CLASP, University of Gothenburg
March 26, 2025

Not (yet) the whole story

The need for human-like evaluation of visual LLMs in multimodal communicative tasks

Sandro Pezzelle
ILLC, University of Amsterdam

s.pezzelle@uva.nl
sandropezzelle.github.io

Overview

In this talk, I will focus on how current visual LLMs deal with a fundamental element of human multimodal communication: **the ability of narrating events that happen to us and around us.**

This is a fundamental and pervasive element of human language communication, widely regarded as **a hallmark of human intelligence** [Heider & Simmel, 1944; Wiessner, 2014].

Overview

We—human speakers—narrate what happens in the surrounding multimodal (primarily visual) context in **many communicative scenarios** and **for very different purposes**.

“A ***story*** *describes a sequence of events involving one or more people.”*

–Antoniak et al., 2024

A hallmark of human intelligence

Telling stories requires **many complex skills**, including:

- understanding sequences of visual actions;
- leveraging common sense and previous knowledge;
- understanding others’ feelings (theory of mind);
- making inferences and predictions;
- targeting an audience: expertise, POV, style, etc.

Beyond standard language-vision tasks

This goes well beyond the set of abilities required for **standard tasks** like Visual Question Answering or Image Captioning

… Where the focus is typically on **objects and attributes grounded** (i.e., fully visible) in a static image!

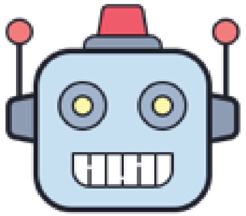
Beyond standard language-vision tasks

Visual Question Answering

- What color is the raincoat?
- What does the person on the right hold?

Beyond standard language-vision tasks

Visual Question Answering



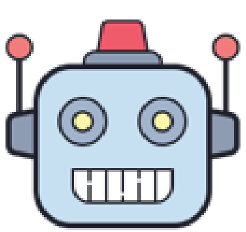
- What color is the raincoat? **Yellow** ✓
- What does the person on the right hold? **Book** ✗

Beyond standard language-vision tasks

Image Captioning

Beyond standard language-vision tasks

Image Captioning



- “A couple of tourists on the lakeside on a rainy day”
- “Two people waiting at the edge of a parking lot”

Real-life language can be different

Real-life language can be different

*"The flowers are **there**"*

Real-life language can be different

*"The flowers are **there***"*

*semantically underspecified

Real-life language can be different

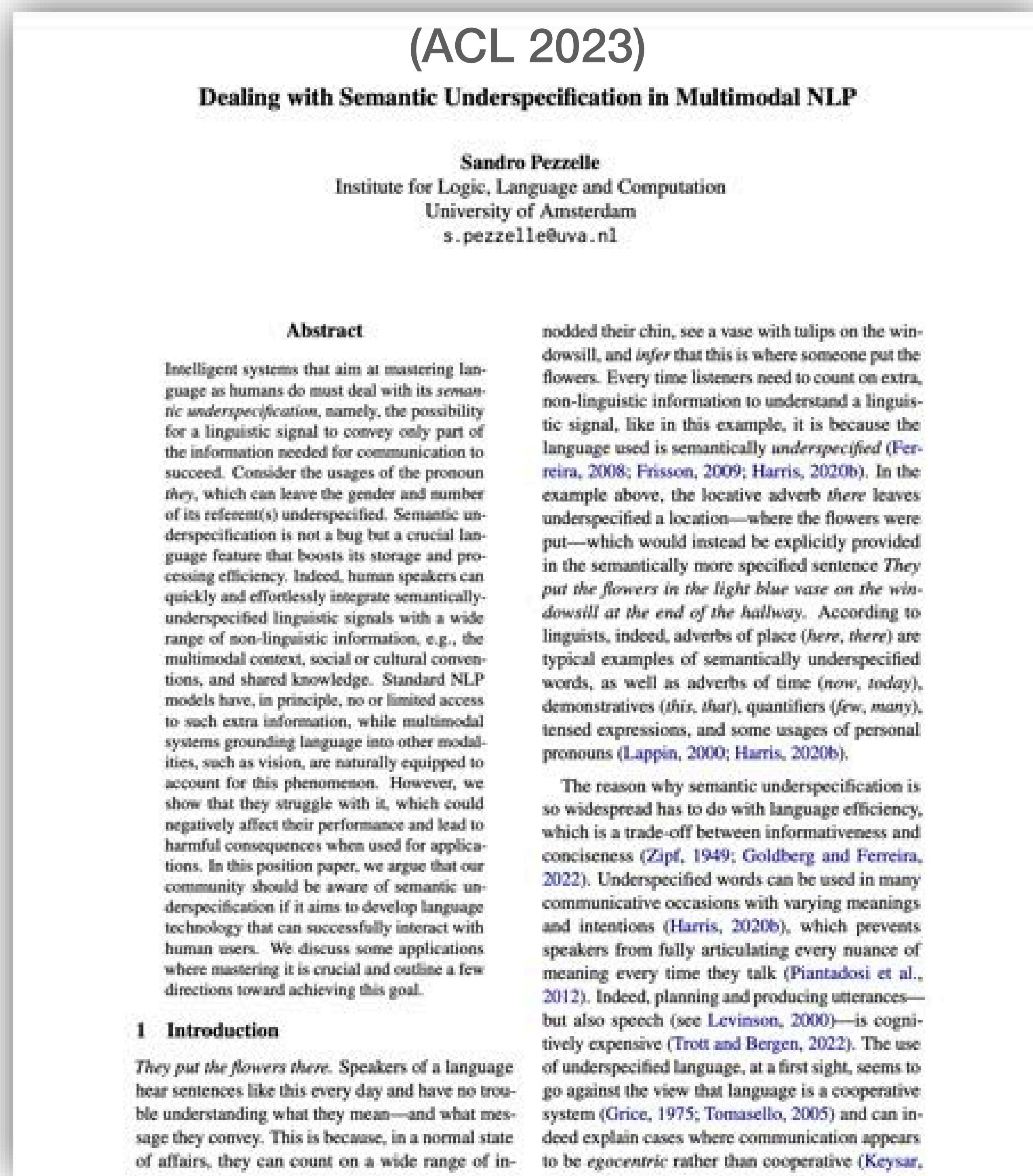
*"The flowers are **there****"

*semantically underspecified

Real-life language can be different

*“The flowers are **there***”*

(ACL 2023)



Dealing with Semantic Underspecification in Multimodal NLP

Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
s.pezzelle@uva.nl

Abstract

Intelligent systems that aim at mastering language as humans do must deal with its *semantic underspecification*, namely, the possibility for a linguistic signal to convey only part of the information needed for communication to succeed. Consider the usages of the pronoun *they*, which can leave the gender and number of its referent(s) underspecified. Semantic underspecification is not a bug but a crucial language feature that boosts its storage and processing efficiency. Indeed, human speakers can quickly and effortlessly integrate semantically-underspecified linguistic signals with a wide range of non-linguistic information, e.g., the multimodal context, social or cultural conventions, and shared knowledge. Standard NLP models have, in principle, no or limited access to such extra information, while multimodal systems grounding language into other modalities, such as vision, are naturally equipped to account for this phenomenon. However, we show that they struggle with it, which could negatively affect their performance and lead to harmful consequences when used for applications. In this position paper, we argue that our community should be aware of semantic underspecification if it aims to develop language technology that can successfully interact with human users. We discuss some applications where mastering it is crucial and outline a few directions toward achieving this goal.

1 Introduction

They put the flowers there. Speakers of a language hear sentences like this every day and have no trouble understanding what they mean—and what message they convey. This is because, in a normal state of affairs, they can count on a wide range of information about the surrounding context, personal knowledge and experience, social or cultural conventions, and so on. Upon hearing this sentence, for example, they would know that flowers go into vases, look in front of them, infer that this is where someone put the flowers. Every time listeners need to count on extra, non-linguistic information to understand a linguistic signal, like in this example, it is because the language used is semantically *underspecified* (Ferreira, 2008; Frisson, 2009; Harris, 2020b). In the example above, the locative adverb *there* leaves underspecified a location—where the flowers were put—which would instead be explicitly provided in the semantically more specified sentence *They put the flowers in the light blue vase on the windowsill at the end of the hallway.* According to linguists, indeed, adverbs of place (*here*, *there*) are typical examples of semantically underspecified words, as well as adverbs of time (*now*, *today*), demonstratives (*this*, *that*), quantifiers (*few*, *many*), tensed expressions, and some usages of personal pronouns (Lappin, 2000; Harris, 2020b).

The reason why semantic underspecification is so widespread has to do with language efficiency, which is a trade-off between informativeness and conciseness (Zipf, 1949; Goldberg and Ferreira, 2022). Underspecified words can be used in many communicative occasions with varying meanings and intentions (Harris, 2020b), which prevents speakers from fully articulating every nuance of meaning every time they talk (Piantadosi et al., 2012). Indeed, planning and producing utterances—but also speech (see Levinson, 2000)—is cognitively expensive (Trott and Bergen, 2022). The use of underspecified language, at a first sight, seems to go against the view that language is a cooperative system (Grice, 1975; Tomasello, 2005) and can indeed explain cases where communication appears to be *egocentric* rather than cooperative (Keysar,

CLIP and underspecified descriptions

Type	Description	Score
Original	The woman is standing above the two packed suitcases.	0.8565

CLIP and underspecified descriptions

TYPE	DESCRIPTION	SCORE
Original	The woman is standing above the two packed suitcases.	0.8565
Quantity	The woman is standing above **some** packed suitcases.	
Gender	The **person** is standing above the two packed suitcases.	
Gender+Number	**They** are standing above the two packed suitcases.	
Location	The woman is standing **here**.	
Object	The woman is standing above **this**.	
Full	**They are doing something here.**	

manually perturbed descriptions to make them underspecified

CLIP and underspecified descriptions

Type	Description	Score
Original	The woman is standing above the two packed suitcases.	0.8565
Quantity	The woman is standing above **some** packed suitcases.	0.8275
Gender	The **person** is standing above the two packed suitcases.	0.7608
Gender+Number	**They** are standing above the two packed suitcases.	0.7435
Location	The woman is standing **here**.	0.5537
Object	The woman is standing above **this**.	0.4931
Full	**They are doing something here.**	0.4646

(much) lower scores assigned to underspecified captions!

Takeaways

Contrastive models (*à la* CLIP) struggle with communicatively plausible—underspecified—descriptions.

We need to shift the paradigm to a scenario where models use language as humans do, that is, **with a communicative goal.**

Visual storytelling is one such task!


Visual Storytelling

From encoders to generative VLMs

In recent years, huge advancements in the development of visual LLMs (VLMs): **GPT-4V**, **LLAVA**, **Gemini**, **Molmo**…

- impressive capabilities on virtually any language and vision tasks—IC and VQA mentioned above are basically solved
- no fine-tuning: these big models can be used via prompting for any tasks, including **generating narrations for visual events**, being them (sequences of) images or videos

Generative VLMs

What could be a good description of this event by the journalist?

"Good evening, this is Sarah Chen reporting live from the scene of a catastrophic disaster. What we're witnessing here is nothing short of a catastrophe. The street before me has been torn apart by a massive sinkhole, leaving behind a chaotic landscape of rubble and destruction.

The entire building to my right has collapsed, its walls crumbled into a heap of debris. In the distance, we can see emergency vehicles, including a yellow excavator and a red fire truck, working tirelessly to manage the situation.

A firefighter is standing nearby, assessing the danger and coordinating the rescue efforts. The atmosphere is heavy with tension and the threat of further collapses.

This disaster has left many in the community displaced and in urgent need of assistance. As we wait for more information, I want to emphasize the scale of this event and the challenges faced by local authorities and residents.

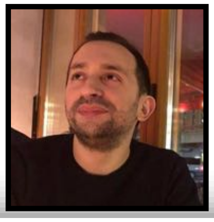
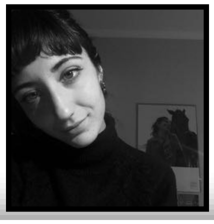
Stay tuned as we continue to bring you updates on this developing situation."

Big questions

1. How good are these models at generating stories—a task requiring tons of (linguistic, pragmatic, creative) abilities?

2. How can we evaluate models—as narrating visual events is not something that can be measured via accuracy?

Back to the (recent) past

(EMNLP Findings 2020)

Be Different to Be Better!
A Benchmark to Leverage the Complementarity of Language and Vision

Sandro Pezzelle[1], **Claudio Greco**[2], **Greta Gandolfi**[2], **Eleonora Gualdoni**[2], **Raffaella Bernardi**[2,3]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]CIMeC, [3]DISI, University of Trento
s.pezzelle@uva.nl,
{greta.gandolfi|eleonora.gualdoni}@studenti.unitn.it,
{claudio.greco|raffaella.bernardi}@unitn.it

Abstract

This paper introduces BD2BB, a novel language and vision benchmark that requires multimodal models combine *complementary* information from the two modalities. Recently, impressive progress has been made to develop universal multimodal encoders suitable for virtually any language and vision tasks. However, current approaches often require them to combine *redundant* information provided by language and vision. Inspired by real-life communicative contexts, we propose a novel task where either modality is necessary but not sufficient to make a correct prediction. To do so, we first build a dataset of images and corresponding sentences provided by human participants. Second, we evaluate state-of-the-art models and compare their performance against human speakers. We show that, while the task is relatively easy for humans, best-performing models struggle to achieve similar results.

1 Introduction

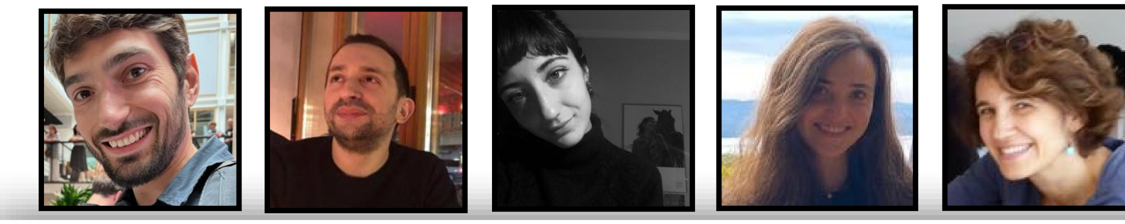
Human communication, in real-life situations, is multimodal (Kress, 2010): To convey and understand a message uttered in natural language, people build on what is present in the multimodal context surrounding them. As such, speakers do not need to "repeat" something that is already provided by the environment; similarly, listeners leverage information from various modalities, such as vision, to interpret the linguistic message. Integrating information from multiple modalities is indeed crucial for attention and perception (Partan and Marler, 1999) since combined information from concurrent modalities can give rise to different messages (McGurk and MacDonald, 1976).

roni, 2016; Beinborn et al., 2018). However, computational approaches to language and vision typically do not fully explore this complementarity. To illustrate, given an image (e.g., the one depicted in Figure 1), popular tasks involve describing it in natural language, e.g., *"A tennis player about to hit the ball"* (Image Captioning; see Bernardi et al., 2016); answering questions that are grounded in it, e.g., Q: *"What sport is he playing?"*, A: *"Tennis"* (Visual Question Answering; see Antol et al., 2015); having a dialogue on its entities, e.g., Q: *"Is the person holding a racket?"*, A: *"Yes."* (visually-grounded dialogue; see De Vries et al., 2017; Das et al., 2017). While all these tasks challenge models to perform visual grounding, i.e., an effective *alignment* of language and vision, none of them require a genuine *combination* of complementary information provided by the two modalities. All the information is fully available in the visual scene, and language is used to describe or retrieve it.

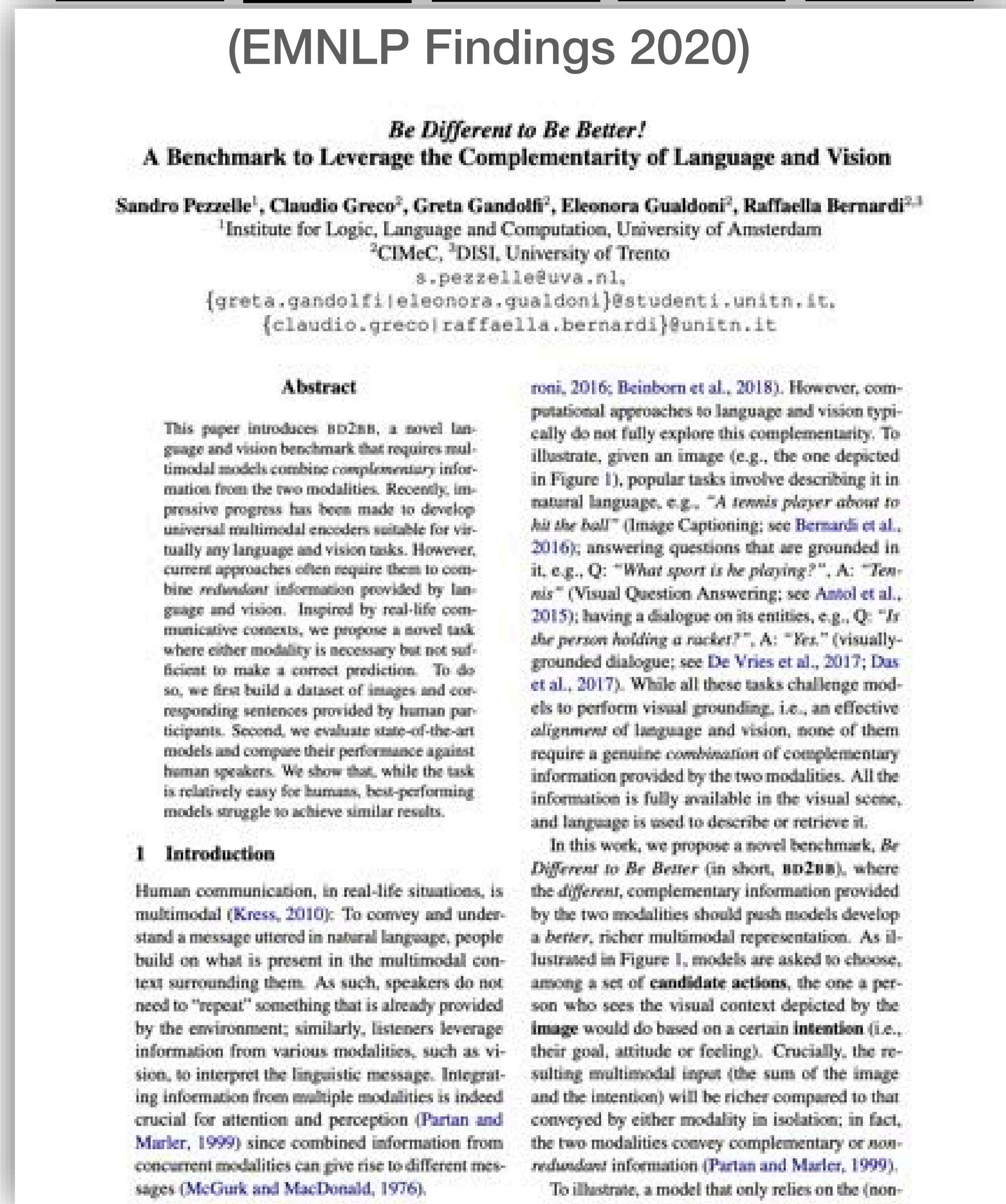
In this work, we propose a novel benchmark, *Be Different to Be Better* (in short, BD2BB), where the *different*, complementary information provided by the two modalities should push models develop a *better*, richer multimodal representation. As illustrated in Figure 1, models are asked to choose, among a set of **candidate actions**, the one a person who sees the visual context depicted by the **image** would do based on a certain **intention** (i.e., their goal, attitude or feeling). Crucially, the resulting multimodal input (the sum of the image and the intention) will be richer compared to that conveyed by either modality in isolation; in fact, the two modalities convey complementary or *non-redundant* information (Partan and Marler, 1999).

To illustrate, a model that only relies on the (non-

Single-image events



(EMNLP Findings 2020)



Be Different to Be Better!
A Benchmark to Leverage the Complementarity of Language and Vision

Sandro Pezzelle[1], Claudio Greco[2], Greta Gandolfi[2], Eleonora Gualdoni[2], Raffaella Bernardi[2,3]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]CIMeC, [3]DISI, University of Trento
s.pezzelle@uva.nl,
{greta.gandolfi|eleonora.gualdoni}@studenti.unitn.it,
{claudio.greco|raffaella.bernardi}@unitn.it

Abstract

This paper introduces BD2BB, a novel language and vision benchmark that requires multimodal models combine *complementary* information from the two modalities. Recently, impressive progress has been made to develop universal multimodal encoders suitable for virtually any language and vision tasks. However, current approaches often require them to combine *redundant* information provided by language and vision. Inspired by real-life communicative contexts, we propose a novel task where either modality is necessary but not sufficient to make a correct prediction. To do so, we first build a dataset of images and corresponding sentences provided by human participants. Second, we evaluate state-of-the-art models and compare their performance against human speakers. We show that, while the task is relatively easy for humans, best-performing models struggle to achieve similar results.

1 Introduction

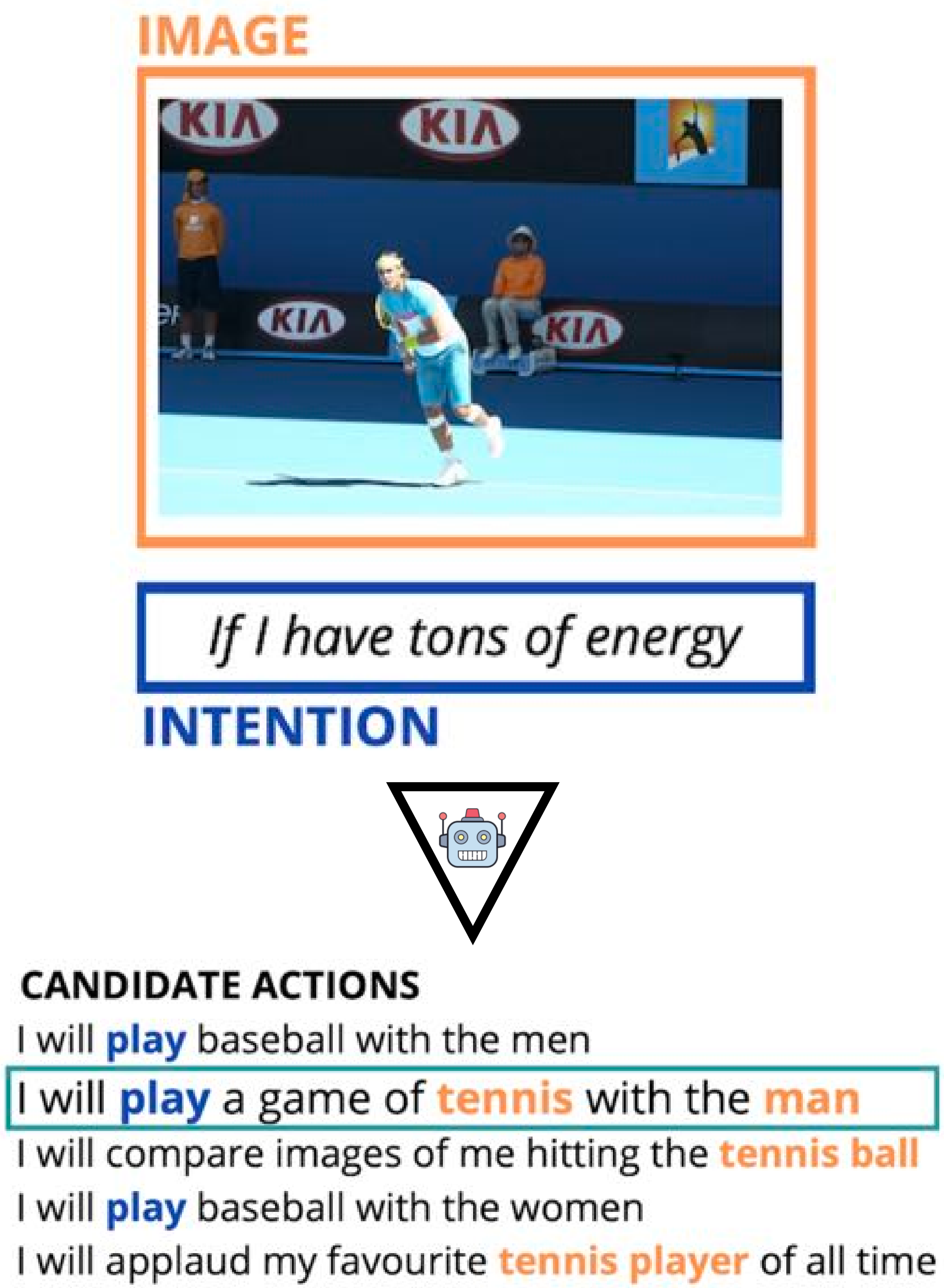
Human communication, in real-life situations, is multimodal (Kress, 2010): To convey and understand a message uttered in natural language, people build on what is present in the multimodal context surrounding them. As such, speakers do not need to "repeat" something that is already provided by the environment; similarly, listeners leverage information from various modalities, such as vision, to interpret the linguistic message. Integrating information from multiple modalities is indeed crucial for attention and perception (Partan and Marler, 1999) since combined information from concurrent modalities can give rise to different messages (McGurk and MacDonald, 1976).

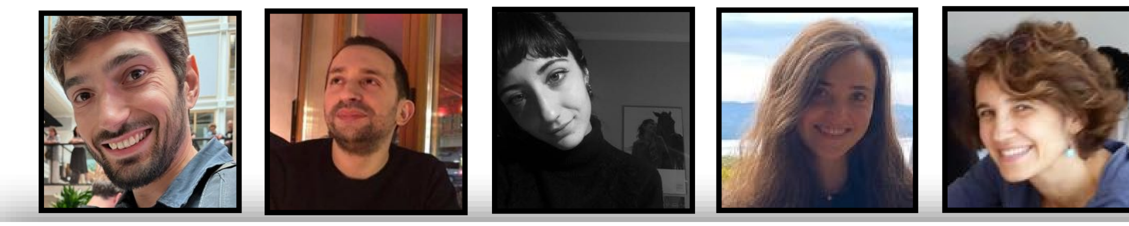
roni, 2016; Beinborn et al., 2018). However, computational approaches to language and vision typically do not fully explore this complementarity. To illustrate, given an image (e.g., the one depicted in Figure 1), popular tasks involve describing it in natural language, e.g., *"A tennis player about to hit the ball"* (Image Captioning; see Bernardi et al., 2016); answering questions that are grounded in it, e.g., Q: *"What sport is he playing?"*, A: *"Tennis"* (Visual Question Answering; see Antol et al., 2015); having a dialogue on its entities, e.g., Q: *"Is the person holding a racket?"*, A: *"Yes."* (visually-grounded dialogue; see De Vries et al., 2017; Das et al., 2017). While all these tasks challenge models to perform visual grounding, i.e., an effective *alignment* of language and vision, none of them require a genuine *combination* of complementary information provided by the two modalities. All the information is fully available in the visual scene, and language is used to describe or retrieve it.

In this work, we propose a novel benchmark, *Be Different to Be Better* (in short, BD2BB), where the *different*, complementary information provided by the two modalities should push models develop a *better*, richer multimodal representation. As illustrated in Figure 1, models are asked to choose, among a set of **candidate actions**, the one a person who sees the visual context depicted by the **image** would do based on a certain **intention** (i.e., their goal, attitude or feeling). Crucially, the resulting multimodal input (the sum of the image and the intention) will be richer compared to that conveyed by either modality in isolation; in fact, the two modalities convey complementary or *non-redundant* information (Partan and Marler, 1999).

To illustrate, a model that only relies on the (non-

Single-image events





(EMNLP Findings 2020)

Be Different to Be Better!

A Benchmark to Leverage the Complementarity of Language and Vision

Sandro Pezzelle[1], Claudio Greco[2], Greta Gandolfi[2], Eleonora Gualdoni[2], Raffaella Bernardi[2,3]
[1]Institute for Logic, Language and Computation, University of Amsterdam
[2]CIMeC, [3]DISI, University of Trento
s.pezzelle@uva.nl,
{greta.gandolfi|eleonora.gualdoni}@studenti.unitn.it,
{claudio.greco|raffaella.bernardi}@unitn.it

Abstract

This paper introduces BD2BB, a novel language and vision benchmark that requires multimodal models combine *complementary* information from the two modalities. Recently, impressive progress has been made to develop universal multimodal encoders suitable for virtually any language and vision tasks. However, current approaches often require them to combine *redundant* information provided by language and vision. Inspired by real-life communicative contexts, we propose a novel task where either modality is necessary but not sufficient to make a correct prediction. To do so, we first build a dataset of images and corresponding sentences provided by human participants. Second, we evaluate state-of-the-art models and compare their performance against human speakers. We show that, while the task is relatively easy for humans, best-performing models struggle to achieve similar results.

1 Introduction

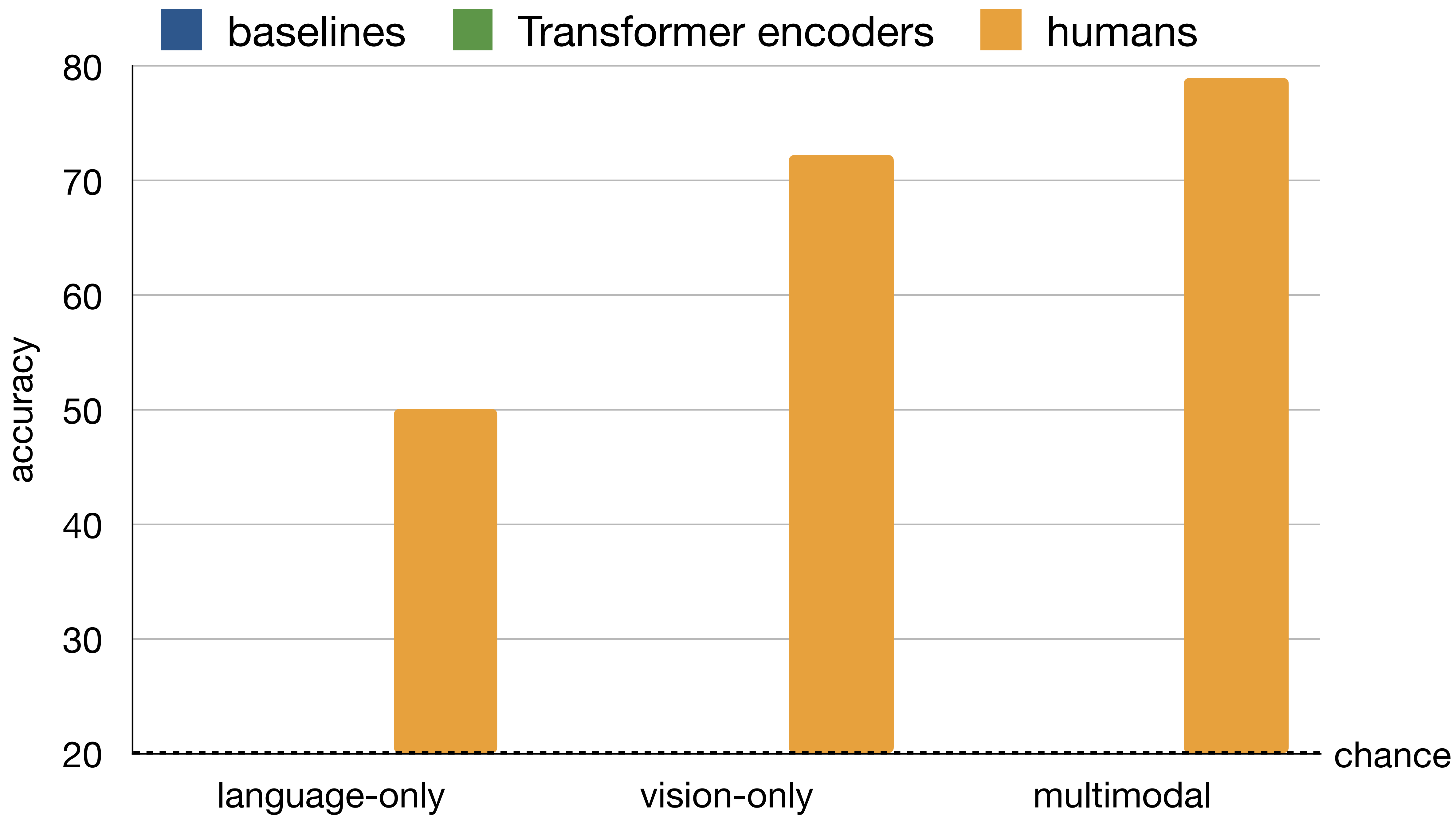
Human communication, in real-life situations, is multimodal (Kress, 2010): To convey and understand a message uttered in natural language, people build on what is present in the multimodal context surrounding them. As such, speakers do not need to "repeat" something that is already provided by the environment; similarly, listeners leverage information from various modalities, such as vision, to interpret the linguistic message. Integrating information from multiple modalities is indeed crucial for attention and perception (Partan and Marler, 1999) since combined information from concurrent modalities can give rise to different messages (McGurk and MacDonald, 1976).

roni, 2016; Beinborn et al., 2018). However, computational approaches to language and vision typically do not fully explore this complementarity. To illustrate, given an image (e.g., the one depicted in Figure 1), popular tasks involve describing it in natural language, e.g., *"A tennis player about to hit the ball"* (Image Captioning; see Bernardi et al., 2016); answering questions that are grounded in it, e.g., Q: *"What sport is he playing?"*, A: *"Tennis"* (Visual Question Answering; see Antol et al., 2015); having a dialogue on its entities, e.g., Q: *"Is the person holding a racket?"*, A: *"Yes."* (visually-grounded dialogue; see De Vries et al., 2017; Das et al., 2017). While all these tasks challenge models to perform visual grounding, i.e., an effective *alignment* of language and vision, none of them require a genuine *combination* of complementary information provided by the two modalities. All the information is fully available in the visual scene, and language is used to describe or retrieve it.

In this work, we propose a novel benchmark, *Be Different to Be Better* (in short, **BD2BB**), where the *different*, complementary information provided by the two modalities should push models develop a *better*, richer multimodal representation. As illustrated in Figure 1, models are asked to choose, among a set of **candidate actions**, the one a person who sees the visual context depicted by the **image** would do based on a certain **intention** (i.e., their goal, attitude or feeling). Crucially, the resulting multimodal input (the sum of the image and the intention) will be richer compared to that conveyed by either modality in isolation; in fact, the two modalities convey complementary or *non-redundant* information (Partan and Marler, 1999).

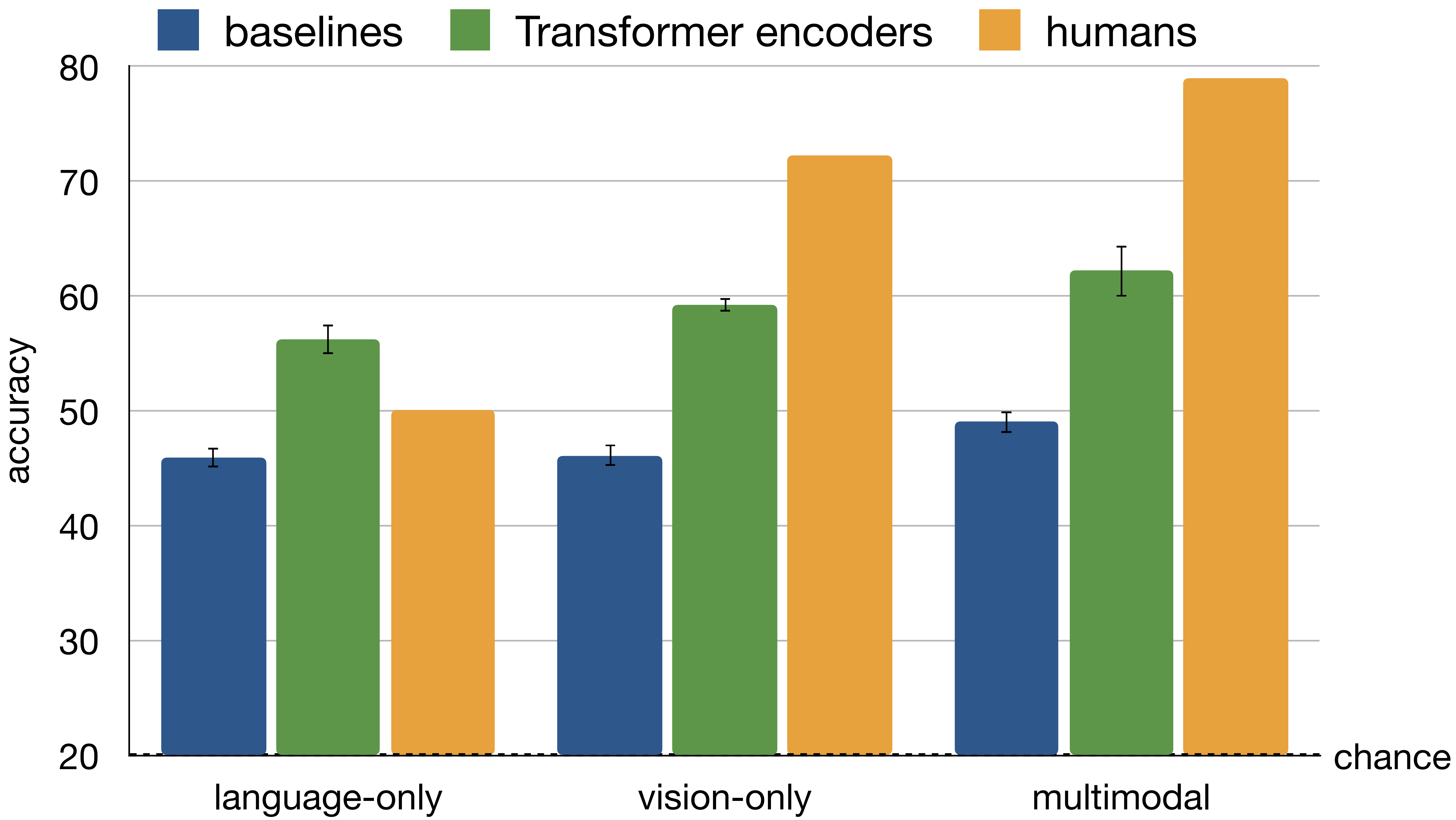
To illustrate, a model that only relies on the (non-

Results



baselines
Transformer encoders
humans
80
70
60
50
40
30
20
accuracy
chance
language-only
vision-only
multimodal

Results



baselines
Transformer encoders
humans
80
70
60
50
40
30
20
accuracy
chance
language-only
vision-only
multimodal

Results


baselines
Transformer encoders
humans
80
70
60
50
40
30
20
accuracy
chance
language-only
vision-only
multimodal

Takeaways

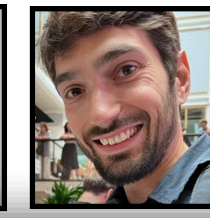
Simplified task: single-event prediction (single-event “story”)

Simplified evaluation: multiple-choice (also, no good generative VLMs in 2020)

Big gap with human performance: ~20-accuracy points!

What about generative VLMs?

Let's generate!

(EvalMG @COLING 2025)

If I feel smart, I will do the right thing: Combining Complementary Multimodal Information in Visual Language Models

Yuyu Bai
University of Amsterdam
stacybai1122@gmail.com

Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
s.pezzelle@uva.nl

Abstract

Generative visual language models (VLMs) have recently shown potential across various downstream language-and-vision tasks. At the same time, it is still an open question whether, and to what extent, these models can properly understand a multimodal context where language and vision provide *complementary* information—a mechanism routinely in place in human language communication. In this work, we test various VLMs on the task of generating action descriptions consistent with both an image's visual content and an intention or attitude (not visually grounded) conveyed by a textual prompt. Our results show that BLIP-2 is not far from human performance when the task is framed as a generative multiple-choice problem, while other models struggle. Furthermore, the actions generated by BLIP-2 in an open-ended generative setting are better than those by the competitors; indeed, human annotators judge most of them as plausible continuations for the multimodal context. Our study reveals substantial variability among VLMs in integrating complementary multimodal information, yet BLIP-2 demonstrates promising trends across most evaluations, paving the way for seamless human-computer interaction.

1 Introduction

In recent years, transformer-based generative visual language models (VLMs) have shown outstanding results in many downstream tasks. Similar to what has happened in NLP, where pre-trained generative models have supplanted previous architectures thanks to their flexibility and portability, VLMs have proven effective in solving language-and-vision tasks by turning them into generative problems. This is possible thanks to their massive

If I feel athletic…

I will…

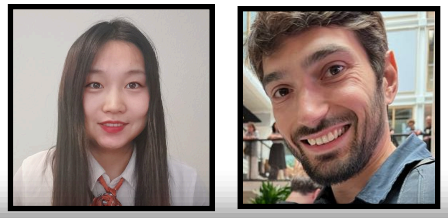
(a) stand and take a break with the baseball players ✗
(b) play baseball with friends ✓
(c) play tennis with friends ✗

Figure 1: We test generative visual language models' (VLMs) abilities to combine *complementary* information brought into context by the two modalities. In this example from the BD2BB dataset (Pezzelle et al., 2020) (slightly edited for space reasons), only one of the actions on the right, (b), is consistent with both the textual prompt and the image on the left. As for (a) and (c), they are plausible based on the image or the textual prompt, respectively, but not on the combination of both.

questions about it, or engage in a dialogue (see Caffagni et al., 2024, for an overview). This might suggest that VLMs have skills similar to those needed for meaningful multimodal communication.

In real-life multimodal communication, human speakers continuously integrate complementary information from various modalities, including language and vision, to understand and convey messages and properly act in various situations (Partan and Marler, 1999; Benoit et al., 2000; Forceville, 2020). An example of such complementarity is shown in Figure 1: If someone observing the scene depicted in the image *feels athletic*, they would likely take an action that is consistent with both the visual content and their attitude or intention, i.e., *play baseball with friends*. In contrast, actions that

Let's generate!



(EvalMG @COLING 2025)

***If I feel smart, I will do the right thing*: Combining Complementary Multimodal Information in Visual Language Models**

Yuyu Bai
University of Amsterdam
stacybai1122@gmail.com

Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
s.pezzelle@uva.nl

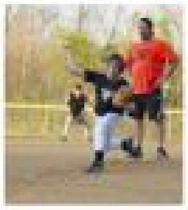
Abstract

Generative visual language models (VLMs) have recently shown potential across various downstream language-and-vision tasks. At the same time, it is still an open question whether, and to what extent, these models can properly understand a multimodal context where language and vision provide *complementary* information—a mechanism routinely in place in human language communication. In this work, we test various VLMs on the task of generating action descriptions consistent with both an image's visual content and an intention or attitude (not visually grounded) conveyed by a textual prompt. Our results show that BLIP-2 is not far from human performance when the task is framed as a generative multiple-choice problem, while other models struggle. Furthermore, the actions generated by BLIP-2 in an open-ended generative setting are better than those by the competitors; indeed, human annotators judge most of them as plausible continuations for the multimodal context. Our study reveals substantial variability among VLMs in integrating complementary multimodal information, yet BLIP-2 demonstrates promising trends across most evaluations, paving the way for seamless human-computer interaction.

1 Introduction

In recent years, transformer-based generative visual language models (VLMs) have shown outstanding results in many downstream tasks. Similar to what has happened in NLP, where pre-trained generative models have supplanted previous architectures thanks to their flexibility and portability, VLMs have proven effective in solving language-and-vision tasks by turning them into generative problems. This is possible thanks to their massive

If I feel athletic…



I will…

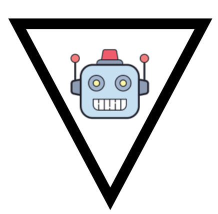
(a) stand and take a break with the baseball players ✗
(b) play baseball with friends ✓
(c) play tennis with friends ✗

Figure 1: We test generative visual language models' (VLMs) abilities to combine *complementary* information brought into context by the two modalities. In this example from the BD2BB dataset (Pezzelle et al., 2020) (slightly edited for space reasons), only one of the actions on the right, (b), is consistent with both the textual prompt and the image on the left. As for (a) and (c), they are plausible based on the image or the textual prompt, respectively, but not on the combination of both.

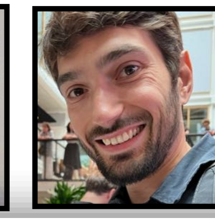
questions about it, or engage in a dialogue (see Caffagni et al., 2024, for an overview). This might suggest that VLMs have skills similar to those needed for meaningful multimodal communication.

In real-life multimodal communication, human speakers continuously integrate complementary information from various modalities, including language and vision, to understand and convey messages and properly act in various situations (Partan and Marler, 1999; Benoit et al., 2000; Forceville, 2020). An example of such complementarity is shown in Figure 1: If someone observing the scene depicted in the image *feels athletic*, they would likely take an action that is consistent with both the visual content and their attitude or intention, i.e., *play baseball with friends*. In contrast, actions that

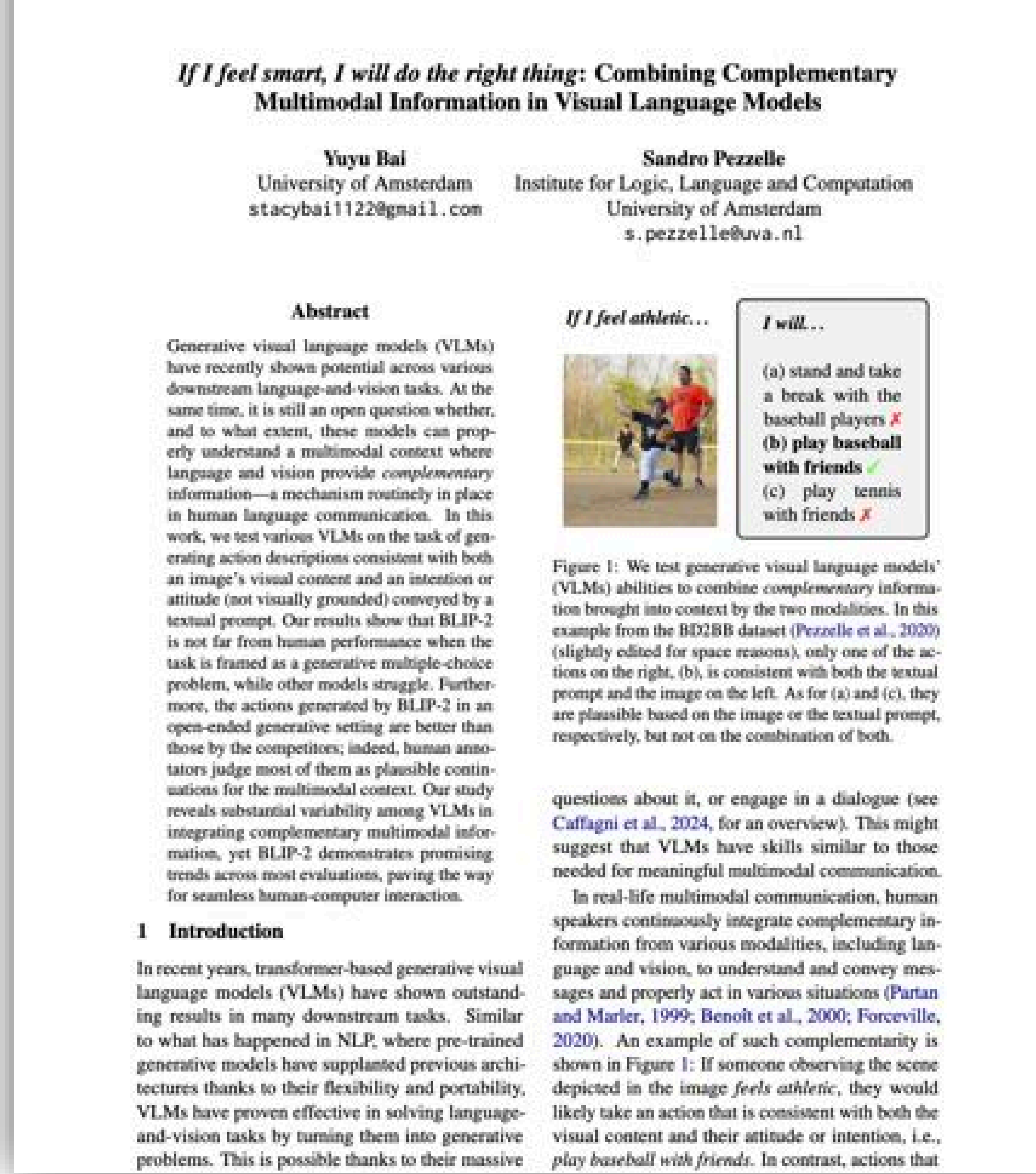
Let's generate!



`prompt:` If I have tons of energy, what will I do?
`answer:` "I will challenge the guy on the court"



(EvalMG @COLING 2025)



If I feel smart, I will do the right thing: Combining Complementary Multimodal Information in Visual Language Models

Yuyu Bai
University of Amsterdam
stacybai1122@gmail.com

Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
s.pezzelle@uva.nl

Abstract

Generative visual language models (VLMs) have recently shown potential across various downstream language-and-vision tasks. At the same time, it is still an open question whether, and to what extent, these models can properly understand a multimodal context where language and vision provide *complementary* information—a mechanism routinely in place in human language communication. In this work, we test various VLMs on the task of generating action descriptions consistent with both an image's visual content and an intention or attitude (not visually grounded) conveyed by a textual prompt. Our results show that BLIP-2 is not far from human performance when the task is framed as a generative multiple-choice problem, while other models struggle. Furthermore, the actions generated by BLIP-2 in an open-ended generative setting are better than those by the competitors; indeed, human annotators judge most of them as plausible continuations for the multimodal context. Our study reveals substantial variability among VLMs in integrating complementary multimodal information, yet BLIP-2 demonstrates promising trends across most evaluations, paving the way for seamless human-computer interaction.

1 Introduction

In recent years, transformer-based generative visual language models (VLMs) have shown outstanding results in many downstream tasks. Similar to what has happened in NLP, where pre-trained generative models have supplanted previous architectures thanks to their flexibility and portability, VLMs have proven effective in solving language-and-vision tasks by turning them into generative problems. This is possible thanks to their massive

If I feel athletic…

I will…

(a) stand and take a break with the baseball players ✗
(b) play baseball with friends ✓
(c) play tennis with friends ✗

Figure 1: We test generative visual language models' (VLMs) abilities to combine *complementary* information brought into context by the two modalities. In this example from the BD2BB dataset (Pezzelle et al., 2020) (slightly edited for space reasons), only one of the actions on the right, (b), is consistent with both the textual prompt and the image on the left. As for (a) and (c), they are plausible based on the image or the textual prompt, respectively, but not on the combination of both.

questions about it, or engage in a dialogue (see Caffagni et al., 2024, for an overview). This might suggest that VLMs have skills similar to those needed for meaningful multimodal communication.

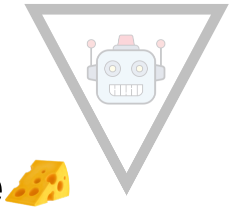
In real-life multimodal communication, human speakers continuously integrate complementary information from various modalities, including language and vision, to understand and convey messages and properly act in various situations (Partan and Marler, 1999; Benoit et al., 2000; Forceville, 2020). An example of such complementarity is shown in Figure 1: If someone observing the scene depicted in the image *feels athletic*, they would likely take an action that is consistent with both the visual content and their attitude or intention, i.e., *play baseball with friends*. In contrast, actions that

Let's generate!

IMAGE

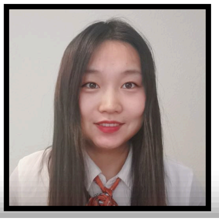
If I have tons of energy

INTENTION

MAPL FROMAGe  BLIP-2 IDEFICS

salesforce

`prompt:` If I have tons of energy, what will I do?
`answer:` "I will challenge the guy on the court"

(EvalMG @COLING 2025)

If I feel smart, I will do the right thing: Combining Complementary Multimodal Information in Visual Language Models

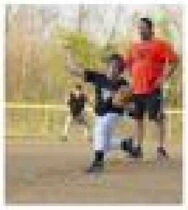
Yuyu Bai
University of Amsterdam
stacybai1122@gmail.com

Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
s.pezzelle@uva.nl

Abstract

Generative visual language models (VLMs) have recently shown potential across various downstream language-and-vision tasks. At the same time, it is still an open question whether, and to what extent, these models can properly understand a multimodal context where language and vision provide *complementary* information—a mechanism routinely in place in human language communication. In this work, we test various VLMs on the task of generating action descriptions consistent with both an image's visual content and an intention or attitude (not visually grounded) conveyed by a textual prompt. Our results show that BLIP-2 is not far from human performance when the task is framed as a generative multiple-choice problem, while other models struggle. Furthermore, the actions generated by BLIP-2 in an open-ended generative setting are better than those by the competitors; indeed, human annotators judge most of them as plausible continuations for the multimodal context. Our study reveals substantial variability among VLMs in integrating complementary multimodal information, yet BLIP-2 demonstrates promising trends across most evaluations, paving the way for seamless human-computer interaction.

If I feel athletic...



I will...

(a) stand and take a break with the baseball players ✗
(b) play baseball with friends ✓
(c) play tennis with friends ✗

Figure 1: We test generative visual language models' (VLMs) abilities to combine *complementary* information brought into context by the two modalities. In this example from the BD2BB dataset (Pezzelle et al., 2020) (slightly edited for space reasons), only one of the actions on the right, (b), is consistent with both the textual prompt and the image on the left. As for (a) and (c), they are plausible based on the image or the textual prompt, respectively, but not on the combination of both.

1 Introduction

In recent years, transformer-based generative visual language models (VLMs) have shown outstanding results in many downstream tasks. Similar to what has happened in NLP, where pre-trained generative models have supplanted previous architectures thanks to their flexibility and portability, VLMs have proven effective in solving language-and-vision tasks by turning them into generative problems. This is possible thanks to their massive questions about it, or engage in a dialogue (see Caffagni et al., 2024, for an overview). This might suggest that VLMs have skills similar to those needed for meaningful multimodal communication.

In real-life multimodal communication, human speakers continuously integrate complementary information from various modalities, including language and vision, to understand and convey messages and properly act in various situations (Partan and Marler, 1999; Benoit et al., 2000; Forceville, 2020). An example of such complementarity is shown in Figure 1: If someone observing the scene depicted in the image *feels athletic*, they would likely take an action that is consistent with both the visual content and their attitude or intention, i.e., *play baseball with friends*. In contrast, actions that

How to evaluate the generations?

1. Reference-based evaluation

2. Reference-free evaluation

Reference-based evaluation

Similarity between `generated action` and each of the five `candidates`—BLEU4, CIDER, ROUGE, Meteor, BERTScore

If `sim(generated, target)` is the highest, then ‘correct'

Accuracy-based performance

Reference-based evaluation

model	accuracy *
MAPL	32.9±8.7
FROMAGe	32.7±4.8
BLIP-2	**49.5±2.6**
IDEFICS	31.5±10.9

*Based on BERTScore (similar patterns for other metrics)

Reference-free evaluation

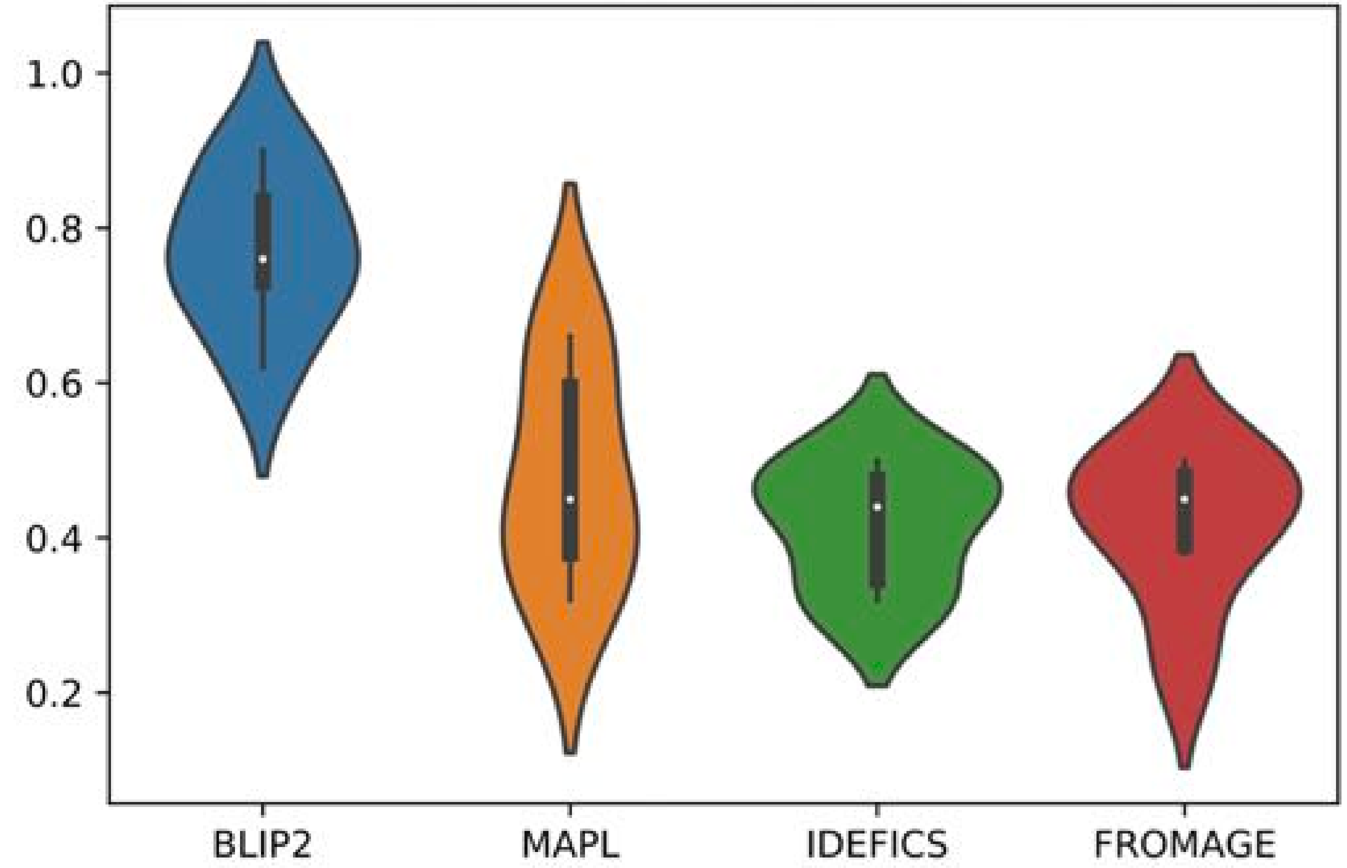
Human evaluation on a sample of `<image, intention, generated action>` datapoints (same for each model)

We asked: *“Is the second part of the sentence [the generated action] a plausible continuation of the first part, based on the contents of the image?”*

Annotators could choose between ‘Yes’ or ‘No’

Performance based on **proportion of ‘Yes’ answers**

Reference-free evaluation



BLIP-2's generations deemed plausible **77% of times**!
Other models in the 40-45% range

A negative example

If I want to socialize...

Ground-truth

I will play the Wii with my friends

BLIP-2

I will play pool with the guys ✗

A negative example: Pragmatics!

If I want to socialize…

Ground-truth
I will play the Wii with my friends

BLIP-2
I will play pool with the guys ✗

There is a pool and there are some guys! But the annotators find this action **pragmatically implausible**, as the people in the image are busy playing video games.

Takeaways

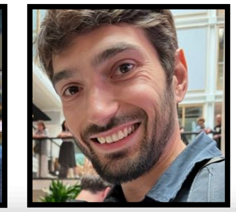
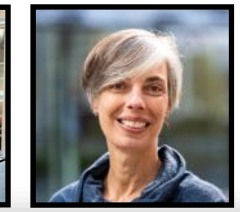
BLIP-2 is the best overall model on both evaluations

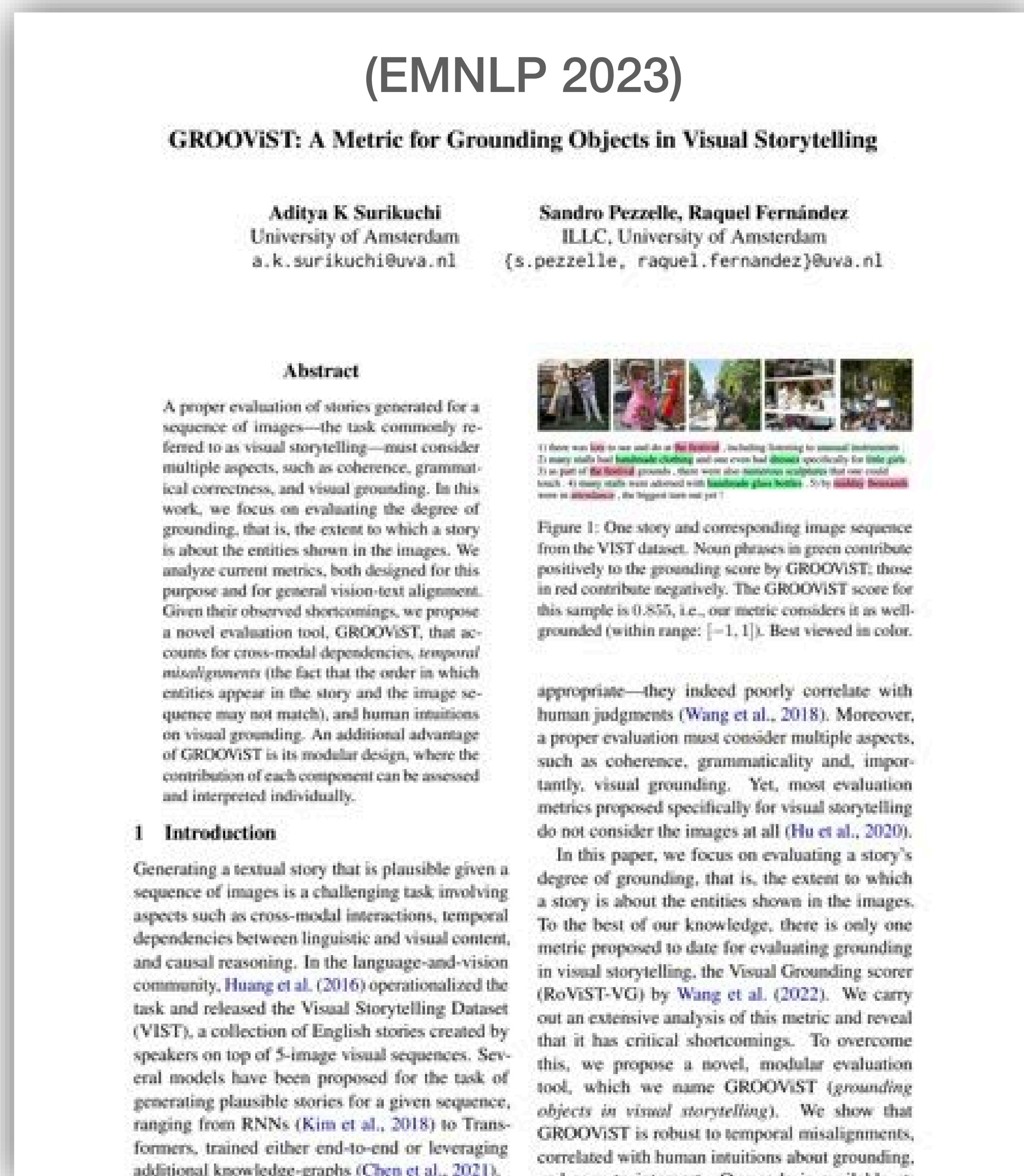
These results suggest this model can generate plausible single-image “stories”—notably, the other tested models can’t (note: BLIP-2 has seen COCO captions in training)

Also, importance of using multiple evaluations!

What about VLMs on actual stories, i.e., sequences of visual events?

Evaluate Visual Storytelling I



(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

1 Introduction

Generating a textual story that is plausible given a sequence of images is a challenging task involving aspects such as cross-modal interactions, temporal dependencies between linguistic and visual content, and causal reasoning. In the language-and-vision community, Huang et al. (2016) operationalized the task and released the Visual Storytelling Dataset (VIST), a collection of English stories created by speakers on top of 5-image visual sequences. Several models have been proposed for the task of generating plausible stories for a given sequence, ranging from RNNs (Kim et al., 2018) to Transformers, trained either end-to-end or leveraging additional knowledge-graphs (Chen et al., 2021).

Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: [−1, 1]). Best viewed in color.

appropriate—they indeed poorly correlate with human judgments (Wang et al., 2018). Moreover, a proper evaluation must consider multiple aspects, such as coherence, grammaticality and, importantly, visual grounding. Yet, most evaluation metrics proposed specifically for visual storytelling do not consider the images at all (Hu et al., 2020).

In this paper, we focus on evaluating a story's degree of grounding, that is, the extent to which a story is about the entities shown in the images. To the best of our knowledge, there is only one metric proposed to date for evaluating grounding in visual storytelling, the Visual Grounding scorer (RoViST-VG) by Wang et al. (2022). We carry out an extensive analysis of this metric and reveal that it has critical shortcomings. To overcome this, we propose a novel, modular evaluation tool, which we name GROOViST (*grounding objects in visual storytelling*). We show that GROOViST is robust to temporal misalignments, correlated with human intuitions about grounding, and easy to interpret. Our code is available at:

Evaluate Visual Storytelling I

(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

1 Introduction

Generating a textual story that is plausible given a sequence of images is a challenging task involving aspects such as cross-modal interactions, temporal dependencies between linguistic and visual content, and causal reasoning. In the language-and-vision community, Huang et al. (2016) operationalized the task and released the Visual Storytelling Dataset (VIST), a collection of English stories created by speakers on top of 5-image visual sequences. Several models have been proposed for the task of generating plausible stories for a given sequence, ranging from RNNs (Kim et al., 2018) to Transformers, trained either end-to-end or leveraging additional knowledge-graphs (Chen et al., 2021).

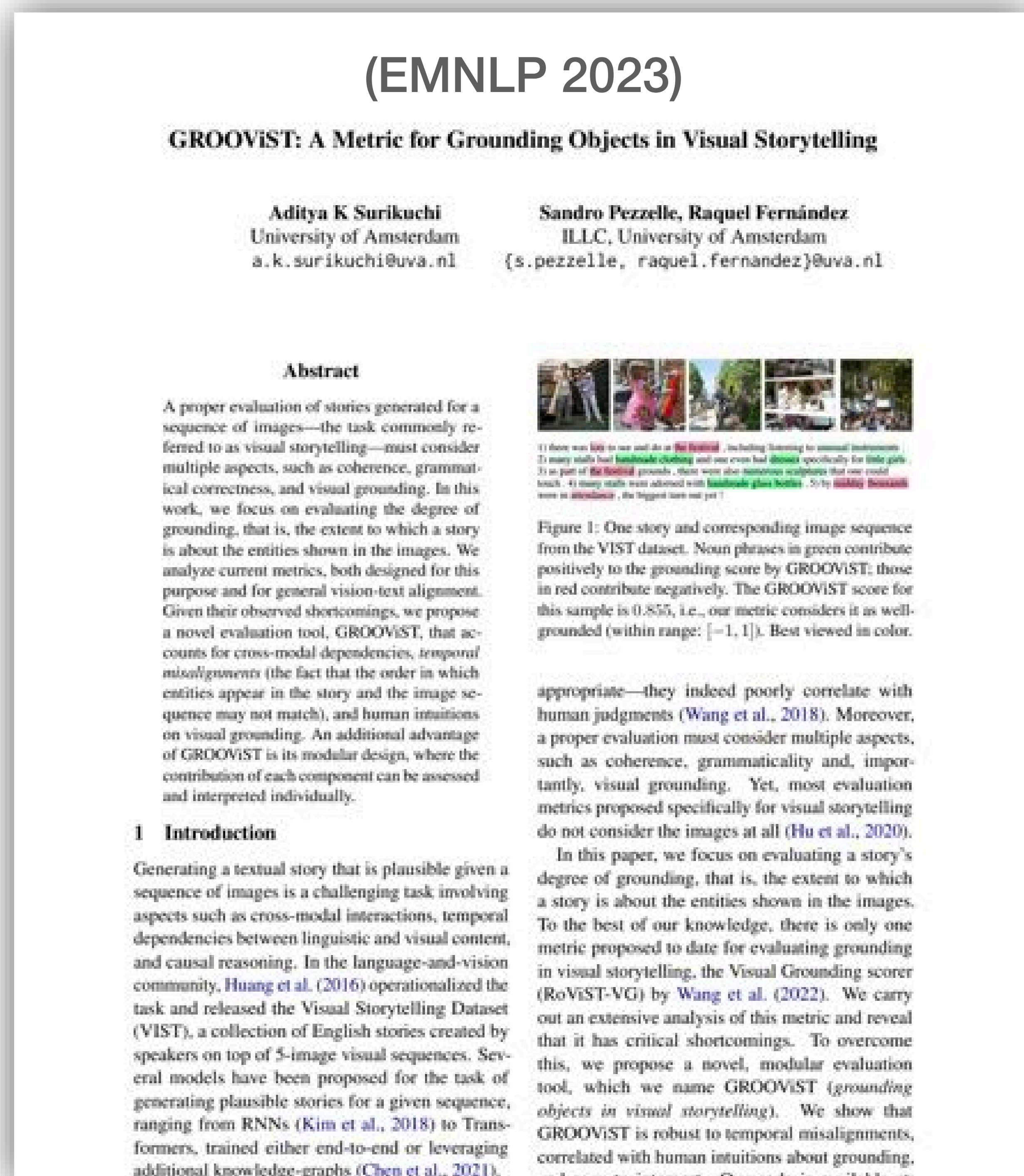
Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: $[-1, 1]$). Best viewed in color.

appropriate—they indeed poorly correlate with human judgments (Wang et al., 2018). Moreover, a proper evaluation must consider multiple aspects, such as coherence, grammaticality and, importantly, visual grounding. Yet, most evaluation metrics proposed specifically for visual storytelling do not consider the images at all (Hu et al., 2020).

In this paper, we focus on evaluating a story's degree of grounding, that is, the extent to which a story is about the entities shown in the images. To the best of our knowledge, there is only one metric proposed to date for evaluating grounding in visual storytelling, the Visual Grounding scorer (RoViST-VG) by Wang et al. (2022). We carry out an extensive analysis of this metric and reveal that it has critical shortcomings. To overcome this, we propose a novel, modular evaluation tool, which we name GROOViST (*grounding objects in visual storytelling*). We show that GROOViST is robust to temporal misalignments, correlated with human intuitions about grounding, and easy to interpret. Our code is available at:

Evaluate Visual Storytelling I

There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!



(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

1 Introduction

Generating a textual story that is plausible given a sequence of images is a challenging task involving aspects such as cross-modal interactions, temporal dependencies between linguistic and visual content, and causal reasoning. In the language-and-vision community, Huang et al. (2016) operationalized the task and released the Visual Storytelling Dataset (VIST), a collection of English stories created by speakers on top of 5-image visual sequences. Several models have been proposed for the task of generating plausible stories for a given sequence, ranging from RNNs (Kim et al., 2018) to Transformers, trained either end-to-end or leveraging additional knowledge-graphs (Chen et al., 2021).

Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: [−1, 1]). Best viewed in color.

appropriate—they indeed poorly correlate with human judgments (Wang et al., 2018). Moreover, a proper evaluation must consider multiple aspects, such as coherence, grammaticality and, importantly, visual grounding. Yet, most evaluation metrics proposed specifically for visual storytelling do not consider the images at all (Hu et al., 2020).

In this paper, we focus on evaluating a story's degree of grounding, that is, the extent to which a story is about the entities shown in the images. To the best of our knowledge, there is only one metric proposed to date for evaluating grounding in visual storytelling, the Visual Grounding scorer (RoViST-VG) by Wang et al. (2022). We carry out an extensive analysis of this metric and reveal that it has critical shortcomings. To overcome this, we propose a novel, modular evaluation tool, which we name GROOViST (*grounding objects in visual storytelling*). We show that GROOViST is robust to temporal misalignments, correlated with human intuitions about grounding, and easy to interpret. Our code is available at:

Evaluate Visual Storytelling I

There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!

Dimensions to evaluate in a story:

(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

1 Introduction

Generating a textual story that is plausible given a sequence of images is a challenging task involving aspects such as cross-modal interactions, temporal dependencies between linguistic and visual content, and causal reasoning. In the language-and-vision community, Huang et al. (2016) operationalized the task and released the Visual Storytelling Dataset (VIST), a collection of English stories created by speakers on top of 5-image visual sequences. Several models have been proposed for the task of generating plausible stories for a given sequence, ranging from RNNs (Kim et al., 2018) to Transformers, trained either end-to-end or leveraging additional knowledge-graphs (Chen et al., 2021).

Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: [−1, 1]). Best viewed in color.

appropriate—they indeed poorly correlate with human judgments (Wang et al., 2018). Moreover, a proper evaluation must consider multiple aspects, such as coherence, grammaticality and, importantly, visual grounding. Yet, most evaluation metrics proposed specifically for visual storytelling do not consider the images at all (Hu et al., 2020).

In this paper, we focus on evaluating a story's degree of grounding, that is, the extent to which a story is about the entities shown in the images. To the best of our knowledge, there is only one metric proposed to date for evaluating grounding in visual storytelling, the Visual Grounding scorer (RoViST-VG) by Wang et al. (2022). We carry out an extensive analysis of this metric and reveal that it has critical shortcomings. To overcome this, we propose a novel, modular evaluation tool, which we name GROOViST (*grounding objects in visual storytelling*). We show that GROOViST is robust to temporal misalignments, correlated with human intuitions about grounding, and easy to interpret. Our code is available at:

Evaluate Visual Storytelling I

There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!

Dimensions to evaluate in a story:

- new information

(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

1 Introduction

Generating a textual story that is plausible given a sequence of images is a challenging task involving aspects such as cross-modal interactions, temporal dependencies between linguistic and visual content, and causal reasoning. In the language-and-vision community, Huang et al. (2016) operationalized the task and released the Visual Storytelling Dataset (VIST), a collection of English stories created by speakers on top of 5-image visual sequences. Several models have been proposed for the task of generating plausible stories for a given sequence, ranging from RNNs (Kim et al., 2018) to Transformers, trained either end-to-end or leveraging additional knowledge-graphs (Chen et al., 2021).

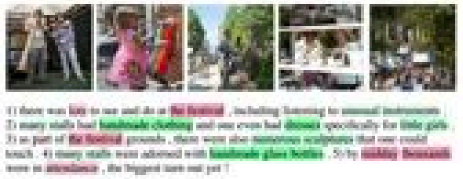
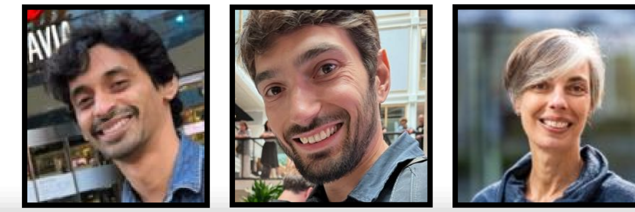


Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: $[-1, 1]$). Best viewed in color.

appropriate—they indeed poorly correlate with human judgments (Wang et al., 2018). Moreover, a proper evaluation must consider multiple aspects, such as coherence, grammaticality and, importantly, visual grounding. Yet, most evaluation metrics proposed specifically for visual storytelling do not consider the images at all (Hu et al., 2020).

In this paper, we focus on evaluating a story's degree of grounding, that is, the extent to which a story is about the entities shown in the images. To the best of our knowledge, there is only one metric proposed to date for evaluating grounding in visual storytelling, the Visual Grounding scorer (RoViST-VG) by Wang et al. (2022). We carry out an extensive analysis of this metric and reveal that it has critical shortcomings. To overcome this, we propose a novel, modular evaluation tool, which we name GROOViST (*grounding objects in visual storytelling*). We show that GROOViST is robust to temporal misalignments, correlated with human intuitions about grounding, and easy to interpret. Our code is available at:

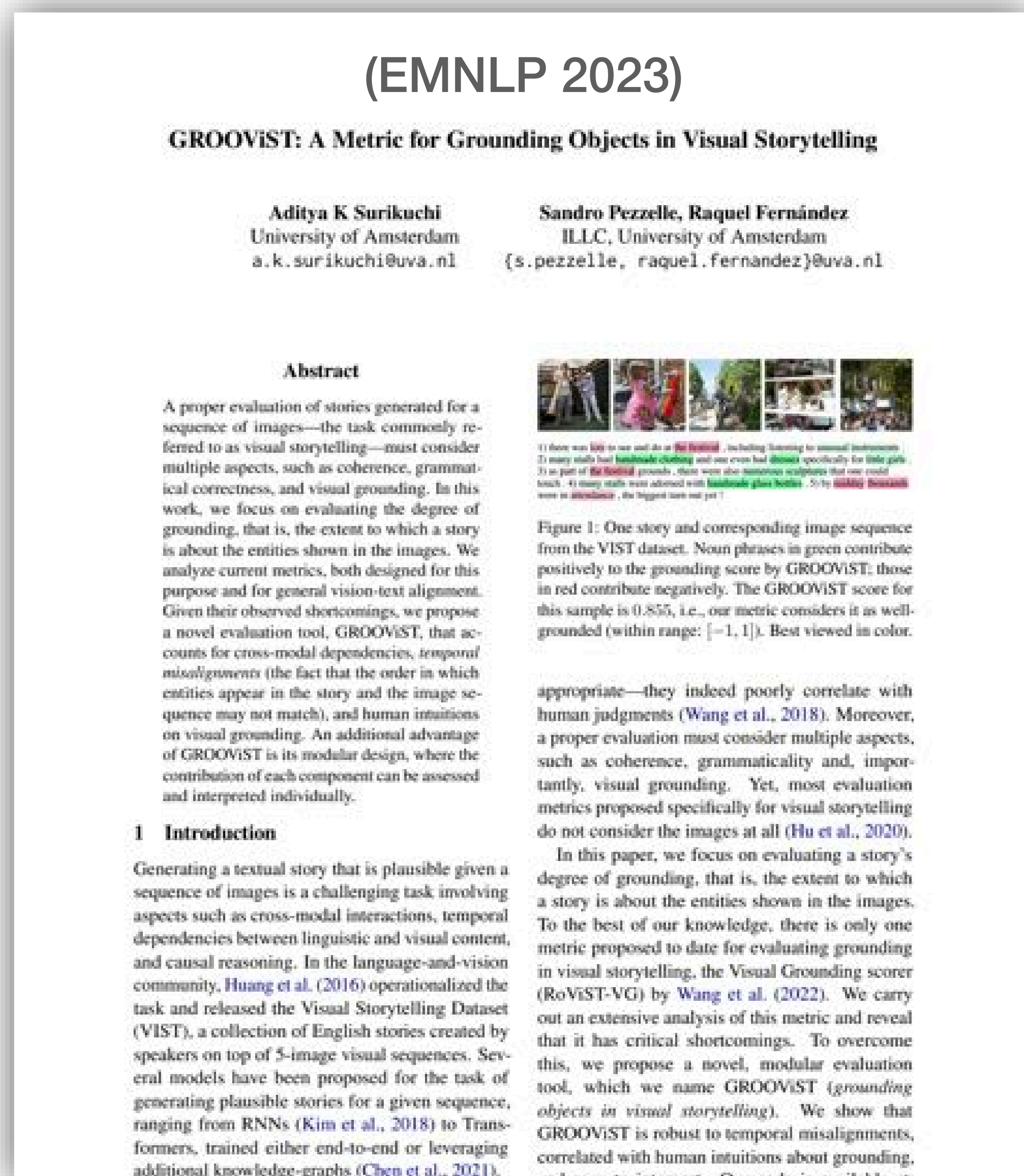
Evaluate Visual Storytelling I



There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!

Dimensions to evaluate in a story:

- new information
- coherence



(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

1 Introduction

Generating a textual story that is plausible given a sequence of images is a challenging task involving aspects such as cross-modal interactions, temporal dependencies between linguistic and visual content, and causal reasoning. In the language-and-vision community, Huang et al. (2016) operationalized the task and released the Visual Storytelling Dataset (VIST), a collection of English stories created by speakers on top of 5-image visual sequences. Several models have been proposed for the task of generating plausible stories for a given sequence, ranging from RNNs (Kim et al., 2018) to Transformers, trained either end-to-end or leveraging additional knowledge-graphs (Chen et al., 2021).

Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: [−1, 1]). Best viewed in color.

appropriate—they indeed poorly correlate with human judgments (Wang et al., 2018). Moreover, a proper evaluation must consider multiple aspects, such as coherence, grammaticality and, importantly, visual grounding. Yet, most evaluation metrics proposed specifically for visual storytelling do not consider the images at all (Hu et al., 2020).

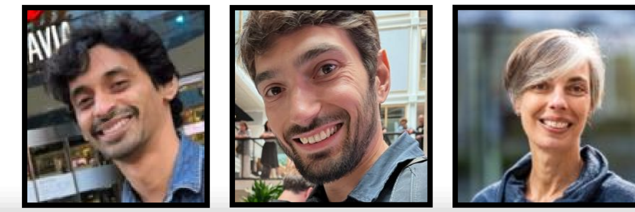
In this paper, we focus on evaluating a story's degree of grounding, that is, the extent to which a story is about the entities shown in the images. To the best of our knowledge, there is only one metric proposed to date for evaluating grounding in visual storytelling, the Visual Grounding scorer (RoViST-VG) by Wang et al. (2022). We carry out an extensive analysis of this metric and reveal that it has critical shortcomings. To overcome this, we propose a novel, modular evaluation tool, which we name GROOViST (*grounding objects in visual storytelling*). We show that GROOViST is robust to temporal misalignments, correlated with human intuitions about grounding, and easy to interpret. Our code is available at:

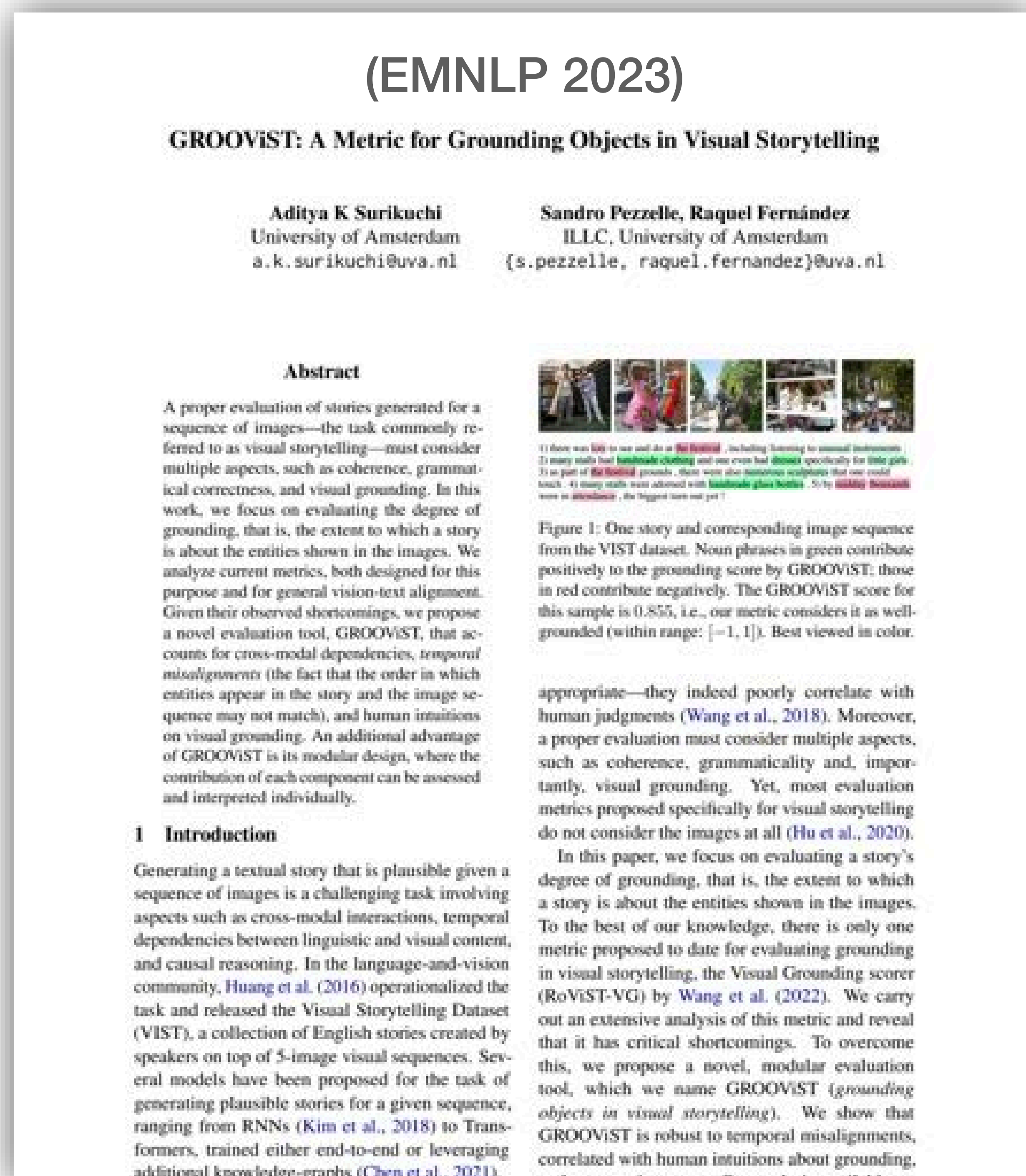
Visual Grounding

There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!

Dimensions to evaluate in a story:

- new information
- coherence
- **visual grounding**





(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

1 Introduction

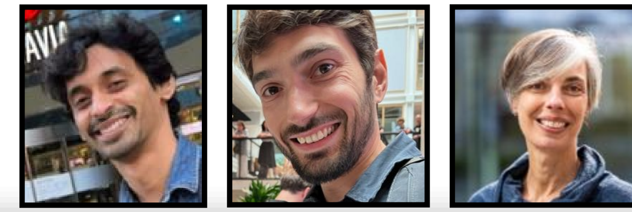
Generating a textual story that is plausible given a sequence of images is a challenging task involving aspects such as cross-modal interactions, temporal dependencies between linguistic and visual content, and causal reasoning. In the language-and-vision community, Huang et al. (2016) operationalized the task and released the Visual Storytelling Dataset (VIST), a collection of English stories created by speakers on top of 5-image visual sequences. Several models have been proposed for the task of generating plausible stories for a given sequence, ranging from RNNs (Kim et al., 2018) to Transformers, trained either end-to-end or leveraging additional knowledge-graphs (Chen et al., 2021).

Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: [−1, 1]). Best viewed in color.

appropriate—they indeed poorly correlate with human judgments (Wang et al., 2018). Moreover, a proper evaluation must consider multiple aspects, such as coherence, grammaticality and, importantly, visual grounding. Yet, most evaluation metrics proposed specifically for visual storytelling do not consider the images at all (Hu et al., 2020).

In this paper, we focus on evaluating a story's degree of grounding, that is, the extent to which a story is about the entities shown in the images. To the best of our knowledge, there is only one metric proposed to date for evaluating grounding in visual storytelling, the Visual Grounding scorer (RoViST-VG) by Wang et al. (2022). We carry out an extensive analysis of this metric and reveal that it has critical shortcomings. To overcome this, we propose a novel, modular evaluation tool, which we name GROOViST (*grounding objects in visual storytelling*). We show that GROOViST is robust to temporal misalignments, correlated with human intuitions about grounding, and easy to interpret. Our code is available at:

Visual Grounding



There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!

Dimensions to evaluate in a story:

- new information
- coherence
- **visual grounding**

(EMNLP 2023)

GROOViST: A Metric for Grounding Objects in Visual Storytelling

Aditya K Surikuchi
University of Amsterdam
a.k.surikuchi@uva.nl

Sandro Pezzelle, Raquel Fernández
ILLC, University of Amsterdam
{s.pezzelle, raquel.fernandez}@uva.nl

Abstract

A proper evaluation of stories generated for a sequence of images—the task commonly referred to as visual storytelling—must consider multiple aspects, such as coherence, grammatical correctness, and visual grounding. In this work, we focus on evaluating the degree of grounding, that is, the extent to which a story is about the entities shown in the images. We analyze current metrics, both designed for this purpose and for general vision-text alignment. Given their observed shortcomings, we propose a novel evaluation tool, GROOViST, that accounts for cross-modal dependencies, *temporal misalignments* (the fact that the order in which entities appear in the story and the image sequence may not match), and human intuitions on visual grounding. An additional advantage of GROOViST is its modular design, where the contribution of each component can be assessed and interpreted individually.

1 Introduction

Generating a textual story that is plausible given a sequence of images is a challenging task involving aspects such as cross-modal interactions, temporal dependencies between linguistic and visual content, and causal reasoning. In the language-and-vision community, Huang et al. (2016) operationalized the task and released the Visual Storytelling Dataset (VIST), a collection of English stories created by speakers on top of 5-image visual sequences. Several models have been proposed for the task of generating plausible stories for a given sequence, ranging from RNNs (Kim et al., 2018) to Transformers, trained either end-to-end or leveraging additional knowledge-graphs (Chen et al., 2021).

Figure 1: One story and corresponding image sequence from the VIST dataset. Noun phrases in green contribute positively to the grounding score by GROOViST; those in red contribute negatively. The GROOViST score for this sample is 0.855, i.e., our metric considers it as well-grounded (within range: $[-1, 1]$). Best viewed in color.

appropriate—they indeed poorly correlate with human judgments (Wang et al., 2018). Moreover, a proper evaluation must consider multiple aspects, such as coherence, grammaticality and, importantly, visual grounding. Yet, most evaluation metrics proposed specifically for visual storytelling do not consider the images at all (Hu et al., 2020).

In this paper, we focus on evaluating a story's degree of grounding, that is, the extent to which a story is about the entities shown in the images. To the best of our knowledge, there is only one metric proposed to date for evaluating grounding in visual storytelling, the Visual Grounding scorer (RoViST-VG) by Wang et al. (2022). We carry out an extensive analysis of this metric and reveal that it has critical shortcomings. To overcome this, we propose a novel, modular evaluation tool, which we name GROOViST (*grounding objects in visual storytelling*). We show that GROOViST is robust to temporal misalignments, correlated with human intuitions about grounding, and easy to interpret. Our code is available at:

Visual Grounding in Visual Storytelling

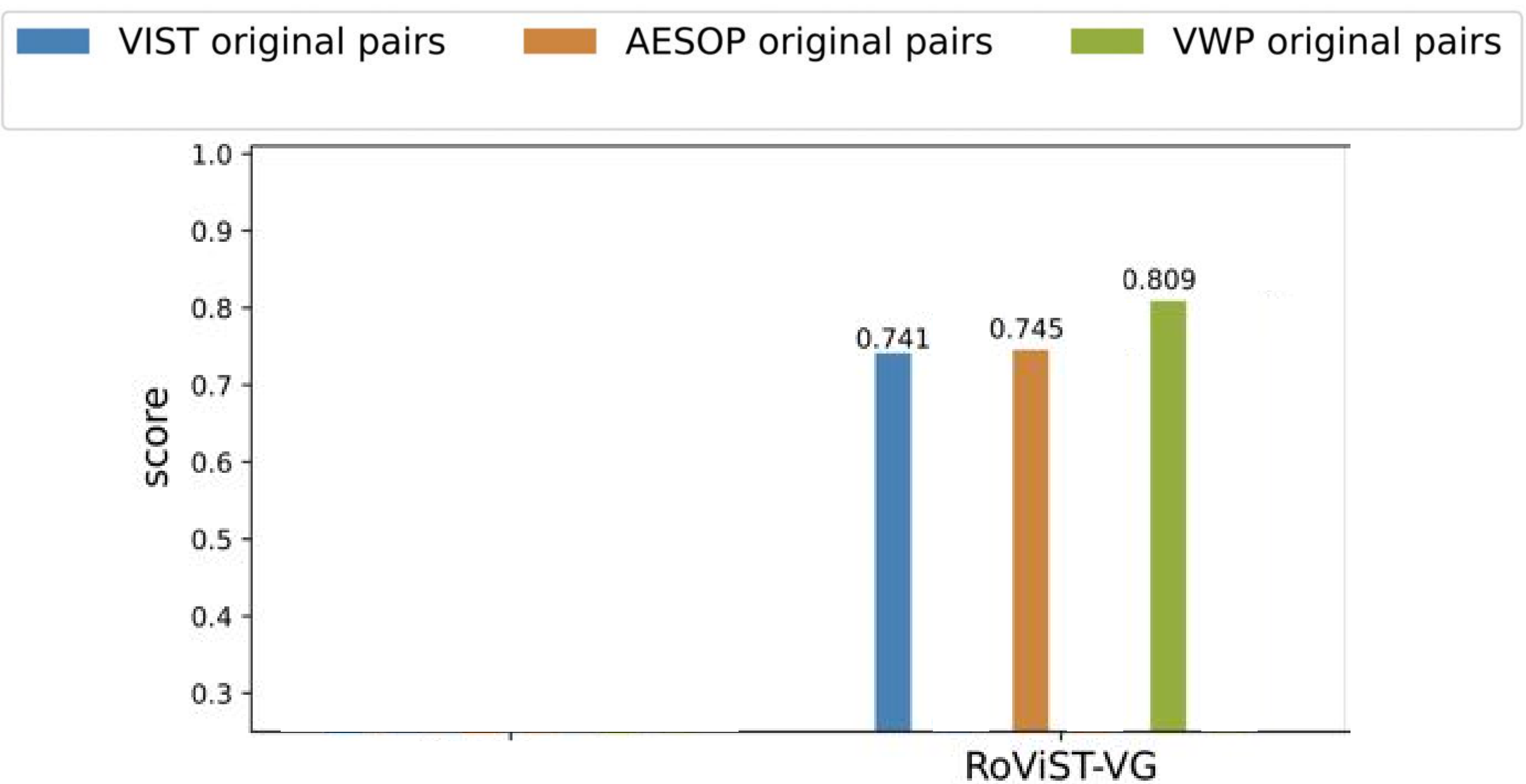
There is just one metric, **RoViST-VG** [Wang et al. 2022], trained to map nouns to an image's regions in Flickr30K Entities

However, we noticed something odd with it (on 3 VST datasets):

Visual Grounding in Visual Storytelling

There is just one metric, **RoViST-VG** [Wang et al. 2022], trained to map nouns to an image's regions in Flickr30K Entities

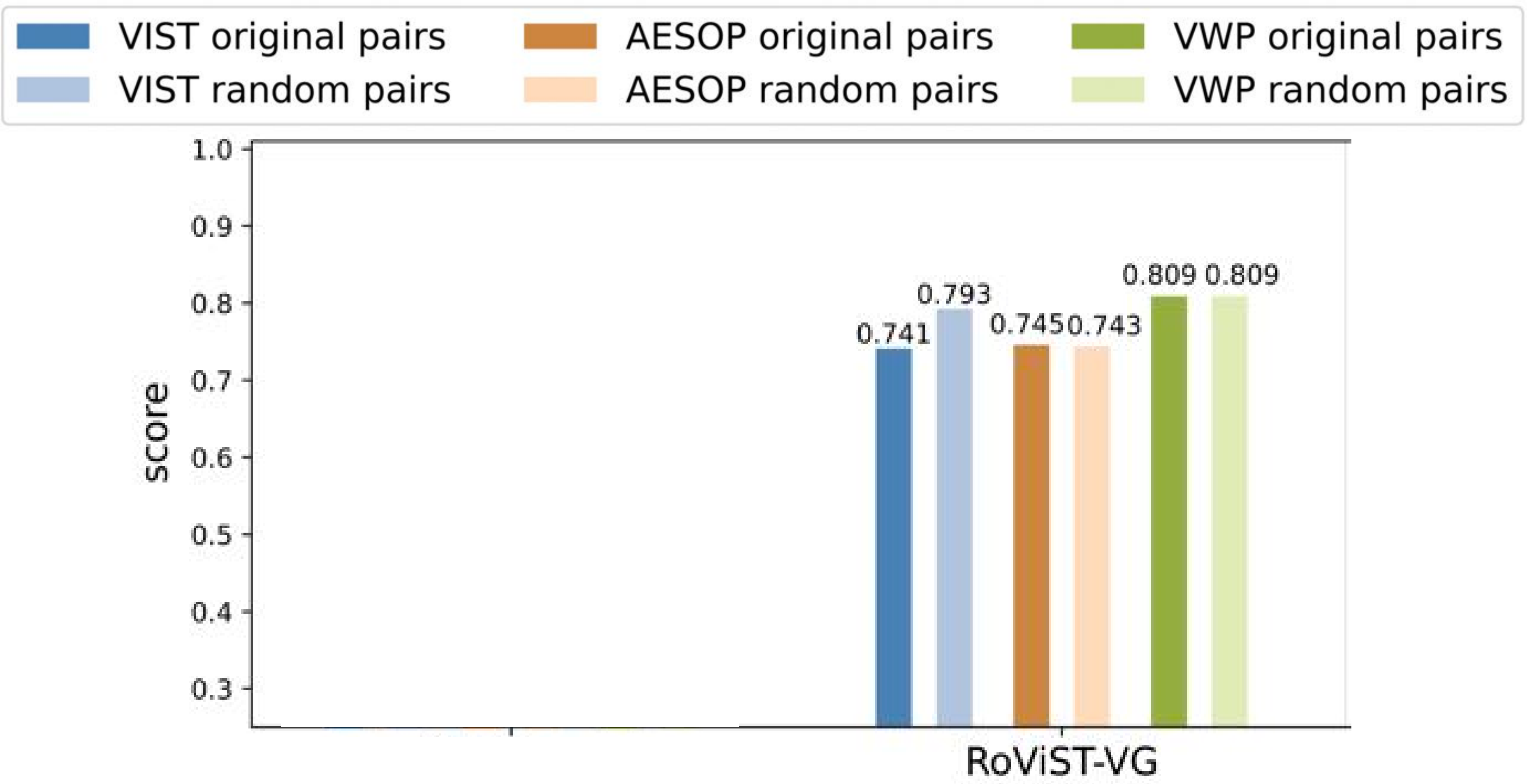
However, we noticed something odd with it (on 3 VST datasets):



Visual Grounding in Visual Storytelling

There is just one metric, **RoViST-VG** [Wang et al. 2022], trained to map nouns to an image's regions in Flickr30K Entities

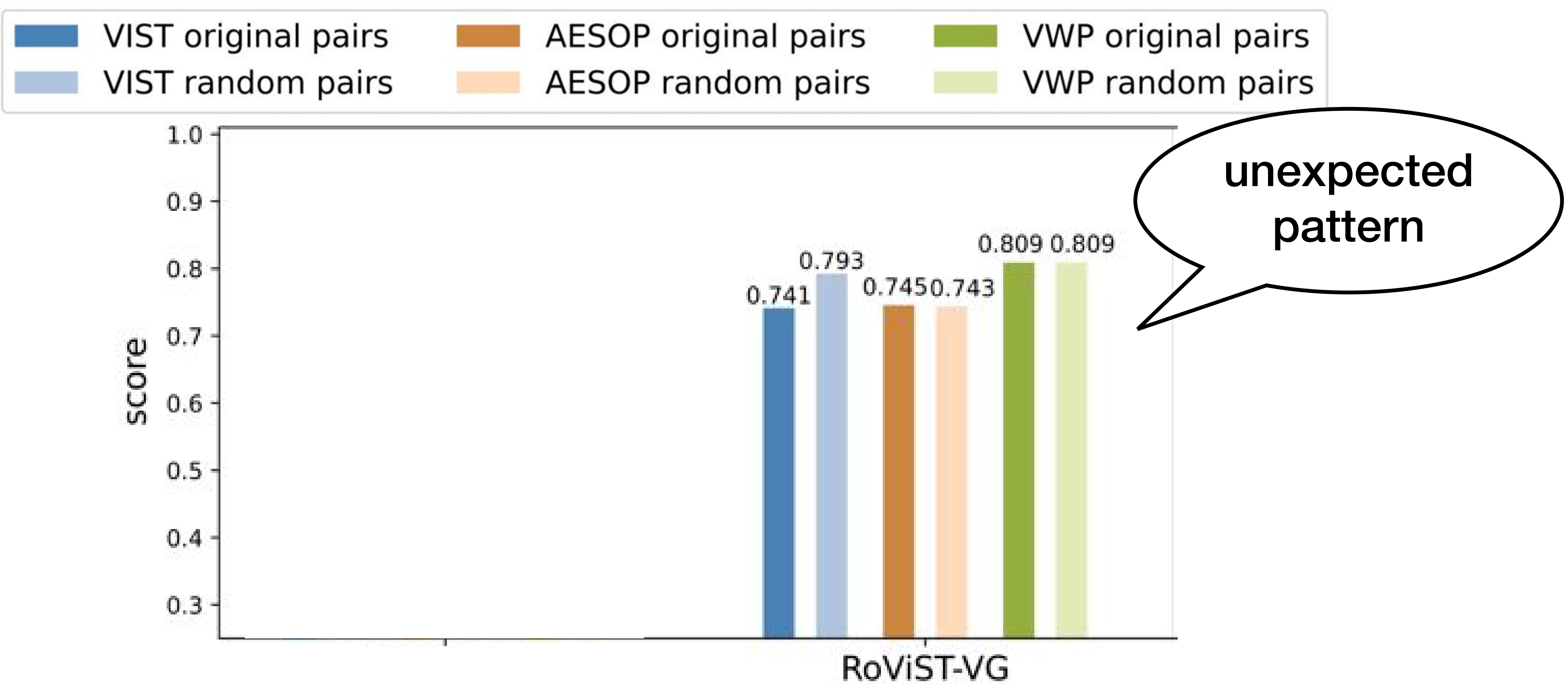
However, we noticed something odd with it (on 3 VST datasets):



Visual Grounding in Visual Storytelling

There is just one metric, **RoViST-VG** [Wang et al. 2022], trained to map nouns to an image's regions in Flickr30K Entities

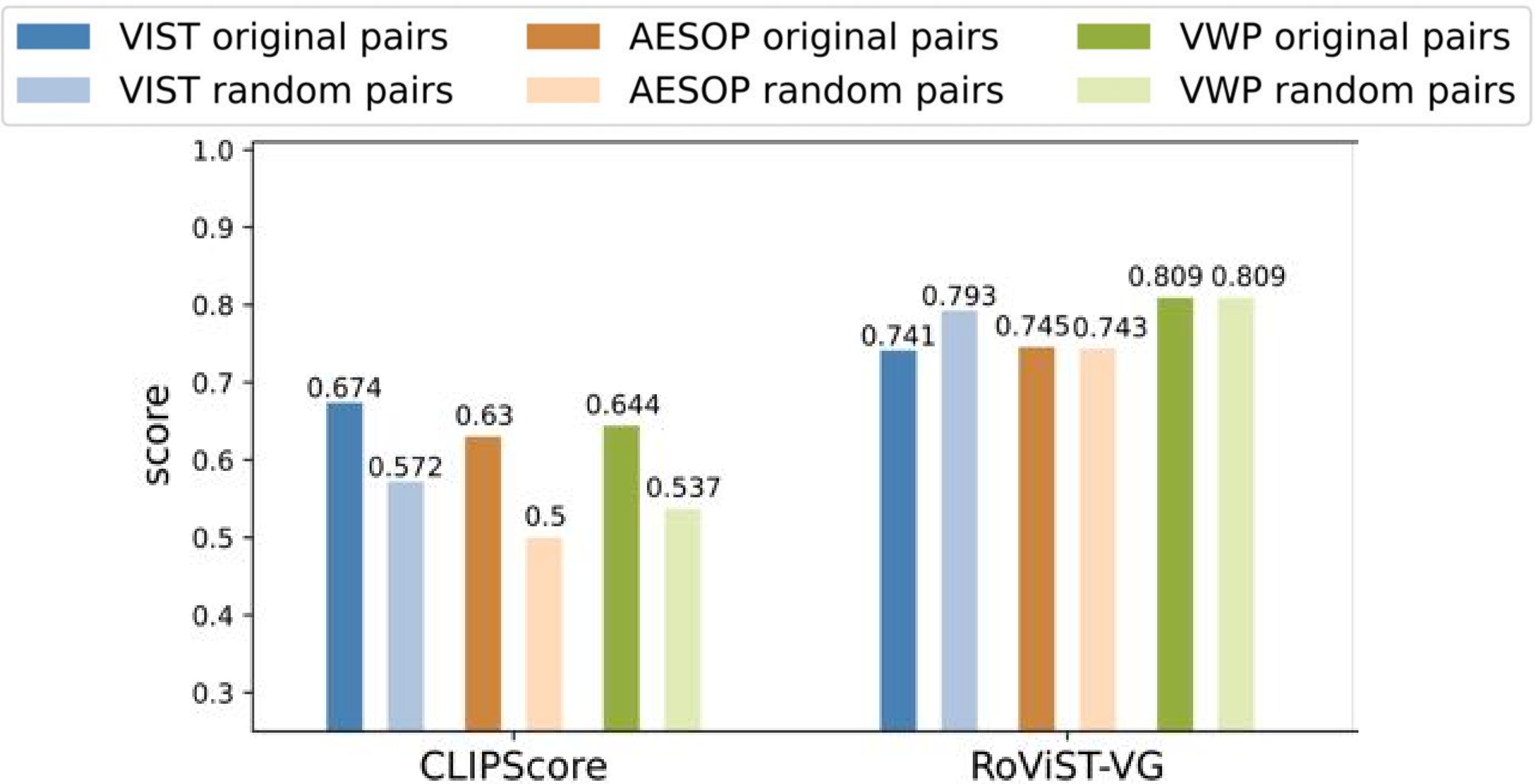
However, we noticed something odd with it (on 3 VST datasets):



Visual Grounding in Visual Storytelling

There is just one metric, **RoViST-VG** [Wang et al. 2022], trained to map nouns to an image's regions in Flickr30K Entities

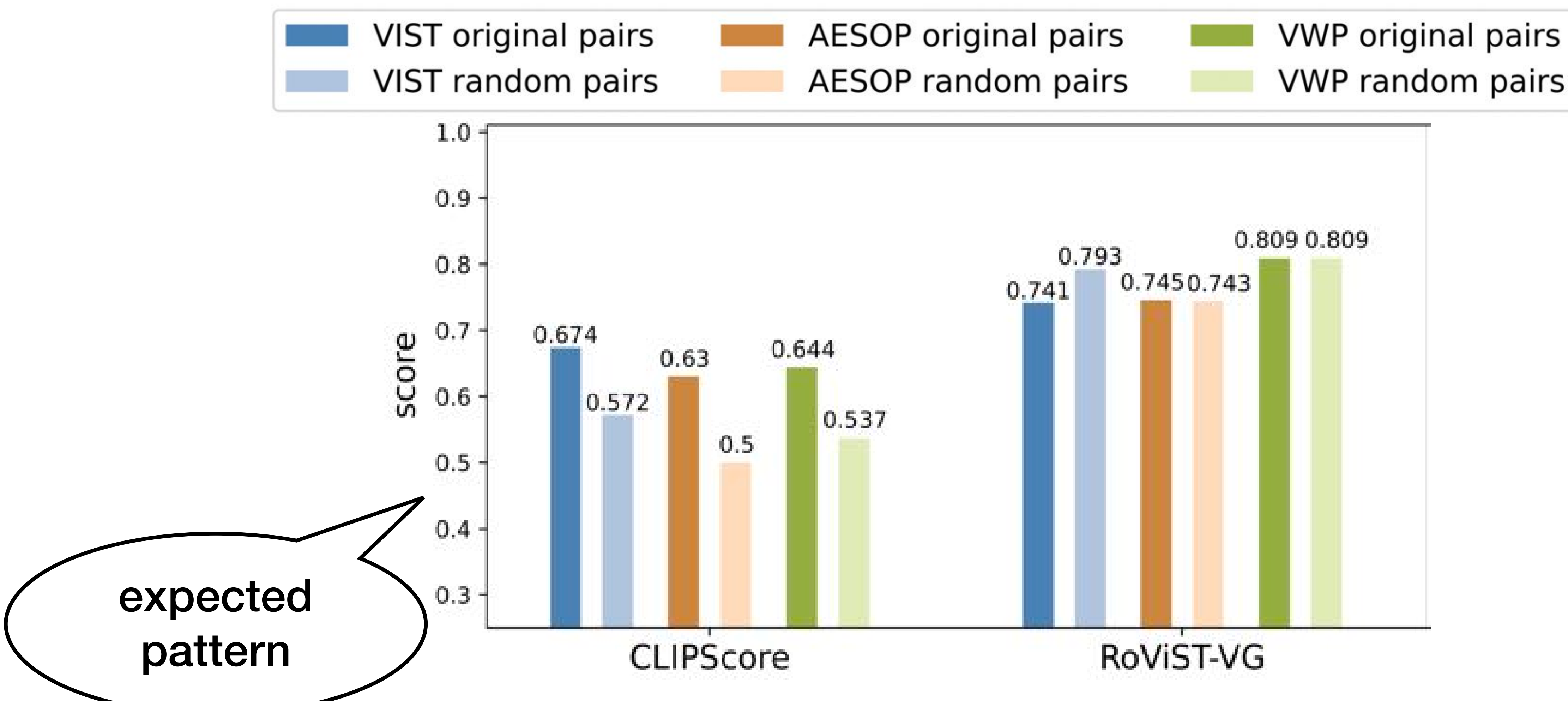
However, we noticed something odd with it (on 3 VST datasets):



Visual Grounding in Visual Storytelling

There is just one metric, **RoViST-VG** [Wang et al. 2022], trained to map nouns to an image's regions in Flickr30K Entities

However, we noticed something odd with it (on 3 VST datasets):



New metric for Visual Grounding

CLIPScore exhibits the expected pattern, but:

- General image-text alignment score: here computed at the individual <sentence, image> level—not at the story level!
- **Unable to capture temporal misalignments** regarding ‘where things are shown’ vs ‘where things are mentioned’
- We propose **a novel metric based on human stories!**

GROOViST

More details in the paper!

- We process a story and extract **NPs** (RoViST-VG uses nouns)
- **Image-text alignment scores** between each NPs and each bounding box (BB) in the sequence w/ CLIP; select highest
- **Penalizing poorly aligned** <NP, BB> pairs based on avg.— they'll contribute negatively to the overall metric

GROOViST

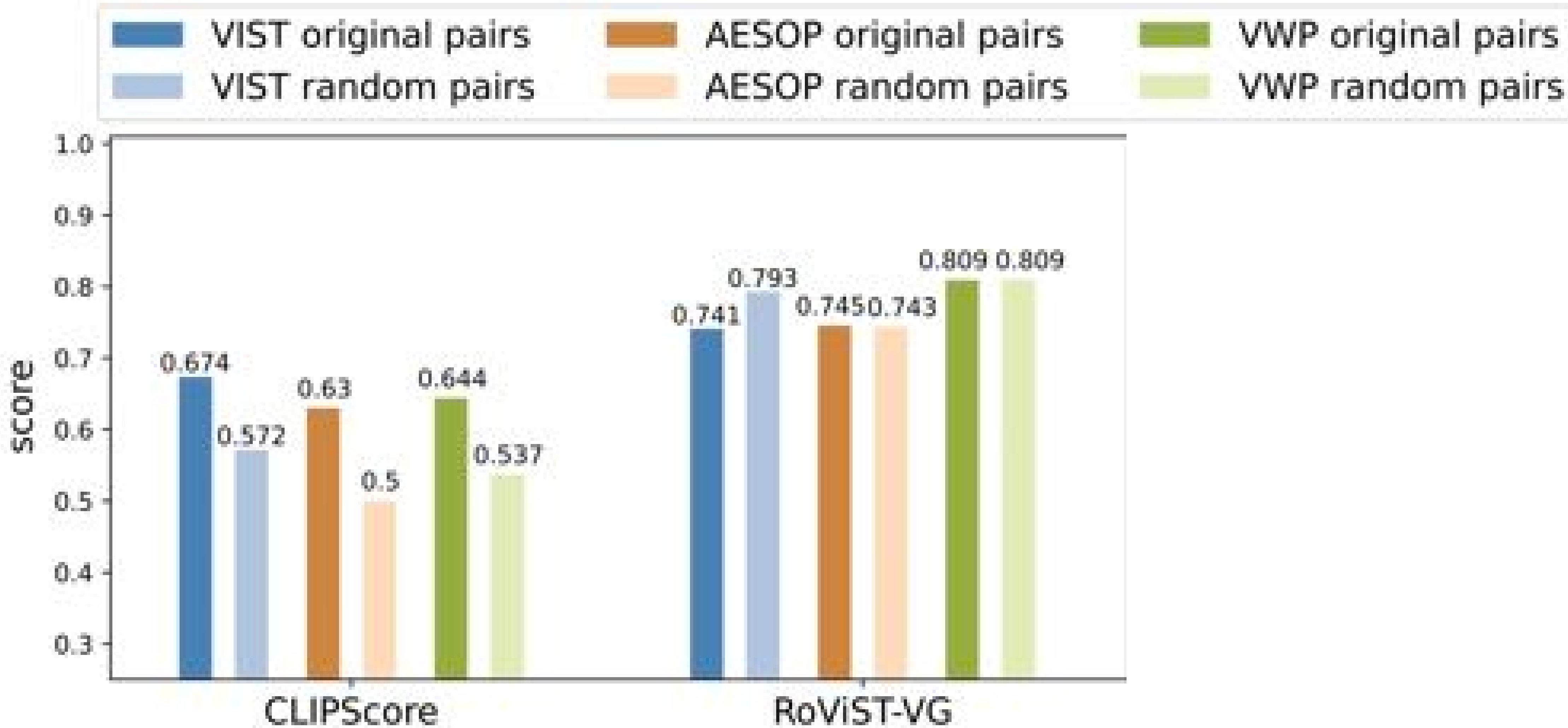
More details in the paper!

- We process a story and extract **NPs** (RoViST-VG uses nouns)
- **Image-text alignment scores** between each NPs and each bounding box (BB) in the sequence w/ CLIP; select highest
- **Penalizing poorly aligned** <NP, BB> pairs based on avg.— they'll contribute negatively to the overall metric

$$\left(\sum_{i=1}^{n} \mathrm{NP}_{pos_i} + \sum_{i=1}^{m} \mathrm{NP}_{neg_i}\right) / \, (n+m)$$

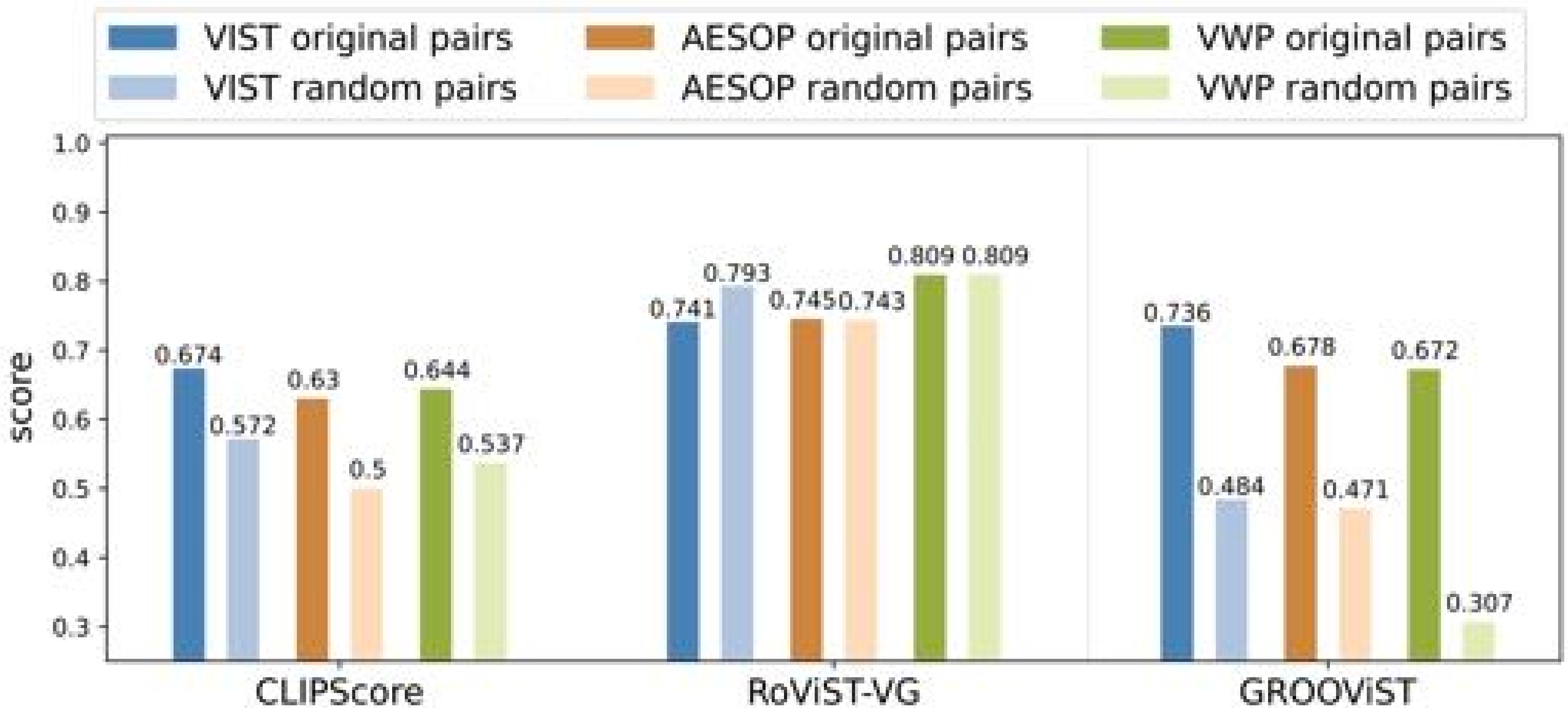
Results

Results

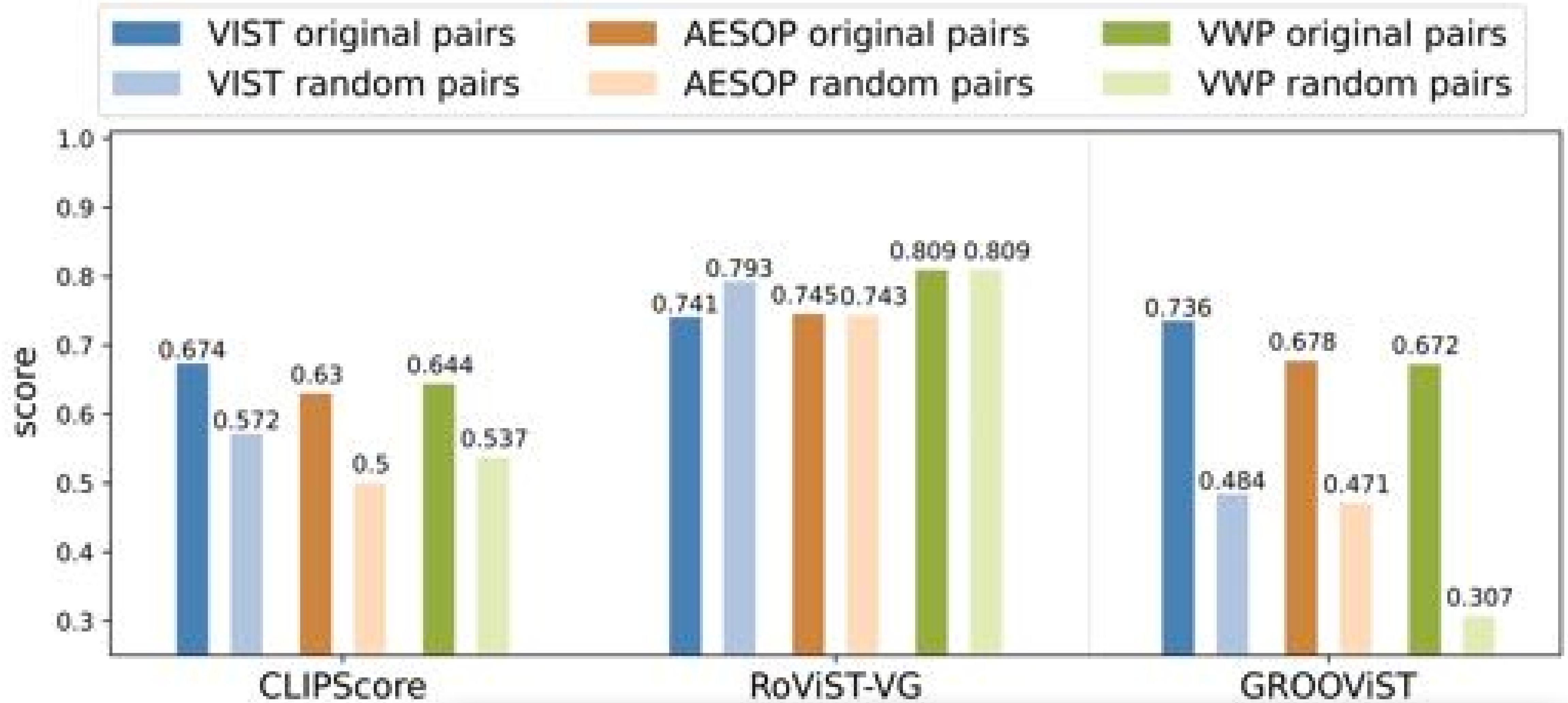


VIST original pairs
VIST random pairs
AESOP original pairs
AESOP random pairs
VWP original pairs
VWP random pairs
score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.674
0.572
0.63
0.5
0.644
0.537
0.741
0.793
0.745
0.743
0.809
0.809
CLIPScore
RoViST-VG

Results


VIST original pairs
VIST random pairs
AESOP original pairs
AESOP random pairs
VWP original pairs
VWP random pairs
score
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.674
0.572
0.63
0.5
0.644
0.537
0.741
0.793
0.745
0.743
0.809
0.809
0.736
0.484
0.678
0.471
0.672
0.307
CLIPScore
RoViST-VG
GROOViST

Results

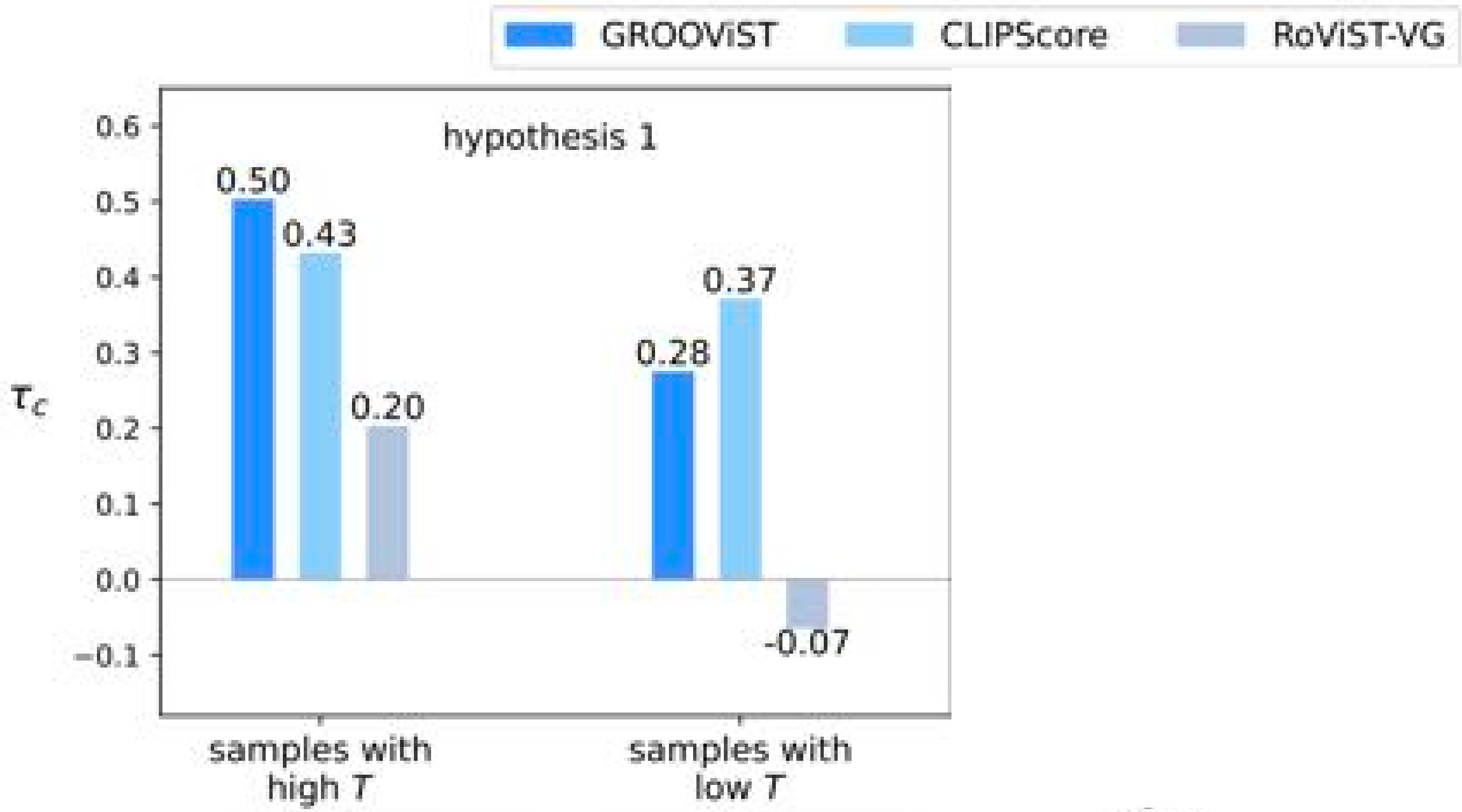


Correlation w/ image-caption grounding in Flickr8kExpert

	RoViST-VG	CLIPScore	GROOViST
τ_c (p-value)	0.019 (0.125)	0.556 (0.0)	0.414 (0.0)

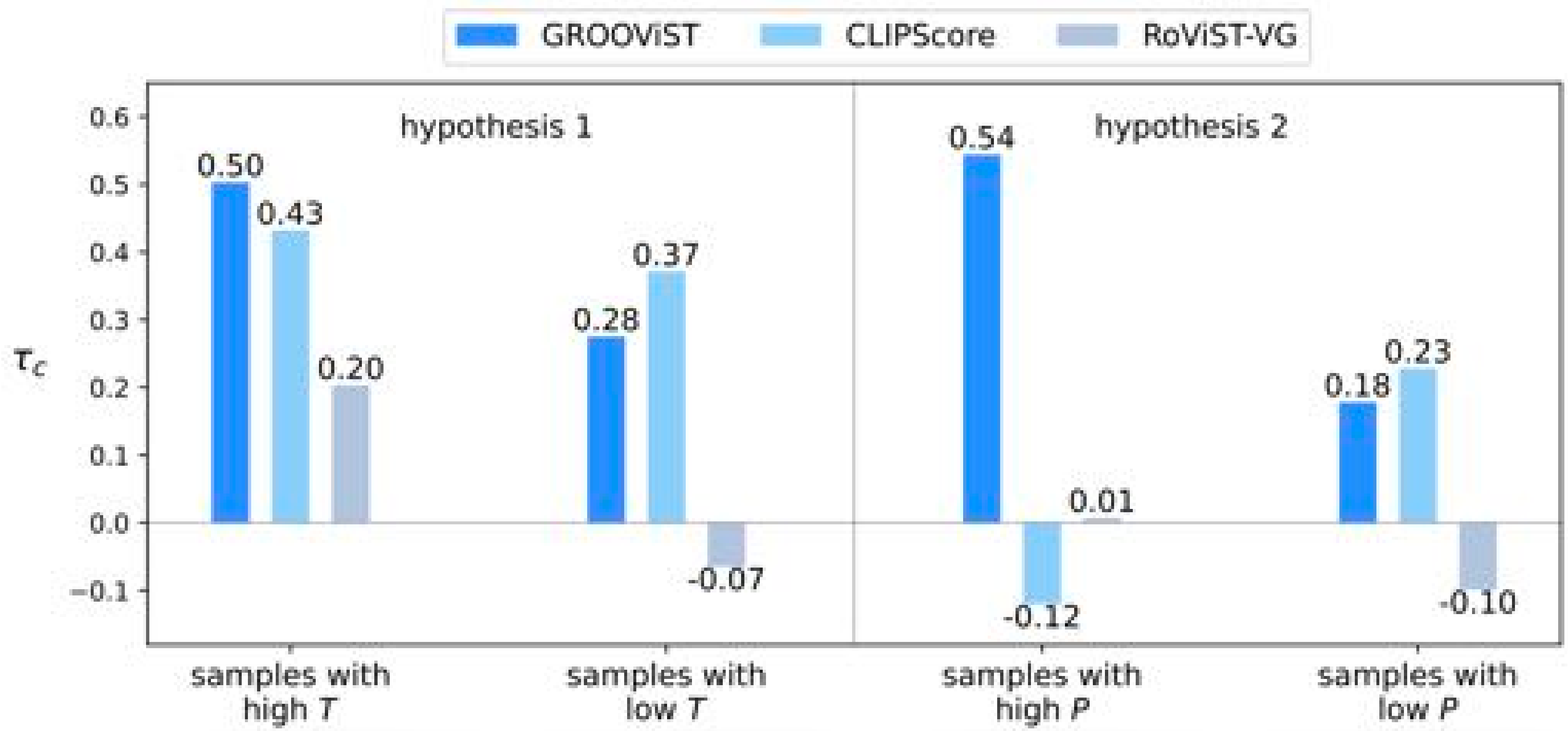
Analysis of GROOViST

Analysis of GROOViST



Vision-language misalignment (T)

Analysis of GROOViST



Vision-language misalignment (T)

Proportion of NPs in the story (P)

Interpretable output (not in CLIP!)

1) there was lots to see and do at the festival , including listening to unusual instruments . 2) many stalls had handmade clothing and one even had dresses specifically for little girls . 3) as part of the festival grounds , there were also numerous sculptures that one could touch . 4) many stalls were adorned with handmade glass bottles . 5) by midday thousands were in attendance , the biggest turn out yet !

Highly grounded story!

GROOViST = 0.855 (in range [-1,1])

Evaluate Visual Storytelling II



(EMNLP Findings 2024)

Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition

Aditya K Surikuchi, Raquel Fernández, Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
{a.k.surikuchi|raquel.fernandez|s.pezzelle}@uva.nl

Abstract

Visual storytelling consists in generating a natural language story given a temporally ordered sequence of images. This task is not only challenging for models, but also very difficult to evaluate with automatic metrics since there is no consensus about what makes a story 'good'. In this paper, we introduce a novel method that measures story quality in terms of human likeness regarding three key aspects highlighted in previous work: visual grounding, coherence, and repetitiveness. We then use this method to evaluate the stories generated by several models, showing that the foundation model LLaVA obtains the best result, but only slightly so compared to TAPM, a 50-times smaller visual storytelling model. Upgrading the visual and language components of TAPM results in a model that yields competitive performance with a relatively low number of parameters. Finally, we carry out a human evaluation study, whose results suggest that a 'good' story may require more than a human-like level of visual grounding, coherence, and repetition.

1 Introduction

Visual storytelling is the task of generating a story for a sequence of several temporally-ordered images or video frames. For both human speakers and machine learning models, the task requires connecting the visual data causally, to generate a narrative consistent with the contents of the images. As for model-generated stories, evaluation is one of the key challenges due to the inherently creative nature of the task. Since human-written stories are coherence, or how repetitive they are. This problem has only been addressed recently, with Wang et al. (2022) and Surikuchi et al. (2023) proposing various metrics to take into account some of these crucial aspects. These methods assess the appropriateness of a generated story independently from its overlap with a ground-truth story for the same image sequence. Given that the same image sequence can possibly give rise to many different stories, this type of higher-level evaluation that does not rely on text overlap is clearly desirable.

Nevertheless, we argue that measuring the degree of coherence or visual grounding of a story may not be sufficiently informative, as there are no standard conventions that determine the preferable level of such properties. To address this issue, in this work, we first propose an evaluation method that assesses the quality of generated stories in terms of their *distance* from human-written stories along several relevant dimensions, each measured by an available reference-free metric. Using this method, we evaluate a range of models on the visual storytelling task, including models specifically designed and trained for this task, as well as—for the first time—foundation models pre-trained to achieve general-purpose language and vision abilities, which we test in a zero-shot manner. We show that LLaVA (Liu et al., 2024), a powerful foundation model, performs best on the task, but only slightly better than TAPM (Yu et al., 2021), a model designed for visual storytelling which is 50 times smaller than LLaVA. Second, given insights derived from our proposed distance-based evalua-

Evaluate Visual Storytelling II



There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!

Dimensions to evaluate in a story:

- new information
- coherence
- visual grounding

(EMNLP Findings 2024)

Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition

Aditya K Surikuchi, Raquel Fernández, Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
{a.k.surikuchi|raquel.fernandez|s.pezzelle}@uva.nl

Abstract

Visual storytelling consists in generating a natural language story given a temporally ordered sequence of images. This task is not only challenging for models, but also very difficult to evaluate with automatic metrics since there is no consensus about what makes a story 'good'. In this paper, we introduce a novel method that measures story quality in terms of human likeness regarding three key aspects highlighted in previous work: visual grounding, coherence, and repetitiveness. We then use this method to evaluate the stories generated by several models, showing that the foundation model LLaVA obtains the best result, but only slightly so compared to TAPM, a 50-times smaller visual storytelling model. Upgrading the visual and language components of TAPM results in a model that yields competitive performance with a relatively low number of parameters. Finally, we carry out a human evaluation study, whose results suggest that a 'good' story may require more than a human-like level of visual grounding, coherence, and repetition.

1 Introduction

Visual storytelling is the task of generating a story for a sequence of several temporally-ordered images or video frames. For both human speakers and machine learning models, the task requires connecting the visual data causally, to generate a narrative consistent with the contents of the images. As for model-generated stories, evaluation is one of the key challenges due to the inherently creative nature of the task. Since human-written stories are coherence, or how repetitive they are. This problem has only been addressed recently, with Wang et al. (2022) and Surikuchi et al. (2023) proposing various metrics to take into account some of these crucial aspects. These methods assess the appropriateness of a generated story independently from its overlap with a ground-truth story for the same image sequence. Given that the same image sequence can possibly give rise to many different stories, this type of higher-level evaluation that does not rely on text overlap is clearly desirable.

Nevertheless, we argue that measuring the degree of coherence or visual grounding of a story may not be sufficiently informative, as there are no standard conventions that determine the preferable level of such properties. To address this issue, in this work, we first propose an evaluation method that assesses the quality of generated stories in terms of their *distance* from human-written stories along several relevant dimensions, each measured by an available reference-free metric. Using this method, we evaluate a range of models on the visual storytelling task, including models specifically designed and trained for this task, as well as—for the first time—foundation models pre-trained to achieve general-purpose language and vision abilities, which we test in a zero-shot manner. We show that LLaVA (Liu et al., 2024), a powerful foundation model, performs best on the task, but only slightly better than TAPM (Yu et al., 2021), a model designed for visual storytelling which is 50 times smaller than LLaVA. Second, given insights derived from our proposed distance-based evalua-

Evaluate Visual Storytelling II

There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!

Dimensions to evaluate in a story:

- new information
- coherence
- visual grounding ✓

(EMNLP Findings 2024)

Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition

Aditya K Surikuchi, Raquel Fernández, Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
{a.k.surikuchi|raquel.fernandez|s.pezzelle}@uva.nl

Abstract

Visual storytelling consists in generating a natural language story given a temporally ordered sequence of images. This task is not only challenging for models, but also very difficult to evaluate with automatic metrics since there is no consensus about what makes a story 'good'. In this paper, we introduce a novel method that measures story quality in terms of human likeness regarding three key aspects highlighted in previous work: visual grounding, coherence, and repetitiveness. We then use this method to evaluate the stories generated by several models, showing that the foundation model LLaVA obtains the best result, but only slightly so compared to TAPM, a 50-times smaller visual storytelling model. Upgrading the visual and language components of TAPM results in a model that yields competitive performance with a relatively low number of parameters. Finally, we carry out a human evaluation study, whose results suggest that a 'good' story may require more than a human-like level of visual grounding, coherence, and repetition.

1 Introduction

Visual storytelling is the task of generating a story for a sequence of several temporally-ordered images or video frames. For both human speakers and machine learning models, the task requires connecting the visual data causally, to generate a narrative consistent with the contents of the images. As for model-generated stories, evaluation is one of the key challenges due to the inherently creative nature of the task. Since human-written stories are

coherence, or how repetitive they are. This problem has only been addressed recently, with Wang et al. (2022) and Surikuchi et al. (2023) proposing various metrics to take into account some of these crucial aspects. These methods assess the appropriateness of a generated story independently from its overlap with a ground-truth story for the same image sequence. Given that the same image sequence can possibly give rise to many different stories, this type of higher-level evaluation that does not rely on text overlap is clearly desirable.

Nevertheless, we argue that measuring the degree of coherence or visual grounding of a story may not be sufficiently informative, as there are no standard conventions that determine the preferable level of such properties. To address this issue, in this work, we first propose an evaluation method that assesses the quality of generated stories in terms of their *distance* from human-written stories along several relevant dimensions, each measured by an available reference-free metric. Using this method, we evaluate a range of models on the visual storytelling task, including models specifically designed and trained for this task, as well as—for the first time—foundation models pre-trained to achieve general-purpose language and vision abilities, which we test in a zero-shot manner. We show that LLaVA (Liu et al., 2024), a powerful foundation model, performs best on the task, but only slightly better than TAPM (Yu et al., 2021), a model designed for visual storytelling which is 50 times smaller than LLaVA. Second, given insights derived from our proposed distance-based evalua-

A comprehensive evaluation

There was lots to see and do at the festival, including listening to unusual instruments. Many stalls had handmade clothing and one even had dresses specifically for little girls... By midday thousands were in attendance, the biggest turn out yet!

Dimensions to evaluate in a story:

- new information
- coherence
- visual grounding

(EMNLP Findings 2024)

Not (yet) the whole story: Evaluating Visual Storytelling Requires More than Measuring Coherence, Grounding, and Repetition

Aditya K Surikuchi, Raquel Fernández, Sandro Pezzelle
Institute for Logic, Language and Computation
University of Amsterdam
{a.k.surikuchi|raquel.fernandez|s.pezzelle}@uva.nl

Abstract

Visual storytelling consists in generating a natural language story given a temporally ordered sequence of images. This task is not only challenging for models, but also very difficult to evaluate with automatic metrics since there is no consensus about what makes a story 'good'. In this paper, we introduce a novel method that measures story quality in terms of human likeness regarding three key aspects highlighted in previous work: visual grounding, coherence, and repetitiveness. We then use this method to evaluate the stories generated by several models, showing that the foundation model LLaVA obtains the best result, but only slightly so compared to TAPM, a 50-times smaller visual storytelling model. Upgrading the visual and language components of TAPM results in a model that yields competitive performance with a relatively low number of parameters. Finally, we carry out a human evaluation study, whose results suggest that a 'good' story may require more than a human-like level of visual grounding, coherence, and repetition.

1 Introduction

Visual storytelling is the task of generating a story for a sequence of several temporally-ordered images or video frames. For both human speakers and machine learning models, the task requires connecting the visual data causally, to generate a narrative consistent with the contents of the images. As for model-generated stories, evaluation is one of the key challenges due to the inherently creative nature of the task. Since human-written stories are coherence, or how repetitive they are. This problem has only been addressed recently, with Wang et al. (2022) and Surikuchi et al. (2023) proposing various metrics to take into account some of these crucial aspects. These methods assess the appropriateness of a generated story independently from its overlap with a ground-truth story for the same image sequence. Given that the same image sequence can possibly give rise to many different stories, this type of higher-level evaluation that does not rely on text overlap is clearly desirable.

Nevertheless, we argue that measuring the degree of coherence or visual grounding of a story may not be sufficiently informative, as there are no standard conventions that determine the preferable level of such properties. To address this issue, in this work, we first propose an evaluation method that assesses the quality of generated stories in terms of their *distance* from human-written stories along several relevant dimensions, each measured by an available reference-free metric. Using this method, we evaluate a range of models on the visual storytelling task, including models specifically designed and trained for this task, as well as—for the first time—foundation models pre-trained to achieve general-purpose language and vision abilities, which we test in a zero-shot manner. We show that LLaVA (Liu et al., 2024), a powerful foundation model, performs best on the task, but only slightly better than TAPM (Yu et al., 2021), a model designed for visual storytelling which is 50 times smaller than LLaVA. Second, given insights derived from our proposed distance-based evalua-

Standard evaluation of VST

Reference-based: comparing generated stories to human ones via METEOR, CIDEr, BERTScore, BLEURT, etc.

Standard evaluation of VST

Reference-based: comparing generated stories to human ones via METEOR, CIDEr, BERTScore, BLEURT, etc.

Reference-free: RoViST [Wang et al. 2022] producing a score for visual grounding, coherence, and non-redundancy

Standard evaluation of VST

Reference-based: comparing generated stories to human ones via METEOR, CIDEr, BERTScore, BLEURT, etc.

- ✦ but, there can be many ‘good’ stories for the same sequence! No “ground-truth”!

Reference-free: RoViST [Wang et al. 2022] producing a score for visual grounding, coherence, and non-redundancy

Standard evaluation of VST

Reference-based: comparing generated stories to human ones via METEOR, CIDEr, BERTScore, BLEURT, etc.

- ✦ but, there can be many ‘good’ stories for the same sequence! No “ground-truth”!

Reference-free: RoViST [Wang et al. 2022] producing a score for visual grounding, coherence, and non-redundancy

- ◆ but, no way to assess the validity of these metrics! E.g., is a high visual grounding in line with how people tell stories?

A reference-free, human-based evaluation

Here, we propose a ‘good’ generated story is similar to a human one with respect to these 3 main components

We operationalize this in terms of a **distance *d* between the scores** achieved by generated and human stories

- The lower the distance, the more human-like the stories!

A reference-free, human-based evaluation

Here, we propose a ‘good’ generated story is similar to a human one with respect to these 3 main components

We operationalize this in terms of a **distance *d* between the scores** achieved by generated and human stories

- The lower the distance, the more human-like the stories!

$$
\begin{aligned}
\mathrm{d}^{C}_{HM} &= |C_H - C_M|, \\
\mathrm{d}^{G}_{HM} &= |G_H - G_M|, \\
\mathrm{d}^{R}_{HM} &= |R_H - R_M|
\end{aligned}
$$

A reference-free, human-based evaluation

Here, we propose a ‘good’ generated story is similar to a human one with respect to these 3 main components

We operationalize this in terms of a **distance *d* between the scores** achieved by generated and human stories

- The lower the distance, the more human-like the stories!

$$\mathrm{d}_{HM} = (\mathrm{d}^{C}_{HM} + \mathrm{d}^{G}_{HM} + \mathrm{d}^{R}_{HM})/3$$

Three components

1. Visual grounding: we use **GROOViST**

2. Coherence: average probability that *s* follows preceding context computed using **ALBERT** [Lan et al., 2020]

3. Repetition (non-redundancy): **Jaccard Similarity** between non-overlapping 4-gram phrases of a sentence

Five models

1. GLAC Net
2. AREL
3. TAPM

VST-trained models

Five models

1. GLAC Net
2. AREL
3. TAPM

VST-trained models

4. BLIP-2
5. LLaVA

general-purpose VLMs

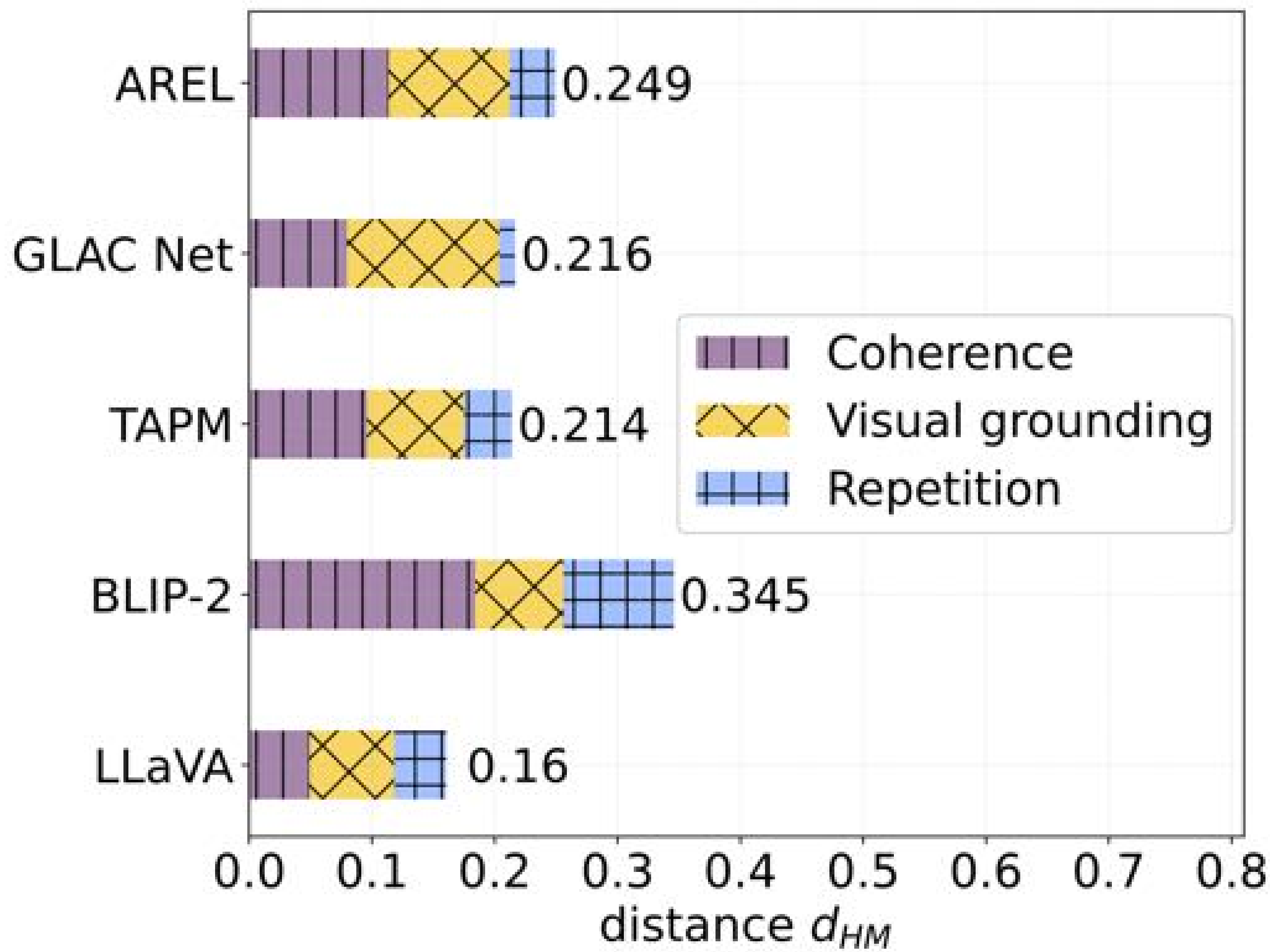
Five models

1. GLAC Net
2. AREL
3. TAPM
4. BLIP-2
5. LLaVA

Transformers

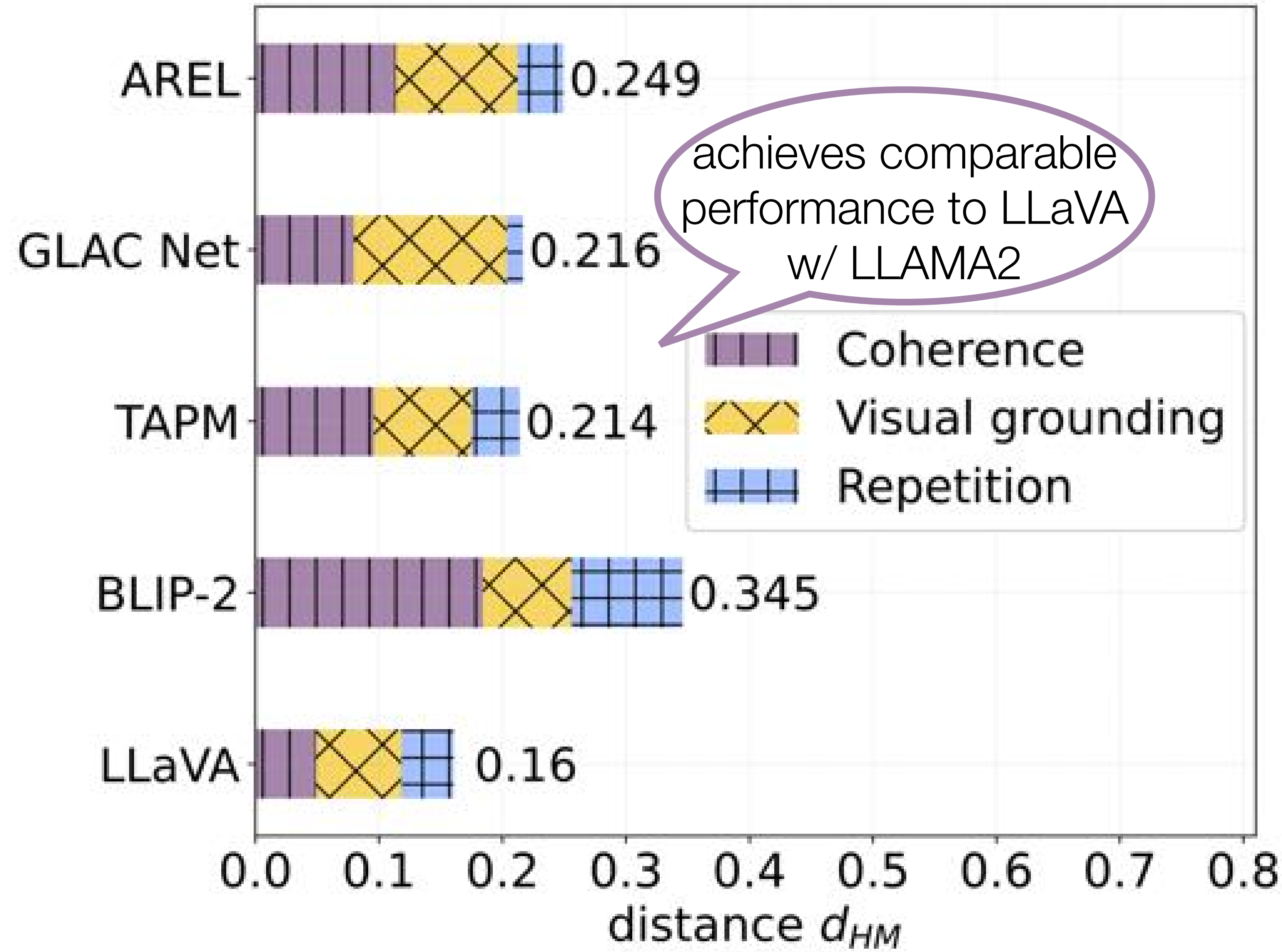
Results

Recall: The lower, the better



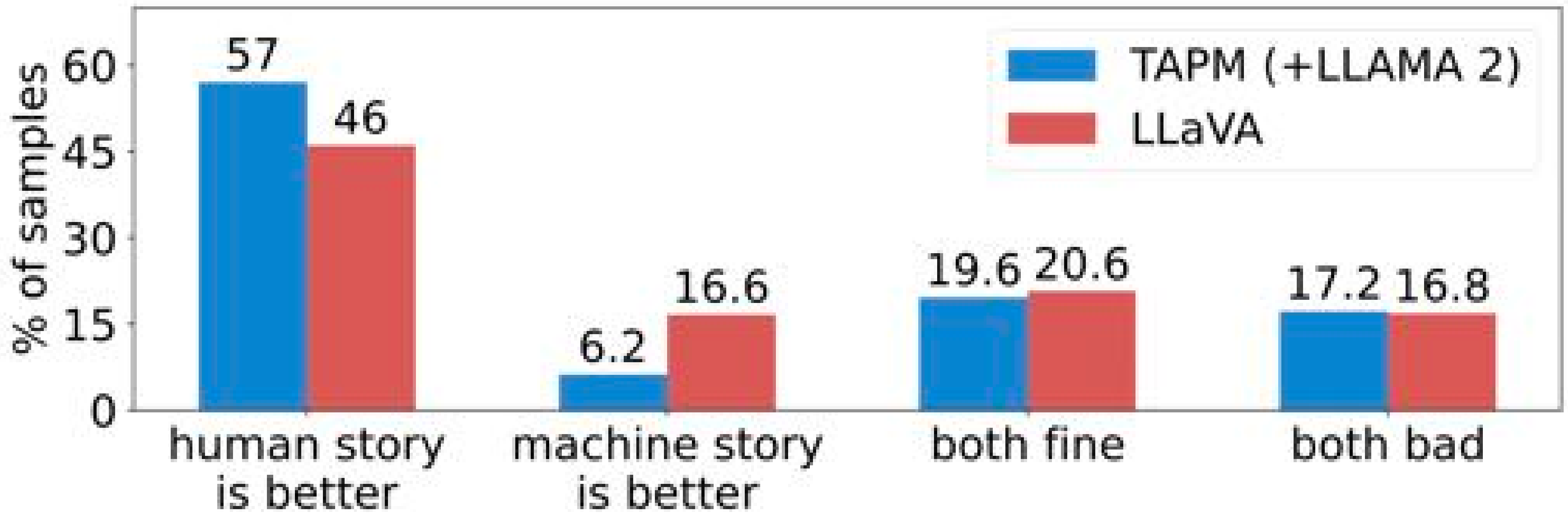
Results

Recall: The lower, the better



Human evaluation

Do these (low) numbers ‘translate’ into high-quality, human-like stories? **Yes, but…**



Takeaways

Our evaluation is sensible: the lower the distance, the higher the chance for the story to be judged as “as good as”

At the same time, human annotators still largely prefer human stories to model-generated ones

Capturing coherence, repetitiveness, and visual grounding is *not the whole story*—creativity, pragmatics, perspective?

Discussion

Evaluating VLMs' outputs in communicative scenarios is an open challenge: complex task, complex evaluation

Reference-based metrics not suitable. We can ask human experts (expensive), or use VLMs as judges (but see Bavaresco et al. 2024!)

Another promising route is to devise **automatic metrics and evaluations based on human(-like) characteristics**—as we did here. We explored three, but there are many more!

Plus, move to even more naturalistic data (videos), often domain-specific and with more implicit and underspecified language

CLASP, University of Gothenburg

March 26, 2025

Questions?

Sandro Pezzelle

ILLC, University of Amsterdam

`s.pezzelle@uva.nl`

`sandropezzelle.github.io`