D. Customized active face extraction: (a)-(e)

(a) Adjust pose/head to focus   (b) Turn to the active speaker   (c) Capture faces by Ameca's camera
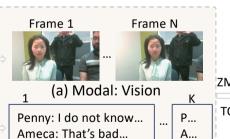
(d) Detect faces

(e) Extract active faces

Frame 1   Frame N

(a) Modal: Vision

1   K

Penny: I do not know…   P…
Ameca: That's bad…   A…

(b) Modal: Language

ZMQ

TCP

Visual encoder   Visual representations

Text encoder   Text representations

Penny: I do not…   P…
Ameca: That's s…   A…

Ameca

(a) Extract active MM features   (b) MM fusion   (c) Emotion response

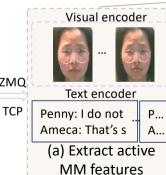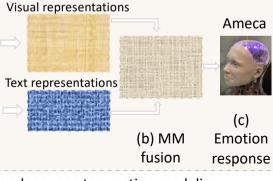**A.** Multimodal, multiparty, affective human-robot conversation

B. On-robot multimodal perception

C. On-edge vision-language to emotion modeling