Dense captioning aims at simultaneously localizing semantic regions and describing these regions-of-interest (ROIs) with short phrases or sentences in natural