We perceive the world through modalities:

- Each provides a distinct view of the same physical reality
- Combined, they allow us to better understand our world
- Enables cross-modal learning as a form of (self) supervision
- Helps with developing more “grounded” models

Goal: Training an any-to-any vision foundation model

- Scaled in terms of number and format of modalities and tasks, model & dataset size

Overview of modalities

1. Pseudo labeling
2. Modality-specific tokenization
3. Masked pre-training

Tokenization

- Unifies representation space for scalable training
- Different modalities require different strategies

Out-of-the-box capabilities

- Fine-grained & controllable multimodal generation & retrieval
- Strong out-of-the-box (zero-shot) performance
- Transfer well to downstream tasks (unimodal, multimodal)
- Maintains the performance of 4M-7 while solving 3x more tasks