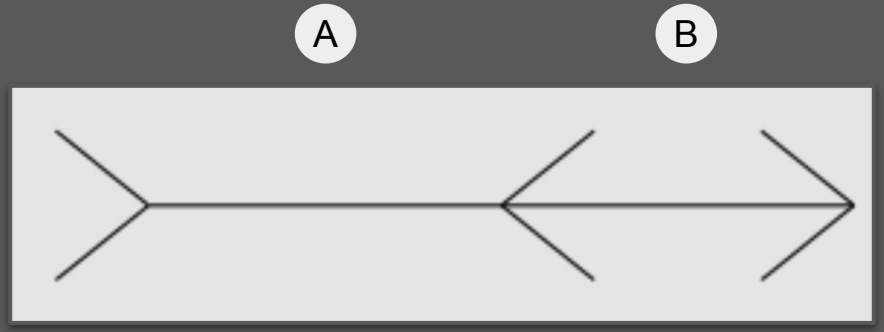
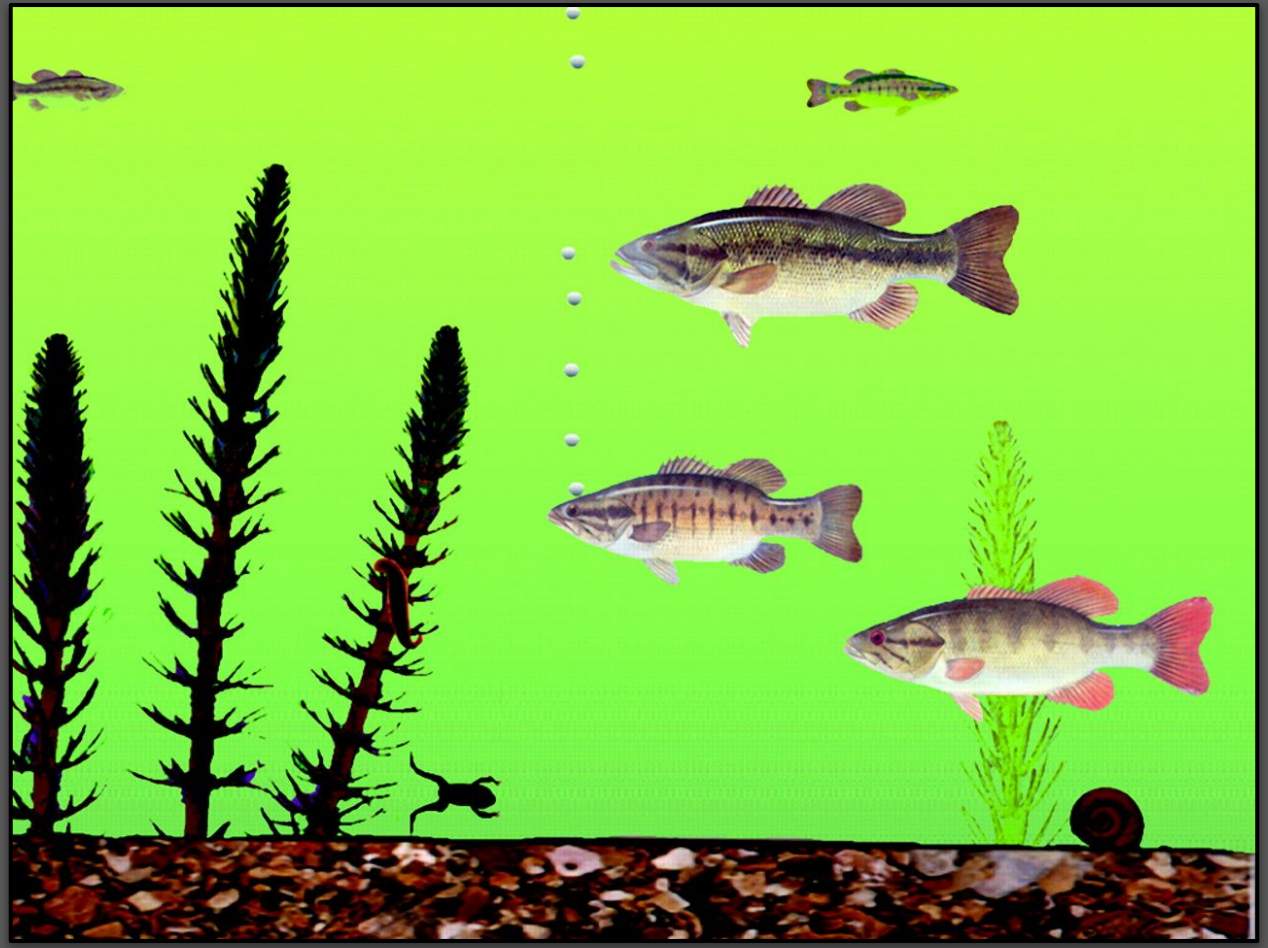
Perceptual Diversity in Computer Vision

Andre Ye, Undergraduate Research Symposium – May 17th, 2024
University of Washington

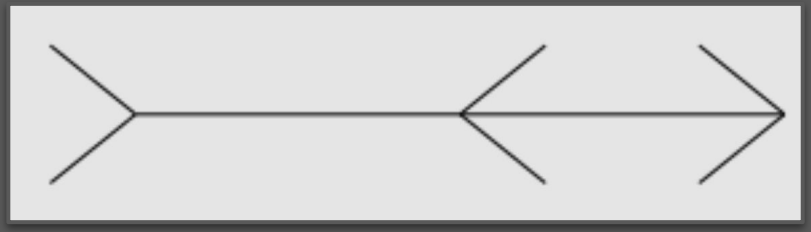
Which line segment is longer?





Different people **see differently**.

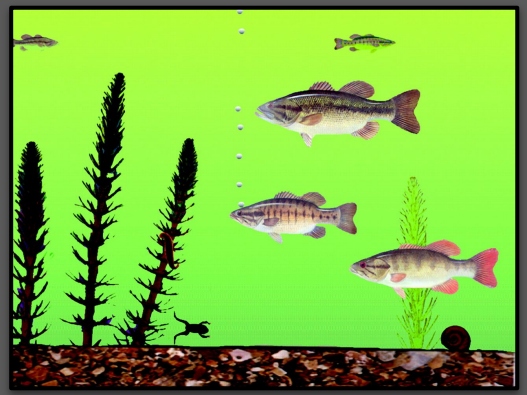
- Müller-Lyer illusion: 1901 — W.H.R. Rivers finds indigenous people on Murray Island (Australia) less susceptible to illusion than Europeans
  - “Carpentered world hypothesis”: rectangular environments primes susceptibility



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

Different people **see differently**.

- Müller-Lyer illusion: 1901 — W.H.R. Rivers finds indigenous people on Murray Island (Australia) less susceptible to illusion than Europeans
  - “Carpentered world hypothesis”: rectangular environments primes susceptibility
- Japanese made more field, relational, and behavioral observations; Americans made more foreground, object, descriptive observations



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

Different people **see differently**.

- Müller-Lyer illusion: 1901 — W.H.R. Rivers finds indigenous people on Murray Island (Australia) less susceptible to illusion than Europeans
  - “Carpentered world hypothesis”: rectangular environments primes susceptibility
- Japanese made more field, relational, and behavioral observations; Americans made more foreground, object, descriptive observations
- Chinese gaze more at background than Americans; Chinese have more context-dependent object recall

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

Different people **see differently**.

- Müller-Lyer illusion: 1901 — W.H.R. Rivers finds indigenous people on Murray Island (Australia) less susceptible to illusion than Europeans
  - “Carpentered world hypothesis”: rectangular environments primes susceptibility
- Japanese made more field, relational, and behavioral observations; Americans made more foreground, object, descriptive observations
- Chinese gaze more at background than Americans; Chinese have more context-dependent object recall

Many parts of seeing (basic functions, eye gaze, memory, description) vary by cultural and linguistic background

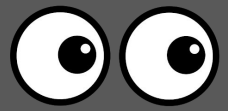
Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1966). The influence of culture on visual perception. Bobbs-Merrill.

Philosophy of **seeing**

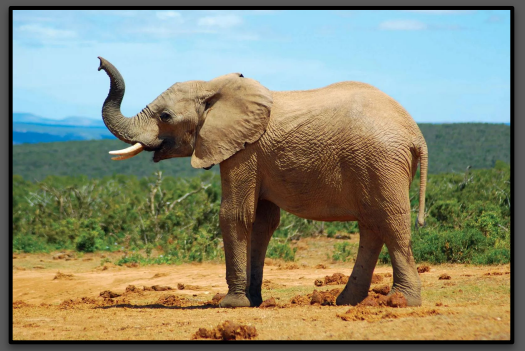
“The eye sees only what the mind is prepared to comprehend”
– Henri Bergson

- Perception is a relation between *subject* and *object*

Perceiver (subject)



Perceived (object)

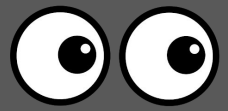


Philosophy of **seeing**

“The eye sees only what the mind is prepared to comprehend”
– Henri Bergson

- Perception is a relation between *subject* and *object*
- The object is perceived *by the subject*

Perceiver (subject)



Perceived (object)

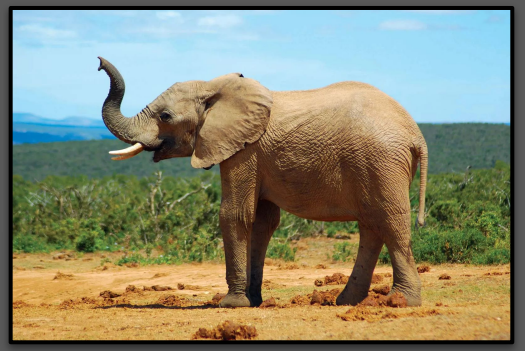
Philosophy of **seeing**

"The eye sees only what the mind is prepared to comprehend"
– Henri Bergson

- Perception is a relation between *subject* and *object*
- The object is perceived *by the subject*
- Is there a "pure object w/out a subject?"
  - Is there an "objective object of vision"?

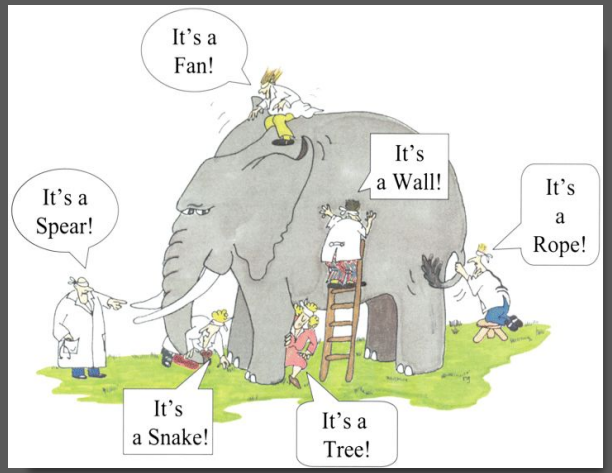
Perceiver (subject)

Perceived (object)



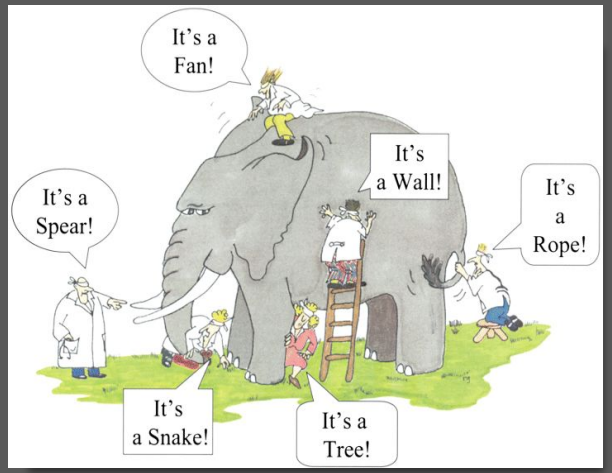
Philosophy of **seeing**

Is there a "pure object w/out a subject?" Practically, no.



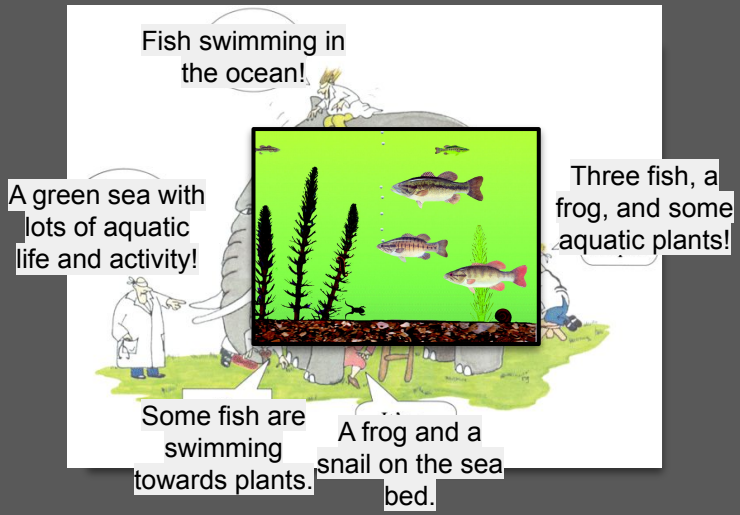
Philosophy of **seeing**

Is there a “pure object w/out a subject?” Practically, no.
There is no “true elephant” to see. We are all “blind” in some way.



Philosophy of **seeing**

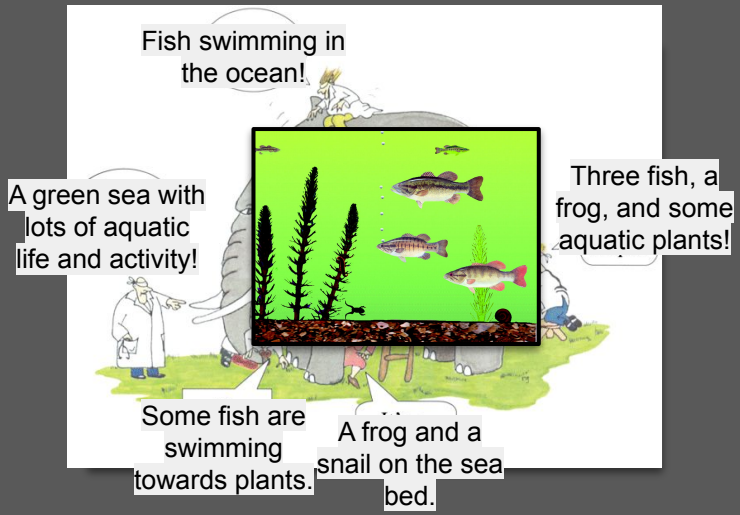
Is there a “pure object w/out a subject?” Practically, no.
There is no “true elephant” to see. We are all “blind” in some way.



Philosophy of **seeing**

Is there a “pure object w/out a subject?” Practically, no.
There is no “true elephant” to see. We are all “blind” in some way.

When we build models to “see”, we should take this into account.



Current state of computer vision models & datasets

- CV datasets are lists of (image, annotation) pairs
- Annotations are treated as generic and true in training
- Annotations do not record annotator background

Microsoft COCO Dataset Captions (item #403981)
“a person sits on a crate in front of bananas.”
“man sitting on a crate near large bunches of bananas.”
“a man sits by bunches of bananas at an outdoor market.”

Current state of computer vision models & datasets

- CV datasets are lists of (image, annotation) pairs
- Annotations are treated as generic and true in training
- Annotations do not record annotator background

CV conceptualizes annotations as the “objective” ways to see an image;
It does not account for **differences in how subjects see**.

Microsoft COCO Dataset Captions (item #403981)
“a person sits on a crate in front of bananas.”
“man sitting on a crate near large bunches of bananas.”
“a man sits by bunches of bananas at an outdoor market.”

Recap: we see differently; CV assumes we don't

Where we currently are

1. Psychological studies: perceptual variation across culture & language
2. Philosophical reflection: the object is always *in relation to* the subject
3. Computer Vision: let's just record the objects & forget about the subject

Recap: we see differently; CV assumes we don't

Where we currently are

1. *Psychological studies:* perceptual variation across culture & language
2. *Philosophical reflection:* the object is always *in relation to* the subject
3. *Computer Vision:* let's just record the objects & forget about the subject

RQ: Does perceptual variation across languages and cultures appear in Computer Vision datasets, models, and APIs?

Methodology

3 data sources, 6 languages (en, fr, de, ru, zh, ja, ko) translated to en

- **XM dataset.** Human-annotated collected captions
- **LLaVA model.** Common open-source image captioning model
- **Vertex API.** Commercial API for image captioning from Google

Methodology

3 data sources, 6 languages (en, fr, de, ru, zh, ja, ko) translated to en

- **XM dataset.** Human-annotated collected captions
- **LLaVA model.** Common open-source image captioning model
- **Vertex API.** Commercial API for image captioning from Google

Methodology: compare captions across 6 languages for the same images

Methodology

3 data sources, 6 languages (en, fr, de, ru, zh, ja, ko) translated to en

- **XM dataset.** Human-annotated collected captions
- **LLaVA model.** Common open-source image captioning model
- **Vertex API.** Commercial API for image captioning from Google

Methodology: compare captions across 6 languages for the same images

How do you compare perception across captions?

Methodology

3 data sources, 6 languages (en, fr, de, ru, zh, ja, ko) translated to en

- **XM dataset.** Human-annotated collected captions
- **LLaVA model.** Common open-source image captioning model
- **Vertex API.** Commercial API for image captioning from Google

Methodology: compare captions across 6 languages for the same images

- **Language-level**: things which are present in the language itself
- **Model-level**: things which are picked up by / relevant to models

Methodology

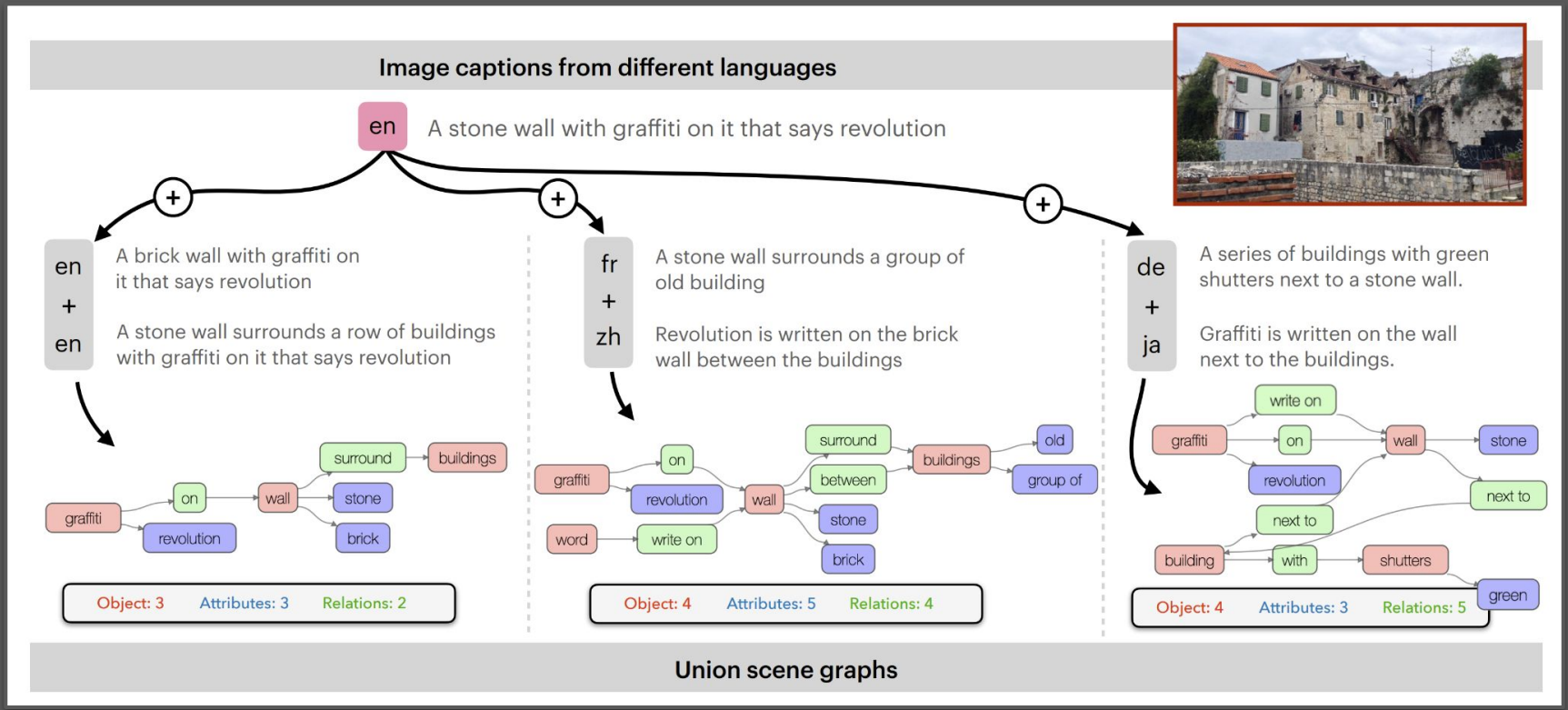
3 data sources, 6 languages (en, fr, de, ru, zh, ja, ko) translated to en

- **XM dataset.** Human-annotated collected captions
- **LLaVA model.** Common open-source image captioning model
- **Vertex API.** Commercial API for image captioning from Google

Methodology: compare captions across 6 languages for the same images

- **Language-level**: things which are present in the language itself
  - **Semantic content:** what actual things are being said in the caption?
  - **Manner of expression:** how is the caption being expressed?
- **Model-level**: things which are picked up by / relevant to models
  - **Model representation:** how do models “read” (represent) the caption?
  - **Model generation:** how does training models on captions affect future captions?

Semantic Content



Semantic Content

Multilingual scene graphs have more objects, relations, and attributes than monolingual scene graphs: **diff. lang captions mention diff. things**

Table 1. **Semantic content evaluation results.** Mean number of objects ("obj.s"), relations ("rel.s"), and attributes ("attr.s") of scene graphs unioned from either a monolingual or multilingual set of three captions. "avg" lists mean results across all mono- and multi-lingual language triplets. "$3 \times X$" means "three captions of language X".

	Caption Source →	Vertex			LLaVA			XM		
	Metric→	Obj.s	Rel.s	Attr.s	Obj.s	Rel.s	Attr.s	Obj.s	Rel.s	Attr.s
Mono	3×en	3.65	2.96	1.67	4.54	3.79	2.75	2.59	1.54	1.27
	3×fr	3.51	2.83	1.67	5.05	4.21	3.47	2.92	1.76	1.66
	3×de	3.60	2.89	1.79	5.26	4.42	3.50	3.16	1.94	1.97
	3×ru	3.86	3.20	1.86	4.52	3.67	2.76	3.03	1.88	1.74
	zh.zh,zh	3.46	2.68	1.66	4.54	3.66	3.25	2.99	1.71	2.01
	ja.ja.ja	3.13	2.37	1.59	-	-	-	3.41	1.99	2.47
	3×ko	3.18	2.47	1.62	-	-	-	2.71	1.59	1.46
	avg	3.48	2.77	1.98	4.78	3.95	3.15	2.98	1.77	1.78
Multi	en,fr,zh	4.13	3.45	2.29	6.14	4.85	4.00	3.71	2.41	2.36
	fr,zh,ru	4.24	3.48	2.40	6.15	4.76	3.99	3.92	2.57	2.59
	de,fr,ru	4.13	3.38	2.33	6.25	4.97	4.07	3.93	2.57	2.55
	avg	4.17	3.40	2.33	5.93	4.54	3.86	4.35	2.94	2.97

Manner of Expression

Multilingual caption sets have more variation in MoE metrics than monolingual caption sets: **diff. lang captions differ in MoE**

Table 2. **Manner of expression evaluation results.** Mean expressive coverage of monolingual and multilingual sets of captions, as measured by the concreteness ("conc."), authenticity ("auth."), and tone lexical measures. "avg" lists mean results across all mono- and multi-lingual language triplets. Two measures excluded for space; see Table 11 for complete data.

Caption Source→		Vertex			LLaVA			XM		
Metric→		Conc.	Auth.	Tone	Conc.	Auth.	Tone	Conc.	Auth.	Tone
Mono	3×en	1.64	23.21	1.85	2.17	35.97	6.79	1.80	33.21	8.62
	3×fr	1.64	22.80	1.84	2.44	38.01	9.53	2.24	40.09	12.16
	3×de	1.67	21.68	2.03	2.53	34.33	16.64	2.11	42.22	9.74
	3×ru	1.66	23.07	2.63	2.27	37.20	16.48	2.03	32.35	10.90
	3×zh	1.51	23.51	2.05	2.33	44.63	11.56	2.26	34.02	10.59
	3×ja	1.50	25.01	1.96	-	-	-	2.25	33.83	9.18
	3×ko	1.56	21.67	2.05	-	-	-	2.14	35.28	9.07
	avg	1.60	22.99	2.06	2.35	38.03	12.20	2.12	35.86	10.04
Multi	en,fr,zh	1.75	40.16	3.98	2.54	54.94	16.74	2.08	53.12	13.78
	fr,zh,ru	1.74	36.94	4.10	2.54	54.64	20.79	2.13	54.41	14.69
	de,fr,ru	1.73	31.85	4.19	2.57	54.82	19.80	2.10	53.35	15.42
	avg	1.81	38.06	3.92	2.56	53.15	16.56	2.17	53.21	15.40

Model Representation

Multilingual caption sets have more variation in how model representation than monolingual caption sets: **diff. lang captions are represented diff.**

Table 4. **Model representation evaluation results.** Mean maximum pairwise cosine distance (caption set width in representation space) of mono- and multi-lingual caption sets. "'Multi → avg" is the mean across *all* triplets of languages.

	Caption Source→	Vertex	LLaVA	XM
Mono	3×en	.19	.22	.38
	3×fr	.18	.29	.42
	3×de	.19	.28	.43
	3×ru	.22	.29	.40
	3×zh	.20	.36	.46
	3×ja	.19	-	.42
	3×ko	.37	-	.43
	avg	.22	.29	.42
Multi	en,fr,zh	.37	.45	.54
	fr,zh,ru	.33	.47	.54
	de,fr,ru	.38	.43	.49
	avg	.49	.47	.52

Model Generation

Models fine-tuned on captions from a particular language perform significantly better on test-set captions from those languages.

Table 5. **Model outputs evaluation results.** 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. 'multi' refers to an even split across all languages. Red: best performance on a split, yellow highlights model fine-tuned on 'multi'.

Fine-tuned on	*Evaluated on* en	de	fr	ru	zh	ja	ko	multi
en	.27	.23	.23	.22	.22	.23	.23	.23
de	.21	.25	.22	.22	.22	.21	.23	.22
fr	.25	.24	.26	.23	.24	.24	.25	.25
ru	.23	.23	.23	.25	.23	.24	.24	.24
zh	.20	.20	.20	.21	.25	.22	.22	.22
ja	.21	.21	.22	.21	.23	.27	.25	.22
ko	.22	.22	.22	.22	.24	.24	.27	.24
multi	.24	.23	.23	.23	.24	.24	.25	.24

Why are these results interesting?

- Multilingual CV datasets encode perceptual differences in captions
- Take an image and ask the same model or API to caption it in different languages: you can expect perceptually different captions
- Take captions written in two languages, translate them into English, and train a model on them: they'll still behave differently

Recap: we see differently; CV assumes we don't

Where we currently are

1. *Psychological studies:* perceptual variation across culture & language
2. *Philosophical reflection:* the object is always *in relation to* the subject
3. *Computer Vision:* let's just record the objects & forget about the subject

RQ: Does perceptual variation across languages and cultures appear in Computer Vision datasets, models, and APIs? **Yes.**

Recap: we see differently; CV assumes we don't

Where we currently are

1. *Psychological studies:* perceptual variation across culture & language
2. *Philosophical reflection:* the object is always *in relation to* the subject
3. *Computer Vision:* let's just record the objects & forget about the subject

RQ: Does perceptual variation across languages and cultures appear in Computer Vision datasets, models, and APIs? **Yes.**

What to do now? Recommendations for CV

- Build datasets that include annotator background information
- Build models and applications that consider the seeing subject
- Multilinguality encodes more information than "just language"!

Thank you!

tl;dr: Perception varies across languages and cultures in Computer Vision datasets, models, and APIs.

paper link

Sebastin Santy

Jena D. Hwang

Amy X. Zhang

Ranjay Krishna

+ thank you to the Mary Gates Endowment for Students!