Visual dialog is a challenging vision-language task, which requires the agent to answer multi-round questions about an image. It typically needs to address two m