Modality-Invariant Targets

Throughout the previous experiments, we have seen that the misalignment between image patches and text tokens leads to problems when regressing the image features of the teacher.

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

Memory Bank