4M: Massively Multimodal Masked Modeling

David Mizrahi[1,2*] Roman Bachmann[1*] Oğuzhan Fatih Kar[1] Teresa Yeo[1] Mingfei Gao[2] Afshin Dehghan[2] Amir Zamir[1]

[1]EPFL [2]Apple

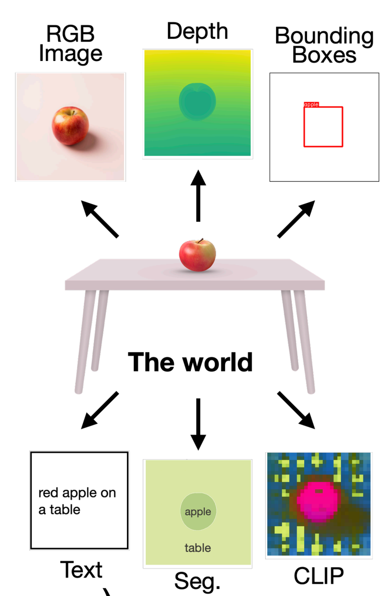
4m.epfl.ch

NEURAL INFORMATION PROCESSING SYSTEMS 2023

Motivation

We perceive the world through various modalities:

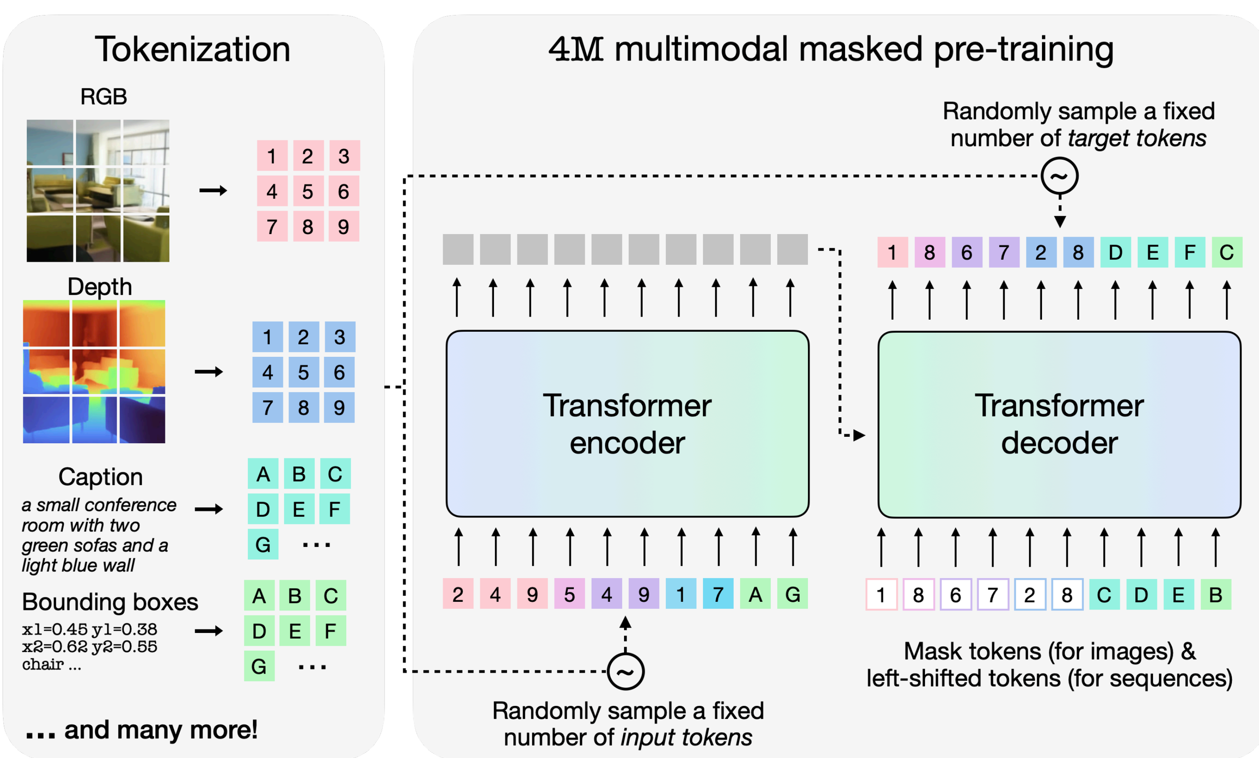
- Each provides a distinct view of the same physical reality
- Combined, they allow us to better understand our world



Goal: A training framework for multimodal foundation models

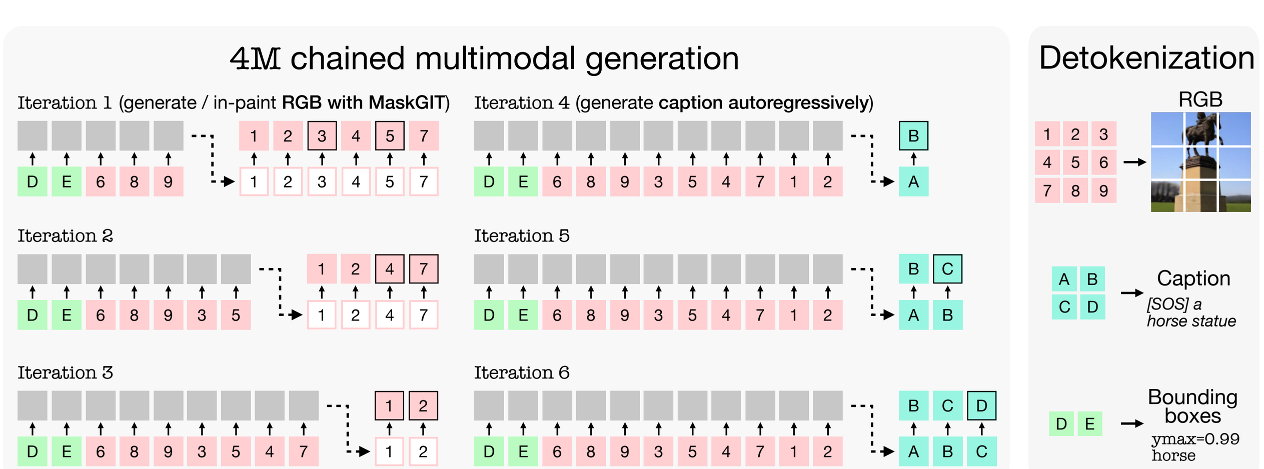
- **Scalable** in terms of **number of modalities & tasks, model size, and dataset size**
- Anything in, anything out **(any-to-any)**

Approach



Training framework:

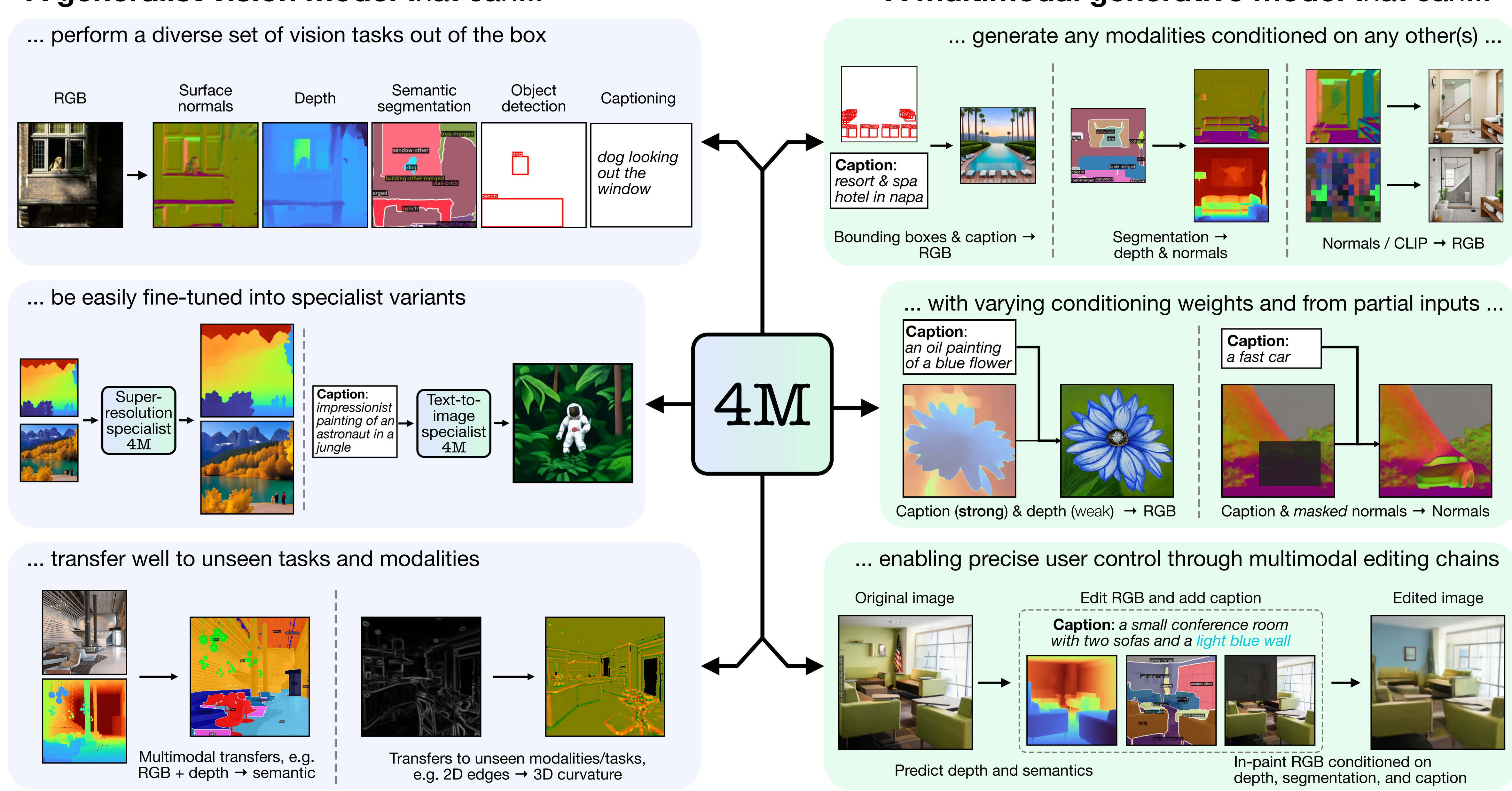
1. **Pseudo labeling:** Start from image-text pairs, then use specialized networks to generate an *aligned multimodal dataset*
2. **Tokenization:** Unify the representation space by mapping all modalities into sets or sequences of *discrete tokens* = *cross-entropy loss for everything*
3. **Multimodal 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

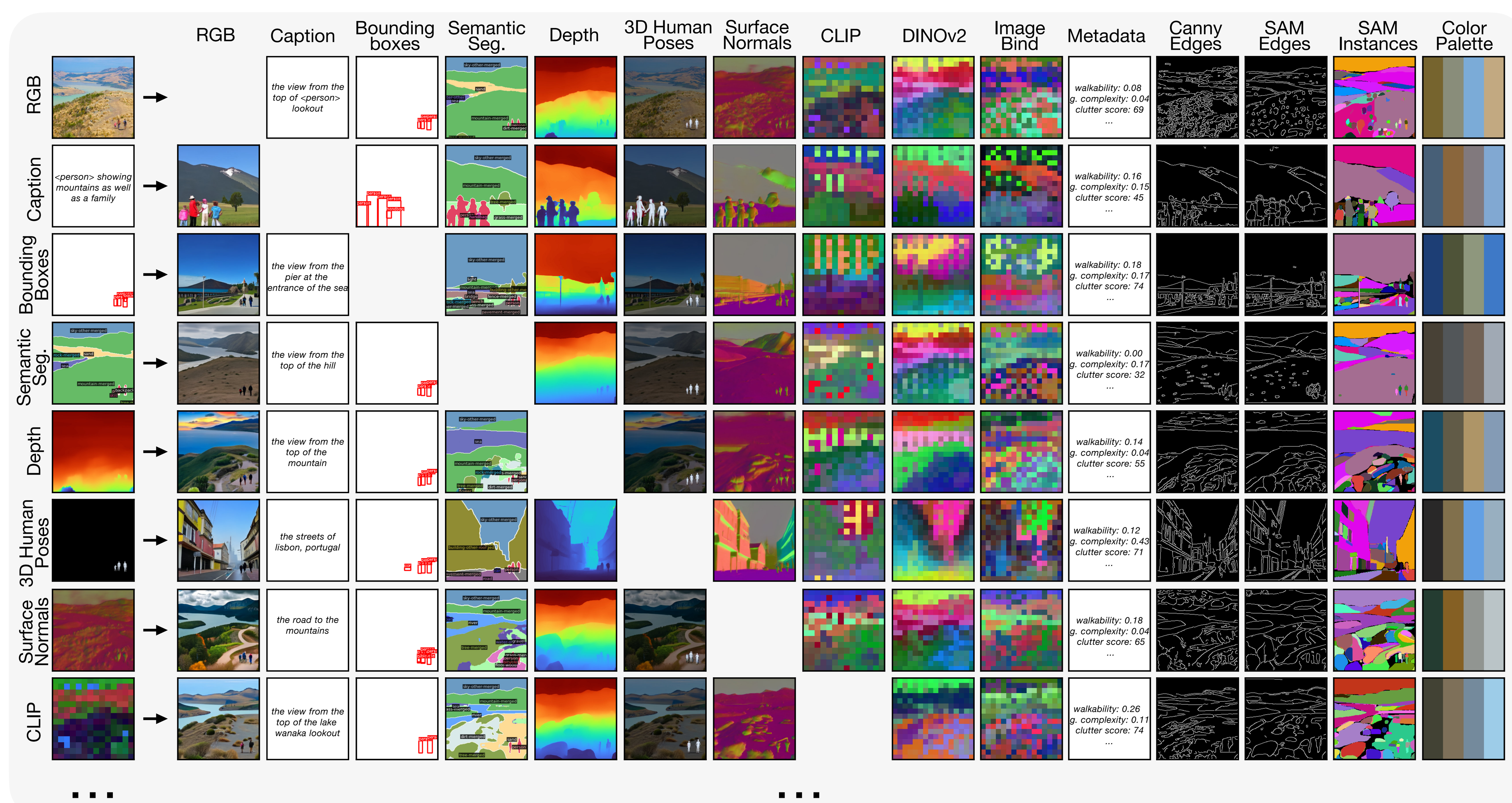
Generation scheme depends on the modality (MaskGIT for 2D/images, autoregressive for sequences)

4M: A framework for training versatile multimodal models

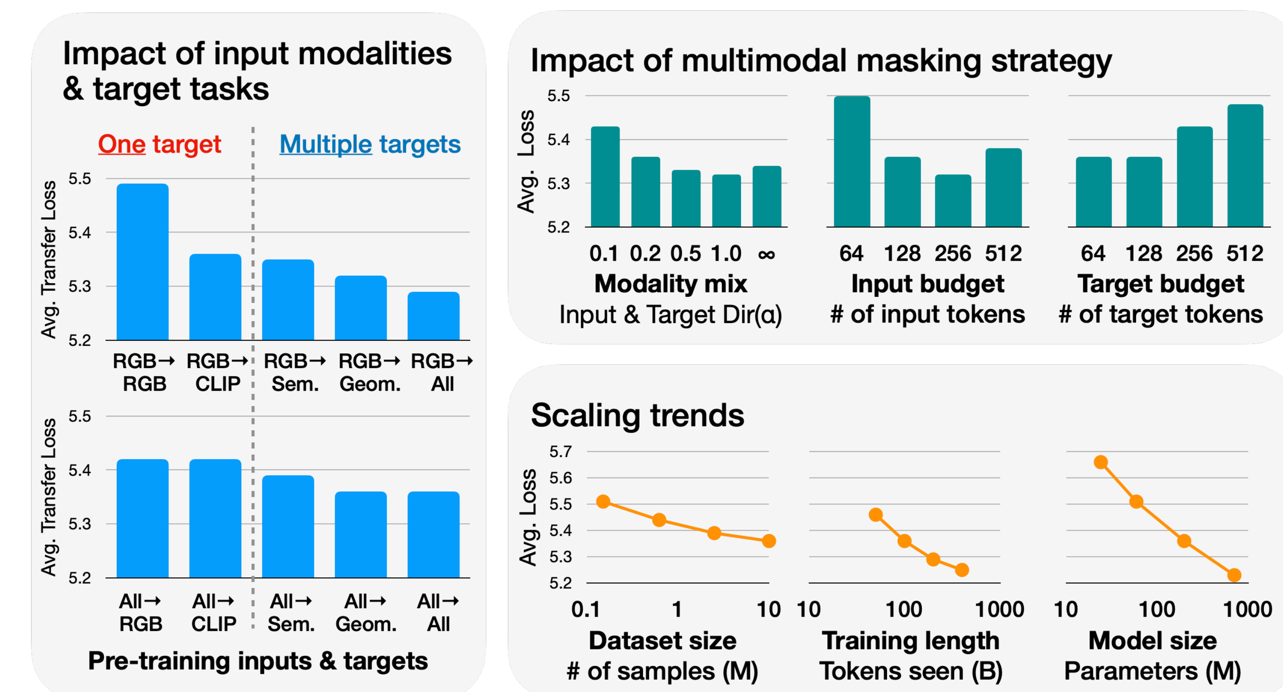


Anything in, anything out

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



Analysis & comparisons



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

Method	Pre-training data	IN-1K Classif.	COCO Det. / Inst. Seg.		ADE20K Sem. Seg.	NYUv2 Depth
		T1 Acc. ↑	AP^{box} ↑	AP^{mask} ↑	mIoU ↑	δ1 ↑
MAE B	IN-1K	84.2	48.3	41.6	46.1	89.1
DeiT III B	IN-21K	**85.4**	46.1	38.5	49.0	87.4
MultiMAE B	IN-1K	84.0	44.1	37.8	46.2	89.0
4M-B	CC12M	84.5	**49.7**	**42.7**	**50.1**	**92.0**
MAE L	IN-1K	86.8	52.8	45.3	51.8	93.6
DEiT III L	IN-21K	**87.0**	48.7	41.1	52.0	89.6
4M-L	CC12M	86.6	**53.7**	**46.4**	**53.4**	**94.4**

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**