Modality-Invariant Targets

Contrastive Target Loss

- had severe problems with misalignment of text tokens and image patches
- mse used as loss -> required embeddings to be exactly the same
- cls token still contains image-specific information (show loss), so mse might not be the best choice