- unimodal kd (data2vec)
- unaligned multimodal kd (data2vec) -> can only be unaligned of mome is used for all layers
  - mome with shared attention
  - all shared -> this would mean alignment -> problem is layers would regress two teachers at once

  -> might lead to a convergence somewhere in the middle, where nothing useful is learned, or training would never converge
- aligned multimodal (See, Hear, and Read: Deep Aligned Representations + data2vec)
  - no e.g. nlvr2 possible (?)
- referencing aligned multimodal kd (vlmo)
  - only vision language possible