

It is still challenging to build an AI system that can perform tasks that involve vision and language at human level. So far, researchers have singled out individual