Validation:

* Normal validation sets of datasets used for training

Zero shot:

* CIFAR-10/100
* SpeechCommands
* IMDB

Testing (-> Fine-tuning):

* CTC
* GLUE
* CIFAR100 zero shot (KNN) (imagenet does not really make sense, as is has been trained on this, so creating representations out of training data would be “cheating”)
* (Imagenet validation top-1 -> Already in “Validation”)
* CIFAR10/100
* VQAv2, NLVR2, COCO Captions
* Retrieval (Image->Text, Text->Image, Text->Audio, Audio-> Text, potentiell interessant: Image -> Audio, Audio -> Image: Nicht explizit gelernt, wie idee in: <https://arxiv.org/pdf/1706.00932.pdf>)

Bonus:

* SETI (zero-shot and/or fine-tuning)

Validation zero-shot KNN for: Cifar10/100, SpeechCommands, IMDB

* Performance mit diesen auch nur durch anwendung von modality specific encoders möglich (auch mal testen), aber bei der verwendung dieser bei der validation geht es mehr darum zu schauen, wie sich das modell verhält, ob es wichtige features produziert und ob es sich während des trainings bessert.