Bridging the Modalities: Investigating the Integration of Information from Language and Vision in Multimodal models

Thesis presentation

Author: Yuyu Bai
Supervisor: Sandro Pezzelle(UVA)
Date: August 24th, 2023
Vrije University

Table of contents

Introduction

Introduction - Motivation

1.The rapid evolution of deep learning in NLP and CV field has given rise to multimodal models.

2. A research gap in evaluating their true capability to integrate multimodal information.

Introduction-Research Question

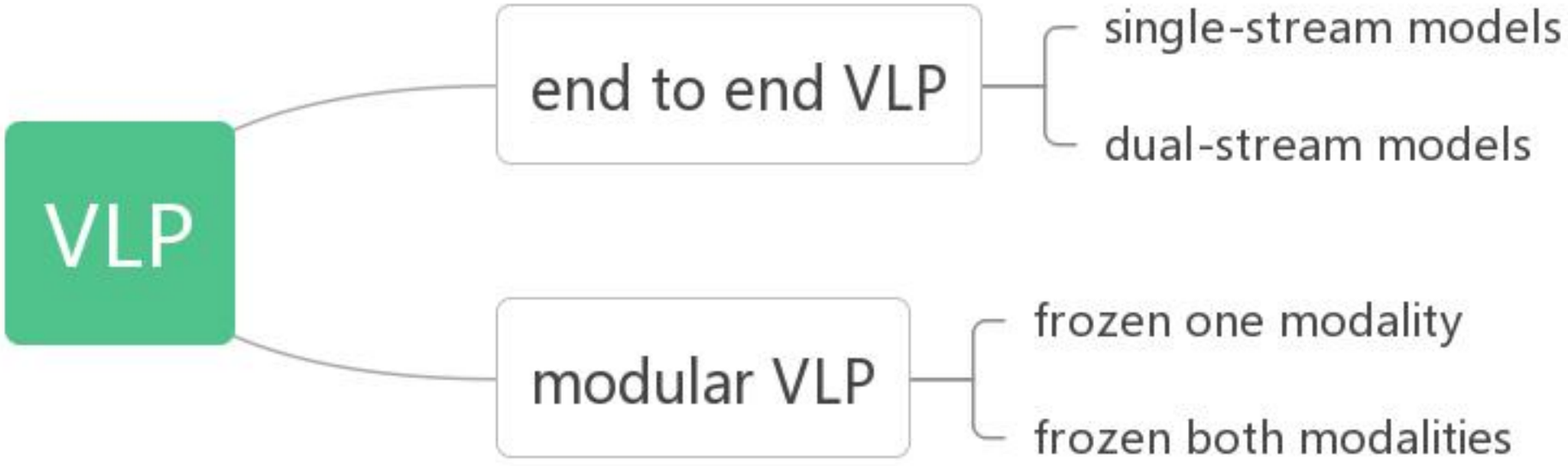
How good recent generative multimodal models are on integrating information from language and vision?

How can their outputs be evaluated? and what are the challenges?

What is the impact of prompts on their outputs?

Related Works

Related works-Vision Language Pretraining



Related works-Tasks and Datasets

Related works-Evaluation metrics

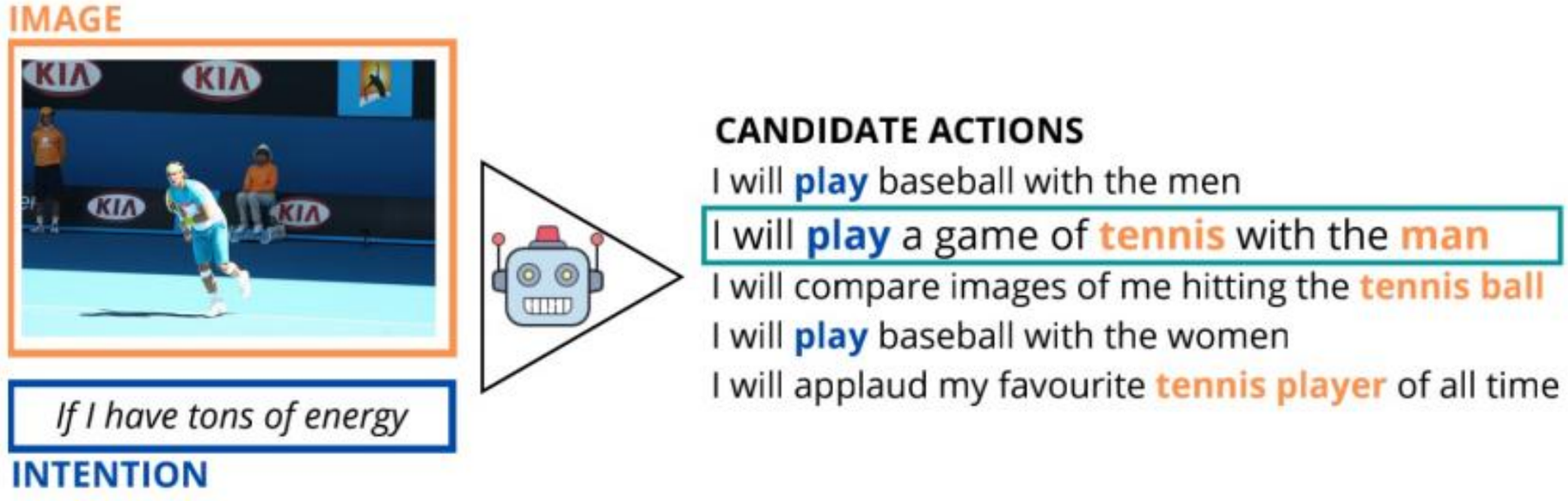
1. Accuracy, Recall, BLEU, ROUGE, R-precision.

2. SPICE, ANLS.

3. VL Checklist , BD2BB.

Method

Method-BD2BB dataset



Method-Models

Model	FROMAGe	MAPL	BLIP2
Year	2023	2022	2023
Is generative?	yes	yes	yes
Is vision model frozen?	Yes	Yes	Yes
Is language model frozen?	Yes	Yes	Yes
No. of trainable parameter	220M	3.3M	188M
Can the model output images?	Yes	No	No
architechture for bridging two modalities	two Linear mapping both for image-to-text and text-to-image	Mapping network	Q-Former（transformer）

Method-Two Experiment Settings

How can BD2BB benchmark be used in a generative models?

1. Modifying the prompt by incorporating the options into it and explicitly asking the model to select one -> Discriminative setting

2. Compute a similarity score between the action generated by the model and each of the given options, ultimately selecting the option with the highest score.

-> Generative setting

Results and Discussion

Results and Discussion-Discriminative setting

Discriminative setting: Modifying the prompt by incorporating the options into it and explicitly asking the model to select one.

Example:

prompt : “If I feel adventurous, what should I do? Choose the best option from the following ones:

A. I will ride an elephant.

B. I will merely watch my friend fly an animal kite.

C. I will go bird watching on an outdoor public patio.

D. I will ride a horse like the man.

E. I will stand and observe the zebras.”

Results and Discussion-CLIP model

Model		accuracy
Models in original paper	Baseline	49.0±0.9
	LXMERTs	51.3±0.4
	LXMERTs pretrain	62.2±2.2
Models that we tested on	CLIP	53.2
	MAPL	39.0
	FROMAGe	41.3
	BLIP2	72.5
	Human	79.0

Results and Discussion-Discriminative setting

Model	accuracy
BLIP2_{LV}	73.5
BLIP2_{L}	56.0
BLIP2_{V}	53.0
Human_{LV}	79.0
Human_{L}	50.0
Human_{V}	72.3

Results and Discussion-Discriminative setting

Is the prediction correct?	case number	percentage	comments
BLIP_LV: T BLIP_V: T BLIP_L: T	1350	0.3308	No errors were found in these cases, indicating that they may be too easy for the multimodality model to handle.
BLIP_LV: T BLIP_V: T BLIP_L: F	581	0.1424	The model in the L setting gave incorrect predictions due to the absence of image information.
BLIP_LV: T BLIP_V: F BLIP_L: T	808	0.1980	The model in the V setting gave incorrect predictions due to the absence of intention information.
BLIP_LV: T BLIP_V: F BLIP_L: F	222	0.0544	Only multimodality setting can give true predictions.
BLIP_LV: F BLIP_V: T BLIP_L: T	11	0.0027	The model's incorrect predictions can be attributed to the following reasons:1.Poor data quality; 2.Wrong object detection; 3.Failure to understand the intention; 4.Only considering one modality;
BLIP_LV: F BLIP_V: T BLIP_L: F	221	0.0542	
BLIP_LV: F BLIP_V: F BLIP_L: T	117	0.0287	
BLIP_LV: F BLIP_V: F BLIP_L: F	771	0.1889	

Results and Discussion-Discriminative setting

Intention: If I want to celebrate

Options:
A. I will sing happy birthday.
B. I will I would invite my friends to a bar to spend time together.
C. I will hope the men to finish their speeches.
D. I will celebrate my birthday with cake and friends.
E. I will eat some cake like the party girl.

Correct answer: A

BLIP2(LV): D
BLIP2(LV): D
BLIP2(LV): D

Results and Discussion-Generative setting

Generative setting: Compute a similarity score between the action generated by the model and each of the given options, ultimately selecting the option with the highest score.

Example:

prompt : “If I feel adventurous, what should I do?” Answer: I will ride a horse.

A. I will ride an elephant.

B. I will merely watch my friend fly an animal kite.

C. I will go bird watching on an outdoor public patio.

D. I will ride a horse like the man. --highest Similarity Score (BERTscore)

E. I will stand and observe the zebras.

Results and Discussion-Generative setting

prompt	accuracy		
	BLIP2	MAPL	FROMAGE
Question: What am I supposed to do + [intention]? Answer:	53	48	\
Q: What am I supposed to do +[intention]+? \nA:	52	50	\
What am I supposed to do + [intention]+?	53	\	\
What am I supposed to do + [intention]+? I will	53	44	39
What am I supposed to do + [intention]+? Answer in the format "I will".	55	\	41

Results and Discussion-Generative setting

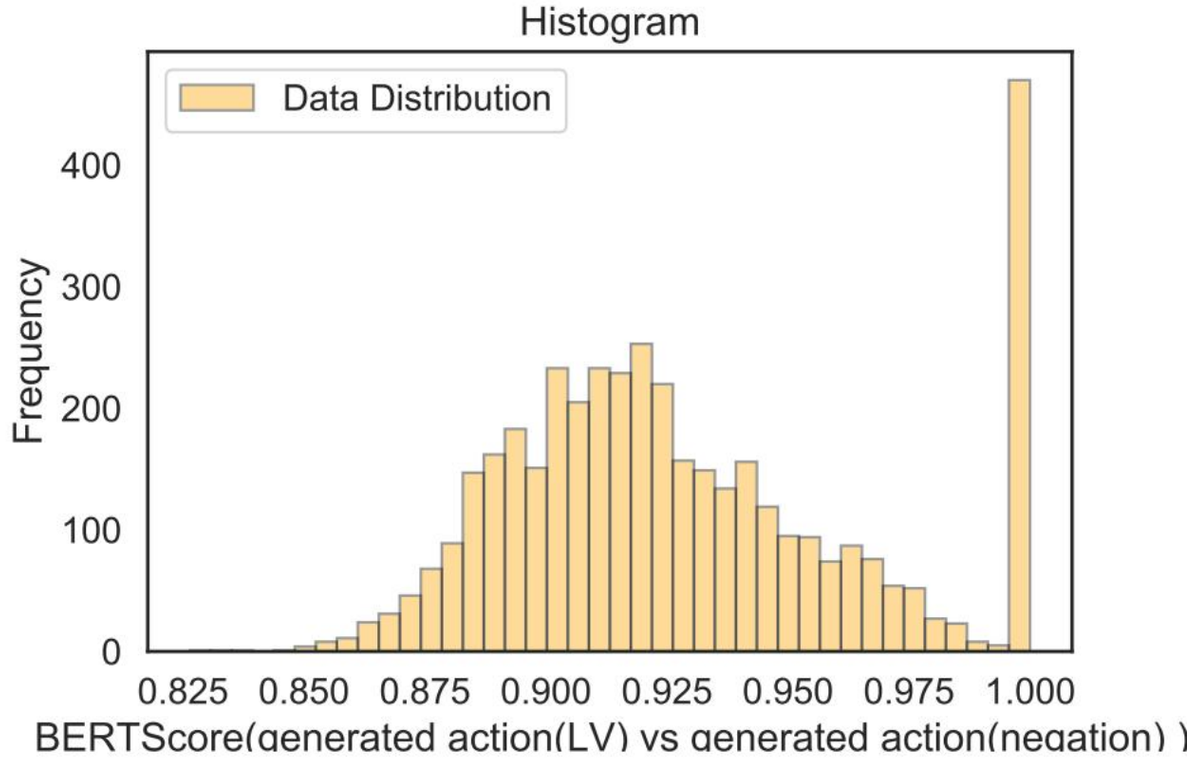
The effect of negation：

Accuracy drops after negation:

54% - > 45%(generative)

72% - > 38%(discriminative)

Results and Discussion-Generative setting



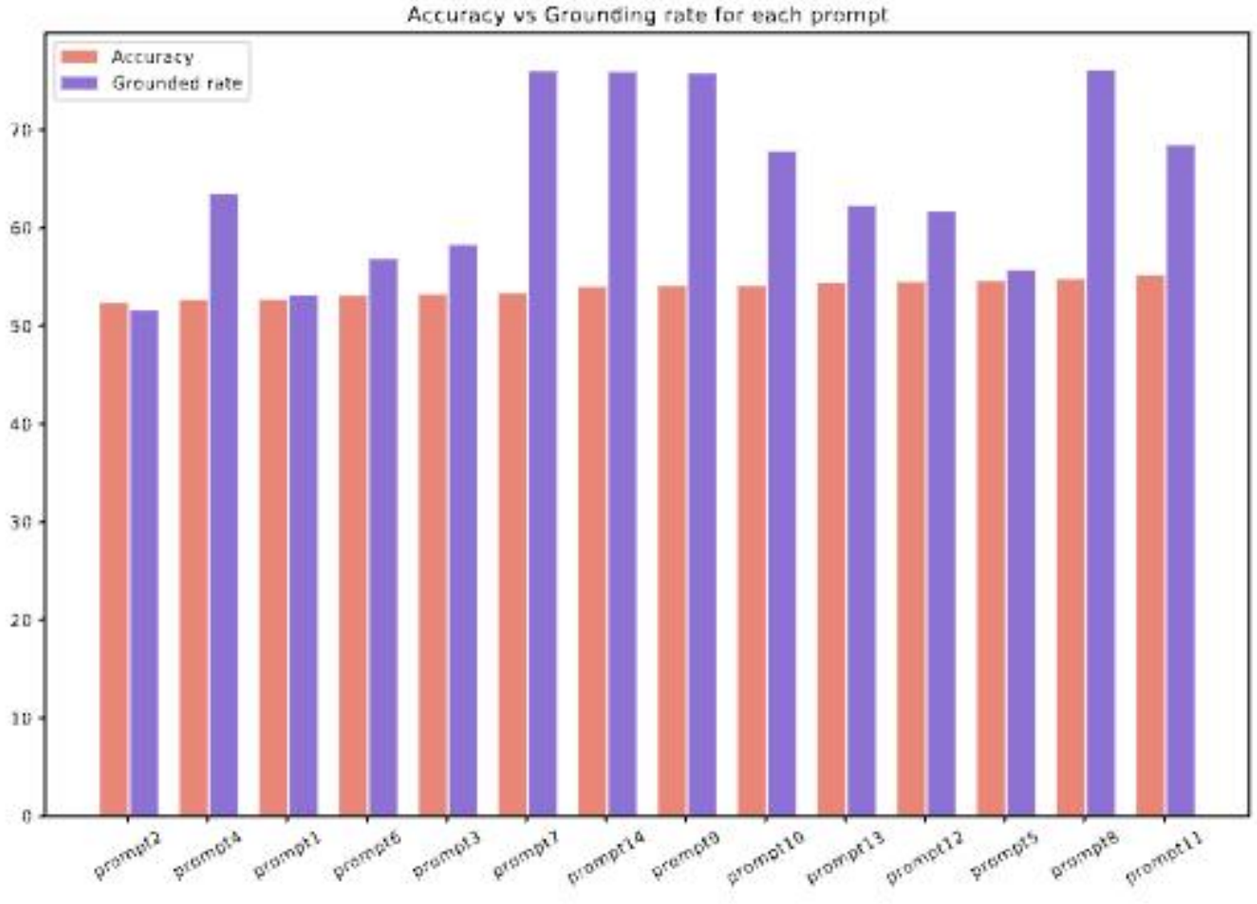
Results and Discussion-Generative setting

Actions	grounding rate
Target action	76%
Generated action(LV)	54%
Generated action(L)	6%
Generated action(V)	69%
Generated action(negation)	38%

Results and Discussion-Generative setting

	Prompt	Accuracy	Grounding rate
prompt1	Question: What am I supposed to do + [intention]? Answer:	52.66	53.10
prompt2	Q: What am I supposed to do + [intention]? \nA:	52.34	51.58
prompt3	What am I supposed to do + [intention]?	53.20	58.25
prompt4	What am I supposed to do + [intention]+? I will	52.63	63.44
prompt5	What am I supposed to do + [intention]+? Answer in the format "I will"	54.57	55.67
prompt6	What am I supposed to do + [intention]+? Please provide an answer based solely on the intention, without considering the image.	53.08	56.82
prompt7	What action should I take +[intention]+? Please base your response solely on the image. Additionally, kindly mention at least one object visible in the image.	53.32	75.97
prompt8	Suppose I am in this specific visual context. What actions should I take+ [intention]+? Please mention at least one object from the image.	54.76	76.06
prompt9	Imagine I am in the given visual scenario. What actions should I take regarding +[intention]+'? Please mention at least one object from the image.	54.06	75.74
prompt10	Imagine yourself in this specific visual context. Considering both the intention and the image, what actions should be taken +[intention]+?	54.06	67.78
prompt11	Considering both the intention and the image, what will you do +[intention]+?	55.16	68.41
prompt12	What will I do +[intention]+?	54.47	61.67
prompt13	What will you do +[intention]+? I will	54.37	62.23
prompt14	What will you do +[intention]+? Please give a a plausible reason by mentioning at lease one object from the image.	53.96	75.89

Results and Discussion-Generative setting



Conclusion and Future work

Conclusion

What have been done?

1. Evaluated several state-of-the-art generative multimodal models using BD2BB benchmark.

2. Performed a series of experiment and analysis

(error analysis, grounding level, prompt analysis and negation experiment)

3. Identified the current limitations of these models and evaluated their robustness in handling diverse scenarios.

Conclusion

What can be concluded?

1.Generative multimodal models are capable of successfully completing BD2BB tasks without fine-tuning.

2. Among them, BLIP2 stands out, outperforming the others in both discriminative and generative settings.

3. The level of grounding can be consider as an evaluative aspect beyond accuracy.

Future work

New datasets-new multimodal dataset that contains more ”true” multimodal data.

New metrics -better determine if the model utilizes complementary information or simply excels at selecting relevant information.

New models - improve the robustness and adaptability.

Thanks for listening!

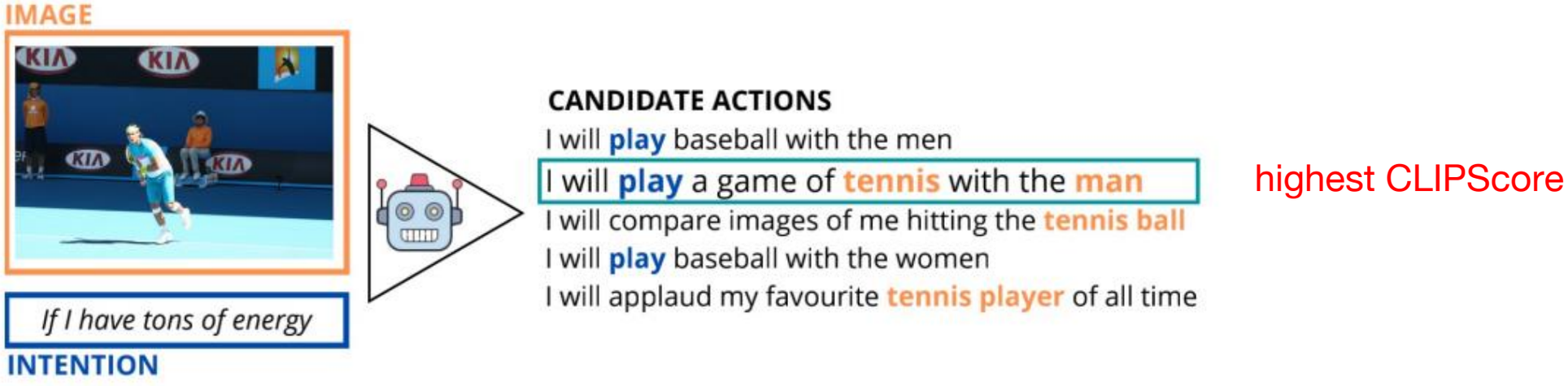
Author: Yuyu Bai
Supervisor: Sandro Pezzelle(UVA)
Date: August 24th, 2023
Vrije University

Results and Discussion-Generative setting

Condition	BERTScore	BLEU-4	CIDER	METEOR	ROUGE
LV	0.53	0.54	0.52	0.48	0.51
L	0.39	0.49	0.38	0.29	0.36
V	0.42	0.37	0.41	0.38	0.4

Results and Discussion-Discriminative setting

- We use CLIP model as a baseline.



Results and Discussion-CLIP model

Intention + target action: If I want to celebrate, I will sing happy birthday.
Intention+ vision decoy action:
If I want to celebrate, I will eat some cake like the party girl.
Intention+ language decoy action:
If I want to celebrate, I will I would invite my friends to a bar to spend time together.
Intention: If I want to celebrate
target action: I will sing happy birthday.
vision decoy action: I will eat some cake like the party girl.
language decoy action: I will I would invite my friends to a bar to spend time together.
COCO Caption: A woman is celebrating her birthday in a nice restaurant.

Results and Discussion-CLIP model