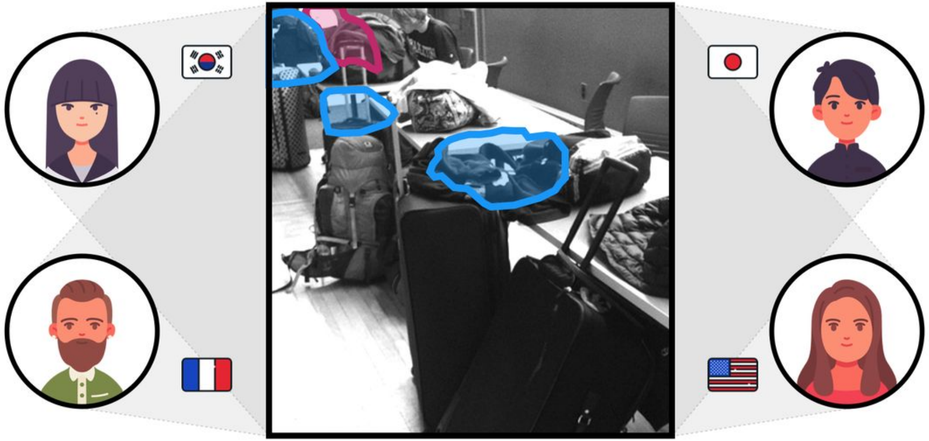
“Cultural and Linguistic Diversity Improves Visual Representations”

Andre Ye, Sebastin Santy, Jena D. Hwang, Amy X. Zhang, Ranjay Krishna

11/28/2023, RAIVN Lab Meeting

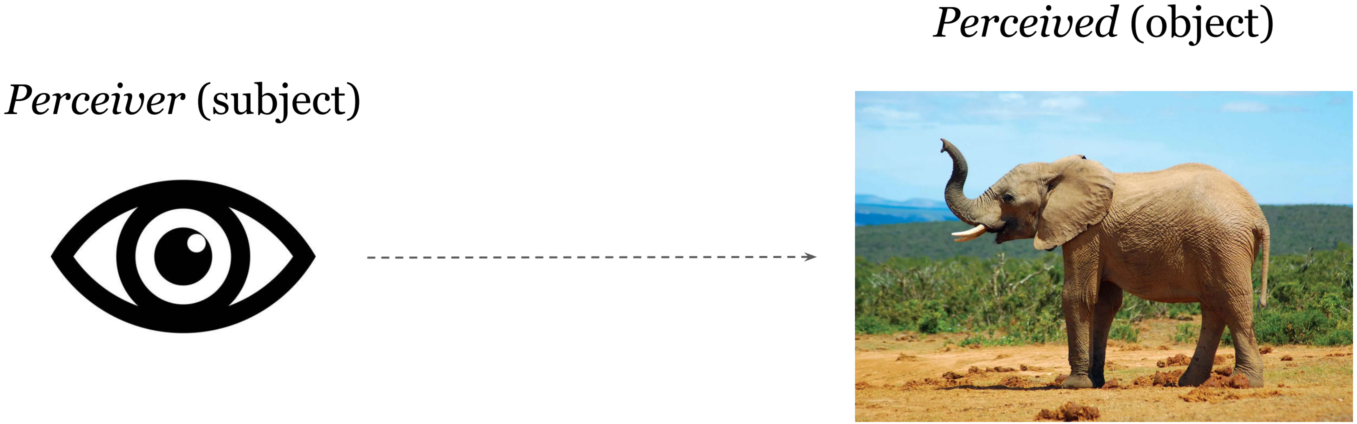


- [1] Objectivity assumptions in CV
- [2] Challenges from the social sciences
- [3] Analysis of datasets and models
- [4] Future ideas

- [1] Objectivity assumptions in CV
- [2] Challenges from the social sciences
- [3] Analysis of datasets and models
- [4] Future ideas

The “Objectivity Assumption”

- Perception is a relation between subject and object
- Assumption: perception is independent of the viewing subject’s position
  - What fundamentally matters in the perception relationship is the object
  - “Perception is objective”, “Collapse of the subject”



Perceived (object)

Perceiver (subject)

The “Objectivity Assumption”

- Perception is a relation between subject and object
- Assumption: perception is independent of the viewing subject’s position
    - What fundamentally matters in the perception relationship is the object
    - “Perception is objective”, “Collapse of the subject”

Perceived (object)

A Preoccupation with Objects

Perceiving generally unambiguous and simple categories

- 1994: MNIST. 10-way classification.

A Preoccupation with Objects

Perceiving more specialized categories

- 1994: MNIST. 10-way classification.
- 2006: ImageNet. 1000-way classification.

A Preoccupation with Objects?

Perceiving open categories

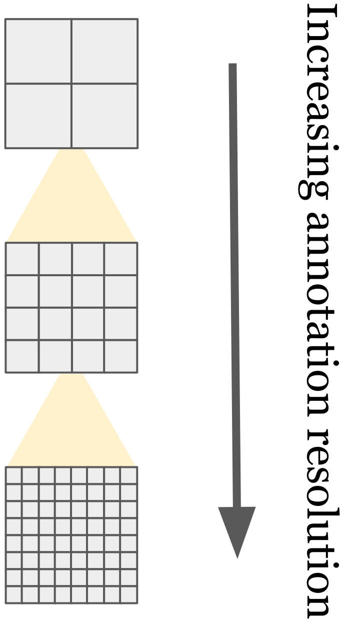
- 1994: MNIST. 10-way classification.
- 2006: ImageNet. 1000-way classification.
- 2021: LAION. Image-text pairs.

A Preoccupation with Objects?

1994: MNIST. 10-way classification.

2006: ImageNet. 1000-way classification.

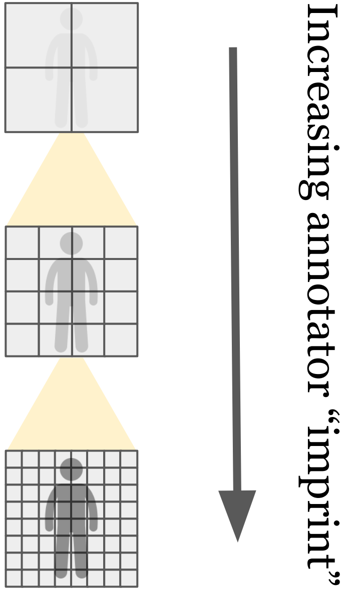
2021: LAION. Image-text pairs.



A Preoccupation with Objects?

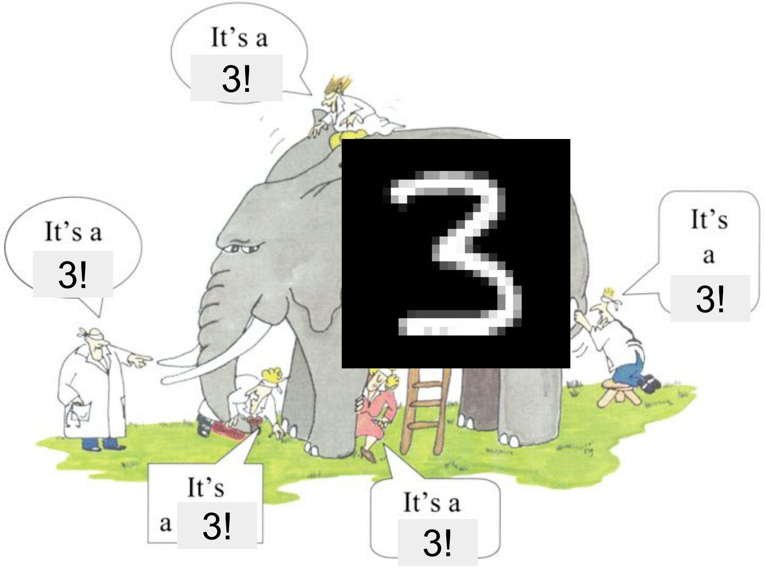
Harder to model pure “objects” and disentangle subject from object

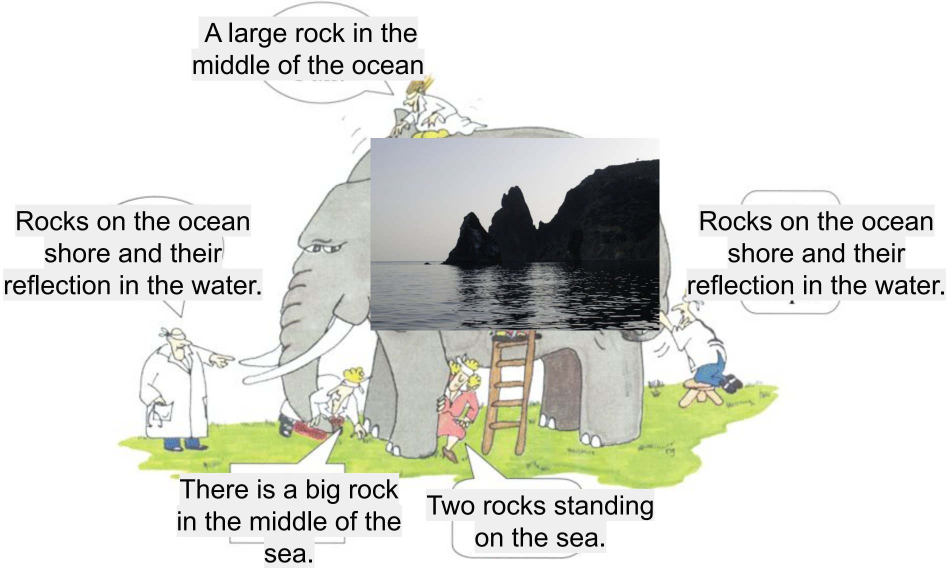
- 1994: MNIST. 10-way classification.
- 2006: ImageNet. 1000-way classification.
- 2021: LAION. Image-text pairs.



A Preoccupation with Objects?

Harder to model pure “objects” and disentangle subject from object




A large rock in the middle of the ocean
Rocks on the ocean shore and their reflection in the water.
Rocks on the ocean shore and their reflection in the water.
There is a big rock in the middle of the sea.
Two rocks standing on the sea.

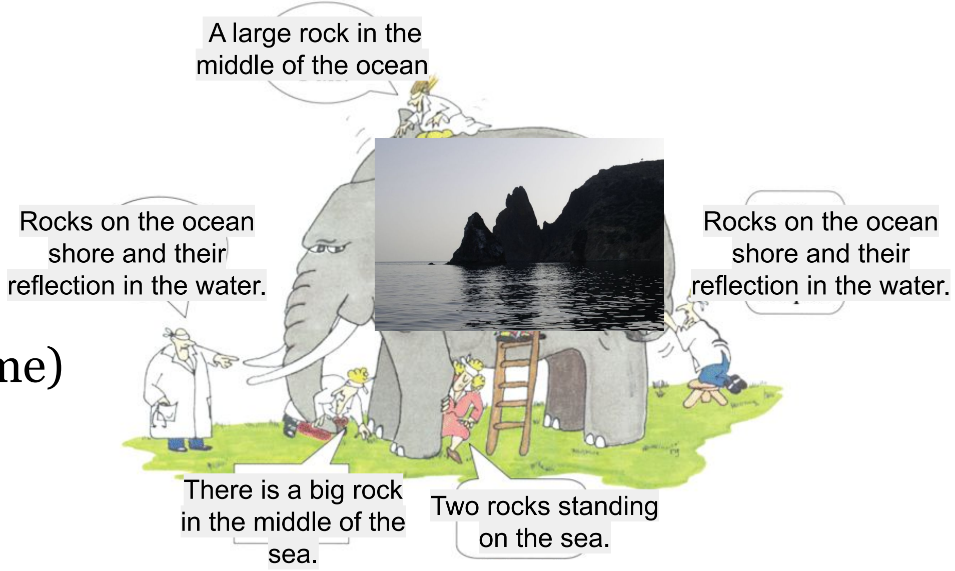
Note: Free text and perception

Complaint: text will naturally produce differences by virtue of being open

Perception is still objective! "*The subjectivity isn't substantive*"

Response:

- We don't want to perceive literally everything in the image
- Free text allows perceiving subjects to make salience judgements (more to come)
- It's important to study what parts of an image a *human* would find visually, semantically relevant



A large rock in the middle of the ocean

Rocks on the ocean shore and their reflection in the water.

Rocks on the ocean shore and their reflection in the water.

There is a big rock in the middle of the sea.

Two rocks standing on the sea.

Object Salience: The Objectivity Assumption in Captioning

“*Because people are often the end consumers of imagery, we need to be able to adopt human-centric views of recognition, especially in user applications such as image or video search.*

*For example, in response to an image search for “tree”, returning an image with a tree that **no person would ever mention** is not desirable.*”

– Berg et al., “Understanding and Predicting Importance in Images” (2012)

Objectivity Assumption: There is one universal mode of perception among humans – strong biases in how people perceive scenes and objects.

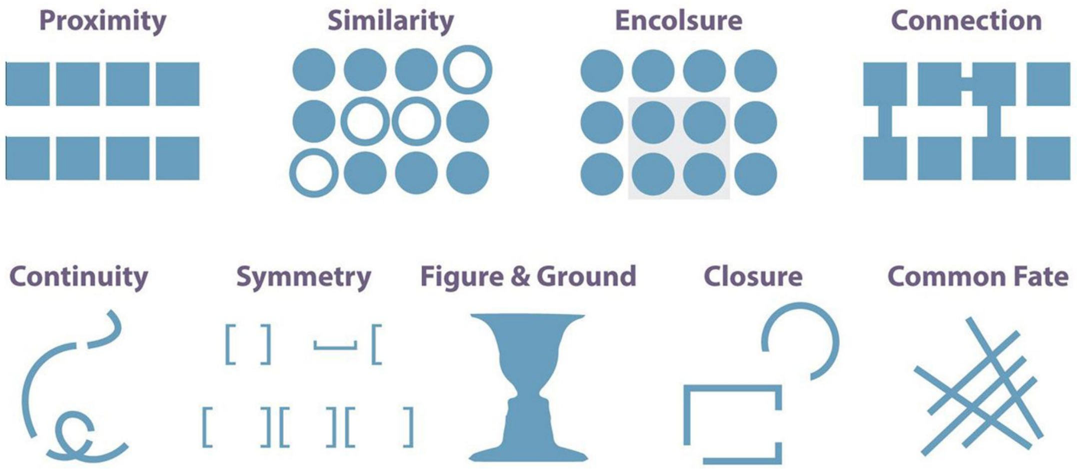
Recap

- Perception is a relation between subject and object
- Objectivity Assumption: there is one general mode of perceiving the object
    - Therefore, we can “collapse” the subject (disregard subjectivity of perception)
- Increasing resolution of perception annotations makes it harder to study pure objects, we need to deal with subject(ive) perception
- **Will the objectivity assumption hold with increasing annotation resolution?**

- [1] Objectivity assumptions in CV
- [2] Challenges from the social sciences
- [3] Analysis of datasets and models
- [4] Future ideas

Some parts of human perception appear innate…

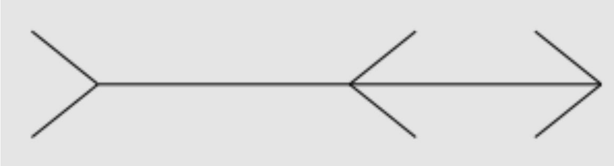
Gestalt psychology, the laws of perception



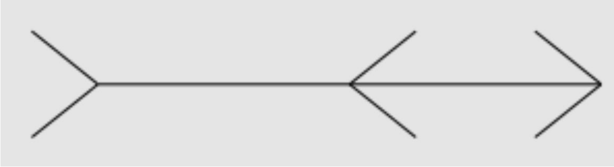
Graphic from *CG Cinematography*, Chapter 3: Gestalt Theory.

…but many parts take shape from experience

Müller-Lyer illusion: which arrow shaft is longer?



…but many parts seem to take shape from experience

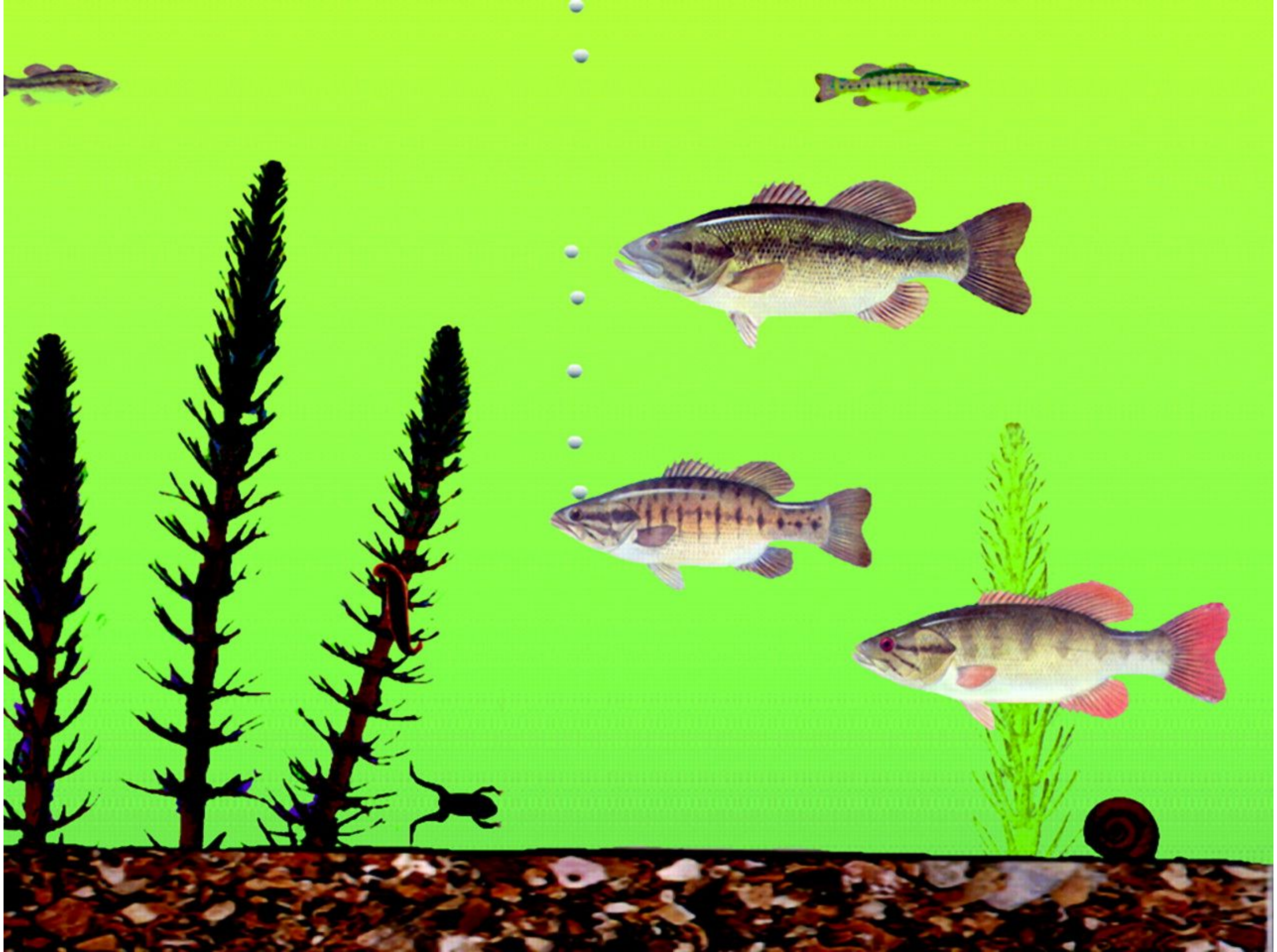


Müller-Lyer illusion: which arrow shaft is longer?

(They’re the same length! But it seems left > right…)

- 1901: W.H.R. Rivers finds indigenous people on Murray Island (Australia) less susceptible to illusion than Europeans
- 1966: Segall et al. show significant differences across 17 cultures
- “Carpentered world hypothesis”: growing up in rectangular environments (straight lines, right angles, etc.) primes susceptibility to the illusion

Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1966). The influence of culture on visual perception. Bobbs-Merrill.



…but many parts seem to take shape from experience

Nisbett & Matsuda 2003:

- Showed 20-second animation to Japanese & American participants
- Participants asked afterwards to describe the scene
- At the beginning of their statements,
  - Americans mentioned salient *objects* far more than Japanese
  - Japanese mentioned field information far more than Americans
- Japanese made 65% more field observations than Americans
- Japanese mentioned ~2x more relations between objects and the field

Nisbett, R. E., & Masuda, T. (2003). Culture and point of view. Proceedings of the National Academy of Sciences of the United States of America, 100(19), 11163–11170. https://doi.org/10.1073/pnas.1934527100

…but many parts seem to take shape from experience

Chua, Boland, Nisbett 2005:

- Tracked eye gaze in American and Chinese participants looking at scenes
- Chinese less likely to recognize previously seen object w/ diff. background
- Americans look at the foreground earlier & longer than Chinese

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. Proceedings of the National Academy of Sciences of the United States of America, 102(35), 12629–12633. https://doi.org/10.1073/pnas.0506162102

…but many parts seem to take shape from experience

What explains these results? Two modes of cognition:

- Analytic: “tendency to focus primarily on objects and their attributes”
- Holistic: “paying attention to relations among objects and their contexts”

Analytic	Individualism	Autonomy	Fulfillment	Independence
Holistic	Collectivism	Harmony	Responsibility	Interdependence

Analytic	Separability of events	Objects are stable	Resolving contradiction (dialectical)	Field-independent
Holistic	Inseparability of events	Objects are subject to change	Tolerating contradiction	Field-dependent

Koo, Minkyung, Jong An Choi, and Incheol Choi, 'Analytic versus Holistic Cognition: Constructs and Measurement', in Julie Spencer-Rodgers, and Kaiping Peng (eds), The Psychological and Cultural Foundations of East Asian Cognition: Contradiction, Change, and Holism (New York, 2018; online edn, Oxford Academic, 18 Jan. 2018), https://doi.org/10.1093/oso/9780199348541.003.0004, accessed 27 Nov. 2023.

The role of language in perception

Generally, two ways of measuring perception in psychology:

- Eye/gaze-tracking – more “subconscious”
- Language (describe/list out what you remember) – more “conscious”

Language is one viable way to measure visual “perception”.

➔ *Did you really perceive something if your eyes flicked over it, but it hasn’t really registered in your mind?*

➔ This means that language, as an elementary structure for storing information / meaning, is a factor in “perception”

Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1966). The influence of culture on visual perception. Bobbs-Merrill.

The role of language in perception

Linguistic relativism

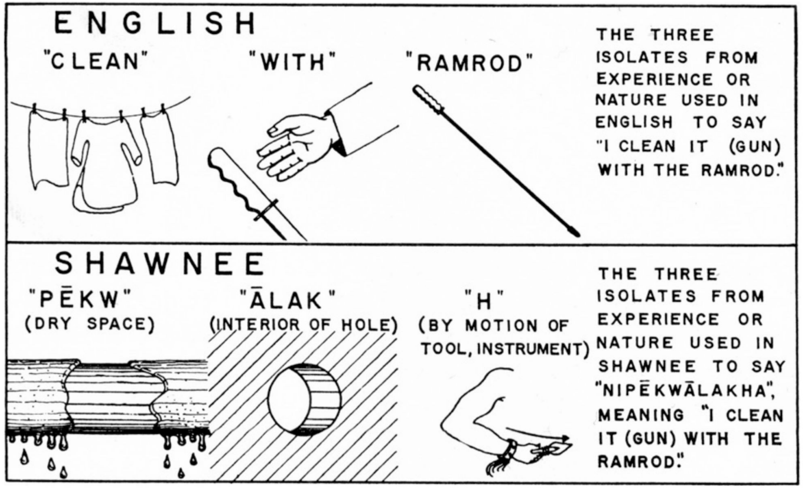
“The diversity of languages is not a diversity of signs and sounds but a diversity of views of the world.” – Wilhelm von Humboldt, 1820

“Formulation of ideas is not an independent process, strictly rational in the old sense, but is part of a particular grammar, and differs, from slightly to greatly, between different grammars.” – Benjamin Whorf, “Science and Linguistics”

Benjamin Lee Whorf, “Science and Linguistics”, Technol. Rev., 42:229-231, 247-248, no. 6 (April 1940).

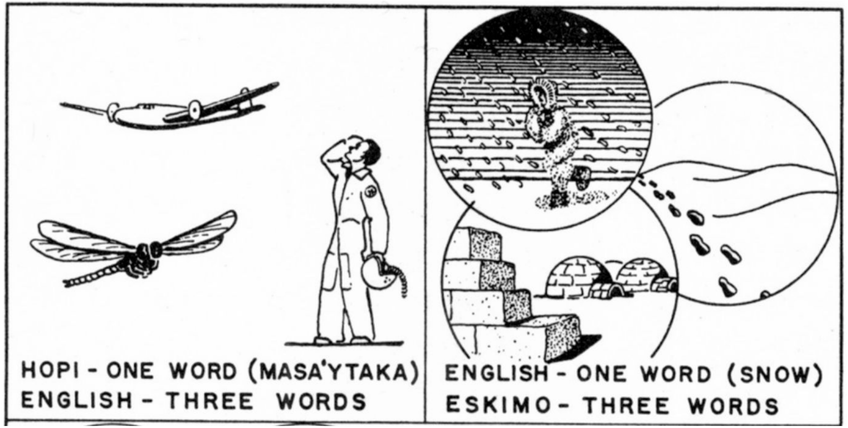
The role of language in perception

“lightning, wave, flame, meteor, puff of smoke, pulsation”
English: nouns. Hopi: verbs.



THE THREE ISOLATES FROM EXPERIENCE OR NATURE USED IN ENGLISH TO SAY "I CLEAN IT (GUN) WITH THE RAMROD."

THE THREE ISOLATES FROM EXPERIENCE OR NATURE USED IN SHAWNEE TO SAY "NIPĒKWĀLAKHA", MEANING "I CLEAN IT (GUN) WITH THE RAMROD."



Benjamin Lee Whorf, “Science and Linguistics”, Technol. Rev., 42:229-231, 247-248, no. 6 (April 1940).

The role of language in perception

Linguistic relativism: not so popular anymore with linguists.

But there are still weaker, balanced views.

- German’s complex morphosyntactic system provides events with nuanced understanding of spatial relationships.
- Russian verbs of motion encode information about means of transport and direction of movement.

Jakob Prange and Nathan Schneider. Draw mir a sheep: A supersense-based analysis of german case and adposition semantics. KI-Kunstliche Intelligenz ¨ , 35(3-4):291–306, 2021.

Recap

- Some aspects of human perception are universal.
- Other aspects, particularly attention to the field, are culturally influenced.
- Perception can be indicated by language descriptions of visual scenes.
- So, the structure & conventions of language can affect perception.

- [1] Objectivity assumptions in CV
- [2] Challenges from the social sciences
- [3] Analysis of datasets and models
- [4] Future ideas

Methodology

Question: Are there differences in the semantic concepts described by vision-language datasets, models, and applications across languages?

By the objectivity assumption, there shouldn’t be.

➔ Subjects seeing the same image should perceive roughly the same objects, beyond natural variation.

Methodology

Question: Are there differences in the semantic concepts described by vision-language datasets, models, and applications **across languages**?

- Language: reasonable and widely available proxy for annotator background
- Differences across languages = culture + linguistic structure / conventions
- Difficult to disentangle, study them indiscriminately / jointly

Methodology

Question: Are there differences in the semantic concepts described by vision-language datasets, models, and applications across languages?

- Datasets: Crossmodal (Thapliyal 2022) and Visual Genome (Krishna 2016)
- Models: Google Vertex API, LLaVA (Liu 2023)
- General approach:
  a. Obtain captions for the same image in different languages
  b. Translate all captions into English*
  c. Identify some measure of semantic content, M
  d. Show var(M(multilingual)) >> var(M(monolingual))

*I can talk more about translation if there are concerns/questions.

Note: distinguishing aims of this work

❌ Investigating geographical representation in images (Ramaswamy 2022)

❌ Investigating cultural knowledge (Liu 2021)

❌ Investigating subjective emotion & aesthetic judgements (Mohamed 2022)

✓ Investigating differences in "objective" perception

From the paper: "even the 'objective concepts' within a scene are, at their root, observed by human subjects with particular perceptual tendencies."

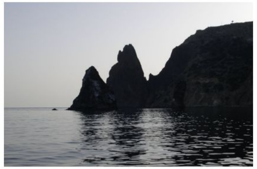
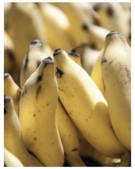
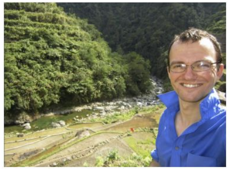
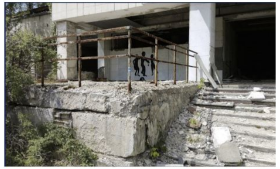
Ramaswamy, V. V., Lin, S. Y., Zhao, D., Adcock, A. B., van der Maaten, L., Ghadiyaram, D., & Russakovsky, O. (2023). Beyond web-scraping: Crowd-sourcing a geographically diverse image dataset. arXiv preprint arXiv:2301.02560.

Liu, Fangyu, et al. "Visually Grounded Reasoning across Languages and Cultures." Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021.

Youssef Mohamed, Mohamed AbdelFattah, Shyma Alhuwaider, Li Fei-Fei, Xiangliang Zhang, Kenneth Ward Church, and Mohamed Elhoseiny. Artelingo: A million emotion annotations of wikiart with emphasis on diversity over language and culture. ArXiv, abs/2211.10780, 2022. URL https://arxiv.org/abs/2211.10780.

Some motivation

Captions from Google Vertex API

Image	Language	Translated Caption
	en	A large body of water with a cliff in the background.
	de	A large rock in the middle of the ocean.
	fr	A mountain rises above the ocean.
	ru	Rocks on the ocean shore and their reflection in the water.
	zh	There is a group of rocks on the water surface.
	ja	There is a big rock in the middle of the sea.
	ko	Two rocks standing on the sea.
	en	A bunch of bananas are stacked on top of each other.
	de	A bunch of bananas with black spots on them.
	fr	A close-up of a bunch of yellow bananas.
	ru	A bunch of bananas with black spots on the skin.
	zh	A close-up of a bunch of yellow bananas.
	ja	Close-up of a bunch of bananas.
	ko	A bunch of bananas is lying on the floor.
	en	A man wearing glasses and a blue shirt smiles in front of a river.
	de	A man with glasses smiles in front of a river.
	fr	A man smiles next to a river and a forest.
	ru	A man in glasses and a blue shirt is smiling in front of the forest.
	zh	A man wearing glasses is standing by the river smiling.
	ja	A man wearing glasses is smiling at the camera in a blue shirt.
	ko	A man wearing glasses and a blue shirt is smiling towards the camera.
	en	A rusty railing with a drawing of two people on it.
	de	A destroyed building with a rusty metal railing in front of it.
	fr	An abandoned building with stairs and a rusty fence.
	ru	Concrete wall with a rusty handrail and stairs.
	zh	On the wall, there is a silhouette of a man and a woman.
	ja	Stairs in front of a building with a rusted handrail.
	ko	Stairs in front of the building with a metal fence.

Some motivation

Captions from GPT-4

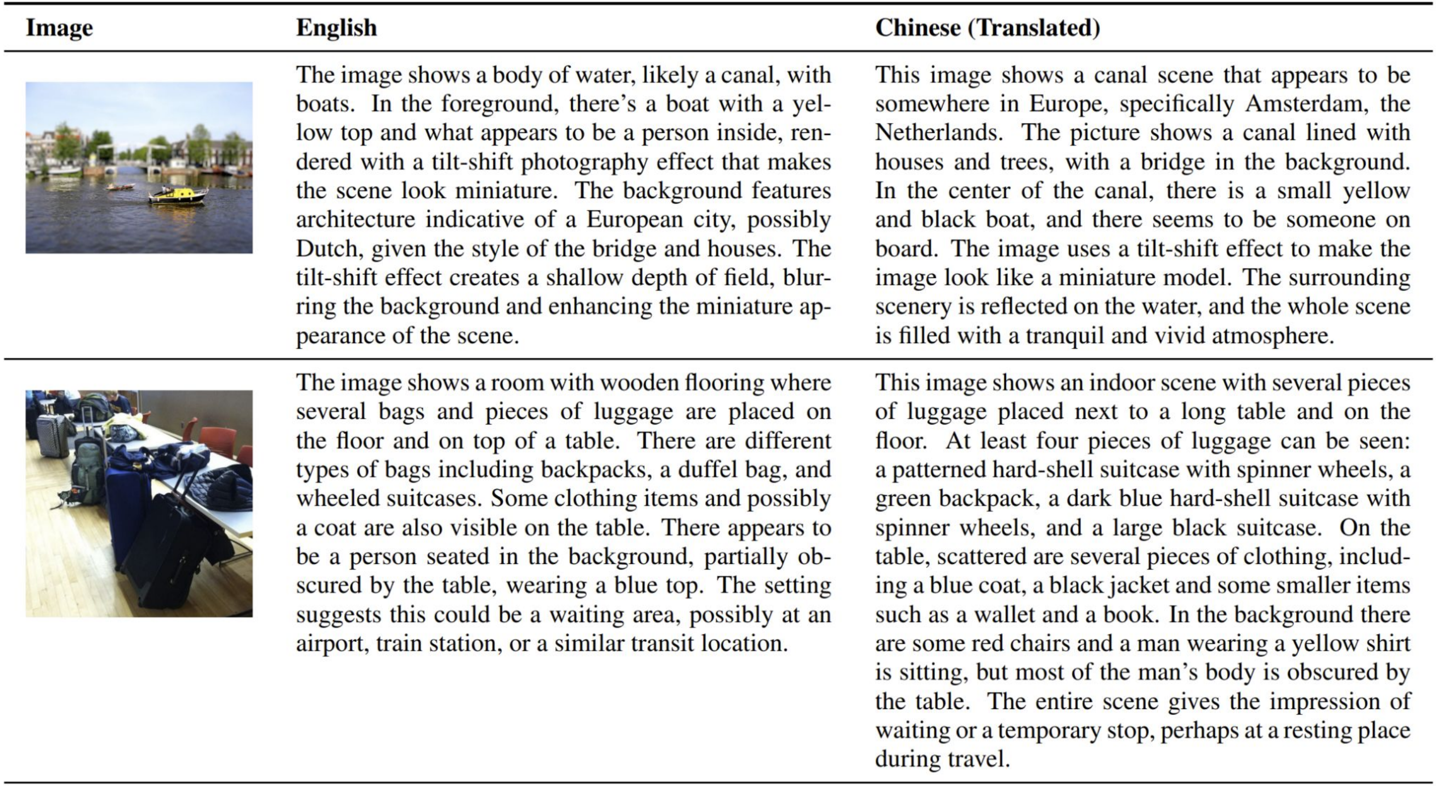
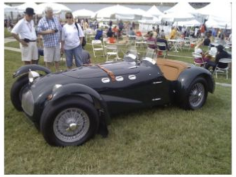
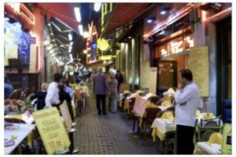


Image	English	Chinese (Translated)
	The image shows a body of water, likely a canal, with boats. In the foreground, there's a boat with a yellow top and what appears to be a person inside, rendered with a tilt-shift photography effect that makes the scene look miniature. The background features architecture indicative of a European city, possibly Dutch, given the style of the bridge and houses. The tilt-shift effect creates a shallow depth of field, blurring the background and enhancing the miniature appearance of the scene.	This image shows a canal scene that appears to be somewhere in Europe, specifically Amsterdam, the Netherlands. The picture shows a canal lined with houses and trees, with a bridge in the background. In the center of the canal, there is a small yellow and black boat, and there seems to be someone on board. The image uses a tilt-shift effect to make the image look like a miniature model. The surrounding scenery is reflected on the water, and the whole scene is filled with a tranquil and vivid atmosphere.
	The image shows a room with wooden flooring where several bags and pieces of luggage are placed on the floor and on top of a table. There are different types of bags including backpacks, a duffel bag, and wheeled suitcases. Some clothing items and possibly a coat are also visible on the table. There appears to be a person seated in the background, partially obscured by the table, wearing a blue top. The setting suggests this could be a waiting area, possibly at an airport, train station, or a similar transit location.	This image shows an indoor scene with several pieces of luggage placed next to a long table and on the floor. At least four pieces of luggage can be seen: a patterned hard-shell suitcase with spinner wheels, a green backpack, a dark blue hard-shell suitcase with spinner wheels, and a large black suitcase. On the table, scattered are several pieces of clothing, including a blue coat, a black jacket and some smaller items such as a wallet and a book. In the background there are some red chairs and a man wearing a yellow shirt is sitting, but most of the man's body is obscured by the table. The entire scene gives the impression of waiting or a temporary stop, perhaps at a resting place during travel.

Some motivation

Captions from GPT-4

Image	English	Chinese (Translated)
	The image displays a vintage black roadster with a tan interior. It has the classic design features of early to mid-20th-century sports cars, such as the rounded bodywork, prominent front wheel arches, wire-spoked wheels, and the absence of a fixed roof. The vehicle is parked on a grassy field with various people and white tents in the background, suggesting this could be at a car show or an outdoor event. There are three individuals standing near the car, two men and one woman, all casually dressed, possibly spectators or attendees of the event.	This picture shows a black retro-style sports car with smooth body lines and spoked wheels. The car is parked on a grassy field with an outdoor event scene in the background, complete with tents and people sitting in folding chairs. There were three people standing next to the car and it looked like they were discussing the car. One of them is wearing a camera, and the other is holding a document, perhaps a brochure for the event. This could be an antique car show or a gathering of car enthusiasts.
	The image shows a bustling European-style alley lined with restaurants and cafes. There are outdoor seating areas with yellow tablecloths, and large menus displayed on stands. Neon signs illuminate the area, suggesting a lively dining atmosphere. People are walking down the alley, some are seated at tables, and a man in the foreground is on a phone call. The cobblestone pavement and architectural details suggest this might be a historic or tourist area. The mood is convivial, typical of a dining district in the evening.	This picture shows a busy restaurant street. The photo shows outdoor seating at several restaurants, with tables and chairs lined with yellow tablecloths on both sides of the street. Some menu boards are placed in front of restaurants, showing prices and gastronomic options, such as ”Moules Marinière”. There were several pedestrians on the street, including a middle-aged couple taking a walk. On the right, a man in a white shirt is talking on the phone. The street is paved with cobblestones, and the buildings on both sides are decorated with neon lights and signboards, creating a cheerful and warm atmosphere. Overall, this photo captures the bustling dining scene of the city at night.

A: Scene Graph Complexity

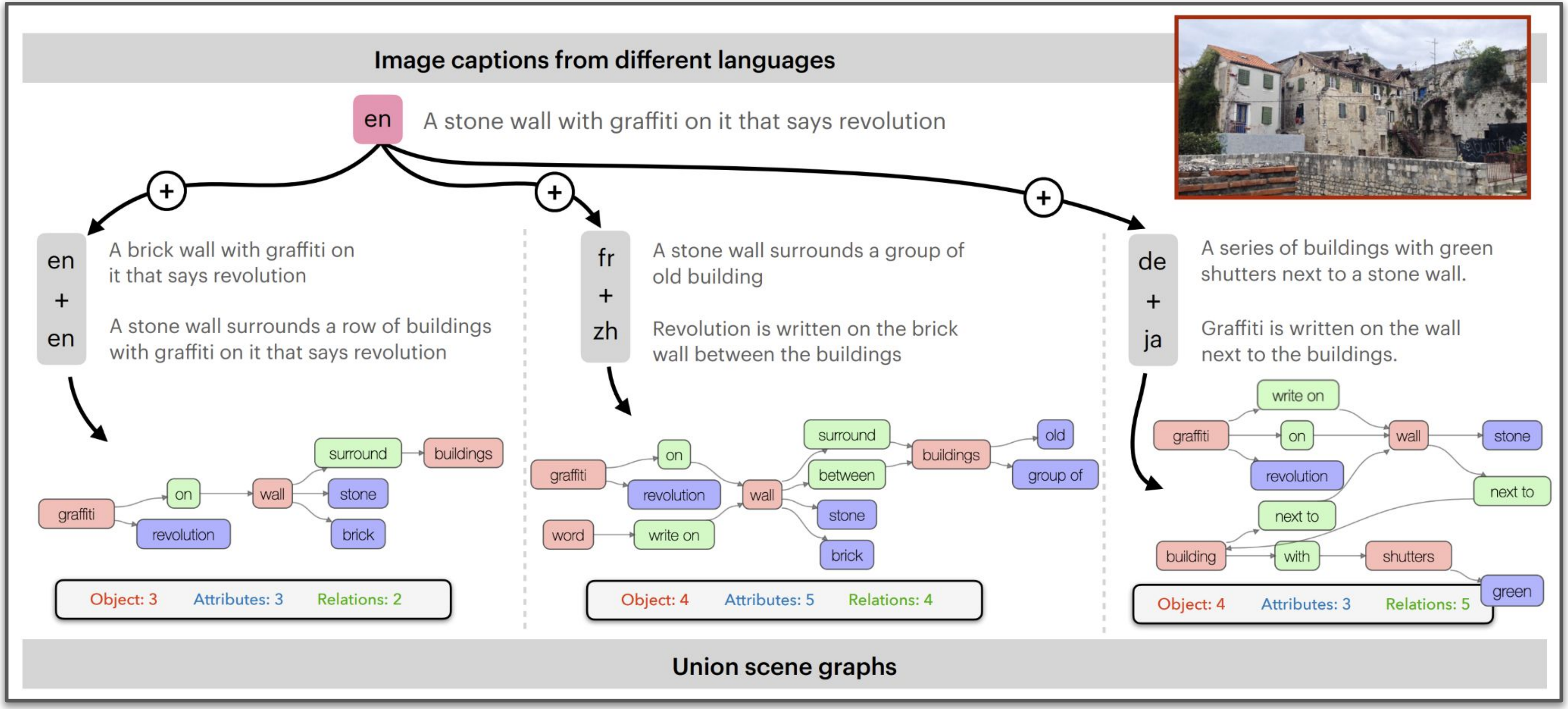


Image captions from different languages

en A stone wall with graffiti on it that says revolution

en + en

A brick wall with graffiti on it that says revolution

A stone wall surrounds a row of buildings with graffiti on it that says revolution

Object: 3 Attributes: 3 Relations: 2

fr + zh

A stone wall surrounds a group of old building

Revolution is written on the brick wall between the buildings

Object: 4 Attributes: 5 Relations: 4

de + ja

A series of buildings with green shutters next to a stone wall.

Graffiti is written on the wall next to the buildings.

Object: 4 Attributes: 3 Relations: 5

Union scene graphs

A: Scene Graph Complexity

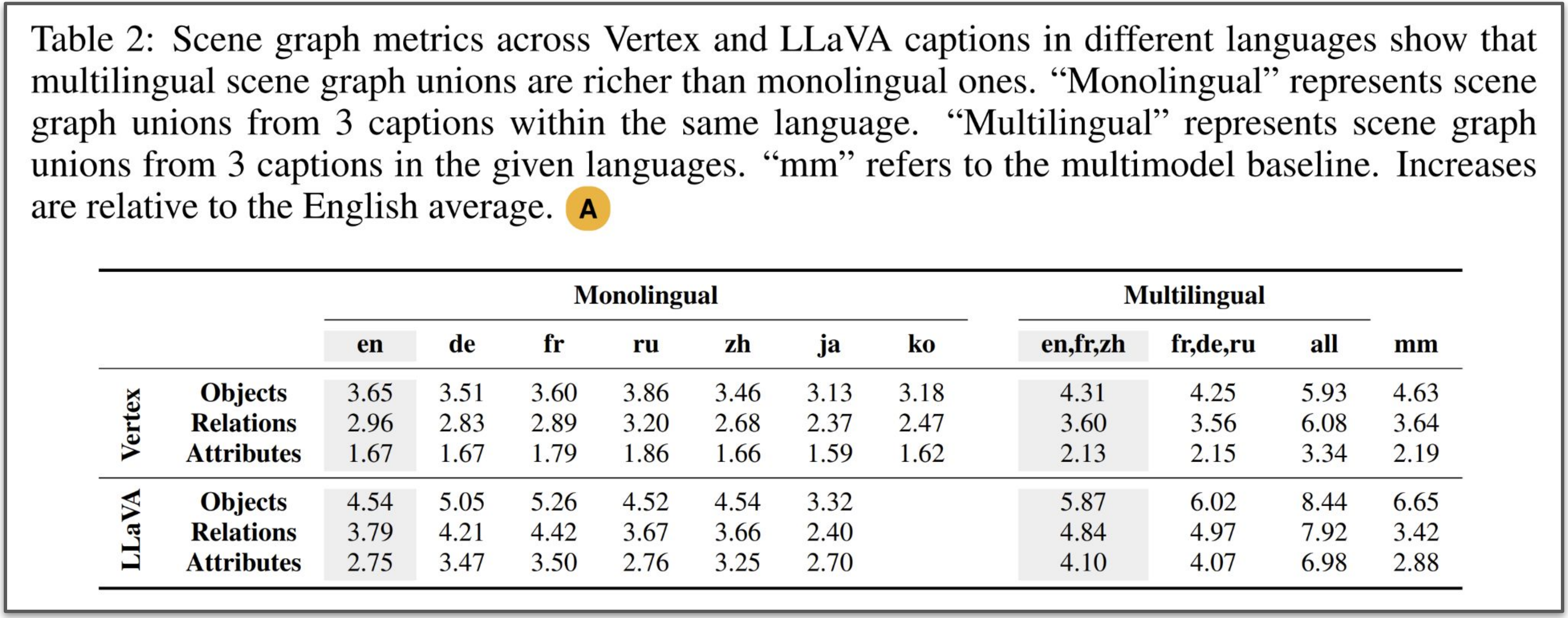


Table 2: Scene graph metrics across Vertex and LLaVA captions in different languages show that multilingual scene graph unions are richer than monolingual ones. "Monolingual" represents scene graph unions from 3 captions within the same language. "Multilingual" represents scene graph unions from 3 captions in the given languages. "mm" refers to the multimodel baseline. Increases are relative to the English average. A

		Monolingual							Multilingual			
		en	**de**	**fr**	**ru**	**zh**	**ja**	**ko**	**en,fr,zh**	**fr,de,ru**	**all**	**mm**
Vertex	**Objects**	3.65	3.51	3.60	3.86	3.46	3.13	3.18	4.31	4.25	5.93	4.63
	Relations	2.96	2.83	2.89	3.20	2.68	2.37	2.47	3.60	3.56	6.08	3.64
	Attributes	1.67	1.67	1.79	1.86	1.66	1.59	1.62	2.13	2.15	3.34	2.19
LLaVA	**Objects**	4.54	5.05	5.26	4.52	4.54	3.32		5.87	6.02	8.44	6.65
	Relations	3.79	4.21	4.42	3.67	3.66	2.40		4.84	4.97	7.92	3.42
	Attributes	2.75	3.47	3.50	2.76	3.25	2.70		4.10	4.07	6.98	2.88

B: Representational Diversity

coverage of text set = maximum pairwise embedding distance

We want to show E[coverage(multilingual)] > E[coverage(monolingual)]

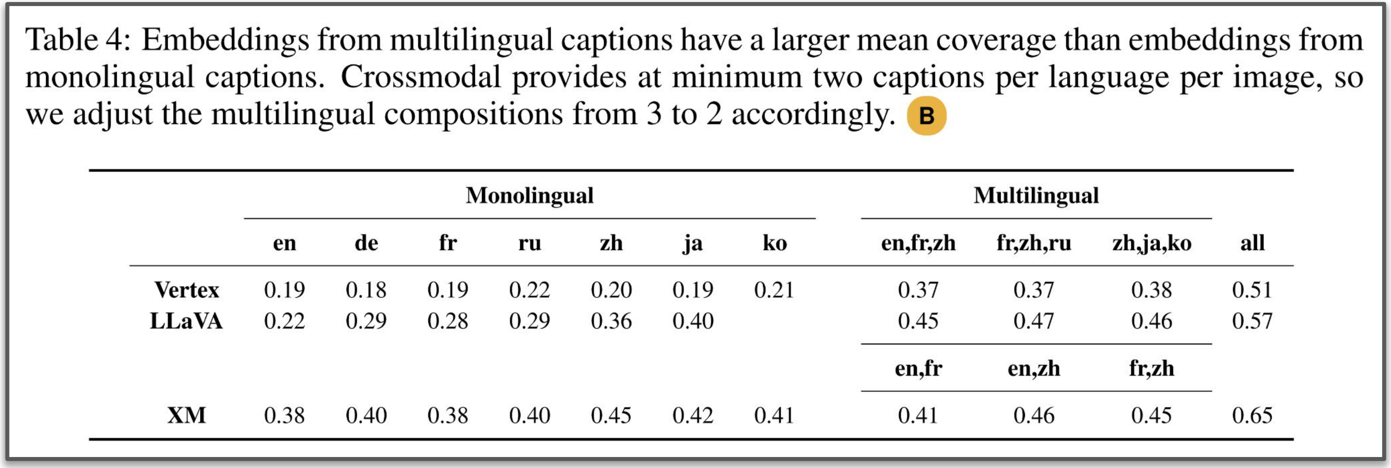


Table 4: Embeddings from multilingual captions have a larger mean coverage than embeddings from monolingual captions. Crossmodal provides at minimum two captions per language per image, so we adjust the multilingual compositions from 3 to 2 accordingly. B

	Monolingual							Multilingual			
	en	**de**	**fr**	**ru**	**zh**	**ja**	**ko**	**en,fr,zh**	**fr,zh,ru**	**zh,ja,ko**	**all**
Vertex	0.19	0.18	0.19	0.22	0.20	0.19	0.21	0.37	0.37	0.38	0.51
LLaVA	0.22	0.29	0.28	0.29	0.36	0.40		0.45	0.47	0.46	0.57
								en,fr	**en,zh**	**fr,zh**	
XM	0.38	0.40	0.38	0.40	0.45	0.42	0.41	0.41	0.46	0.45	0.65

C: Linguistic Diversity

M = some measure of linguistic diversity, range(S) = max S – min S

We want to show E[range(M(multilingual))] > E[range(M(monolingual))]

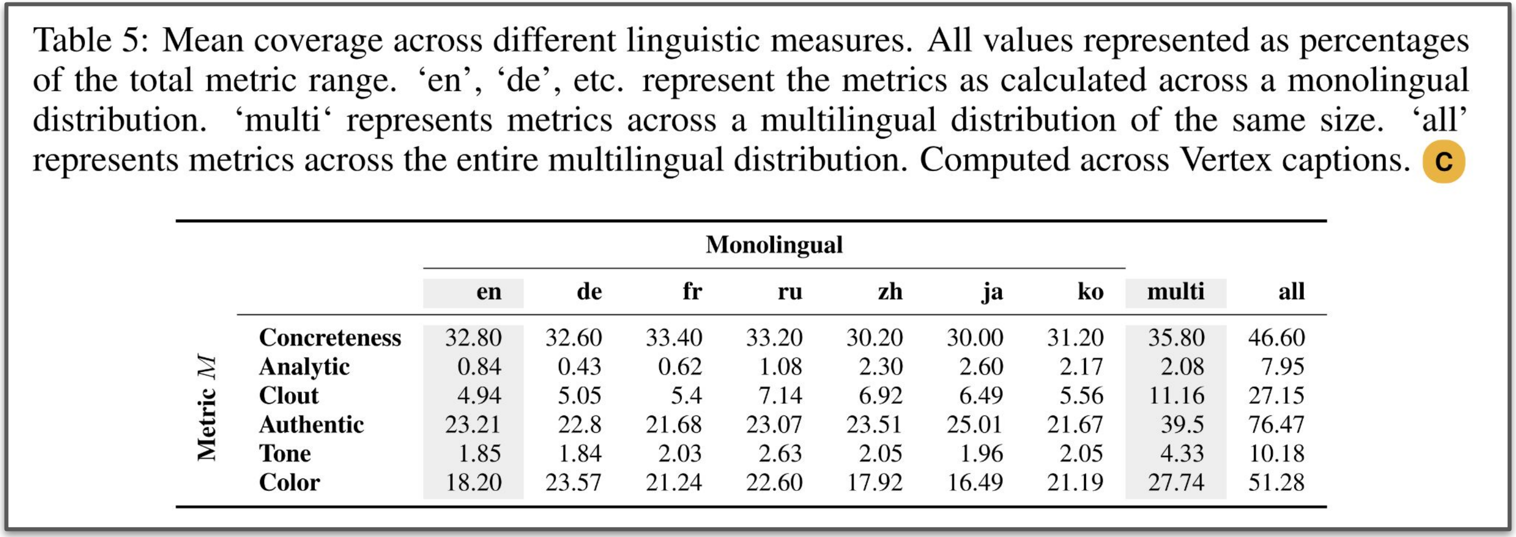


Table 5: Mean coverage across different linguistic measures. All values represented as percentages of the total metric range. ‘en’, ‘de’, etc. represent the metrics as calculated across a monolingual distribution. ‘multi‘ represents metrics across a multilingual distribution of the same size. ‘all’ represents metrics across the entire multilingual distribution. Computed across Vertex captions. C

		Monolingual								
		en	**de**	**fr**	**ru**	**zh**	**ja**	**ko**	**multi**	**all**
Metric M	**Concreteness**	32.80	32.60	33.40	33.20	30.20	30.00	31.20	35.80	46.60
	Analytic	0.84	0.43	0.62	1.08	2.30	2.60	2.17	2.08	7.95
	Clout	4.94	5.05	5.4	7.14	6.92	6.49	5.56	11.16	27.15
	Authentic	23.21	22.8	21.68	23.07	23.51	25.01	21.67	39.5	76.47
	Tone	1.85	1.84	2.03	2.63	2.05	1.96	2.05	4.33	10.18
	Color	18.20	23.57	21.24	22.60	17.92	16.49	21.19	27.74	51.28

D: Ground Truth Coverage

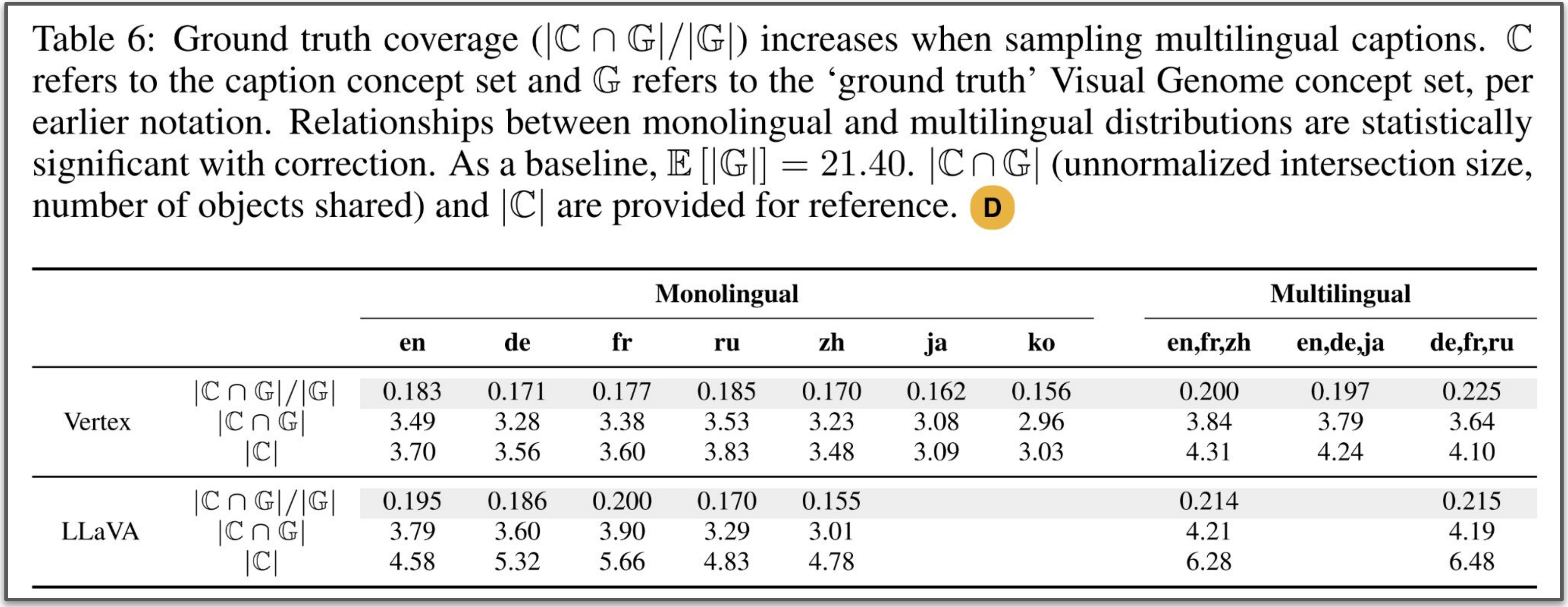


Table 6: Ground truth coverage ($|\mathbb{C} \cap \mathbb{G}|/|\mathbb{G}|$) increases when sampling multilingual captions. $\mathbb{C}$ refers to the caption concept set and $\mathbb{G}$ refers to the 'ground truth' Visual Genome concept set, per earlier notation. Relationships between monolingual and multilingual distributions are statistically significant with correction. As a baseline, $\mathbb{E}[|\mathbb{G}|] = 21.40$. $|\mathbb{C} \cap \mathbb{G}|$ (unnormalized intersection size, number of objects shared) and $|\mathbb{C}|$ are provided for reference. D

		Monolingual							**Multilingual**		
		en	**de**	**fr**	**ru**	**zh**	**ja**	**ko**	**en,fr,zh**	**en,de,ja**	**de,fr,ru**
Vertex	$|\mathbb{C} \cap \mathbb{G}|/|\mathbb{G}|$	0.183	0.171	0.177	0.185	0.170	0.162	0.156	0.200	0.197	0.225
	$|\mathbb{C} \cap \mathbb{G}|$	3.49	3.28	3.38	3.53	3.23	3.08	2.96	3.84	3.79	3.64
	$|\mathbb{C}|$	3.70	3.56	3.60	3.83	3.48	3.09	3.03	4.31	4.24	4.10
LLaVA	$|\mathbb{C} \cap \mathbb{G}|/|\mathbb{G}|$	0.195	0.186	0.200	0.170	0.155			0.214		0.215
	$|\mathbb{C} \cap \mathbb{G}|$	3.79	3.60	3.90	3.29	3.01			4.21		4.19
	$|\mathbb{C}|$	4.58	5.32	5.66	4.83	4.78			6.28		6.48

Fine-tuning on multi/monolingual data

Evaluate how models trained on text from lang *A* evaluate on text from lang *B*.

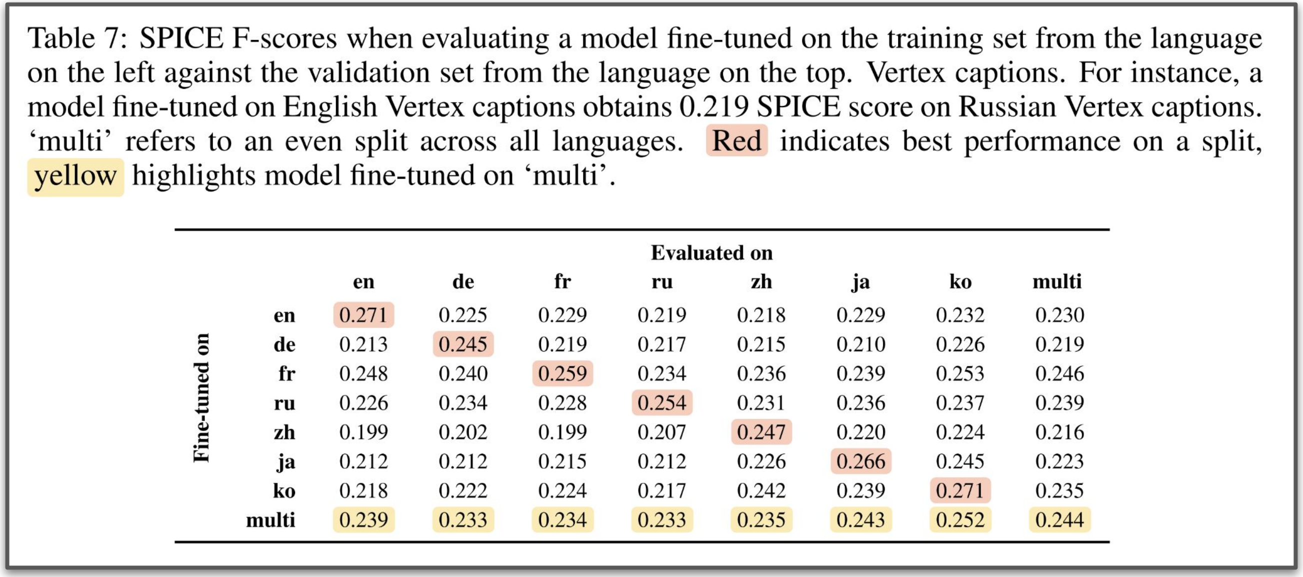


Table 7: SPICE F-scores when evaluating a model fine-tuned on the training set from the language on the left against the validation set from the language on the top. Vertex captions. For instance, a model fine-tuned on English Vertex captions obtains 0.219 SPICE score on Russian Vertex captions. 'multi' refers to an even split across all languages. Red indicates best performance on a split, yellow highlights model fine-tuned on 'multi'.

		Evaluated on							
		en	**de**	**fr**	**ru**	**zh**	**ja**	**ko**	**multi**
Fine-tuned on	**en**	0.271	0.225	0.229	0.219	0.218	0.229	0.232	0.230
	de	0.213	0.245	0.219	0.217	0.215	0.210	0.226	0.219
	fr	0.248	0.240	0.259	0.234	0.236	0.239	0.253	0.246
	ru	0.226	0.234	0.228	0.254	0.231	0.236	0.237	0.239
	zh	0.199	0.202	0.199	0.207	0.247	0.220	0.224	0.216
	ja	0.212	0.212	0.215	0.212	0.226	0.266	0.245	0.223
	ko	0.218	0.222	0.224	0.217	0.242	0.239	0.271	0.235
	multi	0.239	0.233	0.234	0.233	0.235	0.243	0.252	0.244

- [1] Objectivity assumptions in CV
- [2] Challenges from the social sciences
- [3] Analysis of datasets and models
- [4] Future ideas

Reframing multilingual modeling

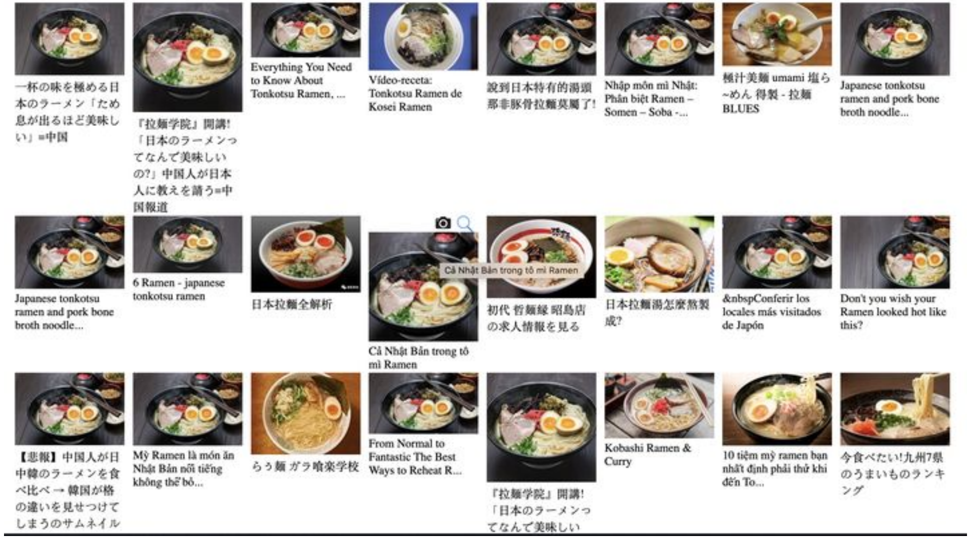
- Multilinguality as accessibility (Hu 2020)
- “Curse of multilinguality” (Conneau 2020)
  - Increasing multilinguality harms monolingual performance after some threshold
  - Accessibility vs. performance trade-off?
- Reframing: Multilingual data can hit both –
  - Accessibility: representing different forms of perception (“a tree no human would see”...)
  - Performance: there is a lot of unused, possibly very rich multilingual data out there

Multilingual data as a source of more captions

- LAION 5B / Common Crawl, WebLI, etc. have large multilingual segments
- Existing multilingual multimodal models – may not be using multilingual data in the best way (very vague and undeveloped)
    - Multilingual (multimodal) models usually encode some information about language
    - Information not shared equally/fully among all languages in multilingual models
    - Aligning across languages: focuses more on vision models “working” in different languages than embracing richness of different languages (“collapsing” vs “opening/sharing”)
- Idea: Translate multilingual LAION to English and train monolingual CLIP
    - What will be differences, if at all, from multilingual CLIP or English CLIP?
- Translation costs are expensive, what can we do about it?

Multilingual data as a source of redundant captions

- Wikipedia Image Text (WIT), LAION 5B
- Possibly very rich information from 2+ captions over 1 image
- Can we exploit shared images across captions?



一杯の味を極める日本のラーメン「ため息が出るほど美味しい」=中国

『拉麺学院』開講! 「日本のラーメンってなんで美味しいの?」中国人が日本人に教えを請う=中国報道

Everything You Need to Know About Tonkotsu Ramen, ...

Vídeo-receta: Tonkotsu Ramen de Kosei Ramen

說到日本特有的湯頭 那非豚骨拉麵莫屬了!

Nhập môn mì Nhật: Phân biệt Ramen – Somen – Soba -...

極汁美麺 umami 塩ら~めん 得製 - 拉麺 BLUES

Japanese tonkotsu ramen and pork bone broth noodle...

Japanese tonkotsu ramen and pork bone broth noodle...

6 Ramen - japanese tonkotsu ramen

日本拉麵全解析

Cả Nhật Bản trong tô mì Ramen

Cả Nhật Bản trong tô mì Ramen

初代 哲麺縁 昭島店の求人情報を見る

日本拉麵湯怎麼熬製成?

 Conferir los locales más visitados de Japón

Don't you wish your Ramen looked hot like this?

【悲報】中国人が日中韓のラーメンを食べ比べ → 韓国が格の違いを見せつけてしまうのサムネイル

Mỳ Ramen là món ăn Nhật Bản nổi tiếng không thể bỏ...

らぅ麺 ガラ喰楽学校

From Normal to Fantastic The Best Ways to Reheat R...

『拉麺学院』開講! 「日本のラーメンってなんで美味しい

Kobashi Ramen & Curry

10 tiệm mỳ ramen bạn nhất định phải thử khi đến To...

今食べたい!九州7県のうまいものランキング

Summary

Theme: turning from the object to the subject when thinking about perception

1. Objectivity assumption: *perception is objective*.
2. As annotation resolution increases, perception data reflects the subject more.
3. The social sciences show that humans from different backgrounds develop and express different perceptual tendencies.
4. We provide some evidence of these differences in CV datasets and models.
5. Using multilingual data may work both towards accessibility and performance.