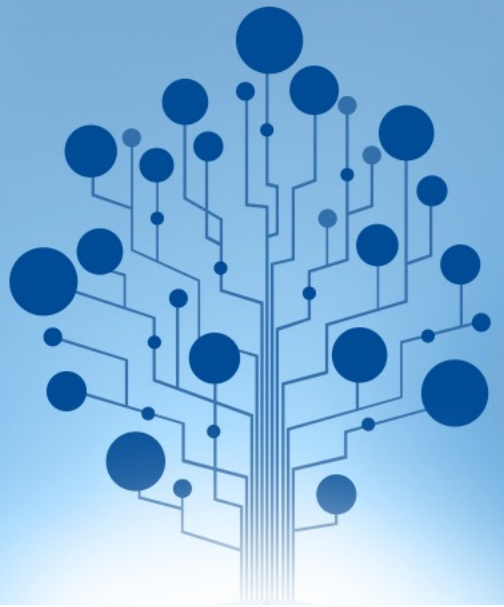
FernUniversität
in Hagen



EViL-Probe

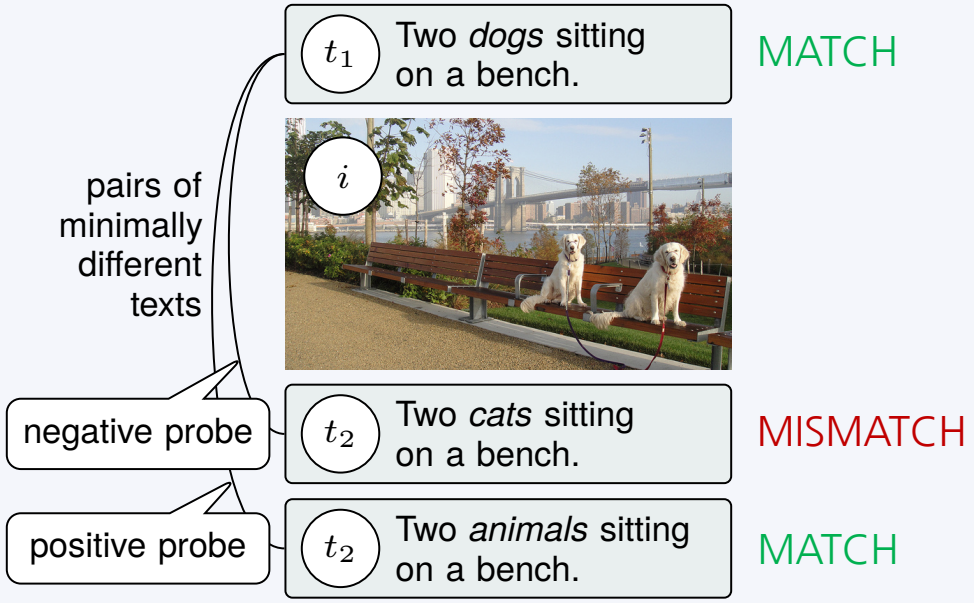
A Composite Benchmark for Extensive Visio-Linguistic Probing

Marie Bexte[1] | **Andrea Horbach**[1,2] | **Torsten Zesch**[1]
[1]FernUniversität in Hagen, Germany | [2]Hildesheim University, Germany

How well can visio-linguistic models align descriptions with photographs ?

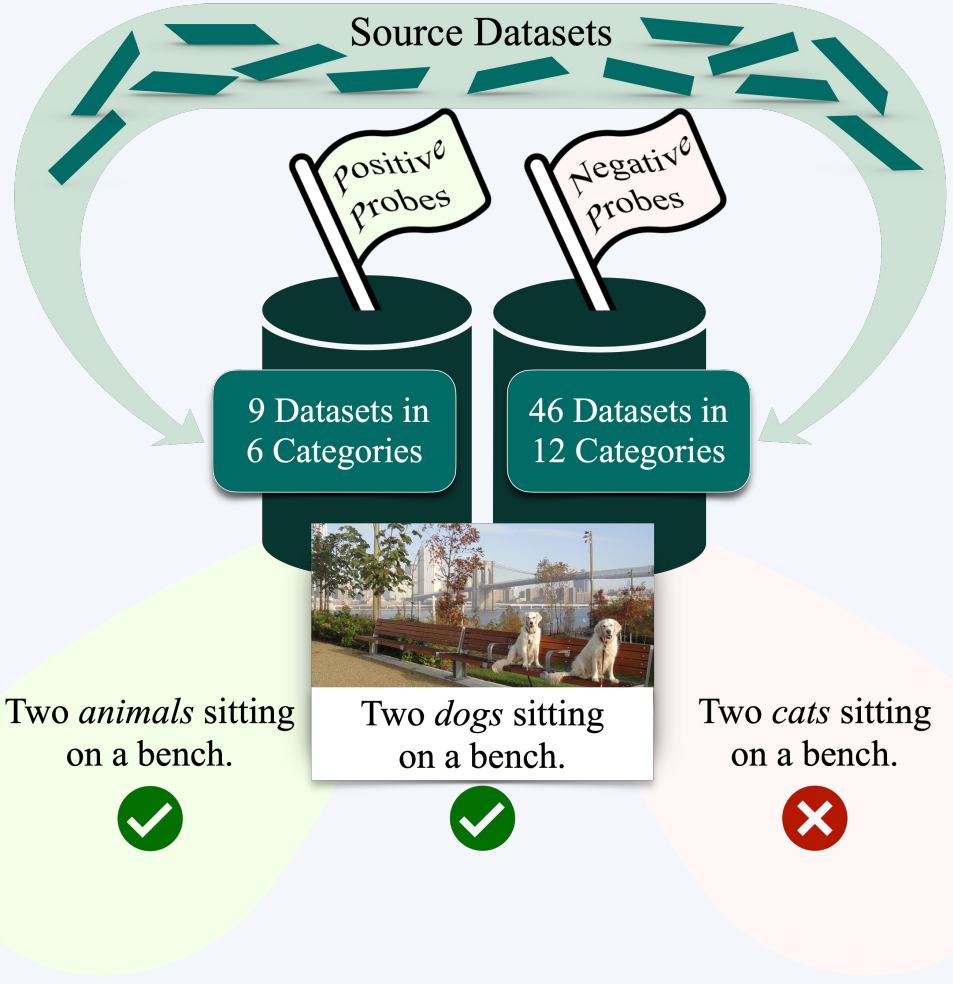
Systematically manipulate target words in image descriptions

Visio-Linguistic Probing



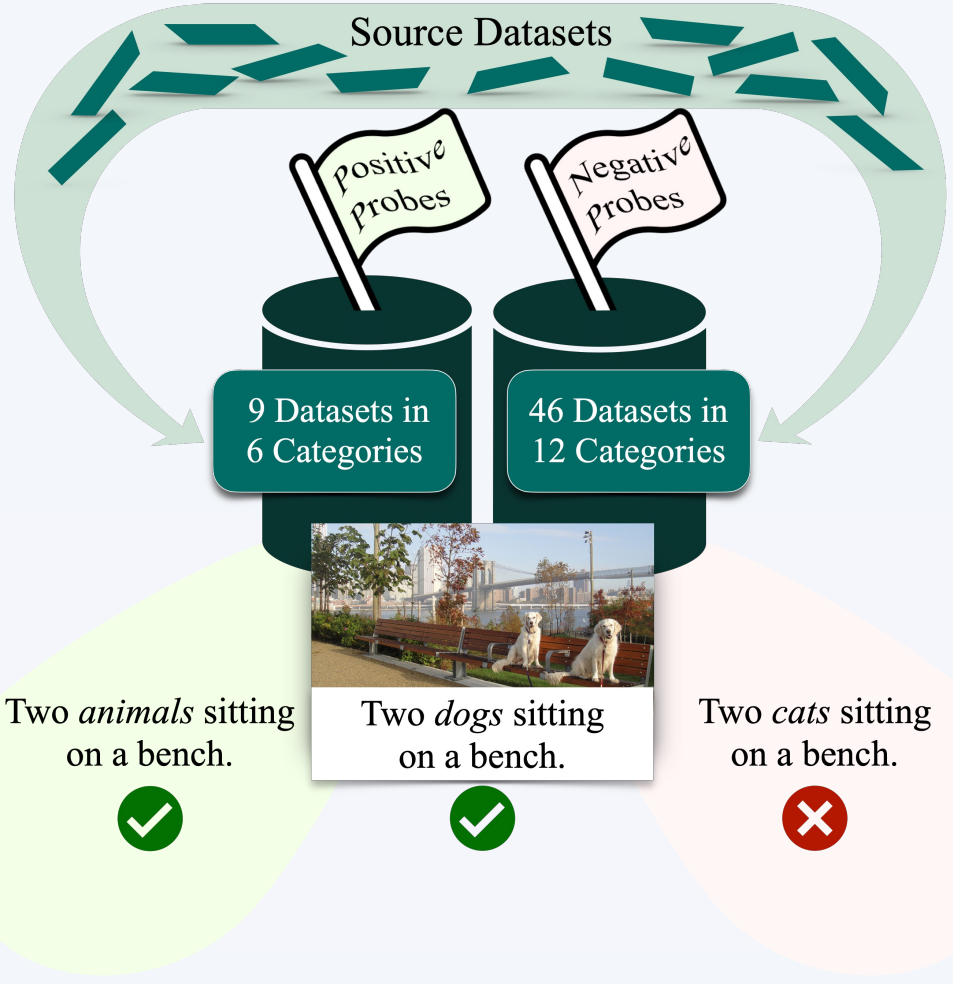
EViL-Probe:
1.5M Examples

EViL-Probe

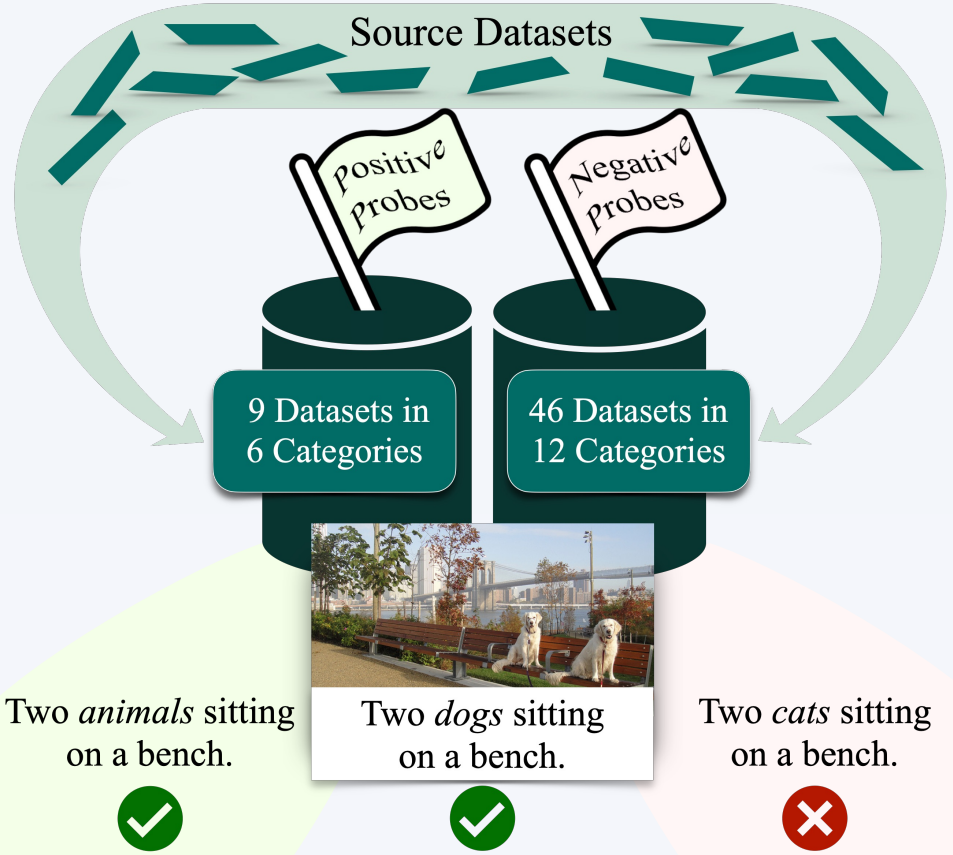

Source Datasets
Positive Probes
Negative Probes
9 Datasets in 6 Categories
46 Datasets in 12 Categories
Two *animals* sitting on a bench.
Two *dogs* sitting on a bench.
Two *cats* sitting on a bench.

Positive Probes

- *Hypernym*
- *Paraphrase*
- *Perspective*
- *Slang*
- *Specificity*
- *Word Order*



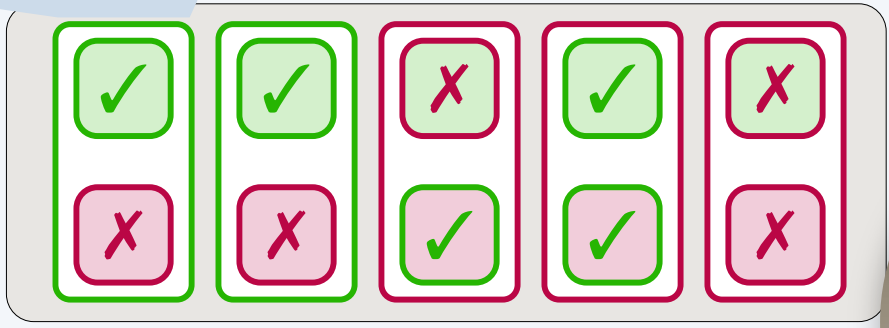
Negative Probes



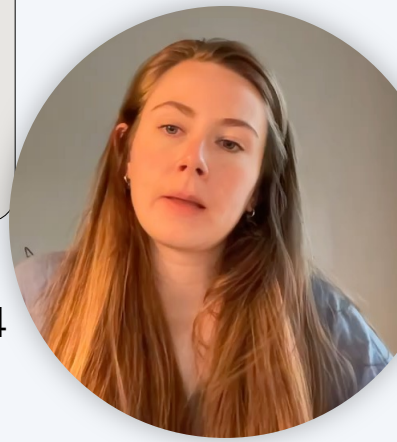
- *Attribute*
- *Color*
- *Image Type*
- *Negation*
- **Noun**
- *Number*
- *Random*
- *Semantic Role*
- *Spatial Relation*
- *Verb*
- *Video-based*
- *Word Order*

Evaluation: Paired Accuracy

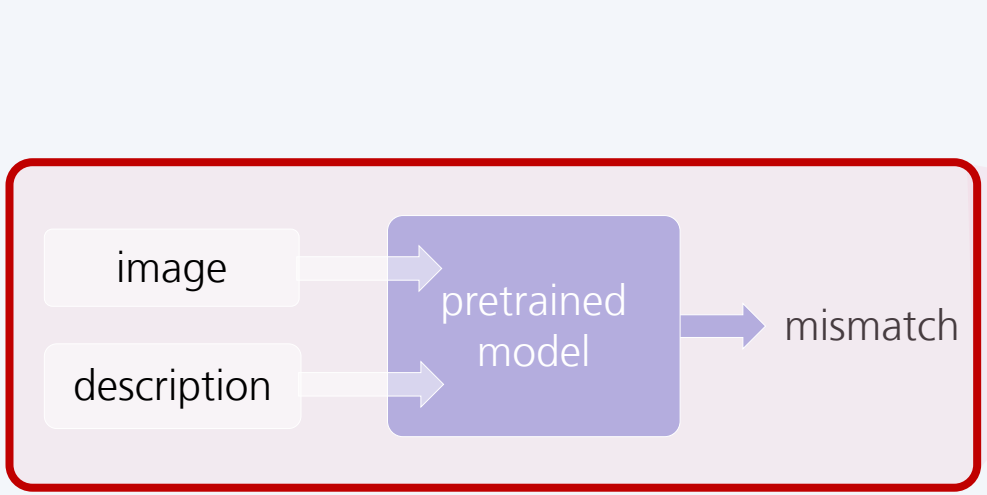
1 probe = 2 descriptions of the same image (1 match, 1 mismatch) processed as 2 separate classifications

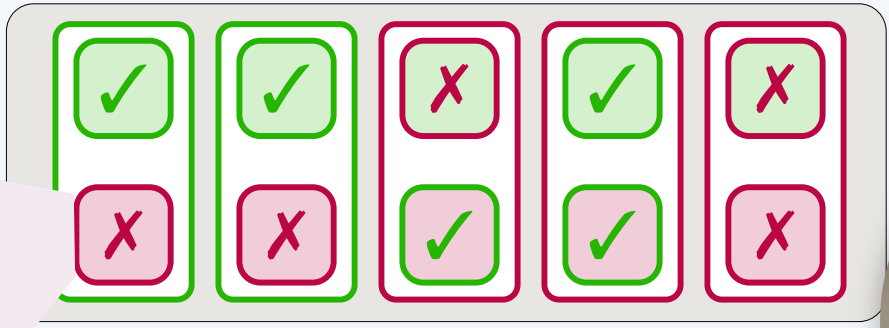


$acc_{paired} = 2/5 = 0.4$

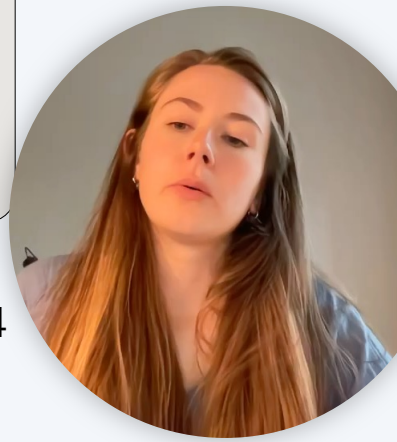


Evaluation: Paired Accuracy



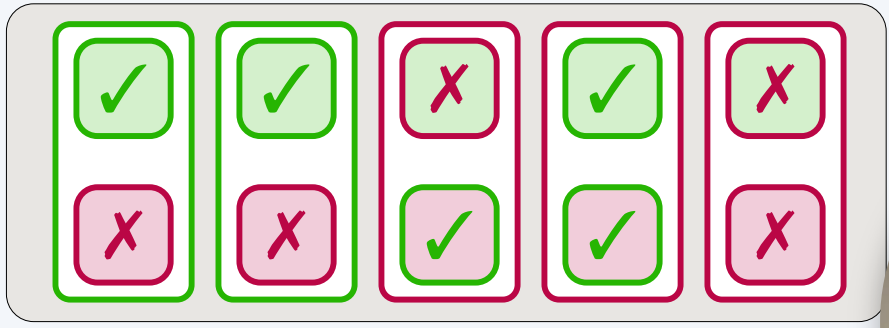


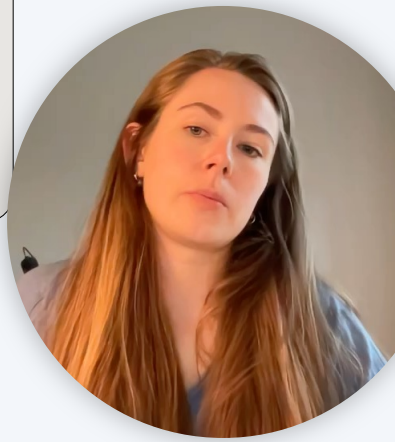
$acc_{paired} = 2/5 = 0.4$



Evaluation: Paired Accuracy

- Not anchored on another text
- Respects text correspondence
- Works with positive probes





Results: Across Categories

	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
Negative Probes							
- Acc	.54	.62	.61	.57	.63	.62	.63
- Acc_{paired}	.27	.26	.24	.20	.29	.27	.29
Positive Probes							
- Acc	.47	.80	.83	.82	.54	.52	.75
- Acc_{paired}	.25	.70	.75	.73	.41	.37	.65

Results: Across Categories

	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
Negative Probes							
- Acc	.54	.62	.61	.57	.63	.62	.63
- Acc_{paired}	.27	.26	.24	.20	.29	.27	.29
Positive Probes							
- Acc	.47	.80	.83	.82	.54	.52	.75
- Acc_{paired}	.25	.70	.75	.73	.41	.37	.65

Results: Across Categories

	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
Negative Probes							
- Acc	.54	.62	.61	.57	.63	.62	.63
- Acc_{paired}	.27	.26	.24	.20	.29	.27	.29
Positive Probes							
- Acc	.47	.80	.83	.82	.54	.52	.75
- Acc_{paired}	.25	.70	.75	.73	.41	.37	.65

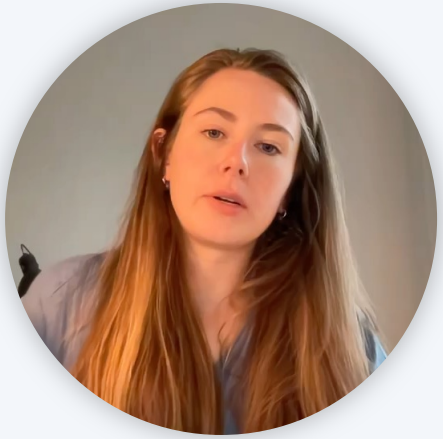
Results: Across Categories

	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
Negative Probes							
- Acc	.54	.62	.61	.57	.63	.62	.63
- Acc_{paired}	.27	.26	.24	.20	.29	.27	.29
Positive Probes							
- Acc	.47	.80	.83	.82	.54	.52	.75
- Acc_{paired}	.25	.70	.75	.73	.41	.37	.65

Results: Across Categories

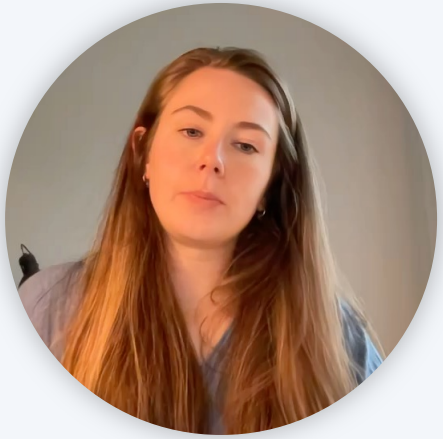
	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
Negative Probes							
- Acc	.54	.62	.61	.57	.63	.62	.63
- Acc_{paired}	.27	.26	.24	.20	.29	.27	.29
Positive Probes							
- Acc	.47	.80	.83	.82	.54	.52	.75
- Acc_{paired}	.25	.70	.75	.73	.41	.37	.65

Results: Individual Categories



	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
	Acc							Acc_{paired}						
Attribute	.52	.61	.60	.55	.62	.60	.62	.26	.25	.23	.15	.27	.24	.26
Color	.55	.66	.65	.58	.68	.66	.72	.30	.34	.32	.23	.38	.35	.46
Image Type	.52	.56	.55	.51	.76	.66	.75	.23	.13	.10	.10	.52	.33	.52
Negation	.51	.53	.54	.52	.50	.51	.51	.23	.09	.09	.09	.03	.05	.08
Noun	.59	.76	.75	.67	.71	.70	.73	.32	.53	.52	.38	.45	.42	.49
Number	.51	.57	.56	.51	.52	.53	.52	.23	.17	.15	.09	.08	.09	.07
Random	.73	.95	.97	.94	.97	.96	.98	.47	.91	.94	.88	.95	.92	.96
Semantic Role	.50	.51	.50	.50	.52	.53	.52	.25	.07	.04	.06	.09	.13	.07
Spatial Relation	.51	.52	.51	.50	.51	.52	.50	.24	.06	.05	.04	.05	.08	.04
Verb	.52	.61	.60	.53	.62	.61	.62	.25	.26	.23	.11	.28	.26	.26
Video-Based	.50	.52	.52	.51	.54	.53	.53	.18	.15	.15	.14	.18	.16	.15
Word Order	.50	.53	.52	.50	.54	.53	.53	.22	.07	.06	.06	.12	.11	.09
Hypernyms	.42	.70	.75	.70	.37	.36	.60	.20	.56	.62	.55	.23	.23	.45
Paraphrase	.46	.78	.82	.80	.45	.44	.71	.24	.73	.76	.72	.38	.37	.65
Perspective	.48	.85	.87	.89	.69	.68	.88	.23	.71	.75	.80	.40	.38	.76
Slang	.42	.69	.71	.70	.30	.28	.62	.19	.52	.55	.57	.16	.14	.47
Specificity	.59	.99	.99	.94	.88	.81	.90	.37	.98	.98	.90	.77	.63	.81
Word Order	.45	.77	.82	.85	.54	.51	.76	.27	.70	.79	.79	.48	.45	.74

Results: Individual Categories

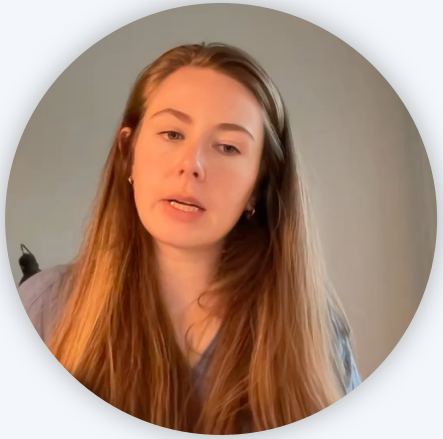


	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
	Acc							Acc_{paired}						
Attribute	.52	.61	.60	.55	.62	.60	.62	.26	.25	.23	.15	.27	.24	.26
Color	.55	.66	.65	.58	.68	.66	.72	.30	.34	.32	.23	.38	.35	.46
Image Type	.52	.56	.55	.51	.76	.66	.75	.23	.13	.10	.10	.52	.33	.52
Negation	.51	.53	.54	.52	.50	.51	.51	.23	.09	.09	.09	.03	.05	.08
Noun	.59	.76	.75	.67	.71	.70	.73	.32	.53	.52	.38	.45	.42	.49
Number	.51	.57	.56	.51	.52	.53	.52	.23	.17	.15	.09	.08	.09	.07
Random	.73	.95	.97	.94	.97	.96	.98	.47	.91	.94	.88	.95	.92	.96
Semantic Role	.50	.51	.50	.50	.52	.53	.52	.25	.07	.04	.06	.09	.13	.07
Spatial Relation	.51	.52	.51	.50	.51	.52	.50	.24	.06	.05	.04	.05	.08	.04
Verb	.52	.61	.60	.53	.62	.61	.62	.25	.26	.23	.11	.28	.26	.26
Video-Based	.50	.52	.52	.51	.54	.53	.53	.18	.15	.15	.14	.18	.16	.15
Word Order	.50	.53	.52	.50	.54	.53	.53	.22	.07	.06	.06	.12	.11	.09
Hypernyms	.42	.70	.75	.70	.37	.36	.60	.20	.56	.62	.55	.23	.23	.45
Paraphrase	.46	.78	.82	.80	.45	.44	.71	.24	.73	.76	.72	.38	.37	.65
Perspective	.48	.85	.87	.89	.69	.68	.88	.23	.71	.75	.80	.40	.38	.76
Slang	.42	.69	.71	.70	.30	.28	.62	.19	.52	.55	.57	.16	.14	.47
Specificity	.59	.99	.99	.94	.88	.81	.90	.37	.98	.98	.90	.77	.63	.81
Word Order	.45	.77	.82	.85	.54	.51	.76	.27	.70	.79	.79	.48	.45	.74

Results: Individual Categories

	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
	Acc							Acc_{paired}						
Attribute	.52	.61	.60	.55	.62	.60	.62	.26	.25	.23	.15	.27	.24	.26
Color	.55	.66	.65	.58	.68	.66	.72	.30	.34	.32	.23	.38	.35	.46
Image Type	.52	.56	.55	.51	.76	.66	.75	.23	.13	.10	.10	.52	.33	.52
Negation	.51	.53	.54	.52	.50	.51	.51	.23	.09	.09	.09	.03	.05	.08
Noun	.59	.76	.75	.67	.71	.70	.73	.32	.53	.52	.38	.45	.42	.49
Number	.51	.57	.56	.51	.52	.53	.52	.23	.17	.15	.09	.08	.09	.07
Random	.73	.95	.97	.94	.97	.96	.98	.47	.91	.94	.88	.95	.92	.96
Semantic Role	.50	.51	.50	.50	.52	.53	.52	.25	.07	.04	.06	.09	.13	.07
Spatial Relation	.51	.52	.51	.50	.51	.52	.50	.24	.06	.05	.04	.05	.08	.04
Verb	.52	.61	.60	.53	.62	.61	.62	.25	.26	.23	.11	.28	.26	.26
Video-Based	.50	.52	.52	.51	.54	.53	.53	.18	.15	.15	.14	.18	.16	.15
Word Order	.50	.53	.52	.50	.54	.53	.53	.22	.07	.06	.06	.12	.11	.09
Hypernyms	.42	.70	.75	.70	.37	.36	.60	.20	.56	.62	.55	.23	.23	.45
Paraphrase	.46	.78	.82	.80	.45	.44	.71	.24	.73	.76	.72	.38	.37	.65
Perspective	.48	.85	.87	.89	.69	.68	.88	.23	.71	.75	.80	.40	.38	.76
Slang	.42	.69	.71	.70	.30	.28	.62	.19	.52	.55	.57	.16	.14	.47
Specificity	.59	.99	.99	.94	.88	.81	.90	.37	.98	.98	.90	.77	.63	.81
Word Order	.45	.77	.82	.85	.54	.51	.76	.27	.70	.79	.79	.48	.45	.74

Results: Individual Categories



	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP	LXMERT	UNITER	VILLA	SOHO	ALBEF	TCL	BLIP
	Acc							Acc_{paired}						
Attribute	.52	.61	.60	.55	.62	.60	.62	.26	.25	.23	.15	.27	.24	.26
Color	.55	.66	.65	.58	.68	.66	.72	.30	.34	.32	.23	.38	.35	.46
Image Type	.52	.56	.55	.51	.76	.66	.75	.23	.13	.10	.10	.52	.33	.52
Negation	.51	.53	.54	.52	.50	.51	.51	.23	.09	.09	.09	.03	.05	.08
Noun	.59	.76	.75	.67	.71	.70	.73	.32	.53	.52	.38	.45	.42	.49
Number	.51	.57	.56	.51	.52	.53	.52	.23	.17	.15	.09	.08	.09	.07
Random	.73	.95	.97	.94	.97	.96	.98	.47	.91	.94	.88	.95	.92	.96
Semantic Role	.50	.51	.50	.50	.52	.53	.52	.25	.07	.04	.06	.09	.13	.07
Spatial Relation	.51	.52	.51	.50	.51	.52	.50	.24	.06	.05	.04	.05	.08	.04
Verb	.52	.61	.60	.53	.62	.61	.62	.25	.26	.23	.11	.28	.26	.26
Video-Based	.50	.52	.52	.51	.54	.53	.53	.18	.15	.15	.14	.18	.16	.15
Word Order	.50	.53	.52	.50	.54	.53	.53	.22	.07	.06	.06	.12	.11	.09
Hypernyms	.42	.70	.75	.70	.37	.36	.60	.20	.56	.62	.55	.23	.23	.45
Paraphrase	.46	.78	.82	.80	.45	.44	.71	.24	.73	.76	.72	.38	.37	.65
Perspective	.48	.85	.87	.89	.69	.68	.88	.23	.71	.75	.80	.40	.38	.76
Slang	.42	.69	.71	.70	.30	.28	.62	.19	.52	.55	.57	.16	.14	.47
Specificity	.59	.99	.99	.94	.88	.81	.90	.37	.98	.98	.90	.77	.63	.81
Word Order	.45	.77	.82	.85	.54	.51	.76	.27	.70	.79	.79	.48	.45	.74

Data:

Conclusions:

- Benchmark is challenging for all tested models
- Positive probes help disambiguate performance