Sources:

1. Generating Diverse and Accurate Captions Based on Generative Adversarial Network
    a. Uses an AI called a generative adversarial network to generate captions. A GAN uses two modules: a generator and a discriminator. The generator creates the captions using a two-step algorithm to pick features from the video then captions them. The discriminator mixes the machine-made captions with human-made samples to make them sound more natural and easy to read.
2. Listening While Speaking and Visualizing: Improving ASR Through Multimodal Chain
    a. Uses an AI called neural machine translation to perform the captioning. An NMT allows for inputs of images, audio, and video, and has an algorithm to caption each type of content. It also has a speial functionality called sequence-to-sequence deep learning that allows the AI to save the captions in its model so it can reference them in the future to create more accurate captions.
3. Dense-Captioning Events in Videos
    a. Uses an AI called a recurrent neural network to generate captions. This AI has a feature called multi-scale detection of events that allows it to caption the video in one single pass. It does this by identifying key events in the video and then creating a window the content might be related to the event. It then uses an algorithm to guess what the caption is based on the context.
4. A Better Use of Audio-Visual Cues: Dense Video Captioning with Bi-modal Transformer
    a. Uses two AIs called recurrent neural networks to generate captions. Each RNN uses multi-scale detection to caption the video in a single pass by identifying key events in the video then guessing the other parts based on those events. This technique is called bi-modal because one RNN captions the video portion and the other captions the audio portion, then combines them for a richer caption.
5. A Multi-Instance Multi-Label Dual Learning Approach For Video Captioning.
    a. Uses three modules called an encoder, decoder, and reconstructor. The encoder selects the important parts of the video to caption, while the decoder does the actual captioning. The reconstructor then takes the caption and tries to recreate the video. If the reconstructed video does not match the actual video, it fixes the caption until it does and records the fixes for future captions.
6. Rich Visual and Language Representation with Complementary Semantics For Video Captioning
    a. Uses an AI with two layers to create captions. One layer picks features out of the video to caption, while the second layer does the actual captioning. To increase performance, the two layers are set up so that they can communicate. The layers can send information to one another while working concurrently so that they can check each other's errors and create accurate captions quickly.
7. Constrained LSTM and Residual Attention For Image Captioning
    a. Uses a normal two-step process to get features from the video and then caption them. However, the first step where features are picked from the video takes two steps: the first process creates a "skeleton graph" with the big objects in the video, and the second process goes back and picks out any small details that were missed. This improves the descriptiveness of the caption.