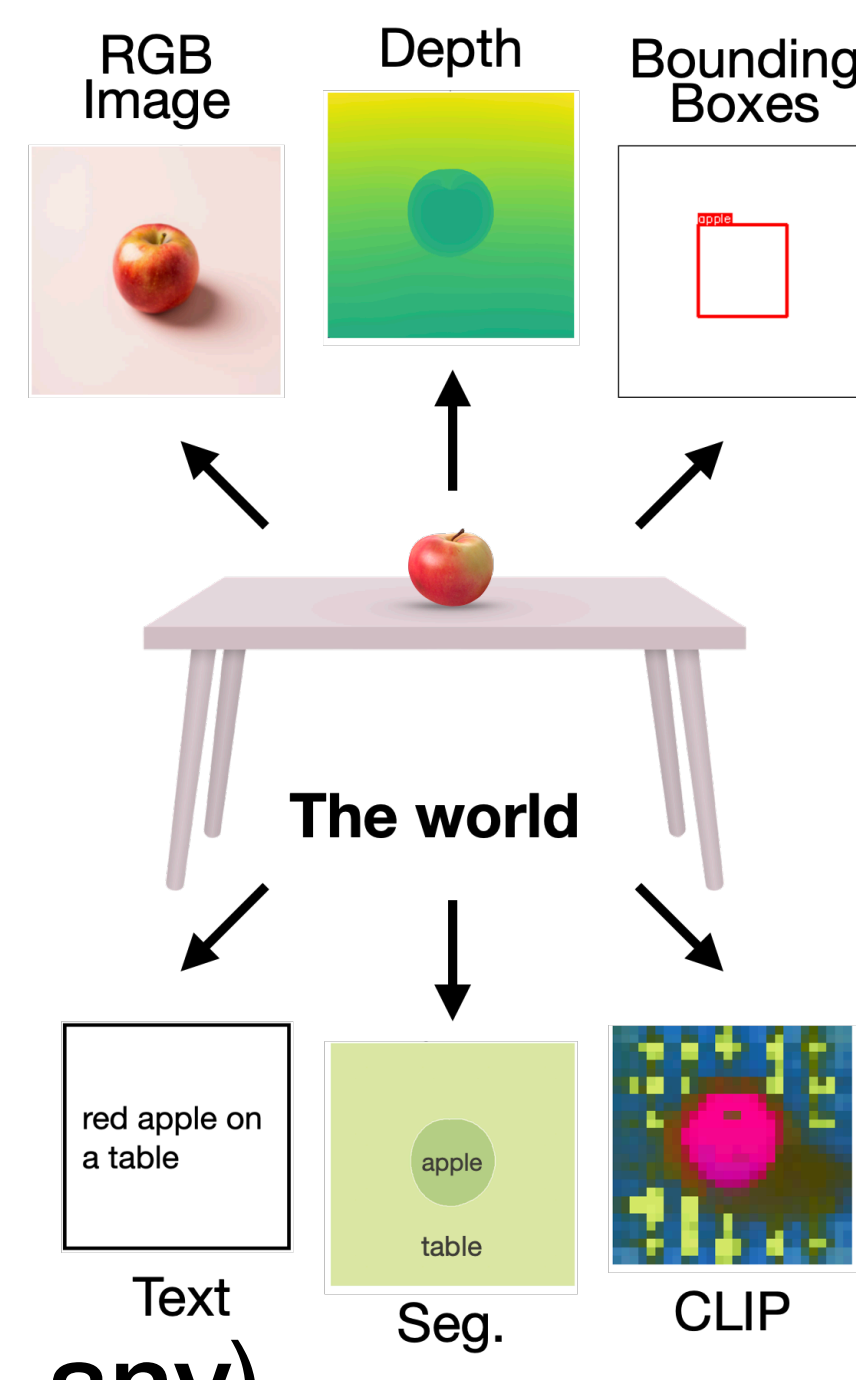


## Motivation

We perceive the world through various modalities:

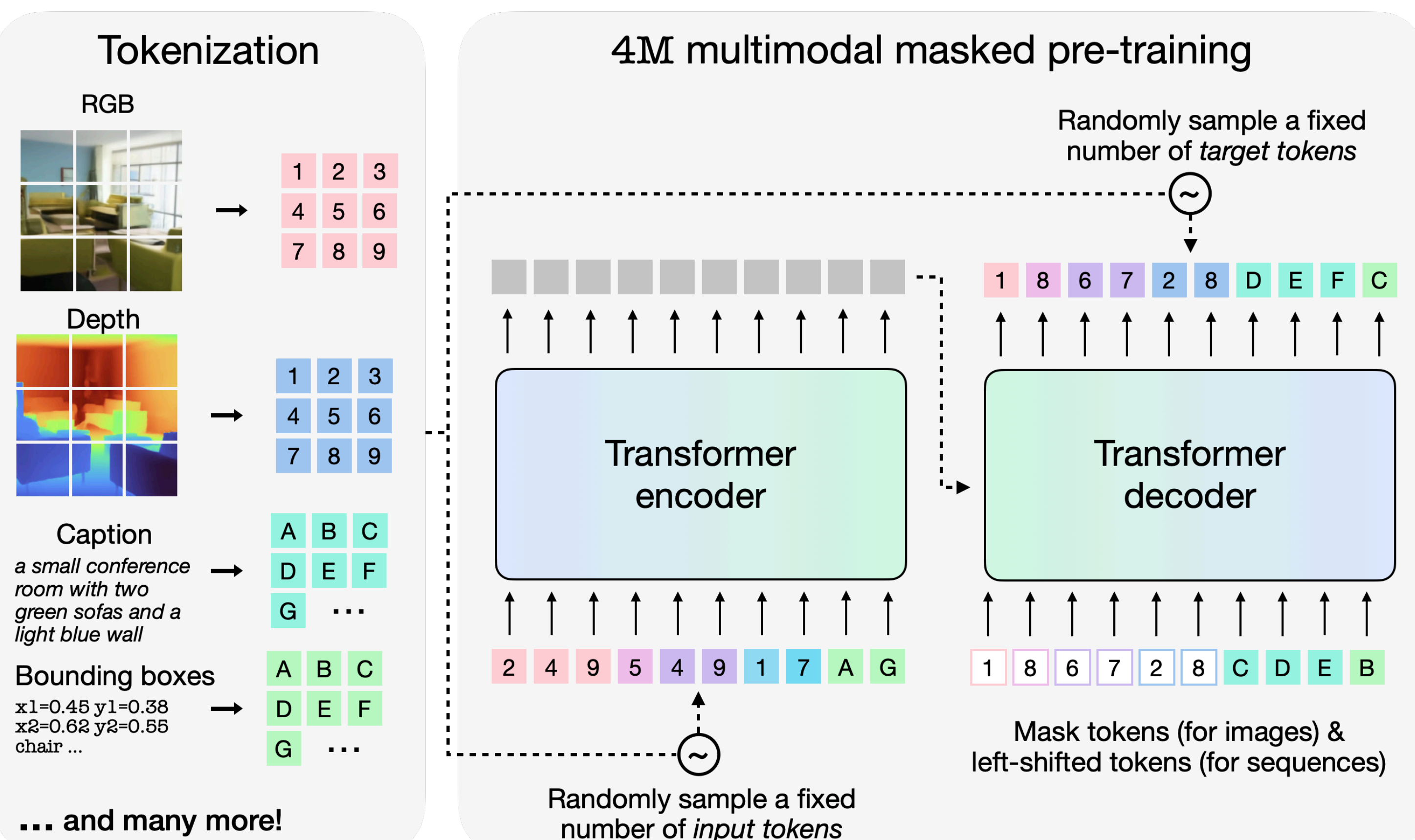
- Each provides a distinct view of the same physical reality
- Combined, they allow us to better understand our world



**Goal:** A training framework for multimodal foundation models

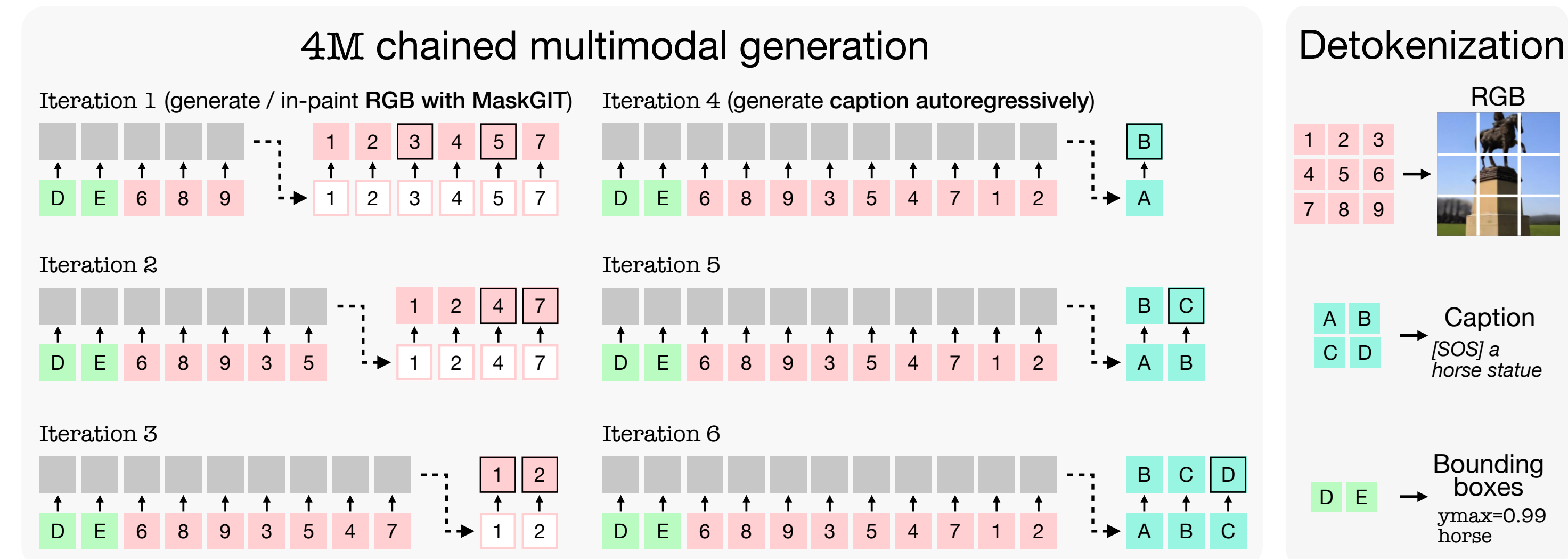
- Scalable in terms of number of modalities & tasks, model size, and dataset size
- Anything in, anything out (any-to-any)

## 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



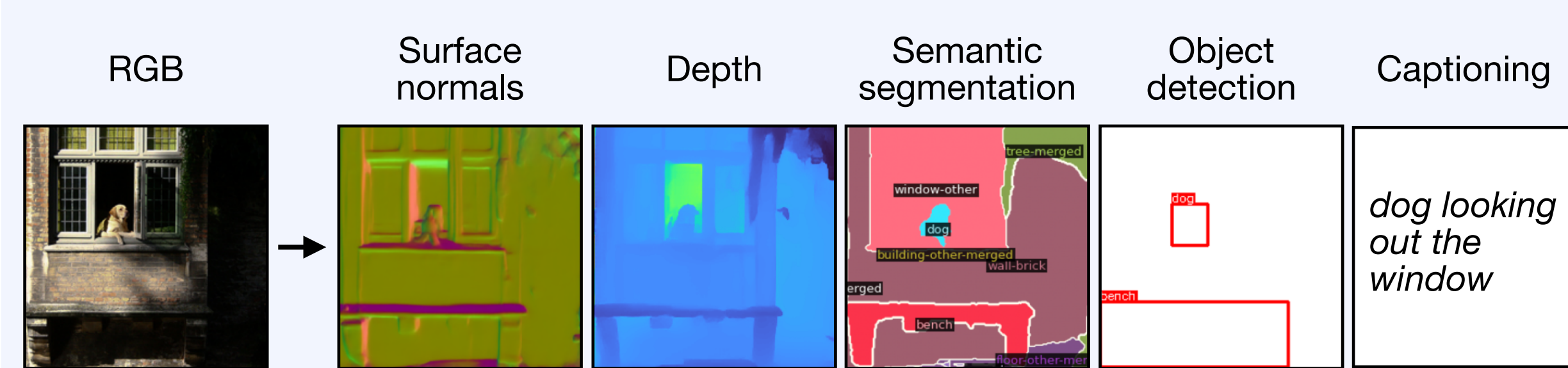
**At inference:** Iteratively predict & sample tokens

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

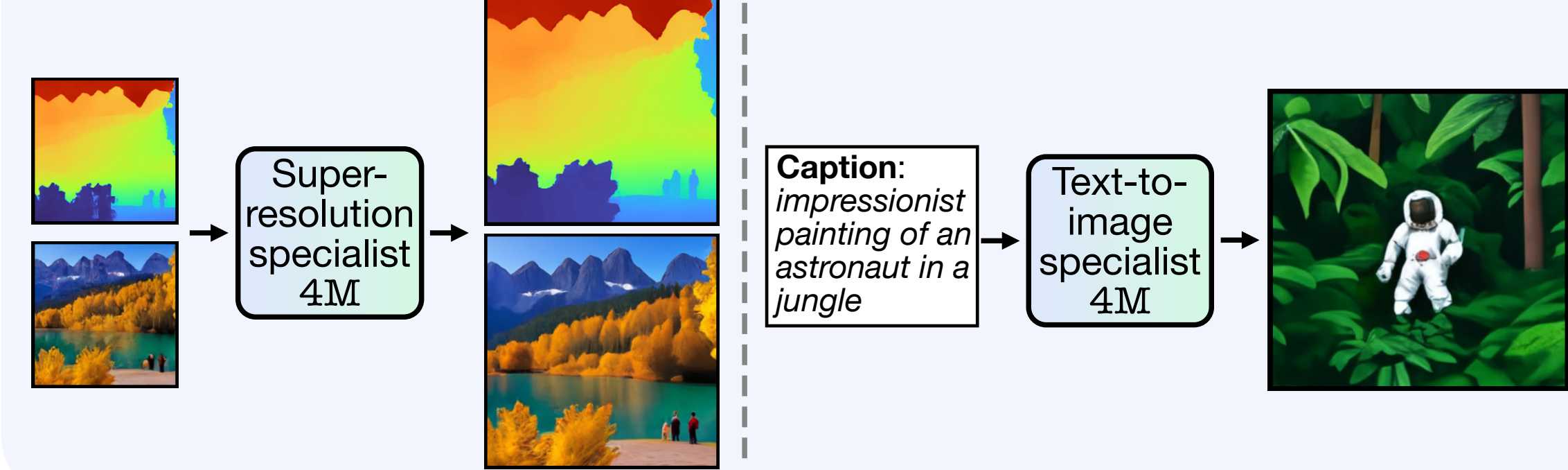
## 4M: A framework for training versatile multimodal models

**A generalist vision model that can...**

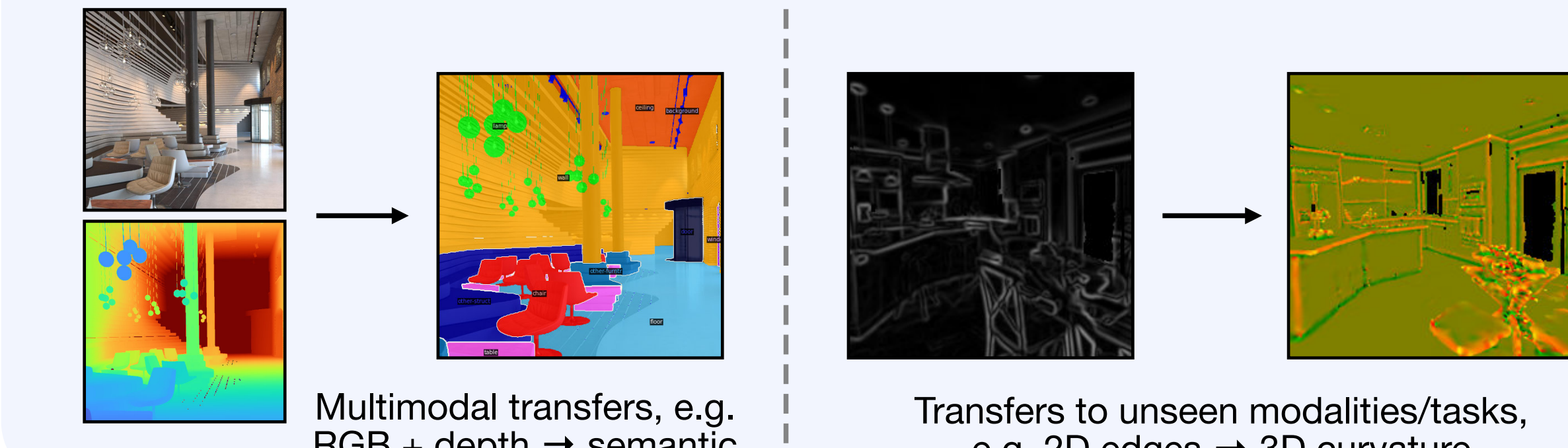
... perform a diverse set of vision tasks out of the box



... be easily fine-tuned into specialist variants

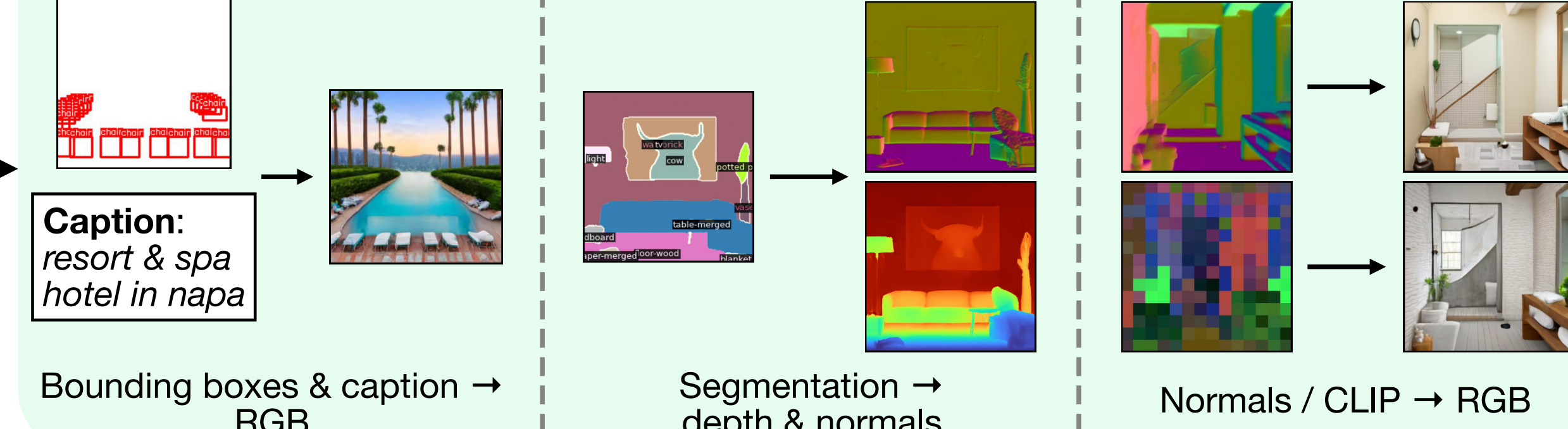


... transfer well to unseen tasks and modalities

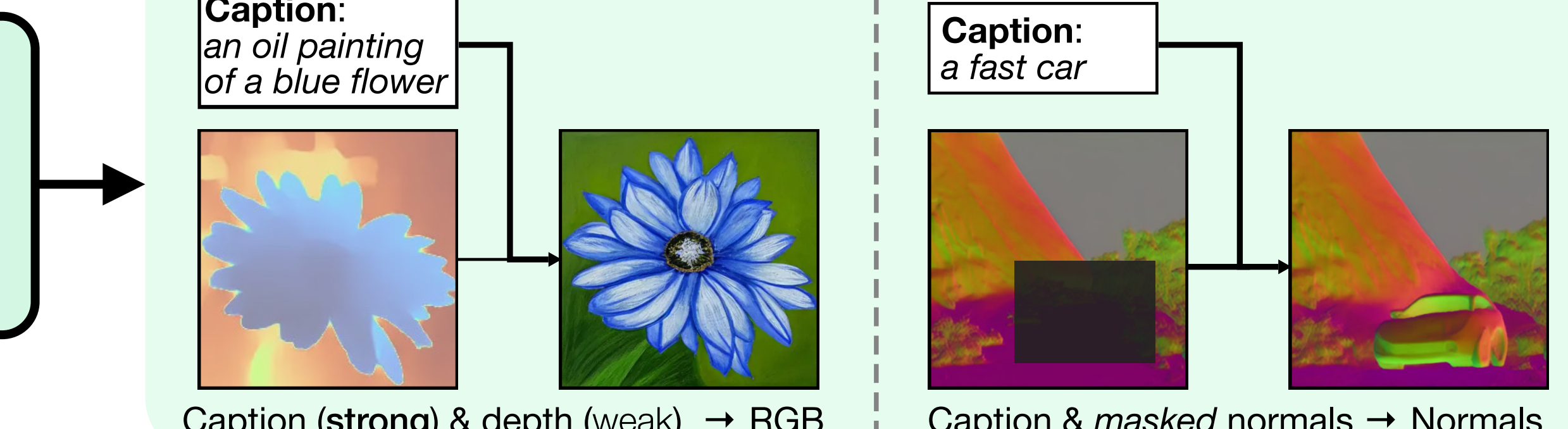


**A multimodal generative model that can...**

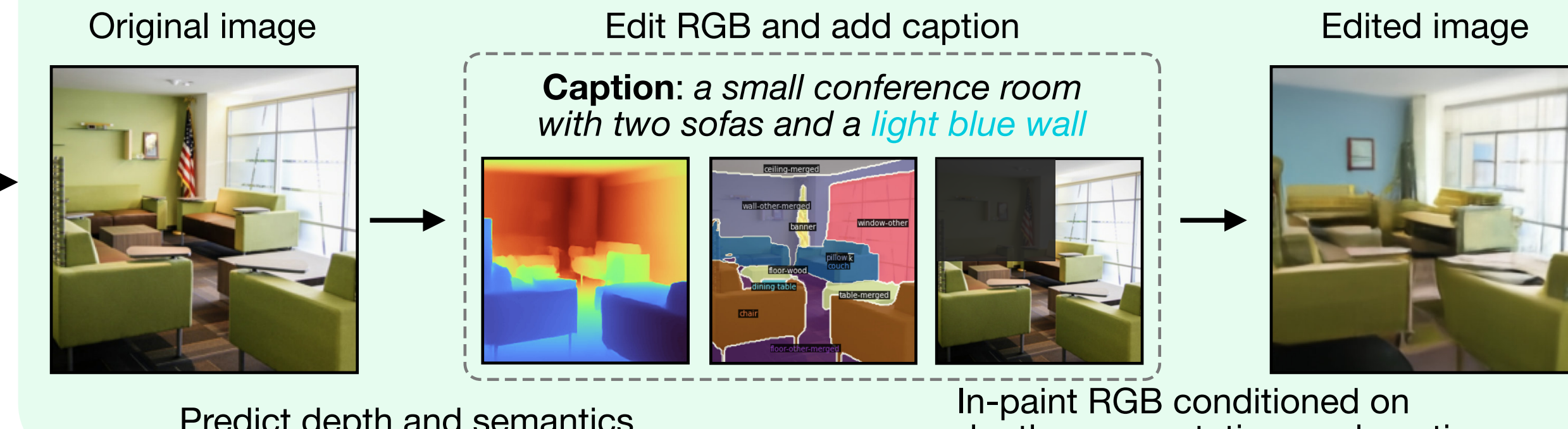
... generate any modalities conditioned on any other(s) ...



... with varying conditioning weights and from partial inputs ...



... enabling precise user control through multimodal editing chains

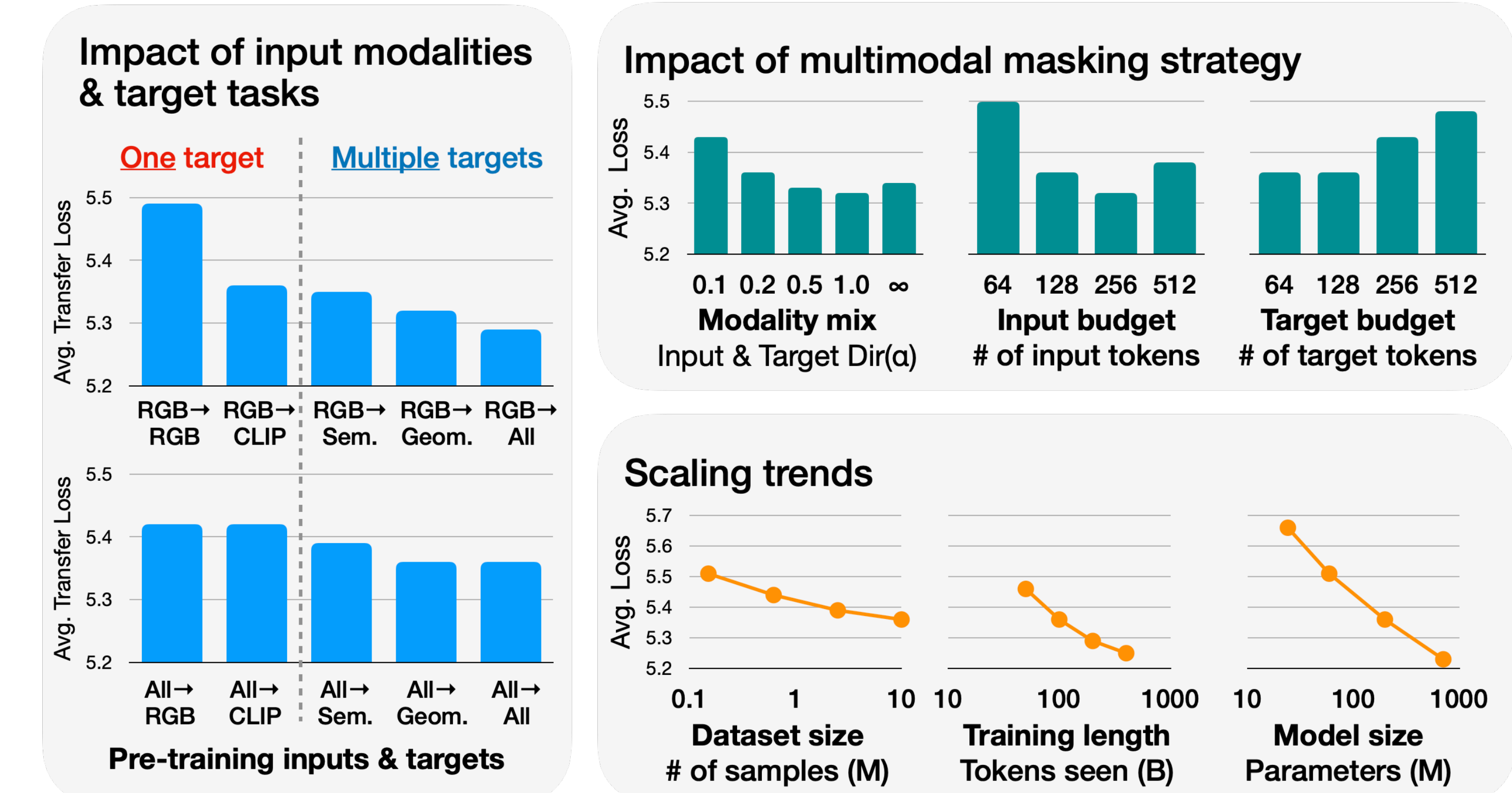


## Anything in, anything out

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

	RGB	Caption	Bounding boxes	Semantic Seg.	Depth	3D Human Poses	Surface Normals	CLIP	DINOv2	Image Bind	Metadata	Canny Edges	SAM Edges	SAM Instances	Color Palette
RGB		the view from the top of open-air lookout									walkability: 0.09 g. complexity: 0.04 clutter score: 69				
Caption		<person> showing mountains as well as a family									walkability: 0.16 g. complexity: 0.15 clutter score: 45				
Bounding Boxes		the view from the pier at the entrance of the sea									walkability: 0.18 g. complexity: 0.17 clutter score: 74				
Semantic Seg.		the view from the top of the hill									walkability: 0.09 g. complexity: 0.17 clutter score: 35				
Depth		the view from the top of the mountain									walkability: 0.14 g. complexity: 0.04 clutter score: 55				
3D Human Poses		the streets of Lisbon, Portugal									walkability: 0.12 g. complexity: 0.48 clutter score: 71				
Surface Normals		the road to the mountains									walkability: 0.18 g. complexity: 0.04 clutter score: 65				
CLIP		the view from the top of the lake wanaika lookout									walkability: 0.28 g. complexity: 0.11 clutter score: 74				

## 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		ADE20K	NYUv2
			T1 Acc. ↑	AP <sub>box</sub> ↑	AP <sub>mask</sub> ↑	mIoU ↑
MAE B	IN-1K	84.2	48.3	41.6	46.1	89.1
DeiT III B	IN-21K	<b>85.4</b>	46.1	38.5	49.0	87.4
MultiMAE B	IN-1K	84.0	44.1	37.8	46.2	89.0
<b>4M-B</b>	CC12M	84.5	<b>49.7</b>	<b>42.7</b>	<b>50.1</b>	<b>92.0</b>
MAE L	IN-1K	86.8	52.8	45.3	51.8	93.6
DeiT III L	IN-21K	<b>87.0</b>	48.7	41.1	52.0	89.6
<b>4M-L</b>	CC12M	86.6	<b>53.7</b>	<b>46.4</b>	<b>53.4</b>	<b>94.4</b>

## 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