Beyond Words: The Multimodal Nature of Concrete and Abstract Concepts

Diego Frassinelli

frassinelli@cis.lmu.de
MaiNLP @ LMU Munich

Göteborg, 12 February 2025

Abstractness

- A central construct in cognitive science (Barsalou, 2003)
- . . . but there is terminological ambiguity
- Barsalou provides six definitions of abstractness
- We focus on one: Abstract vs. Concrete (Brysbaert et al., 2014)
  - Most popular definition in CogSci and NLP
  - Focus on perception
  - $happiness_{2.6} \rightarrow belief_{1.2} \rightarrow banana_{5.0}$

Abstractness: Concrete vs. Abstract

Concrete Concepts

- Clear perceivable referent through the five senses
- Two clear categories: natural objects (e.g., $banana_{5.0}$) and artefacts (e.g., $chair_{4.6}$)
- Very specific defining properties for the entity:
  - Characterising perceptual features: taste, colour, smell, touch, and sound
  - Clear interactions between the entity and other entities
  - Clear situations where the entity can occur
- Concreteness effect: concrete concepts have a cognitive advantage over abstract concepts

Abstractness: Concrete vs. Abstract

Abstract Concepts

- More tricky to precisely define
- Not clear perceivable referents in the real world
- No clear categorization: $number_{3.0}$ vs. $joy_{2.4}$
- More difficult to retrieve properties, interactions, and situations
- **Still a lot to investigate!**

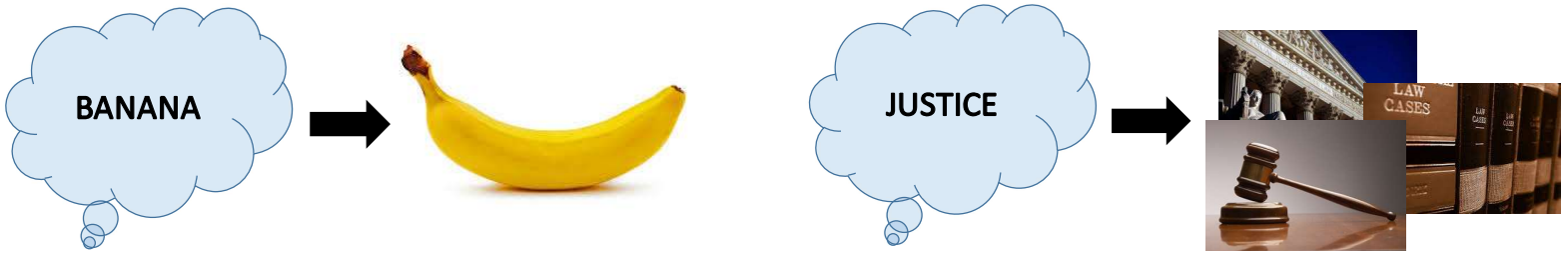
Terminology

- **Concepts:** mental representations
- **Words:** the corresponding linguistic surface forms

Cognitive Evidence: Abstract vs. Concrete Concepts

Grounding Theory:
(Barsalou, 1999)

- concrete concepts are directly grounded in the sensory-motor system
- abstract concepts are **mapped to concrete** concepts in order to be processed



How to Quantify Abstractness?

Concreteness Norms: Brysbaert et al. (2014)

Concreteness Norms: A large scale collection of human judgements via crowd-sourcing

Task

"We want you to indicate how concrete the meaning of each word is for you by using a 5-point rating scale going from abstract to concrete"

Abstract (*language based*) *Concrete* (*experience based*)

1 *2* *3* *4* *5*

N = I do not know this word well enough to give a rating.

---

belief$_{1.2}$ banana $_{5.0}$

- 37,058 English single words
- 4,000 online participants (at least 25 observations per word)

The Brysbaert Concreteness Norms

Brysbaert et al. (2014)

- Extremely important empirical measure in many fields
  - Cognitive science: to better explain specific phenomena
  - Computational linguistics: to better predict specific linguistic phenomena
- Reference collection to automatically quantify abstractness in many studies
- But... it is not perfect!

The Limits of Abstractness Norms

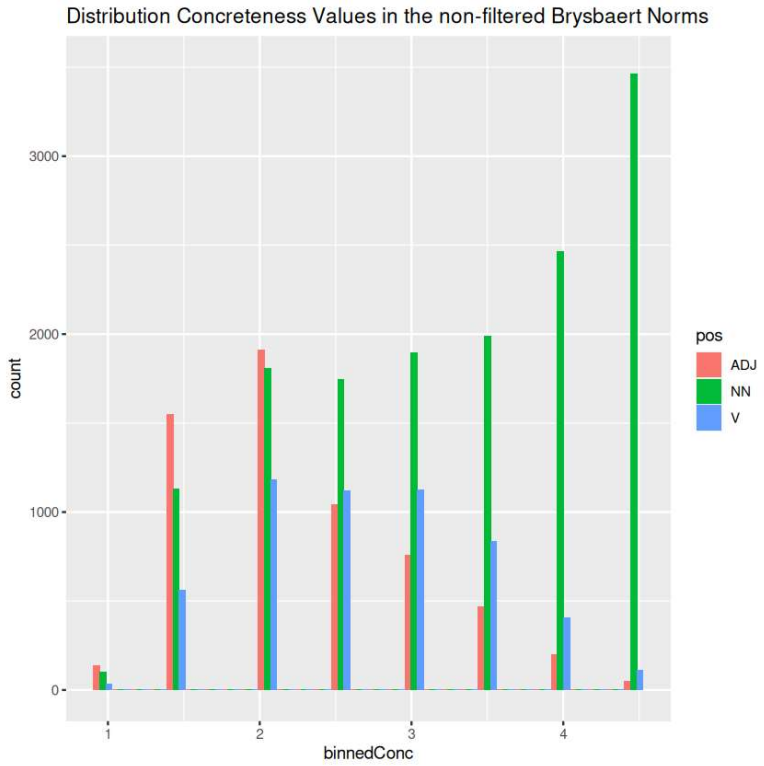
Bimodal Distributions

Really?

- Ratings of [concreteness and abstractness] were bimodally distributed, in agreement with other studies. *Nelson & Schreiber (1992), Wiemer Hastings & Xu (2005), Della Rosa et al. (2010)*

- [...] the distribution of concreteness ratings is bimodal, with separate peaks for concrete and abstract words [...] *Brysbaert et al (2014)*

Corpus Distributions across Word Classes

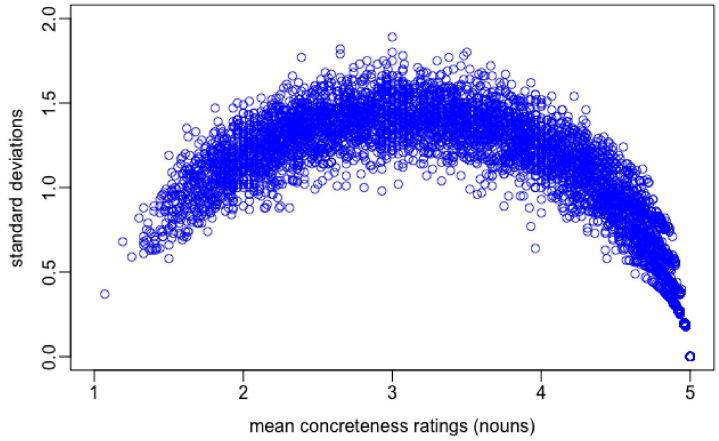


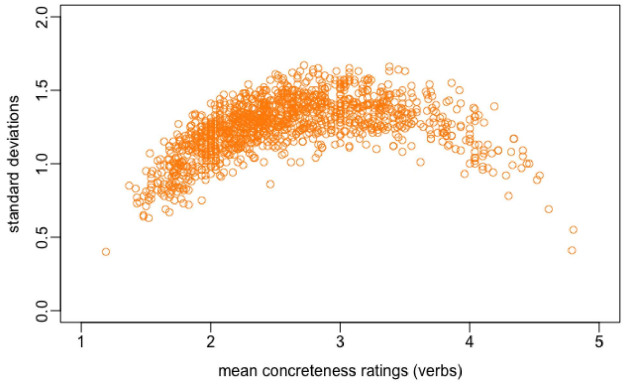
- x-axis: mean concreteness ratings; y-axis: number of targets
- What does this plot tell us?
  - **different distributions** across word classes
  - **nouns** are on average more concrete
  - **verbs** and **adjectives** are on average more abstract
  - no bimodal distribution!

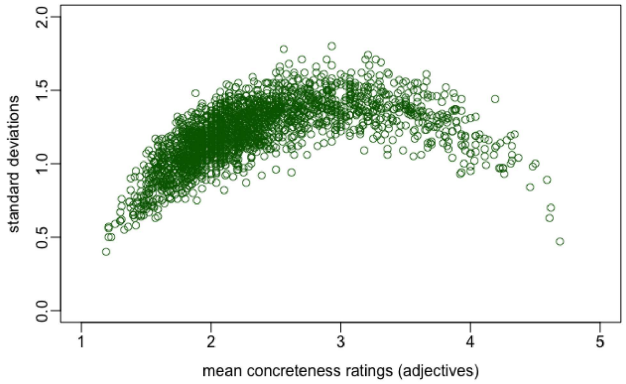
Ratings: Means & Standard Deviations

Croissant Plots

- Usually, we work with average ratings (e.g., out of 25 participants)
- Do humans agree on a specific score?



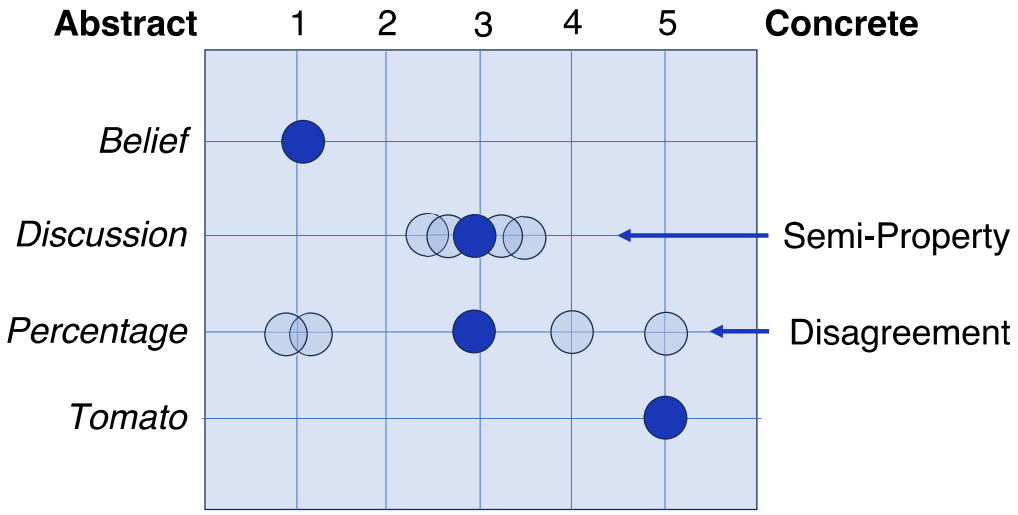




The Characteristics of Mid-Scale Ratings

Knupleš et al. (2023)

- What are the characteristics of concepts with mid-scale ratings?



- Same mean values but very different distributions

Materials

- Concreteness ratings for 1500 English Nouns (Brysbaert Norms)
  - 500 extreme abstract (1.07 - 1.71)
  - 500 mid-scale (2.90 - 3.31)
  - 500 extreme concrete (4.85 - 5.00)
- Two studies:
  1. Classification → average scores
  2. Clustering → individual ratings

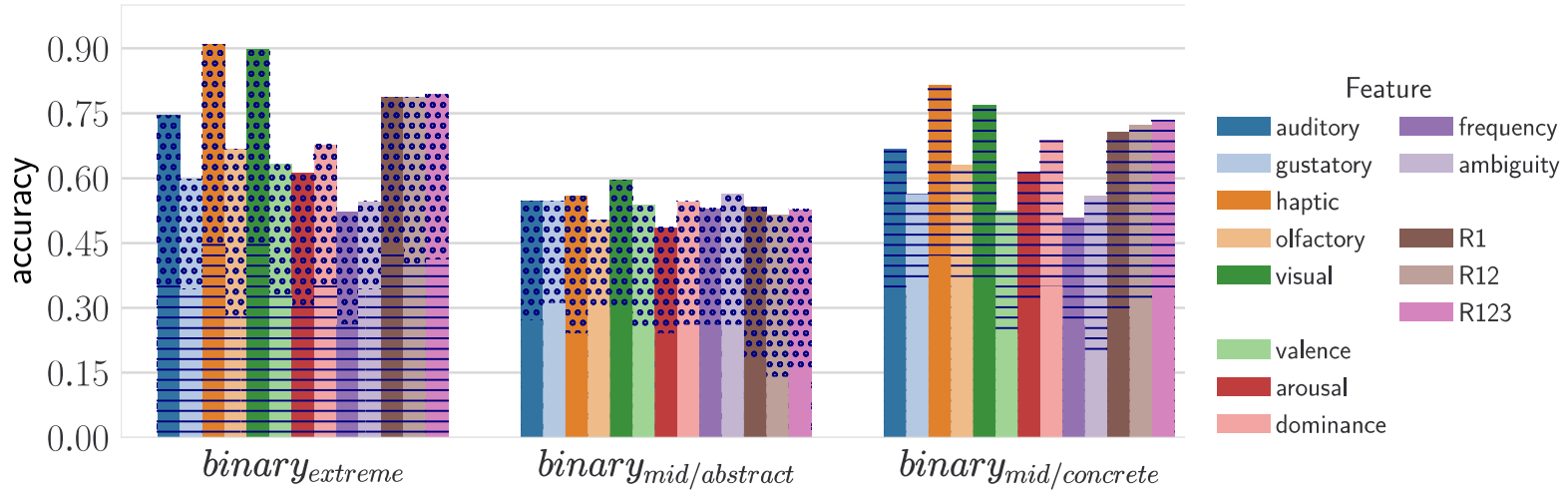
Study 1: Mid-Scale Peculiarities

Setup

- **3 binary classification tasks**
  - concrete vs. abstract (extremes)
  - abstract vs. mid-scale
  - concrete vs. mid-scale
- **Features:**
  - Lancaster Norms: auditory, gustatory, haptic, olfactory, visual
  - NRC VAD Lexicon: valence, arousal, dominance
  - ENCOW Corpus: frequency
  - WordNet: ambiguity
  - Free words associations (SWOW-EN): top 1, top 2, top 3 mostly associated words

Study 1: Mid-Scale Peculiarities

Results: classification using one feature at the time

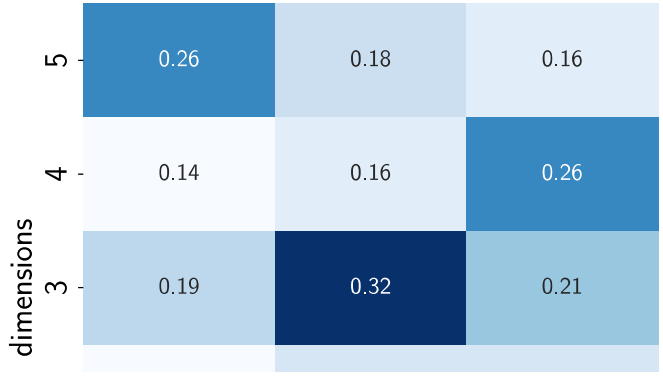


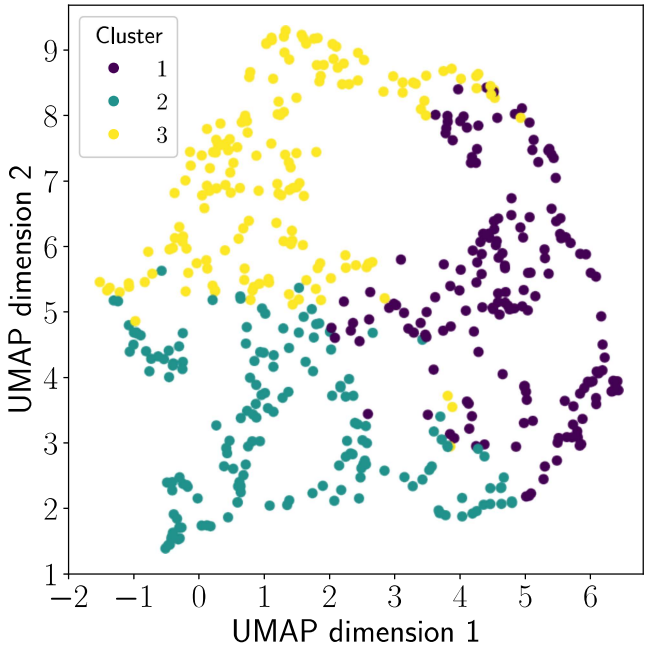
- Extreme & mid/concrete → similar patterns in "feature contribution"
- Mid/abstract → chance level
- Conclusion: Mid-scale concepts similar to abstract concepts

Study 2: Mid-Scale Disagreement Patterns

k-means hard clustering approach → individual rating distribution

C	Target	Distribution
1	*definition*	$\langle 0.32, 0.11, 0.14, 0.11, 0.32 \rangle$
	hero	$\langle 0.22, 0.11, 0.26, 0.19, 0.22 \rangle$
	percentage	$\langle 0.40, 0.03, 0.10, 0.20, 0.27 \rangle$
2	*coward*	$\langle 0.17, 0.20, 0.30, 0.20, 0.13 \rangle$
	discussion	$\langle 0.15, 0.07, 0.48, 0.15, 0.15 \rangle$
	labor	$\langle 0.16, 0.12, 0.40, 0.12, 0.20 \rangle$
3	*booster*	$\langle 0.32, 0.07, 0.14, 0.29, 0.18 \rangle$
	election	$\langle 0.20, 0.10, 0.23, 0.27, 0.20 \rangle$
	hour	$\langle 0.23, 0.07, 0.23, 0.30, 0.17 \rangle$





Discussion

- Mid-scale ratings show very different behaviours
- How can we use them in computational modeling?
  - Exclude them from the study → focus on extremes
  - Fine-tune them according to patterns of disagreement

Concepts: From Words to Images

Motivation

Tater et al. (2024)

- When describing and classifying concepts, all five senses play a critical role → but very strong visual component
- Recent multimodal studies → asymmetric contribution of vision and language
- In this study → exclusive focus on vision
- Can visual information alone distinguish between concrete vs. abstract concepts?

Is it so Obvious?

- Variable **visual similarity**
- Variable **informativeness**
- **No consistency** across concreteness groups!

Visual Information & Abstractness

Research Questions

1. Can visual attributes **differentiate** between concrete and abstract concepts?
2. How **consistent** are visual attributes across multiple images of the same concept?
3. What are **plausible failure categories** for unimodal visual representations?

Materials

Concreteness Norms & Images

- 1,000 target nouns
  - concreteness ratings from Brysbaert et al. (2014)
  - 500 highly abstract (1.07 - 1.96)
  - 500 highly concrete (4.85 - 5.00)
- Images
  - **Bing**
    - 25 images per target
    - Search-based dataset → controlled data
  - **YFCC100M** Multimedia Commons Dataset
    - 25–500 per target → coverage issue!
    - User-tagged images from Flickr → diversity

Models to Extract Visual Attributes

One feature vector for each target/property

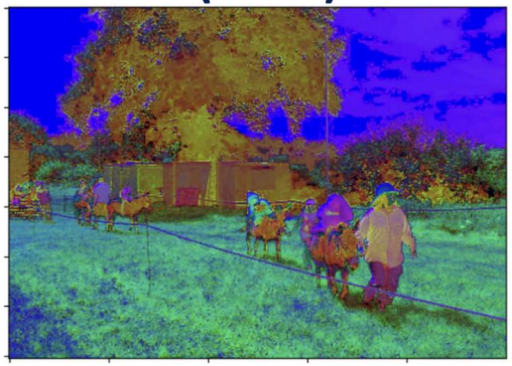
- **Low-level visual features**:
  1. **Color**
  2. **Texture**
  3. **Histogram of Oriented Gradients (HOG)**: For edge and shape information
- **High-level visual features**:
  4. **YOLO**: Object detection and location within images
  5. **GIST**: Scene context for holistic visual representation
  6. **SURF**: Identifies distinct image regions and patterns
- **Complex models for image representations**:
  7. **SimCLR**
  8. **Vision Transformer (ViT)**

Example: Visual Attributes

Target Concept: Mule

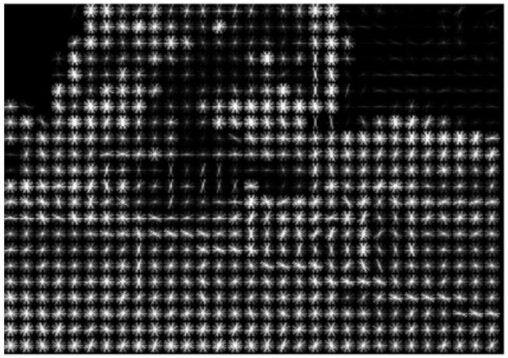
Original Image

HSV Color space (Color)

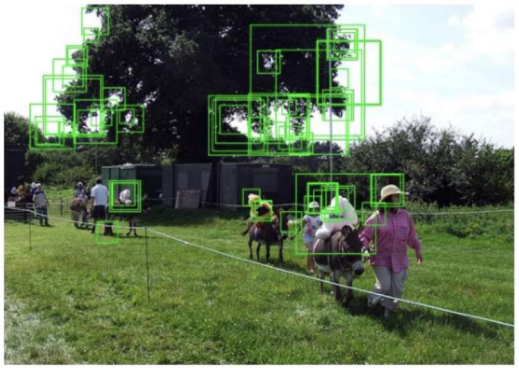


Local Binary Patterns (Texture)

Histogram of Oriented Gradients



SURF



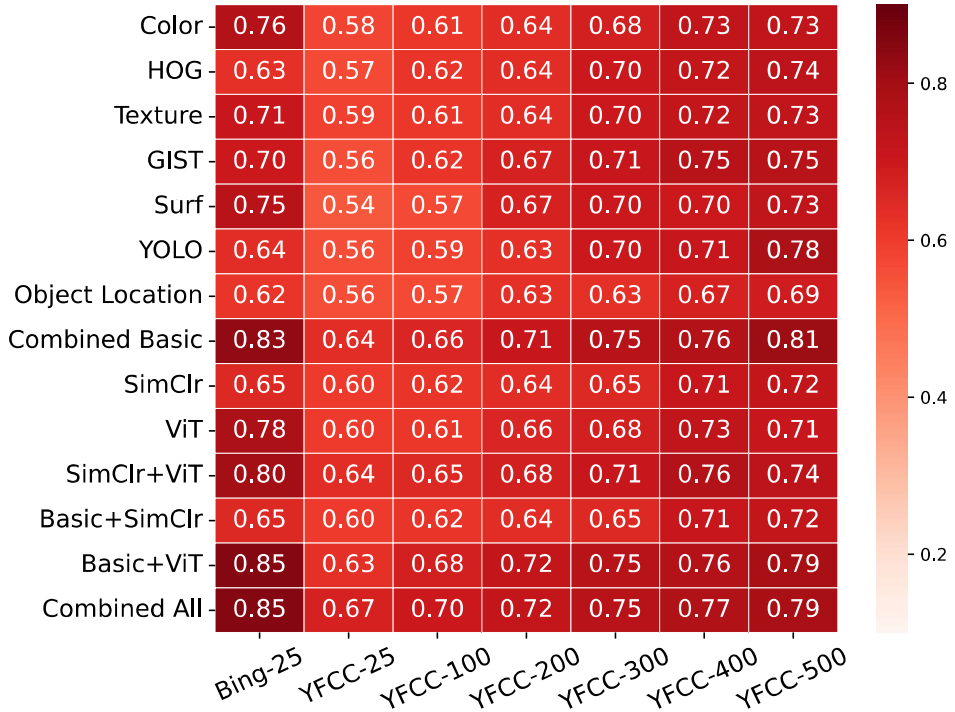
Yolo

Study 1: Classifying Concepts Using Visual Information

- What are the most useful visual features for classifying images of concrete vs. abstract nouns?
- Classifiers: **Random Forest**, Support Vector Machine, and Logistic Regression
- Features: individual and combined visual attributes

Results

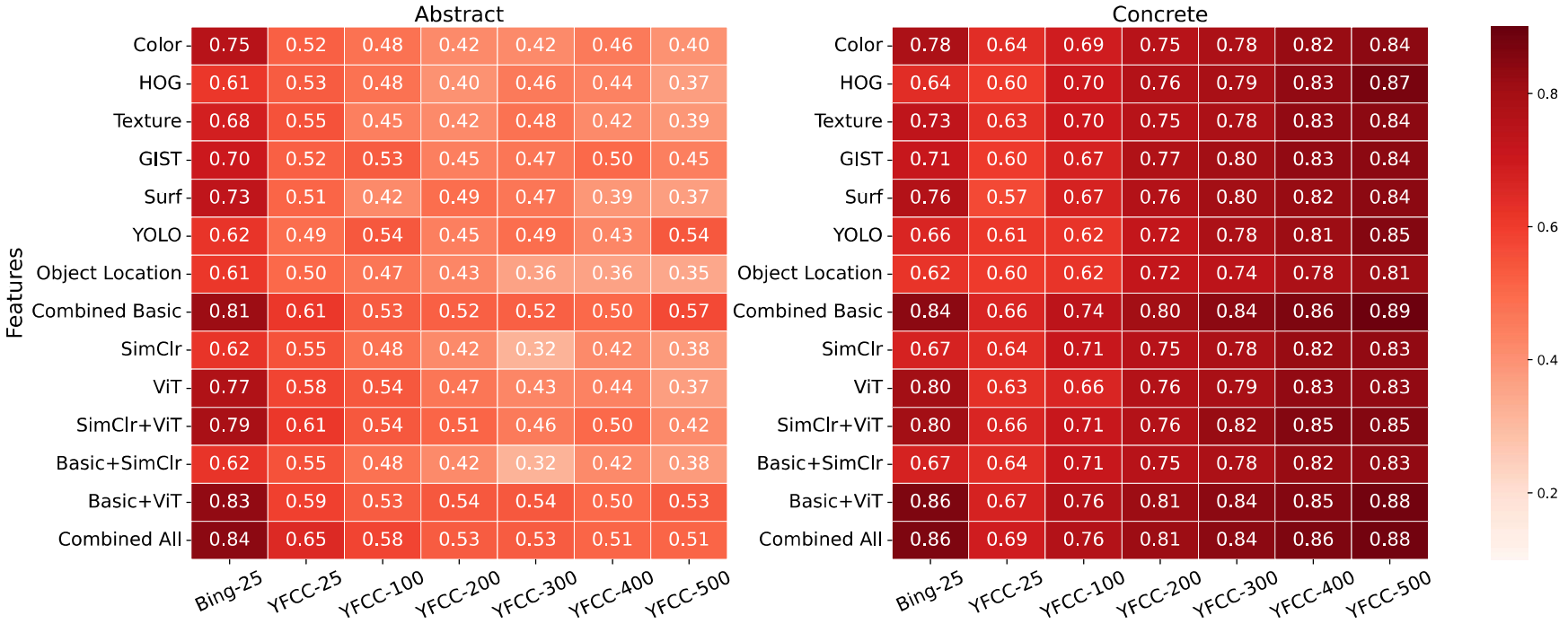
Overall F1 Scores



- BING better than YFCC
- Low level features (independently and combined) → best performance
- ViT & SimCIr → not as good

Results

Abstract vs. Concrete F1 Scores



- Bing: comparable performance for concrete and abstract
- YFCC: big improvements with more features only for concrete → data sparsity!

Study 2: Inspecting Visual Nearest Neighbors

- Assumption: to build stable representations, we need consistent images for the same target
- Method: cosine similarity among visual features → check the top 25 ... 500 neighbors
- How many neighbour images of a **banana** are also **bananas**?

Results

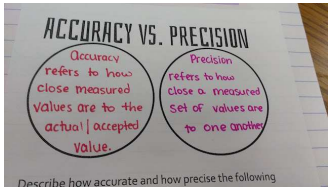
Attribute	Bing-25		YFCC-25		YFCC-500	
	A	C	A	C	A	C
Color	0.68	0.96	1.70	0.95	0.81	0.65
HOG	0.48	1.44	0.68	0.58	0.36	0.44
Texture	0.29	0.33	0.35	0.26	0.28	0.27
GIST	0.55	1.88	1.03	0.76	0.52	0.56
SURF	0.64	1.70	0.93	0.62	0.40	0.38
YOLO	2.25	3.19	1.09	1.03	1.64	1.57
Object Loc.	0.18	0.39	0.15	0.18	0.24	0.27
Combined	0.64	2.14	1.40	0.99	0.69	0.75
Simclr	0.65	1.49	1.15	0.79	0.53	0.55
ViT	2.83	26.44	3.71	6.67	2.27	6.63

- Those are percentages... → catastrophic result!
- ViT (and YOLO) best performing especially for concrete targets
- Images with similar labels share very little visual information (also for concrete targets!)

Study 3: Exploring Factors Behind Visual Diversity

- The biggest challenge in using images of a concept comes from the **diversity of the images** associated with it
- Five most frequent reasons for visual diversity

Multiple Senses

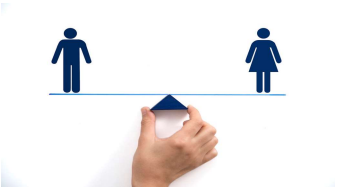


e.g., **accuracy,** generation, cone.

Physical Context

e.g., **banana,** bag, courage.

Subjective Description



e.g., **equality,** paper, laundry.

Popular Culture

e.g., **office,** apple, inception.

Lack of Visual Representation

e.g., **intention,** idealist, paradigm.

Conclusion

- In classification (concrete vs. abstract), low-level features outperform complex models → **Study 1**
- More images improve performance only for concrete nouns → **Study 1**
- Concrete and abstract targets show significant visual variability → **Study 2**
- Multiple reasons for visual variability → **Study 3**

How About Image Generation Models?

Khaliq et al. (2024)

Motivation: Diffusion-based image generation models produce high-quality images based on textual prompts

Research Questions:

1. How well do generated images capture human mental representations?
2. What is represented on the images of abstract events that are not easily depictable (e.g., *speak the truth*)?

Materials

- **Prompts:** 40 phrase-level events → verb + direct-object noun (Frassinelli & Schulte im Walde, 2019)
- **Manipulation:** degree of concreteness (concrete vs. abstract) of verbs and nouns

Verb	Noun	Verb Rating	Noun Rating
serve	food	3.8	4.8
know	man	1.7	4.8
raise	awareness	3.8	1.8
assume	responsibility	1.8	1.4

Image Generation Models

- **Four generation models + Bing**
  1. DALL-E 2
  2. Stable Diffusion 2.1
  3. Stable Diffusion XL 1.0
  4. Midjourney 5.1
  5. (Bing)
- **Settings:** Default parameters, no additional techniques or keywords to enhance the image relevance
- **800 Images:** four images from each models (4+1) $\times$ 40 verb-noun pairs

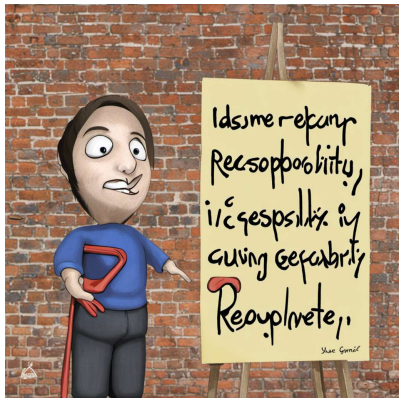
Examples of Generated Images

Example images for the event pair serve$_{3.78}$ food$_{4.8}$:



DALL-E 2 **Midjourney** **Stable Diffusion** **Stable Diffusion XL**

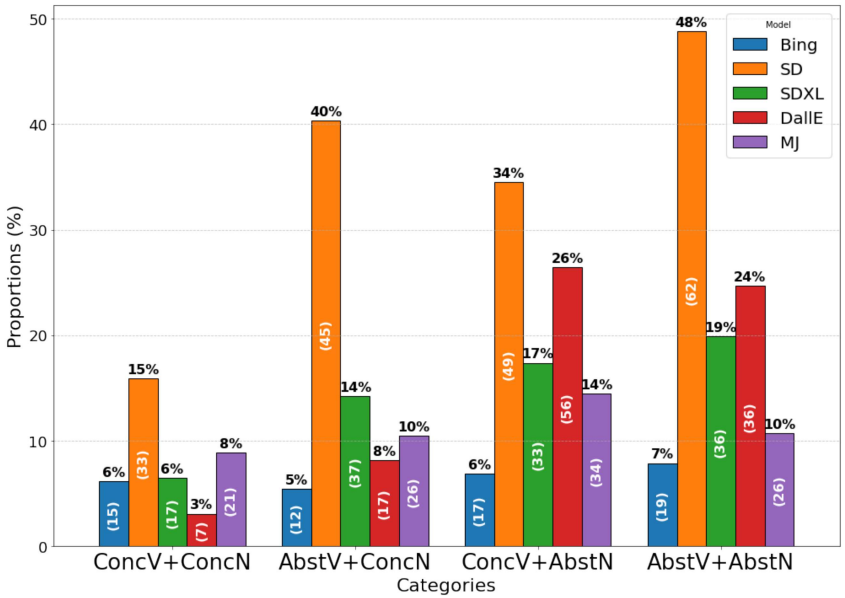
Example images for the event pair assume$_{1.8}$ responsibility$_{1.4}$:

1. Human Ratings (Crowd-sourcing)

Models Evaluation

- **Task:** Rate on a scale from 1 to 6 how well a given image depicts a given event
- **Participants:** 9 unique annotators for each image from MTurk
- **Dataset:** 4,212 ratings (after cleanup)



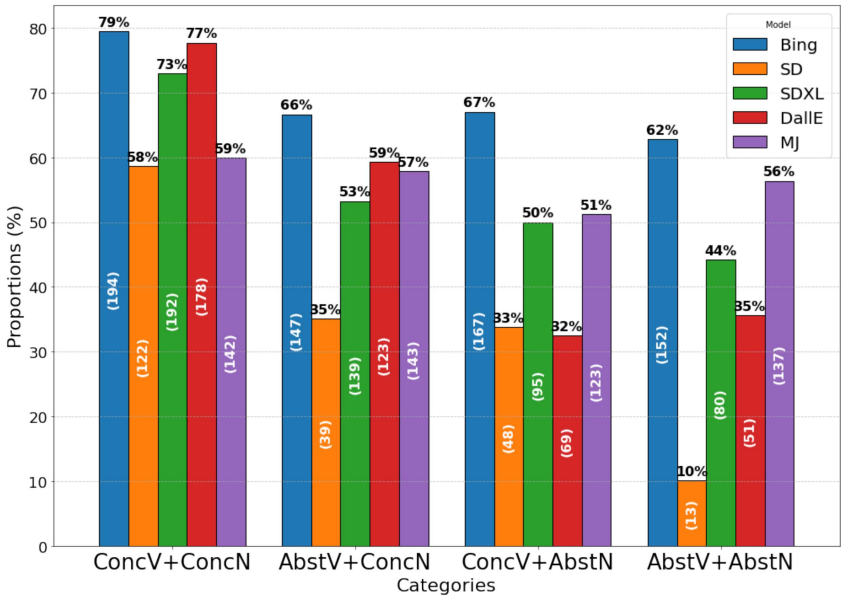
Low Ratings (1–2)

- Stable Diffusion: not good across the board
- Abstract nouns → lowest results

1. Human Ratings (Crowdsourcing)

Models Evaluation

- **Task:** Rate on a scale from 1 to 6 how well a given image depicts a given event
- **Participants:** 9 unique annotators for each image from MTurk
- **Dataset:** 4,212 ratings (after cleanup)

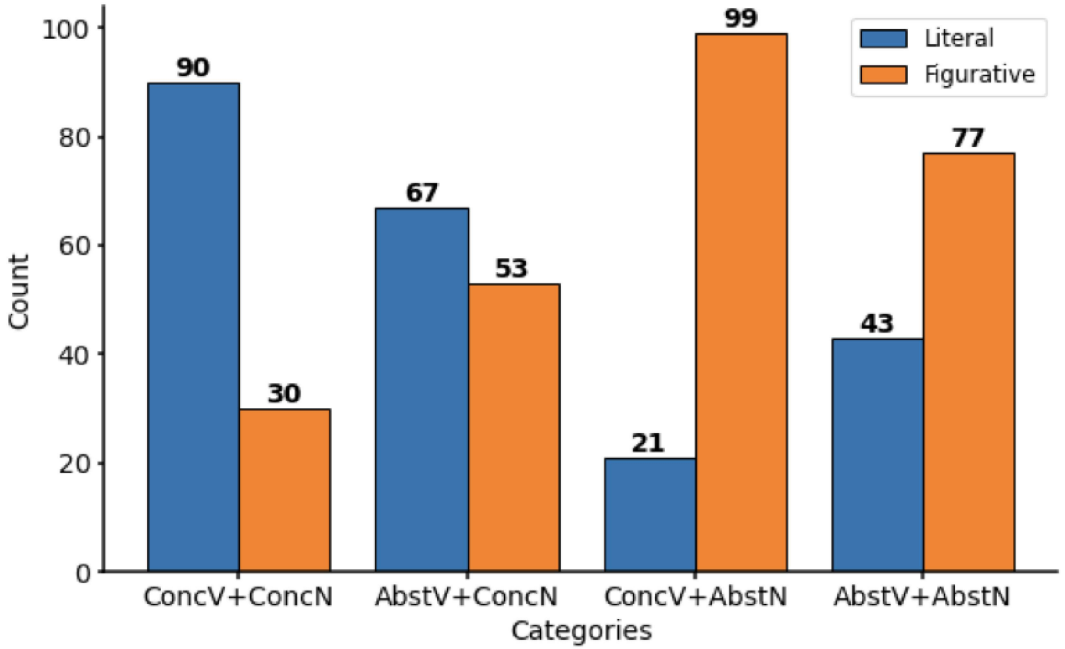


High Ratings (5–6)

- Bing: upper-bound
- DALL-E: strongest for events with concrete nouns (e.g., $pour_{4.1}$ $water_{5.0}$ and $believe_{1.5}$ $person_{4.7}$)
- Midjourney: preferred for events with abstract nouns (e.g., $raise_{3.8}$ $awareness_{1.8}$ and $remain_{1.9}$ $mystery_{2.3}$)
- Stable Diffusion XL: second best generation model in three categories

2. Figurative Language in Abstract Events

- **Task:** Rate (binary: literal or figurative) the event-pairs
- **Participants:** 12 evaluators from MTurk
- **Materials:** 40 verb-noun pairs
- **Results:** ConcV+**AbstN** perceived as strongly figurative



Conclusion

1. Humans have different opinions about model output depending on the concreteness of the targets
2. Abstract concepts are less imaginable than concrete concepts → Models apply different strategies of how to express abstractness (Tater et al., 2022)
3. Humans were most unsatisfied with images of events combining concrete verbs with abstract direct-object nouns → perceived as figurative language

Key Takeaways

1. The limitation of abstractness norms
   - Distributions by POS
   - Mid-scale values
2. Classifying abstract vs. concrete concepts using images
   - Low-level features saved the day
   - Extremely high visual variability
   - Different reasons for visual variability
3. Generating images depicting abstract and concrete concepts
   - Different models for different conditions
   - Figurative language (abstract nouns) → the tricky condition

Thank you!

- **Collaborators from Uni Stuttgart:**
  - Sabine Schulte im Walde
  - Tarun Tater
  - Aylin Wahl (UniKN)
  - Katrin Schmidt
  - Mohammed Abdul Khaliq
  - Urban Knuples

Project on Abstractness

References I

- Barsalou Lawrence W. (2003). Abstraction in Perceptual Symbol Systems. *Philosophical Transactions of the Royal Society London B. 358*.
- Brysbaert Marc, Amy Beth Warriner, Victor Kuperman (2014). Concreteness Ratings for 40 Thousand generally known English Word Lemmas. *Behavior Research Methods.*
- Della Rosa Pasquale A., Eleonora Catricala, Gabriella Vigliocco, Stefano F. Cappa (2010). Beyond the Abstract–Concrete Dichotomy: Mode of Acquisition, Concreteness, Imageability, Familiarity, Age of Acquisition, Context Availability, and Abstractness Norms for a Set of 417 Italian Words. *Behavior Research Methods.*
- Frassinelli Diego, Sabine Schulte im Walde (2019). Distributional Interaction of Concreteness and Abstractness in Verb–Noun Subcategorisation. *In Proceedings of IWCS.*
- Khaliq Mohammed, Diego Frassinelli, Sabine Schulte Im Walde (2024). Comparison of Image Generation Models for Abstract and Concrete Event Descriptions. *Proceedings of FigLang*.
- Knupleš Urban, Diego Frassinelli, Sabine Schulte im Walde (2023). Investigating the Nature of Disagreements on Mid-Scale Ratings: A Case Study on the Abstractness-Concreteness Continuum. *Proceedings of CoNLL.*

References II

- Tater, Tarun, Sabine Schulte im Walde, Diego Frassinelli (2024). Unveiling the mystery of visual attributes of concrete and abstract concepts: Variability, nearest neighbors, and challenging categories. *Proceedings of EMNLP*.
- Tater, Tarun, Diego Frassinelli, Sabine Schulte im Walde (2022). Concreteness vs. abstractness: A selectional preference perspective. *Proceedings of AACL-IJCNLP Student Research Workshop.*
- Wiemer-Hastings Katja, Xu Xu (2005). Content Differences for Abstract and Concrete Concepts. *Cognitive Science.*

Instructions

Some words refer to things or actions in reality, which you can experience directly through one of the five senses. We call these words **concrete words**. Other words refer to meanings that cannot be experienced directly but which we know because the meanings can be defined by other words. These are **abstract words**. Still other words fall in-between the two extremes, because we can experience them to some extent and in addition we rely on language to understand them. We want you to indicate how concrete the meaning of each word is for you by using a 5-point rating scale going from abstract to concrete.

Abstract (language based) *Concrete (experience based)*

1 *2* *3* *4* *5*

N = I do not know this word well enough to give a rating.

Instructions

Concrete Words

A **concrete word** comes with a higher rating and refers to something that exists in reality; you can have immediate experience of it through your senses (smelling, tasting, touching, hearing, seeing) and the actions you do. The easiest way to explain a word is by pointing to it or by demonstrating it (e.g. To explain 'sweet' you could have someone eat sugar; To explain 'jump' you could simply jump up and down or show people a movie clip about someone jumping up and down; To explain 'couch', you could point to a couch or show a picture of a couch).

Instructions

Abstract Words

An **abstract word** comes with a lower rating and refers to something you cannot experience directly through your senses or actions. Its meaning depends on language. The easiest way to explain it is by using other words (e.g. There is no simple way to demonstrate 'justice'; but we can explain the meaning of the word by using other words that capture parts of its meaning). [...]