The Power of Many: Multi-Agent Multimodal Models For Cultural Image Captioning

Longju Bai*, Angana Borah*, Oana Ignat*, Rada Mihalcea

Contact: longju@umich.edu

Motivation

- LMM's effectiveness in **cross-cultural contexts** remains limited - W.E.I.R.D.*(Western, Educated, Industrialized, Rich, Democratic—a concept from psychology)*
- **Multi-agent collaboration** has proven to be highly capable for solving complex tasks.
- Culture is strongly tied to our **group-oriented human nature** - culturally enriched image captioning task as a "social task.

Contributions

1. We create a **multi-agent framework** 'MosAIC' using different cultural personas for better cultural generations.
2. We form a **comprehensive evaluation pipeline** for culturally enriched image captioning task
3. We open source a dataset of: culturally enriched image captions in English and also our **human annotations across cultures**.

Dataset

GDVCR:

- East-Asia, South-Asia, West
- **Movie** scenes
- **Rich** cultural contents

GeoDE:

- China, Romania
- **Real life** objects
- **Less** cultural contents

CVQA:

- China, India, Romania
- **Real life** scenes
- **Rich** cultural contents

Evaluation

Auto-Metrics:

- **Cultural Info**: # cultural words mentioned in the caption
- **Completeness score**: (# objects mentioned in the caption) / (# all objects from RAM)
- **Alignment**: LongCLIP Score on the first sentence of the caption

Humanmetrics:

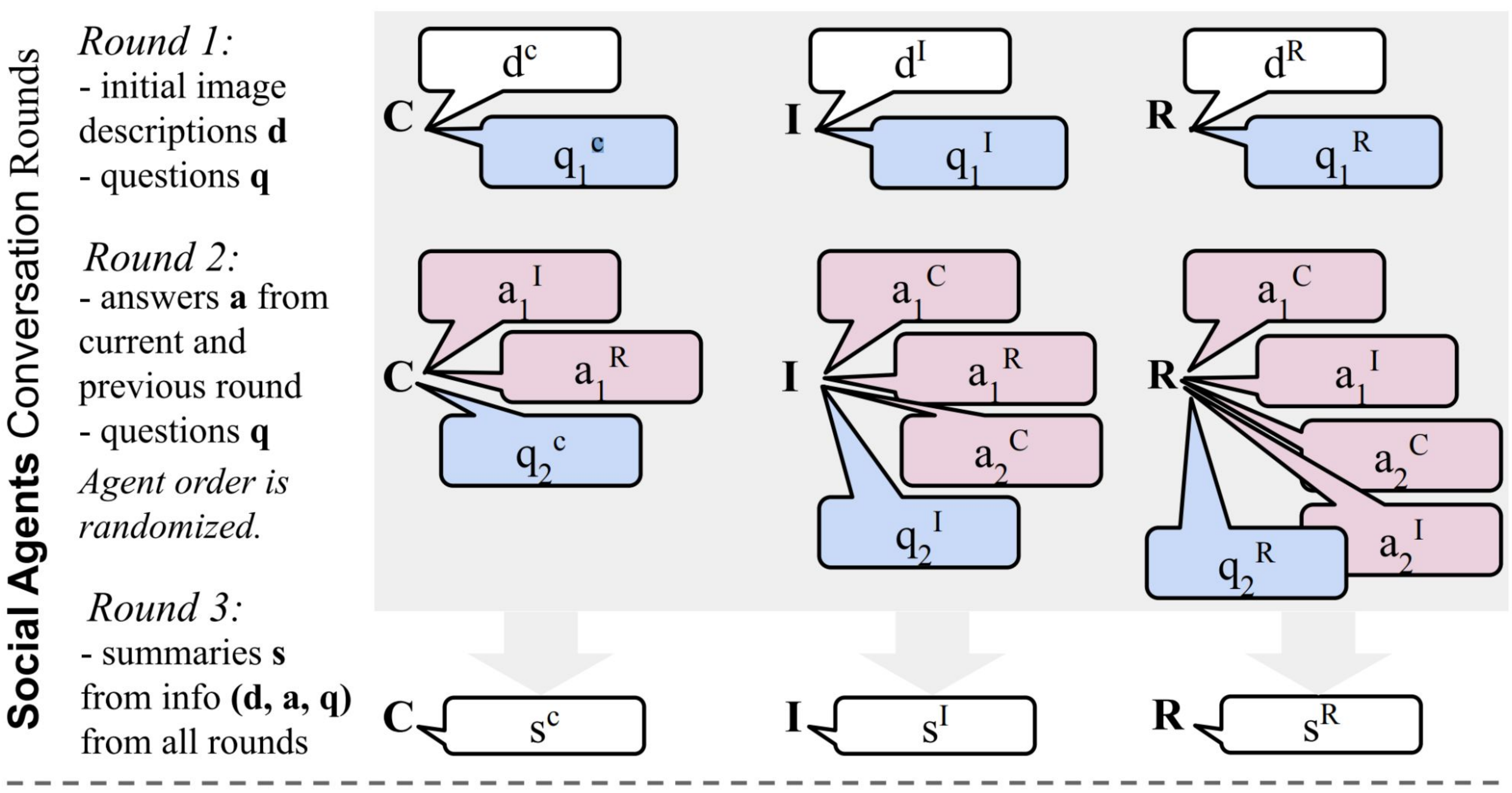
- **Human-likeness**: The accuracy of annotators in distinguishing between human-generated and machine-generated captions
- **Correctness**: Both correctness for image contents caption, and the correctness of the cultural description

https://github.com/MichiganNLP/MosAIC

[Paper Link]

Multi-Agent Collaboration Framework

Moderator Agent generates questions **q** for the **Social Agents C, I, R**



Social Agents Conversation Rounds

Round 1:
- initial image descriptions **d**
- questions **q**

Round 2:
- answers **a** from current and previous round
- questions **q**

Agent order is randomized.

Round 3:
- summaries **s** from info **(d, a, q)** from all rounds

Summarizer Agent summarizes $\mathbf{s^C, s^I, s^R}$ as final image caption.

Fig. The Moderator presents questions to the Social agents, who engage in three conversation rounds. The Summarizer creates the final image caption by compiling the conversation summaries from the Social agents.

Experiment Results and Ablation Study

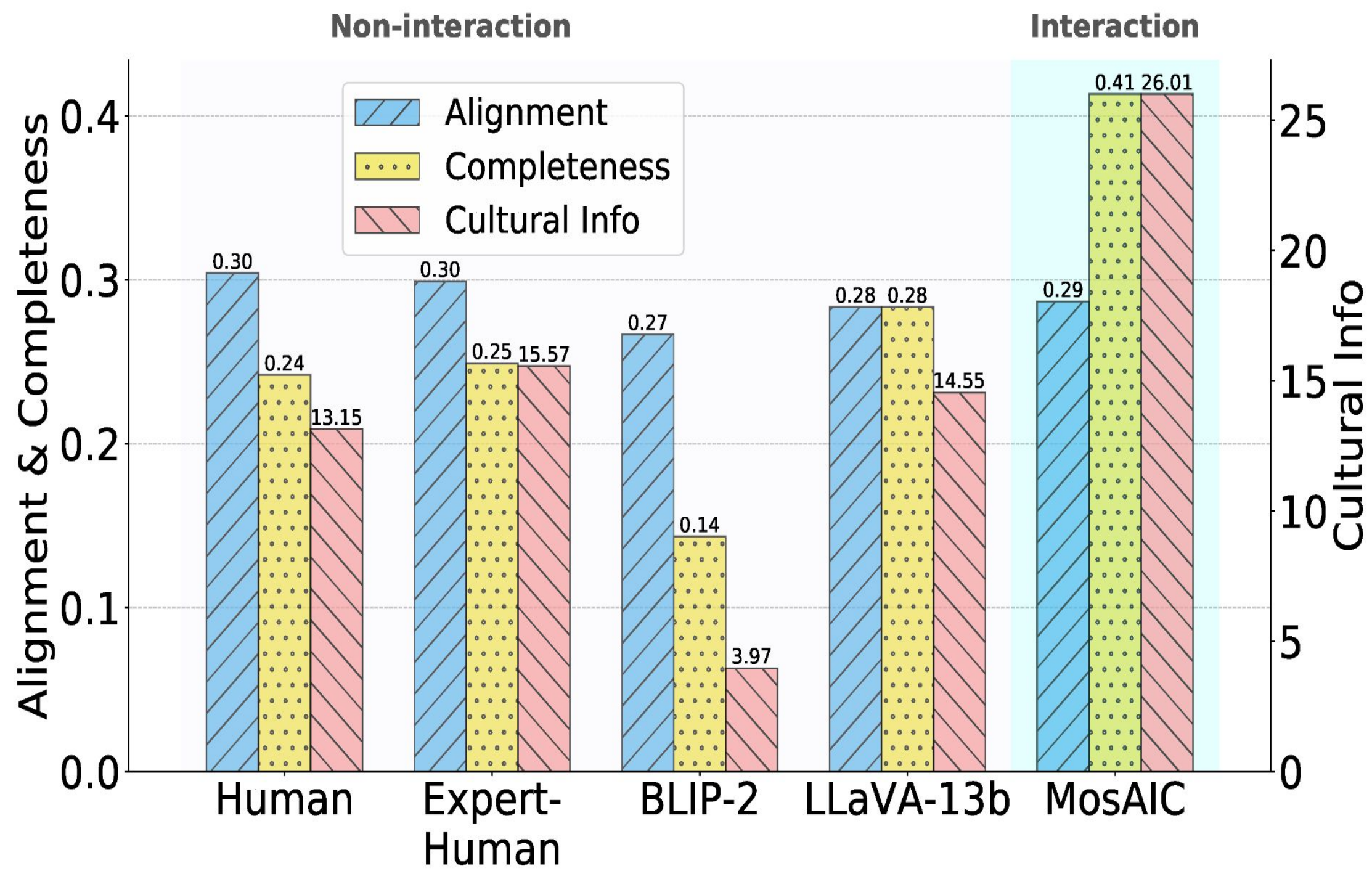


Fig. The multi-agent system surpasses non-interaction models and Humans on Completeness and Cultural Info while performing on par with the other models in Alignment.

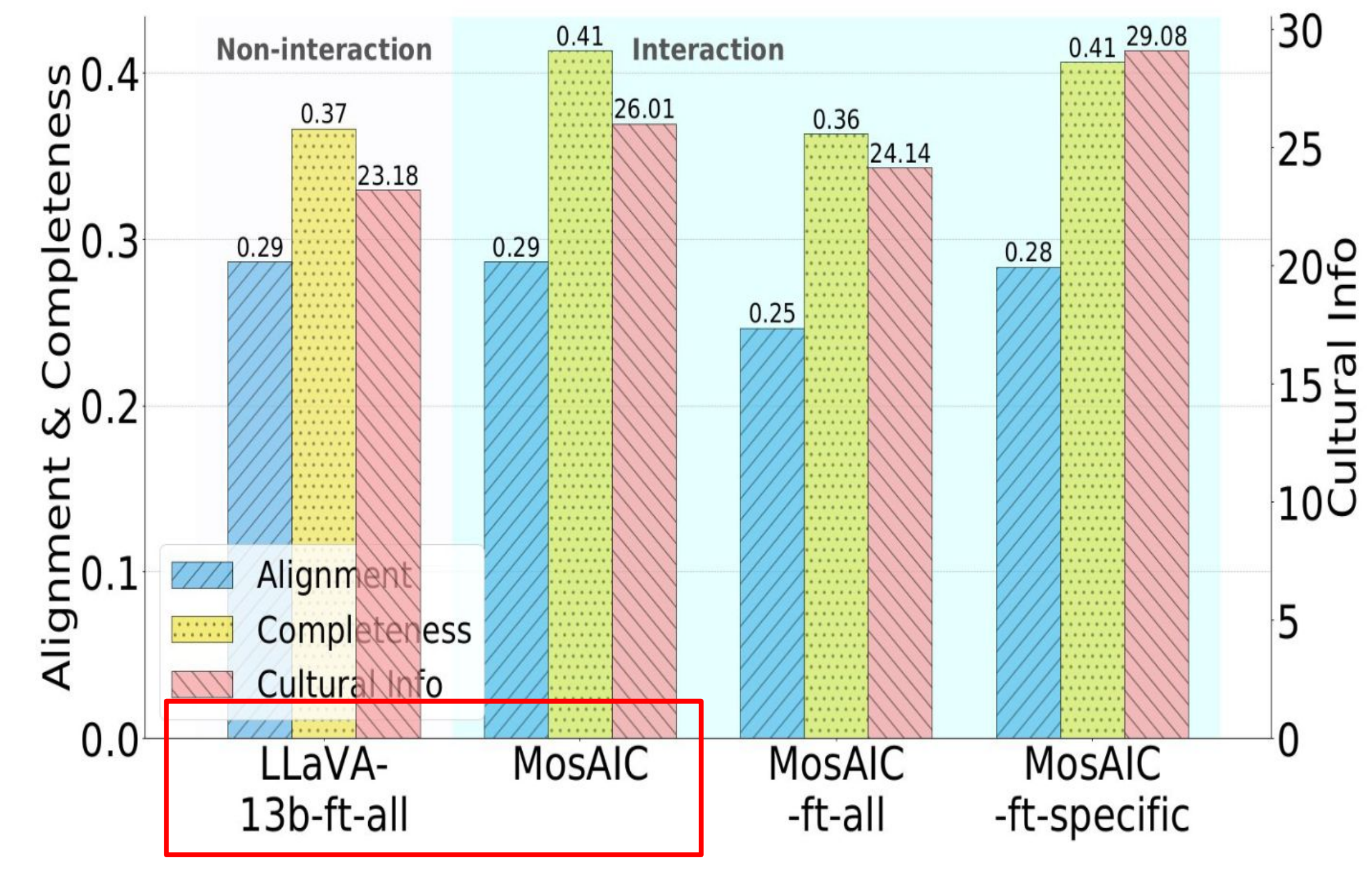


Fig. Ablation study across fine-tuning: The zero-shot multi-agent model MosAIC outperforms the fine-tuned single-agent model LLaVA-13b-ft-all

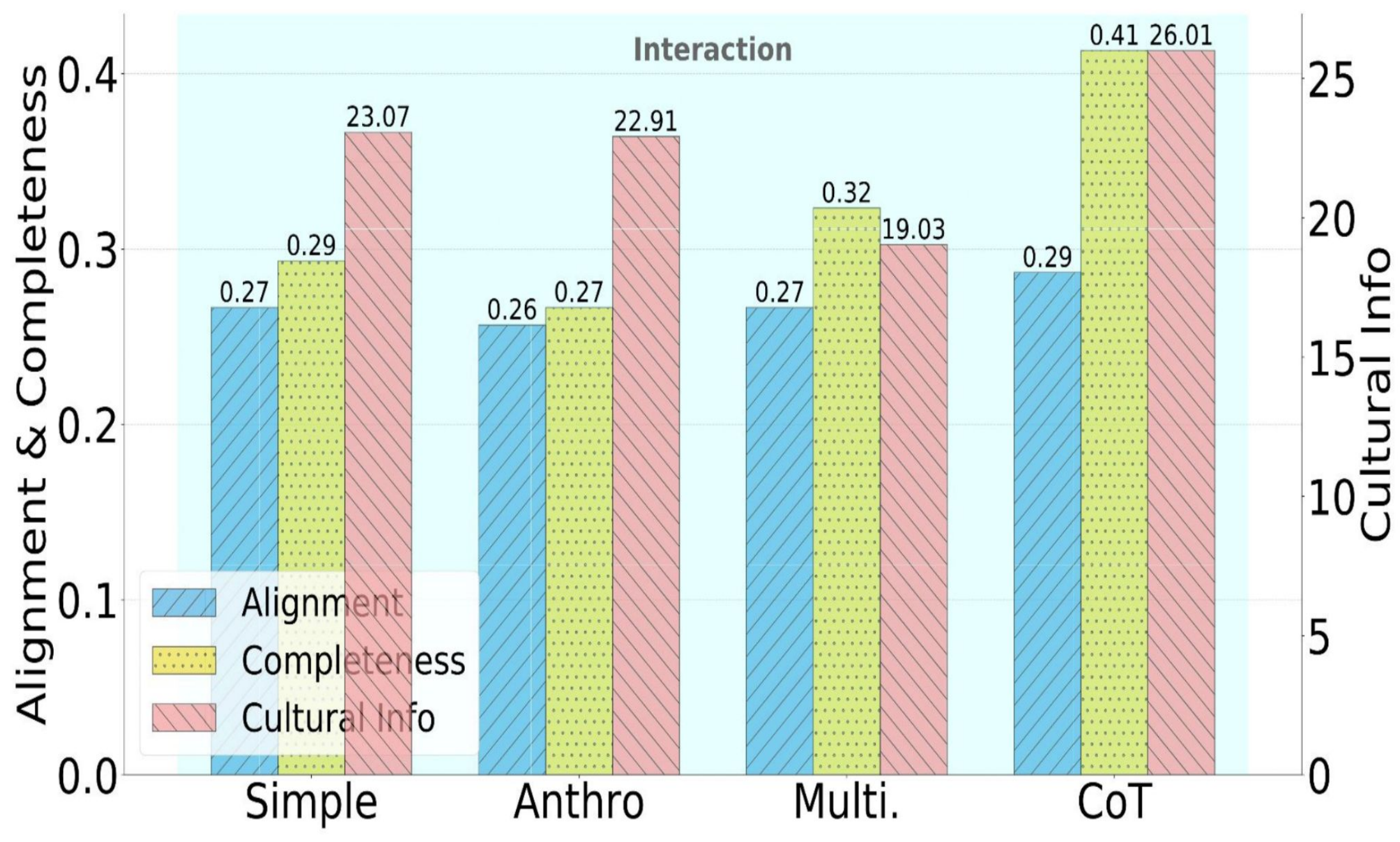


Fig. Ablation study across prompt: The CoT prompt outperforms all other types of prompt

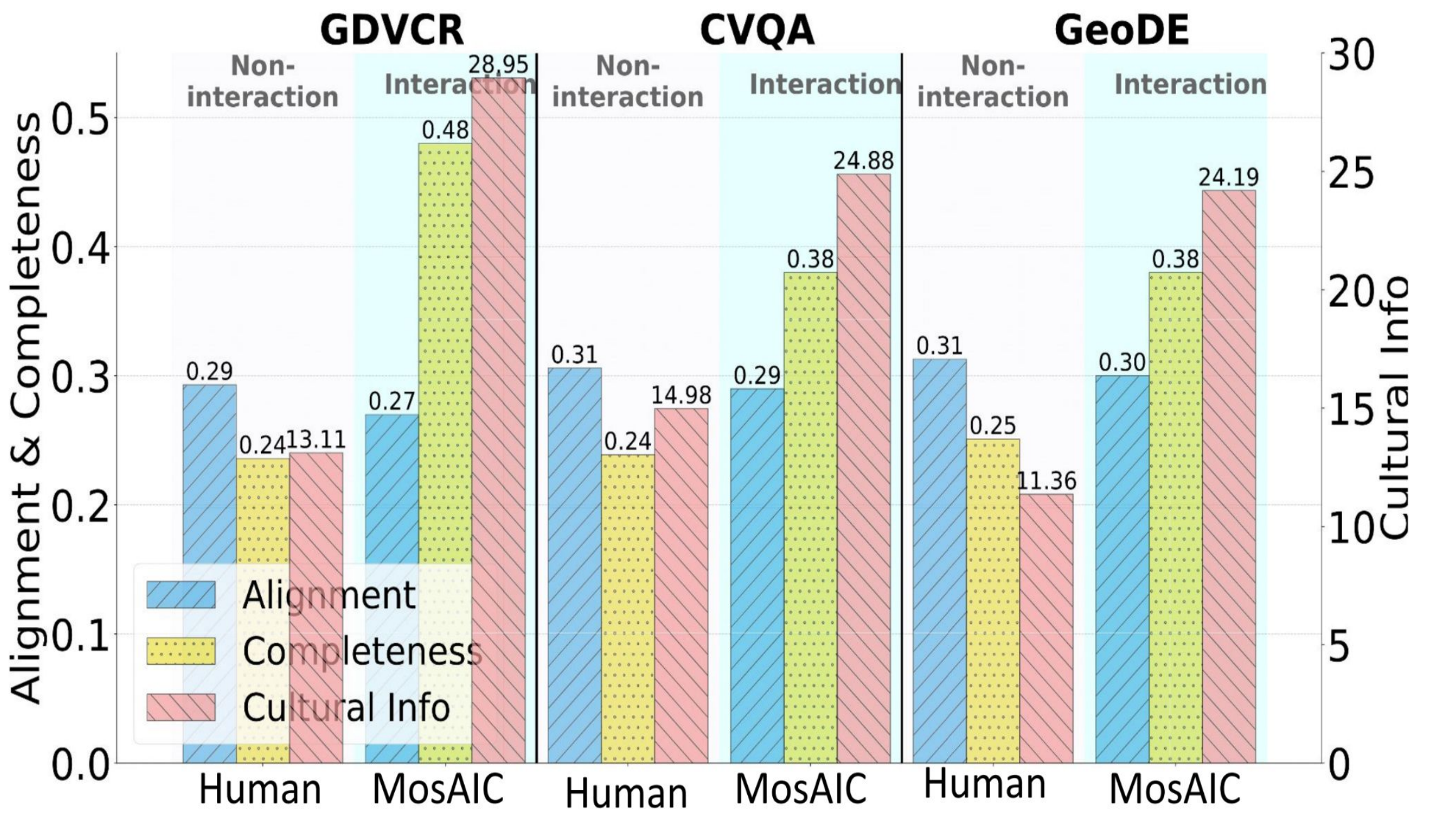


Fig. Ablation study across datasets: The proposed framework performs best on GDVCR and CVQA where the images have more complex cultural contexts

Main Takeaways

- Interaction-based **multi-agent collaboration** system outperforms single model.
- Multi-agent system has **better data and time efficiency** than single model fine-tuning.
- **Task-wise generalization**: multi-agent for simulation or task solving, tool-based agents.