SEOUL NATIONAL UNIV. VISION & LEARNING

NAVER AI LAB

MPCHAT: Towards Multimodal Persona-Grounded Conversation

Jaewoo Ahn, Yeda Song, Sangdoo Yun and Gunhee Kim

Code link: http://vision.snu.ac.kr/projects/mpchat

Motivation

Conversational agents produce inconsistent responses

They contradict the previous utterances. Previous works incorporate **persona** (e.g. self-descriptive sentences) to improve consistency.

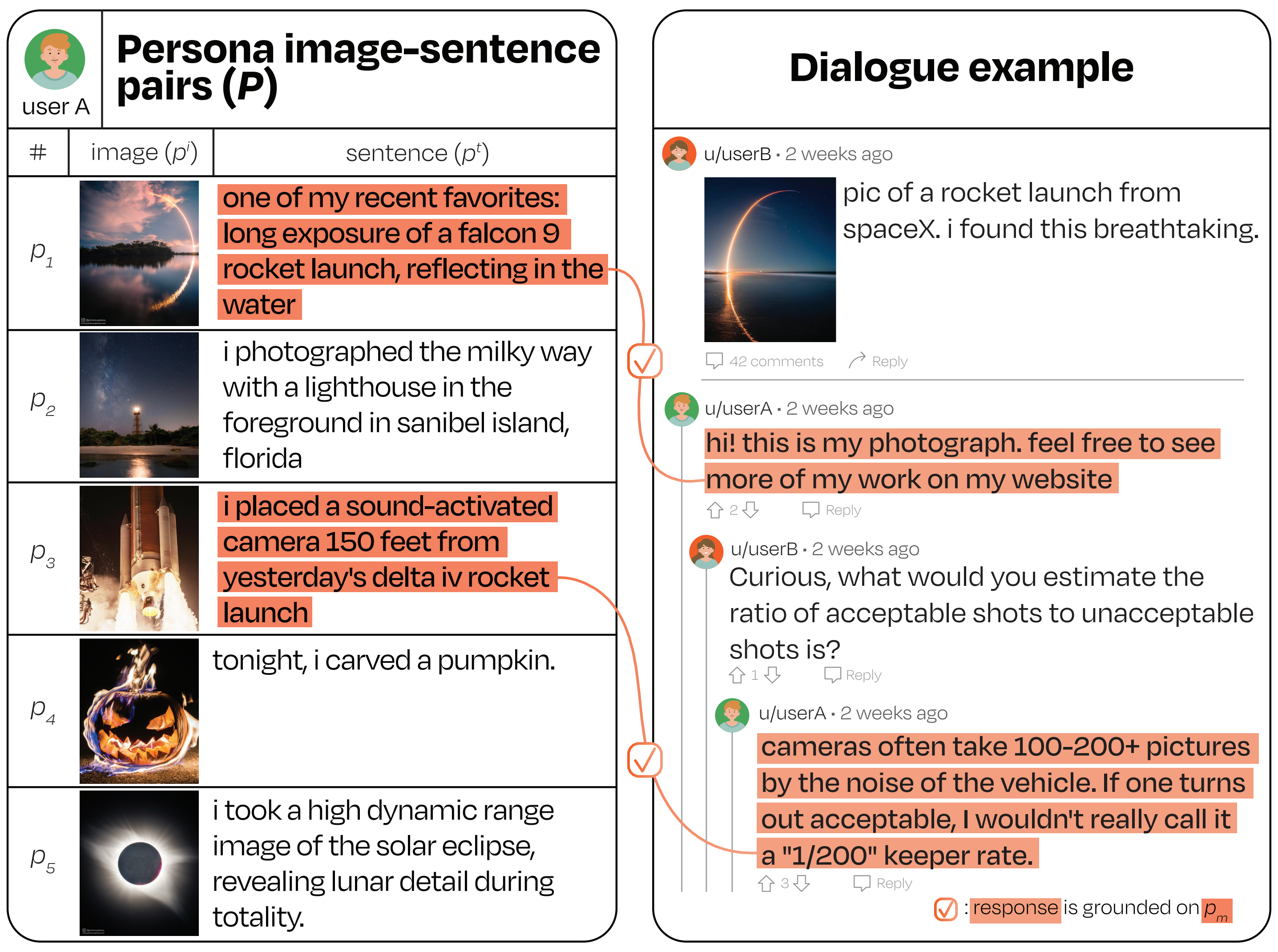
However, previous works focused on *textual* persona

It delivers only personal facts or personalities. But one's persona should be explored in multi-faceted ways.

Episodic memory is a memory of personal experiences, represented in the form of visual images. Since it is crucial in shaping personal identity, it can influence one's persona.

Therefore, we propose *multimodal* persona, a set of persona image-sentence pairs

The MPCHAT Dataset



Persona image-sentence pairs (*P*)

user A

#	image (p^i)	sentence (p^t)
p_1		one of my recent favorites: long exposure of a falcon 9 rocket launch, reflecting in the water
p_2		i photographed the milky way with a lighthouse in the foreground in sanibel island, florida
p_3		i placed a sound-activated camera 150 feet from yesterday's delta iv rocket launch
p_4		tonight, i carved a pumpkin.
p_5		i took a high dynamic range image of the solar eclipse, revealing lunar detail during totality.

Dialogue example

u/userB · 2 weeks ago

pic of a rocket launch from spaceX. i found this breathtaking.

42 comments Reply

u/userA · 2 weeks ago

hi! this is my photograph. feel free to see more of my work on my website

2 Reply

u/userB · 2 weeks ago

Curious, what would you estimate the ratio of acceptable shots to unacceptable shots is?

1 Reply

u/userA · 2 weeks ago

cameras often take 100-200+ pictures by the noise of the vehicle. If one turns out acceptable, I wouldn't really call it a "1/200" keeper rate.

3 Reply

✓ : response is grounded on p_m

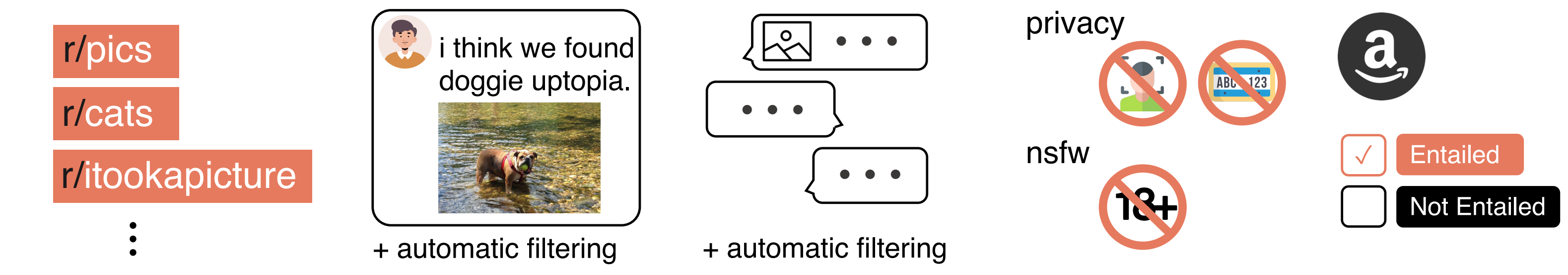
We introduce MPCHAT, a new multimodal persona-grounded dialogue dataset

A 15K multimodal dialogue dataset sourced from **Reddit** including 26K speakers with more than 17 multimodal personas per speaker.

Dataset	# Dialog	Data source	Persona type	Persona modality	Entailment label
LIGHT	11K	Crowd-sourced	Fact	T	No
PD	20.8M	Weibo	Fact	T	No
PEC	355K	Reddit	Thought	T	No
PELD	6.4K	TV shows	Personality	T	No
PersonaChat	13K	Crowd-sourced	Fact	T	Post-Hoc
FoCus	14K	Crowd-sourced	Fact	T	Yes
MPCHAT	15K	Reddit	Episodic memory	V, T	Yes

Dataset construction

1) Subreddit curation 2) Persona collection 3) Dialogue collection 4) Additional filtering 5) Human labeling



Benchmarks

We propose three retrieval-based tasks in MPCHAT

1) Next Response Prediction (NRP)

predict next response based on context and multimodal persona.

- input: context c, multimodal persona P, response candidates R_c
- output: response r

2) Grounding Persona Prediction (GPP)

predict speaker's grounding persona element based on dialogue and remaining persona.

- input: context c, response r (optional), remainder persona set $\bar{P}$, persona candidates P_c
- output: grounding persona element p

3) Speaker Identification (SI)

predict speaker based on dialogue information.

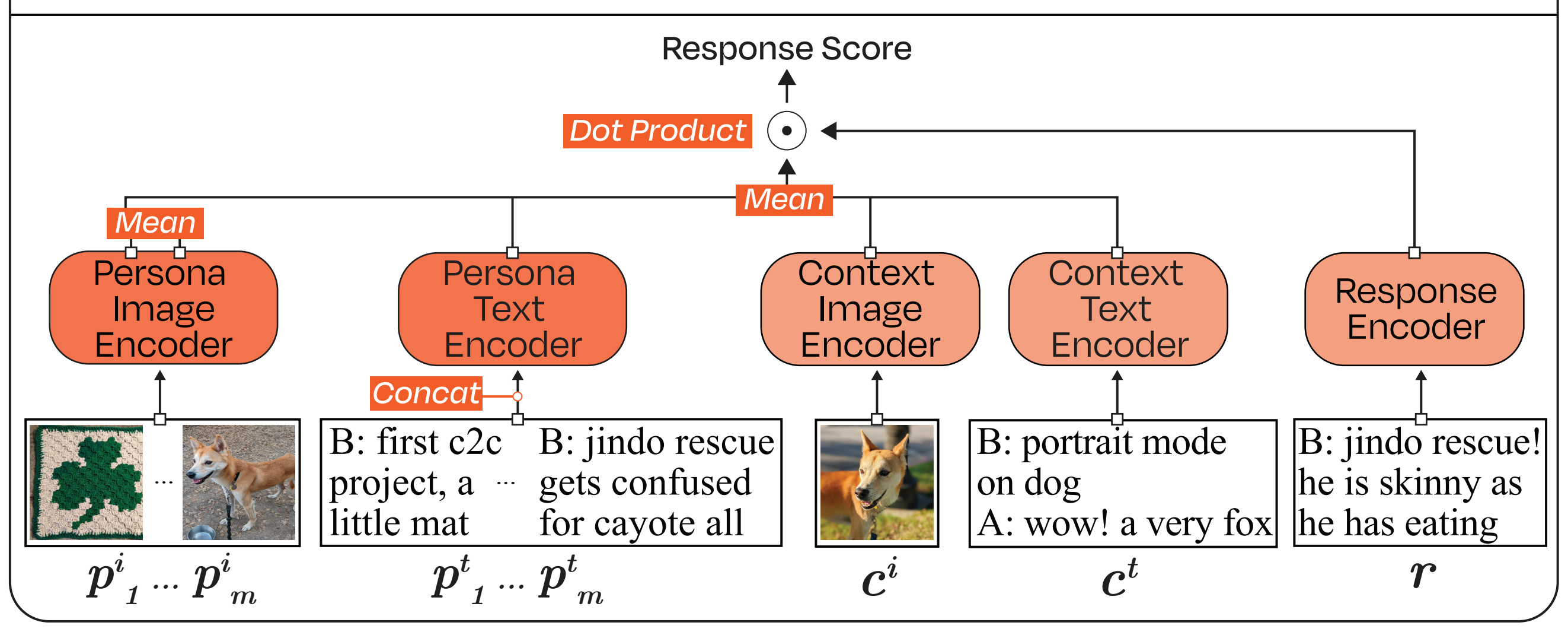
- input: context c, response r, speaker candidates $\mathbb{P}_c$
- output: speaker P

Models

We use separate encoders for each input.

- Image encoder: ViT-B/32, CLIP-ViT-B/32 vision model
- Text encoder: SBERT, CLIP-ViT-B/32 text model

Retrieval-based Dialogue Model

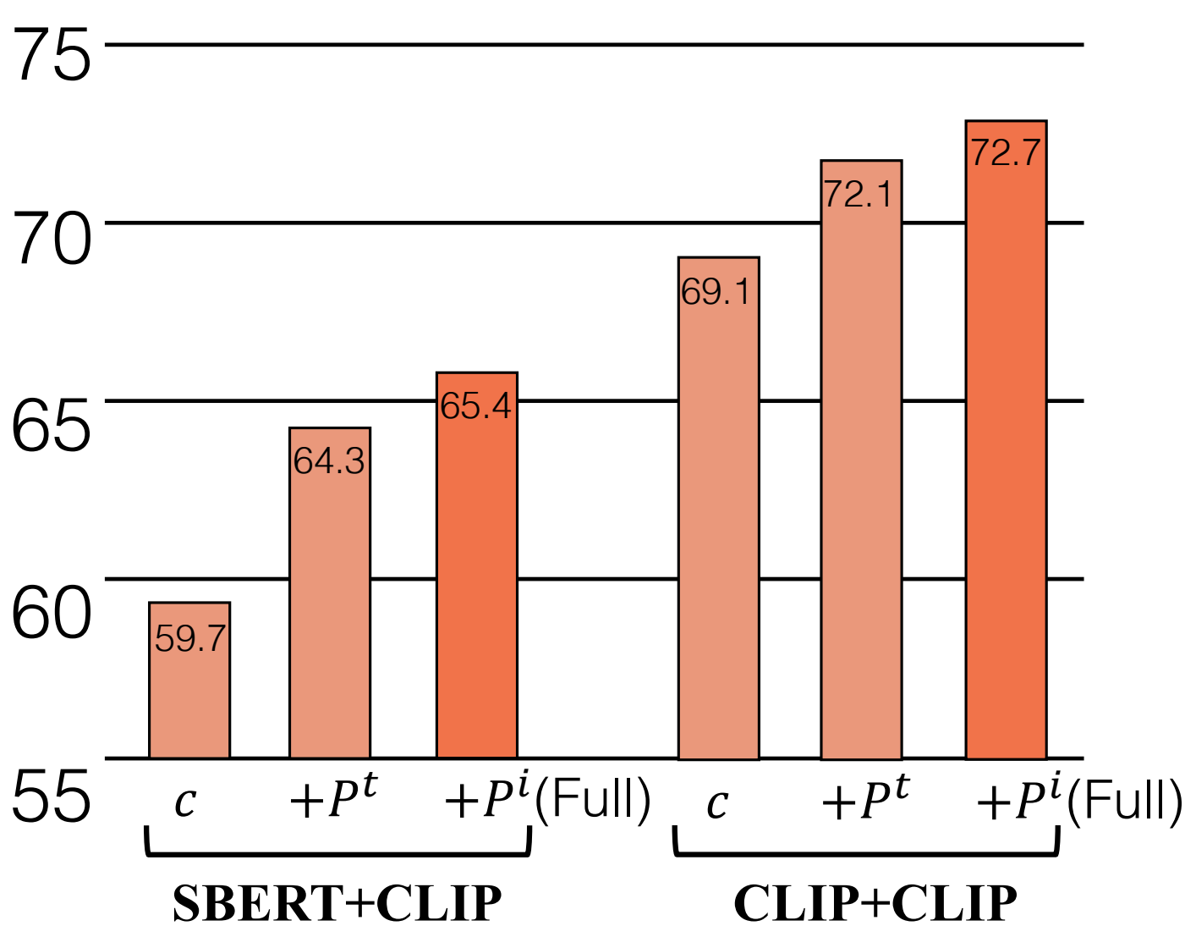


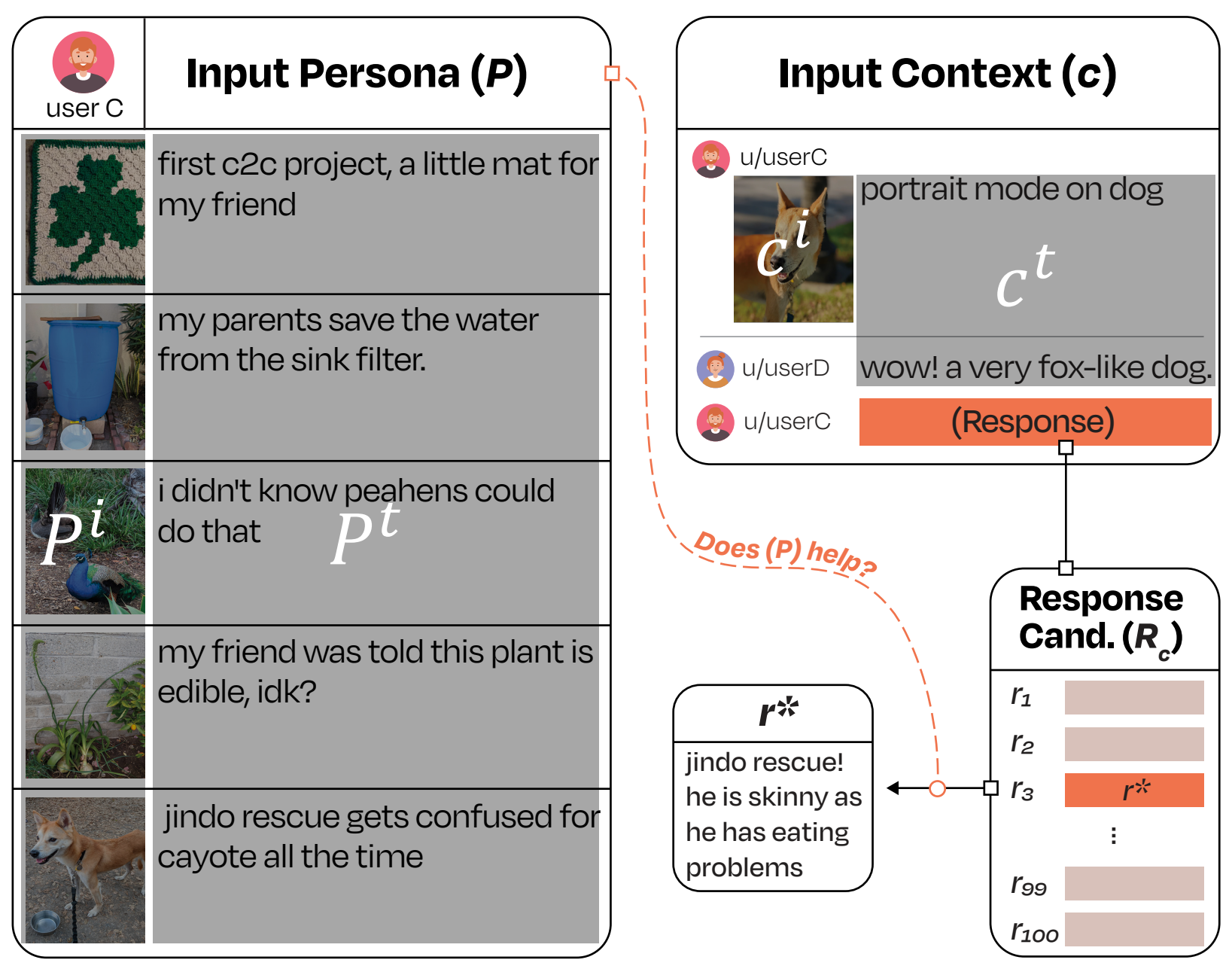
Experimental Results

Results show the superiority of multimodal persona

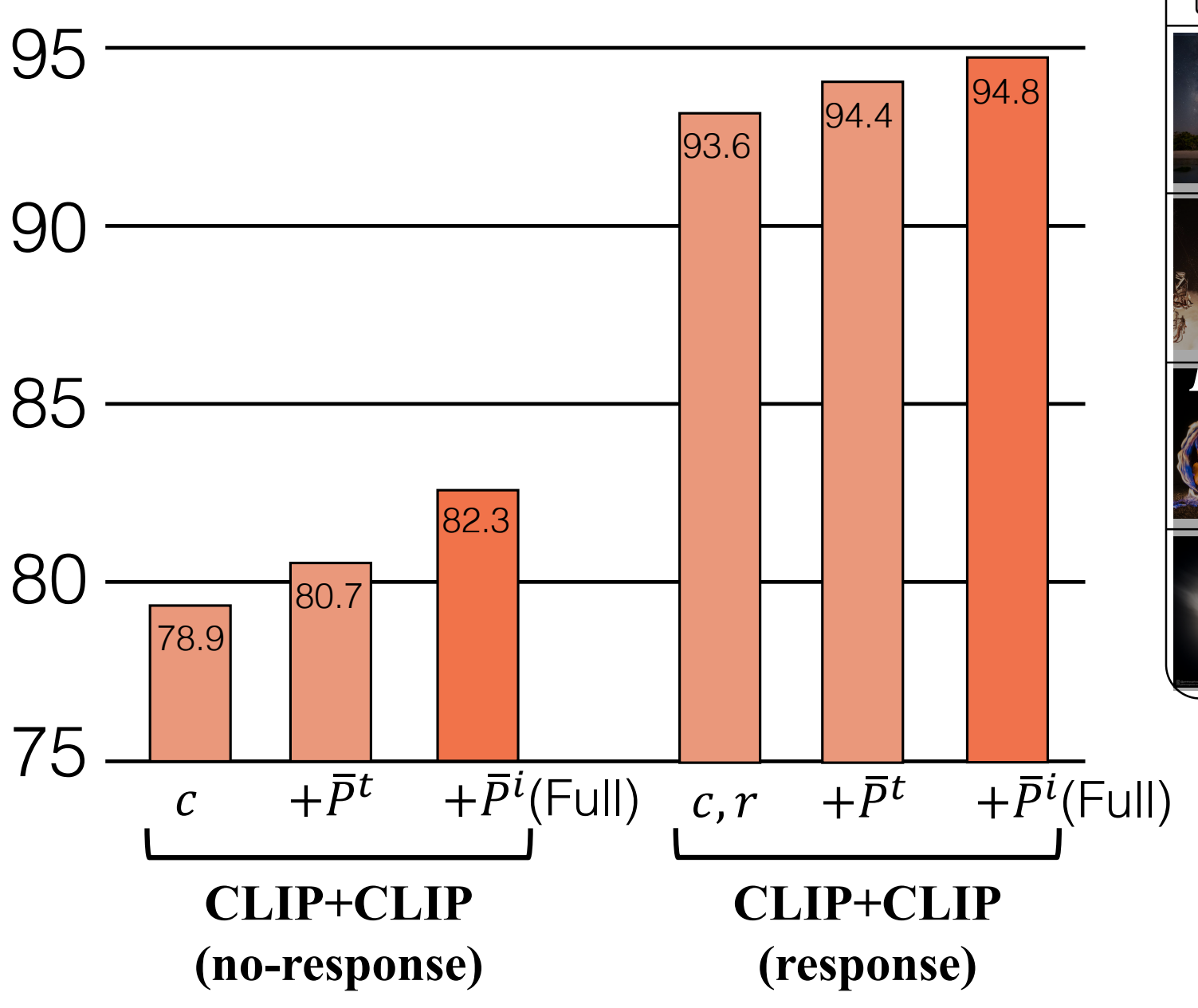
- In all tasks, adding multimodal persona (i.e. P, $\bar{P}$, $\mathbb{P}_c$) leads to statistically significant performance improvement across all models.
- Using either persona images or sentences is consistently better than using only dialogue inputs.

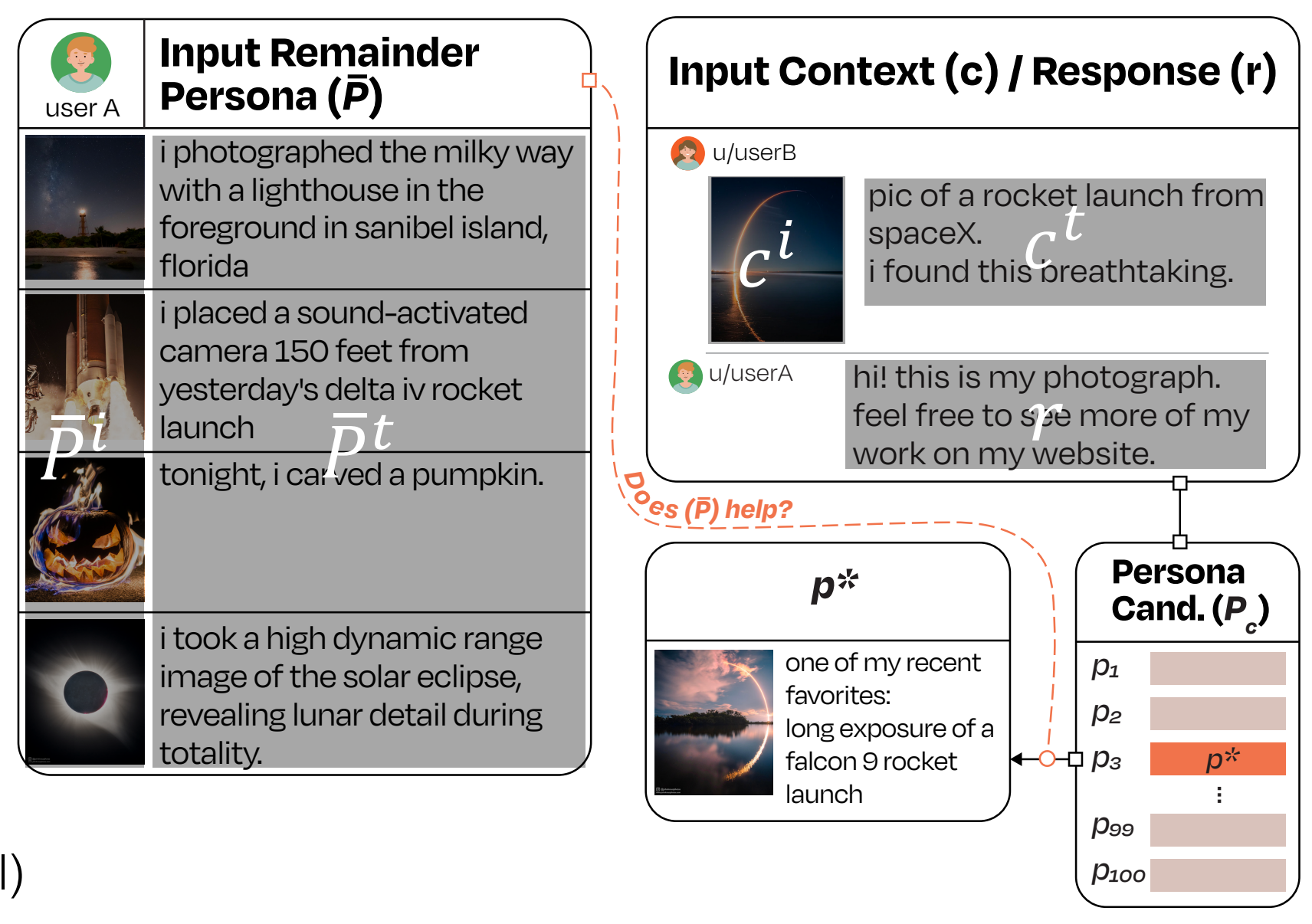
Results on NRP (R@1)



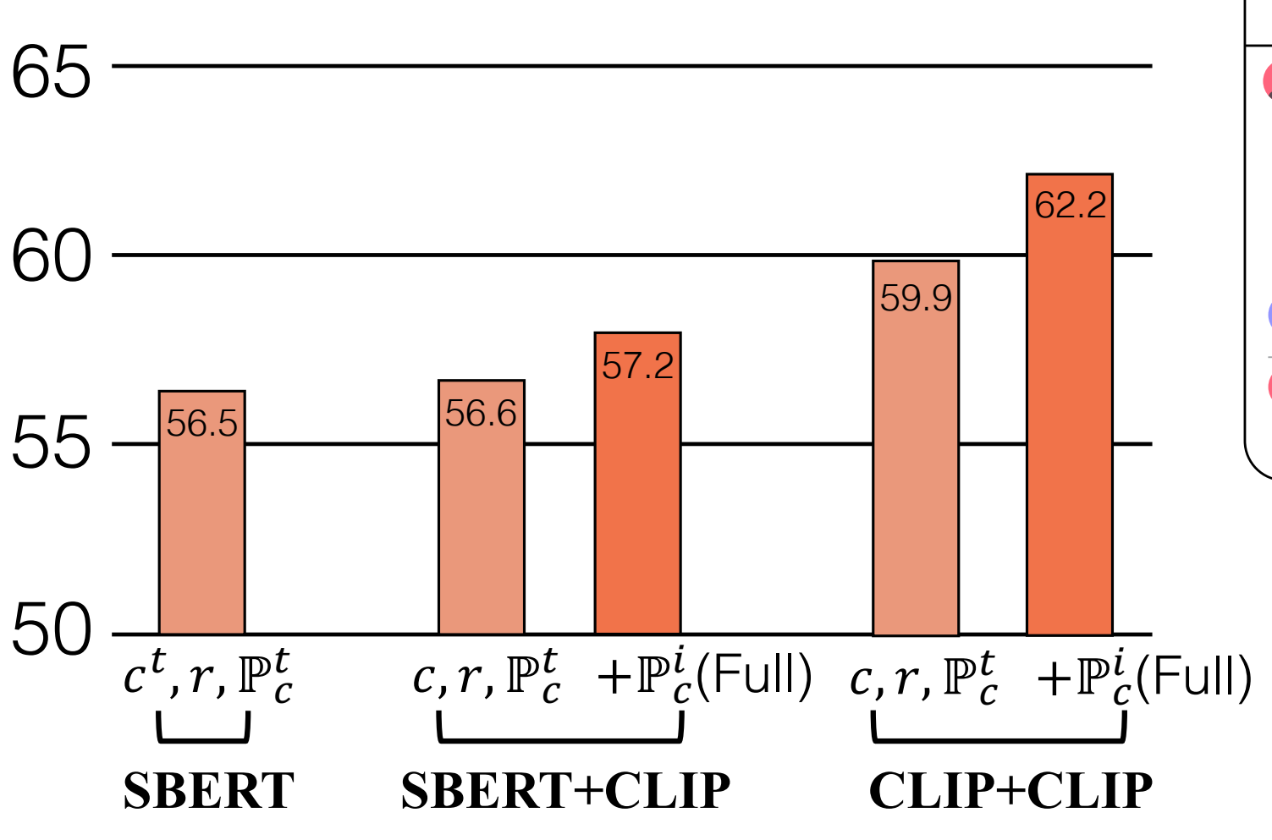

user C
Input Persona (P)
first c2c project, a little mat for my friend
my parents save the water from the sink filter.
i didn't know peahens could do that
my friend was told this plant is edible, idk?
jindo rescue gets confused for cayote all the time
P^i
P^t
Input Context (c)
u/userC
portrait mode on dog
u/userD
wow! a very fox-like dog.
u/userC
(Response)
c^i
c^t
Does (P) help?
Response Cand. (R_c)
r_1
r_2
r_3
r^*
r_{99}
r_{100}
jindo rescue! he is skinny as he has eating problems

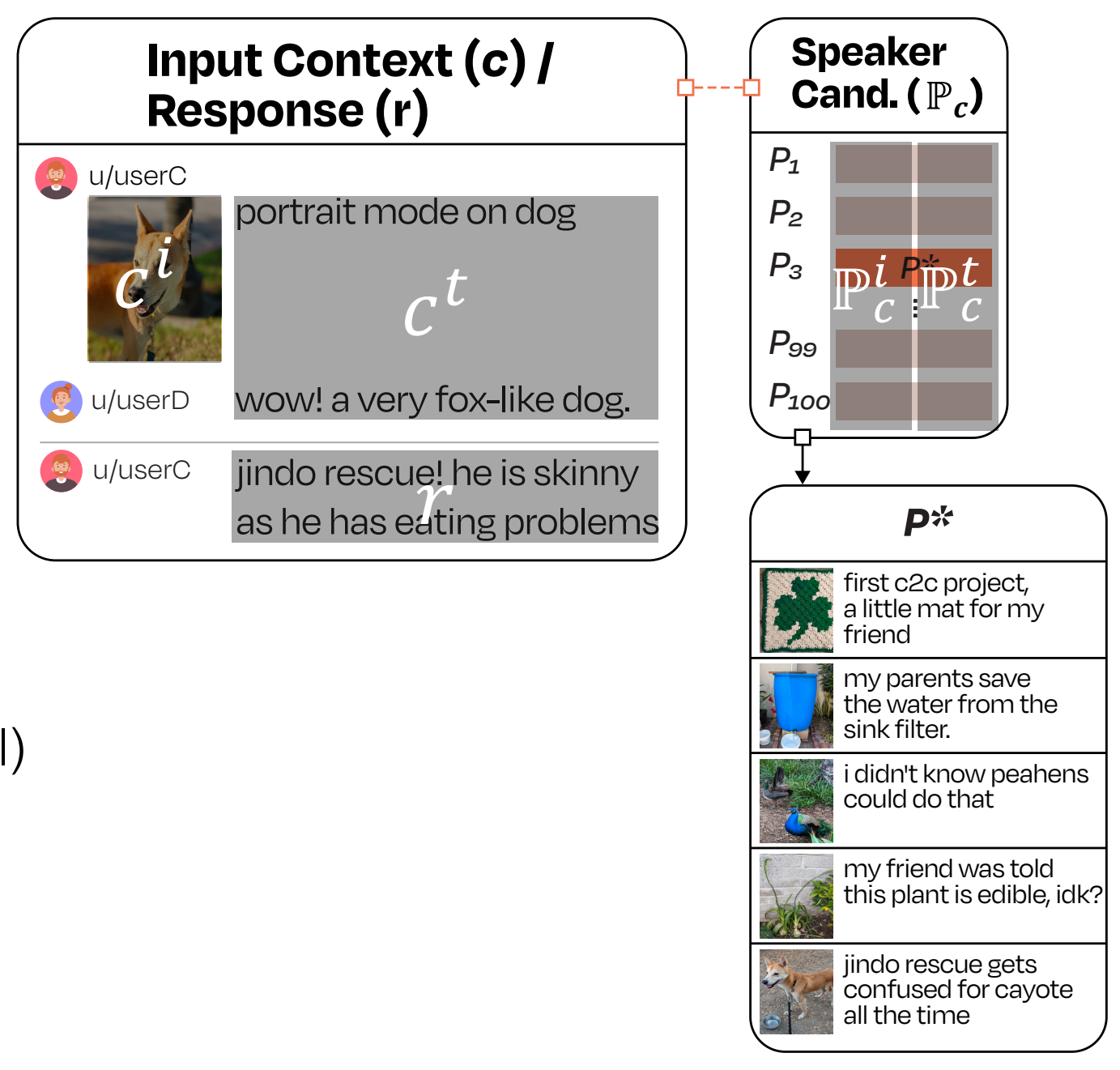
Results on GPP (R@1)




user A
Input Remainder Persona ($\bar{P}$)
i photographed the milky way with a lighthouse in the foreground in sanibel island, florida
i placed a sound-activated camera 150 feet from yesterday's delta iv rocket launch
tonight, i carved a pumpkin.
i took a high dynamic range image of the solar eclipse, revealing lunar detail during totality.
$\bar{P}^i$
$\bar{P}^t$
Input Context (c) / Response (r)
u/userB
pic of a rocket launch from spaceX. i found this breathtaking.
u/userA
hi! this is my photograph. feel free to see more of my work on my website.
c^i
c^t
r
Does ($\bar{P}$) help?
p^*
one of my recent favorites: long exposure of a falcon 9 rocket launch
Persona Cand. (P_c)
p_1
p_2
p_3
p^*
p_{99}
p_{100}

Results on SI (R@1)




Input Context (c) / Response (r)
u/userC
portrait mode on dog
u/userD
wow! a very fox-like dog.
u/userC
jindo rescue! he is skinny as he has eating problems
c^i
c^t
r
Speaker Cand. ($\mathbb{P}_c$)
P_1
P_2
P_3
P_{99}
P_{100}
$\mathbb{P}^i_c$
$\mathbb{P}^t_c$
P^*
first c2c project, a little mat for my friend
my parents save the water from the sink filter.
i didn't know peahens could do that
my friend was told this plant is edible, idk?
jindo rescue gets confused for cayote all the time

Summary

MPCHAT dataset

- The **first** dialogue dataset that supports multimodal persona, representing one's episodic memory
- The responses of speakers are grounded on their multimodal personas

Three new benchmarks

- Incorporating multimodal persona leads to **statistically significant** performance improvements across all tasks