4M-21: AN ANY-TO ANY VISION MODEL FOR TENS OF TASKS AND MODALITIES

Roman Bachmann*[1] Oğuzhan Fatih Kar[1]* David Mizrahi[1,2]* Ali Garjani[1]
Mingfei Gao[2] David Griffiths[2] Jiaming Hu[2] Afshin Dehghan[2] Amir Zamir[1]

[1]EPFL [2]Apple 4m.epfl.ch

Motivation

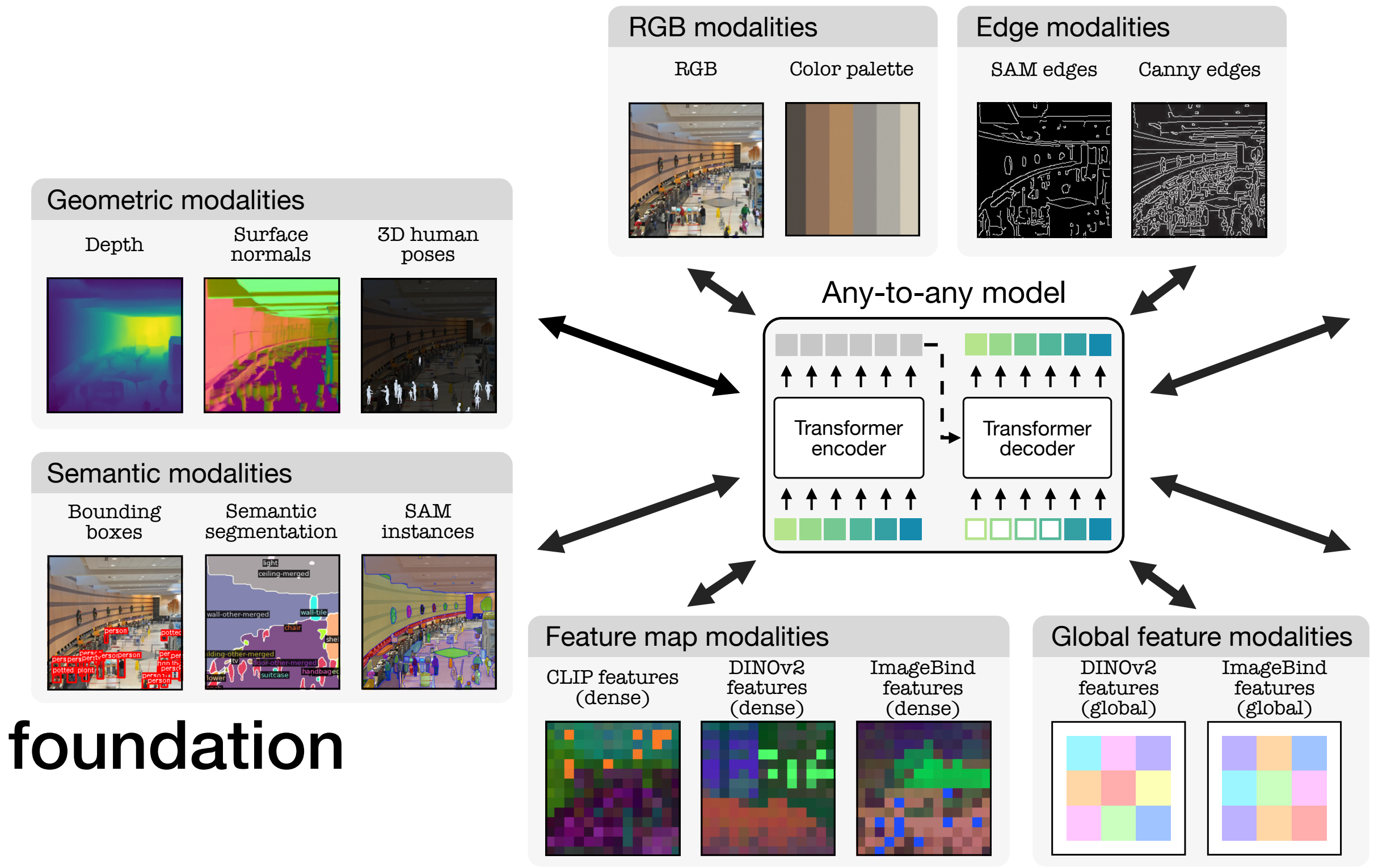
We perceive the world through modalities:

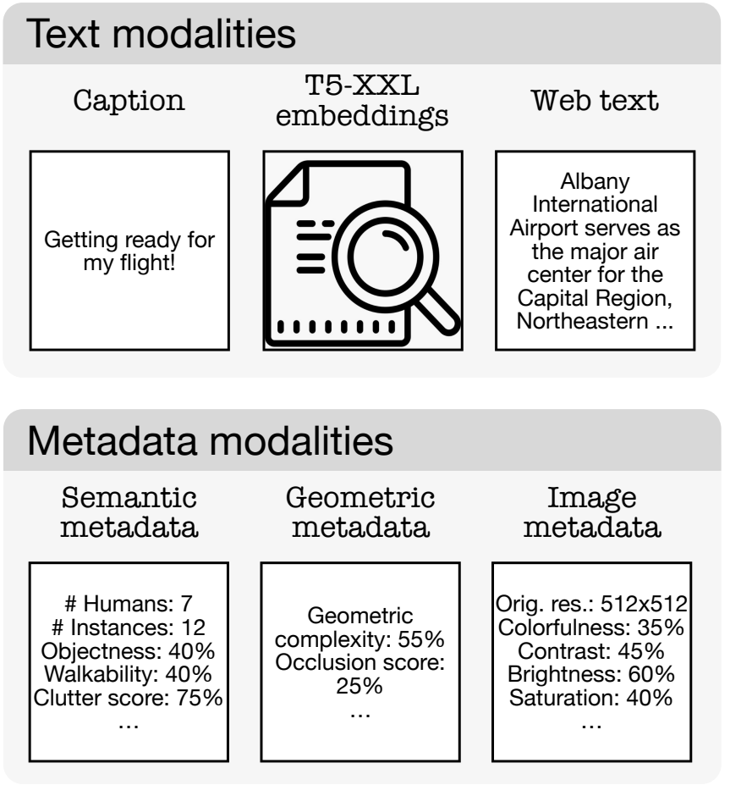
- Each provides a distinct view of the same physical reality
- Combined, they allow us to better understand our world
- Enables cross-modal learning as a form of (self) supervision
- Helps with developing more “grounded” models

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

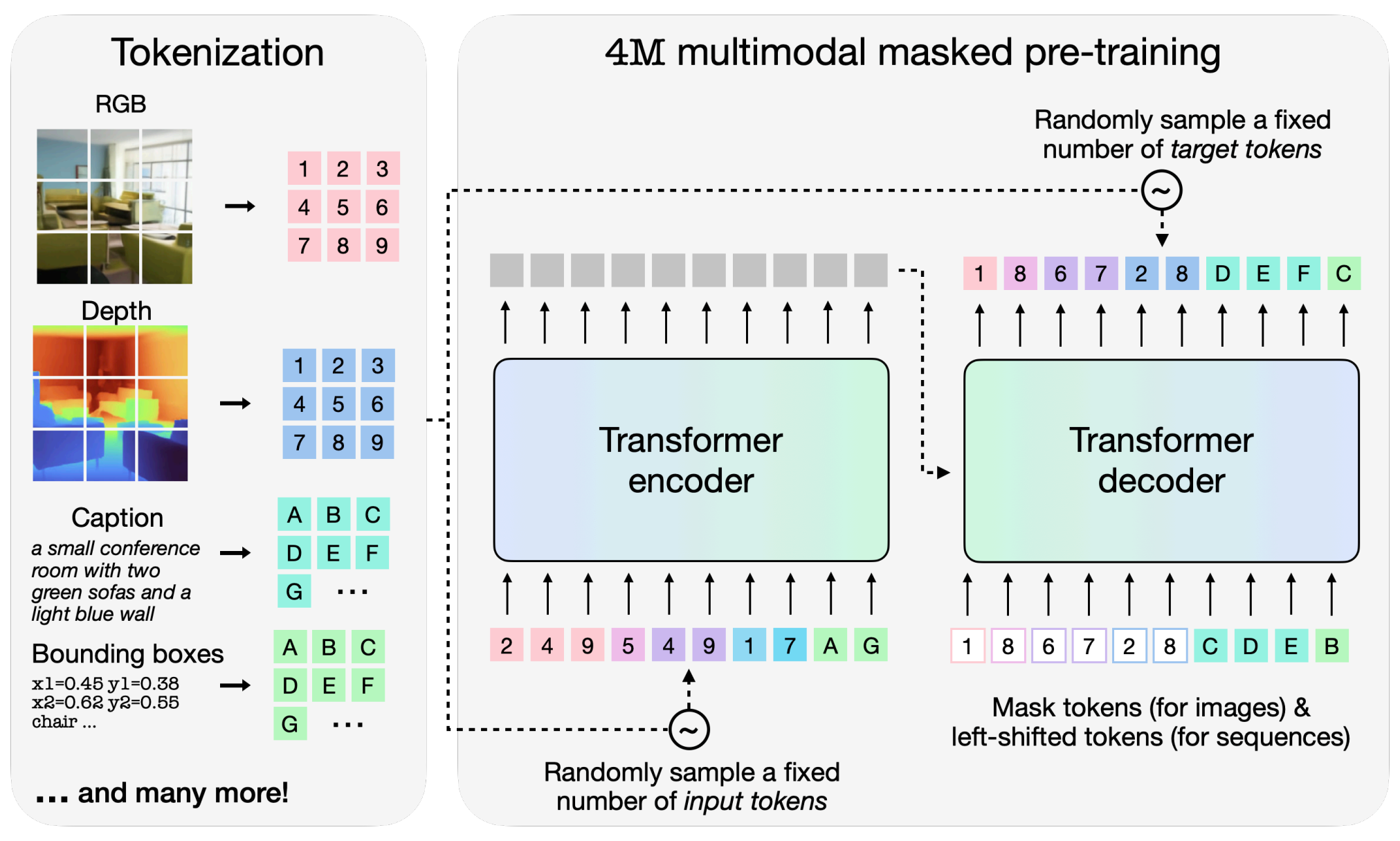
- **Scaled** in terms of number and format of modalities and tasks, model & dataset size

Overview of modalities



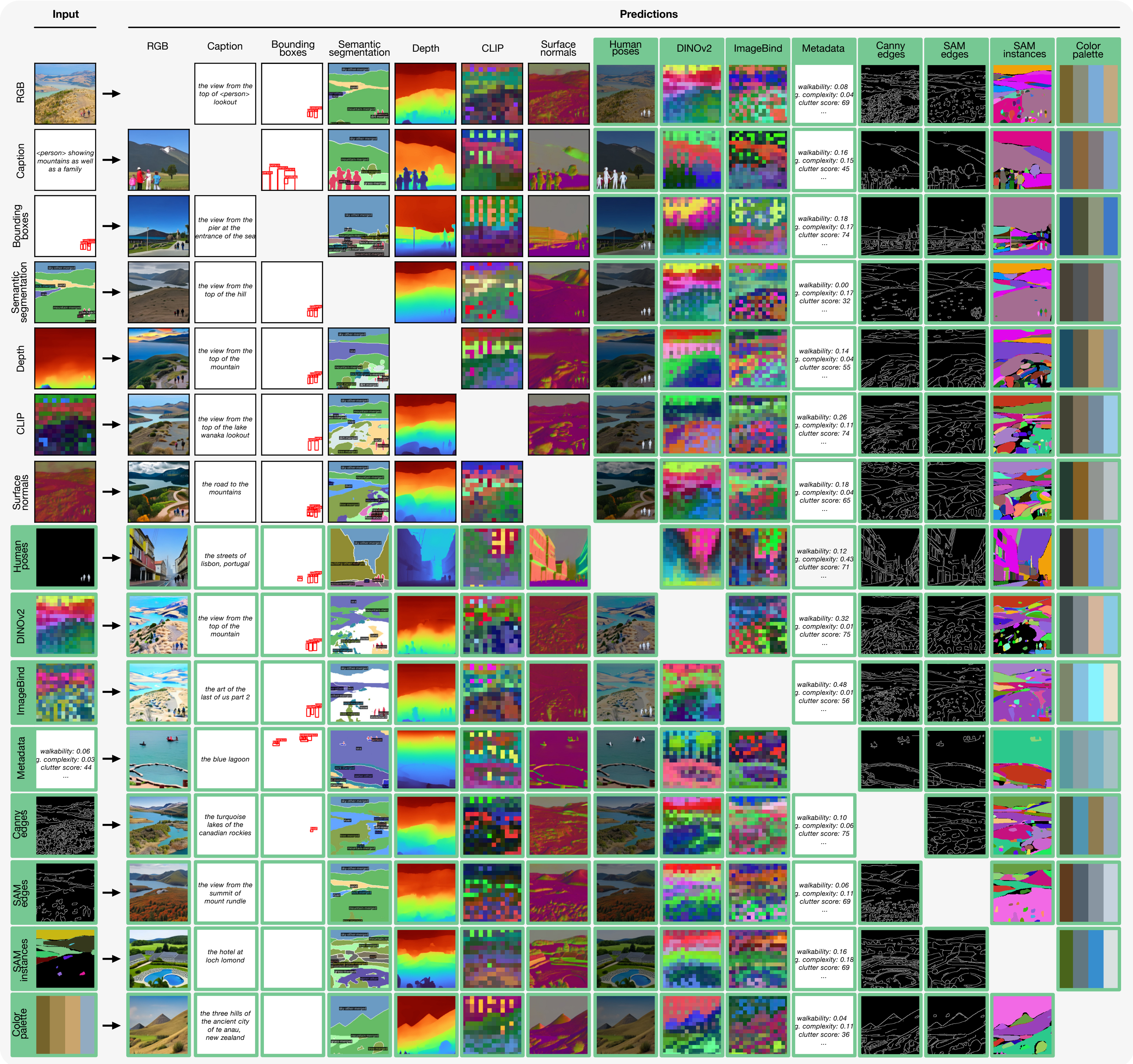


Model

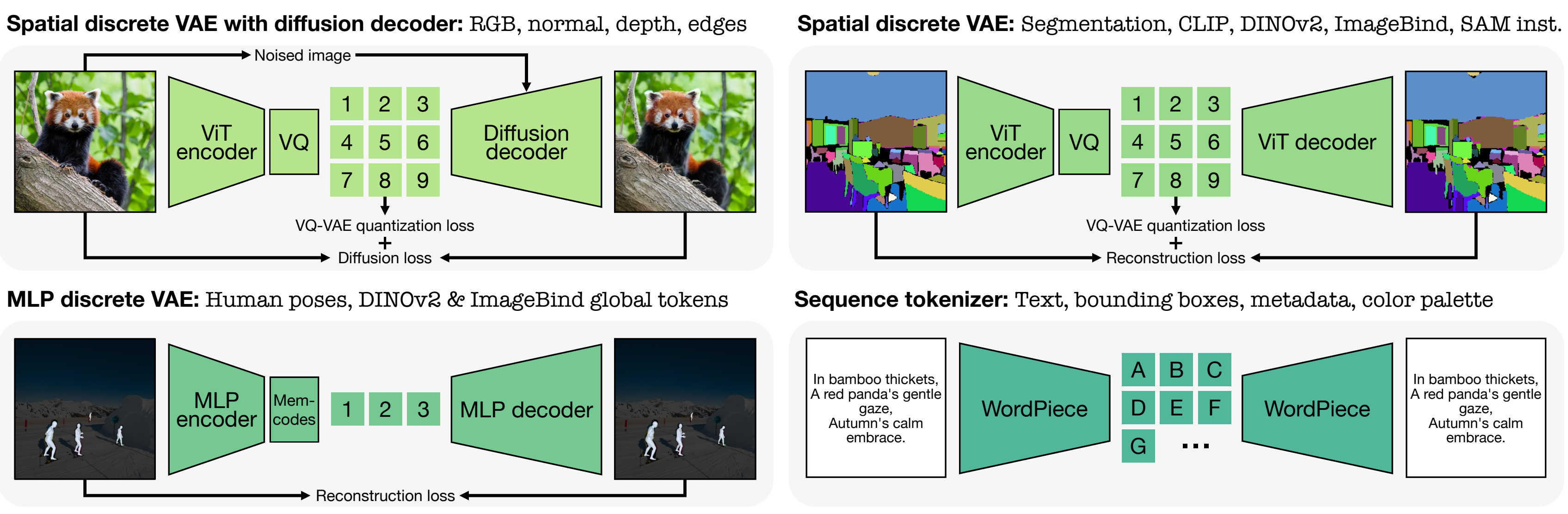

Tokenization
RGB
Depth
Caption
a small conference room with two green sofas and a light blue wall
Bounding boxes
x1=0.45 y1=0.38
x2=0.62 y2=0.55
chair ...
... and many more!
4M multimodal masked pre-training
Randomly sample a fixed number of target tokens
Transformer encoder
Transformer decoder
Randomly sample a fixed number of input tokens
Mask tokens (for images) & left-shifted tokens (for sequences)

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

Anything in, anything out



Tokenization


Spatial discrete VAE with diffusion decoder: RGB, normal, depth, edges

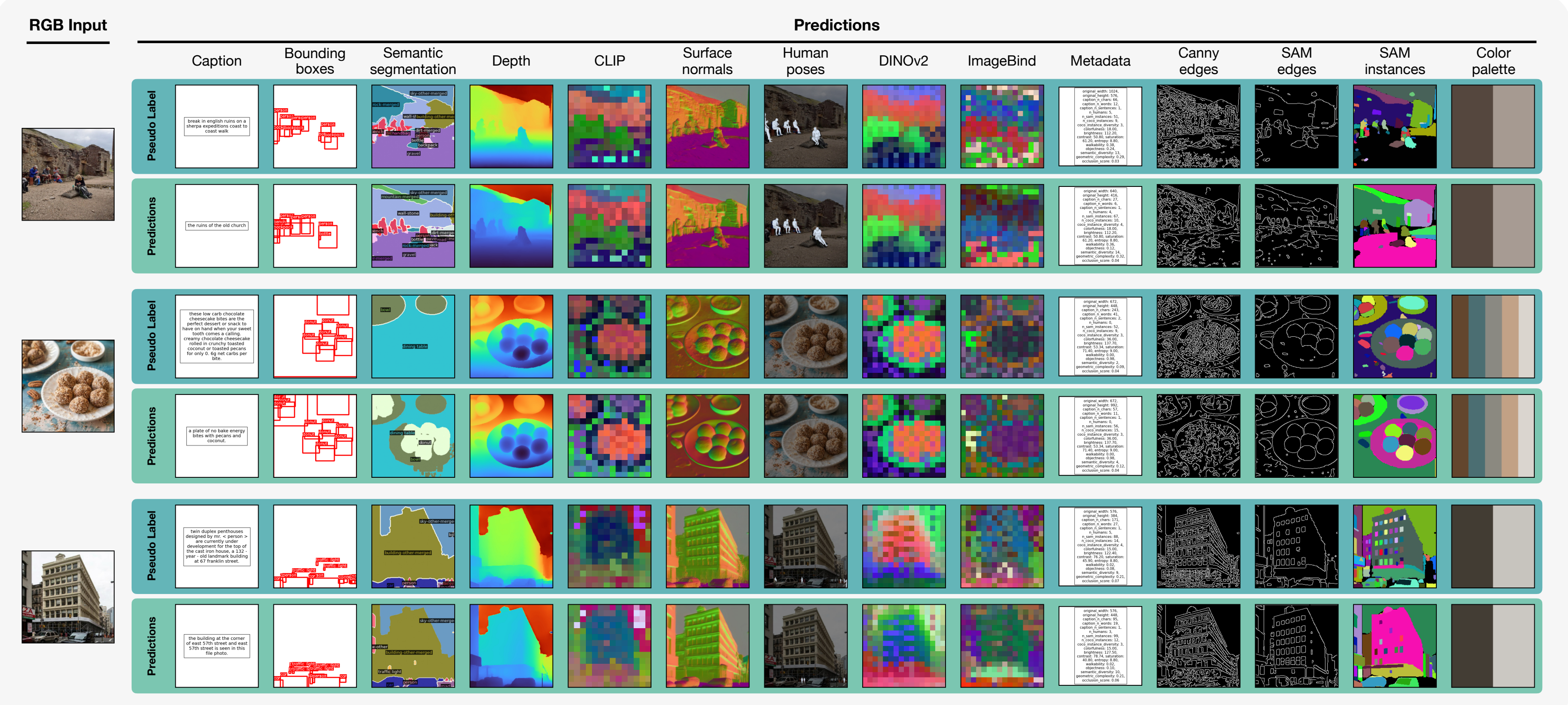
Spatial discrete VAE: Segmentation, CLIP, DINOv2, ImageBind, SAM inst.

MLP discrete VAE: Human poses, DINOv2 & ImageBind global tokens

Sequence tokenizer: Text, bounding boxes, metadata, color palette

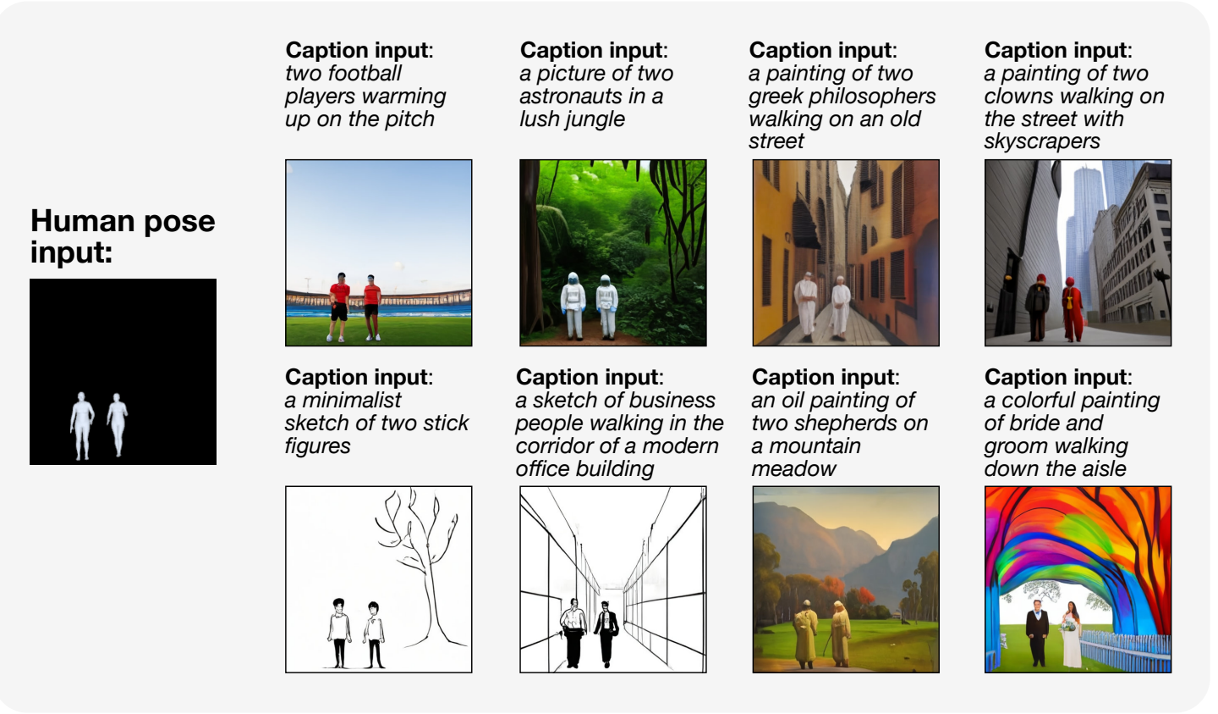
- Unifies representation space for scalable training
- Different modalities require different strategies

Out-of-the-box capabilities

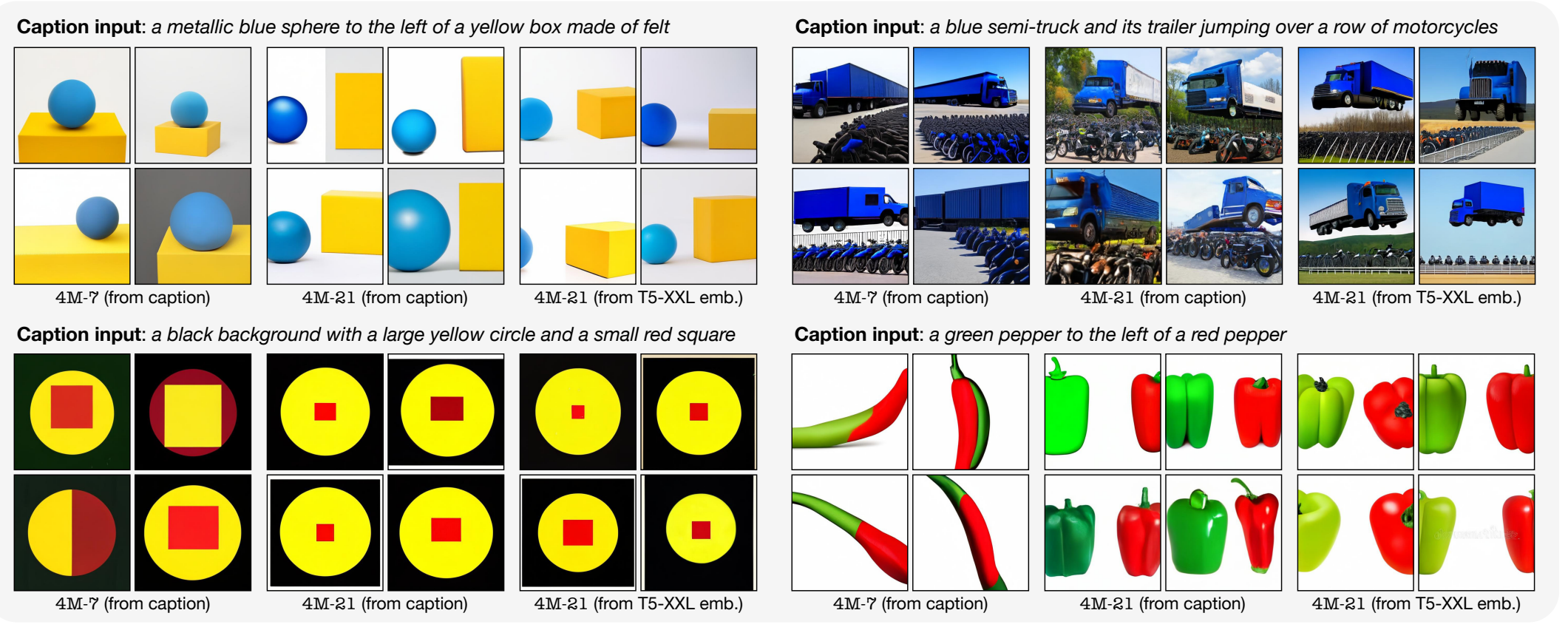


Multimodal generation & retrieval capabilities

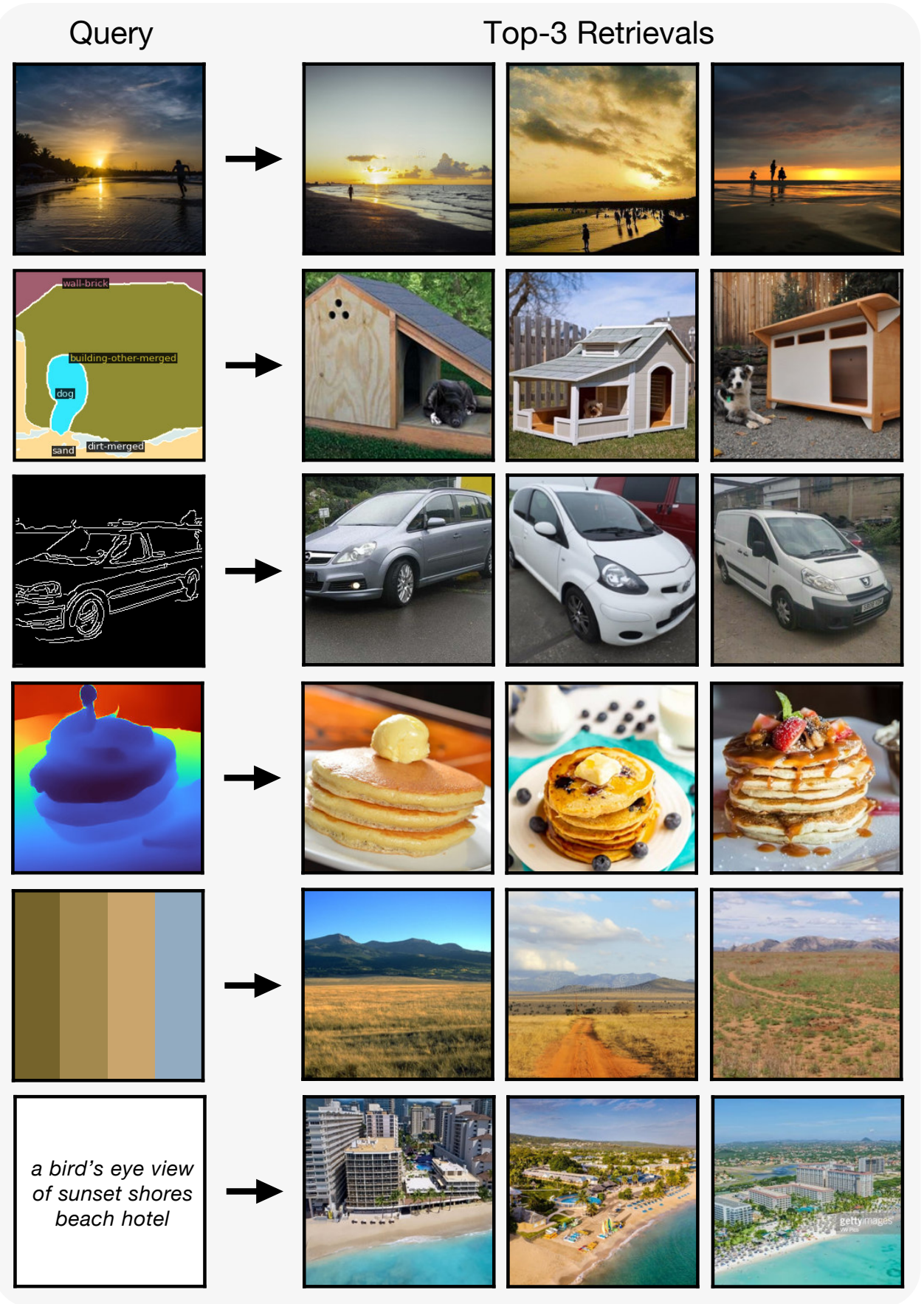
Fine-grained multimodal conditioning control



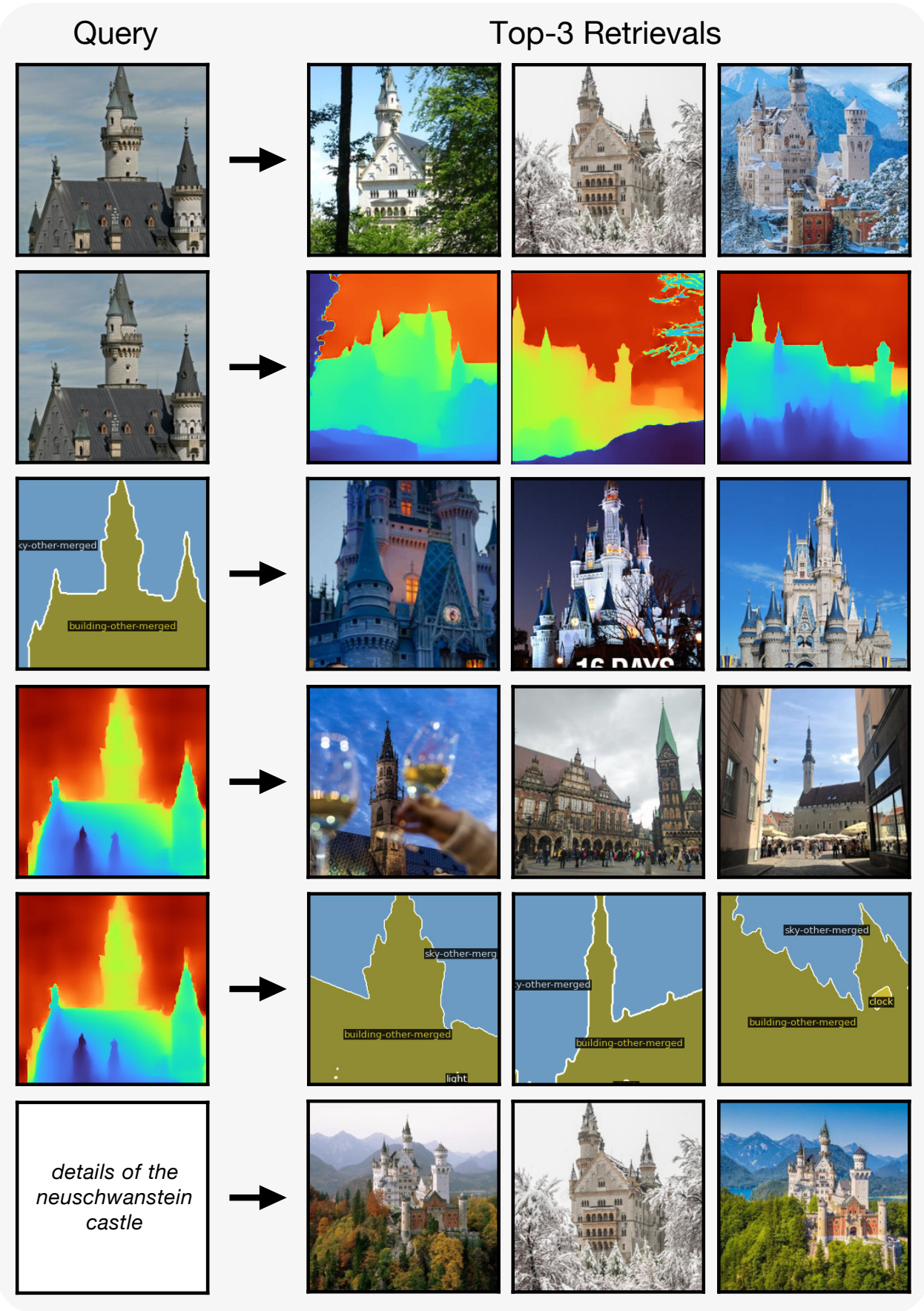
Improved text understanding capabilities



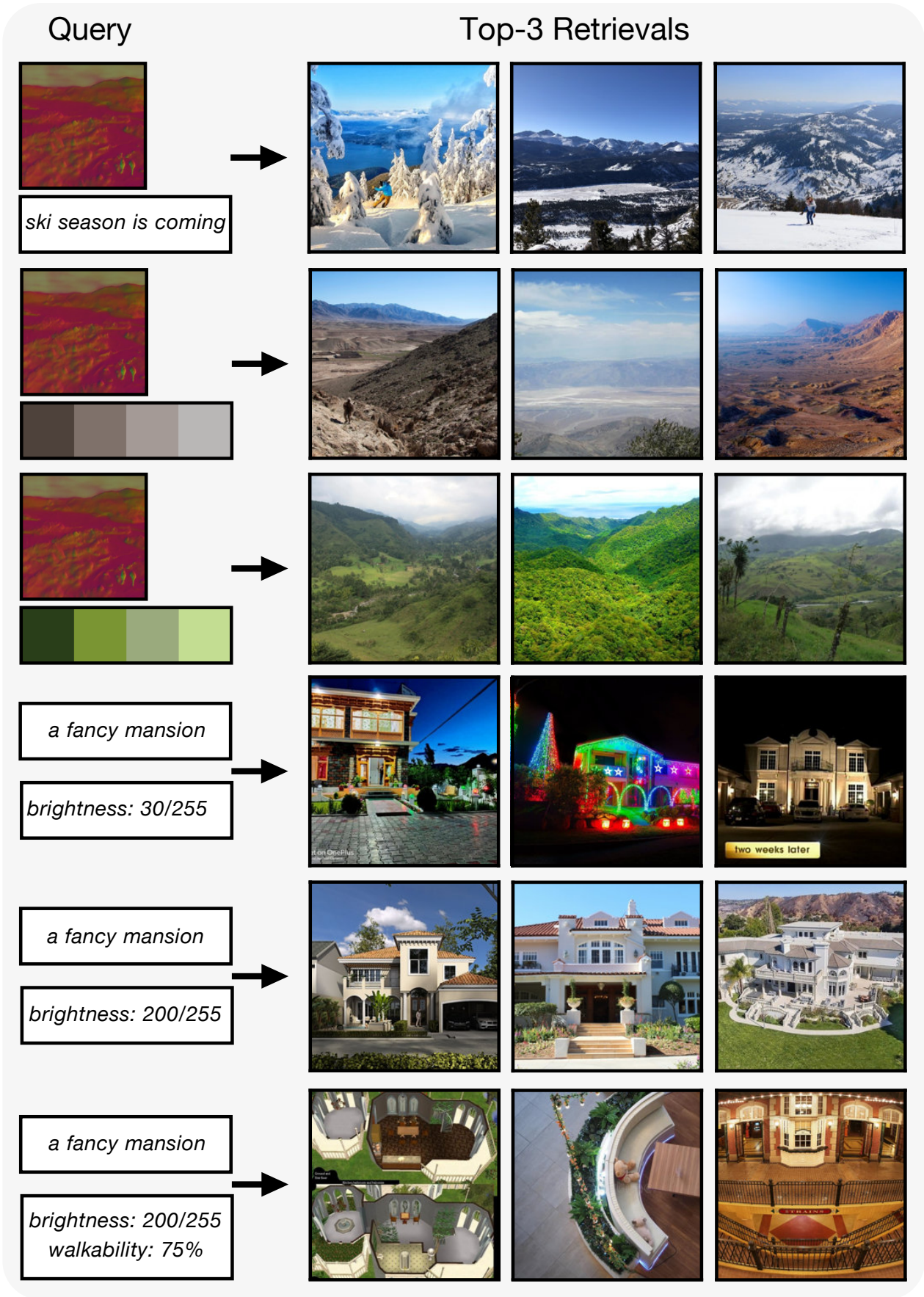
Any-to-RGB retrieval



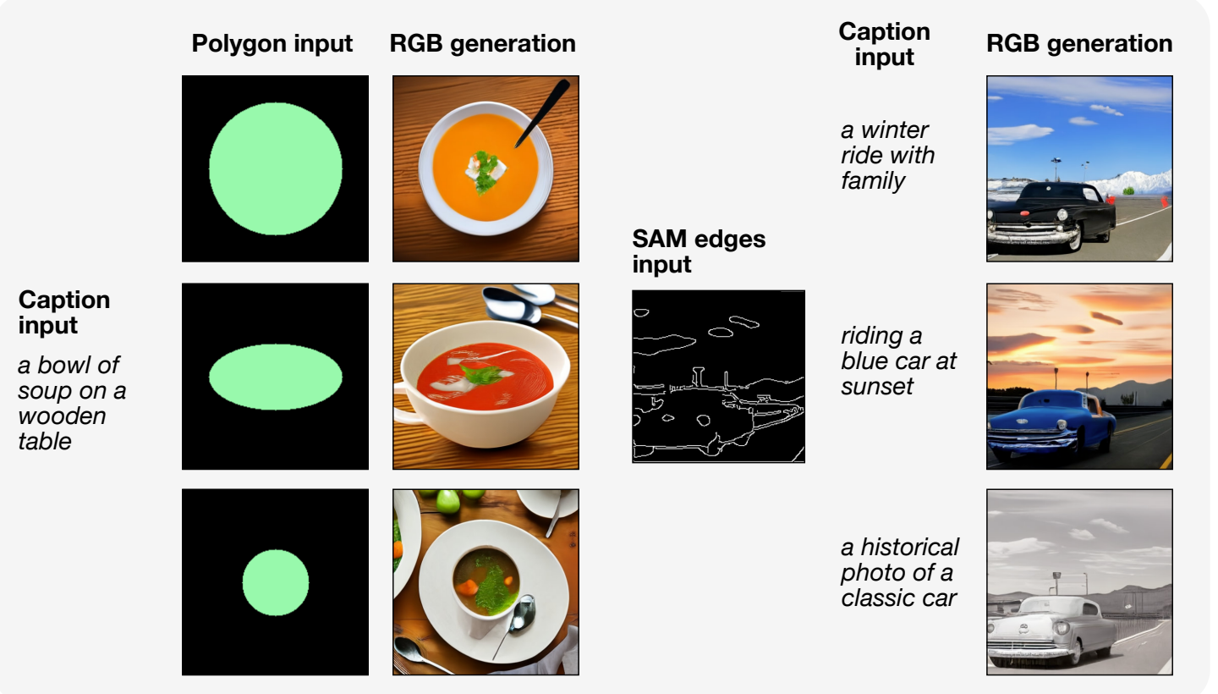
Any-to-any retrieval



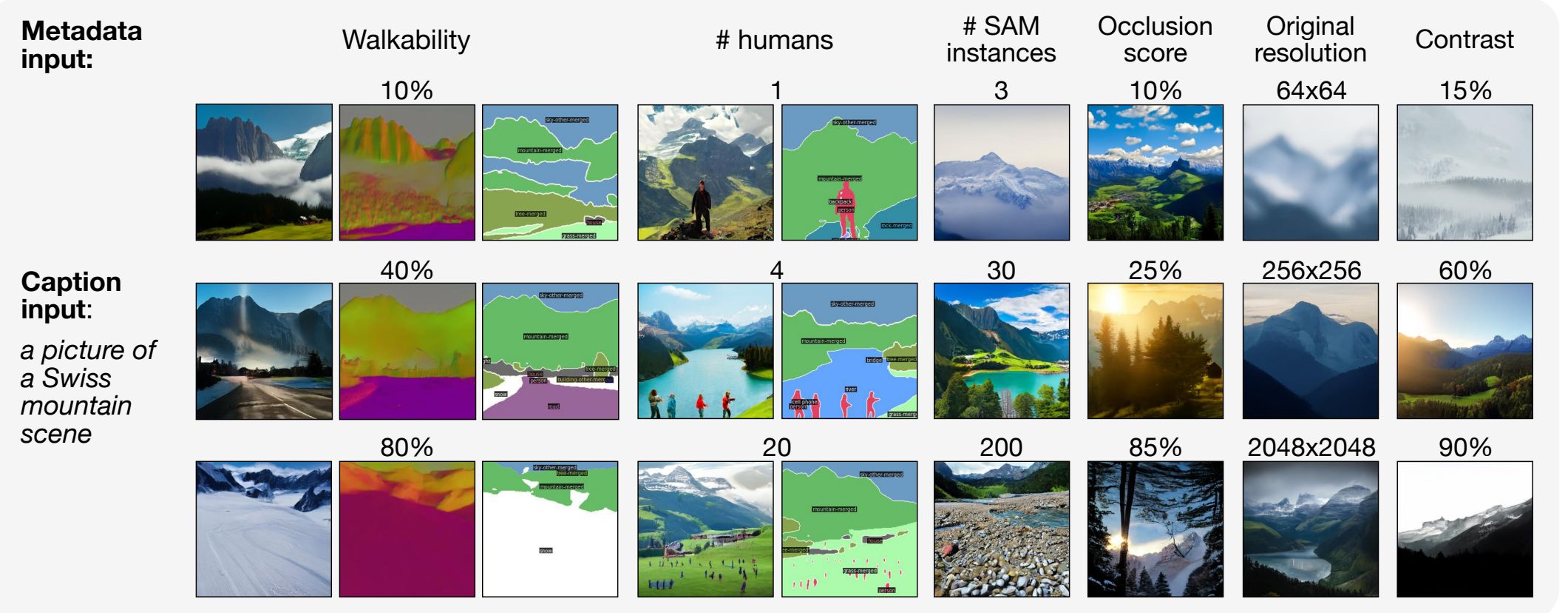
Multimodal retrieval



Probing with grounded generation


Polygon input
RGB generation
Caption input
SAM edges input
Caption input
a bowl of soup on a wooden table
a winter ride with family
riding a blue car at sunset
a historical photo of a classic car

Steerable multimodal data generation from metadata



- 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