Für unimodal training examples masking is necessary, but what about multimodal examples (like text-image)? The data going into the network is different (different modalities), although the concept represented using the training example is the same