4M: Massively Multimodal Masked Modeling

David Mizrahi¹,²* Roman Bachmann¹* Oğuzhan Fatih Kar¹ Teresa Yeo¹ Mingfei Gao² Afshin Dehghan² Amir Zamir¹

¹EPFL ²Apple

4m.epfl.ch

Motivation

We perceive the world through various modalities:
- Each provides a distinct view of the same physical reality
- Combined, they allow us to better understand our world

Goal: A training framework for multimodal foundation models
- Scalable in terms of number of modalities & tasks, model size, and dataset size
- Anything in, anything out (any-to-any)

Approach

4M: A framework for training versatile multimodal models

A generalist vision model that can...
- perform a diverse set of vision tasks out of the box
- be easily fine-tuned into specialist variants
- transfer well to unseen tasks and modalities

A multimodal generative model that can...
- generate any modalities conditioned on any other(s)
- enabling precise user control through multimodal editing chains

Anything in, anything out

- 4M leads to models capable of generating any modality conditioned on any other(s)
- Chained generation leads to self-consistent predictions

Training framework:
1. Pseudo labeling: Start from image-text pairs, then use specialized networks to generate an aligned multimodal dataset
2. Tokenization: Unify the representation space by mapping all modalities into sets or sequences of discrete tokens = cross-entropy loss for everything
3. Multimodal masked pre-training: Train a single Transformer to predict a randomly selected subset of tokens, sampled from all modalities, from another random subset of tokens

At inference: Iteratively predict & sample tokens

Generation scheme depends on the modality (MaskGIT for 2D/images, autoregressive for sequences)

Analysis & comparisons

Impact of input modalities & target tasks
Impact of multimodal masking strategy
Scaling trends

Token-to-token transfer benchmark: Ablation of key design parameters by transferring to 25 different single-modal & multimodal downstream tasks

Key findings:
- More diverse sets of 4M pre-training tasks improve transfer performance
- Masking strategy matters: Multimodal masking over the inputs & targets improves efficiency and performance
- Promising scaling trends in terms of dataset size, training length, and model size

RGB → X transfers:
- 4M models also support pixel inputs (not just tokens)
- Can be used as ViT backbones & significantly outperform MAE and MultiMAE on standard vision tasks

Summary

4M: A framework for training any-to-any multimodal foundation models
- Relies on tokenization & masking to scale to many diverse modalities
- Models trained using 4M can:
  - Perform a wide range of vision tasks out of the box
  - Transfer well to unseen tasks and modalities
  - Function as flexible and steerable multimodal generative models