A generalist vision model that can...

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

RGB → Surface normals | Depth | Semantic segmentation | Object detection | Captioning

dog looking out the window
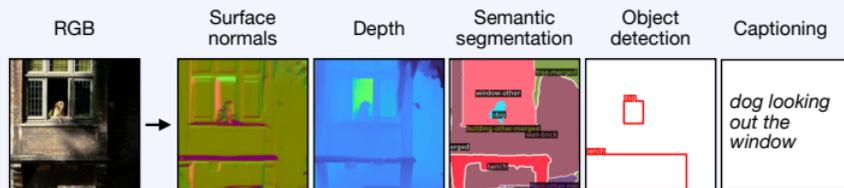
... be easily fine-tuned into specialist variants

Super-resolution specialist 4M

Caption: *impressionist painting of an astronaut in a jungle* → Text-to-image specialist 4M
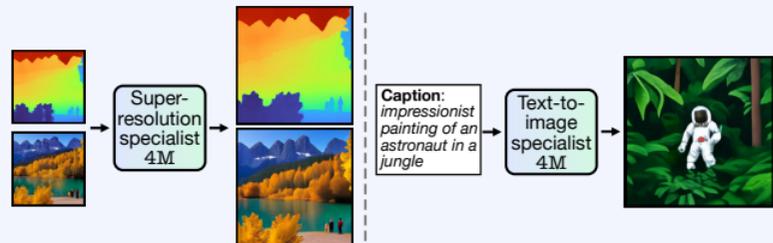
... transfer well to unseen tasks and modalities

Multimodal transfers, e.g. RGB + depth → semantic

Transfers to unseen modalities/tasks, e.g. 2D edges → 3D curvature
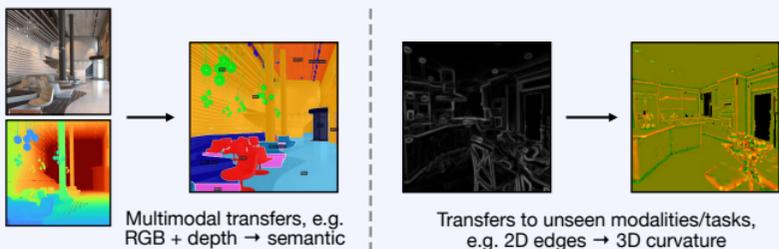
A multimodal generative model that can...

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

Caption: *resort & spa hotel in napa*

Bounding boxes & caption → RGB

Segmentation → depth & normals

Normals / CLIP → RGB
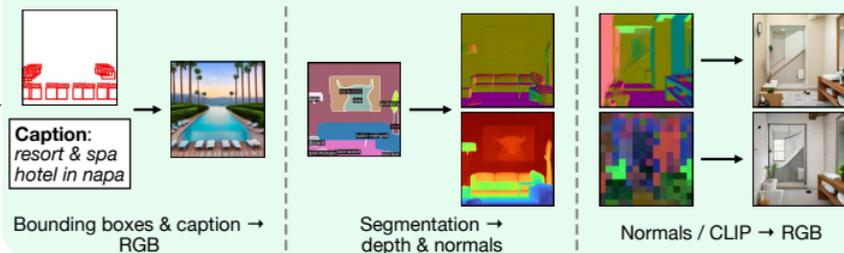
... with varying conditioning weights and from partial inputs ...

Caption: *an oil painting of a blue flower*

Caption (**strong**) & depth (weak) → RGB

Caption: *a fast car*

Caption & *masked* normals → Normals
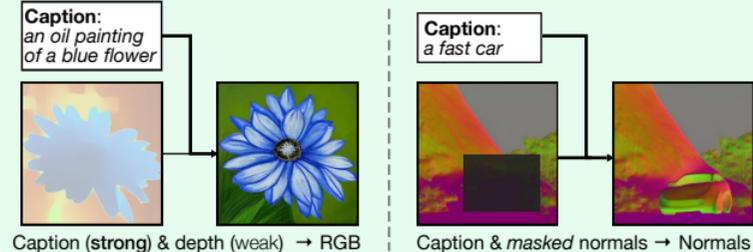
... enabling precise user control through multimodal editing chains

Original image

Edit RGB and add caption

Caption: *a small conference room with two sofas and a light blue wall*

Edited image

Predict depth and semantics

In-paint RGB conditioned on depth, segmentation, and caption

4M