

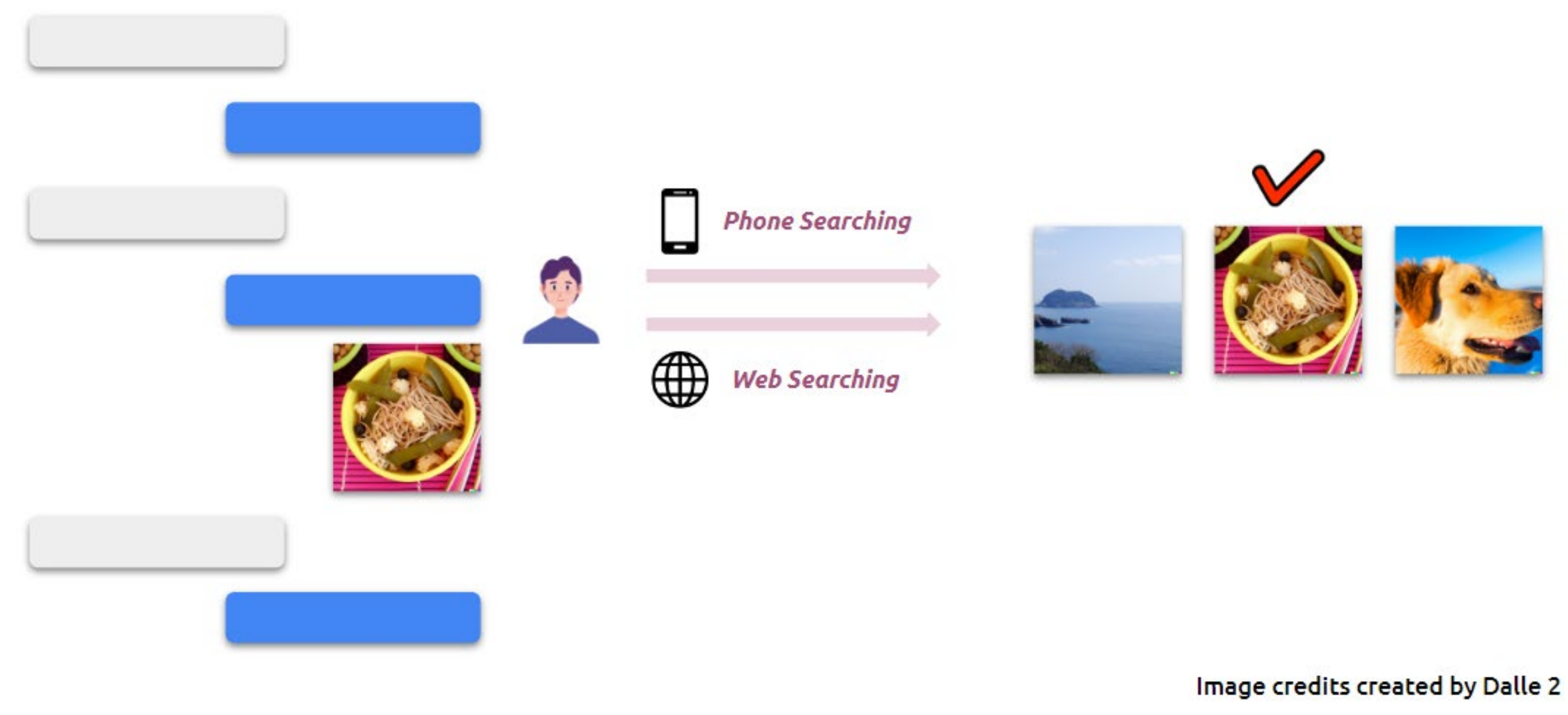
DialogCC: An Automated Pipeline for Creating High-Quality Multi-modal Dialogue Datasets

Young-Jun Lee¹, Byungsoo Ko², Han-Gyu Kim³, Jonghwan Hyeon¹, Ho-Jin Choi¹

¹School of Computing, KAIST ²NAVER Vision ³NAVER Cloud AI Works

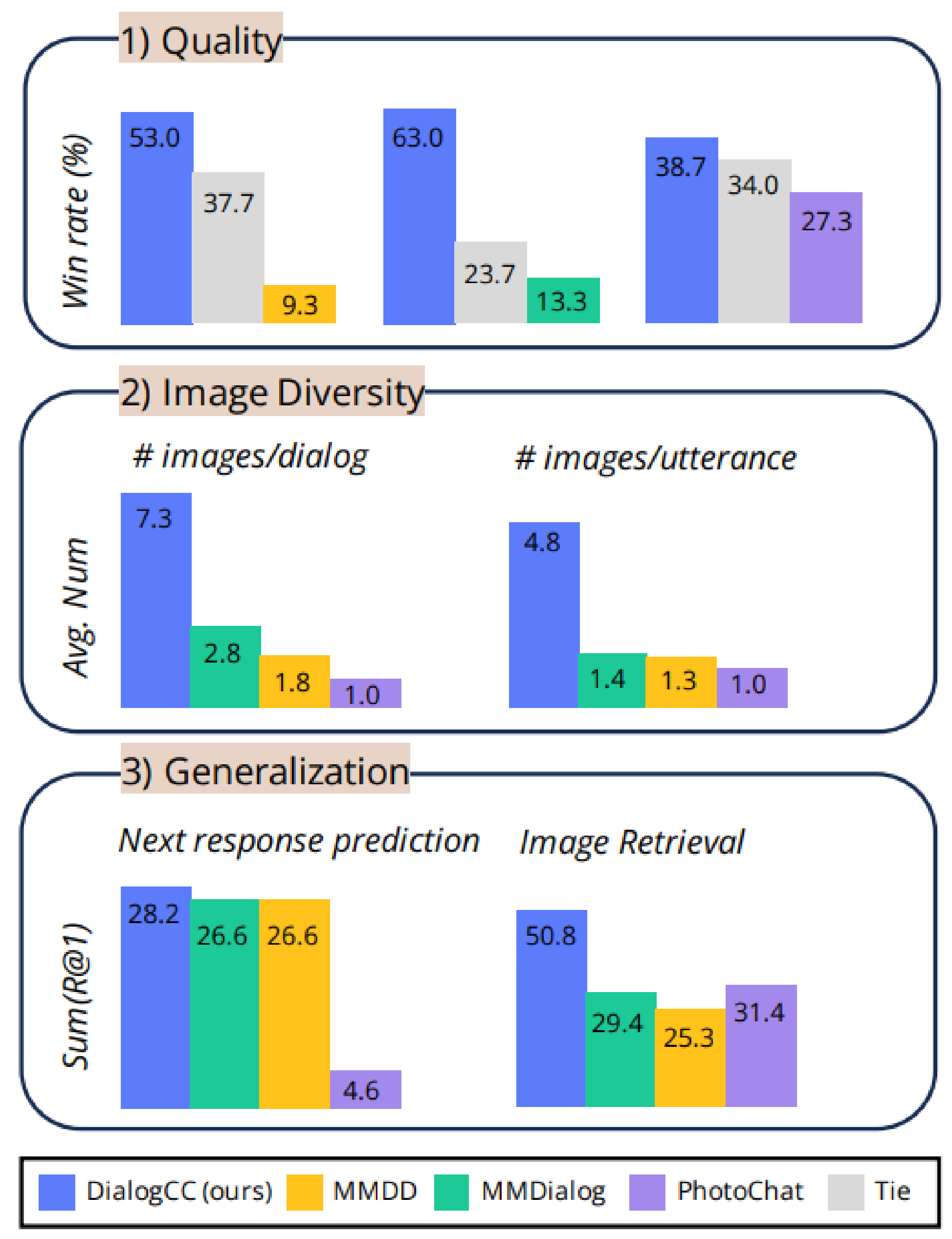
Background: Image-Sharing Behavior

People often share a variety of images during interactions via instant messaging tools



Motivation

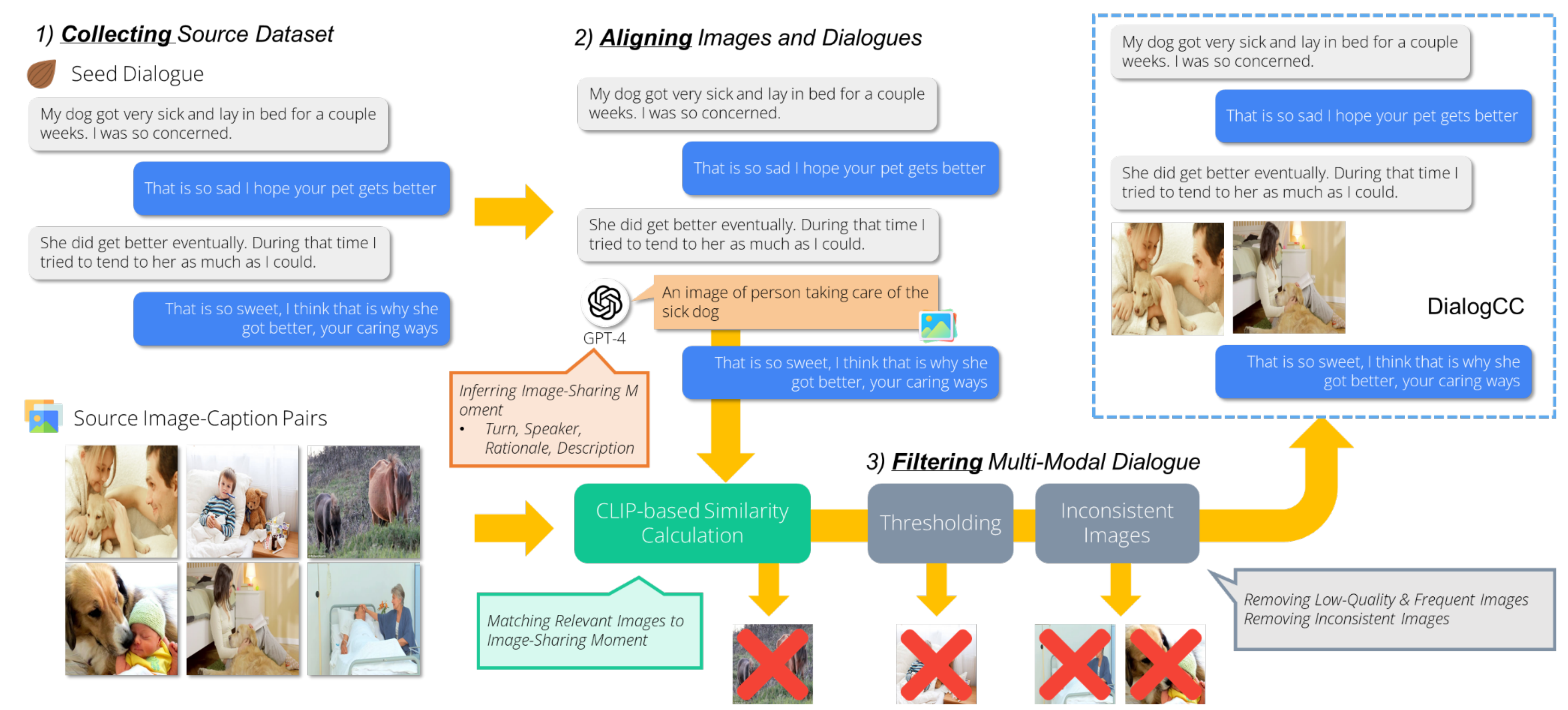
Existing multi-modal dialogue datasets have three significant limitations: Quality, Image Diversity, Generalization



Overview of Proposed Automatic Pipeline

We propose a fully automatic framework for creating a multi-modal dialogue dataset that involves three main steps: collecting, aligning, and filtering.

- Collecting: We collect source datasets (e.g., PersonaChat, CC3M)
- Aligning: We ask GPT-4 to infer all possible image-sharing moments and leverage CLIP to increase the aligned image relevancy
- Filtering: We eliminate inappropriate images based on CLIP similarity



DialogCC: High-Quality and Diverse Multi-Modal Dialogue Dataset

I recently had a long weekend with some old friends. It was fantastic.

It was. We've spent a lot of time together and apart now, so it was good to catchup.

Well I hope you guys continue to stay in touch.

It must have been fun to catch up with them.

Hello the plants in my garden I water them, I move the earth, I try to keep them happy

Especially when you feel something special by nature

Gardening can be peaceful

Nature is a wonderful thing

Analysis of DialogCC

To assess the quality of DialogCC, we conduct the human evaluations based on five criteria:

- Image-Sharing Turn Relevance: 3.68
- Image-Sharing Speaker Relevance: 95.1%
- Image-Sharing Rationale Relevance: 3.41
- Aligned Image Relevance: 3.30
- Image Consistency: 3.57

* Inter-rater agreement (Krippendorff's alpha): 0.39 (fair agreement)

To assess the quality gap between DialogCC and real-world scenarios, we conduct head-to-head human evaluations by comparing DialogCC with existing datasets.

