MPCHAT: Towards Multimodal Persona-Grounded Conversation

ACL 2023


Jaewoo Ahn


Yeda Song


Sangdoo Yun


Gunhee Kim

SEOUL NATIONAL UNIV.
VISION & LEARNING

NAVER AI LAB

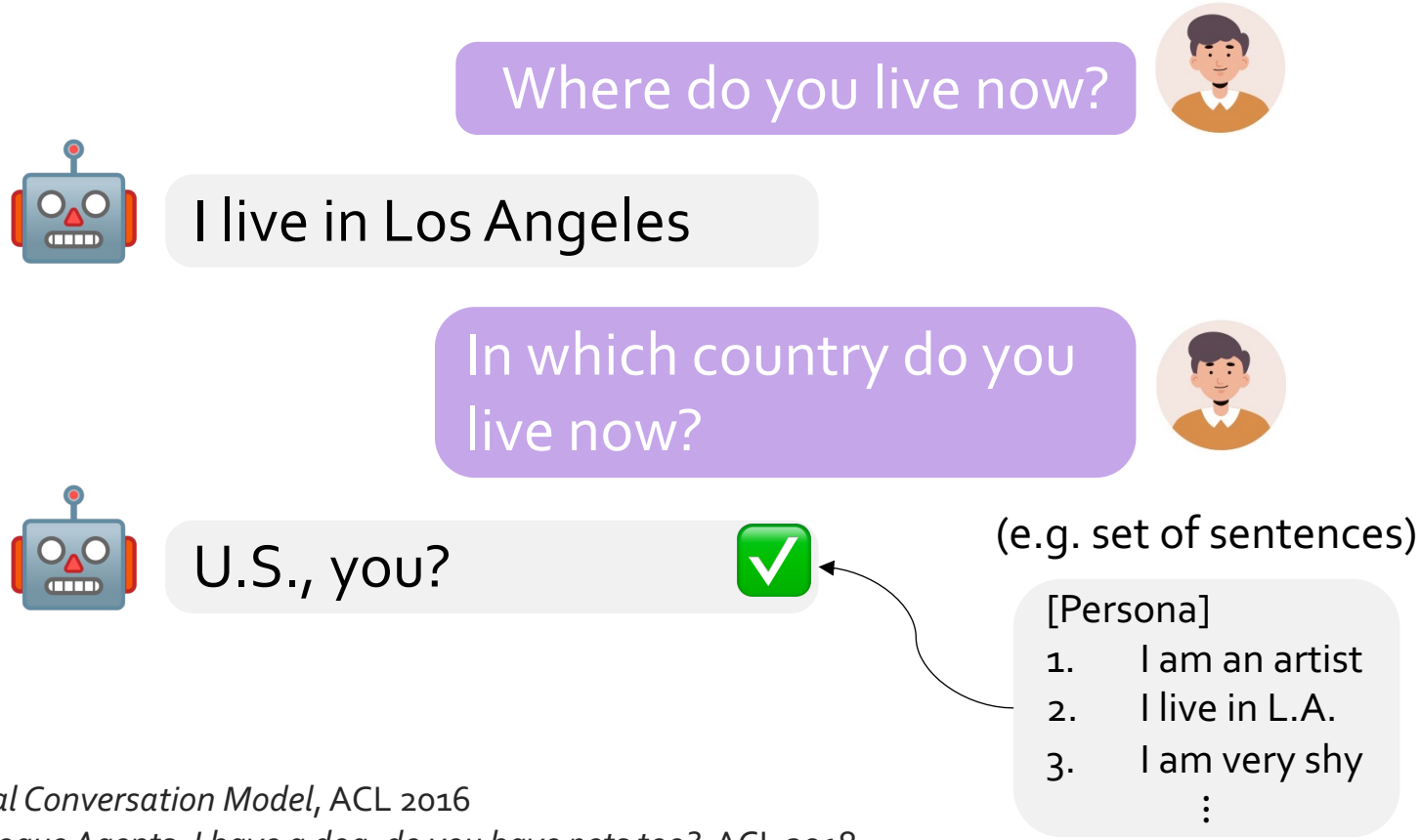
Persona-Grounded Dialogue?

- Dialogue models tend to produce inconsistent responses[1]

[1] Li et al., *A Persona-Based Neural Conversation Model*, ACL 2016
[2] Zhang et al., *Personalizing Dialogue Agents: I have a dog, do you have pets too?*, ACL 2018

Persona-Grounded Dialogue?

- Dialogue models tend to produce inconsistent responses[1]
- Incorporating persona to generate consistent responses[2]



[1] Li et al., *A Persona-Based Neural Conversation Model*, ACL 2016
[2] Zhang et al., *Personalizing Dialogue Agents: I have a dog, do you have pets too?*, ACL 2018

Persona Type

- Previous works have focused on textual persona
  - Personal Facts
  - Personalities

Persona Type (Dataset)	Personal Facts (PersonaChat[2])	Personalities (PELD[3])
Format	Character description using 5 sentences	Strength of big-five personality: Openness, Conscientiousness, Extraversion, Agreeableness, Neuroticism
Example	1. I like to ski 2. My wife doesn't like anymore 3. I am an artist 4. I am on a diet now 5. I have a cat	[0.648, 0.375, 0.386, 0.58, 0.477]

[2] Zhang et al., *Personalizing Dialogue Agents: I have a dog, do you have pets too?*, ACL 2018
[3] Wen et al., *Automatically Select Emotion for Response via Personality-affected Emotion Transition*, ACL Findings 2021

Persona Type

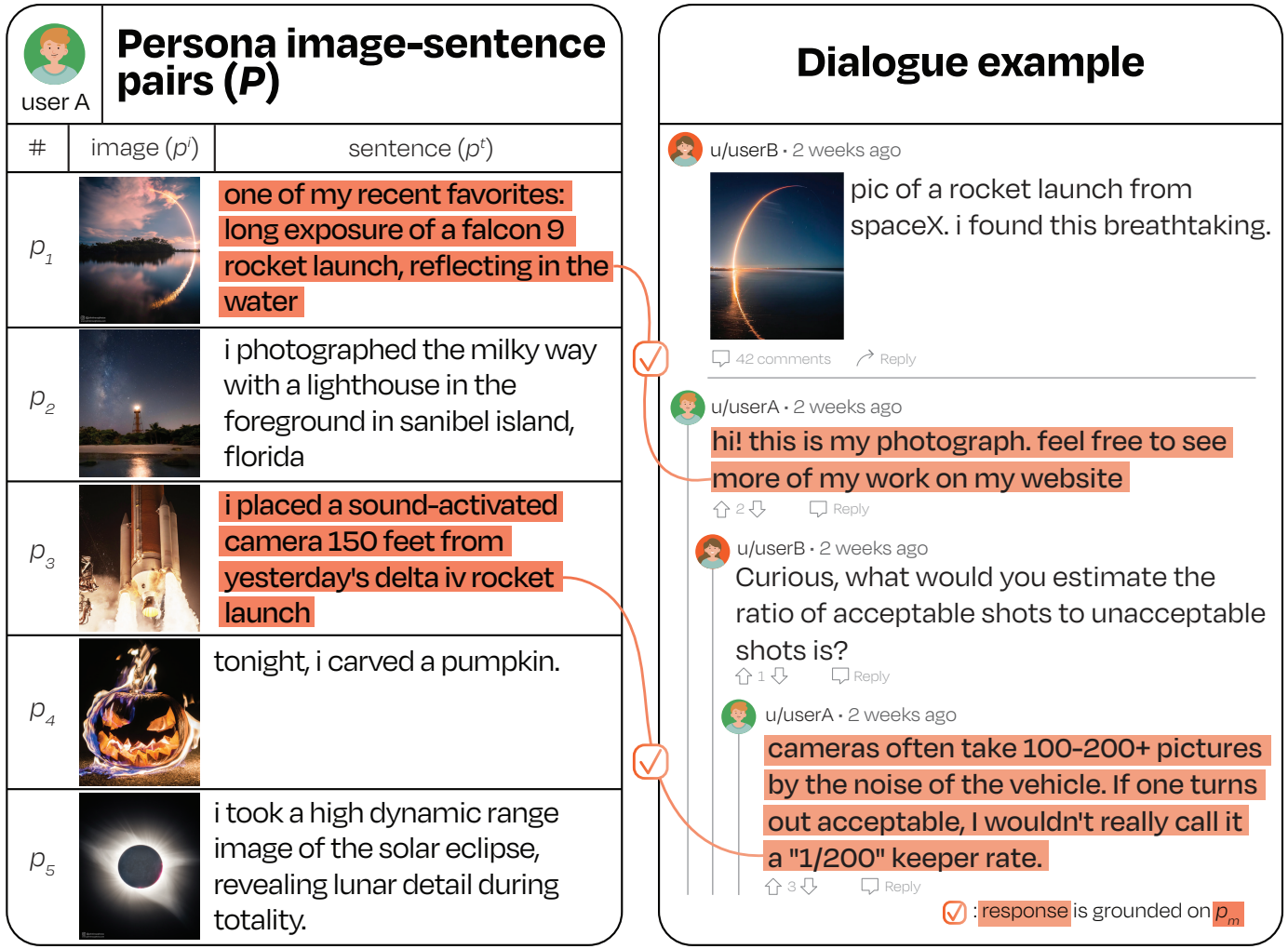
- However, persona should be explored in multi-faceted ways[4]
  - Episodic memory is important in shaping personal identity[5]
    - Memory of everyday events or personal experiences[6]
    - Represented in the form of visual images[7]
- We propose multimodal persona, a set of image-sentence pairs

MPCHAT	PersonaChat
• i gave my computer setup a christmas themed overhaul 	• i love computers • i work as a computer programmer • i work at home on my computer • i love rpg computer games ⋮
• i think we found doggie uptopia. 	• i have a dog • i love dogs • i walk dogs for a living • i enjoy log walks with my dog ⋮

[4] Moore et al., *Five dimensions of online persona*, Persona Studies 2017
[5] Wilson and Ross, *The identity function of autobiographical memory: Time is on our side*, Memory 2003
[6] Tulving, *Episodic and Semantic Memory*, Organization of Memory 1972
[7] Conway., *Episodic memories*, Neuropsychologia 2009

Towards Multimodal Persona-Grounded Dialogue

- MPCHAT dataset
  - Sourced from reddit
  - Multimodal persona reveals one's episodic memories
  - Responses are grounded on persona image-sentence pairs



MPCHAT: Statistics

- Total of 15K multi-turn dialogues
- Avg. # of persona: 17.87
- Avg. length of persona sent.: 10.14
- Avg. length of utterances: 18.49

	Train	Valid	Test
# Dialogue	11,975	1,516	1,509
# Speaker	21,197	2,828	2,797
# Utterance	34,098	4,189	4,244
# Persona Speaker	8,891	1,193	1,162
# Grounded Response	6,628	709	676
# Avg. Persona	15.89	25.6	30.76
# Avg. Subreddits	4.2	5.97	5.88
Avg. Utterance Length	18.39	18.74	19.05
Avg. Persona Length	10.16	10.23	10.02

MPCHAT: Multimodal Persona

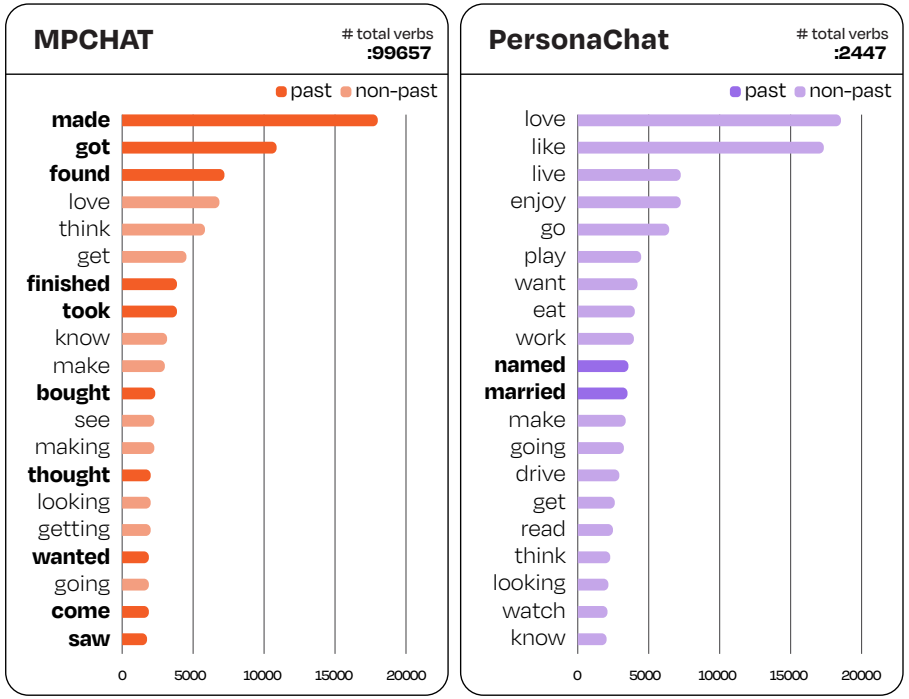
- Only MPCHAT supports both textual and visual persona
- MPCHAT provides persona entailment labels

Dataset	# Dialogue	Data source	Persona type	Persona modality	Entailment label
LIGHT[8]	11K	Crowd-sourced	Fact	T	No
PD[9]	20.8M	Weibo	Fact	T	No
PEC[10]	355K	Reddit	Thought	T	No
PELD[3]	6.4K	TV shows	Personality	T	No
PersonaChat[2]	13K	Crowd-sourced	Fact	T	Post-Hoc
FoCus[11]	14K	Crowd-sourced	Fact	T	Yes
MPCHAT	15K	Reddit	Episodic memory	V, T	Yes

[2] Zhang et al., *Personalizing Dialogue Agents: I have a dog, do you have pets too?*, ACL 2018
[3] Wen et al., *Automatically Select Emotion for Response via Personality-affected Emotion Transition*, ACL Findings 2021
[8] Urbanek et al., *Learning to speak and act in a fantasy text adventure game*, EMNLP 2019
[9] Zheng et al., *Personalized dialogue generation with diversified traits*, arXiv 2019
[10] Zhong et al., *Towards persona-based empathetic conversational models*, EMNLP 2020
[11] Jang et al., *Call for customized conversation: Customized conversation grounding persona and knowledge*, AAAI 2022

MPCHAT: Persona Statistics

- Episodic-memory-based persona
  - Lots of past tense verbs
  - Lexically diverse



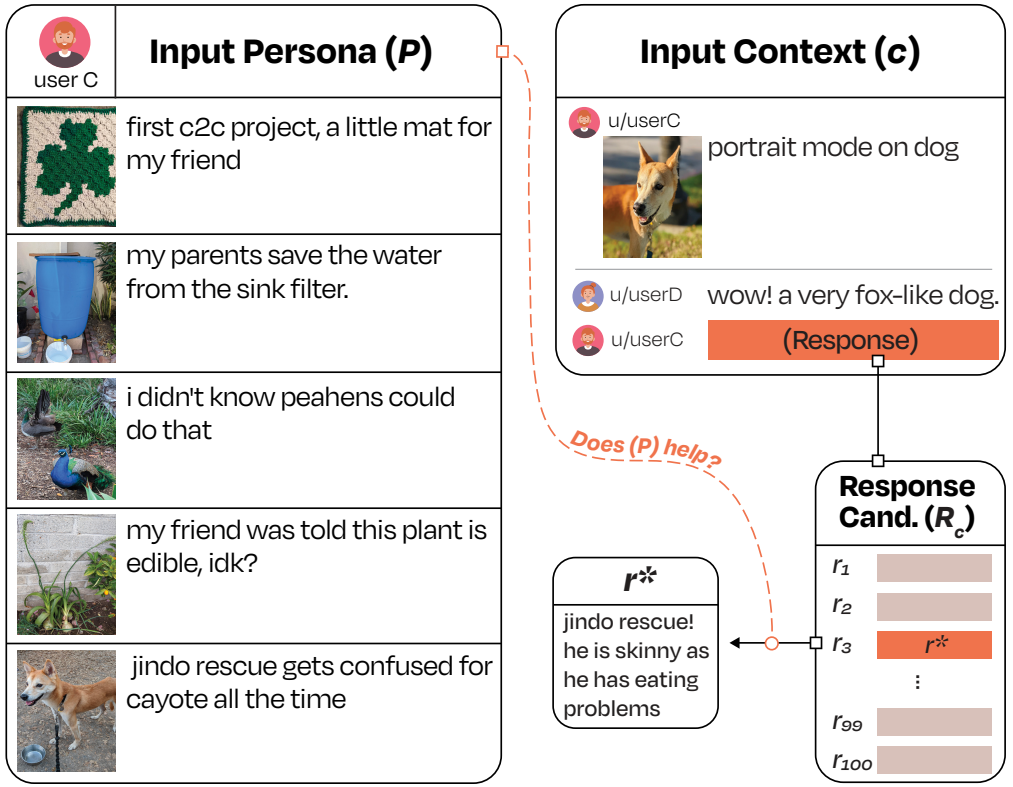
Dataset	# 2-grams	# 3-grams	# 4-grams	MTLD	MATTR	HD-D
PersonaChat[2]	15,263	27,631	36,063	78.08	0.7791	0.7945
PEC[10]	34,051	54,649	62,290	111.39	0.811	0.8315
MPCHAT	**39,694**	**60,199**	**66,732**	**171.91**	**0.8534**	**0.8674**

[2] Zhang et al., *Personalizing Dialogue Agents: I have a dog, do you have pets too?*, ACL 2018
[10] Zhong et al., *Towards persona-based empathetic conversational models*, EMNLP 2020

MPChat: Three Benchmarks

1) Next Response Prediction (NRP)

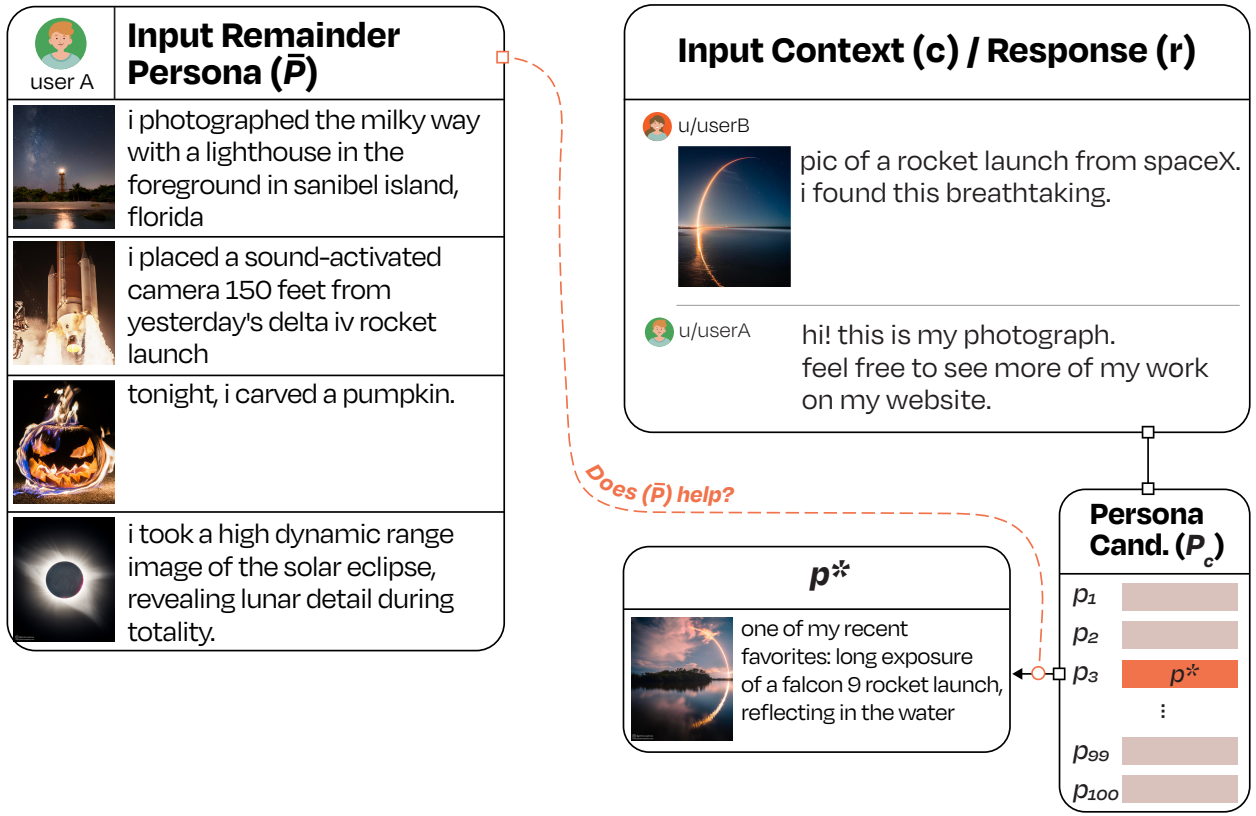
- Input: context c, multimodal persona P, response candidates R_c
- Output: response r



MPCHAT: Three Benchmarks

2) Grounding Persona Prediction (GPP)

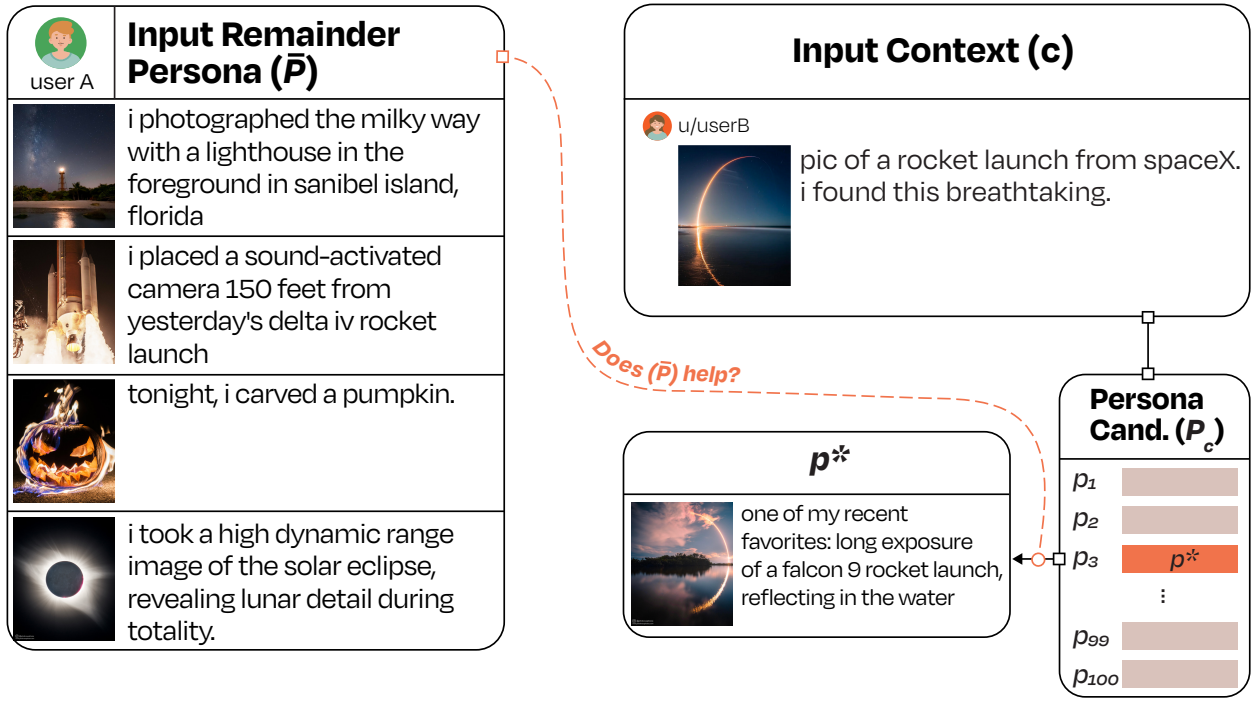
- Predict speaker's grounding persona element based on dialogue info
- "response" case
  - Input: context c, response r, remainder persona set $\bar{P}$, persona candidates P_c
  - Output: persona element p



MPCHAT: Three Benchmarks

2) Grounding Persona Prediction (GPP)

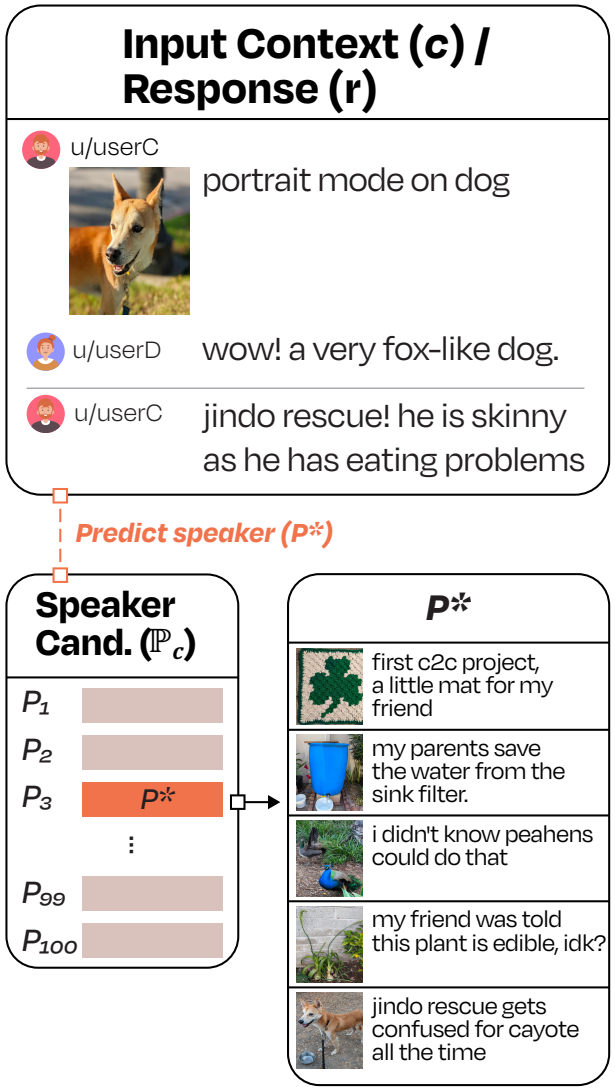
- Predict speaker's grounding persona element based on dialogue info
- "no-response" case
  - Input: context c, ~~response r~~, remainder persona set $\bar{P}$, persona candidates P_c
  - Output: persona element p



MPCHAT: Three Benchmarks

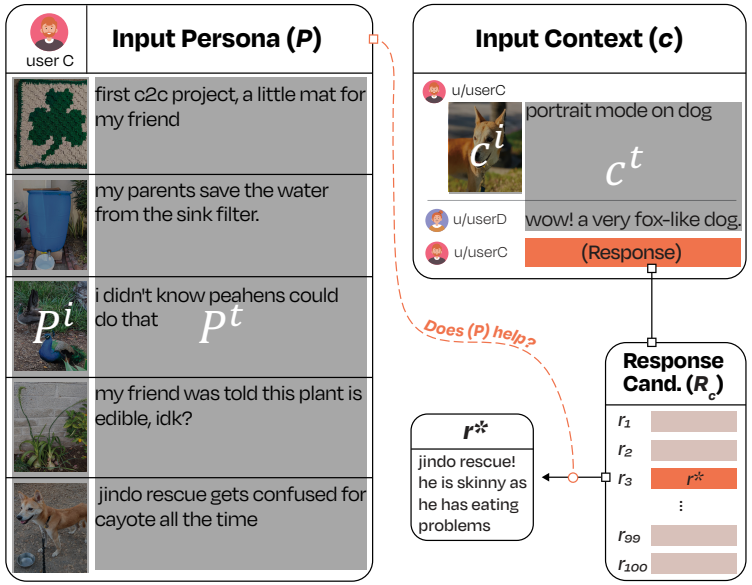
3) Speaker Identification (SI)

- Predict speaker based on dialogue info
- Input: context c, response r, speaker candidates $\mathbb{P}_c$
- Output: speaker P



Quantitative Results on NRP

- Model w/ multimodal persona outperforms baseline



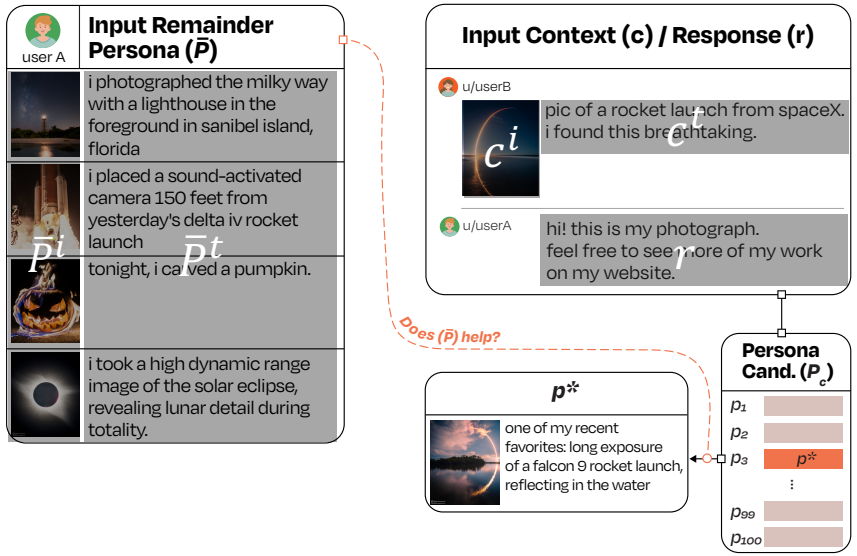
P^i: persona images
P^t: persona sentences
P: $P^i \cup P^t$

c^i: context image
c^t: context text
c: $c^i \cup c^t$

Model	R@1↑	MRR↑
Text Only (c^t)		
IR Baseline	10.69	18.06
SBERT (zero-shot)	35.67	45.75
SBERT	51.32±1.32	64.76±0.92
SBERT+ViT (text+image encoder)		
c	57.7±0.71	69.39±0.4
c, P^i	58.55±0.7	70.17±0.45
c, P^t	64.32±0.64	74.3±0.45
c, P (Full)	**65.29±0.66****	**75.08±0.43****
SBERT+CLIP		
c	59.68 ±0.7	70.99 ±0.49
c, P^i	60.3±0.5	71.47±0.27
c, P^t	64.32±0.75	74.33±0.57
c, P (Full)	**65.43±0.42****	**75.19±0.32****
CLIP+CLIP		
c^i (zero-shot)	39.38	54.06
c^i	40.85±0.64	54.32±0.3
c	69.11±0.74	78.22±0.49
c, P^i	69.87±0.4	78.85±0.27
c, P^t	72.13±0.61	80.72±0.38
c, P (Full)	**72.65±0.38***	**81.12±0.26***

Quantitative Results on GPP

- Model w/ multimodal persona outperforms baseline



$\bar{P}^i$: remainder persona images
$\bar{P}^t$: remainder persona sentences
$\bar{P}$: $\bar{P}^i \cup \bar{P}^t$

c^i: context image
c^t: context text
c: $c^i \cup c^t$

Model	no-response		response (+r)	
	R@1↑	MRR↑	R@1↑	MRR↑
SBERT+ViT				
c	70.91±0.7	79.26±0.47	95.06±0.32	97.12±0.17
$c, \bar{P}^i$	70.7±0.9	79.17±0.57	95.16±0.55	97.21±0.29
$c, \bar{P}^t$	73.87±0.65	81.41±0.34	94.86±1.35	97.09±0.78
$c, \bar{P}$ (Full)	**74.43±0.64***	**82.05±0.39****	**95.75±0.53****	**97.58±0.3****
SBERT+CLIP				
c	70.98±0.94	79.28±0.56	94.99±0.55	97.06±0.31
$c, \bar{P}^i$	70.63±1.03	79.22±0.71	94.91±0.44	97.04±0.24
$c, \bar{P}^t$	74.06±0.68	81.52±0.42	94.92±0.42	97.13±0.26
$c, \bar{P}$ (Full)	**74.69±0.62***	**82.24±0.41****	**95.55±0.58***	**97.48±0.32****
CLIP+CLIP				
c	78.85±1.04	85.96±0.67	93.56±0.56	96.21±0.37
$c, \bar{P}^i$	82.02±0.89	88.31±0.58	94.62±0.48	96.86±0.32
$c, \bar{P}^t$	80.69±0.8	87.28±0.55	94.43±0.45	96.79±0.23
$c, \bar{P}$ (Full)	**82.32±0.75**	**88.52±0.46**	**94.79±0.5**	**96.94±0.28**

Quantitative Results on SI

- Model w/ multimodal persona outperforms baseline

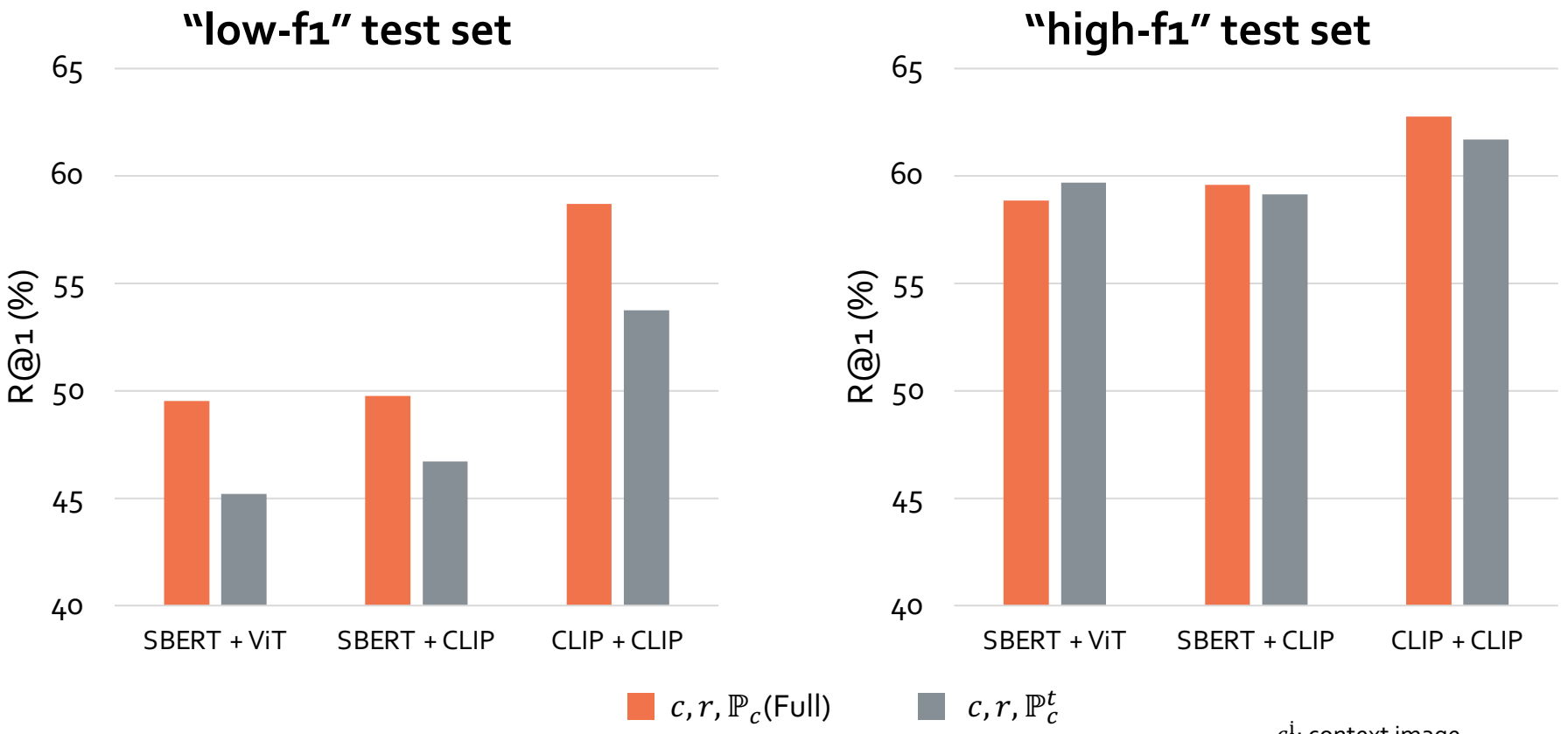
Model	R@1↑	MRR↑
Text Only $(c^t, r, \mathbb{P}^t_c)$		
SBERT	56.47±0.58	67.92±0.52
SBERT+ViT		
$c, r, \mathbb{P}^i_c$	19.56±0.64	35.84±0.45
$c, r, \mathbb{P}^t_c$	56.87±0.6	68.33±0.37
$c, r, \mathbb{P}_c$ (Full)	**57.28±0.44**	**68.86±0.3****
SBERT+CLIP		
$c, r, \mathbb{P}^i_c$	25.71±0.49	42.47±0.34
$c, r, \mathbb{P}^t_c$	56.63±0.66	68.15±0.42
$c, r, \mathbb{P}_c$ (Full)	**57.24±0.63***	**68.69±0.39****
CLIP+CLIP		
$c, r, \mathbb{P}^i_c$	44.27±0.66	59.04±0.35
$c, r, \mathbb{P}^t_c$	59.89±0.71	70.87±0.53
$c, r, \mathbb{P}_c$ (Full)	**62.17±0.56****	**73.08±0.35****

$\mathbb{P}^i_c$: speakers' persona images
$\mathbb{P}^t_c$: speakers' persona sentences
$\mathbb{P}_c$: $\mathbb{P}^i_c \cup \mathbb{P}^t_c$

c^i: context image
c^t: context text
c: $c^i \cup c^t$
r: response

When is multimodal persona helpful?

- SI: Larger gap in "low-f1" test set (same trend in NRP)



c^i: context image
c^t: context text
c: $c^i \cup c^t$
r: response

$\mathbb{P}_c^i$: speakers' persona images
$\mathbb{P}_c^t$: speakers; persona sentences
$\mathbb{P}_c$: $\mathbb{P}_c^i \cup \mathbb{P}_c^t$

Concluding Remarks

- Limitations of persona type and modality

  → Represent personal facts or personalities through textual persona

- Towards episodic-memory-based multimodal persona

  → MPCHAT: Multimodal persona-grounded dialogue dataset
  & propose three benchmarks: NRP, GPP, SI

- Outperforms baselines on all tasks w/ multimodal persona

  → MPCHAT is a high-quality resource, given its well-grounded dialogues on multimodal personas

Thank you

Code https://github.com/ahnjaewoo/mpchat

Paper https://arxiv.org/abs/2305.17388

Contact jaewoo.ahn@vision.snu.ac.kr