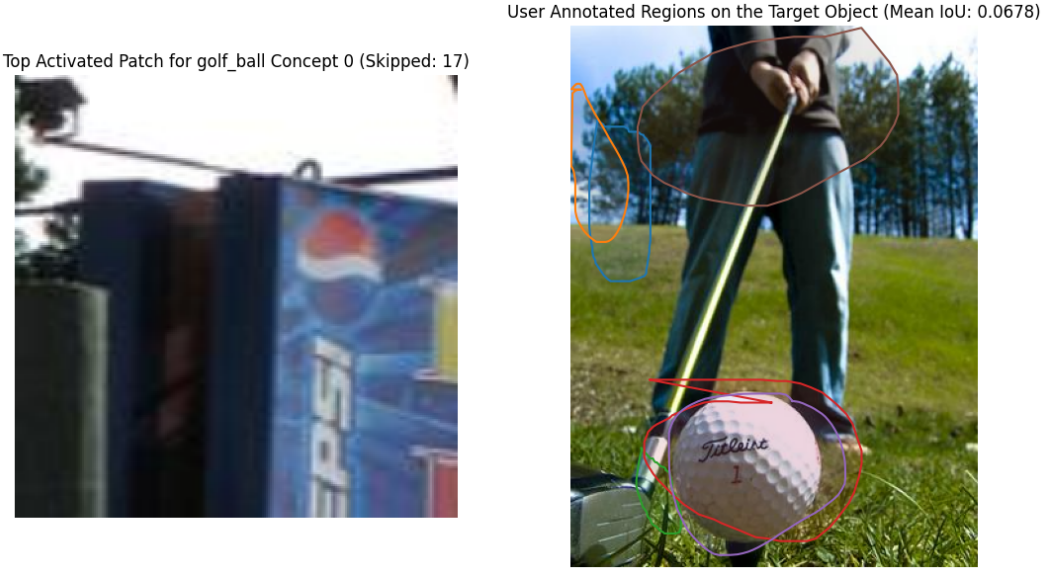

Figure 9: Overview of participants responses to the localization of a *golf_ball* patch (left side) on a reference image (right side). Given the content of the patch, participants expectedly chose to skip the task in 17 out of 23 responses. As indicated by the low IoU of 6.78%, the remaining submissions show measurable disagreement on the correct location. Together, both measures provide complementary insights into the comprehensibility of the patch and consequently its underlying concept.

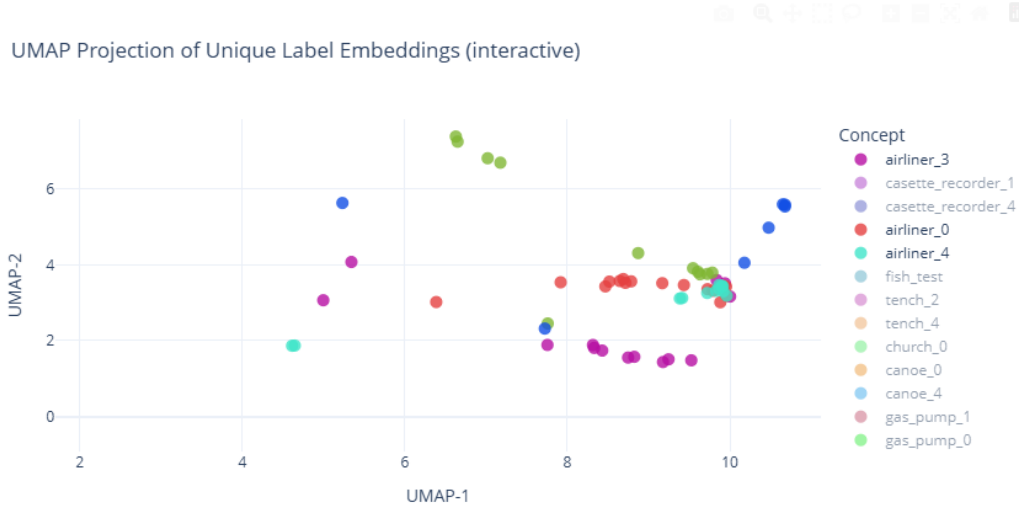

Figure 10: The plot shows a UMAP reduced visualization of text-embedded labels corresponding to all evaluated airliner concepts. It is evident that the labels associated with each concept form distinct global structures, where four out of the five groups share a common airliner-themed anchorpoint at $(3.7, 9.9)$. The purple outlier labels belong to a sky-themed concept and therefore plausibly posses no direct relation to the airliner semantics of the other labels.

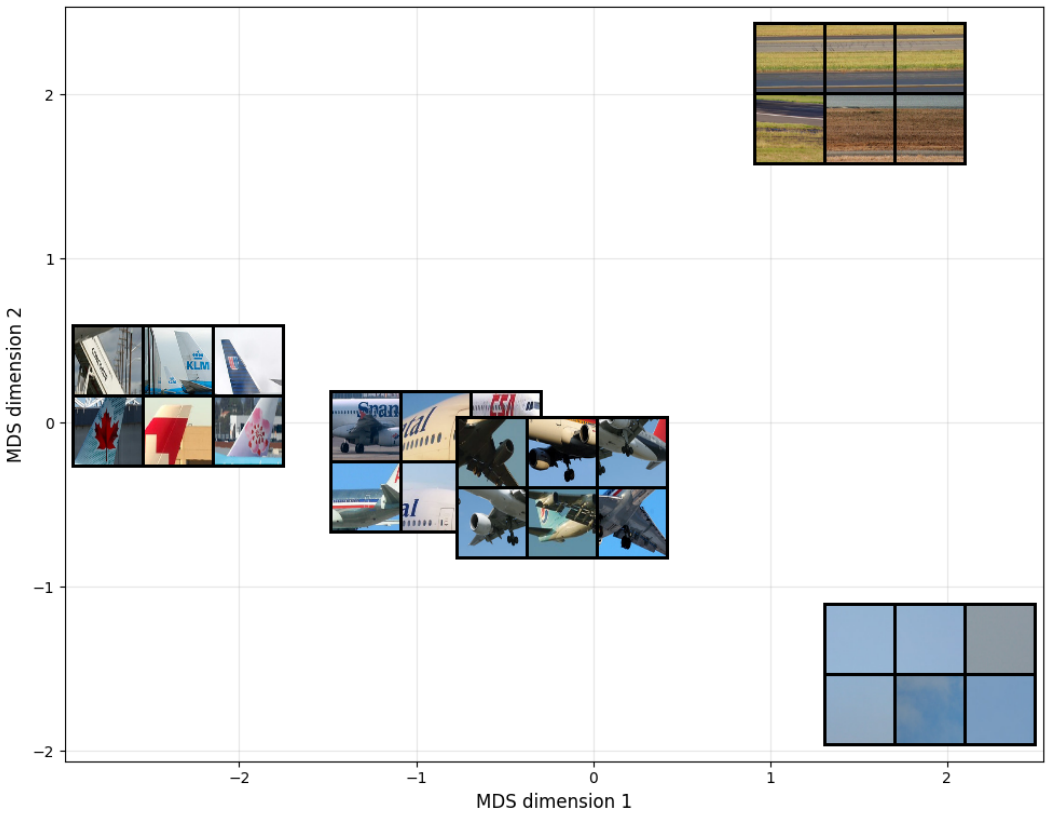

Figure 11: Visualization of airliner concepts in a low-dimensional embedding space obtained via Multidimensional Scaling (MDS) [Torgerson, 1965]. The spatial arrangement is derived from pairwise distinguishability ratings, such that conceptually similar aircraft are positioned closer together, while more dissimilar ones appear farther apart.

K INTERFACE SCREENSHOTS

The following figures exhibit screenshots of the web interface which the participants were confronted with. As outlined in Figure 1, all four tasks were operationalized with different interfaces, offering various options for interaction.

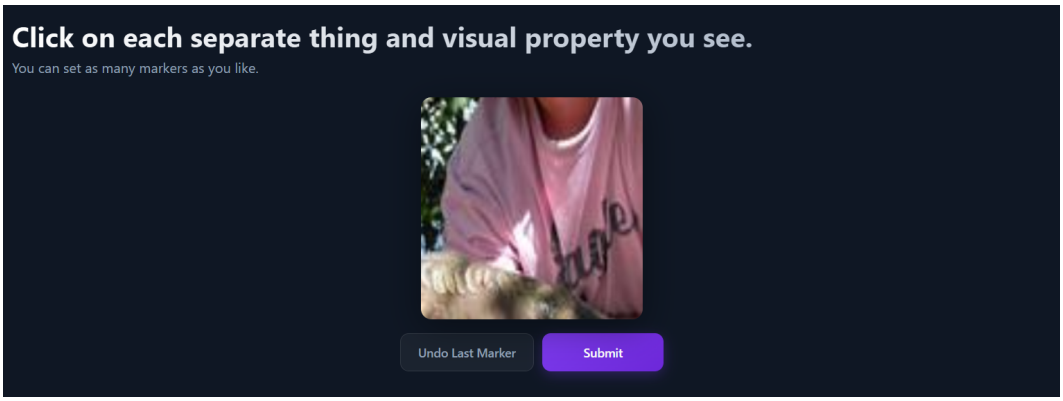

Figure 12: The screenshot shows the interface for Task 1. Participants can use the cursor to place markers on the patch and press the buttons below to either undo the last marker or submit their response.

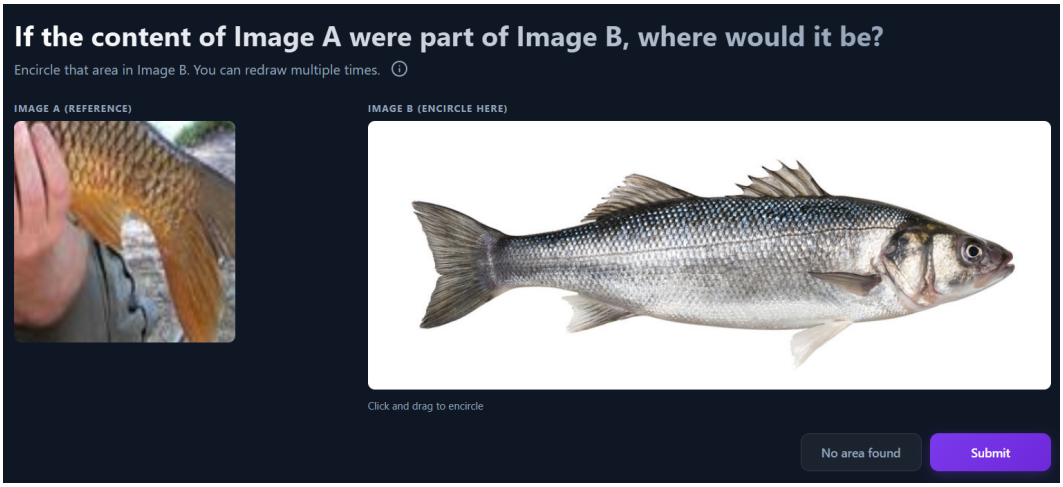

Figure 13: Screenshot of Task 2: participants were provided with a drawing tool to encircle the matching region on the target image. Lifting the mouse while encircling causes the polygon to close by connecting the first and last drawing points. Encircling a new area when a polygon is already present causes the old polygon to be deleted. Redrawing is therefore allowed. The buttons in the lower right can both be used to submit answers; however, pressing "No area found" submits the response with a skip attribute, omitting any drawn region.

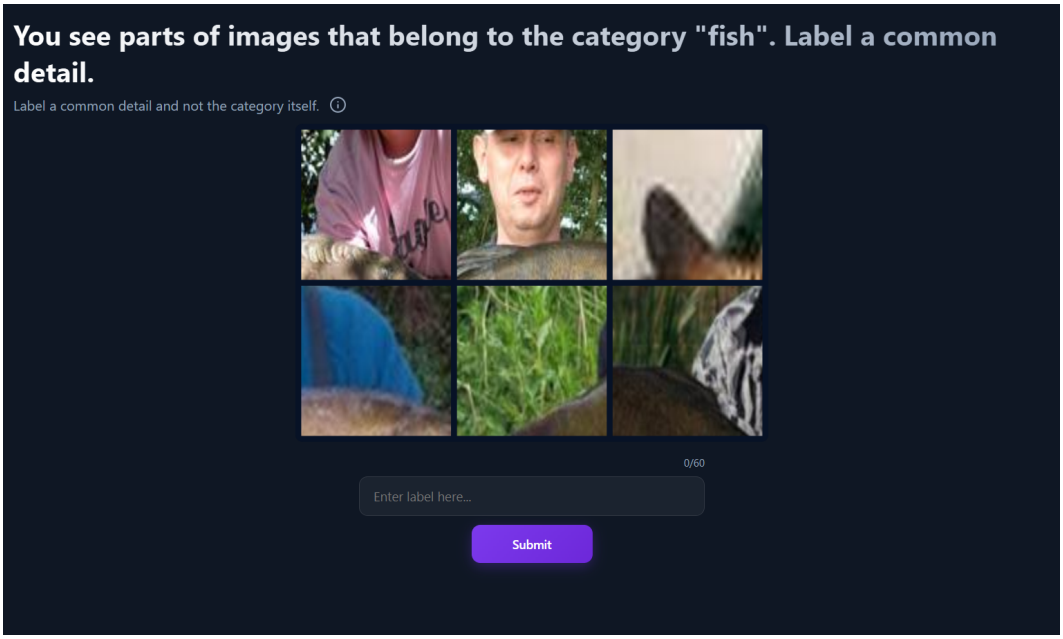

Figure 14: This screenshot of Task 3 shows how participants are presented with six representative concept patches. They are asked to submit a label with a maximum length of 60 characters, allowing for single-word responses as well as short phrases.

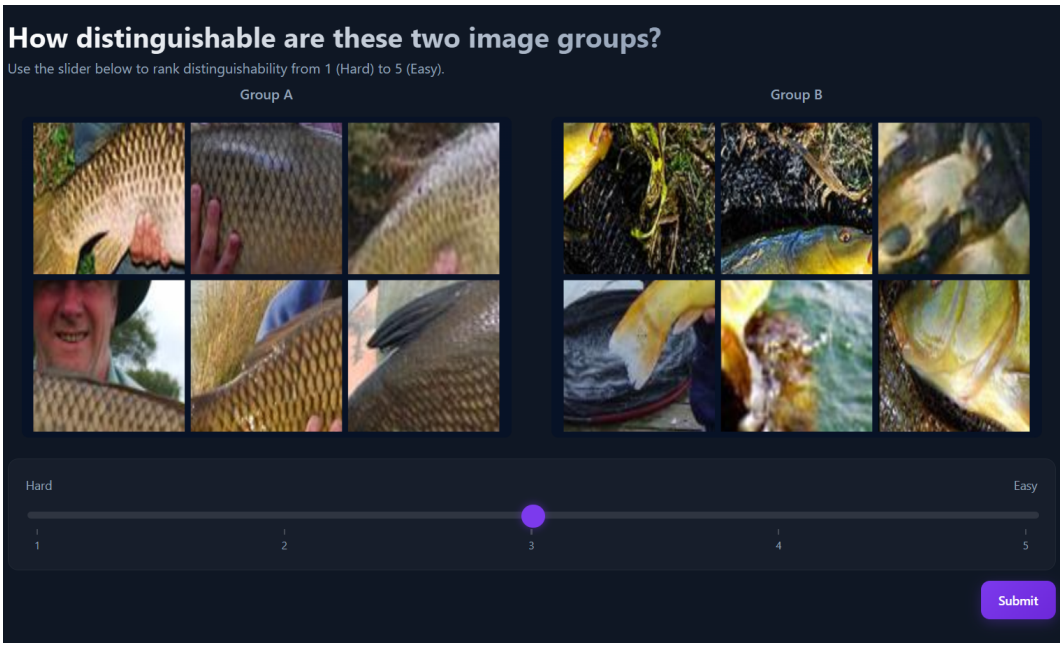

Figure 15: Task 4 shows a pair of concepts together with an interactive ordinal Likert scale. Participants can drag the slider point to express how visually distinguishable they perceive the concepts to be.