Vision Banana

Image Generators are Generalist Vision Learners

Songyou Peng

On behalf of the Vision Banana team

Google DeepMind

Team


Valentin Gabeur*


Shangbang Long*


Songyou Peng*


Paul Voigtlaender


Shuyang Sun


Yanan Bao


Karen Truong


Zhicheng Wang


Wenlei Zhou


Jon Barron


Kyle Genova

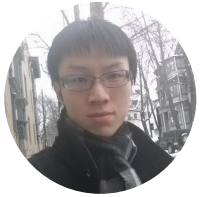

Nithish Kannen


Sherry Ben


Yandong Li


Mandy Guo


Suhas Yogin


Yiming Gu


Huizhong Chen


Oliver Wang


Saining Xie


Howard Zhou

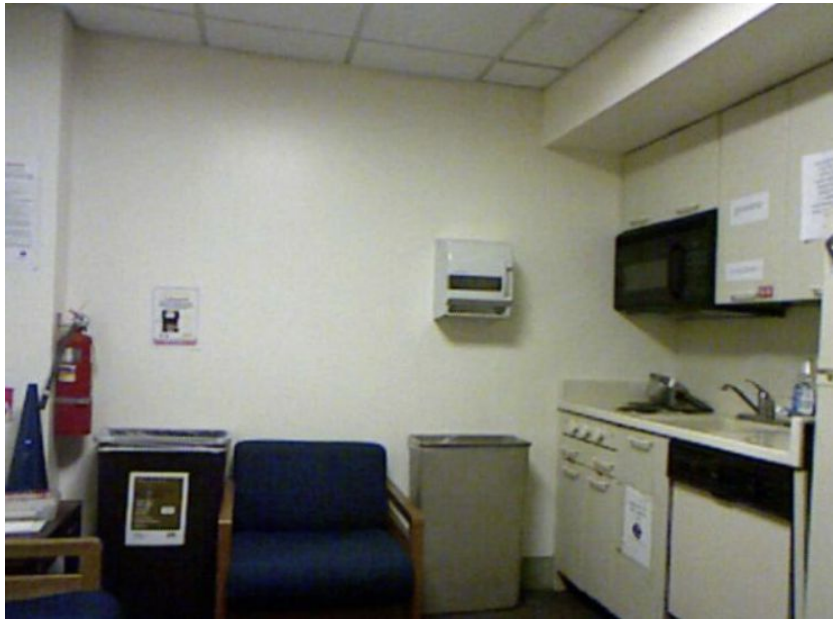

Kaiming He

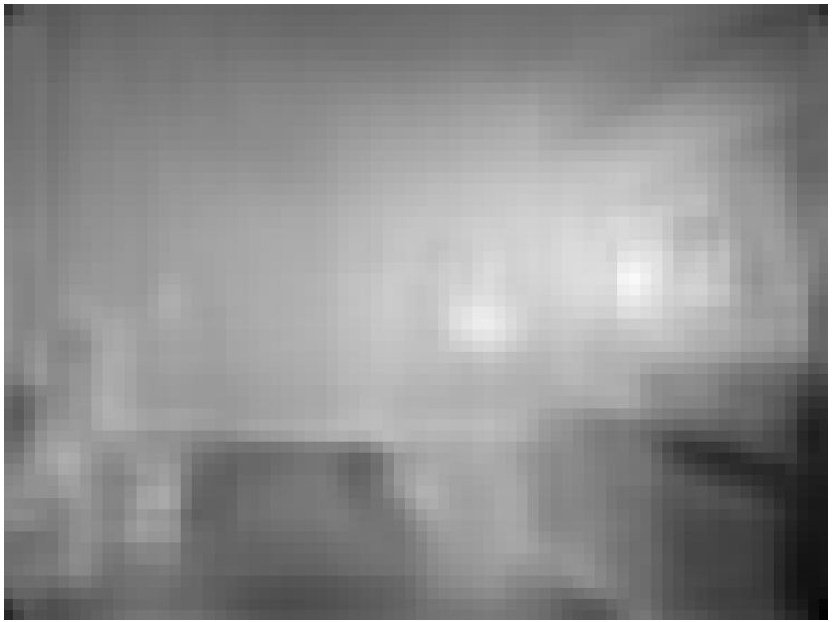

Tom Funkhouser


Jean-Baptiste Alayrac


Radu Soricut

Computer Vision Has Come a Long Way

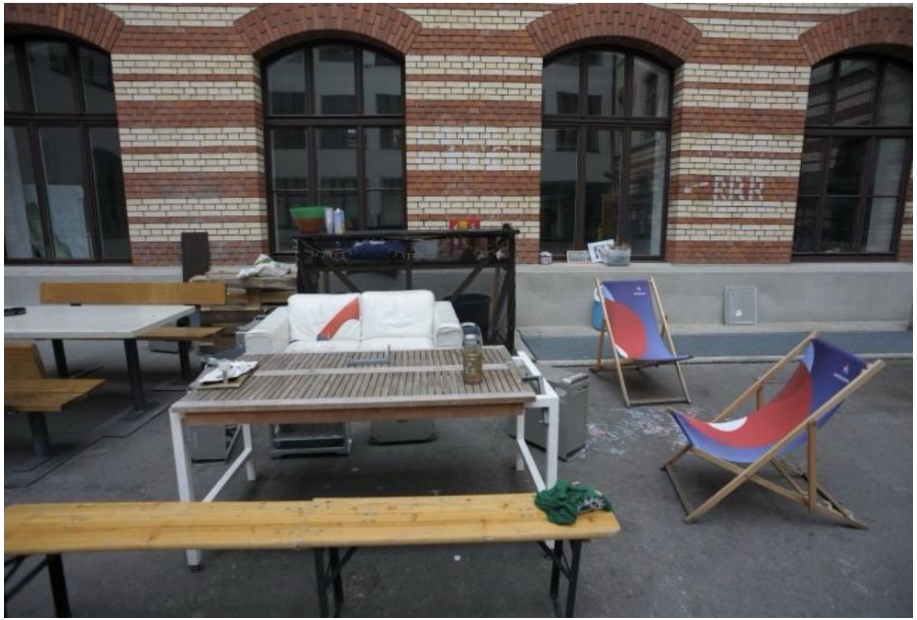




Depth Prediction in 2014

[Eigen and Fergus, 2014]

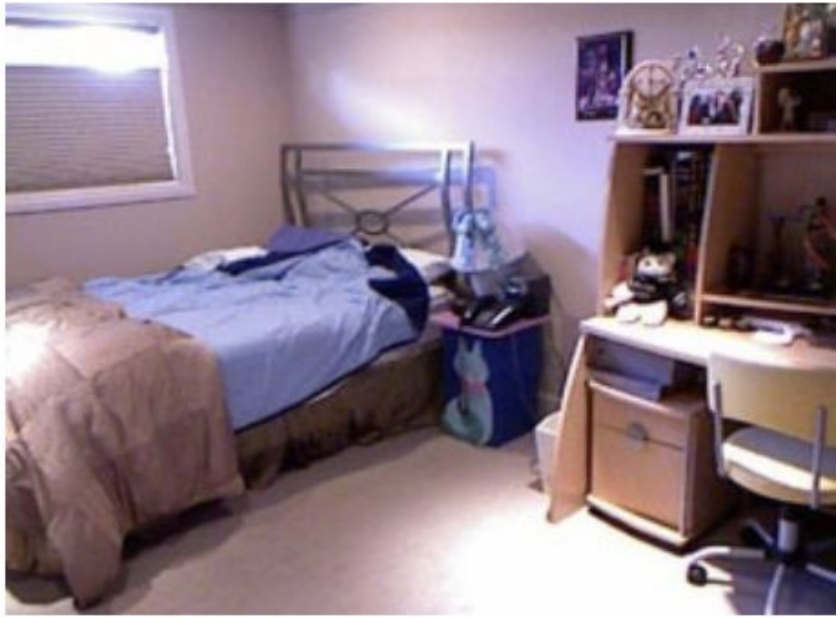
Computer Vision Has Come a Long Way

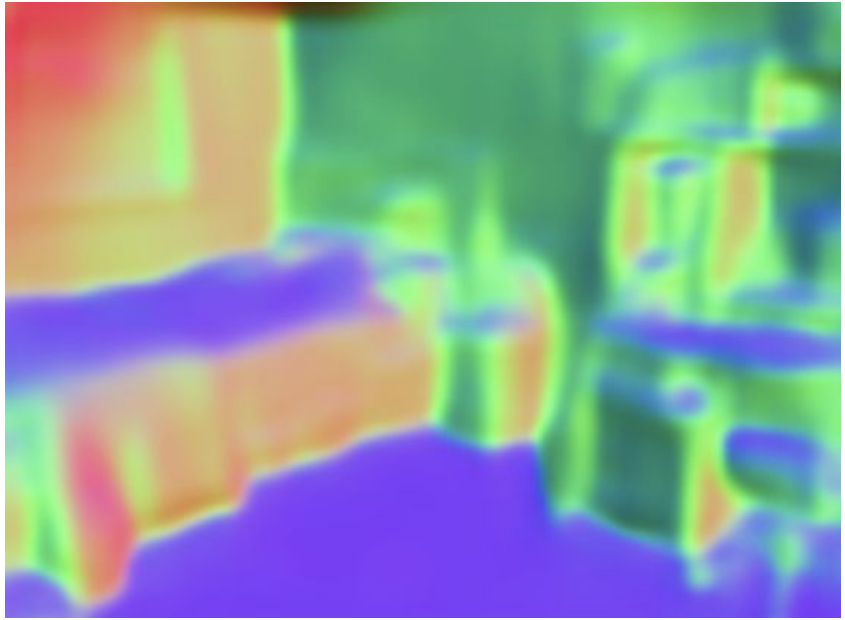


Depth Prediction in 2025

[Depth Anything 3, 2025]

Computer Vision Has Come a Long Way

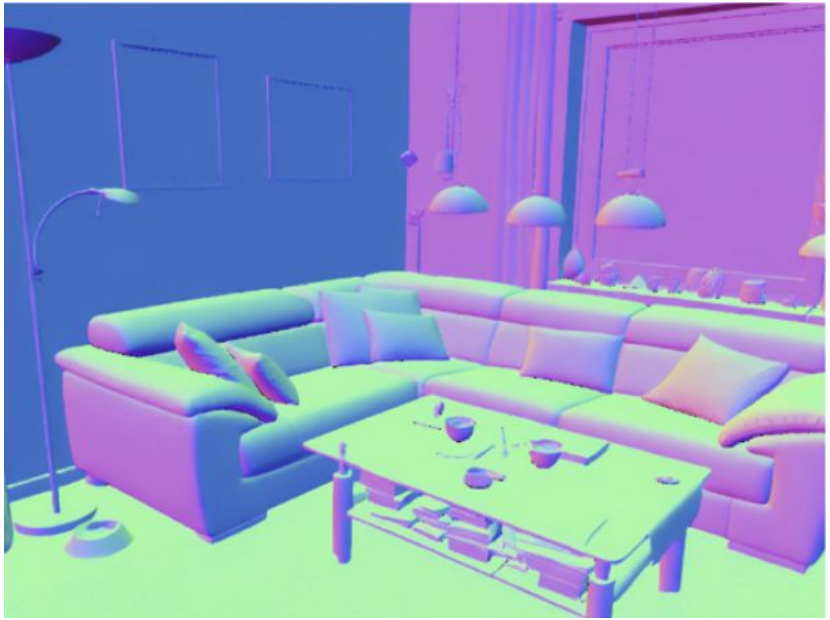




Surface Normal Estimation in 2014

[Eigen and Fergus, 2014]

Computer Vision Has Come a Long Way



Surface Normal Estimation in 2025

[Lotus-2, 2025]

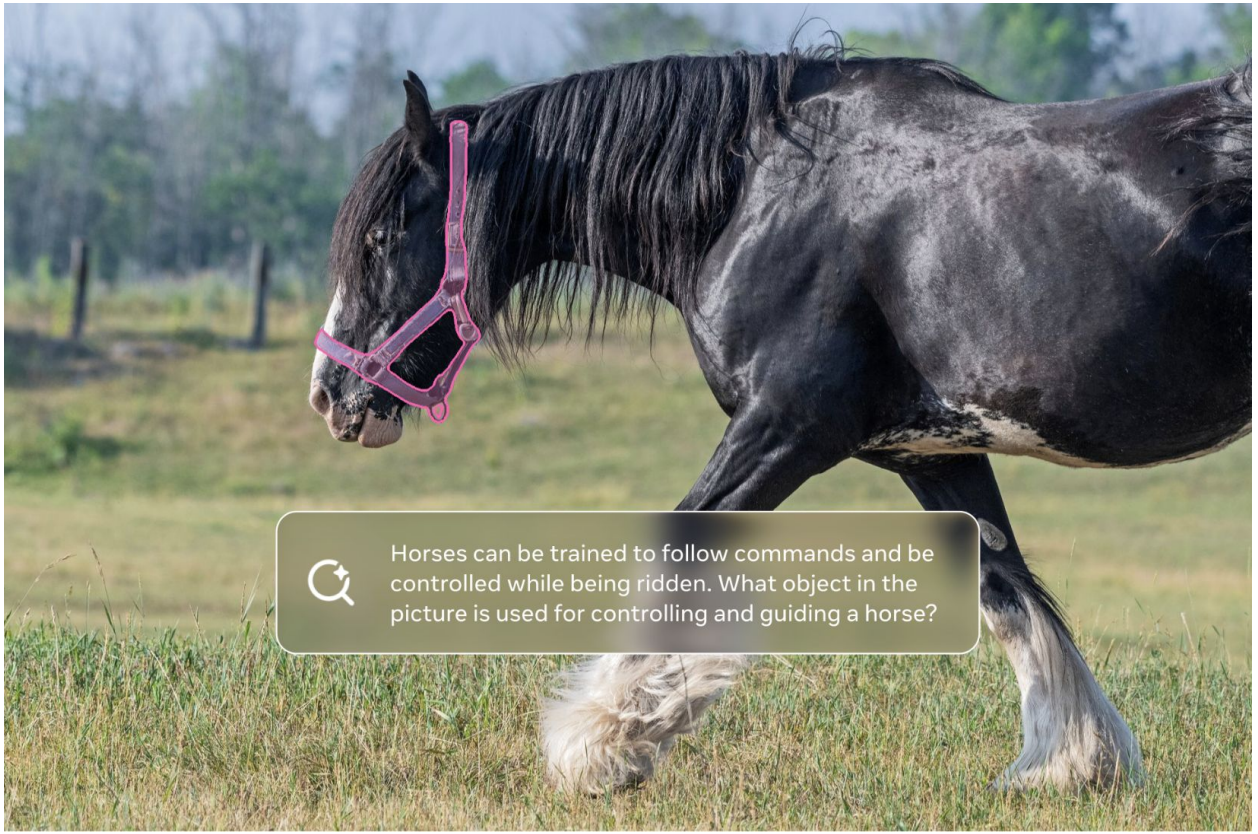
Computer Vision Has Come a Long Way

Input

Semantic Segmentation in **2014**

[Long et al., 2014]

Computer Vision Has Come a Long Way



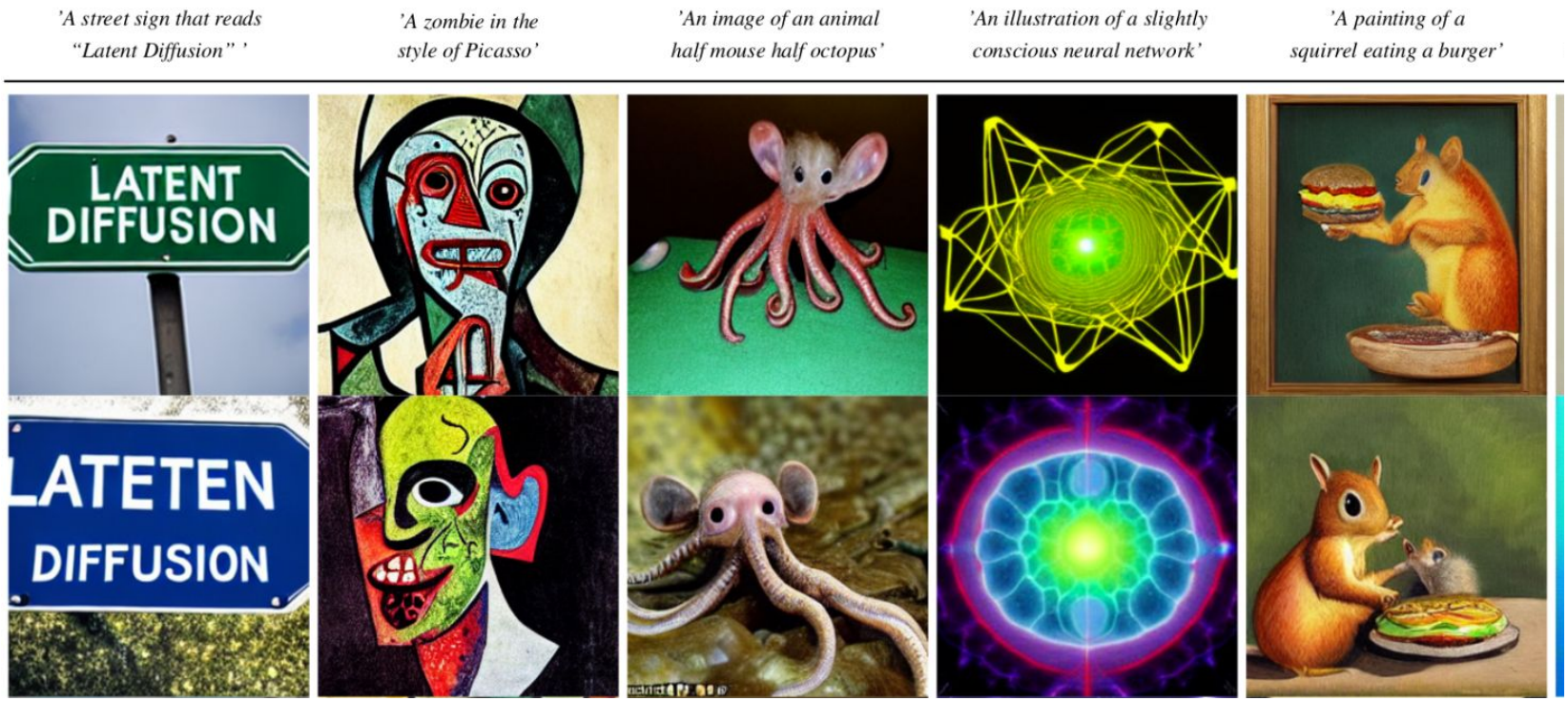
Open-Vocabulary Segmentation in 2025

[SAM 3., 2025]

They are awesome, but all “Specialists”

On the other hand...

Image Generation Has Come a Long Way



Latent Diffusion [Rombach, et al, 2021]

Image Generation Has Come a Long Way

Nano Banana [Google DeepMind, 2025]

Image generators are “Generalist”!

They can generate anything from a text prompt

Moreover...

Image generators already understand vision!

SF background?

Finishing line?

No trash on hand?

@ SF Half Marathon, July, 2025

Editing w/ Nano Banana 1

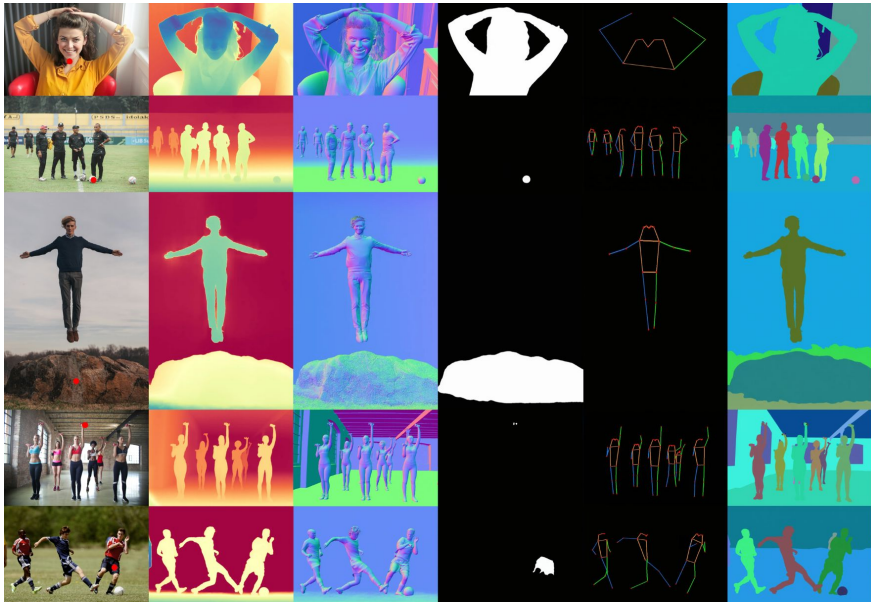
Image Generators For Vision

Related Work

Specialist Fine-Tuning

*[**Marigold**, 2024; **StableNormal**, 2024; **GeoWizard**, 2024 ...]*

Multi-Task Generalization



*[**DICEPTION**, 2025; **GenPercept**, 2024 ...]*

What Is Vision Banana?

Image Generator as Generalist Vision Learner

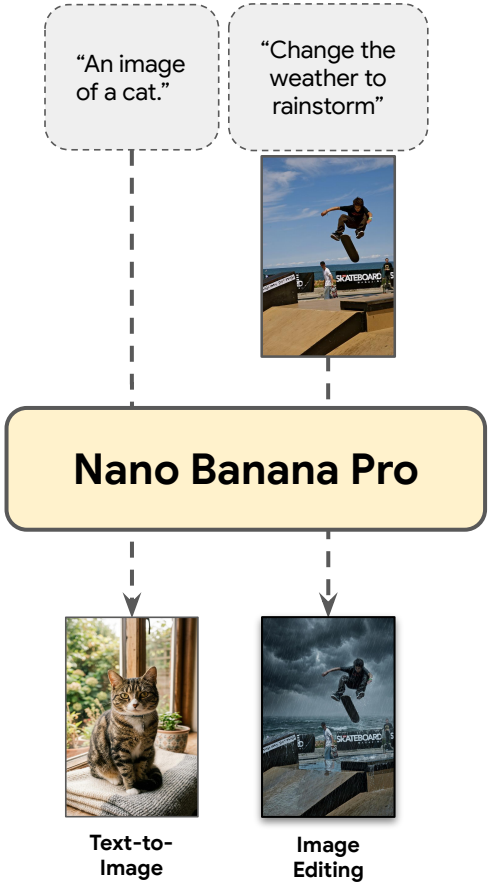


Image Generator as Generalist Vision Learner

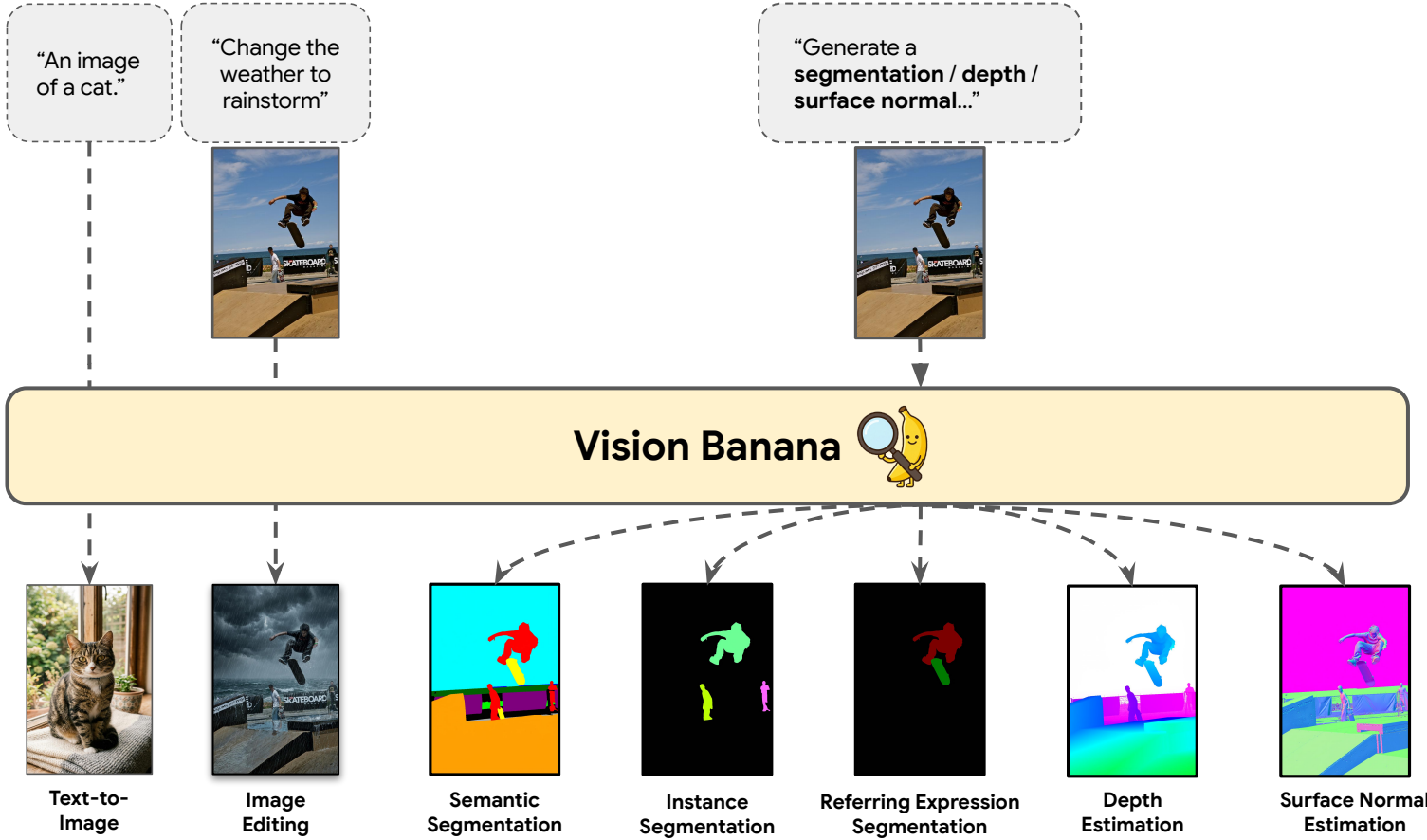
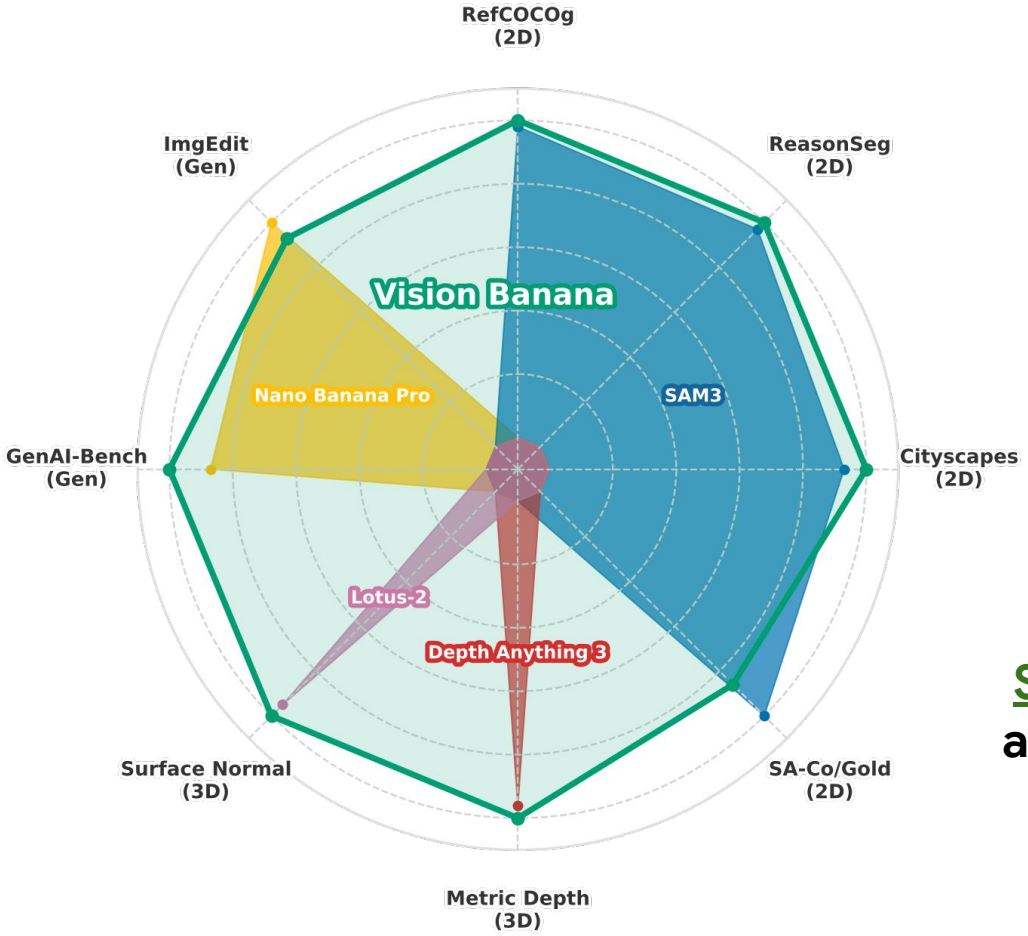
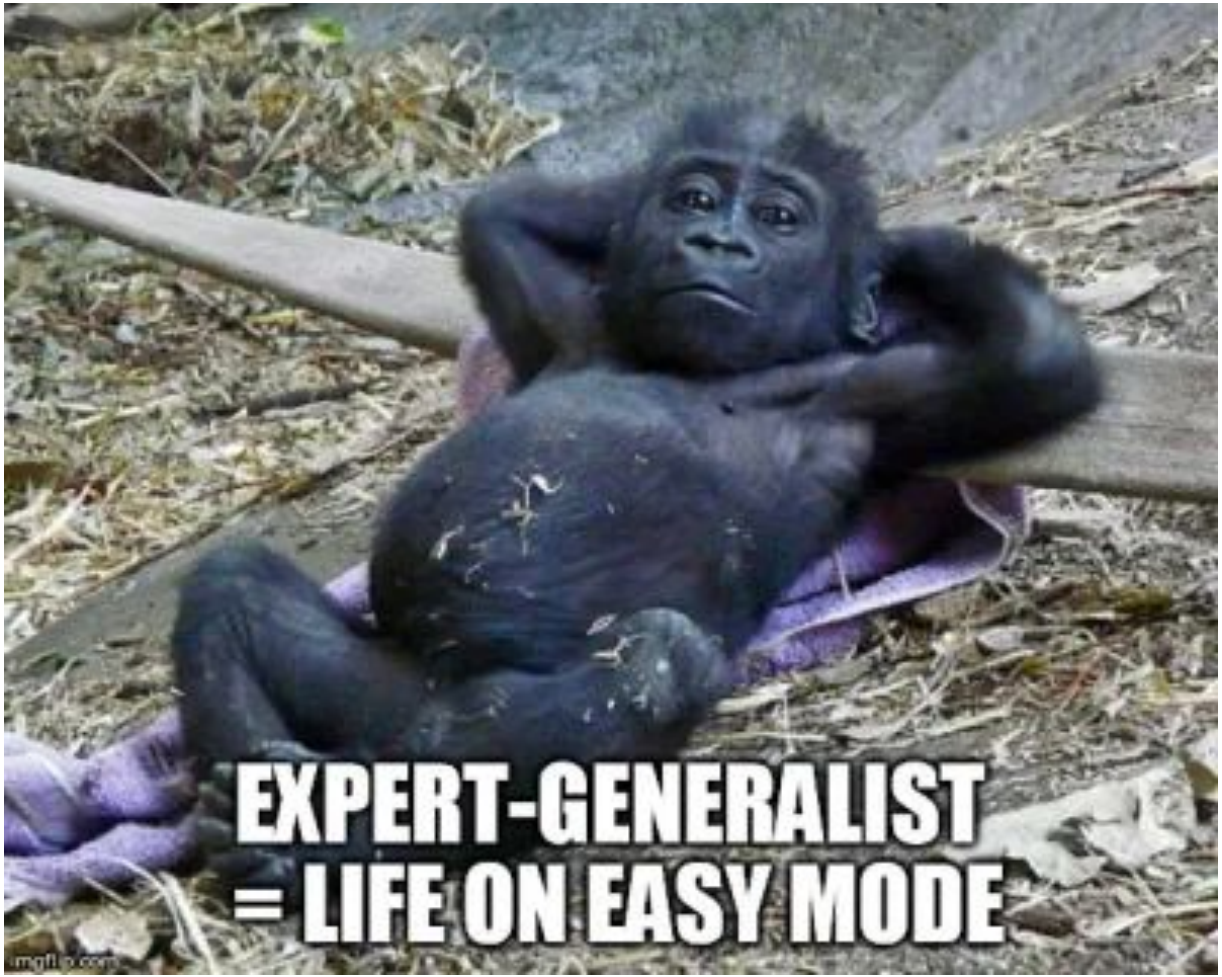


Image Generator as Generalist Vision Learner



State-of-the-Art
across tasks from a single model!


EXPERT-GENERALIST
= LIFE ON EASY MODE

How Did We Do it?

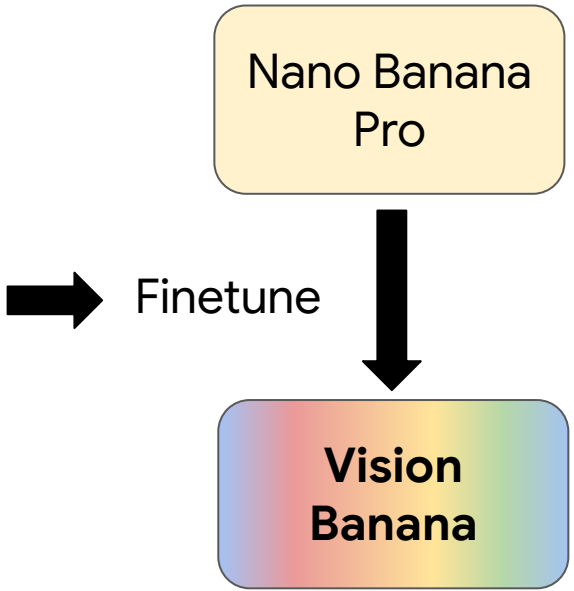
Method

1. **Generative Pre-trained Prior:** Nano Banana Pro as the base model

2. **Instruction Tuning**: Finetune Nano Banana Pro with a Data Mixture

Sample Datasets	Type	Tasks
Internal NB Pro Datasets	-	Text to image, etc
HyperSim	Synth 🤖	Depth, Normal, Edges, Semantic segmentation, Instance segmentation, Any-to-Any modality
Tartan Air	Synth 🤖	Depth
Internal Vision Datasets	Synth 🤖 & Real 🎥	Depth, Normal, Semantic segmentation, Instance segmentation...

Data Mixtures



Every Task is Treated as an Image Generation Task

Method

All tasks are formulated as image generation, so we can use the same training objective for all of them

Most tasks can be represented as RGB naturally

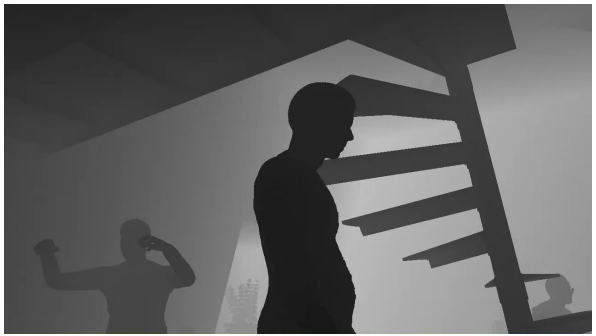
Surface Normal

Segmentation Map

Method

All tasks are formulated as image generation, so we can use the same training objective for all of them

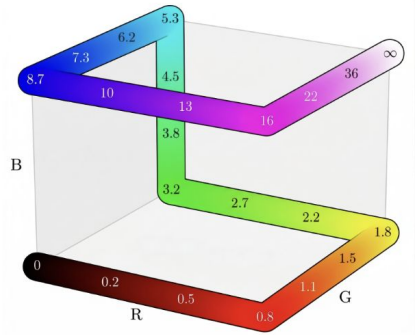
Metric depth map is single channel, unbounded



An ordinary grayscale depth map

1. Power transform [Barron, 2025] to map metric distances from [0, ∞) to [0, 1)

2. Apply a bijective colormap to the RGB space



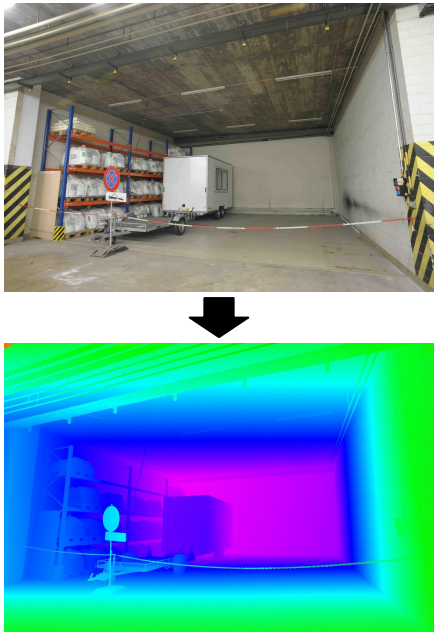
Colored depth map

Method

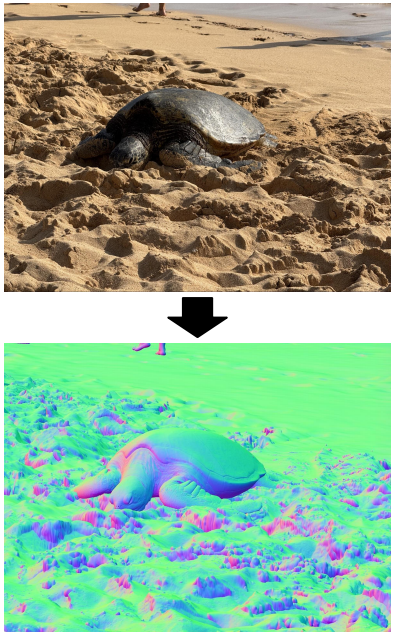
All tasks are driven purely by text prompts

Prompt: `Segment the image with humans in (255, 0, 0), trees in (0, 255, 0)...`

Prompt: `generate a metric depth map for the image`



Prompt: `Predict the surface normal for the input image`

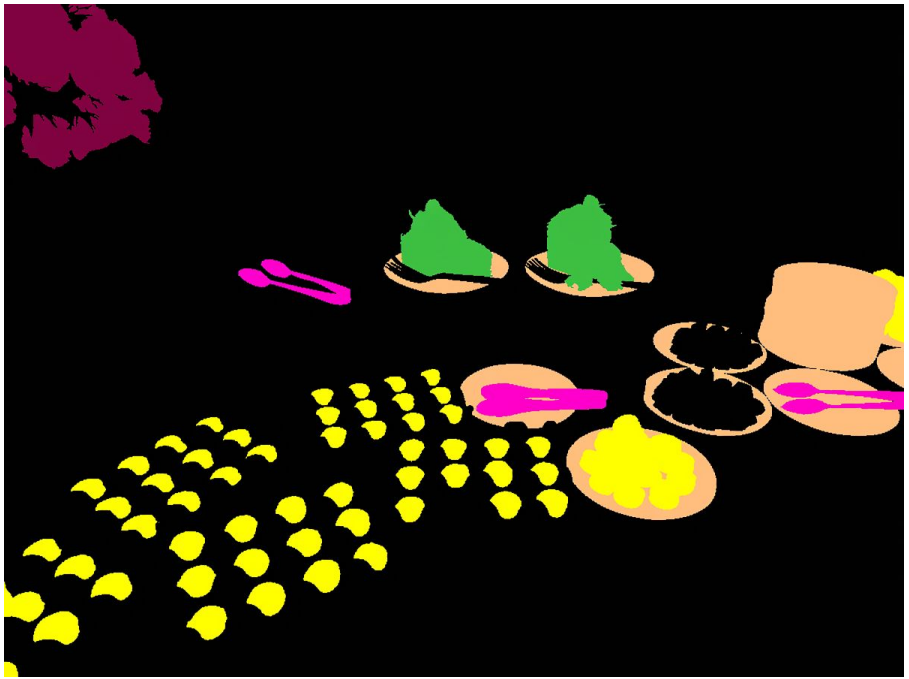


Results

Semantic Segmentation

“This image is a per-pixel class labeling of the input. The macaron cakes are represented by (255, 255, 0). The round plates are represented by (255, 192, 128). The slice cakes are depicted in (64, 192, 64). The flowers are shown in (128, 0, 64). The tongs are (255, 0, 192).”

Semantic Segmentation

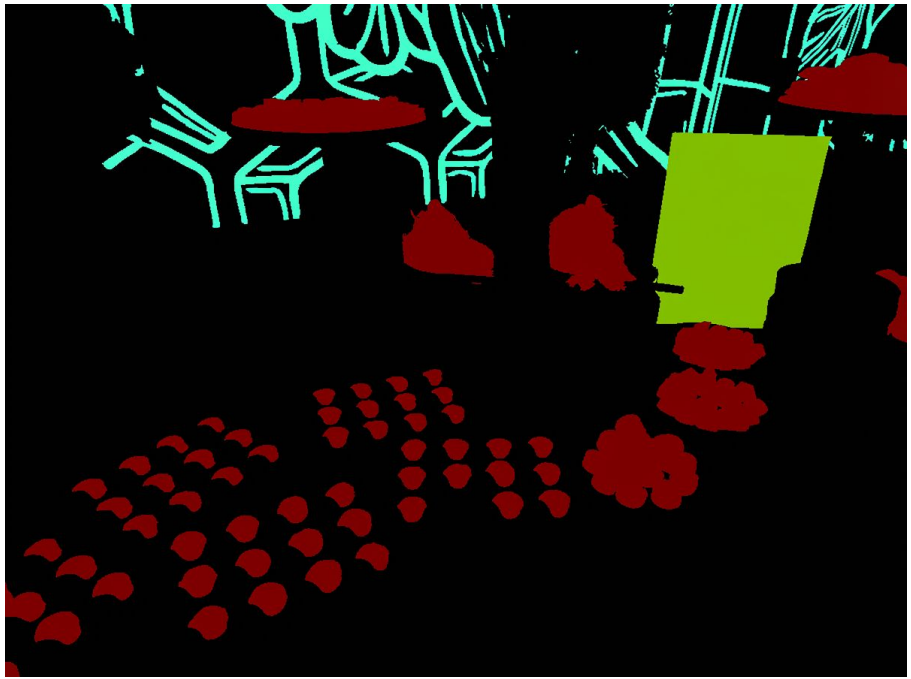


“This image is a per-pixel class labeling of the input. The macaron cakes are represented by (255, 255, 0). The round plates are represented by (255, 192, 128). The slice cakes are depicted in (64, 192, 64). The flowers are shown in (128, 0, 64). The tongs are (255, 0, 192).”

Semantic Segmentation

“Generate a semantic segmentation visualization of the input. The menu is #80C000. The dessert is #800000. The patterns on the wall is #40FFC0””

Semantic Segmentation



“Generate a semantic segmentation visualization of the input. The menu is #80C000. The dessert is #800000. The patterns on the wall is #40FFC0””

Instance Segmentation

Generate an instance segmentation visualization of this image. **Each piece of garlic is colored differently.**

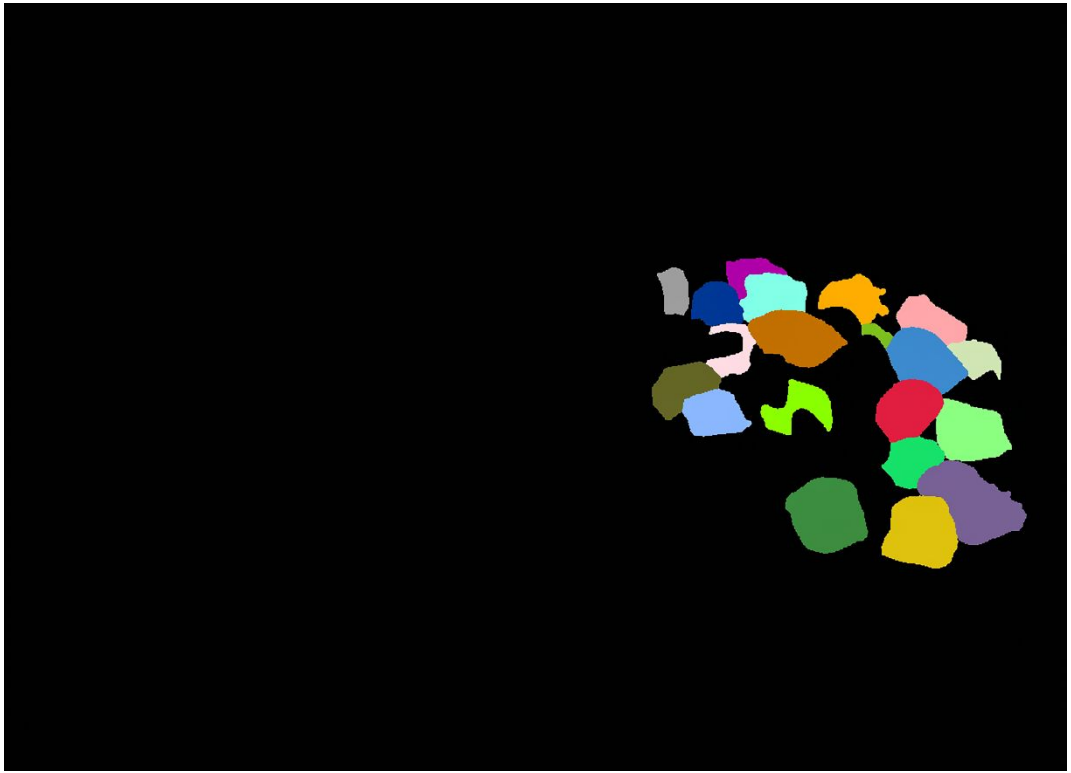
Instance Segmentation

Generate an instance segmentation visualization of this image. **Each piece of garlic is colored differently**.

Instance Segmentation

Generate an instance segmentation visualization of this image. **Each piece of beef is colored differently.**

Instance Segmentation



Generate an instance segmentation visualization of this image. **Each piece of beef is colored differently.**

Referring Expression

…… Segment the chef’s name in both Chinese and English as cyan color.

Depth Prediction

Depth Prediction

Super-Human **Metric** Depth Prediction

An image taken in Monterey, CA

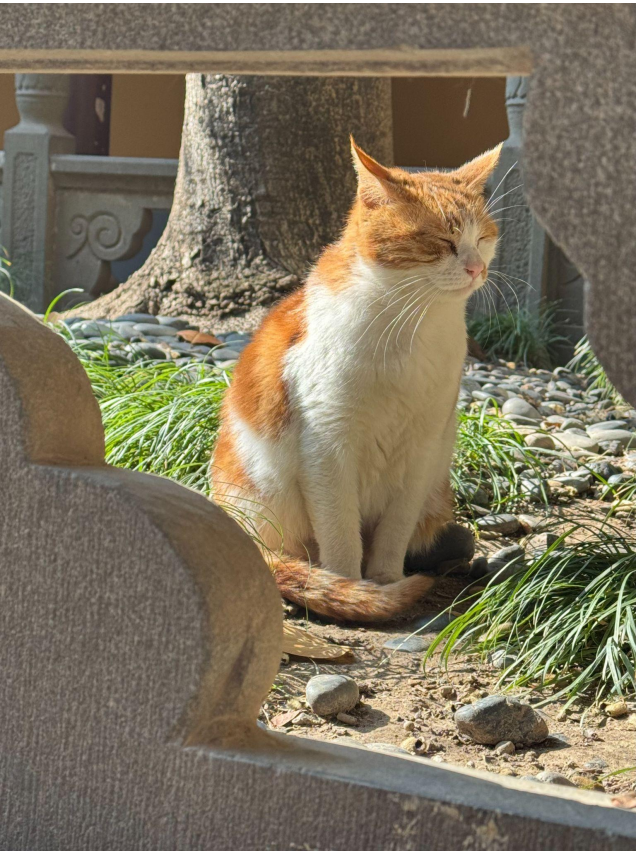
Vision Banana Predicted Depth

- Depth to the motorcycle billboard: **68m**
- Depth to the race car billboard: **123m**

Estimated distance between the 2 billboards: **55m**

Actual distance on Google Map: **57m**

Surface Normal Estimation

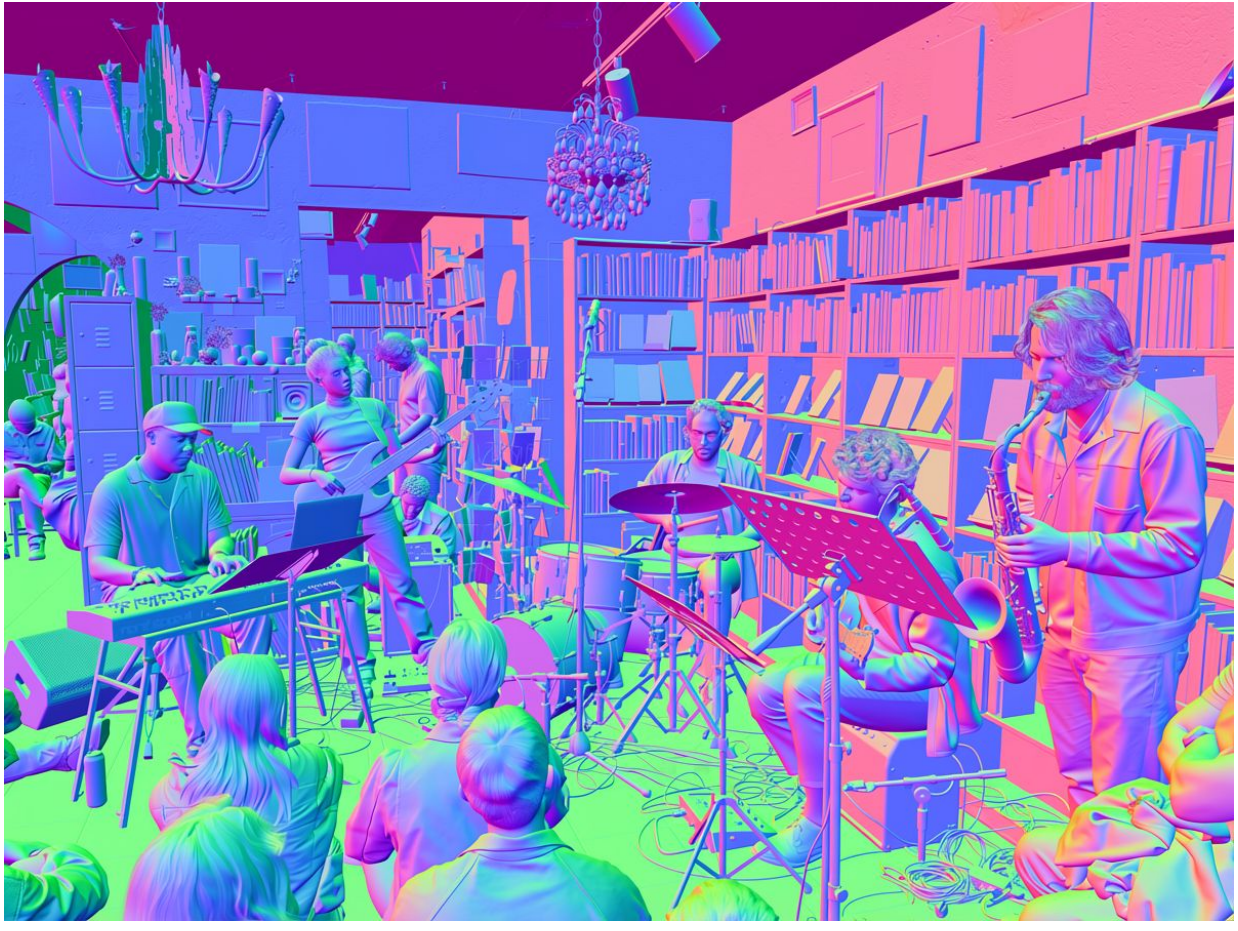


Vision Banana

Lotus-2 (external SOTA)

Surface Normal Estimation

Surface Normal Estimation



Vision Banana still Does Image Generation!

Text prompt: `A lantern casting dim light in a haunted forest.`

T2I

Vision Banana

Nano Banana Pro

Vision Banana still Does Image Generation!

Text prompt: `Change the vehicle's color to red`

I2I

Input

Vision Banana

Nano Banana Pro

Vision Banana: A Generalist Vision Model

Capabilities	Benchmarks and Metrics	Vision Banana	Best Counterpart
2D Understanding	Referring segmentation: RefCOCOg UMD val (cIoU ↑)	**73.8**	73.4 (*SAM3 Agent*)
	Referring segmentation: ReasonSeg val (gIoU ↑)	**79.3**	77.0 (*SAM3 Agent*)
	Semantic segmentation: Cityscapes val (mIoU ↑)	**69.9**	65.2 (*SAM3*)
	Instance segmentation: SA-Co/Gold (cgF_1 ↑)	**47.5**	24.6 (*OWLv2*)
3D Understanding	Metric depth estimation: average of 4 datasets (δ_1↑)	**0.929**	0.918 (*Depth Anything 3*)
	Surface normal estimation: average of 4 datasets (mean angle error ↓)	**18.928**	19.642 (*Lotus-2*)
Visual Generation	Text-to-image: GenAI-Bench (win rate against the other ↑)	**53.5%**	46.5% (*Nano Banana Pro*)
	Image editing: ImgEdit (win rate against the other ↑)	47.8%	**52.2%** (*Nano Banana Pro*)

Live Demo Time!

go/densepix_demo

Why Should We Care About Vision Banana?

Take A Look at the Brief History of LLM

GPT-1 / BERT

Improving Language Understanding by Generative Pre-Training

GPT-2 / T5

Language Models are Unsupervised Multitask Learners

GPT-3 / Instruct GPT

Language Models are Few-Shot Learners

Potential Paradigm Shift for Vision Community?

- Image generators already understand computer vision
- **Generative pretraining** + **instruction tuning** might be enough for CV tasks
- Vision banana unifies vision tasks through raw RGB pixels + text prompts
- Vision community should consider adapting powerful generative models rather than training new specificist models, because...

Image Generators are Generalist Vision Learners

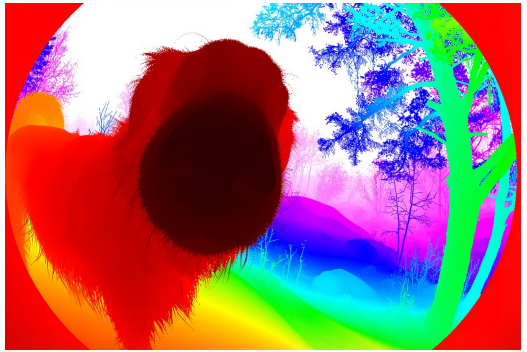
Next Steps?

Future Steps

- **Extending to Multi-View and Video Inputs**
- **Scaling Task Diversity**: increase variety of task to unlock emergent cross-task generation
- **Integration with LLMs**: Enhance cross-modality reasoning
- **Compute Efficiency**: Accelerate NB Pro, making it deployable

Demo @ Google Booth

June 5
11:00 - 11:30
Demo



June 5
16:00 - 17:00
Interactive Demo

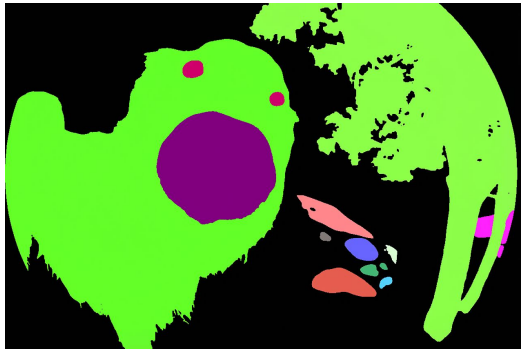


Image credit

Google

Vision Banana

Image Generators are Generalist Vision Learners

vision-banana.github.io

Songyou Peng

Reach out
vision-banana@google.com