Multimodal CL and NLP: Combining Language and Vision for (Computational) Semantics

Carina Silberer

IMS, University of Stuttgart

CL Fall School 2024, Passau
16 - 20 September 2024

Outline

Analysis of VL-Models

- To what extent do they model **linguistic** phenomena?
- To what extent do they possess a fine-grained **visual–linguistic** understanding?
- Phenomena:
  - verb meaning; situation awareness (actions and actors) e.g., "selling" vs. "buying"
  - spatial relationships e.g., "below" vs. "on"
  - composition e.g., "fire truck" vs. "truck fire"
  - action–effects
  - counting
  - negation

Analysis of VL-Models

How many legs does a table have?

Analysis of VL-Models

How many legs does the table have?

Analysis of VL-Models

How many legs does the table have?

https://en.tvar-kt.cz

Analysis Schemes

Test or probe models in a zero-shot setting, to analyse their inherent capabilities obtained during pretraining.

- ***Foiling*** Create foils: image–caption pairs where the target phenomenon (e.g., spatial relations) deems the pair incorrect (e.g., wrong preposition)
  Methods:
  1. Linguistic: A caption is turned into a counterfactual (foil) by minimal edits (cf. minimal pair in linguistics)
  2. Visual: An image is replaced by another, incorrect image that does not match the caption
- ***Probing***, e.g., through classification [Salin et al., 2022] or Visual Question Answering
  $\rightarrow$ Train a very simple classifier on top of representations
- **Guided Masking** [Beňová et al., 2024]

Vision-Language Models: Capabilities

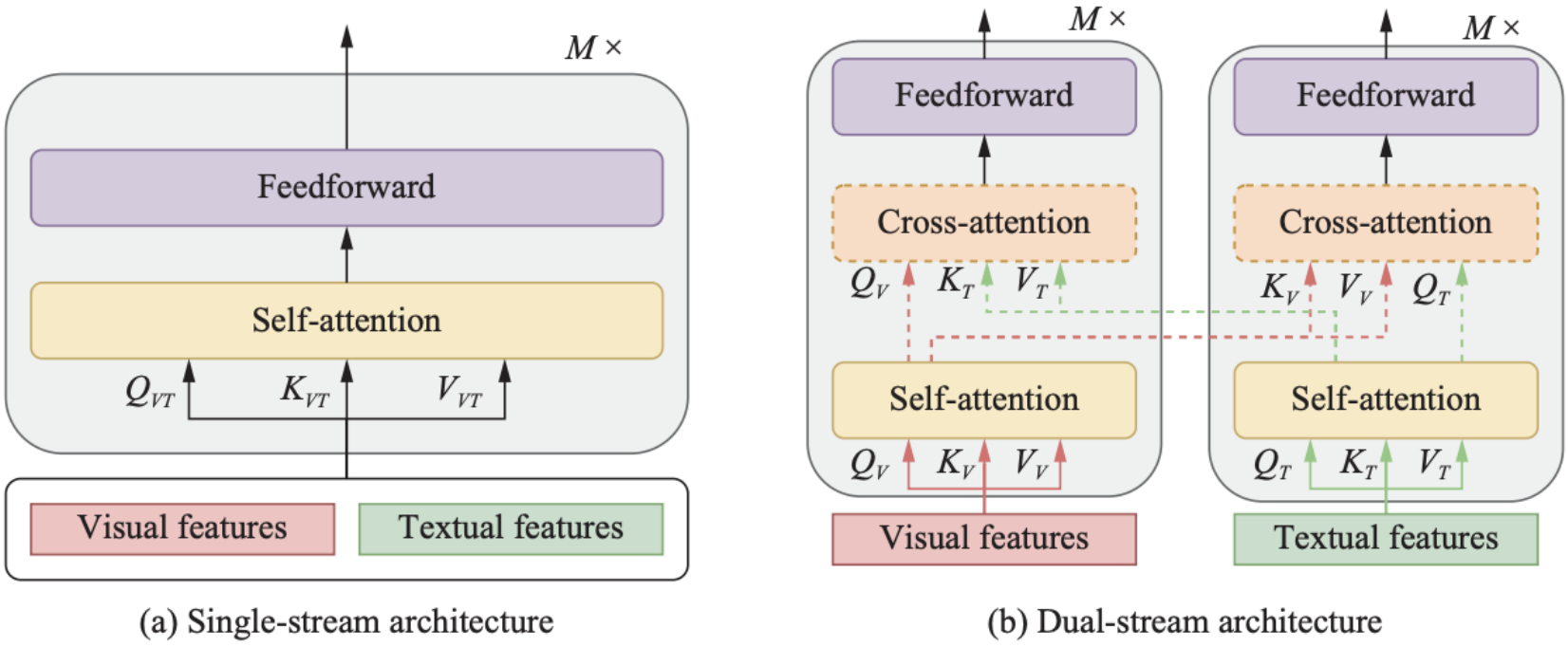
Aspects

- **Architecture:** How strong do they *fuse* the two modalities?
- **Modality:** What is the relative *contribution*/importance of the two modalities?
  Discussion: Dependence on task
- **Linguistic abilities:** Which linguistic phenomena do they (not) learn/encode?

Outline

Recap: Vision–Language Models

BERT-like



(a) Single-stream architecture

(b) Dual-stream architecture

source: [Chen et al., 2022]

Investigating Negation in Pre-trained Vision-and-language Models

[Dobreva and Keller, 2021]

Investigating Negation in Pre-trained Vision-and-language Models

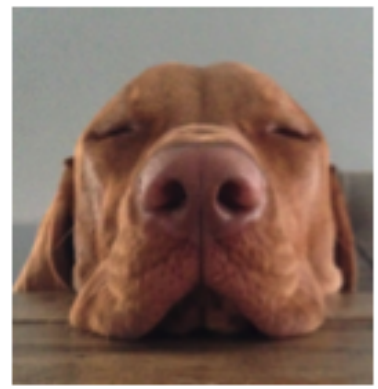
[Dobreva and Keller, 2021]

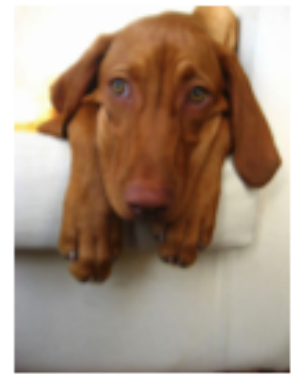
Negation: Dataset

Original: Every dog is wearing a collar. → False

Negated: Not every dog is wearing a collar. → True

(c) NP (nonexistential)





Original: A dog is resting its head on something. → True

Negated: No dog is resting its head on something. → False

(d) NP (existential)

- Creation of incorrect caption (foil) for each image–caption pair
- Categories: verbal and non-verbal negation
  "did not see anything" and "saw nothing"

Investigating Negation in Pre-trained Vision-and-language Models

[Dobreva and Keller, 2021]

Negation: Dataset

Original: Four or fewer television screens are visible. → True

Negated: No television screens are visible. → False

(e) NP (number-to-none)

Original: Three or fewer goats are visible. → False

Negated: It is not true that three or fewer goats are visible. → True

(f) Sentence-wide

- Creation of incorrect caption (foil) for each image–caption pair
- Categories: verbal and non-verbal negation
  "did not see anything" and "saw nothing"

Investigating Negation in Pre-trained Vision-and-language Models

[Dobreva and Keller, 2021]

Negation: Experiments

- Task: Given two images, is the caption true or false?
- Models: LXMERT and UNITER (dual- and single-stream, resp.)
- UNITER: two variants for combining image embeddings
- Fine-tuning of models on NLVR2 task

Investigating Negation in Pre-trained Vision-and-language Models

[Dobreva and Keller, 2021]

Negation: Experiments – Results

	LXMERT		$\text{UNITER}_{paired-attn}$		$\text{UNITER}_{triplet}$	
	negative	positive	negative	positive	negative	positive
Verbal (content)	28.72	69.23	43.62	73.63	43.62	71.43
Verbal (existential)	30.56	82.41	50.0	77.77	44.44	66.66
NP (nonexistential)	44.83	67.86	48.28	64.29	55.17	50.0
NP (existential)	34.55	80.0	50.91	85.45	32.73	87.27
NP (number-to-none)	54.17	72.22	51.39	77.77	55.56	76.39
Sentence-wide	38.55	66.27	31.33	69.87	38.55	65.06
Overall	36.96	73.5	45.35	76.5	44.22	71.5

Table 3: Accuracy on the negation test set and the corresponding non-negated (positive) examples.

- Models: LXMERT and UNITER (dual- and single-stream)
- $\text{UNITER}_{\text{paired-attn}}$: attention layer on individual embeddings
- Fine-tuning of models on NLVR2 task

Investigating Negation in Pre-trained Vision-and-language Models

[Dobreva and Keller, 2021]

Negation: Experiments – Results

	LXMERT		$\text{UNITER}_{paired-attn}$		$\text{UNITER}_{triplet}$	
	o. correct	o. incorrect	o. correct	o. incorrect	o. correct	o. incorrect
Verbal (content)	15.38	58.62	40.58	52.0	30.3	75.0
Verbal (existential)	21.35	73.68	46.43	62.5	29.17	75.0
NP (nonexistential)	30.0	77.78	42.11	60.0	40.0	71.43
NP (existential)	36.36	27.27	55.32	25.0	27.08	71.43
NP (number-to-none)	46.15	75.0	51.79	50.0	47.27	82.35
Sentence-wide	25.45	64.29	8.62	84.0	9.26	93.1
Overall	27.38	63.79	40.54	60.19	29.35	79.39

Table 4: Accuracy for the negated examples for whose original (unnegated) version the model makes a correct/incorrect prediction ("o. correct"/"o. incorrect").

- Models: LXMERT and UNITER (dual- and single-stream)
- $\text{UNITER}_{\text{paired-attn}}$: attention layer on individual embeddings
- Fine-tuning of models on NLVR2 task

Probing Image–Language Transformers for Verb Understanding

[Hendricks and Nematzadeh, 2021]

Focus

Verb comprehension

- Subject–Verb–Object probes

Probing Image–Language Transformers for Verb Understanding

[Hendricks and Nematzadeh, 2021]

SVO-Probing

- ***Do semantic models understand verbs?***
- **Task:** Do image-sentence pairs of SVO-Probes match?
  $\Rightarrow$ Model requires verb, not only nouns, to make decision
- **Findings:**
  – Models worst on verbs, then subjects, then objects
  – Models cannot handle mismatching image-sentence pairs
  $\rightarrow$ Cannot reason multimodally robustly

Probing Image–Language Transformers for Verb Understanding

[Hendricks and Nematzadeh, 2021]

SVO-Probing: Dataset

Do semantic models understand verbs?

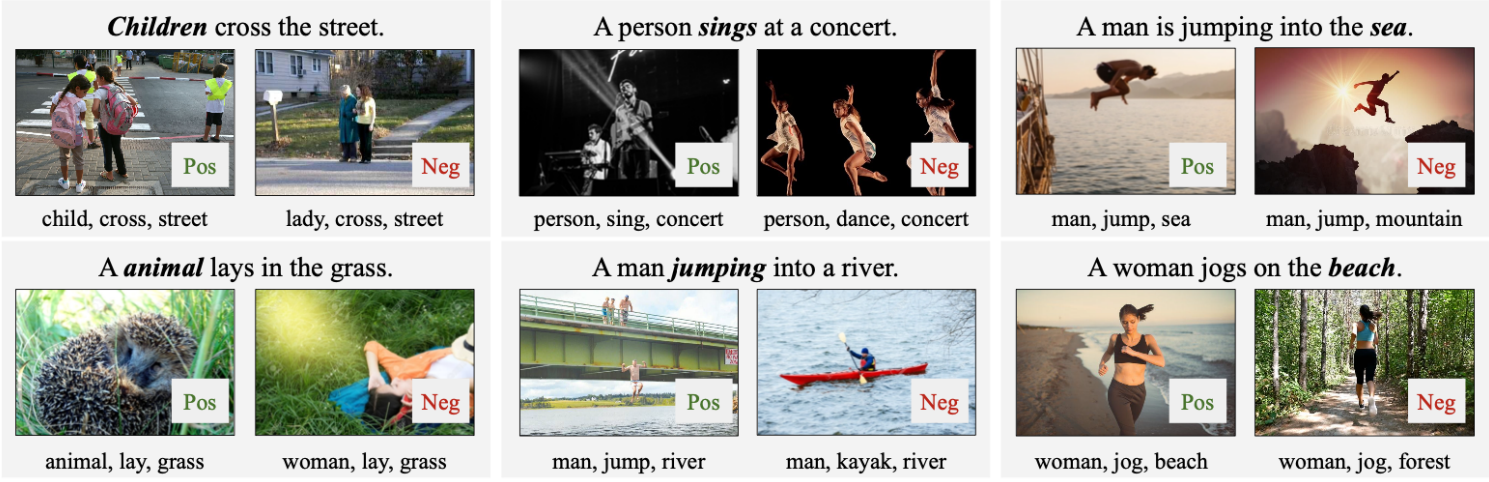


Figure 1: Examples from SVO-Probes. Images on the left and right show positive and negative image examples for each sentence. Below each image is the ⟨subject, verb, object⟩ triplet corresponding to the image.

source: `https://www.deepmind.com/blog/probing-image-language-transformers-for-verb-understanding`

Probing Image–Language Transformers for Verb Understanding

[Hendricks and Nematzadeh, 2021]

SVO-Probing: Dataset

Do semantic models understand verbs?

Figure 1: Examples from SVO-Probes. Images on the left and right show positive and negative image examples for each sentence. Below each image is the ⟨subject, verb, object⟩ triplet corresponding to the image.

source: `https://www.deepmind.com/blog/probing-image-language-transformers-for-verb-understanding`

- SVO-Triples: from Conceptual Captions, with frequent verbs, subjects, objects (>50 times)
- Images: from Google Image Search API
- Sentences: Collected through AMT – write sentence given positive and negative SVO-Triplets

Probing Image–Language Transformers for Verb Understanding

[Hendricks and Nematzadeh, 2021]

SVO-Probing: Dataset

Do semantic models understand verbs?

source: `https://www.deepmind.com/blog/probing-image-language-transformers-for-verb-understanding`

- Sentences: Collected through AMT – write sentence given positive and negative SVO-Triplets

Probing Image–Language Transformers for Verb Understanding

[Hendricks and Nematzadeh, 2021]

SVO-Probing: Experiments

- MMT model: (MultiModal Transformer) *only cross-modal* attention; trained on MRM, MLM, ITM (masked region/language modeling, img-text matching)

Name	Multimodal Attention	Similar Model	MLM	MRM	*ZS Flickr*
MMT	Queries from L (I) take values and keys from *only* I (L)	ViLBERT; LXMERT	✓	✓	**41.9**
Merged–MMT	Queries from L (I) take values and keys from *both* L and I	UNITER	✓	✓	**40.0**
Lang–MMT	Queries are *only* from L (Hendricks et al., 2021)		✓	✓	33.6
Image–MMT	Queries are *only* from I (Hendricks et al., 2021)		✓	✓	31.6
SMT	Single-Modality Transformers without multimodal attention		✓	✓	16.9
No-MRM–MMT	The same as MMT		✓	✗	**41.1**
No-MLM–MMT	The same as MMT		✗	✓	20.2

Table 2: Different variants of the image–language transformer architecture we test. L and I stand for language and image, respectively. We note that models with Merged attention (like UNITER) are also referred to as single-stream models. ViLBERT: Lu et al. (2019); LXMERT: Tan and Bansal (2019); UNITER: Chen et al. (2020)

Probing Image–Language Transformers for Verb Understanding

[Hendricks and Nematzadeh, 2021]

	Overall			Subj. Negative			Verb Negative			Obj. Negative		
	Avg	Pos.	Neg.	Avg	Pos.	Neg.	Avg	Pos.	Neg.	Avg	Pos.	Neg.
# Examples	48k	12k	36k	8k	3k	5k	34k	11k	23k	11k	3k	8k
MMT	64.3	*93.8*	34.8	67.0	*94.4*	39.5	60.8	*93.8*	27.8	73.4	*94.4*	52.4
Merged–MMT	64.7	**94.4**	35.0	69.1	**94.9**	43.2	60.7	**94.4**	27.0	74.1	**94.9**	53.3
Lang–MMT	*68.1*	80.2	**56.0**	*71.5*	82.1	**60.9**	*64.5*	80.2	*48.9*	*77.7*	81.4	**74.1**
Image–MMT	64.3	91.6	37.0	68.2	92.1	44.2	59.7	91.6	27.8	75.6	91.5	59.6
SMT	52.4	49.1	*55.6*	52.6	47.7	57.5	51.8	49.1	**54.6**	53.9	50.7	57.0
No-MRM–MMT	**69.5**	85.4	53.7	**73.5**	87.4	*59.7*	**65.5**	85.6	45.5	**80.1**	86.2	**74.1**
No-MLM–MMT	60.8	92.3	29.3	64.8	93.9	35.8	57.4	92.5	22.4	69.5	93.6	45.5

Table 4: Results on SVO-Probes on different models for subject, verb, and object negatives. Best results are shown in bold; second best results are italicized.

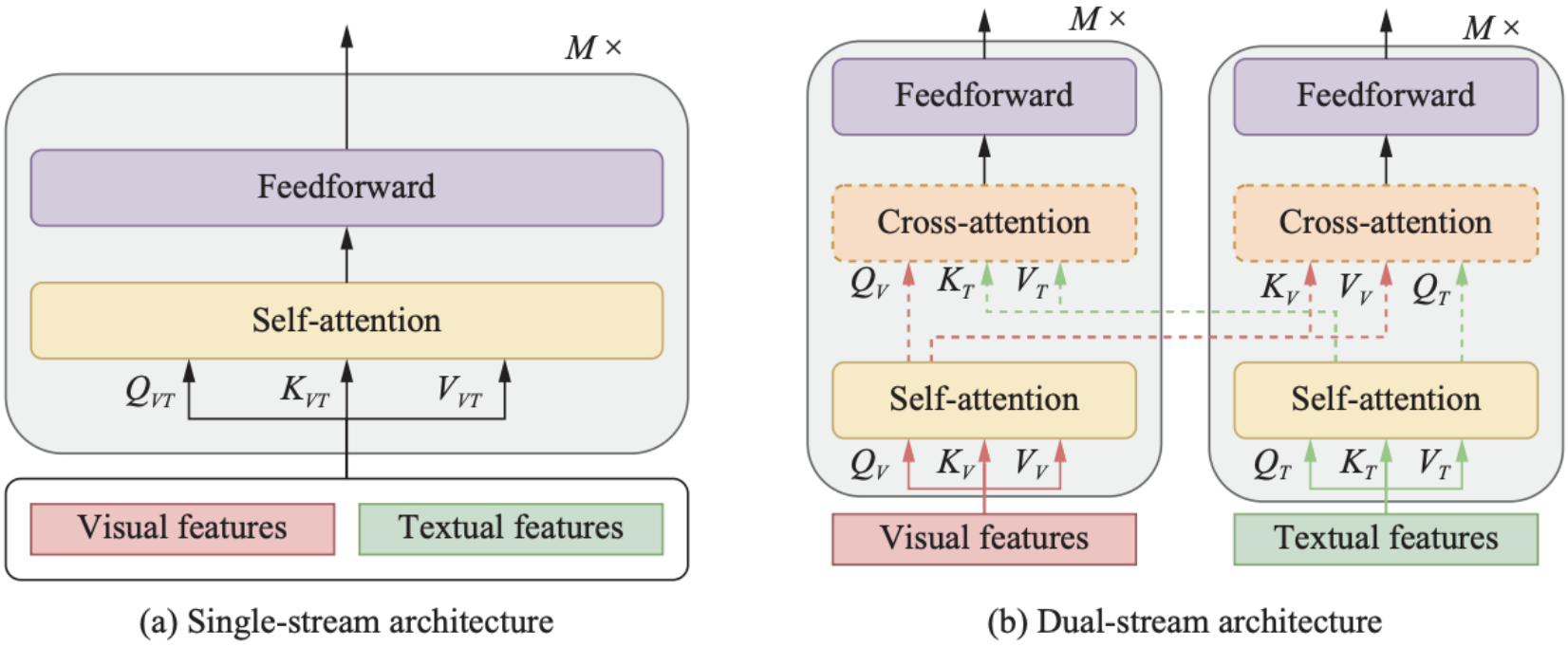
SVO-Probing: Results

- Models cannot handle mismatching image-sentence pairs
  Not dependent on # examples in training data
  Noise in dataset affects model more than domain-mismatch
- Higher accuracy on subject seems to be an artifact
  $\leftarrow$ noun similarity of subjects of pos vs. neg triplets higher (.49) than that of verbs and objects (.29 and .27, resp.)

Outline

Recap: Vision–Language Models

BERT-like



(a) Single-stream architecture

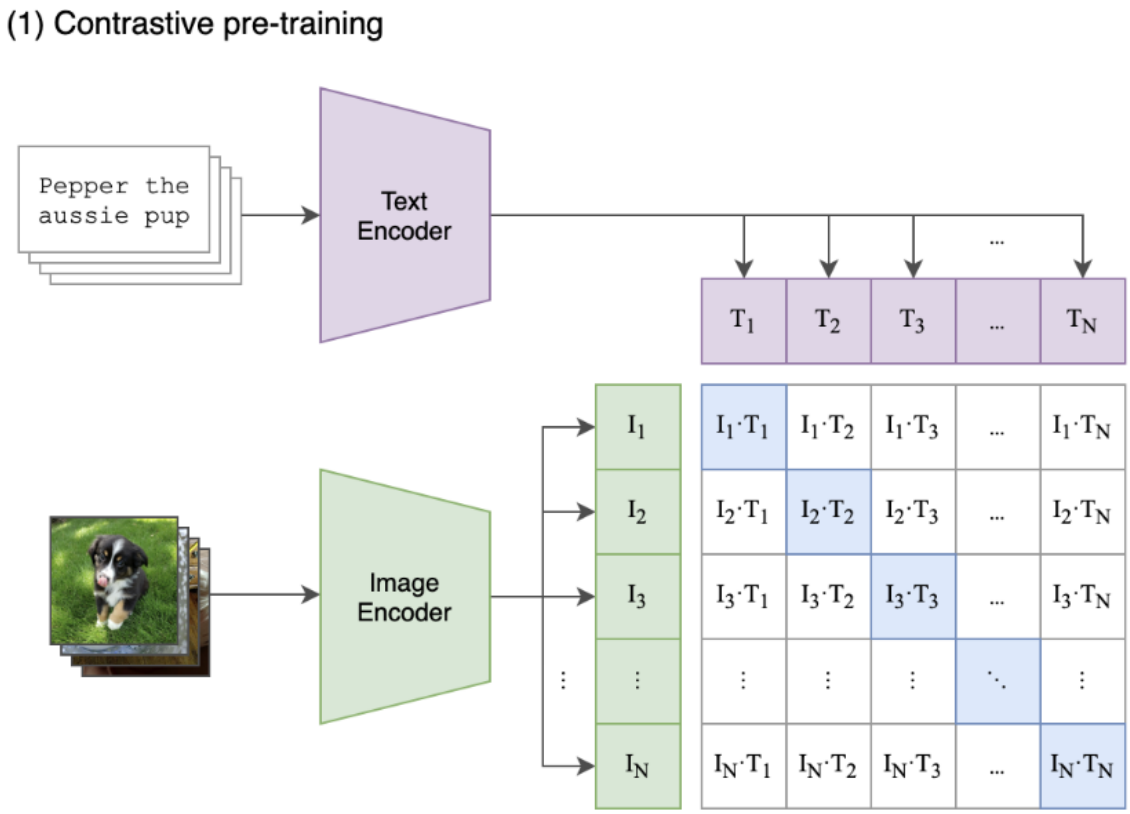
(b) Dual-stream architecture

source: [Chen et al., 2022]

Recap: Vision–Language Models

CLIP

[Radford et al., 2021]



VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena

[Parcalabescu et al., 2022]

Focus

In how far do models learn to ground linguistic phenomena during pretraining?

- Existence
- Plurality
- Counting
- Relations
- Actions
- Coreference
- Uses BERT-like and CLIP

VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena

[Parcalabescu et al., 2022]

- VALSE benchmark: Task-independent, zero-shot benchmark
- Test visual grounding capabilities through linguistic constructs
- In how far do models learn to ground linguistic phenomena during pretraining (or fine-tuning)?
- Test items: Captions and edited foils

VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena

[Parcalabescu et al., 2022]

- VALSE benchmark: Task-independent, zero-shot benchmark
- Test visual grounding capabilities through linguistic constructs
- In how far do models learn to ground linguistic phenomena during pretraining (or fine-tuning)?
- Test items: Captions and edited foils
  - Is image-sentence pair foiled? or
  - Which image-sentence pair is the correct / foiled one?

VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena

[Parcalabescu et al., 2022]

- VALSE benchmark: Task-independent, zero-shot benchmark
- Test visual grounding capabilities through linguistic constructs
- In how far do models learn to ground linguistic phenomena during pretraining (or fine-tuning)?
- Test items: Captions and edited foils
  - Is image-sentence pair foiled? or
  - Which image-sentence pair is the correct / foiled one?
- Comparison Models:
  - LM: GPT1 and GPT2
  - BERT-like VL models (BERT_{VL}): LXMERT, ViLBERT, 12-in-1 (dual-stream); VisualBERT (single-stream)
  - CLIP

VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena

	pieces	**existence**	**plurality**	**counting**	**relations**	**actions**	**coreference**
Data collection & metadata	instruments	*existential quantifiers*	*semantic number*	*balanced, adversarial, small numbers*	*prepositions*	*replacement, actant swap*	*standard, clean*
	#examples†	505	851	2,459	535	1,633	812
	foil generation method	*nothing* ↔ *something*	NP replacement (`sg2pl`; `pl2sg`) & quantifier insertion	numeral replacement	SpanBERT prediction	action replacement, actant swap	*yes* ↔ *no*
	MLM	✗	✗	✗	✓	✓	✗
	GRUEN	✗	✓	✗	✓	✗	✗
	NLI	✗	✓	✗	✓	✗	✗
	src. dataset	Visual7W	MSCOCO	Visual7W	MSCOCO	SWiG	VisDial v1.0
	image src.	MSCOCO	MSCOCO	MSCOCO	MSCOCO	SituNet	MSCOCO
Example data	caption (blue) / foil (orange)	*There are no animals / animals shown.*	*A small copper vase with some flowers / exactly one flower in it.*	*There are four / six zebras.*	*A cat plays with a pocket knife on / underneath a table.*	*A man / woman shouts at a woman / man.*	*Buffalos walk along grass. Are they in a zoo? No / Yes.*
	image						

Table 1: Overview of pieces and instruments in VALSE, with number of examples per piece; the foil generation method used; whether masked language modelling (MLM), GRUEN, and NLI filtering are used; dataset and image sources; and image-caption-foil examples. †The number of examples is the sum of the examples available for each instrument in the piece. In Table 5 (in the Appendix) we list the number of examples in each individual instrument.

VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena

[Parcalabescu et al., 2022]

Metric	Model	Existence	Plurality	Counting			Sp.rel.‡	Action		Coreference		Foil-it!	Avg.
		quantifiers	number	balanced	sns.†	adv.†	relations	repl.†	actant swap	standard	clean		
	Random	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0	50.0
acc_r	GPT1*	61.8	53.1	51.2	48.7	69.5	**77.2**	65.4	72.2	45.6	45.2	77.5	60.7
	GPT2*	58.0	51.9	51.6	49.8	45.3	75.0	66.8	**76.9**	54.5	50.0	80.7	60.1
	CLIP	66.9	56.2	62.1	62.5	57.5	64.3	**75.6**	68.6	52.1	49.7	**88.8**	64.0
	LXMERT	78.6	64.4	62.2	69.2	42.6	60.2	54.8	45.8	46.8	44.2	87.1	59.6
	ViLBERT	65.5	61.2	58.6	62.9	73.7	57.2	70.7	68.3	47.2	48.1	86.9	63.7
	12-in-1	**95.6**	**72.4**	**76.7**	**80.2**	**77.3**	67.7	65.9	58.9	**75.7**	**69.2**	86.9	**75.1**
	VisualBERT	39.7	45.7	48.2	48.2	50.0	39.7	49.2	44.4	49.5	47.6	48.5	46.4
acc	LXMERT	55.8	55.1	52.0	55.4	49.9	50.8	51.1	48.5	49.8	49.0	70.8	53.5
	ViLBERT	2.4	50.3	50.7	50.6	51.8	49.9	52.6	50.4	50.0	50.0	55.9	51.3
	12-in-1	**89.0**	**62.0**	**64.9**	**69.2**	**66.7**	**53.4**	**57.3**	**52.2**	**54.4**	**54.3**	**71.5**	**63.2**
	VisualBERT	49.3	46.5	48.3	47.8	50.0	49.3	48.8	49.7	50.0	50.0	46.6	48.8
$\min(p_c, p_f)$	LXMERT	41.6	**42.2**	50.9	50.0	37.3	**28.4**	35.8	36.8	**18.4**	**17.3**	**69.3**	38.9
	ViLBERT	47.9	2.1	24.4	24.7	17.5	1.5	11.9	7.1	1.3	1.9	12.9	13.9
	12-in-1	**85.0**	33.4	**64.3**	**61.7**	**59.5**	13.3	**47.8**	**37.6**	15.8	13.5	48.8	**43.7**
	VisualBERT	1.3	0.3	0.0	0.0	0.0	1.3	0.0	0.0	0.0	0.0	0.2	0.3
$AUROC$ $\times 100$	LXMERT	60.5	57.3	53.8	57.7	50.5	51.9	52.1	47.6	49.8	49.5	76.9	55.2
	ViLBERT	52.5	54.1	50.8	51.6	53.5	51.2	57.2	**57.8**	49.9	49.9	75.2	54.9
	12-in-1	**96.3**	**67.4**	**72.0**	**77.8**	**75.1**	**55.8**	**61.3**	55.0	**59.8**	**59.6**	**81.0**	**69.2**
	VisualBERT	28.9	29.0	24.5	16.5	20.9	45.2	17.7	36.3	45.3	46.3	28.5	30.8

Table 2: Performance of unimodal and multimodal models on the VALSE benchmark according to different metrics. We bold-face the best overall result per metric, and underscore all results below (or at) the random baseline. acc_r is a pairwise ranking accuracy where a prediction is considered correct if $p(caption, img) > p(foil, img)$. Precision p_c and foil precision p_f are *competing* metrics where naïvely increasing one can decrease the other: therefore *looking at the smaller number among the two gives a good intuition of how informed is a model prediction.* †**sns.** Counting small numbers. **adv.** Counting adversarial. **repl.** Action replacement. ‡ **Sp.rel.** Spatial relations. *Unimodal text-only models that do not use images as input. CLIP is only tested in pairwise ranking mode (fn. 6).

VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena

[Parcalabescu et al., 2022]

- VALSE benchmark: test visual grounding capabilities through linguistic constructs
- Foiling tasks
- V&L models are **good**: object identification + existence in images
- V&L models **fail**: ground their interdependence and relationships in visual scenes when forced to respect linguistic indicator

VALSE

[Parcalabescu et al., 2022]

- VL-models > LM: Object existence, plurality, counting

There is a truck pictured.

There is no truck pictured.

Exactly one row of motorcycles parked together on a grass yard area with a house in the background.

A number of rows of motorcycles parked together on a grass yard area with a house in the background.

There are exactly 8 horses.

There are exactly 5 horses.

VALSE Test: Spatial Relations

[Parcalabescu et al., 2022]

LM > VL-models

	A cow stands on a sidewalk outside a building.	A cow stands on a sidewalk in a building.
	Three giraffes banding down to drink water with trees in the background.	Three giraffes banding up to drink water with trees in the background.

VALSE Test: Actions I

- LM > VL-models: Actant swap (roles): [Parcalabescu et al., 2022]

The people unveil the prize. A prize unveils people.

A baby drools over clothing. A clothing drools over the baby.

VALSE: BERT$_{VL}$ models vs. CLIP

[Parcalabescu et al., 2022]

- Dual-stream BERT$_{VL}$ > single-stream BERT$_{VL}$:
  - single-stream VisualBERT fails across all tasks: never better than baseline
  - dual-stream 12-in-1 (aka Multi-task ViLBERT) overall best, followed by LXMERT
- CLIP > BERT$_{VL}$
  - actions (action replacement)
- BERT$_{VL}$ > CLIP
  - existence, plurality, counting (object existence and number in images)
  - coreference (object identification)
  - (spatial relations (prepositions))

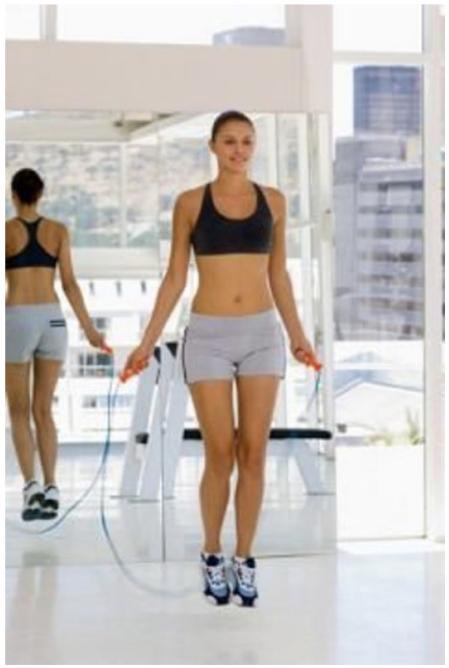
VALSE Test: Actions II

- CLIP $>$ BERT$_{VL}$
- Action replacement:

[Parcalabescu et al., 2022]

A figure climbs the stairs. A figure descends the stairs.

actions



A woman skips a jump rope. A woman releases a jump rope.

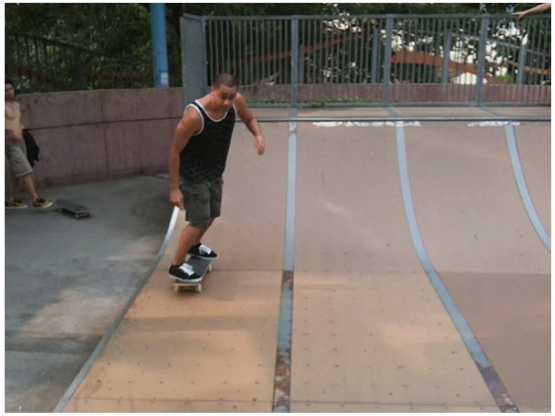
VALSE Test: Coreference

- $\text{BERT}_{VL} > \text{CLIP}$

[Parcalabescu et al., 2022]

A close up of a hot dog with onions. Is it a big hot dog? Yes.

A close up of a hot dog with onions. Is it a big hot dog? No.



A skateboarding man is on a half pipe. Does he wear a helmet? No.

A skateboarding man is on a half pipe. Does he wear a helmet? Yes.

Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality

[Thrush et al., 2022] https://huggingface.co/datasets/facebook/winoground

Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality

[Thrush et al., 2022] https://huggingface.co/datasets/facebook/winoground

Compositional Reasoning

- Observation: Textual transformers often insensitive to word order
- Test: Given two images and two captions, match them correctly
  - Captions contain the same words, but in different order
  - Three settings: Text, Image, Group (is text/image ranked higher than incorrect one given image/text, resp.)
- Challenge: Text and images need to be encoded well, and integrated well
- Benchmark was hand-crafted by VL and CL researchers
- 1600 image–caption pairs (800 correct, 800 incorrect)

Winoground: Probing VL Models for VL Compositionality

[Thrush et al., 2022] `https://huggingface.co/datasets/facebook/winoground`

Examples

(a) there is [a mug] in [some grass]

(b) there is [some grass] in [a mug]

(c) a person [sits] and a dog [stands]

(d) a person [stands] and a dog [sits]

(e) it's a [truck] [fire]

(f) it's a [fire] [truck]

Object *Relation* *Both*

Winoground: Probing VL Models for VL Compositionality

[Thrush et al., 2022] https://huggingface.co/datasets/facebook/winoground

Examples

(a) the kid [with the magnifying glass] looks at them []

(c) the person with the ponytail [packs] stuff and other [buys] it

(e) there are [three] people and [two] windows

(b) the kid [] looks at them [with the magnifying glass]

(d) the person with the ponytail [buys] stuff and other [packs] it

(f) there are [two] people and [three] windows

Pragmatics *Series* *Symbolic*

Winoground: Probing VL Models for VL Compositionality

[Thrush et al., 2022]

Model	Text	Image	Group
MTurk Human	**89.50**	**88.50**	**85.50**
Random Chance	25.00	25.00	16.67
VinVL	**37.75**	17.75	14.50
UNITER$_{large}$	**38.00**	14.00	10.50
UNITER$_{base}$	**32.25**	13.25	10.00
ViLLA$_{large}$	**37.00**	13.25	11.00
ViLLA$_{base}$	**30.00**	12.00	8.00
VisualBERT$_{base}$	15.50	2.50	1.50
ViLT (ViT-B/32)	**34.75**	14.00	9.25
LXMERT	19.25	7.00	4.00
ViLBERT$_{base}$	23.75	7.25	4.75
UniT$_{ITMfinetuned}$	19.50	6.25	4.00
CLIP (ViT-B/32)	**30.75**	10.50	8.00
VSE++$_{COCO}$ (ResNet)	22.75	8.00	4.00
VSE++$_{COCO}$ (VGG)	18.75	5.50	3.50
VSE++$_{Flickr30k}$ (ResNet)	20.00	5.00	2.75
VSE++$_{Flickr30k}$ (VGG)	19.75	6.25	4.50
VSRN$_{COCO}$	17.50	7.00	3.75
VSRN$_{Flickr30k}$	20.00	5.00	3.50

- In most cases, models perform no better than chance

VSR: Visual Spatial Reasoning

[Liu et al., 2023] https://github.com/cambridgeltl/visual-spatial-reasoning

VSR: Visual Spatial Reasoning

[Liu et al., 2023] https://github.com/cambridgeltl/visual-spatial-reasoning

Overview

- 10k image–test pairs
- 66 spatial relation types
- Data creation: annotators are given 2 images and need to complete a caption, making it correct for only one image; subsequent manual validation
- Used dataset: MS COCO

Category	Spatial Relations
Adjacency	Adjacent to, alongside, at the side of, at the right side of, at the left side of, attached to, at the back of, ahead of, against, at the edge of
Directional	Off, past, toward, down, deep down*, up*, away from, along, around, from*, into, to*, across, across from, through, down from
Orientation	Facing, facing away from, parallel to, perpendicular to
Projective	On top of, beneath, beside, behind, left of, right of, under, in front of, below, above, over, in the middle of
Proximity	By, close to, near, far from, far away from
Topological	Connected to, detached from, has as a part, part of, contains, within, at, on, in, with, surrounding, among, consists of, out of, between, inside, outside, touching
Unallocated	Beyond, next to, opposite to, after*, among, enclosed by

VSR: Visual Spatial Reasoning

[Liu et al., 2023] https://github.com/cambridgeltl/visual-spatial-reasoning

Examples

Figure 1: Caption: *The potted plant is at the right side of the bench.* Label: `True`.

Figure 2: Caption: *The cow is ahead of the person.* Label: `False`.

VSR: Visual Spatial Reasoning

[Liu et al., 2023] https://github.com/cambridgeltl/visual-spatial-reasoning

model↓	random split	zero-shot split
human ceiling	95.4	
CLIP (w/ prompting)	56.0	54.5
VisualBERT	$55.2_{\pm 1.4}$	$51.0_{\pm 1.9}$
ViLT	$\mathbf{69.3}_{\pm 0.9}$	$\mathbf{63.0}_{\pm 0.9}$
LXMERT	$\mathbf{70.1}_{\pm 0.9}$	$61.2_{\pm 0.4}$

- Models fine-tuned on training data
- CLIP zero-shot setting
- Zero-shot split: unseen objects

VSR: Visual Spatial Reasoning

[Liu et al., 2023] https://github.com/cambridgeltl/visual-spatial-reasoning

model↓	random split	zero-shot split
human ceiling	95.4	
CLIP (w/ prompting)	56.0	54.5
VisualBERT	$55.2_{\pm 1.4}$	$51.0_{\pm 1.9}$
ViLT	$\mathbf{69.3}_{\pm 0.9}$	$\mathbf{63.0}_{\pm 0.9}$
LXMERT	$\mathbf{70.1}_{\pm 0.9}$	$61.2_{\pm 0.4}$

- Models fine-tuned on training data
- CLIP zero-shot setting
- Zero-shot split: unseen objects

Findings

- **Positional** encodings are very relevant for VSR
- Poor generalisation ability to **unseen objects**
- Challenging relation types: **orientation** of objects and **proximity** (difficult concept, depends on object)
- Size of training data is not a factor

VSR: Visual Spatial Reasoning

[Liu et al., 2023] `https://github.com/cambridgeltl/visual-spatial-reasoning`

Example: Orientation of object

(a) Caption: *The hair drier is facing away from the person*. Label: `False`.

Outline

Measuring Progress in Fine-grained Vision-and-Language Understanding

[Bugliarello et al., 2023]

Measuring Progress in Fine-grained Vision-and-Language Understanding

[Bugliarello et al., 2023]

Overview

- Comparison of state-of-the-art vision–language models with corss-modal attention and trained with contrastive loss and (masked) language modelling
- Evaluation benchmarks
  - SVO: verb comprehension [Hendricks and Nematzadeh, 2021]
  - VALSE: linguistic capabilities [Parcalabescu et al., 2022]
  - Winoground: composition [Thrush et al., 2022]
  - VSR: spatial relations [Liu et al., 2023]

Measuring Progress in Fine-grained Vision-and-Language Understanding

[Bugliarello et al., 2023]

Compared Models

Model	Loss			Data		Downstream		
	CL	Text	Obj Det	Unsupervised	Supervised	VQAv2	NLVR2	RefCOCO+
ALBEF$_{4M}$	✔	MLM	-	4M: COCO+SBU+VG+CC$_{3M}$	-	74.7	80.5	-
ALBEF$_{14M}$	✔	MLM	-	14M: 4M + CC$_{12M}$	-	76.0	83.1	-
BLIP$_{14M}$	✔	LM	-	CapFilt/B(14M)	-	77.6	82.3	-
BLIP$_{129M}$	✔	LM	-	CapFilt/B(14M + LAION)	-	78.2	83.1	-
BLIP$_{129M}$-CapFilt/L	✔	LM	-	CapFilt/L(14M + LAION)	-	78.3	82.2	-
BLIP-ViT/L$_{129M}$	✔	LM	-	CapFilt/L(14M + LAION)	-	-	-	-
PEVL$_{14M}$	✔	MLM	MLM	14M	RefCOCO{,+,g}+F30KE+GQA+VCR+VG	-	-	74.5
X-VLM$_{4M}$	✔	MLM	Regress	4M	COCO + VG	78.1	84.2	71.0
X-VLM$_{16M}$	✔	MLM	Regress	14M	COCO + VG + Objects365 + OpenImages	78.4	84.4	76.9

Table 2: Overview of core evaluated models. All the models use contrastive learning (CL), cross-attention and a (masked) language modelling objective. Fine-grained models also predict object locations from supervised data.

Measuring Progress in Fine-grained Vision-and-Language Understanding

[Bugliarello et al., 2023]

	Model	SVO Avg.	VALSE Avg.	VSR Test Avg.	Winoground Text	Winoground Image	Winoground Group
	Random	50.0	50.0	50.0	25.0	25.0	12.5
	CLIP_{400M}	81.6	64.0	N/A	30.7	10.5	8.0
	BLIP-2_{129M}	86.5	74.0	61.5	43.0	22.0	18.2
1	ALBEF_{4M}	87.6	69.1	57.3	29.2	15.5	11.0
2	$\text{X-VLM}_{4M}{}^{\sharp}$	88.9	72.4	63.0	44.0	**26.7**	**21.5**
3	ALBEF_{14M}	88.6	69.4	58.3	32.5	16.2	12.7
4	BLIP_{14M}	48.7	67.8	49.7	36.5	18.5	14.5
5	$\text{PEVL}_{14M}{}^{\sharp}$	86.2	68.9	57.5	33.2	15.7	12.2
8	$\text{X-VLM}_{16M}{}^{\sharp}$	**90.0**	**74.5**	**64.3**	**46.7**	24.5	21.2
9	BLIP_{129M}	51.4	68.8	46.9	35.5	15.0	11.7
10	BLIP_{129M}-CapFilt/L	51.2	68.2	48.7	34.7	15.2	12.2
11	BLIP-ViT/L_{129M}	50.8	70.3	50.3	34.7	14.5	12.2

- X-VLM & PEVL: fine-grained models (fine-grained training tasks)
- X-VLM: object localisation, region descriptions

Outline

Improving VL Models

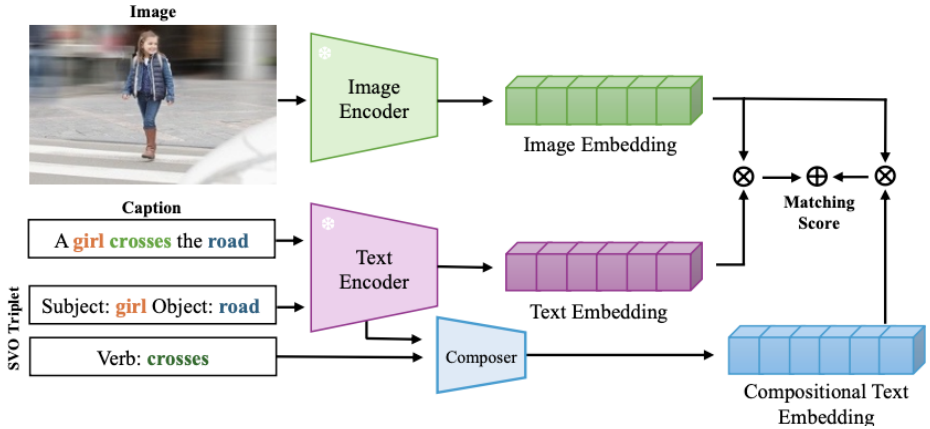
How to fix the limitations of VL Models?

- VerbCLIP [Wazni et al., 2024]
- HNC: Hard Negative Captions [Dönmez et al., 2023]
- When and why Vision-Language Models behave like Bags-of-Words, and what to do about it? [Yuksekgonul et al., 2023]
  `github.com/mertyg/vision-language-models-are-bows`

Improving VL Models: VerbCLIP [Wazni et al., 2024]

Improving Verb Understanding in VL Models with Compositional Structures

- CLIP (dual encoder) found to be object-focused
- Idea: Combine CLIP with verb-focused compositional models



- Best method: **Copy-Add**

$$\overrightarrow{T_{sent}} \cdot \overrightarrow{I_{img}} + \overrightarrow{svo}$$

$$\overrightarrow{svo} = \alpha[\vec{s} \odot (\bar{v} \times \vec{o})] + \beta[(\vec{s} \times \bar{v}) \odot \vec{o})] \cdot \overrightarrow{I_{img}}$$

$$\bar{v} = \vec{v} \otimes \vec{v} \text{ (outer product; Kronecker)}$$

Improving VL Models: VerbCLIP [Wazni et al., 2024]

Improving Verb Understanding in VL Models with Compositional Structures

	VALSE	VL-Checklist	SVO-Probes
VerbCLIP	75.24	60.41	77.30
CLIP	72.86	57.27	77.43

The goat *stands* in the grass.		
The goat *lies* in the grass.		
	Positive	Negative
CLIP	28.71	**28.73** ✗
VerbCLIP	**37.28** ✓	37.12

A baby *speaks* on the telephone.		
A baby *sits* on the telephone.		
	Positive	Negative
CLIP	28.01	**28.11** ✗
VerbCLIP	**36.51** ✓	36.06

A person *holding* ski-poles.		
A person *crossing* ski-poles.		
	Positive	Negative
CLIP	28.65	**28.68** ✗
VerbCLIP	**35.16** ✓	34.87

A man *threw* the ball.		
A man *holding* the ball.		
	Positive	Negative
CLIP	18.50	**19.54** ✗
VerbCLIP	**5.095** ✓	4.759

- But: computationally expensive
- But: limited to subject-verb-object format

References I

Beňová, I., Košecká, J., Gregor, M., Tamajka, M., Veselỳ, M., and Šimko, M. (2024).

Beyond image-text matching: Verb understanding in multimodal transformers using guided masking.

arXiv preprint arXiv:2401.16575.

Bugliarello, E., Sartran, L., Agrawal, A., Hendricks, L. A., and Nematzadeh, A. (2023).

Measuring progress in fine-grained vision-and-language understanding.

In Rogers, A., Boyd-Graber, J., and Okazaki, N., editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1559–1582.

Chen, F., Zhang, D., Han, M., Chen, X., Shi, J., Xu, S., and Xu, B. (2022).

Vlp: A survey on vision-language pre-training.

arXiv preprint arXiv:2202.09061.

Dobreva, R. and Keller, F. (2021).

Investigating negation in pre-trained vision-and-language models.

In Bastings, J., Belinkov, Y., Dupoux, E., Giulianelli, M., Hupkes, D., Pinter, Y., and Sajjad, H., editors, *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 350–362, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Dönmez, E., Tilli, P., Yang, H.-Y., Vu, N. T., and Silberer, C. (2023).

Hnc: Leveraging hard negative captions towards models with fine-grained visual-linguistic comprehension capabilities.

In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 364–388.

Hendricks, L. A. and Nematzadeh, A. (2021).

Probing image-language transformers for verb understanding.

In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3635–3644.

References II

Liu, F., Emerson, G., and Collier, N. (2023).
Visual spatial reasoning.
Transactions of the Association for Computational Linguistics, 11:635–651.

Parcalabescu, L., Cafagna, M., Muradjan, L., Frank, A., Calixto, I., and Gatt, A. (2022).
VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena.
In Muresan, S., Nakov, P., and Villavicencio, A., editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8253–8280.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021).
Learning Transferable Visual Models From Natural Language Supervision.
CoRR, abs/2103.00020.

Salin, E., Farah, B., Ayache, S., and Favre, B. (2022).
Are vision-language transformers learning multimodal representations? a probing perspective.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 11248–11257.

Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., and Ross, C. (2022).
Winoground: Probing vision and language models for visio-linguistic compositionality.
In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Wazni, H., Lo, K. I., and Sadrzadeh, M. (2024).
VerbCLIP: Improving verb understanding in vision-language models with compositional structures.

In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pages 195–201.

References III

Yuksekgonul, M., Bianchi, F., Kalluri, P., Jurafsky, D., and Zou, J. (2023).
When and why vision-language models behave like bags-of-words, and what to do about it?
In *International Conference on Learning Representations*.