

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

Tarun Tater¹, Diego Frassinelli², Sabine Schulte im Walde¹
¹University of Stuttgart, ²LMU Munich



Abstract vs Concrete Concepts

- Concrete concepts can be strongly experienced through human senses i.e., things that can be *seen, heard, felt, smelled, or tasted* as opposed to abstract concepts.



AbsVis Dataset

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

	Lamp (concrete)	Curvature (mid-range)	Contemplation (abstract)
Human-exclusive attributions	<p>H1 illuminate: the light shines at night to brighten the street</p> <p>H2 spring: color of leaves on the trees</p>	<p>cold: bare trees, frozen lake, probably winter.</p> <p>empty: the environment is empty and there's no people.</p>	<p>admiration: seems to be admiring the artwork</p> <p>contemplative: sat taking in the surroundings</p>
VLM-exclusive attributions	<p>V1 autumn: the image features vibrant orange and red leaves, which are characteristic of the autumn season.</p> <p>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.</p>	<p>tranquility: the stillness of the water and the calmness of the scene convey a sense of peace.</p> <p>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.</p>	<p>solitude: the person is alone in a spacious room, which emphasizes the feeling of solitude and introspection.</p> <p>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.</p>
Overlapping attributions	<p>O1 tranquility (H) the stillness of the scene feels calm and peaceful.</p> <p>O2 (V) the solitary street lamp against a blurred background of autumn leaves conveys a sense of calm and peacefulness.</p>	<p>winter: (H) along both sides of a curving walkway with turquoise railings - the bare branches of the trees in the distance also suggest winter. (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.</p>	<p>reflection: (H) a space to see beautiful new images and muse upon them in relation to ones own life. (V) the image conveys a sense of reflection, as the person is sitting alone and looking at a painting, possibly contemplating the scene depicted.</p>

- 675 images (3 images × 225 nouns: 75 abstract, 75 mid-range, 75 concrete).
- Human Annotations: 5 per image × 3 pairs.
- VLM Annotations: Qwen-VL (7B) & LLaVA-Next (7B), 3 pairs per image.
- Total ≈ 10k human + 4k VLM concept-explanation pairs.

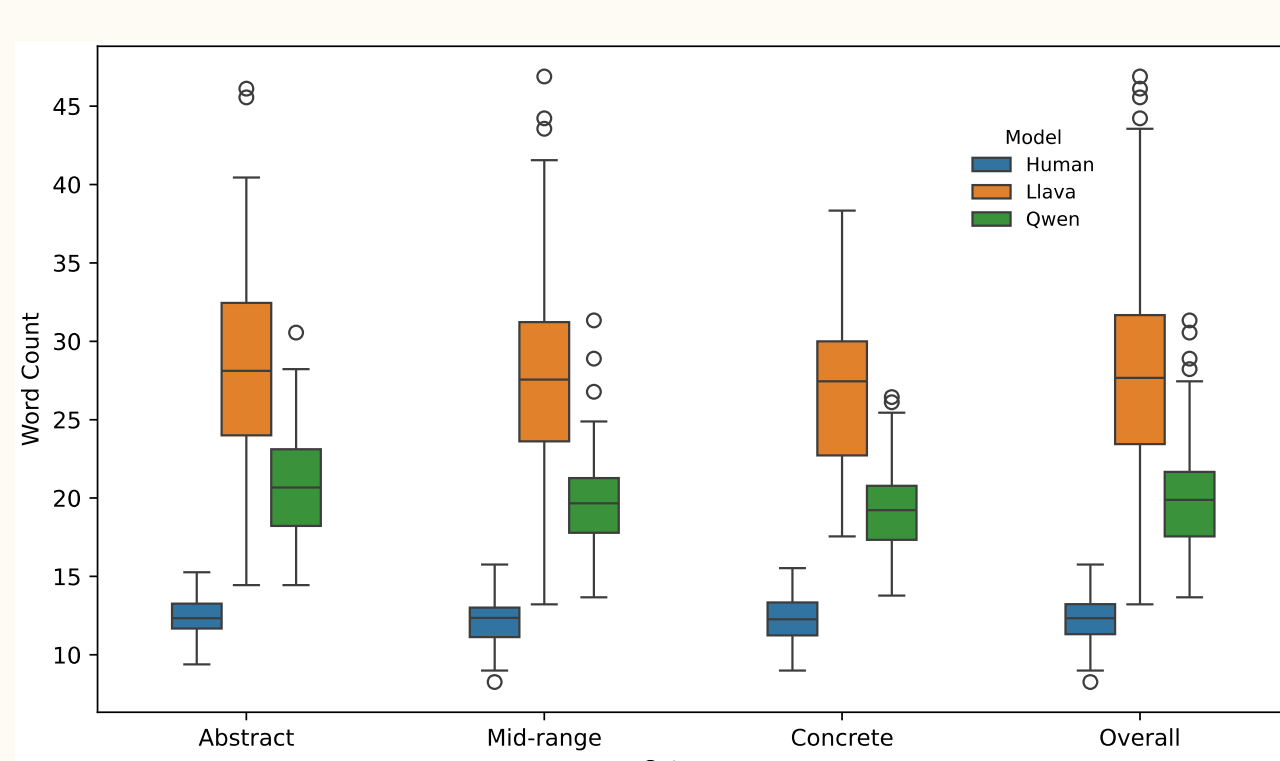
Humans vs VLMs: Properties of Attributions

Does everyone see the same concept?

- Humans: highly diverse (≈ 10 % overlap).
- VLMs: more internally consistent (≈ 20 – 25 % overlap).
- Word2vec similarity of concepts within each group is low between [0.1, 0.2].

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

Percentage of overlapping concepts across annotator groups.



Average word counts for human and VLM explanations by noun category.

How do explanations differ?

- Human explanations: shorter (≈ 12 words)
- VLM explanations: longer (≈ 20 – 28 words).

Which attributions do humans and VLMs prefer?

- Human preferences collected using **Best-Worst scaling**.
- 63 images - 3 images × 21 nouns.
- Total ≈ 2.7k human preferences.

Which concepts do humans prefer?

- Overlapping concepts (human ∩ VLM) are **most preferred**.
- VLM-exclusive > human-exclusive.

Category	Concept		
	Overlap	VLM	Human
Abstract	5	10	6
Mid-range	11	6	4
Concrete	10	8	3
Overall	26	24	13

Human preferences for **Top-1** concept out of 63 images.

Category	Explanation		
	Overlap	VLM	Human
Abstract	11	7	3
Mid-range	12	6	3
Concrete	16	5	—
Overall	39	18	6

Human preferences for **Top-1** explanation out of 63 images.

Which explanation pairs do humans prefer?

- Explanations with overlapping concepts highly preferred.
- VLM-exclusive > human-exclusive.

How can we scale collection of preferences?

Can VLMs approximate human preferences?

- Prompted Qwen and LLaVA to rank concept-explanation pairs.
- VLMs can approximate preferences ($\rho \approx 0.7 - 0.8$) - promising for scalable evaluation.

	Concept-level	Explanation-level
LLaVA-Human	0.71	0.75
Qwen-Human	0.78	0.79

Spearman's correlation (ρ) between human and VLM preferences.

Can DPO improve VLM alignment with humans for abstract concept attributions?

- Used Qwen preferences to supervise LLaVA and vice versa, avoiding self-bias and simulating external human judgment.
- Direct Preference Optimization (DPO) to fine-tune Qwen and LLaVA.
- Training pairs → a *preferred* 👍 and *less preferred* 👎 concept-explanation.
- Considerably improves the alignment of VLMs with human attributions.

Takeaways

Abstract concept attributions are **diverse** and **subjective** – no single ground truth.

Overlapping attributions are the most interpretable and **preferred**.

VLMs moderately approximate human preferences.

Using VLM preferences as proxies with DPO offers a path toward **human-aligned multimodal understanding**.