VALSE

Phenomenon-centered testing of Vision and Language models

UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Letitia Parcalabescu

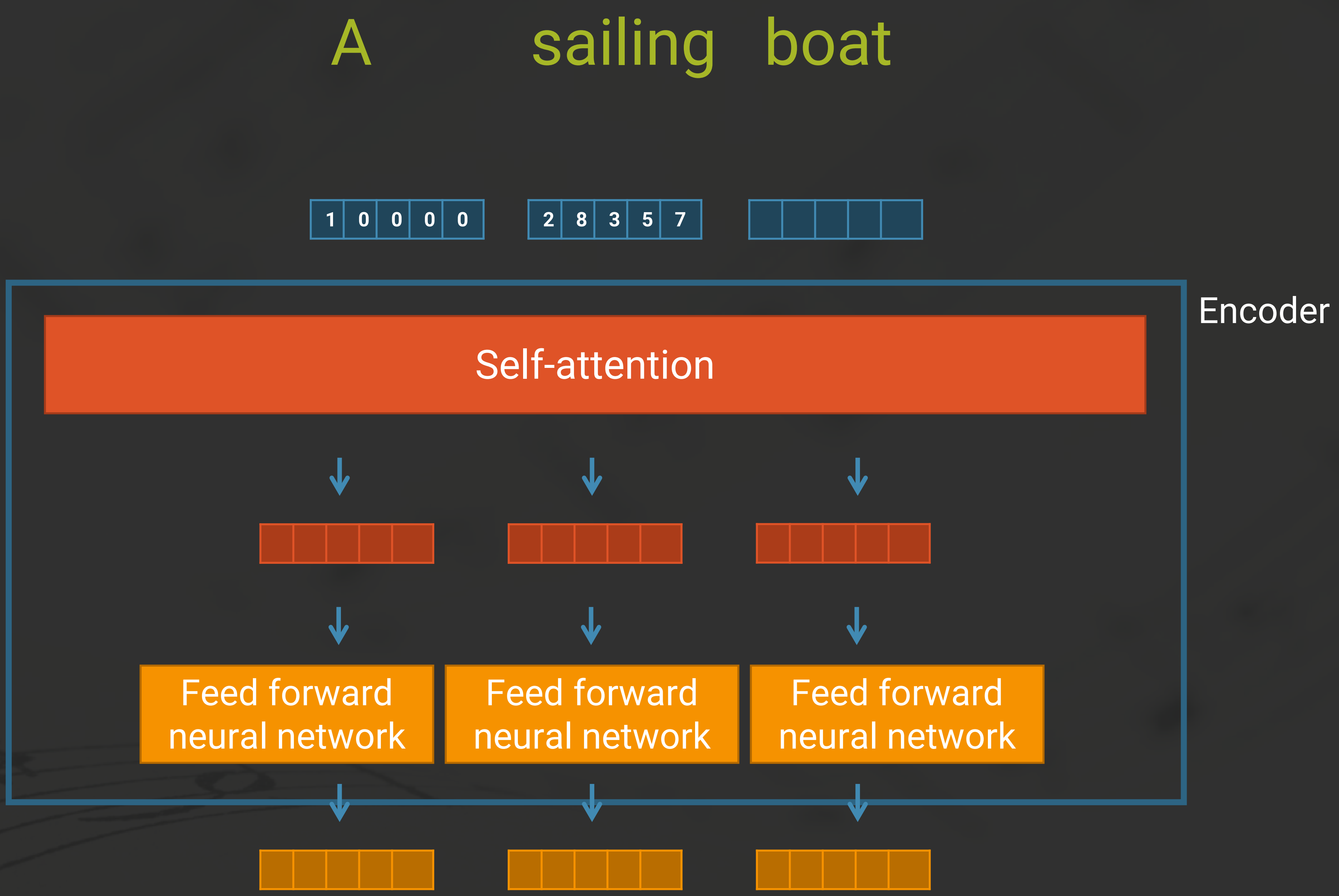
Computational Linguistics Department
Heidelberg University

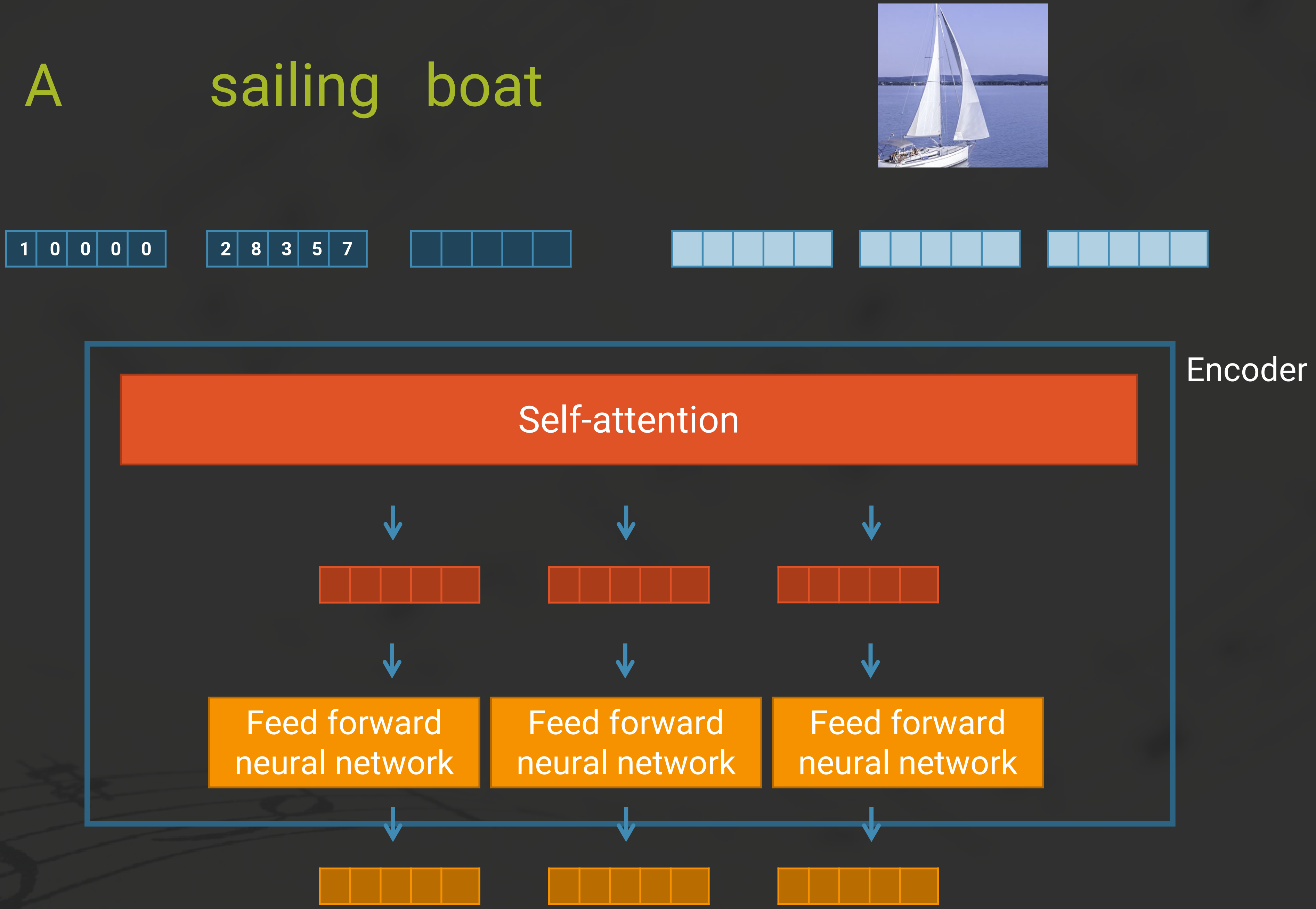
Heidelberg Natural Language Processing Group

What is the state of SoTA in V&L?

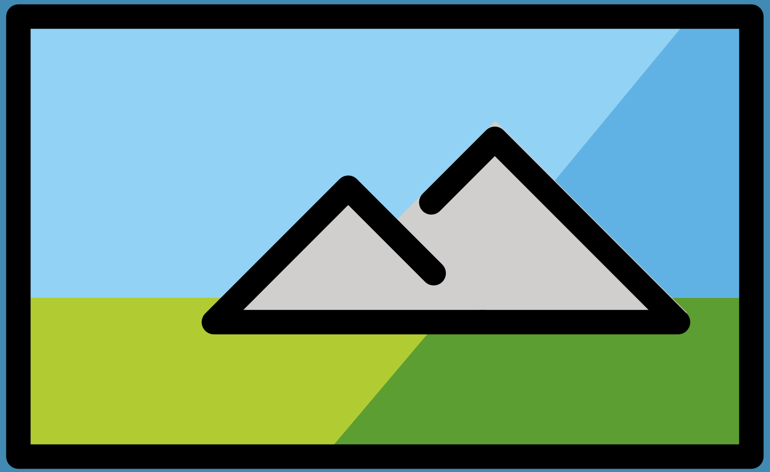
Vision and Language (V&L) models

Multimodal Transformers


A
sailing
boat
1 0 0 0 0
2 8 3 5 7
Encoder
Self-attention
Feed forward neural network
Feed forward neural network
Feed forward neural network


A
sailing
boat
1 0 0 0 0
2 8 3 5 7
Encoder
Self-attention
Feed forward neural network
Feed forward neural network
Feed forward neural network

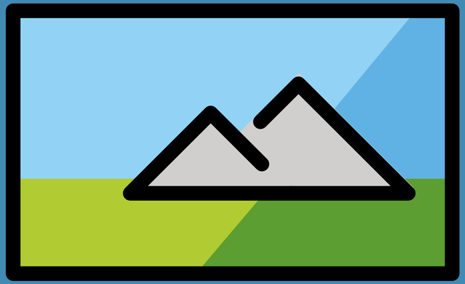
Vision and Language Model



There are mountains in the image.

Vision and Language Model

There are mountains in the image.



LXMERT

ViLBERT

ViLBERT 12-in-1

VL-BERT

UNITER

VisualBERT

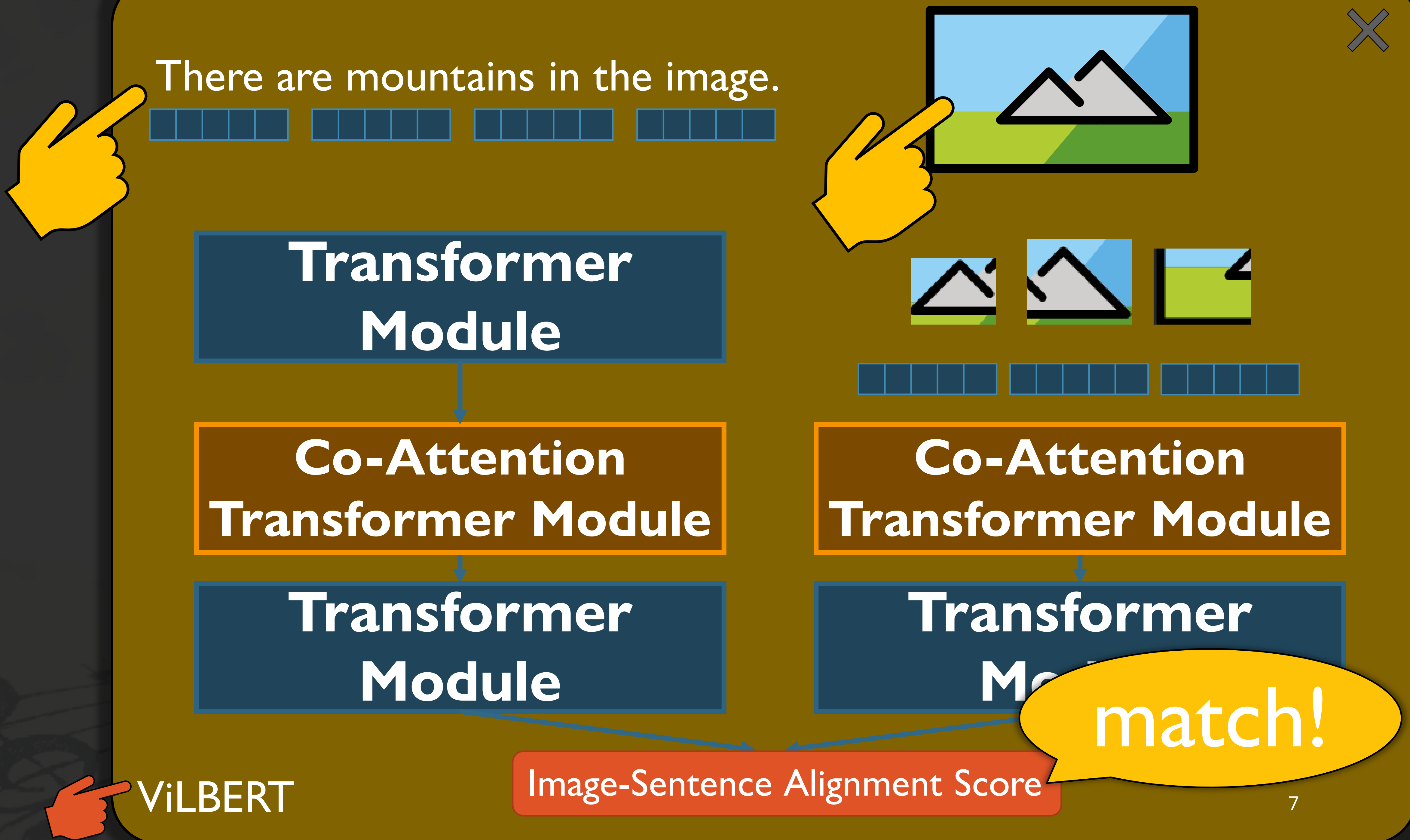
VinLV

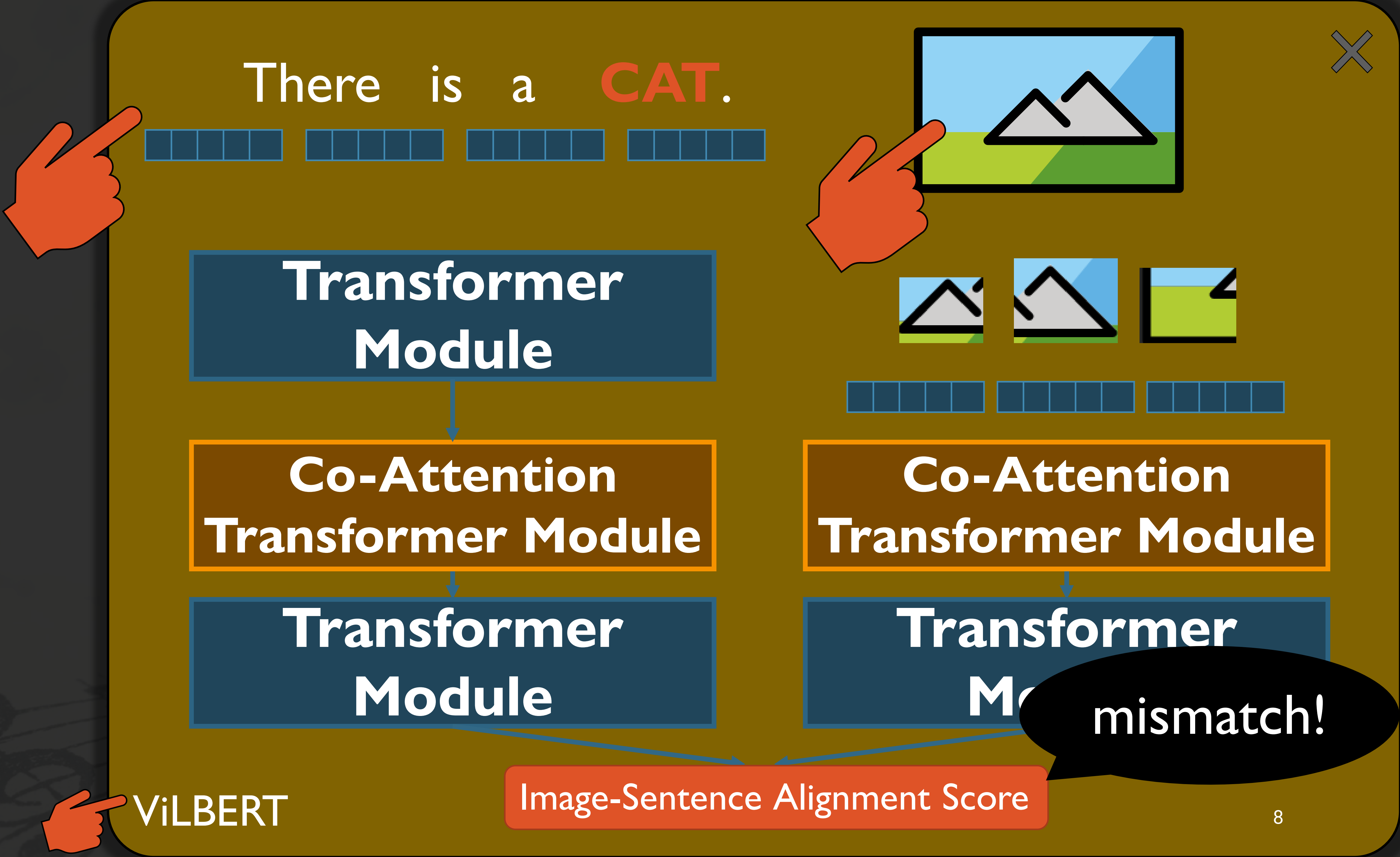
Unicoder-VL

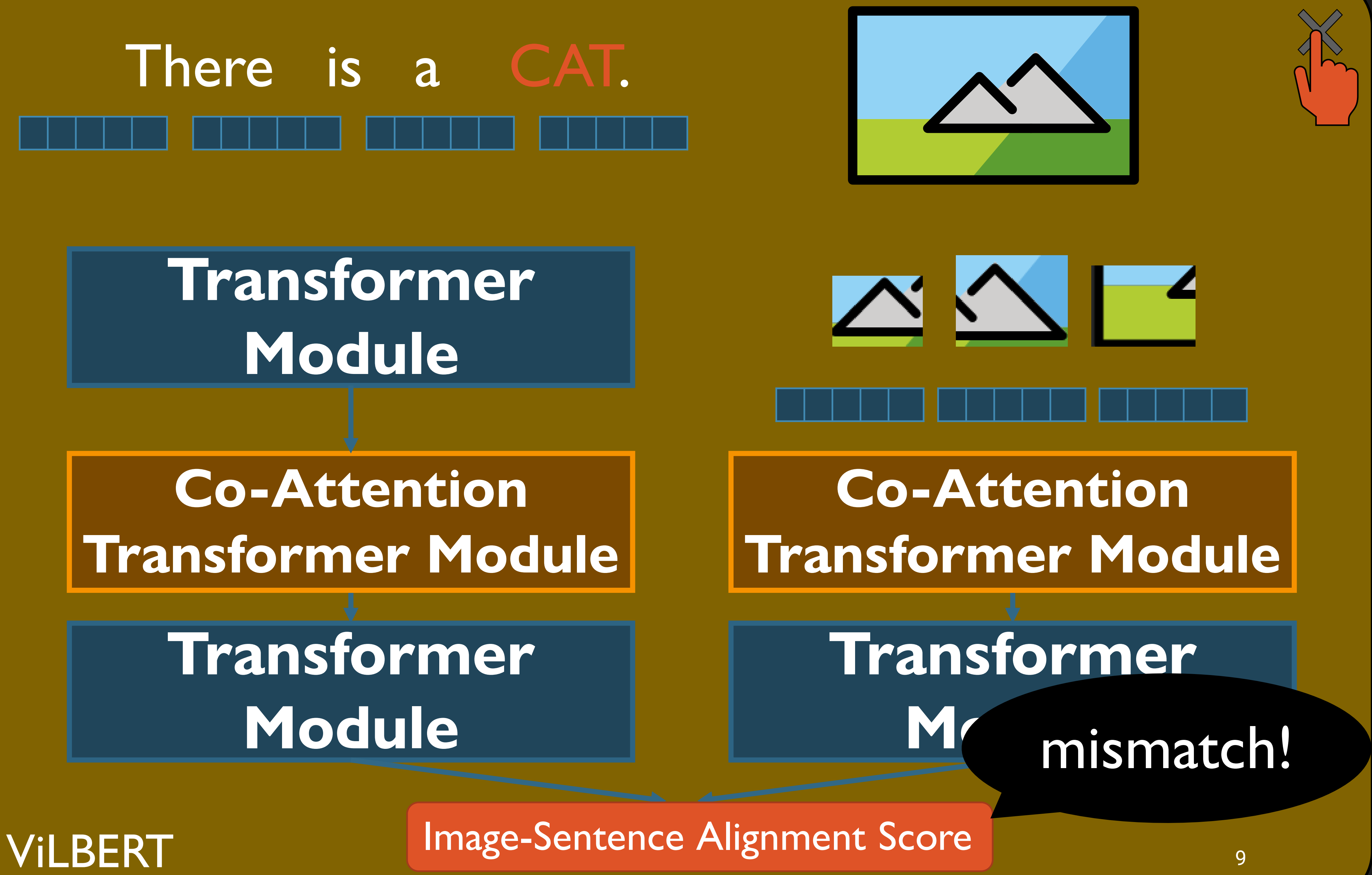
Oscar

X-LXMERT

…

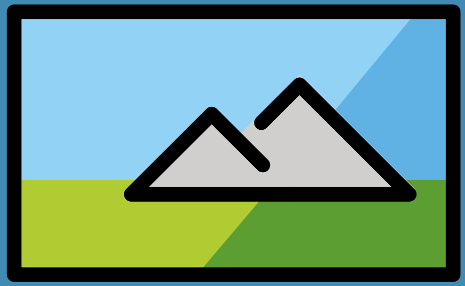

There are mountains in the image.
Transformer Module
Co-Attention Transformer Module
Transformer Module
Co-Attention Transformer Module
Transformer Module
match!
Image-Sentence Alignment Score
ViLBERT


There is a CAT.
Transformer Module
Co-Attention Transformer Module
Transformer Module
Co-Attention Transformer Module
Transformer Mo
mismatch!
Image-Sentence Alignment Score
ViLBERT


There is a CAT.
Transformer Module
Co-Attention Transformer Module
Transformer Module
Co-Attention Transformer Module
Transformer Mo
mismatch!
Image-Sentence Alignment Score
ViLBERT

Vision and Language Models

There are mountains in the image.



LXMERT

ViLBERT

ViLBERT 12-in-1

VL-BERT

UNITER

VisualBERT

VinLV

Unicoder-VL

X-LXMERT

Oscar

...

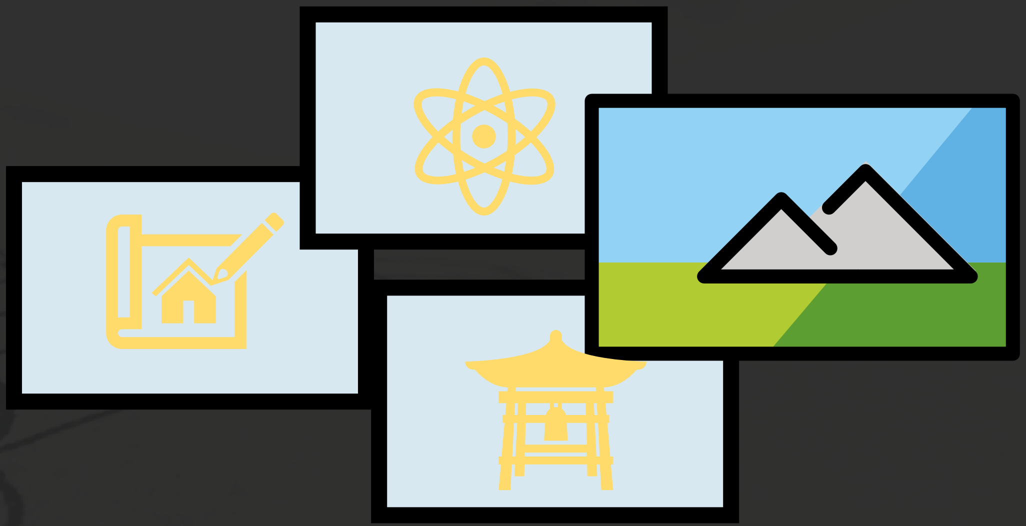
VQA

Is this a mountain? Yes.
How many mountains? Two.

VCR

Is this a mountain?
Yes.
Because it is taller than the horizon.

Image Retrieval



Phrase Grounding

Where are **mountains**?

Task-centrism in the V&L community

task A, task B, …., task Z.

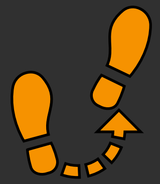
~~Vision~~ and Language Model

How many mountains are there in the image?


FOIL it!
Visual D
Daniela Mass
Elevating the R
Yash Goyal*1
1Virginia T
1{ygoyal, tjskhot}@vt
Does my
Alle
jack
Problems at the inters
are of significant importan
questions and for the rich
However, inherent structu
language tend to be a sin
sual modalities, resulting
mation, leading to an infl
We propose to counter
of Visual Question Answe
in VQA) matter! Specifi
dataset [3] by collectin
every question in our b
not just a single image,
that result in two diffe
dataset is by construc
nal VQA dataset and
of image-question pai
is publicly available
part of the 2nd iterati
Dataset and Challenge (VQA v2.0).
We characterise s
(VD)—a sequenti
are related throu
method based on
near state-of-the
and over-param
ignores the visu
off-the-shelf fe
learns in practi
approaches to v
effects of over-
1 Introductio
Recent years have
AI, enabling natur
machines, early
(Weizenbaum, 1
resurgence of in
neural-network-b
in the perceptual
A particularly thriving
that of visually grounded dialogue, termed visual dialogue
involving an AI agent conversing with a human about visual
Modeling express
seems crucial in
visual question a
times high-perfo
turn out to be mo
nals in the data.
tic tool, empiric
tion projection (
or not cross-mo
formance for a
This function p
dictions so tha
eliminated, iso
structure. For
tasks (on each
the-art bench
cases, removi
sults in little
Surprisingly,
In this
whether
(LaVi)
tion betw
end, we
COCO
ciates in
captions
age tha
ones, b
word').
els fall into the traps of this data
Behind the Scene: Revealing the Secrets of
Pre-trained Vision-and-Language Models
Jize Cao*1, Zhe Gan2, Yu Cheng2, Licheng Yu*3
Yen-Chun Chen2, and Jingjing Liu2
1 University of Washington
caojize@cs.washington.edu
2 Microsoft Dynamics 365 AI Research
{zhe.gan,yu.cheng,yen-chun.chen,jingjl}@microsoft.com
3 Facebook AI
lichengyu@fb.com
Abstract. Recent Transformer-based large-scale pre-trained models have revolutionized vision-and-language (V+L) research. Models such as ViL-BERT, LXMERT and UNITER have significantly lifted state of the art across a wide range of V+L benchmarks. However, little is known about the inner mechanisms that destine their impressive success. To reveal the secrets behind the scene, we present VALUE (Vision-And-Language Understanding Evaluation), a set of meticulously designed probing tasks (e.g., Visual Coreference Resolution, Visual Relation Detection) generalizable to standard pre-trained V+L models, to decipher the inner workings of multimodal pre-training (e.g., implicit knowledge garnered in individual attention heads, inherent cross-modal alignment learned through contextualized multimodal embeddings). Through extensive analysis of

Phenomenon-centrism

Let’s VALSE!



UNIVERSITÄT HEIDELBERG ZUKUNFT SEIT 1386

Heidelberg Natural Language Processing Group

L-Università ta' Malta

NYU

VALSE

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

Letitia Parcalabescu
Computational Linguistics Department
Heidelberg University

Michele Cafagna
Institute of Linguistics and Language Technology
University of Malta

Lilitta Muradjan
Computational Linguistics Department
Heidelberg University

Anette Frank
Computational Linguistics Department
Heidelberg University

Iacer Calixto
New York University
ILLC, University of Amsterdam

Albert Gatt
Institute of Linguistics and Language Technology
University of Malta

Phenomenon-centrism

VALSE: a FOIL concerto of 6 pieces

Plurality

The greenhouse has many plants.

The greenhouse has a single plant.

Counting

The man wears one pair of glasses.

The man wears two pairs of glasses.

Existence

There is a man in the image.

There is no man in the image.

Relations

There is a sink behind the man.

There is a sink to the right of the man.

Coreference

The apron looks clean. Is it white? No.

The apron looks clean. Is it white? Yes.

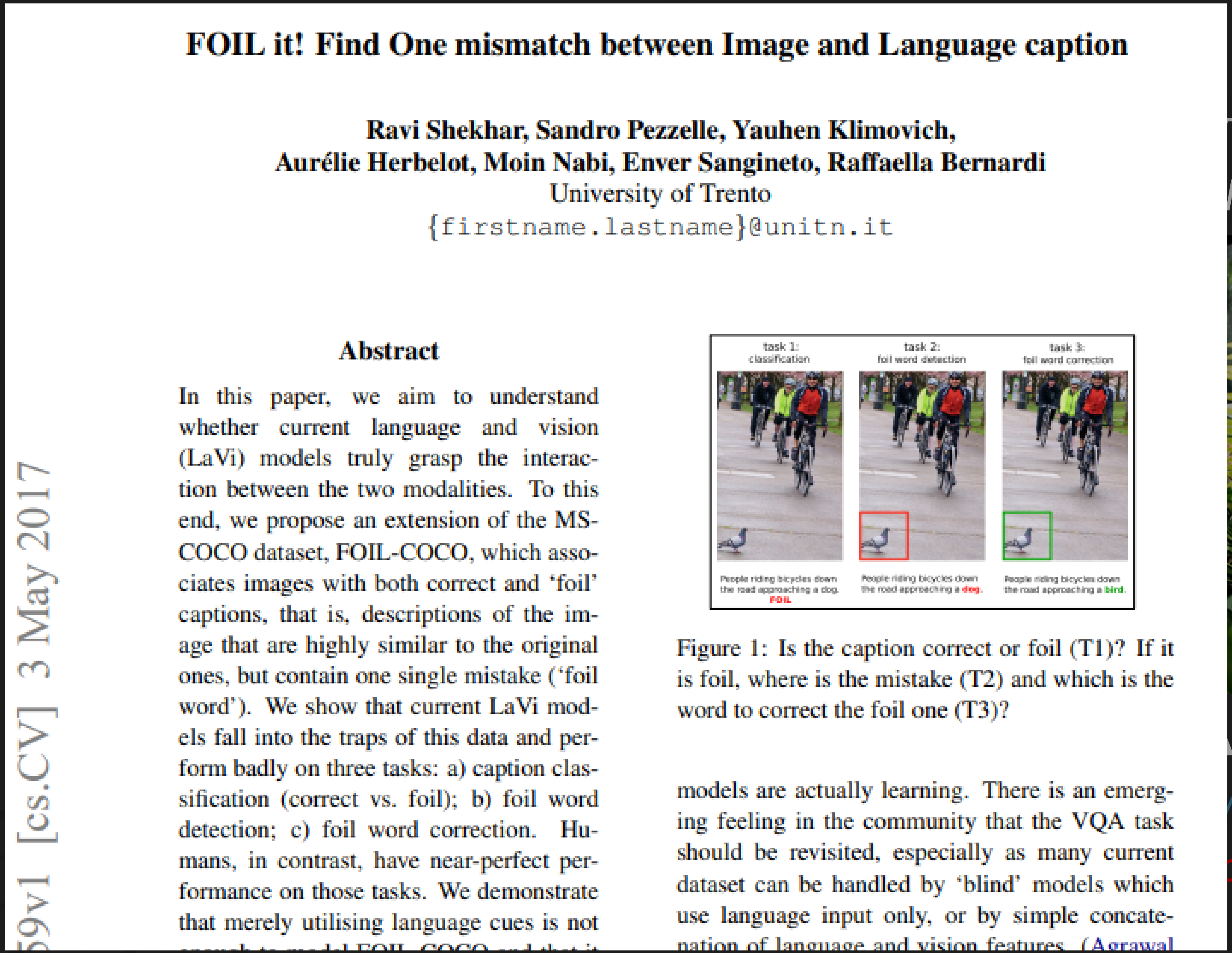
Actions

The man is watering the plants.

The man is cutting the plants.

Phenomenon-centrism

VALSE: a FOIL concerto of 6 pieces



59v1 [cs.CV] 3 May 2017

FOIL it! Find One mismatch between Image and Language caption

Ravi Shekhar, Sandro Pezzelle, Yauhen Klimovich,
Aurélie Herbelot, Moin Nabi, Enver Sangineto, Raffaella Bernardi
University of Trento
{firstname.lastname}@unitn.it

Abstract

In this paper, we aim to understand whether current language and vision (LaVi) models truly grasp the interaction between the two modalities. To this end, we propose an extension of the MS-COCO dataset, FOIL-COCO, which associates images with both correct and 'foil' captions, that is, descriptions of the image that are highly similar to the original ones, but contain one single mistake ('foil word'). We show that current LaVi models fall into the traps of this data and perform badly on three tasks: a) caption classification (correct vs. foil); b) foil word detection; c) foil word correction. Humans, in contrast, have near-perfect performance on those tasks. We demonstrate that merely utilising language cues is not

task 1: classification | task 2: foil word detection | task 3: foil word correction

People riding bicycles down the road approaching a dog. FOIL
People riding bicycles down the road approaching a dog.
People riding bicycles down the road approaching a bird.

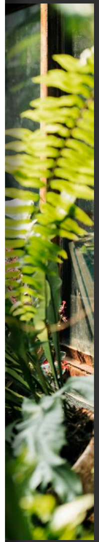
Figure 1: Is the caption correct or foil (T1)? If it is foil, where is the mistake (T2) and which is the word to correct the foil one (T3)?

models are actually learning. There is an emerging feeling in the community that the VQA task should be revisited, especially as many current dataset can be handled by 'blind' models which use language input only, or by simple concatenation of language and vision features (Agrawal

Counting

The man wears one pair of glasses.

The man wears two pairs of glasses.



Relations

There is a sink behind the man.

There is a sink to the right of the man.

Actions

watering the plants.

cutting the plants.

Phenomenon-cen

VALSE: a F

FOIL it! Find One mis

Ravi Shekh
Aurélie Herbelot, M

{firs

Abstract

In this paper, we aim to u
whether current language a
(LaVi) models truly grasp th
tion between the two modalitie
end, we propose an extension o
COCO dataset, FOIL-COCO, w
ciates images with both correct
captions, that is, descriptions o
age that are highly similar to th
ones, but contain one single mis
word'). We show that current LaVi mod-
els fall into the traps of this data and per-
form badly on three tasks: a) caption clas-
sification (correct vs. foil); b) foil word
detection; c) foil word correction. Hu-
mans, in contrast, have near-perfect per-
formance on those tasks. We demonstrate
that merely utilising language cues is not

59v1 [cs.CV] 3 May 2017

word to correct the foil one (T3)?

models are actually learning. There is an emerg-
ing feeling in the community that the VQA task
should be revisited, especially as many current
dataset can be handled by 'blind' models which
use language input only, or by simple concate-
nation of language and vision features (Agrawal

Seeing past words: Testing the cross-modal capabilities of pretrained V&L models on counting tasks

Letitia Parcalabescu[1] Albert Gatt[2] Anette Frank[1] Iacer Calixto[3,4]

[1]Heidelberg University, Department of Computational Linguistics
[2]University of Malta, Institute of Linguistics and Language Technology
[3]New York University [4]ILLC, University of Amsterdam
{parcalabescu,frank}@cl.uni-heidelberg.de
albert.gatt@um.edu.mt, iacer.calixto@nyu.edu

17 Jun 2021

Abstract

We investigate the reasoning ability of pre-trained vision and language (V&L) models in two tasks that require multimodal integration: (1) discriminating a correct image-sentence pair from an incorrect one, and (2) counting entities in an image. We evaluate three pre-trained V&L models on these tasks: ViLBERT, ViLBERT 12-in-1 and LXMERT, in zero-shot and finetuned settings. Our results show that

tasks, e.g. visual question answering (VQA); visual commonsense reasoning; grounding referring expressions; and image retrieval, among others.

Pretrained V&L models use a combination of masked multimodal modelling – i.e., masking out words and object bounding boxes from the input and predicting them – and image-sentence alignment, i.e., predicting whether an image-sentence pair is correctly aligned or not. Such models hold the promise of partially addressing the 'meaning

ctions

atering the plants.

cutting the plants.

VALSE: a FOIL concerto of 6 pieces

Plurality

The greenhouse has many plants.

The greenhouse has a single plant.

Counting

The man wears one pair of glasses.

The man wears two pairs of glasses.

Existence

There is a man in the image.

There is no man in the image.

Relations

There is a sink behind the man.

There is a sink to the right of the man.

Coreference

The apron looks clean. Is it white? No.

The apron looks clean. Is it white? Yes.

Actions

The man is watering the plants.

The man is cutting the plants.

The plants are watering the man.

VALSE: a FOIL concerto of 6 pieces

Plurality

The greenhouse has many plants.

The greenhouse has a single plant.

Counting

The man wears one pair of glasses.

The man wears two pairs of glasses.

Existence

There is a man in the image.

There is no man in the image.

Relations

There is a sink behind the man.

There is a sink to the right of the man.

Coreference

The apron looks clean. Is it white? No.

The apron looks clean. Is it white? Yes.

Actions

The man is watering the plants.

The man is cutting the plants.

The plants are watering the man.

plausibility bias

How to obtain valid foils?

- Language models for generating foil words (e.g., SpanBERT)
- Natural Language Inference (NLI)
- Human annotation

Natural Language Inference filtering

neutral / contradiction

Premise → Hypothesis

Caption

The greenhouse has many plants.

Foil

The greenhouse has a single plant.

Premise → Hypothesis

entailment

Natural Language Inference filtering

neutral / contradiction

Premise Hypothesis

Caption

The greenhouse has many plants.

The greenhouse has many plants.

Foil

The greenhouse has a single plant.

The greenhouse has some plants.

entailment

Premise Hypothesis

entailment

Natural Language Inference filtering

neutral / contradiction

Premise

Hypothesis

Caption

The greenhouse has many plants.

Foil

The greenhouse has a single plant.

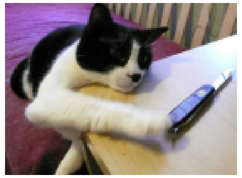
Premise

Hypothesis

entailment

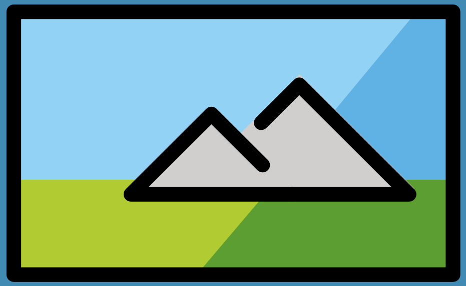
How to obtain valid foils?

- Language models for generating foil words (e.g., SpanBERT)
- Natural Language Inference (NLI)
- Human annotation

	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						

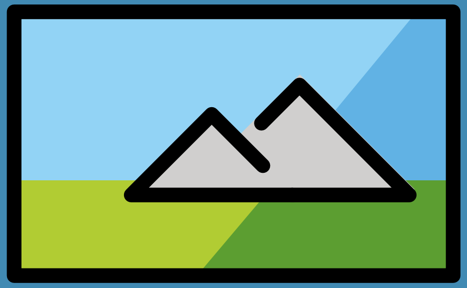
Vision and Language Model

There are mountains in the image.



Vision and Language Model

There are mountains in the image.



LXMERT

ViLBERT

ViLBERT 12-in-1

VisualBERT

CLIP

zero-shot testing

Pairwise accuracy

Caption

The greenhouse has many plants.

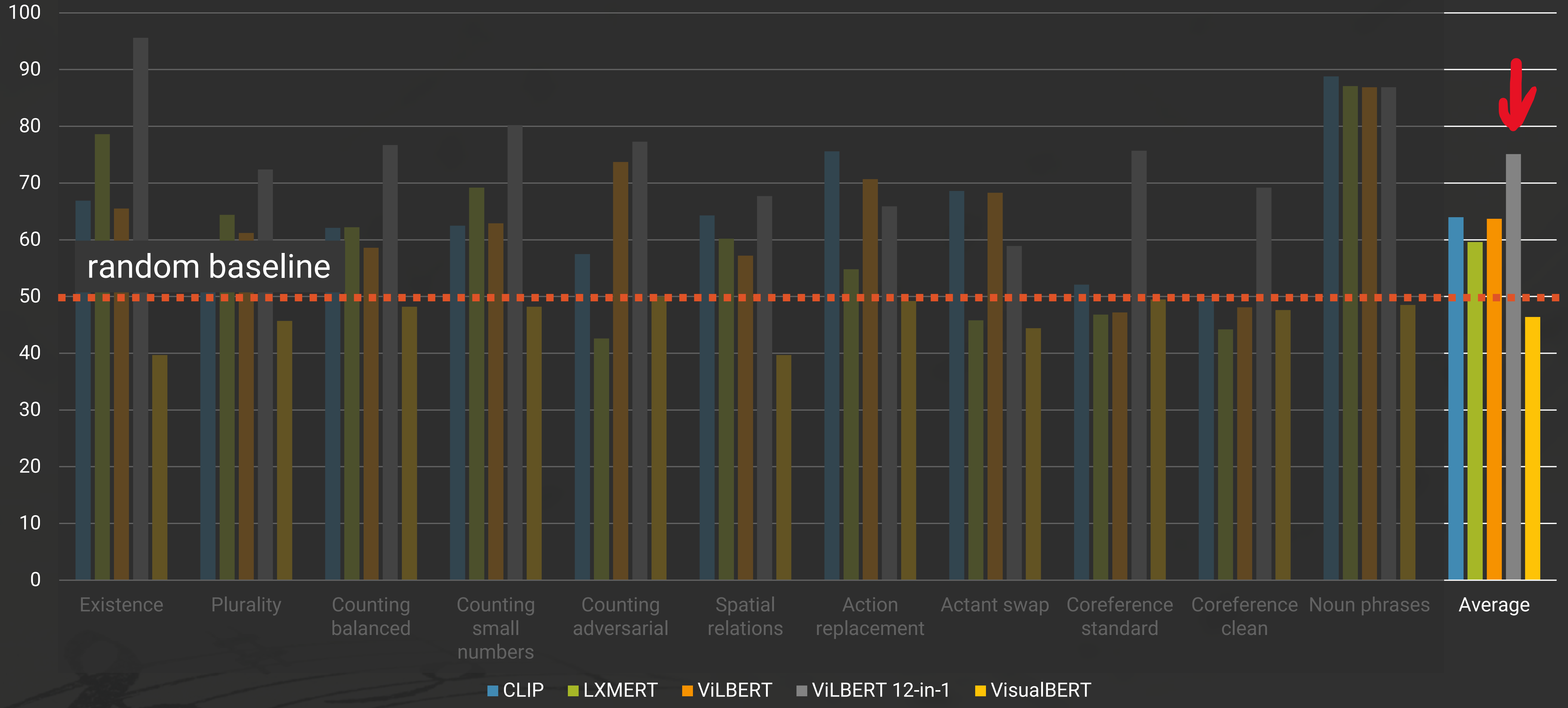
Foil

The greenhouse has a single plant.

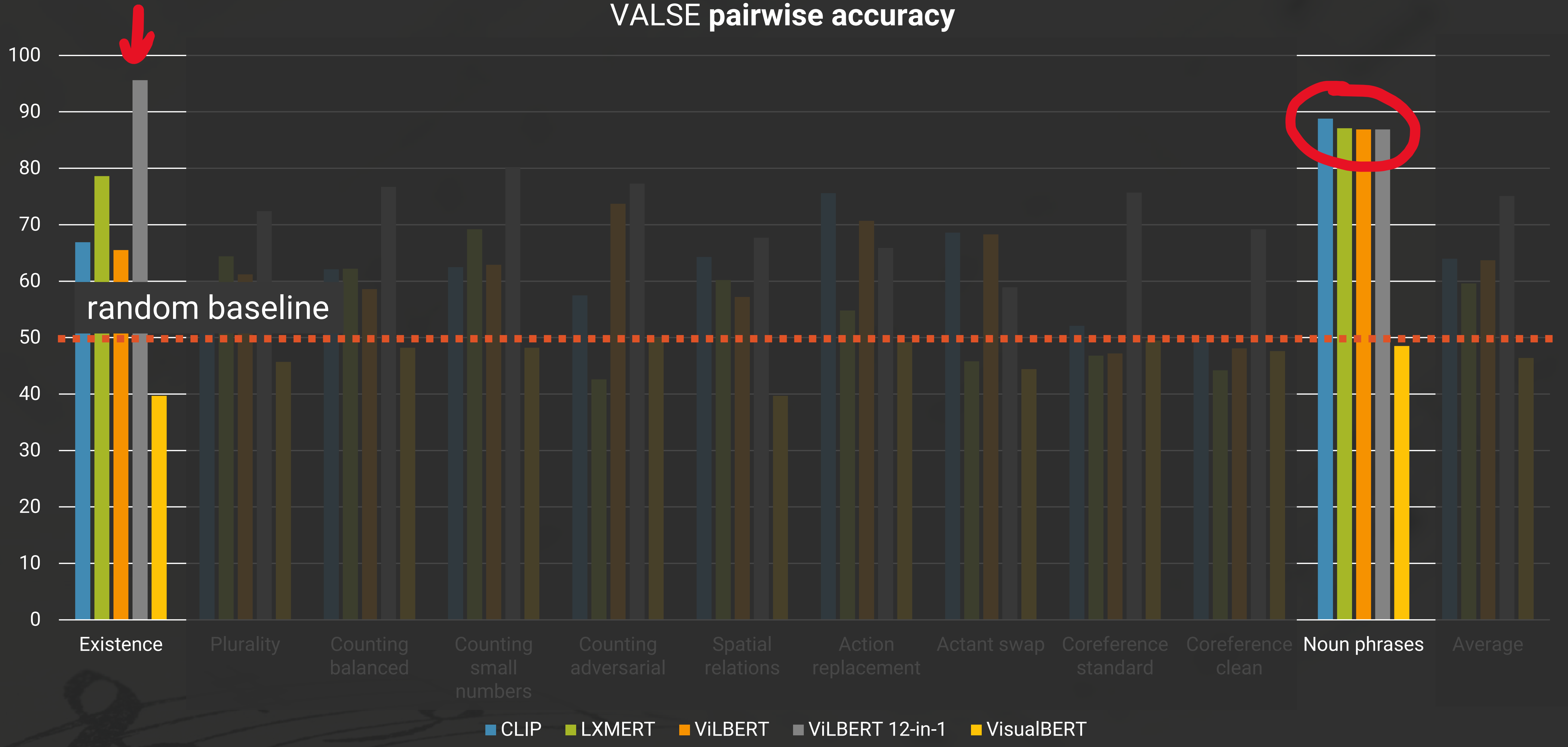
image-sentence alignment score ≥ image-sentence alignment score

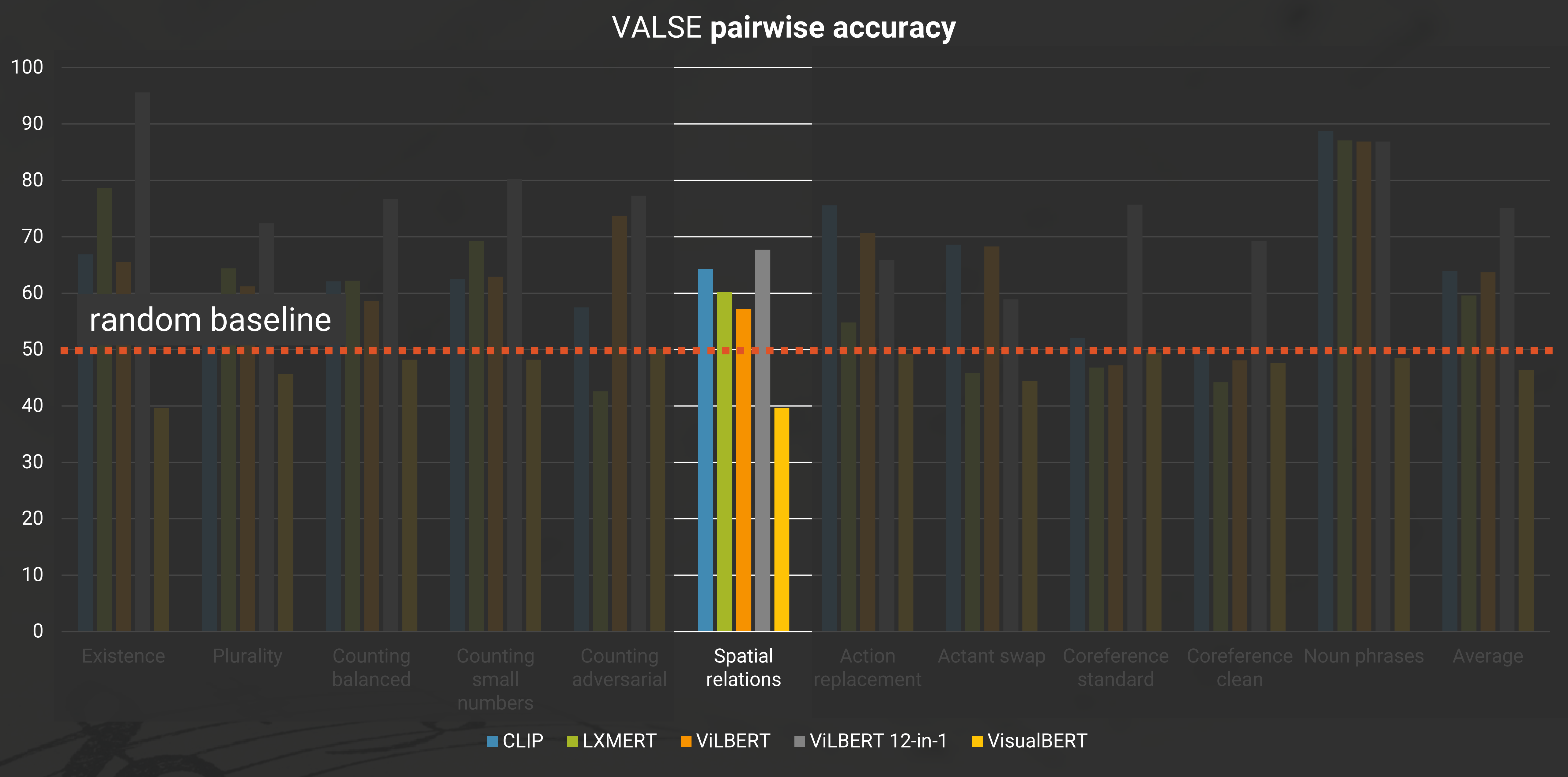
<

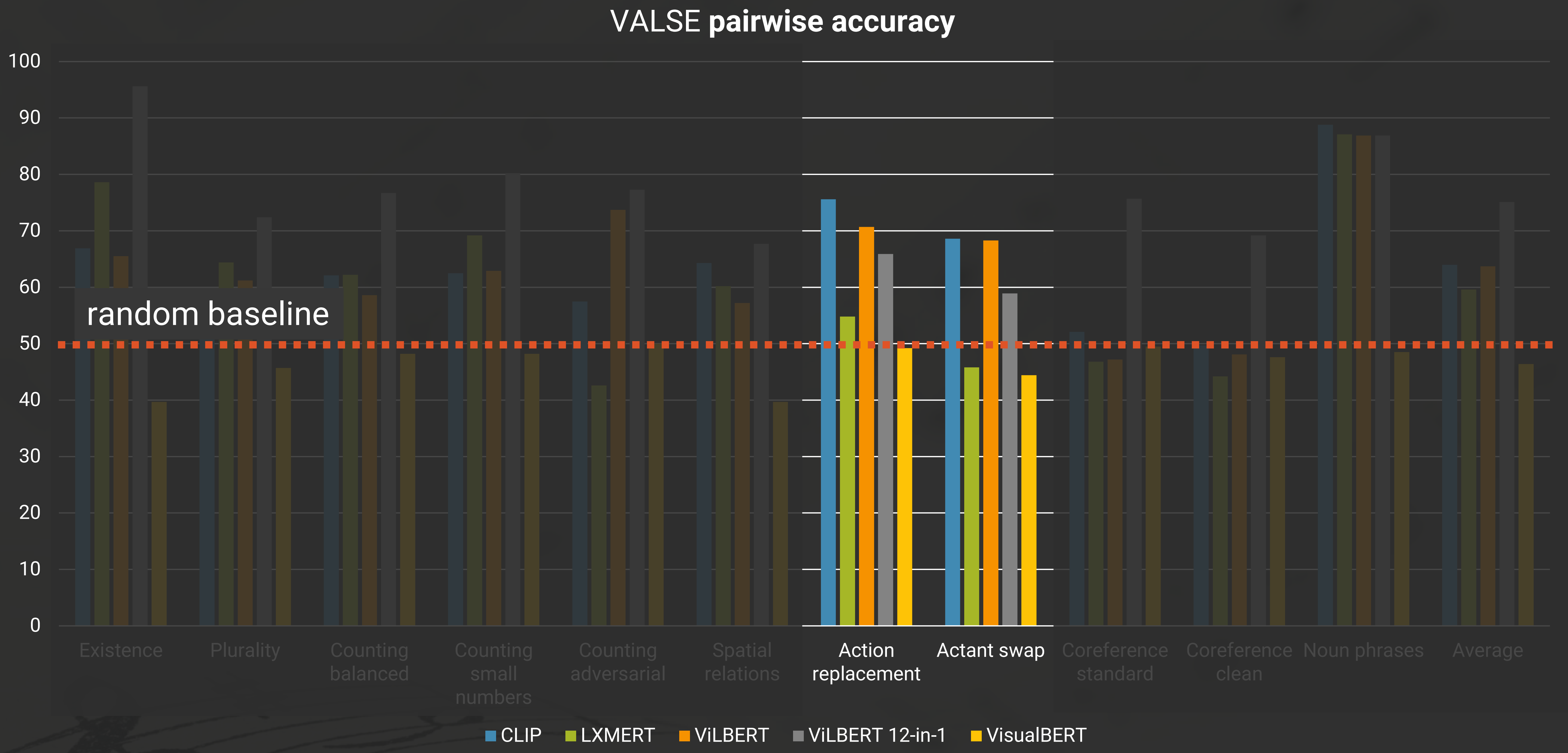
VALSE **pairwise accuracy**

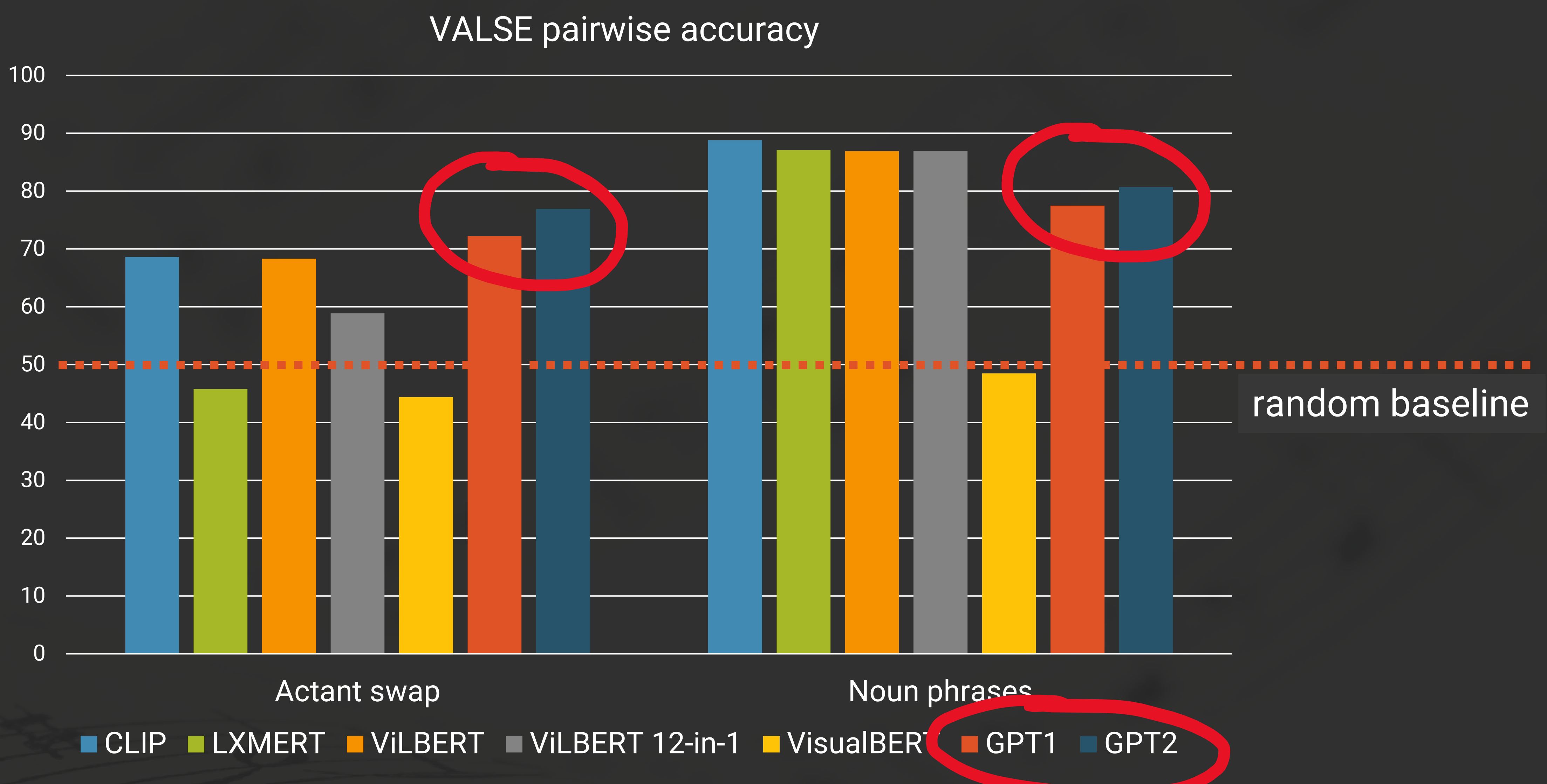


VALSE **pairwise accuracy**




VALSE pairwise accuracy
100
90
80
70
60
50
40
30
20
10
0
random baseline
Existence
Plurality
Counting balanced
Counting small numbers
Counting adversarial
Spatial relations
Action replacement
Actant swap
Coreference standard
Coreference clean
Noun phrases
Average
CLIP
LXMERT
ViLBERT
ViLBERT 12-in-1
VisualBERT


VALSE **pairwise accuracy**
100
90
80
70
60
50
40
30
20
10
0
random baseline
Existence
Plurality
Counting balanced
Counting small numbers
Counting adversarial
Spatial relations
Action replacement
Actant swap
Coreference standard
Coreference clean
Noun phrases
Average
CLIP
LXMERT
ViLBERT
ViLBERT 12-in-1
VisualBERT


VALSE pairwise accuracy
100
90
80
70
60
50
40
30
20
10
0
random baseline
Actant swap
Noun phrases
CLIP
LXMERT
ViLBERT
ViLBERT 12-in-1
VisualBERT
GPT1
GPT2

unimodal!

VALSE: a FOIL concerto of 6 pieces

Plurality

The greenhouse has many plants.

The greenhouse has a single plant.

Counting

The man wears one pair of glasses.

The man wears two pairs of glasses.

Existence

There is a man in the image.

There is no man in the image.

Relations

There is a sink behind the man.

There is a sink to the right of the man.

Coreference

The apron looks clean. Is it white? No.

The apron looks clean. Is it white? Yes.

Actions

The man is watering the plants.

The man is cutting the plants.

https://github.com/Heidelberg-NLP/VALSE