We introduce the task of scene-aware dialog. Our goal is to generate a complete and natural response to a question about a scene, given video and audio of th