MPCHAT: Towards Multimodal Persona-Grounded Conversation

Jaewoo Ahn, Yeda Song, Sangdoo Yun and Gunhee Kim

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

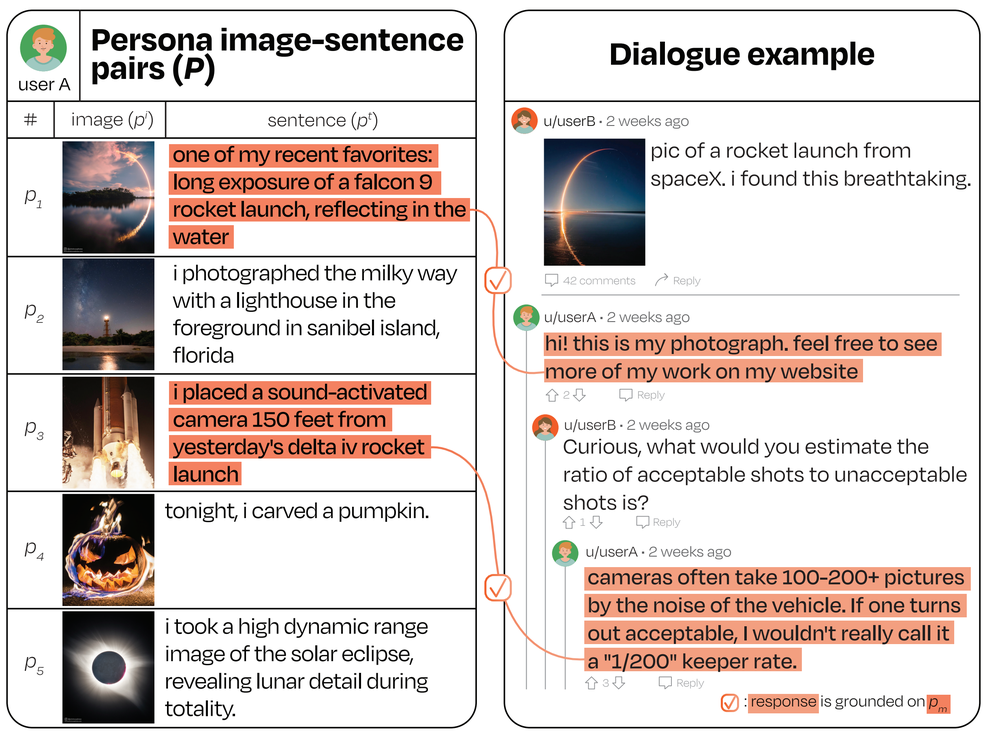
Motivation

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

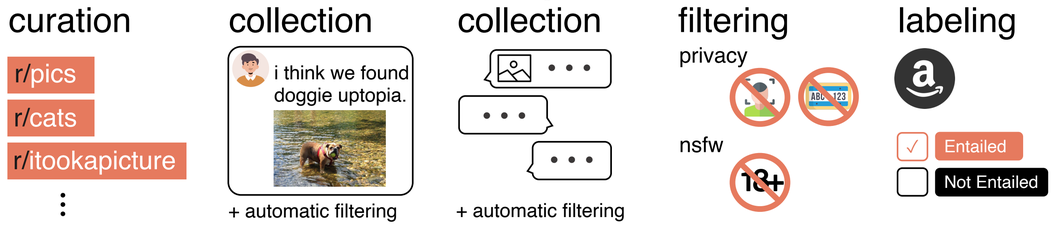


We introduce MPCHAT, a new multimodal persona-grounded dialogue dataset

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



Benchmarks

We propose three retrieval-based tasks in MPCHAT

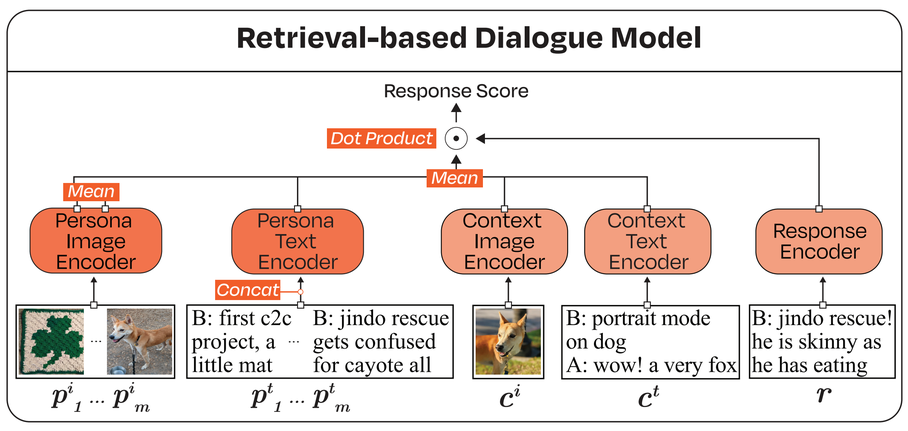
1) Next Response Prediction (NRP)
predict next response based on context and multimodal persona.
- input: context c, multimodal persona P, response candidates R_c
- output: response r

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

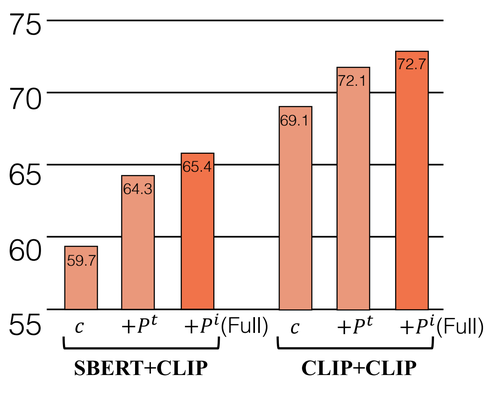
We use separate encoders for each input.
- Image encoder: ViT-B/32, CLIP-ViT-B/32 vision model
- Text encoder: SBERT, CLIP-ViT-B/32 text model

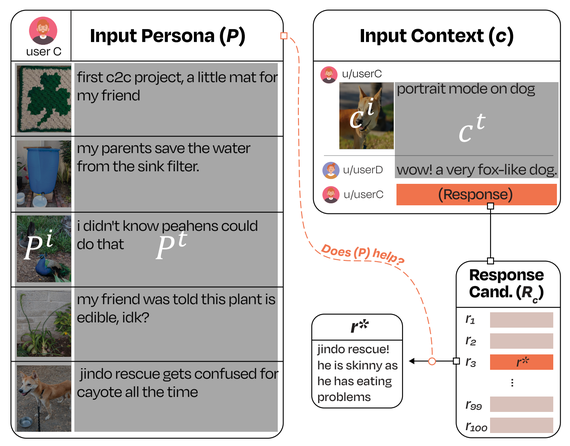


Experimental Results

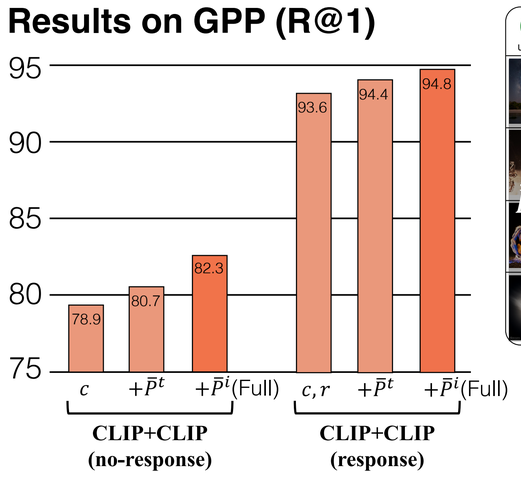
Results show the superiority of multimodal persona
- In all tasks, adding multimodal persona (i.e. P, $\bar{P}$, $\mathbb{P}_c$) leads to statistically significant performance improvement across all models.
- Using either persona images or sentences is consistently better than using only dialogue inputs.

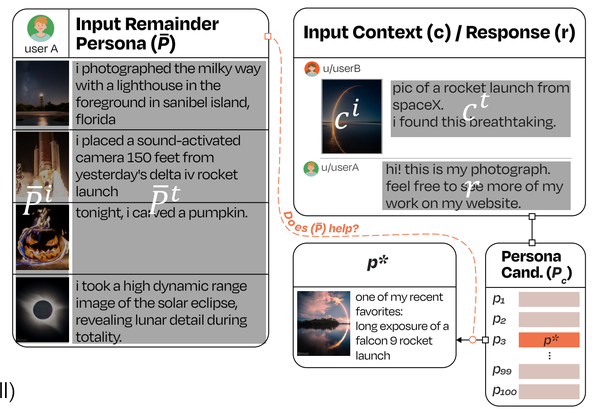
Results on NRP (R@1)



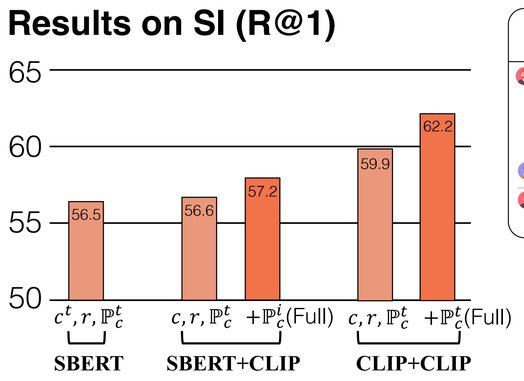


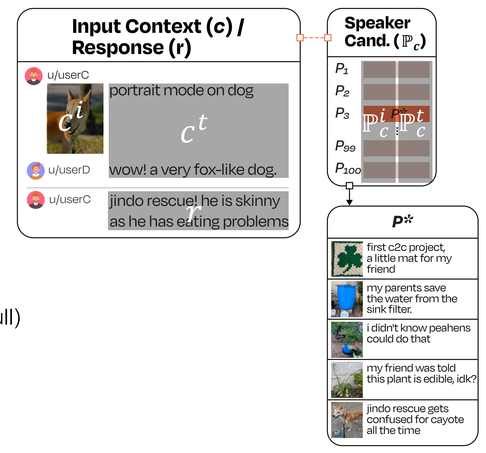
Results on GPP (R@1)





Results on SI (R@1)





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