# AbsVis – Benchmarking How Humans and Vision-Language Models "See" Abstract Concepts in Images

**Tarun Tater**[1], **Diego Frassinelli**[2], **Sabine Schulte im Walde**[1]
[1]Institute for Natural Language Processing, University of Stuttgart, Germany
[2]MaiNLP, Center for Information and Language Processing, LMU Munich, Germany
{tarun.tater, schulte}@ims.uni-stuttgart.de
frassinelli@cis.lmu.de

## Abstract

Abstract concepts like *mercy* and *peace* often lack clear visual grounding, thus making it challenging to study how they are associated with images. To address this, we introduce AbsVis – a dataset of 675 images annotated with 14,175 concept–explanation pairs from humans and two Vision-Language Models (VLMs: Qwen and LLaVA), where each concept is supported by a textual explanation. We compare human and VLM attributions in terms of diversity, abstractness, and alignment, and find that humans attribute more varied concepts. AbsVis also includes 2,680 human preference judgments evaluating the quality of a subset of these annotations, showing that overlapping concepts (attributed by both humans and VLMs) are most preferred. Explanations clarify and strengthen the perceived attributions, both from humans and VLMs. Finally, we show that VLMs can approximate human preferences and use them to fine-tune VLMs via Direct Preference Optimization (DPO), yielding improved alignments with preferred concept–explanation pairs.

## 1 Introduction

Concrete concepts, such as *tree* and *car*, typically refer to objects easily depictable and recognizable in images. In contrast, abstract concepts such as *mercy* and *peace* mostly lack a direct visual representation, except if certain visual scenes may serve as prototypical depictions of abstractness – such as a lone person on a bench evoking *isolation*. However, such attributions are inherently diverse (Paivio et al., 1968; Kastner et al., 2020; Tater et al., 2024b), and shaped by personal experiences and cultural, emotional, and contextual factors. Figure 1 illustrates how the same image may evoke different abstract concept attributions, and why.

The general lack of an approach for dealing with abstract concepts has been a long-standing observation in psycholinguistic and cognitive research (Paivio, 1971; Pecher et al., 2011; McRae et al., 2018; Lynott et al., 2020, i.a.). Prior computational work has demonstrated its importance for visual grounding and multimodal alignment (Hill et al., 2014; Kiela and Bottou, 2014; Bhaskar et al., 2017; Köper and Schulte im Walde, 2017; Hessel et al., 2018; Pezzelle et al., 2021; Tater et al., 2024a), yet agrees on the challenging visual nature of abstract concepts in contrast to concrete concepts. Accordingly, even recent Vision-Language Models (VLMs) perform well on concrete object recognition, but often struggle to reason about visual abstractions, frequently defaulting to literal object interpretations (Menon and Vondrick, 2023; Hsu et al., 2025). Moreover, existing datasets are also limited in scope, or focus on reasoning over objects rather than explicitly evaluating abstract concepts (Chen et al., 2024a; Hsu et al., 2025).


Figure 1: Example of an image with two abstract concept attributions and corresponding explanations.

These limitations highlight the need to examine in depth how humans and VLMs attribute and reason about the depiction of abstract concepts, and whether recent methods like Direct Preference Optimization (DPO; Rafailov et al. (2023)) can align models with human preferences, despite a high diversity in attributions. This leads to the following research questions (RQs):

RQ1 Which abstract concepts do humans and VLMs attribute to an image, and how do they explain these attributions?

RQ2 Which attributions do humans and VLMs prefer, and can VLMs serve as reliable evaluators of human preferences for these attributions?

RQ3 Can DPO improve VLM alignment with human attributions, given the high diversity of abstract concepts?

To address these RQs, we present AbsVis, the first large-scale dataset of abstract concept attributions in images by humans and VLMs. It enables comparison of *which* concepts are attributed and *how* they are justified (explanations), and supports analyses of concept diversity, reasoning patterns, and preference alignment. AbsVis consists of:

1. **Concept – Explanations**: $10,125$ human-written and $4,050$ VLM-generated concept-explanation pairs across $675$ images[1]. The VLM generations are not intended as ground truth or fixed references, but as one set of model attributions for comparative analysis.
2. **Human Preferences**: $2,680$ human preference judgments on both concepts and explanations, enabling an analysis of which annotations are preferred, and how closely VLM preferences align with humans.
3. **DPO Fine-Tuning**: a baseline model to explore whether preference learning can help VLMs to better align with human preferences in reasoning about abstract concepts.

## 2 Related Work

The distinction between concrete and abstract concepts has been widely studied in psychology, linguistics, and computational modeling (Paivio et al., 1968; Barsalou et al., 2003; Frassinelli et al., 2017; Naumann et al., 2018; Frassinelli and Schulte im Walde, 2019; Charbonnier and Wartena, 2019; Schulte im Walde and Frassinelli, 2022; Tater et al., 2022, i.a.). Brysbaert et al. (2014) provided abstractness ratings for over $40k$ English lemmas, while recent work extended this approach to additional perceptual modalities (Lynott et al., 2020). Hessel et al. (2018) introduced a method to compute visual concreteness from multimodal data, finding that visually concrete concepts are easier for models to retrieve. However, VLMs still struggle with abstract concepts. Tater et al. (2024a) evaluate SigLIP, a VLM suitable for multi-label classification on images. They show that while SigLIP often predicts semantically related alternatives to the human-provided tags (e.g., synonyms, hypernyms, or co-hyponyms) of images, it less consistently predicts the exact tags, particularly for abstract concepts. This underscores both the diversity of abstract concept representations and the difficulties VLMs face in aligning with human annotations. Moreover, Hsu et al. (2025) find that VLMs often default to naming concrete objects, overlooking more subjective or conceptual associations. This motivates a closer examination of how abstract concepts are attributed and explained by humans versus VLMs in real-world visual contexts.

**Vision-Language Reasoning.** Modern VLMs such as CLIP (Radford et al., 2021), Qwen (Yang et al., 2024; Wang et al., 2024), and LLaVa (Liu et al., 2023, 2024) are trained to align images and text to generate image captions, answer questions, and classify images. Recently, various studies have explored visual reasoning (Pirsiavash et al., 2014; Park et al., 2018; Suhr et al., 2019; Ghorbani et al., 2019; Park et al., 2020, i.a.), some focussing on DPO and chain-of-thought (CoT) for VLM reasoning (Lu et al., 2022; Zhang et al., 2025). Hessel et al. (2022) introduces *Sherlock*, a dataset of images with {clue, inference} pairs, where human annotators identify visual clues in images to provide plausible inferences about the scene. However, their approach is based on inferencing from visible clues such as objects and actions.

**Datasets Related to Abstract Recognition.** Very few datasets attempt to evaluate abstract concept recognition in VLMs, and they are limited in scale and diversity. Some related datasets include: (a) *Visual Abstractions Dataset* (VAD) (Hsu et al., 2025), a small-scale dataset (180 images) only covering 12 abstract concepts in four pre-fixed categories, highly limiting generalization. (b) *CURE* Benchmark (Chen et al., 2024a): A dataset for CoT reasoning in vision models. It focuses on structured reasoning rather than free-form abstract concept explanations. (c) *CogBench* (Song et al., 2025): A benchmark for evaluating cognitive reasoning in LVLMs using semantically rich images. While it includes reasoning dimensions like mental state and event inference, it is not designed for concept attribution. (d) *ArtEmis* (Achlioptas et al., 2021), a dataset with human explanations of emotions evoked by images, however limited to art.

[1]The AbsVis dataset can be found here: AbsVis Dataset.

## 3 AbsVis Dataset

This section introduces AbsVis, a dataset capturing how abstract concepts are attributed and explained for images, especially by humans, without relying on predefined categories or reasoning templates as used in prior work. AbsVis includes two layers of annotation: (i) *concept-explanation* pairs and (ii) *preference judgments*. Section 3.1 outlines the selection of target nouns and associated images. Section 3.2 details the collection of abstract concept-explanation pairs from both humans and two VLMs. Section 3.3 analyses the diversity and abstractness of attributed concepts, and explanations across annotators. Section 3.4 evaluates image-concept representativeness, where humans assess how well an image fits an attributed concept.

### 3.1 Target Nouns and Images

We select 225 nouns – 75 abstract, 75 mid-range, and 75 concrete ones – based on their abstractness ratings from Brysbaert et al. (2014), which were collected on a scale of 1 (abstract) to 5 (concrete). To avoid biases and ensure a meaningful benchmark, we exclude nouns from the $1,000$ ILSVRC-2012 ImageNet categories (Russakovsky et al., 2015), as they are commonly used for VLM training. We also remove nouns with explicit content and those with high ambiguity ($> 4$ senses) using WordNet (Miller, 1995) as Tater et al. (2024b) find polysemy to be one of the reasons of significant diversity in images. After filtering, the selected nouns are in the ranges of 1.07-2 (abstract), 2.75-3.44 (mid-range), and 4.63-5 (concrete). The full list of 225 nouns is in Table 12 in the Appendix.

The images are sourced from the YFCC100M Multimedia Commons Dataset (*YFCC*; Thomee et al. (2016)), which is the largest publicly available user-tagged dataset containing $\sim$ 99 million images that exhibit diverse content and annotations. We extract images for each target noun from YFCC where the target noun is one of the user tags for the image. Then, we manually select three images per noun, ensuring that this selected subset of images reflect the same or closely related sense of the noun across its images. This results in a total of 675 images ($225 * 3$). Only images under permissive licenses are included (see Section 9.10 in the Appendix).

### 3.2 Concept - Explanation Annotation

Given an image, the aim is to obtain associated abstract concepts, along with explanations of why the image evokes these concepts. While the images are associated with abstract, mid-range, and concrete nouns, annotators are explicitly asked to provide abstract concepts, rather than concrete entities.

**Task Description.** Similarly to McRae et al. (2018), we instruct the annotators as follows: *Identify the concepts conveyed by the images. Avoid naming specific people or objects; instead, focus on summarizing the situation or the ideas, emotions, or feelings conveyed by the image. For each concept, provide a brief explanation of how it is conveyed by the image.*

**Annotators.** We collect annotations from both human annotators and two chat-based VLMs. For human annotations, we use Google Forms and recruit 220 participants via Prolific[2]. Each image was annotated by five participants, with each of them providing three concept-explanation pairs totaling 15 annotations per image. To ensure diversity and avoid repetition, each form only contained 18 images, and no participant saw more than one form. Further selection criteria are detailed in Section 9.10 in the Appendix.

**VLMs.** Using the same instructions as for humans, we include annotations from two recent open-source VLMs: LLaVa-next (llava-v1.6-mistral-7b-hf) (Liu et al., 2024) and Qwen2-VL-7B-Instruct (Wang et al., 2024), hereafter referred to as Llava and Qwen. These models[3] were selected for their performance relative to other models of similar scale, open-source availability, and wide usage. Full configuration details and justification for model selection are provided in Section 9.4 in the Appendix.

The final **AbsVis** dataset contains $10,125$ human-written concept-explanation pairs across 675 images. In addition, we include $4,050$ annotations generated by Qwen and Llava.

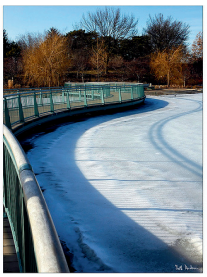
### 3.3 Dataset Analysis

**How do humans and VLMs attribute concepts?** We manually analyzed the annotations and observed that human annotators often drew on personal experiences, or emotions, contributing to the

[2] https://www.prolific.com/

[3] Throughout this paper, when we refer to "VLMs", we specifically mean Llava and Qwen. Future work should evaluate a broader set of VLMs to test the robustness of our findings.

Lamp
(concrete)



Curvature
(mid-range)

Contemplation
(abstract)

| | | | |
|---|---|---|---|
| H1 | **illuminate**: the light shines at night to brighten the street. | **cold**: bare trees, frozen lake, probably winter. | **admiration**: seems to be admiring the artwork. |
| H2 | **spring**: color of leaves on the trees. | **empty**: the environment is empty and there's no people. | **contemplative**: sat taking in the surroundings. |
| V1 | **autumn**: the image features vibrant orange and red leaves, which are characteristic of the autumn season. | **tranquility**: the stillness of the water and the calmness of the scene convey a sense of peace. | **solitude**: the person is alone in a spacious room, which emphasizes the feeling of solitude and introspection. |
| V2 | **light**: the street lamp is a central object in the image, and it is lit, which conveys the concept of light. the light from the lamp illuminates the surrounding area, creating a contrast with the darker background. | **isolation**: the bridge in the image appears to be empty, with no people or vehicles visible. this gives a sense of isolation, as there is no immediate sign of human activity. | **appreciation**: the man is taking the time to appreciate the art, indicating an interest in and respect for the creative expression displayed in the painting. |
| O1 | **tranquility**: **(H)** the stillness of the scene feels calm and peaceful. | **winter**: **(H)** along both sides of a curving walkway with turquoise railings - the bare branches of the trees in the distance also suggest winter. | **reflection**: **(H)** a space to see beautiful new images and muse upon them in relation to ones own life. |
| O2 | **(V)** the solitary street lamp against a blurred background of autumn leaves conveys a sense of calm and peacefulness. | **(V)** the image shows a snow-covered landscape, indicating that it is winter. the snow on the ground and the presence of trees without leaves suggest a cold season. | **(V)** the image conveys a sense of reflection, as the person is sitting alone and looking at a painting, possibly contemplating the scene depicted. |

Figure 2: Example annotations from the AbsVis dataset showing human-exclusive concepts (H1, H2), VLM-exclusive concepts (V1, V2), and overlapping concepts (O1, O2) – concepts attributed by both humans and VLMs, along with their explanations – attributed to three representative images corresponding to an abstract (*contemplation*), mid-range (*curvature*), and concrete noun (*lamp*).

diversity in human annotations. Quantitatively, Table 2 shows that overlap across human annotations is very low (10.12% overall), compared to higher overlaps for VLMs. Beyond this, humans and VLMs also differ in how they justify their attributions. As Table 3 indicates, human explanations are substantially shorter on average, whereas VLMs tend to produce longer, more elaborate descriptions. Example annotations illustrating these differences are shown in Section 9.7 in the Appendix. Some participants reported difficulties writing three abstract concepts for certain images. Notably, VLMs occasionally produce two concepts[4], followed by a single explanation addressing them (for e.g., *courage and dedication* as one concept for an image of a soldier). Figure 2 illustrates examples of human and VLM annotations, with concepts attributed by only one group (human-exclusive or VLM-exclusive), as well as overlapping cases – where both groups identify the same concept[5], reflecting both aligned and divergent interpretations.

We next compare human and VLM annotations quantitatively in terms of (i) the abstractness of attributed concepts, (ii) differences and overlap between the groups, and (iii) consistency within each group.

**How abstract are the concept labels?** Table 1 compares the abstractness of attributed concepts. Since abstractness ratings are only available for

[4]We consider responses of the form *concept A and concept B* as a single annotation.

[5]We treat singular and plural forms of the same base as equivalent.

| | % Concepts in ratings | | | Avg. rating ($\mu \pm \sigma$) | | |
|---|---|---|---|---|---|---|
| Category | Human | Llava | Qwen | Human | Llava | Qwen |
| Abstract | 92.81 | 66.81 | 84.44 | $2.67 \pm 0.31$ | $2.12 \pm 1.11$ | $2.64 \pm 0.94$ |
| Mid-range | 92.72 | 65.93 | 82.52 | $2.76 \pm 0.31$ | $2.17 \pm 1.20$ | $2.75 \pm 0.95$ |
| Concrete | 92.96 | 74.96 | 91.70 | $2.85 \pm 0.31$ | $2.37 \pm 1.03$ | $3.10 \pm 0.83$ |
| Overall | 92.73 | 69.00 | 86.33 | $2.76 \pm 0.32$ | $2.22 \pm 1.12$ | $2.83 \pm 0.93$ |

Table 1: Percentage of annotated concepts that could be directly matched to entries in the Brysbaert norms (37,058 English words), and average abstractness ratings ($\mu \pm \sigma$) for concepts from Humans, Llava, and Qwen annotations.

concepts that exactly match words in the Brysbaert norms, we report statistics over this subset. Human annotations have the highest coverage in the norms ($\sim 93\%$), compared to Qwen ($\sim 86\%$) and Llava ($69\%$). This lower coverage is due to VLMs producing more varied word forms or multi-word concepts, that do not match words in Brysbaert norms. Moreover, on the Brysbaert $1-5$ scale (where lower scores indicate greater abstractness), Qwen's concepts are slightly more concrete ($2.83$) than humans ($2.76$), while Llava's are considerably more abstract ($2.22$). VLMs also exhibit greater variability ($\sigma = [0.83, 1.12]$), as also shown in the box plot in Figure 3.

**How aligned are humans and VLMs in attributing concepts?** We assess alignment by computing the percentage of overlapping concepts per image, as shown in Table 2. Human-human overlap is defined as the proportion of the $15$ concepts ($5$ annotators $* 3$) that were attributed by more than one annotator. Since annotators differ across forms and overall agreement is low, we report overlap instead of inter-annotator agreement. For human-VLM and VLM-VLM comparisons, we compute the percentage of VLM concepts (out of 3) that appear in the other group's concepts (i.e., overlapped concepts $/3$). VLMs show higher internal overlap ($21-24\%$), likely due to shared pretraining approaches. In contrast, human annotators exhibit lower internal overlap ($9.63\%$ to $11.11\%$), reflecting greater diversity and subjectivity in human interpretations, where individuals often associate different abstract concepts with the same image. These values should be interpreted in light of the difference in annotation volume: humans attribute $15$ concepts per image, while each VLM attributes only three concepts.

We also compute pairwise word2vec similarity between concepts in each group to assess how semantically close they are. As detailed in Section 9.3 in the Appendix, VLM concepts are only slightly more similar to each other and marginally more aligned with the target noun than human concepts.

| Category | Human-Human | Human-Llava | Human-Qwen | Llava-Qwen |
|---|---|---|---|---|
| Abstract | 9.63 | 22.37 | 29.78 | 21.33 |
| Mid-range | 9.63 | 20.89 | 27.70 | 21.33 |
| Concrete | 11.11 | 29.63 | 31.56 | 23.70 |
| Overall | 10.12 | 24.30 | 29.67 | 22.12 |

Table 2: Percentage of overlapping concepts across annotator groups.

**How do human explanations compare to VLM ones?** While the previous analysis focused on concepts, we now examine their explanations along two dimensions: length (number of words) and semantic similarity. As shown in Table 3, human explanations are substantially shorter, averaging $12.25 \pm 1.46$ words. In contrast, Llava and Qwen produce much longer explanations ($27.91 \pm 6.04$ and $19.91 \pm 3.22$ words, respectively). This pattern holds across all abstractness categories (abstract, mid-range, concrete). Figure 4 shows the distribution of lengths. Second, we assess the semantic similarity of explanations of overlapping concepts using sentence transformers (Reimers and Gurevych, 2019), Similarity scores range from $-1$ (complete dissimilarity) to $1$ (identical). Despite differences in length, humans and VLMs produce reasonably similar explanations for overlapping concepts, with average similarity scores around $[0.56, 0.64]$. However, these scores may be influenced by explanation length, since they are not normalized for length: longer texts can mention more concepts or share more words, inflating similarity without stronger semantic alignment.

### 3.4 Image - Concept Representativeness

While the previous sections examined what concepts an image evokes, we now investigate the reverse perspective: *whether an image is a good representation of an attributed concept*. This allows us to examine how well human and VLM attributed

| | Avg. exp. length | | | Avg. exp. sim. | |
|---|---|---|---|---|---|
| Category | Human | Llava | Qwen | H–L | H–Q |
| Abstract | 12.42 | **28.80** | 20.90 | 0.60 | 0.56 |
| Mid-range | 12.12 | **28.07** | 19.64 | 0.64 | 0.63 |
| Concrete | 12.21 | **26.87** | 19.19 | 0.63 | 0.65 |
| Overall | 12.25 | **27.91** | 19.91 | 0.62 | 0.59 |

Table 3: Average explanation length and average similarity of explanations of overlapping concepts. (H: Human, Q: Qwen, L: Llava, exp.: explanation, sim.: similarity)

concepts align with an image or how *stretched* or *subjective* they are.

To evaluate this, we collect human judgments for a subset of 63 images across 21 nouns (7 each from abstract, mid-range and concrete). Five annotators rate whether an image is *strongly*, *moderately*, *weakly*, or *not at all* representative of an attributed concept. For each image, we evaluate five concepts: two randomly sampled from human-exclusive, two from VLM-exclusive, and one overlapping (see Section 4). Inter-annotator agreement, measured using Fleiss' Kappa (Fleiss, 1971), is 0.19, indicating high subjectivity in image-abstract concept mappings. The diversity in annotations and the low agreement (as expected) suggest that defining a gold standard for *good* attributions or defining a single best attribution is inherently challenging.

Hence, we use preference judgments for ranking attributions (see Section 4). We also assess how VLMs rate image-concept representativeness for all 675 images. We find that both Qwen and Llava struggle to strongly associate human concepts to images, highlighting the subjectivity of human interpretations that VLMs fail to capture. Detailed results are reported in Section 9.5 in the Appendix.

## 4 Human Preferences

We analyze whether humans prefer annotations from other humans or from VLMs, at both the concept level and the explanations level. Since defining a single *best* attribution is inherently subjective, we rely on preferences to evaluate which annotations are perceived as *more* or *less* representative of an image. In addition, we test whether VLMs can act as evaluators for this task. If their preferences align closely with those of humans, they could serve as scalable proxies for preferences, potentially reducing the need for costly manual annotations.

**Collecting Preference Judgments.** We use best-worst scaling (BWS) (Kiritchenko and Mohammad, 2017) to collect comparative human judgments on concepts and explanations. In each BWS tuple, annotators are shown an image and a set of four candidate attributions (either concepts or explanations) and asked to select the most and least representative options. This relative ranking approach avoids the inconsistencies of absolute scoring and enables more reliable preferences. For (a) concept preferences, participants assess: *which concept best and least represents the core idea of the image*, and (b) for explanations, they assess *which one best describes the image and which one provides the worst description*, out of 4 candidates. The phrasing was kept consistent with the original concept-explanation attribution prompts to avoid any bias.

To ensure sufficient coverage and comparisons, we construct $2N$ annotation tuples per image, where $N = \{9, 10\}$ for concepts and $N = \{10, 11, 12, 13\}$ for explanations. Each candidate attribution appears in $7 - 8$ different tuples, and each tuple is rated by three annotators. In total, we have $1,242$ concept-level tuples and $1,438$ explanation-level tuples, across 63 images, spanning 21 nouns (7 each from abstract, mid-range, and concrete), with the list of nouns provided in the Section 9.10.1. Exact annotation prompts and screenshots of the BWS forms are also provided in Sections 9.8 and 9.10.2 in the Appendix.

**Concept Selection.** For each image, we randomly sample 10 concepts: 6 human-exclusive, 2 VLM-exclusive, and 2 overlapping. Only concepts with ratings in the Brysbaert norms are included. As shown in Table 1, coverage is slightly lower for VLM-attributed concepts, especially from Llava, meaning some VLM concepts are excluded from evaluation. If two VLM-exclusive concepts are not available, we substitute additional overlapping concepts, and vice versa[6]. For 9 out of 63 images, we could only retrieve 3 concepts from the VLM-exclusive and overlapping pools combined.

**Explanation Selection.** Explanations are chosen for the 10 selected concepts per image. We have $10 - 13$ total explanations, since overlapping concepts have one each from a human and a VLM. If multiple human annotators provided the same concept, we randomly sample one explanation. Only explanations between 5 and 40 words are included. Since they are inherently tied to their concepts,

[6]We always include 6 distinct human-exclusive concepts.

and to ensure consistency in formatting, all explanations are standardized as: "The image conveys [Concept] because: [Explanation]."

Following Kiritchenko and Mohammad (2017), we convert best-worst annotations into continuous scores by calculating the number of times an attribution was chosen as "best" minus the number of times it was chosen as "worst" for an image. The resulting scores are normalized between $-1$ to 1. We use these scores to examine which attributions are preferred and whether explanations can influence concept-level preferences. We hypothesize that overlapping concepts will be preferred, as they may reflect more immediately recognizable associations. Additionally, human-exclusive concepts may initially appear more subjective, making them less preferred. However, when accompanied by explanations, annotators may view them as more justified, potentially shifting preferences.

**Annotators.** We recruited 186 participants via Prolific[7] and collected responses using Google Forms. To ensure diversity and avoid repetition, each annotator was assigned a form containing either 51 concept-level tuples or $38-40$ explanation-level tuples. No annotator saw the same image or concept more than once, and no annotator was assigned more than one form.

### 4.1 Results

We examine which annotations – human-exclusive, VLM-exclusive, or overlapping are most preferred by humans, by evaluating the top-1 ranking. Table 4 presents human preferences both at the concept and at the explanation level.

**Concept Preferences.** Consistent with our hypothesis, we find that overlapping concepts are most preferred overall ($26/63$ images). Notably, VLM-exclusive concepts (24) are preferred more often than human-exclusive ones (13), suggesting that VLMs generate highly representative concepts even when they do not overlap with human concepts. In particular, out of 21 images for abstract nouns, VLM-exclusive concepts (10) are selected more frequently than overlapping (5) or human (6) concepts. One possible reason for the lower preference for human concepts is that the most representative ones may already be present in the overlapping category, and the remaining human-exclusive ones may be more subjective, drawing heavily from personal experiences, that, while valid, are less universally interpretable without additional context. Examples are provided in Section 9.7 in the Appendix. We consider only top-1 preferences because the number of candidates from each group (overlap, human-, VLM-exclusive) differs.

| | Concept | | | Explanation | | |
|---|---|---|---|---|---|---|
| Category | Overlap | VLM | Human | Overlap | VLM | Human |
| Abstract | 5 | **10** | 6 | **11** | 7 | 3 |
| Mid-range | **11** | 6 | 4 | **12** | 6 | 3 |
| Concrete | **10** | 8 | 3 | **16** | 5 | – |
| Overall | **26** | 24 | 13 | **39** | 18 | 6 |

Table 4: Human preferences for top-ranked concept and explanation annotations (out of 63 images), comparing overlapping, human-, and VLM-exclusive options.

**Explanation Preferences.** When explanations were included, we see a notable shift in human preferences. Overlapping concept-explanation pairs were even more strongly preferred ($39/63$ overall), reinforcing the idea that these represent more recognizable associations. Moreover, VLM explanations (18) were also preferred over human explanations (6). Among the 39 overlapping top-1 preferences, VLM explanations were favoured 31 times, compared to 8 from humans. When considering all ranks (not just top-1) VLM explanations were preferred in 80 out of 98 overlapping cases across 63 images. One contributing factor may be explanation fluency. VLMs produce longer sentences, as earlier discussed in Table 3, which annotators may find more structured or easier to follow, potentially contributing to their higher preference scores whereas human annotators might produce less structured descriptions.

To test this, we compute a Spearman correlation between explanation length and preference rank, where a higher rank (numerically lower number) indicates a stronger preference (i.e., rank 1 = most preferred). We find a negative correlation ($\rho = -0.44, p < .0001$), which means that longer explanations tend to receive higher ranks. This moderately indicates that as preference becomes stronger (i.e., rank is higher - numerically lower number), explanation length tends to increase.

Another factor may be that human explanations sometimes rely on implicit or personal associations, which could make them less consistently preferred in a comparative setting. We also test whether more abstract concepts are preferred. At the concept level, we find a weak negative corre-

[7] See Section 9.10 in the Appendix for selection criteria, annotation forms and demographic details.

lation between concept abstractness and human preference ranks ($\rho = -0.11$, $p = .0069$). A $t$-test further confirmed this trend, with top-1 preferred concepts rated as significantly more abstract on average ($t = 4.16$, $p < .001$). However, this effect disappears when considering preferences over concept-explanation pairs.

**Preference Reliability** To assess the reliability of our BWS annotations, we compute split-half reliability following (Kiritchenko and Mohammad, 2017). For each image, we split the annotations into two equal halves randomly, compute BWS scores for each set, and calculate the Pearson correlation between them. The average Pearson correlation across all images, using Fisher's z-transformation shows strong internal consistency: 0.88 for concept-only scores and 0.87 for concept-explanation scores (with $p$-value $< 10^{-6}$), indicating high reliability of human preference judgments.

**Preference Shifts.** We investigate whether preferences shift when moving from concept-only to concept-explanation pairs. A preference shift is defined as a change in concept-explanation rank greater than one standard deviation from the mean rank change from the previous concept ranks. Since the number of candidates differs across these settings, we use rank (rather than score) as a more comparable measure. Table 5 presents the percentage of preferences that improve or degrade in rank when explanations are added to concepts; we observe rank improvements across all groups. Overlapping concepts show the largest positive shift (36.75%), followed by human-exclusive (26.46%). Notably, human attributions exhibit the lowest degradation rate (4.23%), compared to 13.67% for both overlap and VLM categories. This suggests that explanations often strengthen human-exclusive concepts, making them more interpretable without introducing additional confusion. In contrast, while explanations also improve many overlapping and VLM attributions, they may occasionally overspecify or misalign with the concept, leading to more frequent preference drops.

| Annotator | Improvement ↑ | Degradation ↓ |
|---|---|---|
| Human | 26.46% | **4.23%** |
| Overlap | **36.75%** | 13.67% |
| VLM | 20.51% | 13.67% |

Table 5: Percentage of rank shifts from concept-only to concept-explanation preferences.

### 4.2 VLMs as a Judge

To assess whether VLMs can be reliable evaluators for preferences, we compute the correlation between human and VLM preferences at both concept and explanation levels. As shown in Table 6, we observe moderate to strong correlations overall, with explanation-level preferences showing slightly higher agreement than concept-level preferences ($\sim 0.7 - 0.8$ with $p$-value $< 10^{-6}$). Qwen shows slightly higher alignment with human preferences than Llava, indicating that it more closely reflects human judgments.

| | Concept-level | Explanation-level |
|---|---|---|
| Llava-Human | 0.71 | 0.75 |
| Qwen-Human | 0.78 | 0.79 |

Table 6: Spearman's correlation ($\rho$) between human and VLM preferences at concept and explanation levels.

While these results suggest that despite this subjectivity of human annotations, these two VLMs capture human preference trends reasonably well, they should be used as approximate, and not absolute proxies for human judgment in subjective tasks like these. Additional results, including top-1 preferences (Table 9) and significant rank shifts (Table 10) for VLMs are reported in the Section 9.6 in the Appendix.

## 5 DPO Fine-Tuning

We investigate whether VLMs can be fine-tuned to better align with human attributions for abstract concepts. Building on our previous finding that VLMs' preferences moderately align with human preferences, we use these as a scalable supervision signal for fine-tuning VLMs for human attributions. We then fine-tune Qwen and Llava using DPO, a preference-optimization algorithm that adjusts a generation policy to favor preferred outputs over rejected ones. Specifically, we use Qwen to judge human concept-explanation attributions and train Llava, and vice versa. This cross-model supervision avoids biasing a model with its own preferences, and allows us to simulate a more realistic setup where preference data is external, mimicking how human judgments would be used in practice. This allows us to construct preference pairs for all human attributions for the full dataset of 675 images, substantially expanding the amount of supervision available compared to our original 63 image human-preference subset. Importantly,

DPO was applied only on human-attributed concept–explanation pairs; the VLMs were used solely to provide preference judgments over these human attributions.

Annotations are first sorted by preference scores per image and filtered to retain only the top and bottom four. Preference pairs are then formed by combining each top-ranked annotation with each bottom-ranked one, and a pair is included only if the top item's score exceeds the bottom's by at least a threshold $\delta$, where $\delta = 1.96 \cdot \frac{\sigma}{\sqrt{N}}$, with $\sigma$ as the standard deviation of scores for an image and $N$ the number of human concept-explanation pairs (max: 15). We perform $5-$fold cross-validation using an $80 - 20$ train-test split on images, and fine-tune both models using LoRA adapters (Hu et al., 2022) for three epochs with varying hyper-parameters. We evaluate our DPO-fine-tuned models (Llava and Qwen) against their respective base models on proxy preference pairs derived from each other: Llava is evaluated on Qwen-generated preferences, and Qwen on Llava-generated ones. We report the average log-probability margin, defined as the difference in log-probability the model assigns to the chosen explanation versus the rejected one. The base Llava's mean margin is $-1.89 \pm 1.74$, while the DPO-fine-tuned Llava's mean margin is $0.70 \pm 1.78$ (paired $t$-test: $t(4) = -26.59$, $p < 10^{-3}$), showing a significant shift from negative to positive. For Qwen, the base model's mean margin is $0.63 \pm 1.88$, which increases to $1.11 \pm 1.86$ after DPO fine-tuning (paired $t$-test: $t(4) = -11.31$, $p < 10^{-3}$), indicating consistent improvements. These results suggest that DPO fine-tuning improves alignment with proxy preferences from another VLM, and highlights its potential for scaling to subjective tasks when more preference data becomes available.

## 6 Conclusion

We presented AbsVis, a dataset designed to study how humans vs. VLMs attribute abstract concepts to images and justify these attributions. AbsVis contains $10,125$ human and $4,050$ VLM-generated annotations, along with human preference judgments over a curated subset. Our analysis shows that humans and VLMs often attribute different concepts to the same image, with relatively low overlap even among humans. VLMs tend to produce longer explanations, which are typically preferred over human-written ones; concepts attributed by both humans and VLMs are the most preferred. We further evaluate VLMs as preference judges and find moderate alignment with human preferences, particularly for explanations. Leveraging this, we fine-tune VLMs via Direct Preference Optimization for human attributions, and observe consistent improvements in reward margins.

## 7 Limitations

We took several measures to maintain the quality and reliability of our dataset, but some limitations remain. First, although we filtered human responses which looked like AI-generated responses, where concept-explanation pairs showed patterns like excessive verbosity, unrelated object references, multiple concepts (e.g., *concept A and concept B*), or extremely quick responses, we cannot entirely rule out instances where some annotators may have used AI assistance. Second, VLM outputs can be sensitive to prompt variations, which may affect response consistency. Although our dataset provides a strong benchmark, our preference evaluations are limited to concepts with Brysbaert ratings, which may miss some annotations. Moreover, variations in explanation formatting may have influenced human preferences, with some explanations receiving lower scores due to readability rather than content quality. To mitigate this, annotators were explicitly instructed to '*ignore minor grammatical errors and focus on the content of description*' in the annotation forms. Third, we focus on two VLMs to facilitate comparisons with humans. While these models work well for our tasks and show moderate alignment with human preferences, other VLMs could reveal different behaviors and correlation patterns on similar attribution tasks. Finally, we apply DPO for fine-tuning due to its compatibility with preference data. However, other fine-tuning approaches are worth exploring in future work to better capture subjectivity.

We find a lot of diversity in associated concepts and explanations across annotators, reflecting the subjective nature of this task. This raises a broader question of what the optimization or alignment goal should be for this type of task – should models replicate individual interpretations, or aim for broader conceptual consensus.

## 8 Ethics Statement

We do not see any ethical issues related to this work. All annotations involving human participants were

fairly compensated (£9 per hour), and participation was voluntary. We did not collect any information that can link the data back to the participants, and they were fully informed why the annotations were being collected. All modeling experiments were conducted using open-sourced libraries, which received proper citations.

**Use of AI Assistants.** The authors acknowledge the use of AI assistants solely for correcting grammatical errors, formatting table boundaries, enhancing the coherence of the final manuscripts, and providing assistance with coding.

## Acknowledgments

This research was supported by the DFG Research Grant SCHU 2580/4-1 *Multimodal Dimensions and Computational Applications of Abstractness*.

## References

Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. 2021. Artemis: Affective Language for Visual Art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Lawrence W Barsalou, W Kyle Simmons, Aron K Barbey, and Christine D Wilson. 2003. Grounding Conceptual Knowledge in Modality-Specific Systems. *Trends in Cognitive Sciences*, 7(2).

Sai Abishek Bhaskar, Maximilian Köper, Sabine Schulte im Walde, and Diego Frassinelli. 2017. Exploring Multi-Modal Text+Image Models to Distinguish between Abstract and Concrete Nouns. In *Proceedings of the IWCS Workshop on Foundations of Situated and Multimodal Communication*.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand generally known English Word Lemmas. *Behavior Research Methods*, 64.

Jean Charbonnier and Christian Wartena. 2019. Predicting Word Concreteness and Imagery. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*.

Yangyi Chen, Karan Sikka, Michael Cogswell, Heng Ji, and Ajay Divakaran. 2024a. Measuring and Improving Chain-of-Thought Reasoning in Vision-Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How Far Are We to GPT-4V? Closing the Gap to Commercial Multimodal Models with Open-Source Suites. *Science China Information Sciences*, 67(12).

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024c. Internvl: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

Joseph L Fleiss. 1971. Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76(5).

Diego Frassinelli, Daniela Naumann, Jason Utt, and Sabine Schulte im Walde. 2017. Contextual Characteristics of Concrete and Abstract Words. In *Proceedings of the 12th International Conference on Computational Semantics*.

Diego Frassinelli and Sabine Schulte im Walde. 2019. Distributional Interaction of Concreteness and Abstractness in Verb–Noun Subcategorisation. In *Proceedings of the 13th International Conference on Computational Semantics*.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards Automatic Concept-Based Explanations. *Advances in Neural Information Processing Systems*, 32.

Jack Hessel, Jena D Hwang, Jae Sung Park, Rowan Zellers, Chandra Bhagavatula, Anna Rohrbach, Kate Saenko, and Yejin Choi. 2022. The Abduction of Sherlock Holmes: A Dataset for Visual Abductive Reasoning. In *European Conference on Computer Vision*.

Jack Hessel, David Mimno, and Lillian Lee. 2018. Quantifying the Visual Concreteness of Words and Topics in Multimodal Datasets. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Felix Hill, Roi Reichart, and Anna Korhonen. 2014. Multi-Modal Models for Concrete and Abstract Concept Meaning. *Transactions of the Association for Computational Linguistics*, 2(1).

Joy Hsu, Jiayuan Mao, Joshua B. Tenenbaum, Noah Goodman, and Jiajun Wu. 2025. What Makes a Maze Look Like a Maze? In *The Thirteenth International Conference on Learning Representations*.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Marc A Kastner, Ichiro Ide, Frank Nack, Yasutomo Kawanishi, Takatsugu Hirayama, Daisuke Deguchi, and Hiroshi Murase. 2020. Estimating the Imageability of Words by Mining Visual Characteristics from Crawled Image Data. *Multimedia Tools and Applications*, 79.

Douwe Kiela and Léon Bottou. 2014. Learning Image Embeddings using Convolutional Neural Networks for Improved Multi-Modal Semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*.

Maximilian Köper and Sabine Schulte im Walde. 2017. Complex Verbs are Different: Exploring the Visual Modality in Multi-Modal Models to Predict Compositionality. In *Proceedings of the 13th Workshop on Multiword Expressions*.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. LLaVA-NeXT: Improved Reasoning, OCR, and World Knowledge. https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. *Advances in Neural Information Processing Systems*, 36.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to Explain: Multimodal Reasoning via Thought Chains for Science Question Answering. *Advances in Neural Information Processing Systems*, 35.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: Multidimensional Measures of Perceptual and Action Strength for 40,000 English Words. *Behavior Research Methods*, 52.

Ken McRae, Daniel Nedjadrasul, Raymond Pau, Bethany Pui-Hei Lo, and Lisa King. 2018. Abstract Concepts and Pictures of Real-World Situations Activate One Another. *Topics in Cognitive Science*, 10.

Sachit Menon and Carl Vondrick. 2023. Visual Classification via Description from Large Language Models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, May 1-5, 2023*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *arXiv preprint arXiv:1301.3781*.

George A Miller. 1995. WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11):39–41.

Daniela Naumann, Diego Frassinelli, and Sabine Schulte im Walde. 2018. Quantitative Semantic Variation in the Contexts of Concrete and Abstract Words. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*.

Allan Paivio. 1971. Imagery and Language. In *Imagery: Current Cognitive Approaches*. Academic Press.

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, Imagery, and Meaningfulness Values for 925 Nouns. *Journal of Experimental Psychology*, 76(1p2):1.

Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2018. Multimodal Explanations: Justifying Decisions and Pointing to the Evidence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. 2020. Visualcomet: Reasoning about the Dynamic Context of a Still Image. In *Computer Vision–ECCV 2020: 16th European Conference, August 23–28, 2020, Proceedings, Part V 16*.

Diane Pecher, Inge Boot, and Saskia Van Dantzig. 2011. Abstract Concepts. Sensory-Motor Grounding, Metaphors, and Beyond. *Psychology of Learning and Motivation – Advances in Research and Theory*, 54.

Sandro Pezzelle, Ece Takmaz, and Raquel Fernández. 2021. Word Representation Learning in Multimodal Pre-trained Transformers: An Intrinsic Evaluation. *Transactions of the Association for Computational Linguistics*, 9.

Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. 2014. Inferring the Why in Images. *arXiv preprint arXiv:1406.5472*, 2.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning Transferable Visual Models from Natural Language Supervision. In *International Conference on Machine Learning*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct Preference Optimization: your Language Model is Secretly a Reward Model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3).

Sabine Schulte im Walde and Diego Frassinelli. 2022. Distributional Measures of Abstraction. *Frontiers in Artificial Intelligence: Language and Computation 4:796756. Alessandro Lenci and Sebastian Pado (topic editors): "Perspectives for Natural Language Processing between AI, Linguistics and Cognitive Science"*.

Xiujie Song, Mengyue Wu, Kenny Q. Zhu, Chunhao Zhang, and Yanyi Chen. 2025. A Cognitive Evaluation Benchmark of Image Reasoning and Description for Large Vision-Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*.

Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. 2019. A Corpus for Reasoning about Natural Language Grounded in Photographs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Tarun Tater, Diego Frassinelli, and Sabine Schulte im Walde. 2022. Concreteness vs. Abstractness: A Selectional Preference Perspective. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research Workshop*.

Tarun Tater, Sabine Schulte Im Walde, and Diego Frassinelli. 2024a. Evaluating Semantic Relations in Predicting Textual Labels for Images of Abstract and Concrete Concepts. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*.

Tarun Tater, Sabine Schulte Im Walde, and Diego Frassinelli. 2024b. Unveiling the Mystery of Visual Attributes of Concrete and Abstract Concepts: Variability, Nearest Neighbors, and Challenging Categories. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Bart Thomee, Benjamin Elizalde, David Shamma, Karl Ni, Gerald Friedland, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The New Data in Multimedia Research. *Communications of the ACM*, 59.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution. *arXiv preprint arXiv:2409.12191*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Ruohong Zhang, Bowen Zhang, Yanghao Li, Haotian Zhang, Zhiqing Sun, Zhe Gan, Yinfei Yang, Ruoming Pang, and Yiming Yang. 2025. Improve Vision Language Model Chain-of-thought Reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

# 9 Appendix

The appendix offers additional methodological details, analyses, and examples that support the results and discussions presented in the main text. It includes dataset statistics (Section 9.1 and 9.2), similarity between attributed concepts (Section 9.3), selection criteria for VLMs (Section 9.4), extended analysis of image-concept representativeness (Section 9.5), VLM as a judge for preferences (Section 9.6), additional annotation examples (Section 9.7), prompt templates for VLMs (Section 9.8), implementation configurations (Section 9.9), and annotation details (Section 9.10) used in our experiments.

## 9.1 Abstractness Ratings of Attributed Concepts

This section presents a box plot in Figure 3 that compares the abstractness ratings of the concept labels attributed by humans, Llava, and Qwen across noun categories: abstract, mid-range, concrete, and overall. As discussed in Section 3.3, the figure highlights the higher standard deviation in VLM outputs, particularly for Llava, indicating greater variation in the abstractness ratings of the concepts they attribute.

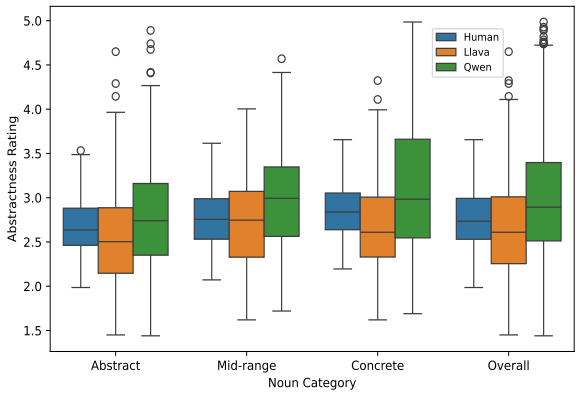



Figure 3: Box plot of abstractness ratings for concept annotations by humans, LLaVA, and Qwen, grouped by target noun category. Lower ratings indicate more abstract concepts.

## 9.2 Distribution of Explanation Lengths

This section presents a box plot in Figure 4 that compares the number of words in the explanations attributed by humans, Llava, and Qwen, across abstract, mid-range, concrete, and overall categories. While Section 3.3 discusses average lengths, the figure additionally shows that VLMs, especially Llava, exhibits high standard deviation in explanation length, indicating huge variability in the verbosity of its outputs. This difference in length is further analyzed in relation to human preferences in Section 4.1.

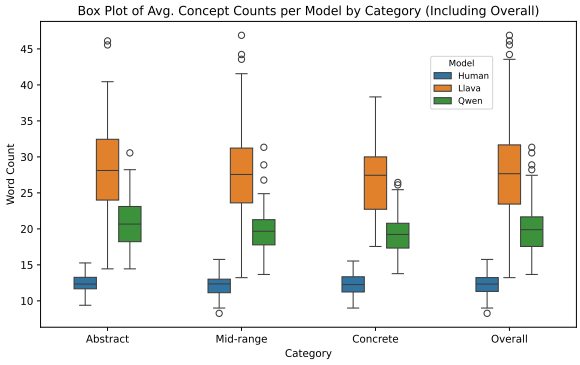



Figure 4: Distribution of explanation lengths (in words) across abstractness categories and annotator groups.

## 9.3 Do annotators assign semantically similar concepts?

To assess the consistency of assigned concepts for the same image, we compute pairwise cosine similarity between word embeddings of concepts within each annotator group using word2Vec[8] (Mikolov et al., 2013) We evaluate static embeddings since concepts are analyzed independently of context. As shown in Table 7, human-generated concepts exhibit slightly lower internal similarity (overall: 0.15), than those from VLMs.

| | Similarity within concepts | | | Similarity with target noun | | |
|---|---|---|---|---|---|---|
| Category | Human | Llava | Qwen | Human | Llava | Qwen |
| Abstract | 0.19 | 0.23 | 0.24 | 0.17 | 0.20 | 0.20 |
| Mid-range | 0.12 | 0.23 | 0.22 | 0.14 | 0.17 | 0.19 |
| Concrete | 0.14 | 0.04 | 0.21 | 0.13 | 0.14 | 0.17 |
| Overall | 0.15 | 0.17 | 0.22 | 0.15 | 0.17 | 0.19 |

Table 7: Average concept word2vec similarity and average word2vec similarity of concepts with target noun across annotator groups. (Higher values indicate greater alignment.)

To further assess whether assigned concepts remain semantically close to the meaning of the target noun associated with the image, we compute pairwise cosine similarity between word2Vec embedding of each concept and the target noun. Here,

[8]Trained on Google News dataset with 300-dimensional vectors.

humans again show slightly lower similarity to the noun compared to VLMs.

### 9.4 VLM Selection Criteria

We selected Qwen2-VL-7B-Instruct and LLaVA-Next (llava-v1.6-mistral-7b-hf) as representative open-source vision-language models for our tasks. These models were chosen for their small size, recent release, wide usage, and strong performance relative to models of similar scale, based on benchmarks relevant to attribution and reasoning tasks.

**Performance.** LLaVA-Next-1.6-7B performs comparably to its 13B variant and outperforms GPT-4V on VQAv2 (Liu et al., 2024). Qwen2-VL-7B also performs better than contemporary models like InternVL2-8B on multiple benchmarks (MMMU val, Blink val, MMT val, RealWorldQA), and is competitive with InternVL2-40B, and even InternVL2-76B on tasks such as Text-VQA and DocVQA (Chen et al., 2024b,c). We do not include closed-source models like GPT-4V, whose training data and design are not publicly available, limiting the interpretability of their attributions.

**Compute.** Scaling to larger 70B+ models remains computationally infeasible, requiring $150 - 200$ GB of GPU memory even with quantized formats (8/16-bit).

Our results support the adequacy of these models for our task: they performed well on abstract attributions, were often preferred by human raters, and served as reasonable judges of human preferences (correlation scores $0.7 - 0.8$).

### 9.5 VLM Judgments of Image - Concept Representativeness

This section extends the representativeness analysis discussed in Section 3.4 by evaluating how VLMs judge the alignment between attributed concepts and corresponding images for all annotations of 675 images. Specifically, Qwen and Llava were prompted to rate whether an image is *strongly*, *moderately*, *weakly*, or *not at all* representative of an attributed concept. Table 8 summarizes their judgments across human-, Llava-, and Qwen-attributed concepts. The corresponding prompt is presented in Section 9.8.

The results show notable differences in how humans and VLMs assess representativeness. Qwen assigns *Strongly Representative* ratings very frequently across all annotations. It classifies $\sim 81 - 84$ % of VLM concepts as *Strongly Representative*, compared to (62.32%) of human concepts. Notably, Qwen assigns almost no *Not Representative* concepts, especially for VLM concepts ($0.05$ and $0.10\%$), suggesting a strong self-reinforcement bias. In contrast, Llava acts as a stricter judge. However, both Qwen and Llava struggle to strongly associate human concepts to images, reinforcing the need for human oversight in subjective annotation tasks.

### 9.6 VLM Judgments of Concept and Explanation Preferences

This section presents additional results on how VLMs (Llava and Qwen) rank concepts and explanations when acting as judges in the preference evaluation task. Tables 9 and 10 report the number of times VLMs preferred concepts and explanations from different sources (human-exclusive, VLM-exclusive, and overlapping), and how these preferences changed when explanations were added.

When ranking top concept and explanation preferences (Table 9), both Qwen and Llava tend to favor overlapping concepts, similar to human preferences. When considering all ranked comparisons (not limited to top-1), VLM-generated explanations were preferred over human explanations in 79 and 78 respectively by Llava and Qwen, out of 98 overlapping concept cases across 63 images. Table 10 shows the percentage of significant rank changes between concept-only and concept-explanation preferences and is similar to that of human annotators. The Spearman correlation between explanation length and preference rank (where lower rank indicates stronger preference), is $\rho = -0.43$ ($p < .0001$) for both Llava and Qwen. Comparing abstractness ratings of concepts with preference ranks, we found no significant preference at the concept level for Llava ($\rho = 0.03$, $p = .44$) or Qwen ($\rho = -0.01$, $p = .83$). Similarly, there was no significant correlation at explanation level.

### 9.7 Example Annotations for Contrasting Annotations

This section presents qualitative examples that illustrate the subjectivity of concept attributions and how explanations can help clarify why a particular concept was assigned to an image. Figure 5 shows an example where different annotators interpret the same image in contrasting ways, evoking distinct abstract concepts based on personal perspectives.

| Rating | Llava as judge | | | Qwen as judge | | |
|---|---|---|---|---|---|---|
| | Human concepts | Llava concepts | Qwen concepts | Human concepts | Llava concepts | Qwen concepts |
| Strongly | 39.76 | 59.16 | 53.60 | 62.32 | 80.89 | 83.56 |
| Moderately | 32.27 | 33.38 | 33.38 | 30.81 | 18.42 | 14.86 |
| Weakly | 23.16 | 5.87 | 10.77 | 5.65 | 0.64 | 1.48 |
| Not at all | 4.82 | 1.58 | 2.24 | 1.21 | 0.05 | 0.10 |

Table 8: Percentage of images rated as strongly, moderately, weakly, or not at all representative of an attributed human, Llava, or Qwen concept by Llava and Qwen as judges.

| Category | Llava as judge | | | | | | Qwen as judge | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Concept | | | Explanation | | | Concept | | | Explanation | | |
| | Overlap | VLM | Human | Overlap | VLM | Human | Overlap | VLM | Human | Overlap | VLM | Human |
| Abstract | 5 | **9** | 7 | **10** | 9 | 2 | 8 | **10** | 3 | **10** | 7 | 4 |
| Mid-range | **12** | 7 | 2 | 9 | **12** | −− | **11** | 7 | 3 | **12** | 7 | 2 |
| Concrete | **15** | 3 | 3 | **11** | 8 | 2 | **13** | 8 | −− | **16** | 5 | −− |
| Overall | **32** | 19 | 12 | **30** | 29 | 4 | **32** | 25 | 6 | **38** | 19 | 6 |

Table 9: Top-1 preference counts for concepts and explanations as judged by LLaVA and Qwen. Each cell shows how often the most preferred attribution (concept or explanation) came from overlapping sources (used by both human and model), VLM-exclusive sources, or human-exclusive sources. Overall, both VLMs show a consistent preference for overlapping attributions.

| Origin | **Llava** (%) | | **Qwen** (%) | |
|---|---|---|---|---|
| | Rank ↑ | Rank ↓ | Rank ↑ | Rank ↓ |
| Human | 24.60 | 7.67 | 29.37 | 7.14 |
| Overlap | 44.44 | 17.09 | 38.46 | 14.52 |
| VLM | 17.09 | 17.94 | 19.66 | 13.68 |

Table 10: Change in concept preference rankings with added explanations, as judged by Llava and Qwen. Values show the percentage of cases where concept ranks increased (↑) or decreased (↓) when explanations were included.


**Jealousy:** I want to go to universal studious again and to go to this location.

**Happy**: brings back happy feelings from when I first watched and read harry potter

Figure 5: Example of an image annotations with contrasting concepts.

Figure 6 highlights an example where the attributed concept may not seem very representative of the image in isolation, but the accompanying explanation provides a rationale that may make the concept more relatable and potentially more preferred. These examples align with observations in Section 4.1, emphasizing the role of explanation and interpretation in abstract concept attributions.

## 9.8 VLM Prompts

This section details the prompts used for Llava and Qwen across different setups of this work. To ensure consistency across human and model annotations, the prompts were identical to those shown to human annotators, with the only difference being minor formatting instructions for the models (e.g., returning outputs in a dictionary format).

### 9.8.1 Concept - Explanation Annotations

For the concept-explanation annotation task, both Llava and Qwen were prompted to generate abstract concepts and the corresponding explanations for each image using the following prompt:

**Farewell:** the city skyline fading into the background reminds me of a saying goodbye to a city to move to another one.
**Familiar:** I go to the beach often in the summer and can see the skylines of the city across from it

Figure 6: Example of an image's annotations where concepts are strongly derived from personal experience.

Model Prompt

Identify 3 concepts conveyed by the image. Each concept should be a single word. Avoid naming specific people or objects; instead, focus on summarizing the situation or the ideas, emotions, or feelings conveyed by the image. For each concept, provide a brief explanation of how it is conveyed by the image. Return the concepts and their explanations in a python dictionary format: {concept1: explanation1, concept2: explanation2, concept3: explanation3}.

#### 9.8.2 Image - Concept Representativeness

For the image-concept representativeness task, both Llava and Qwen were prompted to rate how well an image represents a given concept, using the same four-point scale as in the human evaluation (Section 3.4). The prompt was:

Model Prompt

Your task is to evaluate if this image is a good representation of "concept"?. Focus on the overall situation, ideas, emotions, and feelings conveyed by the image instead of specific people or objects.
Review each image and concept carefully and choose the appropriate option:
- 1 - Not at all representative: The image is not associated with the concept. There are little to no visual or contextual cues connecting it to the concept.
- 2 - Weakly representative: The image has vague or minimal association with the concept. Significant inference or imagination is required to establish any connection.
- 3 - Moderately representative: The image moderately represents the concept. There are some visual or contextual cues, but the representation could be even stronger.
- 4 - Strongly representative: The image clearly and meaningfully evokes the concept. It strongly conveys the essence of the concept with little room for doubt
Return your answer in a python dictionary format exactly as: "Rating":

#### 9.8.3 Concept Preferences

For the concept preference task, the prompt presented an image along with four candidate concepts and asked the model to select the most and least representative ones. The prompt was:

Model Prompt

Your task is to evaluate which of the given concepts best represents the core idea conveyed by the image and which concept is the least accurate. Focus on the overall situation, ideas, emotions, and feelings conveyed by the image rather than specific people or objects.
1. {concept1}
2. {concept2}
3. {concept3}
4. {concept4}
Instructions:
Identify the Most accurate: The concept that best captures the core idea of the image. Identify the Least accurate: The concept that has the weakest connection to the image.
Return your answer in a python dictionary format exactly as:
{{ "Most": , "Least": }}

#### 9.8.4 Concept-Explanation Preferences

For the explanation preference task, the prompt presented an image with four concept-explanation pairs, and the model was asked to choose the best and worst descriptions based on how well they described the image. The prompt was:

> Model Prompt
>
> Your task is to evaluate which of the given descriptions best describes the image and which one provides the worst description. Focus on the overall situation, ideas, emotions, and feelings conveyed by the image rather than specific people or objects.
> 1. {explanation1}
> 2. {explanation2}
> 3. {explanation3}
> 4. {explanation4}
>
> Instructions:
> Identify the Best Description: The description that most accurately represents the image. Identify the Worst Description: The description that is the least accurate.
> Note: Do not focus on minor grammar issues, instead focus on the content of the description.
> Return your answer in a python dictionary format exactly as: "Best": 1..4 , "Worst": 1..4

#### 9.8.5 DPO Finetuning

For DPO fine-tuning, the base prompt presented to the Llava and Qwen model was:

> Model Prompt
>
> Which of the given descriptions best describes the image? Focus on the overall situation, ideas, emotions, and feelings conveyed by the image rather than specific people or objects.

This prompt was paired with two candidate responses - one marked as *preferred* and the other as *rejected*, based on human rankings from the explanation preference task. These {*prompt, preferred, rejected*} triplets formed the training data for computing the DPO loss. The image input and both responses were encoded as part of the conversational context during fine-tuning.

### 9.9 Compute and Experimental Details

**Compute and Inference Setup** All VLM-based inferences were performed using an NVIDIA RTX A6000 GPU. For both Qwen and Llava, each prompt took approximately $3-5$ seconds to process. We used the default publicly available model checkpoints for both models from huggingface[9], without any additional modifications for inference.

**DPO Fine-tuning Configuration** For DPO fine-tuning, each training instance consists of a prompt $x$ (question + image) and a *chosen (preferred)* ($y_w$) attribution and a *rejected (less preferred)* ($y_l$) attribution. For such a preference dataset $D$, we fine-tune our VLM $\pi_\theta$ with the loss function $\mathcal{L}_{\text{DPO}}$

$$\mathcal{L}_{\text{DPO}}(\pi_\theta \,;\, \pi_{\text{ref}}) = - \mathbb{E}_{(x,y_w,y_l)\sim D}\Big[\log \sigma\Big(\beta \log \frac{\pi_\theta(y_w \mid x)}{\pi_{\text{ref}}(y_w \mid x)} - \beta \log \frac{\pi_\theta(y_l \mid x)}{\pi_{\text{ref}}(y_l \mid x)}\Big)\Big]. \quad (1)$$

where $\pi_{\text{ref}}$ is the reference model (i.e., the base model), and $\beta$ is a scaling parameter, and the optimization is KL-constrained to prevent excessive divergence from the original model. For finetuning llava, we applied LoRA adapters to the following projection layers: *q_proj*, *k_proj*, *v_proj* of the last 4 layers of the language model, along with *multi_modal_projector.linear_1*, *multi_modal_projector.linear_2* layers. For finetuning qwen, we applied LoRA adapters to the following projection layers: *q_proj*, *k_proj*, *v_proj* of the last 4 layers. We initialized training with the default DPO settings and conducted a hyperparameter sweep. Specifically, we varied the learning rate from $5\text{e}{-6}$ to $5\text{e}{-5}$, the DPO scaling parameter $\beta$ from $0.01$ to $0.1$ in increments of $0.02$, and the weight decay in $\{0.01, 0.02\}$. We also experimented with LoRA rank values of $2$, $4$, and $8$. All experiments used a fixed random seed of $42$.

### 9.10 Annotation Details

This section describes the setup for human annotations used throughout the study, including concept-explanation attribution, image-concept representativeness study, and preference judgments. We report details on participant recruitment, compensation, filtering criteria, noun selections, and licensing of the images used.

[9] https://huggingface.co/

**Compensation and Annotation Cost.** Each participant was paid at a rate of £9 per hour. The total cost of the annotations was $\sim$ £3000.

**Consent and Personal Information.** Participation was voluntary and participants were informed of the time, compensation, and purpose of the study before they could continue. We do not collect any information that can associate the participants with the data back to the participants.

**Participant Filtering.** Due to the nature of our study, we only allowed participants who had listed English as their first language and were based in Australia, Canada, Ireland, New Zealand, UK, and US. Moreover, their approval rating was between 99-100 on Prolific and had made $\geq$ 20 submissions.

**Participant Demographics.** We conducted a post hoc analysis of the demographics of the participants for the concept-explanation annotations using metadata provided by Prolific. The participants' age ranged from 18 to 77 (mean = 36.1, SD = 14.3), with balanced gender distribution (51% male, 49% female). The majority reported White ethnicity (59%), followed by Black (21%), Asian (12%), and others. Participants were primarily based in the United Kingdom (53%), United States (29%), and Canada (13%). English was the primary language for over 93% of participants.

While demographic data was not part of our core analysis, we include this summary for completeness. Our study is not designed to systematically investigate demographic variation, but we recognize that individual background may play a role in subjective interpretations. We view this as a complementary direction for future work, particularly in studies focused on cross-cultural or population-level comparisons.

**Image Licenses.** We used only images released under permissive Creative Commons licenses, specifically: *Attribution*, *Attribution-NonCommercial*, *Attribution-ShareAlike*, and *Attribution-NonCommercial-ShareAlike*.

#### 9.10.1 Selected Nouns for Preference Judgments

We present the 21 nouns for which we collected preference judgments in Table 11. We also present the complete set of 225 nouns for which we collected concept-explanation attributions in Table 12.

| Category | Nouns (Abstractness rating) |
|---|---|
| Abstract | Forgiveness (1.44), Enlightenment (1.5), Courage (1.52), Tradition (1.69), Contemplation (1.83), Mischief (1.9), Patriotism (2.0) |
| Mid-range | Deployment (2.9), Bedtime (2.9), Curvature (2.93), Witchcraft (2.96), Adoption (3.03), Malnutrition (3.1), Royalty (3.11) |
| Concrete | Penguin (5.0), Lamp (4.97), Dolphin (4.96), Wasp (4.96), Wreath (4.93), Guitar (4.9), River (4.89), |

Table 11: Subset of 21 Nouns Categorized by Abstractness Ratings

#### 9.10.2 Annotation Forms

We present example instances of the annotation forms used in our study. Figure 7 shows an instance of the form used to collect concept-explanation annotations. Figure 8 shows the form used to rate image-concept representativeness. Figure 9 presents the form for concept preference judgments, and Figure 10 shows the form used to collect explanation preference judgments. Each figure displays the actual format and content as seen by annotators. We ensured that the annotation phases followed a consistent setup. Annotators were explicitly instructed to "focus on the overall situation, ideas, emotions, and feelings conveyed by the image rather than specific people or objects". This alignment in task context was intentional to ensure comparability between how concepts and explanations were initially generated and how they were later evaluated.

| Category | Nouns |
|---|---|
| **Abstract** | priority, spirituality, gloom, humility, holiness, oblivion, ignorance, opportunity, patriotism, contemplation, serenity, injustice, censorship, fraud, intimacy, capitalism, heaven, disappointment, luck, awareness, intricacy, generosity, belief, encouragement, curiosity, strategy, desperation, endurance, courage, fantasy, splendor, nostalgia, exception, bureaucracy, aftermath, anticipation, patience, bravery, accuracy, creativity, repression, imagination, bliss, duality, probability, purity, identity, innocence, economics, democracy, compassion, mysticism, tradition, sarcasm, perseverance, qualification, hardship, irony, downtime, fun, comparison, authentication, forgiveness, admiration, eternity, privacy, adversity, enlightenment, sadness, explanation, legacy, amazement, artistry, kindness, mischief |
| **Mid-range** | hobby, unemployment, oath, captivity, accident, deforestation, census, math, astronomy, procession, physics, deployment, adoption, countdown, crime, quotation, announcement, proposal, addiction, rehearsal, bedtime, altitude, montage, flexibility, statistic, symbol, artisan, piracy, discussion, transaction, royalty, simulation, nutrition, refusal, nightmare, summertime, excursion, evacuation, recreation, oxidation, curvature, cleanup, geometry, fairytale, visualization, entertainment, communication, catastrophe, donation, symmetry, motherhood, surveillance, finale, legend, science, witchcraft, geography, winner, autumn, meditation, malnutrition, recycling, departure, promotion, district, zodiac, stamina, tourism, adrenaline, civilization, measurement, nightlife, commuter, fitness, ancestry |
| **Concrete** | kayak, escalator, shelf, waterfall, beach, aquarium, penguin, helmet, sushi, elephant, cigar, bonfire, bulldozer, footbridge, pollen, toothbrush, thunderstorm, airplane, mango, postcard, motorcycle, policeman, classroom, mountain, wasp, lamp, tomato, stroller, grenade, island, bride, firewood, suitcase, puppet, alley, spacecraft, laundry, arrow, swimmer, salad, microscope, river, bamboo, toad, medal, windmill, cyclist, meteor, sprinkler, ocean, chick, yacht, fridge, chess, apartment, crayon, warship, scaffolding, meadow, playground, container, kitchen, hammock, statue, billboard, firefighter, wheelchair, orchestra, owl, motorway, donkey, dolphin, guitar, wreath, iceberg |

Table 12: List of 75 nouns per category used to build our dataset totaling 225 nouns.

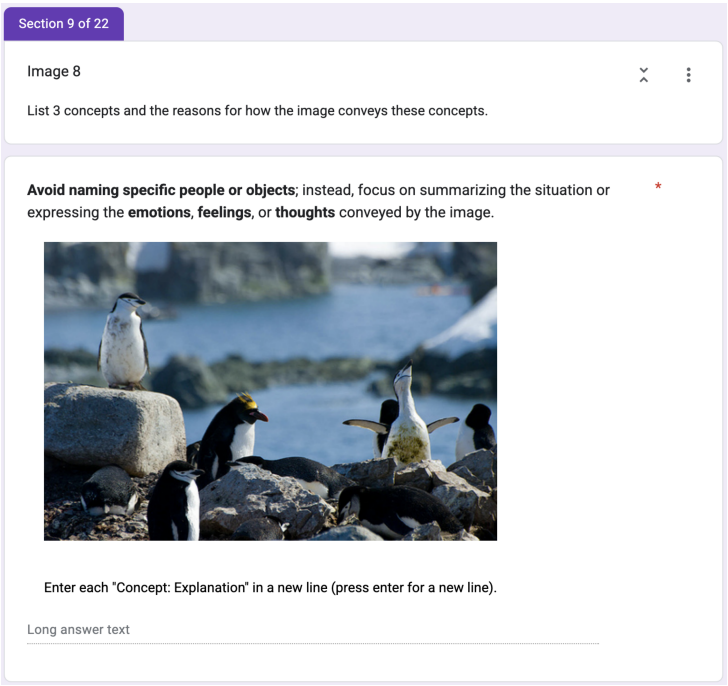



Figure 7: Example annotations question for collecting abstract concept and explanations.

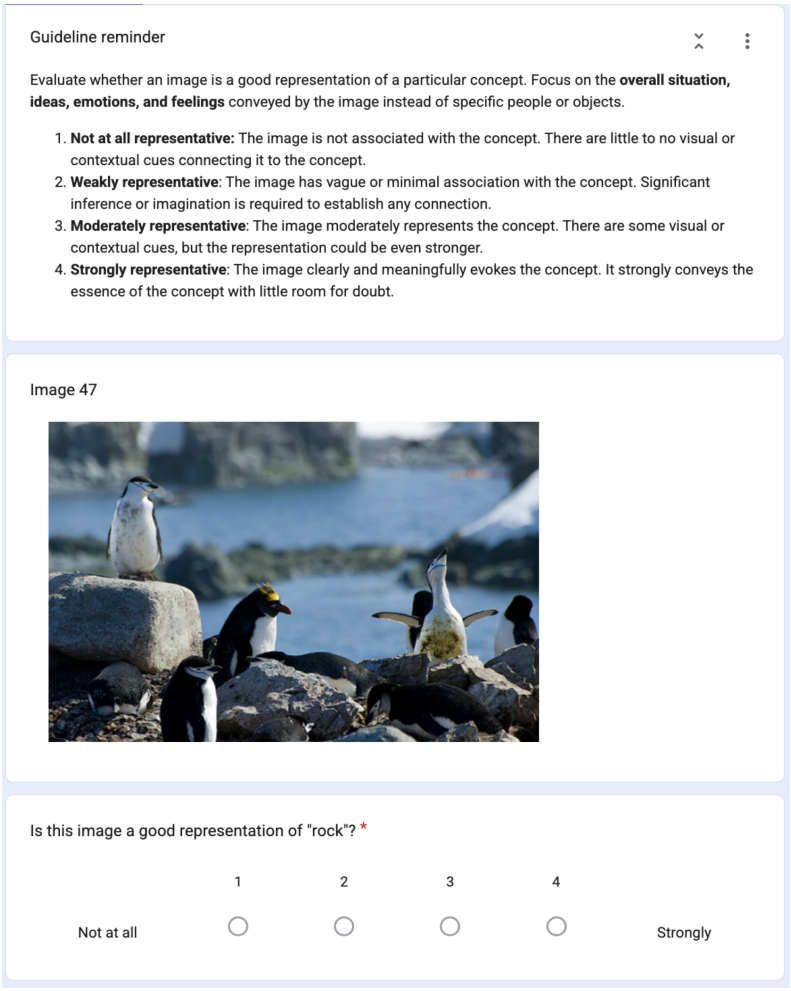



Figure 8: Example of annotation instance for evaluating image representativeness of a concept.

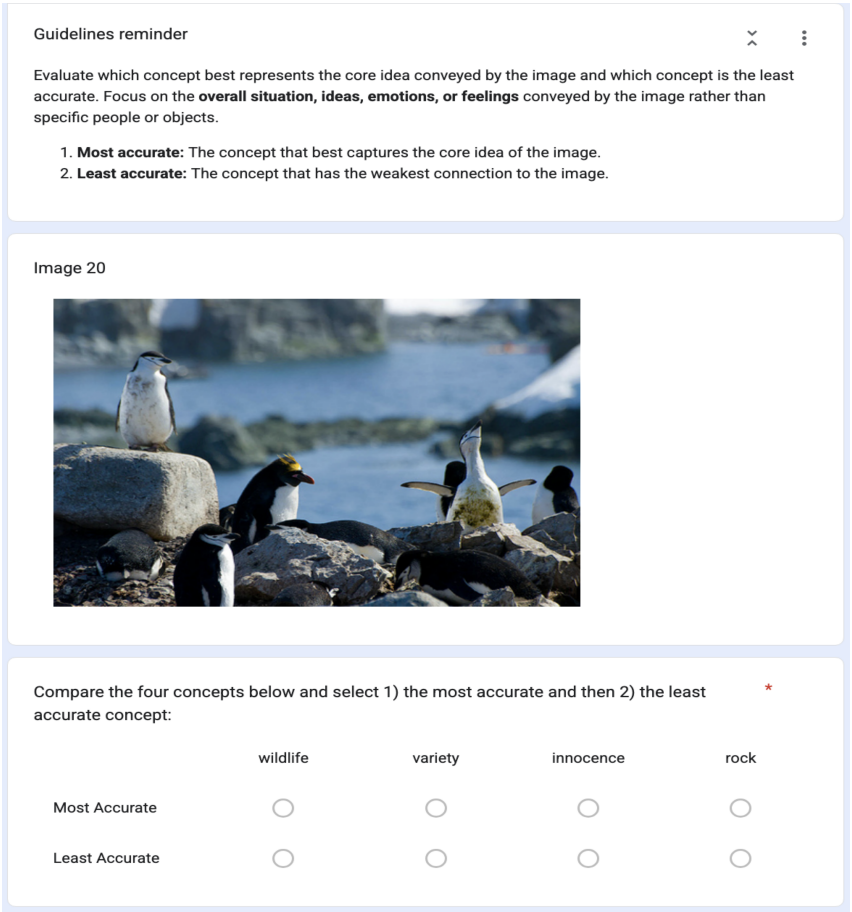



Figure 9: Example of annotation instance for concept preferences.

**Guidelines reminder**

Evaluate which description **best** describes the image and which one provides the **worst** description. Focus on the **overall situation, ideas, emotions, or feelings** conveyed by the image rather than specific people or objects.

Please do **NOT** focus on minor grammar issues, instead **focus on the content of the description**.

1. **Best description:** The description that most accurately represents the image.
2. **Worst description:** The description that is the least accurate.

Image 22

Compare the four descriptions below: *

| | The image conveys Freedom because: on of the penguins in the picture has its wings spread which could express the feeling of being free | The image conveys Variety because: the picture shows different types of a species of animal | The image conveys Color because: a black and white pigeons on lake . | The image conveys Innocence because: the animals convey the feeling of innocence as they sit together seemingly away from harm. |
|---|---|---|---|---|
| best description | ○ | ○ | ○ | ○ |
| worst description | ○ | ○ | ○ | ○ |

Figure 10: Example of annotation instance for explanation preferences.